Lin Wan, Fangyi Jiang, Chenmeng Ma, Haoyu Lu

*DePaul University | DSC 423 Data Analysis and Regression*

*BIKE SHARING IN WASHINGTON D.C.*

**Table of Contents**

[**Abstract:**](#_heading=h.2et92p0) 2

[**Introduction:**](#_heading=h.tyjcwt) 2

[**Methodology:**](#_heading=h.3dy6vkm) 2

Dataset description 2

Steps of approach 3

Dependent variable 3

Independent variables 3

[**Analysis:**](#_heading=h.3rdcrjn) 4

Data Pre-processing 4

Data Exploratory Stage 5

Data Analysis Stage 6

Model Validation 8

Model Selection 8

Fitted Final Model 10

Prediction 12

[**Results and Findings:**](#_heading=h.1ksv4uv) 13

[**Future work:**](#_heading=h.44sinio) 13

[**Research References:**](#_heading=h.2jxsxqh) 15

[**Appendix:**](#_heading=h.z337ya) 16

**Research on Bike-Sharing System in Washington D.C.**

**Abstract**

The bike-sharing system is a new generation of transportation nowadays. Many cities in the world have established bike-sharing systems for urban commuting. The purpose of this project is to analyze how different factors influence the bike-sharing system and to predict biking sharing system performance. The dataset we used for bike-sharing in Washington D.C. is a statistical dataset from Kaggle. Linear regression model is used to predict the total users of bike-sharing in Chicago in summer and winter. Based on our analysis, the three most important predictors are year, temperature and ds1(spring - Dec 21th to Mar 21th). From the year 2011 to 2012, the total bike-sharing users increased by 2067 in Washington D.C. When outside temperature increases by 1 °C, total bike-sharing users will increase 121. We illustrate the use of our analysis to identify the total number of users in different cities and different time periods.

## Introduction

Today, across the world, unusual traffic congestion and poor air quality threaten people's daily life as urban development growing rapidly. The increasing amount of bicycle usage for urban commuting can help relieve these problems (Kabra 2016). The whole process of bike sharing system is self-service. Users can rent and return bikes by themselves in any pick-up and drop-off locations. For short distance trips, people are more willing to use bikes to solve the last-mile problem (Guo 2017). In 2012, America had more than 22 public bikes-haring systems, and there were about 884442 users and 7549 bikes. Canada had about four bike-sharing organizations, and there were approximately 197419 users and 6115 bikes. Mexico had two bike-sharing organizations, and there were more than 71611 users and 6115 bikes (Shaheen 2014). With a gathering user to use bike-sharing systems, bike sharing systems play an important role in many different aspects such as traffic, environmental and health issues.

**Methodology**

**Dataset description**

Our group found the Bike sharing in Washington D.C. datasets from Kaggle (<https://www.kaggle.com/marklvl/bike-sharing-dataset>), dataset (day.csv) contains 731 observations cross 2 years, from 2011 to 2012. There are 15 variables as follows: Dteday, Season, Yr, Mnth, Holiday, Weekday, Workingday, Weathersit, Temp, Atemp, Hum, Windspeed, casual, registered, cnt.

**Steps of approach**

a. Clean dataset:

Delete unrelated variables.

b. Import data

c. Create dummy variables:

Season, Weekday, Weathersit are created as dummy variables.

d. Data description:

Histogram, boxplots, scatterplots will be applied here to check symmetric, outliers, range of variables, and multicollinearity.

e. Create full model to check collinearity, assumptions, outliers and influential points

f. Split training set and testing set

g. Model selection

h. Create final model

i. Check residuals, outliers, influential points

j. Make prediction (gathering research data from future work)

**Dependent variable**

One dependent variable - cnt is chosen from the dataset:

Cnt represents count of daily total rental bikes, including both casual and registered users.

**Independent variables**

Nine independent variables are chosen from the dataset:

1. Season:

1 is assigned for spring; 2 is assigned for summer; 3 is assigned for fall; and 4 is assigned for winter.

1. Holiday (extracted from holiday schedule):

1 is assigned for holiday; 2 is assigned for not.

1. Weekday:

Assigned from 1 to 7 for Monday to Sunday.

1. Workingday:

If day is neither weekend nor holiday is 1; otherwise is 0.

1. Weathersit (weather situations extracted from Freemeteo):

1 is assigned for clear,few clouds, partly cloudy; 2 is assigned for mist and cloudy, mist and broke clouds; 3 is assigned for light snow and light rain, thunderstorm; 4 is assigned for heavy rain and ice pallets, thunderstorm and mist, snow and fog.

1. Temp:

Normalized temperature in Celsius.

1. Atemp:

Normalized feeling temperature in Celsius.

1. Hum:

Normalized humidity.

1. Windspeed:

Normalized wind speed.

From the original Bike-sharing dataset description:

The goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the ranges of values. Thus, to normalized temperature in Celsius, all values are derived via (t-t\_min)/( t\_max-t\_min), t\_min= - 8, t\_max= + 39 ; to normalized feeling temperature in Celsius, all values are derived via (t-t\_min)/(t\_max-t\_min), t\_min= - 16, t\_max=+50; to normalized humidity, all values are divided to 100 (max); to normalized wind speed, all values are divided to 67 (max).

**Analysis**

**Data Pre-processing**

1. **Data Cleaning**

Since we noticed that variables Dteday and Mnth are similar with Season, and the amount of cnt are the sum of casual and registered users for each observation, so we dropped the data of Dteday, Mnth, casual, and registered which would not affect the accuracy of the model. Moreover, there are no missing values and all variables are numeric.

1. **Import data**

We imported dataset in SAS using proc import statement.

1. **Create dummy variables**

Season, Weekday, Weathersit are created as dummy variables.

1. Season:

Winter from 9/23 to 12/20 is set up as our baseline (season winter = 0); ds1 is assigned for spring from 12/21 to 3/20 (season = 1); ds2 is assigned for summer from 3/21 to 6/20 (season = 2); ds3 is assigned for fall from 6/21 to 9/22 (season = 3).

1. Weekday:

Sunday is set up as our baseline (sunday = 0); dw1 is assigned for Monday (weekday = 1); dw2 is assigned for Tuesday (weekday = 2); dw3 is assigned for Wednesday (weekday = 3); dw4 is assigned for Thursday (weekday = 4); dw5 is assigned for Friday (weekday = 5); dw6 is assigned for Saturday (weekday = 6).

1. Weather situation:

Weather situation 4 which is heavy rain and ice pallets, thunderstorm and mist, snow and fog is set up as our baseline (weathersit = 0); dwe1 is assigned for weather situation 1 which is 1 clear,few clouds, partly cloudy (weathersit = 1); dwe2 is assigned for weather situation 2 which is mist and cloudy, mist and broke clouds (weathersit = 2); dwe3 is assigned for weather situation 3 which is light snow and light rain, thunderstorm (weathersit = 3).

***Data Exploratory Stage***

1. Histogram (A-3):

The distribution of cnt (total bike-sharing users) would be symmetric so that we do not need to use transformation. The median and mean are very close to each other. We can see there is no outlier from the histogram graph. We use the mean value 4504 later in our report to analyze the percentage of the total population who used the bike-sharing system in Washington D.C. We can find out from the graph that there is no outlier.

The range of histogram is 8692, which is a wild range and we should pay some attention to analyze what caused this large difference between maximum and minimum. The bars tend to increase and pile up around 4500 and then tend to decrease gradually. From descriptive statistics, we noticed that the mid-fifty range is 3141 to 5976, and the interval of the histogram from around 3600 to 5400 has the highest bars, showing that half of daily total users are in this range. Furthermore, combined with the normal distribution I created above, I noticed that 68.3% of total users are in range 2567 to 6441 and 95.45% of total users are in range 630 to 8379.

1. Scatter Plot & correlation table (A-4, A-5, A-6):

We created the scatterplots and correlation table to evaluate the relationships among different variables. However, most of our predictors are dummy variables with only values 0 and 1, and it is meaningless to look at these scatterplots(A-4). So, we then ran the scatter plots and correlation table for other numeric values. From scatterplots A-5, we can see that there is a quite high positive linear relationship between total rental bikes and temperature/ atemp (feeling temperature) -- the line has an upward trend and the correlation coefficient parameters are both higher than 0.62. Whereas there is a quite low negative linear relationship between total rentals and humidity/windspeed -- even though the trends of lines are not clear from the scatter plots, but the correlation coefficient parameters are -0.23 and -0.1, respectively. In addition, temperature and feeling temperature has a very high correlation value of 0.99, so we flag them as nearly perfect collinearity. As one of the biggest issues in the regression model, the way we solve this problem is to create a new interaction variable. We add an interaction variable of atemp and windspeed later in the analysis.

1. Boxplots (A-7, A-8, A-9):

We also created boxplots to see the distribution of total rental bikes by year, season and workingday. From the boxplot of total users by year (A-7), it is obvious to see that year 2 has way more total rentals than the first year. The maximum, average, median and mid-fifty range of users are all higher than the first year. It indicates that there was a significant development in the biking sharing system from year 2011 to 2012, or maybe the advertisements or promotion/ discount policies attracted more people, and the users of sharing bikes were almost double.

From the boxplot of total users by season (A-8), according to the bars of mid-fifty range, the bar of 1, which starts from December 21st to March 20th is the lowest. 2 and 3 starts from March 21st to Sep 22nd have similar and the highest bars, and then the value decreases in bar 4, but it is still higher than 1. We believe that this makes a lot of sense because the season from late December to March 20th in Washington D.C. is the coldest time during a year, therefore, few people would use bikes. In contrast, more people were willing to rent a bike during the summer and fall season when the temperature increases instead of other public transportation methods, such as the subway or bus, and this made the boxplot of summer and fall to be the highest. And the distribution of total users has almost followed the trend of temperature. As a result, we assumed that temperature should be a very important predictor for total users.

The distribution of total users by workingday (A-9) does not have a large difference between two years. However, the mid-fifty range is shorter in the working day than the weekend/holiday. We supposed that the reason would be: working days have a fixed group of users, such as employees and students who will use bikes normally to go to work/school. So, the total number of users will not change a lot or have a large range. However, weekends or holidays are very different compared to working days. Some unfixed groups will join and use bikes such as visitors, or someone who usually drive/ take buses instead of bikes.

***Data Analysis Stage***

Get full model to check collinearity, assumptions, outliers and influential points.

1. Full regression model with all predictors (A-10):

Our full model has an adj-R2 of 0.8234. And we checked the goodness of fit, as F-value is 201 with p-value less than 0.05, showing that we can reject the null hypothesis and there are associations between Y and X variable, and F-test gives strong support to the model.

From parameter estimates table, because dw1 and atemp are insignificant, we flag them and will check if they would be changed or removed in model selection. And based on standardized estimates, we can see that year, weathersit and temperature are most significant predictors, we will also check them in our final model to see if the most significant predictors will be the same or similar.

1. Check Goodness-of-fit (A-10):

Null hypothesis H0: there is no relationship between dependent variable and independent variables.

Alternate hypothesis Ha: there is a relationship between dependent variable and independent variables.

From the Analysis of Variance, the F-value is 201.26 with a very small P-value (<.0001). So, we reject H0 and conclude that there are variables that can significantly predict the variation in Y (Cnt).

1. Check the Diagnostic (A-11):
2. Collinearity: Our model has Collinearity problem as temp and atemp has a very high correlation of 0.99. (A-6)
3. Outliers and Influential Points: Outliers and influential points exist because there are some points out of the range of 3 and -3 band in the studentized residual plot. (A-11)
4. Check Assumptions (A-11, A-12, A-13):
5. Linearity: satisfied. The predictors have a linear relationship with the dependent variable.
6. Normality: satisfied. The line in normal probability plot shows a 45-degree diagonal line, which indicates linear and normally distributed.
7. Independence: satisfied. The points look like randomly scattered around the zero line and there’s no specific pattern of the residuals.
8. Constant variance: satisfied because the points look like randomly scattered around the zero line and there’s no specific pattern of the residuals.

Even though the residual plots of temp/atemp look like have a pattern at the first place with a peak (or upward trending) during the middle part. However, because the dependent variable is the total number of users, which will change by time/ season/ temperature. As temperature goes up, more people will be more likely to use bikes. That’s why most of our users are clustered in summer and fall season. More observations mean more outliers and influential points. These points will change the line of residual plots and make the trend line curved and show a pattern. In addition, we have only 763 observations, which is a quite small dataset, and this kind of small dataset will be easily affected, and the pattern will be much more obvious compared to a large dataset. Therefore, once we can collect more data, or as the increasing of total number of users, this pattern will be eliminated, and the residual points will be scattered randomly around the zero line with no specific pattern.

**Model Validation**

1. Split training set and testing set (A-14):

We split our dataset into 70% training set and 30% testing set, and we used the training set, which has 512 observations to build a new model. We will also use the left of 30% of data -- testing set to test our model performance after we build our final model. The new model has a quite similar Adj-R2 with our full model, which is 0.8256.

1. 5- Fold Cross Validation (A-20, A-21):

In order to utilize our data better, we also used 5-Fold Cross Validation to evaluate model performance over different folds.

**Model Selection** (A-15, A-16, A-17, A-18, A-19, A-20, A-21)

1. Training and Testing sets (A-16, A-17, A-18, A-19):

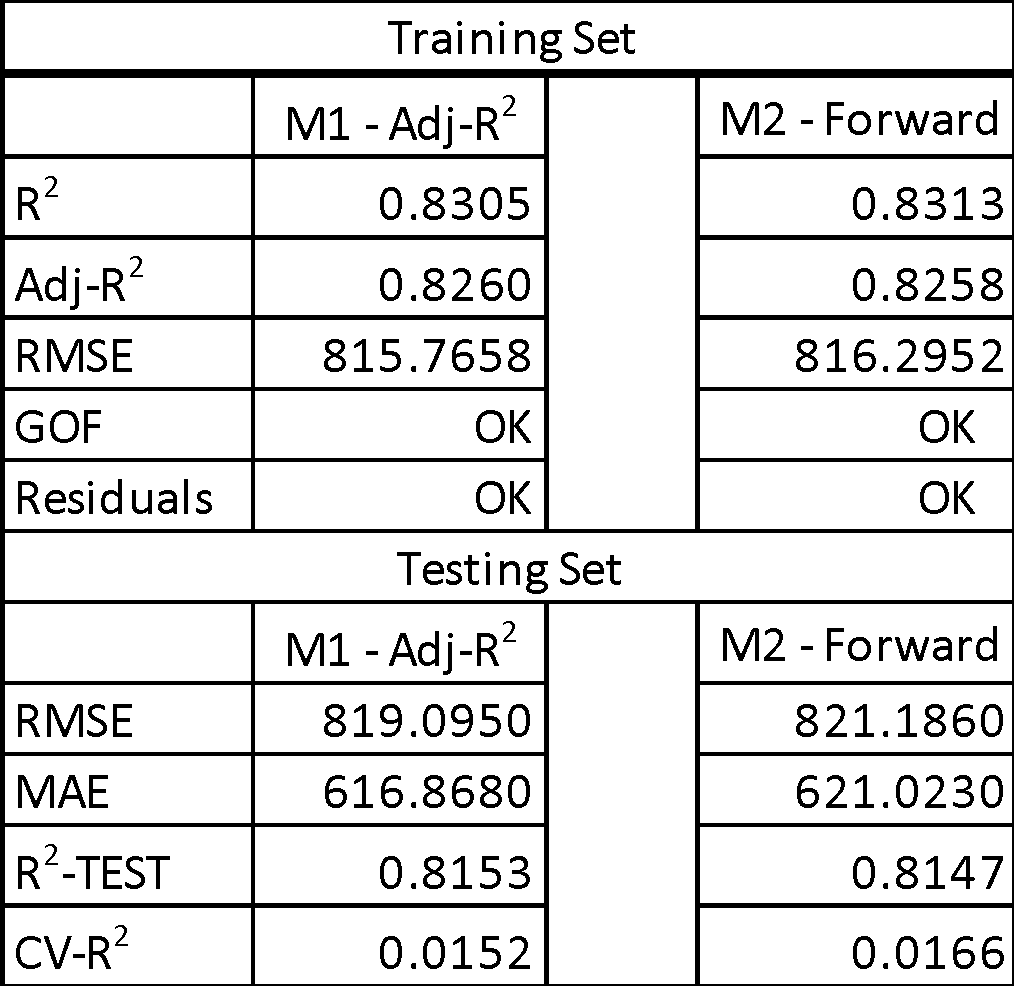
In order to find the best model, we applied two different Model Selections. The first one we chose Adj-R2 Model Selection, Adjusted R2 model selection comes up with a list of subsets who have similar R2. Even though our optimal choice should be the one with the highest adjusted R2 value, but we finally decided to choose the subset of 13 predictors as it’s adjusted R2 is 0.8260, just 0.0002 less than the highest adjusted R2 0.8262, which is ignorable. But it only has 13 predictors compared to 15 predictors for the highest Adjusted R2. Therefore, to combine these two reasons together, we decided to choose the subset of 13 predictors. The second method we chose Forward Model Selection. We also used forward model selection and got a total of 16 variables and the R2 is around 0.83.

After doing Model Selection, we use the testing set to evaluate how well the model performance with data outside the training set. From the picture we can see the two yhat are both around 90.2%.

1. 5- Fold Cross Validation (A-20, A-21):

From the pictures (A-20, A-21) we can see the ASE of training set and testing set are pretty close, which means the fitted model performs well with unseen data.

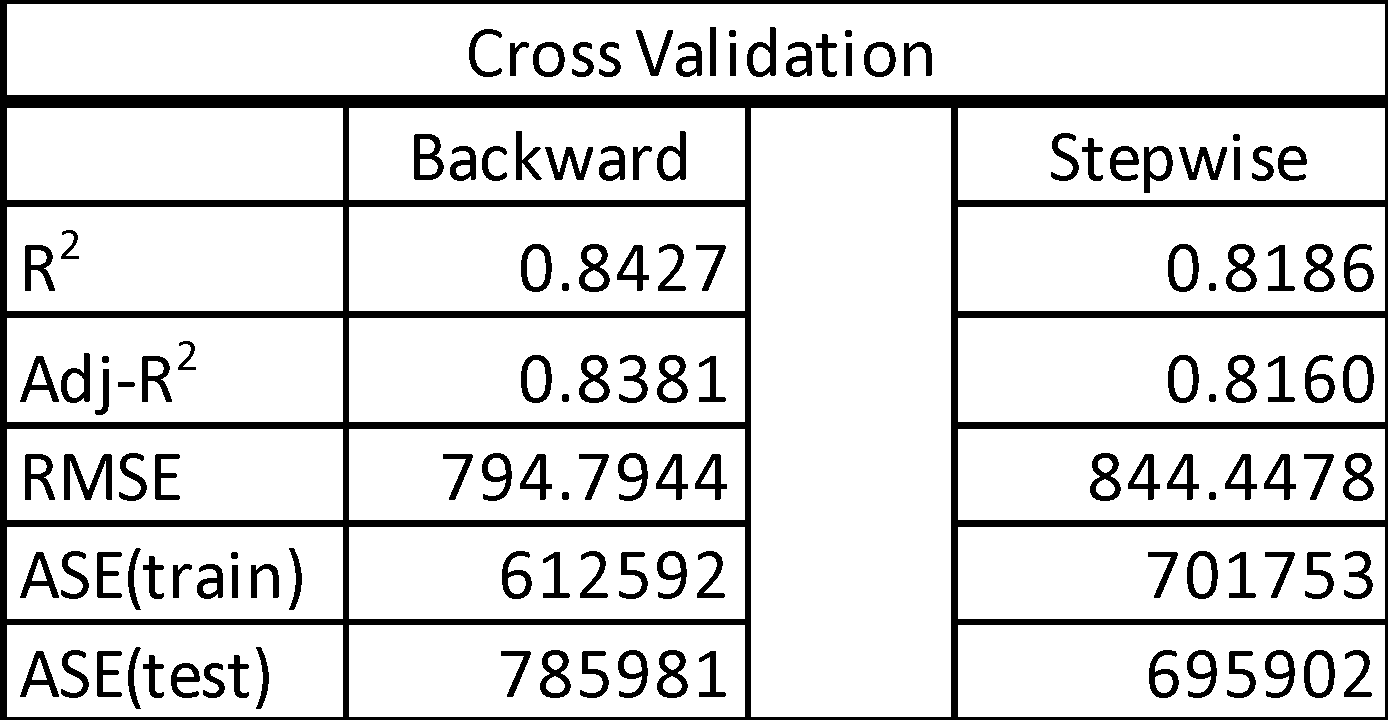
1. Choose better selection method:
   1. Training and Testing sets:

****

According to the testing set performance, they both have high R2 and their CV- R2 are both less than 0.3, which means the two models are pretty good.

We finally decided to choose the Adj-R2 because it has a higher Adj- R2 and lower RMSE, and its testing set performance a higher R2 Test and lower RMSE, showing that it will perform better in the future with another dataset.

* 1. 5- Fold Cross Validation:



Since the Adj-R2 of Backward is higher, we consider that Backward produces a better model.

* 1. Conclusion:

Adj-R2 method is selected because it has a quite high adj-R2 with less predictors.

**Fitted Final Model**

1. Check the Diagnostic:
2. Collinearity:

Collinearity problem has solved as we removed the predictor atemp, which is highly correlated with temp. Even though we tried to build interaction variable, as discussed before, the best way to solve collinearity in our model is to remove atemp and keep temp only. All variables’ VIF values are less than 10 (A-22).

1. Outliers and Influential Points (A-24):

There are still a few residuals that almost out of range of -3 and 3 band.

1. Goodness of fit:

Null hypothesis H0: there is no relationship between dependent variable and independent variables.

Alternate hypothesis Ha: there is a relationship between dependent variable and independent variables.

From the Analysis of Variance, we can see that F-value is high (361.10). P-value associated with the F-value is almost zero (<.0001). So, we can reject H0. There are associations between Y and X variable, so that at least one of the predictors can significantly predict the variation in Y(Cnt), and F-test gives strong support to the model.

1. Check assumptions (A-23, A-24, A-25):
   1. Linearity: satisfied. The predictors have linear relationship with dependent variable.
   2. Normality: satisfied. The line in normal probability plot shows a 45-degree diagonal line, which indicates linear and normally distributed (A-24).
   3. Independence: satisfied. The points look like randomly scattered around the zero line and there’s no specific pattern of the residuals.
   4. Constant variance: satisfied because the points look like randomly scattered around the zero line and there’s no specific pattern of the residuals.
2. Test Final model using testing set:

We tested the final model using testing set, and the R2-Test is 0.8153, and CV-R2 is less than 0.3. So, we believe this is a good model, and it will also perform well in other dataset in the future.



1. Check outliers and influential points(A-22, A-26, A-27, A-28, A-29, A-30, A-31):

After determining our final model (A-22), we start to remove outliers and influential points. In order to reduce as much bias as possible, firstly, we remove observations, which have both red and blue arrows for both outliers and influential points, which are 442, 554, 668, 669, 692, and 694 (A-26). We can see that adj-R2 increases from 0.8237 to 0.8448 and it is about 2.56% increase (A-27). Then we rerun the model and continue to remove observations which have both red and blue arrows, which are 204, 499, 553, and 689 (A-28). We can see that adj-R2 increases from 0.8448 to 0.8556 and it is about 1.28% increase (A-29). Afterword, we rerun the model and find only one observation which have both red and blue arrows. We decide to remove it as well, which is observation 238 (A-30). In the end, after removing 11 observations by three times, the adj-R2 increases from 0.8237 to 0.8573 (A-31). Finally, we found out that there is one variable dw2 is non-significant (p-value is greater than 0.05). We decided to remove it and get the final model (A-32).

1. Final Model Equation:

Cnt = 2927.32 – 1685.67\*ds1 – 470.76\*ds2 – 692.06\*ds3 + 2067.61 \* yr – 161.40\*dw1 + 494.00\*dw6 + 364.96\*workingday – 454.47\*dwe2 – 1874.59\*dwe3 + 4958.88\*temp – 1249.16\*hum – 2461.76\*windspeed

**Prediction**

After we built the final model, we did some research and used the final model to predict how bike sharing system would perform in Chicago on summer time and winter season.

1. Predict the total number of users in Chicago, non-working Saturday during August in 2012 (A-33).

According to our research, the temperature in Chicago during August would be 81° / 70° F (21 - 28 °C), and because our model used normalized (divided by 41) temperature, we did the same process for prediction and got an average normalized temperature of 0.6.

The humidity in Chicago during August usually fall in the range from 55 percent to 85 percent. So, we normalized humidity by divided by 100, and got an average humidity of 0.7.

The average hourly wind speed in Chicago during August increases from 8.5 miles per hour to 9.3 miles per hour. We also normalized (divided by 67) wind speed and got an average value of 0.13.

Through these numbers, we computed the predicted total users, 95% confidence interval and prediction interval for our estimation using final model. The predicted Value is 6578, the 95% confidence interval is (6374, 6782) and 95% prediction interval is (5135, 8020).

Therefore, our conclusion is: In Chicago, a non-working Saturday in August 2012 the total number of users can reach 6578. And the total users can range anywhere between 5135 and 8020 confidence intervals.

1. Predict the total number of users in Chicago, non-working Saturday during January in 2012 (A-34).

According to our research, the temperature in Chicago during January would be 32° / 22° F (0 - 5.6 °C), and because our model used normalized (divided by 41) temperature, we did the same process for prediction and got an average normalized temperature of -0.014.

The humidity in Chicago during January usually around 76 percent. So, we normalized humidity by divided by 100, and got an average humidity of 0.76.

The average hourly wind speed in Chicago is decreasing from 13.9 miles per hour to 13.4 miles per hour. We normalized (divided by 67) the wind speed, and the average is 0.2.

We computed the predicted total users, 95% confidence interval and prediction interval for our estimation. The predicted Value is 2103, the 95% confidence interval is (1686, 2520) and 95% prediction interval is (615, 3591).

Therefore, our **conclusion** is: For Chicago, a non-working Saturday in February 2012 with light snow or rain, the total number of users can reach 2103. And the total users can range anywhere between 1686 and 2520 confidence interval.

From above predictions, we can see that Chicago has a large market as the predicted total users are more than Washington DC no matter in Summer or Winter. So, we would recommend the bike sharing company to set branch office in Chicago.

**Results and Findings:**

1. The final model equation:

Cnt = 2927.32 – 1685.67\*ds1 – 470.76\*ds2 – 692.06\*ds3 + 2067.61 \* yr – 161.40\*dw1 + 494.00\*dw6 + 364.96\*workingday – 454.47\*dwe2 – 1874.59\*dwe3 + 4958.88\*temp – 1249.16\*hum – 2461.76\*windspeed

From absolute value of Standardized Estimate in Parameter Estimates table (A-29), the strongest to weakest independent variables are yr, temp, ds1, dwe3, ds3, dwe2, ds2, windspeed, hum, dw6, workingday and dw1.

From dataset characteristics, four variables are normalized: temp, atemp, hum, windspeed. Thus, if the year increases by 1, the total user will increase by 2067. If humidity increases by 1%, the total user will decrease by 1249. If the temperature increases by 1 °C, total bike-sharing users will increase 121.

1. Results:

From this study, we finalized three important predictors, which are year, temperature and ds1.

* 1. Year variable is the most important predictor from this model. However, we don’t agree with the result, because two years data is very limited, and the year variable should not be a factor affecting bike sharing usage. Thus, we will collect more data that covered a longer time period for further analysis whether the year variable is the most important predictor that affect bike sharing usage or not.
  2. Temperature is the second important predictor from this model, which means that when the temperature gets lower, the total number of bike sharing will decrease. On the contrary, when the temperature gets higher, the total number of bike sharing will increase.
  3. Ds1 is the third important predictor from this model. Ds1 is from Dec 21st to March 21th. This time period is the coldest time in Washington, D.C. Thus, ds1 has the negative effect on total count of users.

**Future work**

In 2012, total population in Washington D.C. was 633427. The average bike-sharing users were 4504 among our data analysis. There was only 0.7% of the population used bike-sharing system. Thus, the potential bike-sharing market in huge and our analysis is valued.

Based on existing dataset that we used, we can analyze how performance is different between casual users and registered users. Also, we can analyze how performance is varied among 24 hours within a day.

During our analysis, we found out that the data collectors focused on two factors: time (season, holiday, weekday, workingday) and temperature (temp, atemp, humidity, windspeed). In order to make the most advantage of bike-sharing system, we can analyze more factors that affecting bike-sharing. For example, the relationship between accessibility (how far the user must walk to reach stations) and bike-availability (the likelihood of finding a bicycle) (Kabra 2016).

In addition, as international students, we are lucky to experience the fast-growing bike sharing system in both China and the United States. However, we noticed that there are still lots of differences between these two countries while implementing bike sharing. According to our research -- the largest bike sharing program worldwide, 20 of 25 cities are in China. This statistic really interests us to analyze what and how makes Chinese bike sharing programs developed so fast in a few years. And we believe that these features can be actually used by other countries to improve the usage, such as computerized bike sharing system in Hangzhou, China, which integrates stations with bus and subway networks and allows the same transit card to be used across all modes and grants extra free bike riding time with a bus transfer (Larsen 2013).

With the development of bike sharing program, Washington D.C, the pioneer of bike sharing system, has launched a new technology in 2017 -- Dockless bikes. The interesting thing is dockless bikes are much more popular in the evening and on the weekends, instead of morning and evening peaks of station-based bikes. This phenomenon indicates that dockless bikes could be a totally different mode than traditional bikes because it could possibly use for fun instead of work. We are so interested and passionate to analyze dockless bikes because if the assumption that it is more useful when relaxing and having fun is true, the combination of dockless bikes and station-based bikes systems will be able to fulfilling all kinds of different needs, which will largely satisfy most of people in the city and nurture a growing ecosystem of mobility options by providing the advantages of both systems (Walker 2018).

**Research References**

Guo Yanyong, Zhou Jibiao,Wu Yao,Li Zhibin. 2017. *Identifying the factors affecting bike-sharing usage and degree of satisfaction in Ningbo, China*.

Kabra Ashish, Belavina Elena, Girotra Karan. 2016. Bike-Share Systems: Accessibility and Availability.

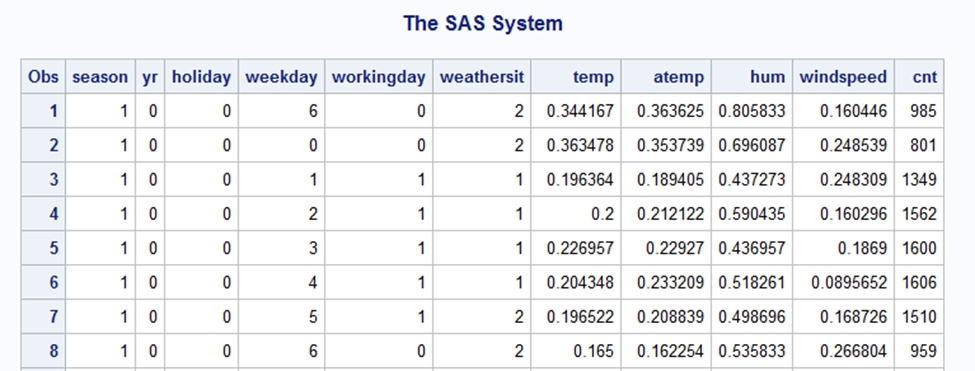
Larsen, Janet. Bike-Sharing Programs Hit the Streets in Over 500 Cities Worldwide. 2013. <http://www.earth-policy.org/plan_b_updates/2013/update112>

Susan A. Shaheen, Elliot W. Martin, Nelson D. Chan, Adam P. Cohen, and Mike Pogodzinski. 2014. Public Bikesharing in North America During a Period of Rapid Expansion: Understanding Business Models, Industry Trends and User Impacts. MINETA TRANSPORTATION INSTITUTE.

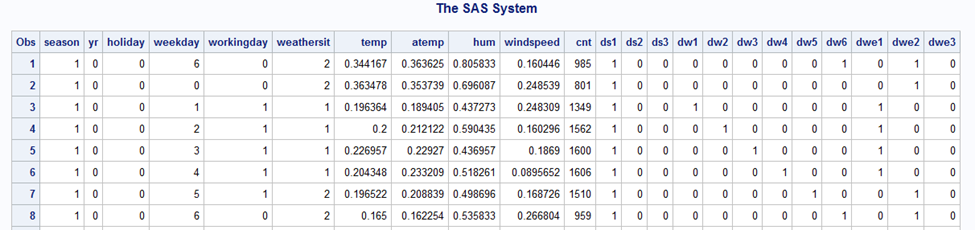
Walker, Alissa. Can dockless and station-based bike-share programs coexist? 2018. <https://www.curbed.com/2018/5/30/17390264/bike-sharing-dockless-bicycles-cities>

**Appendix**

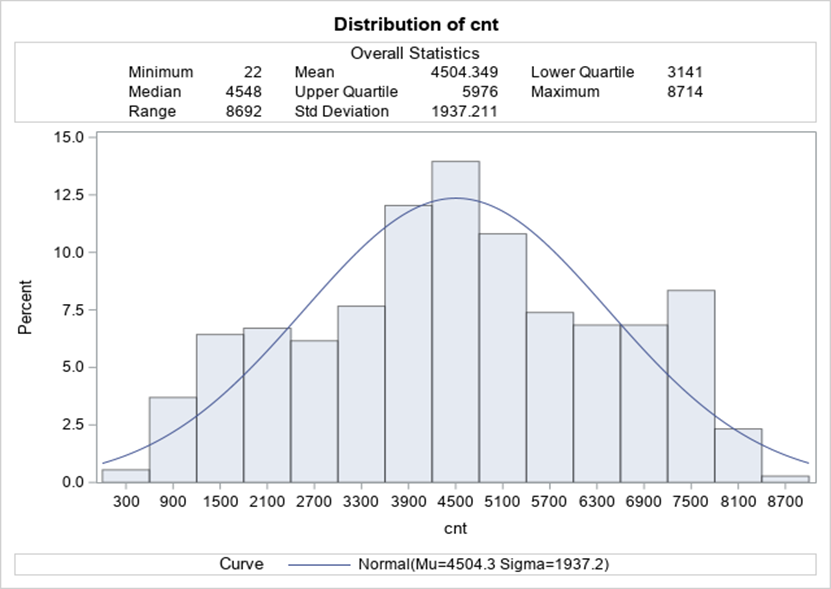
A-1: Full dataset



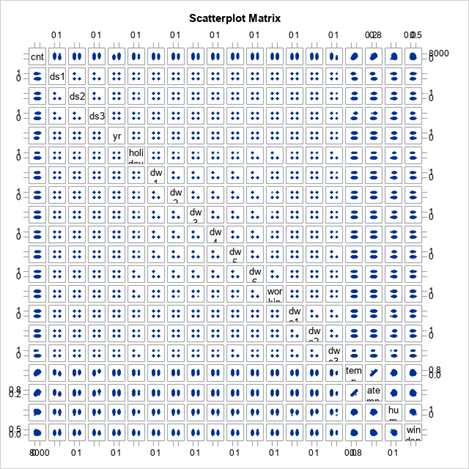
A-2: Create dummy variables



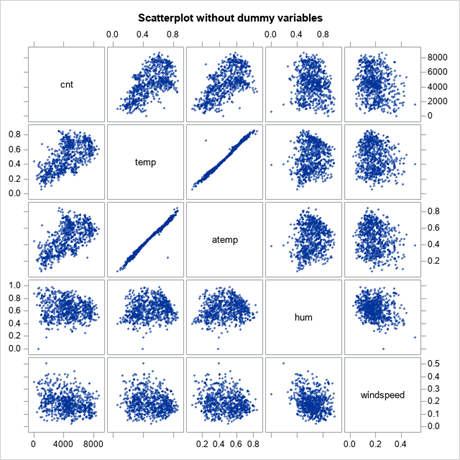
A-3: Histogram



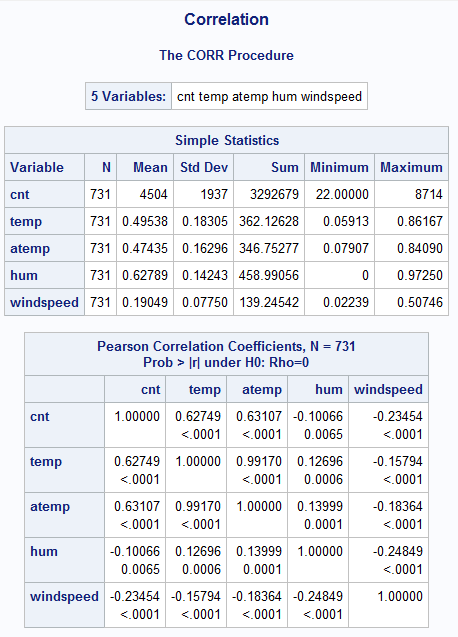
A-4: Scatter Plots for All variables



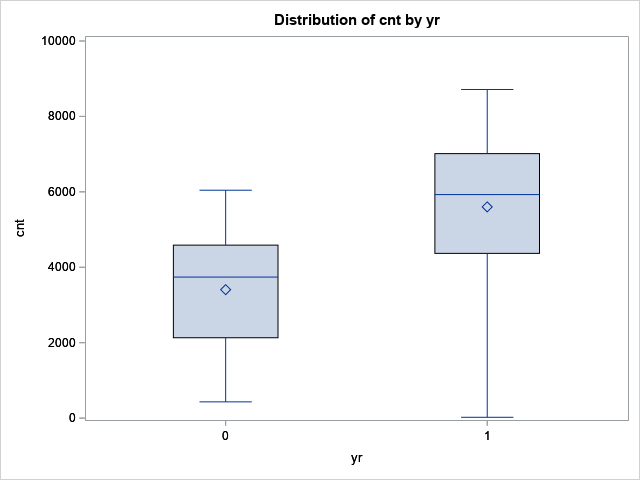
A-5: Scatter Plots without dummy variables

****

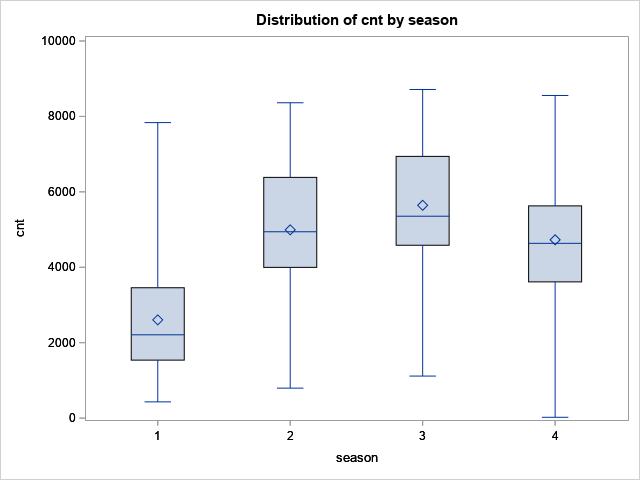
A-6: Correlation without dummy variables



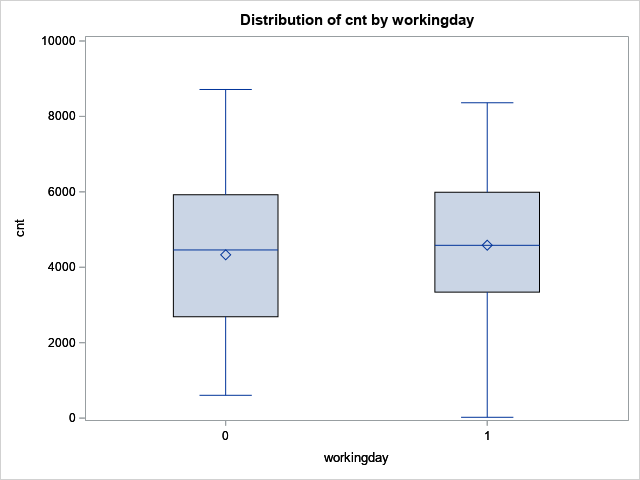
A-7: Boxplot between count and year



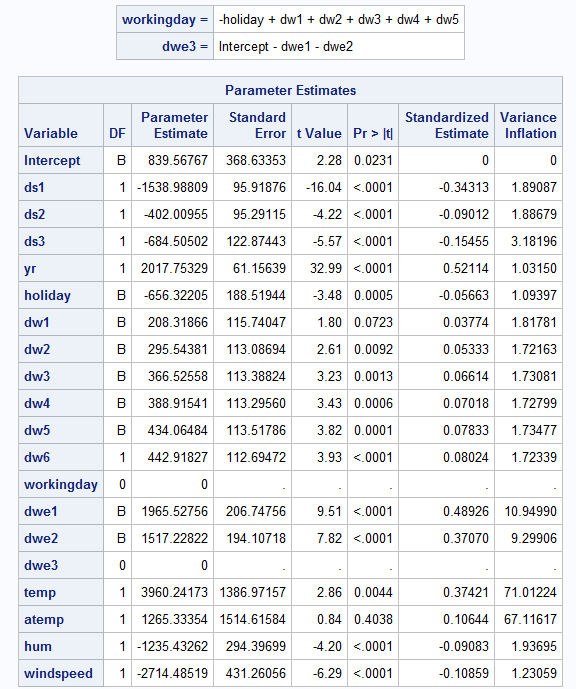
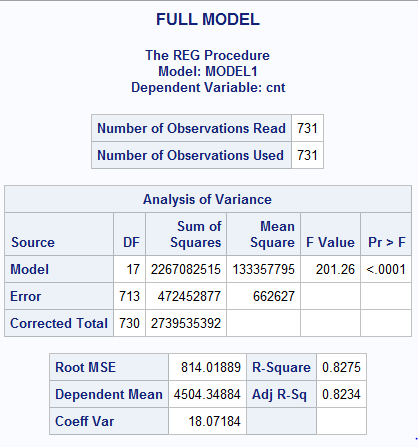
A-8: Boxplot between count and season

****

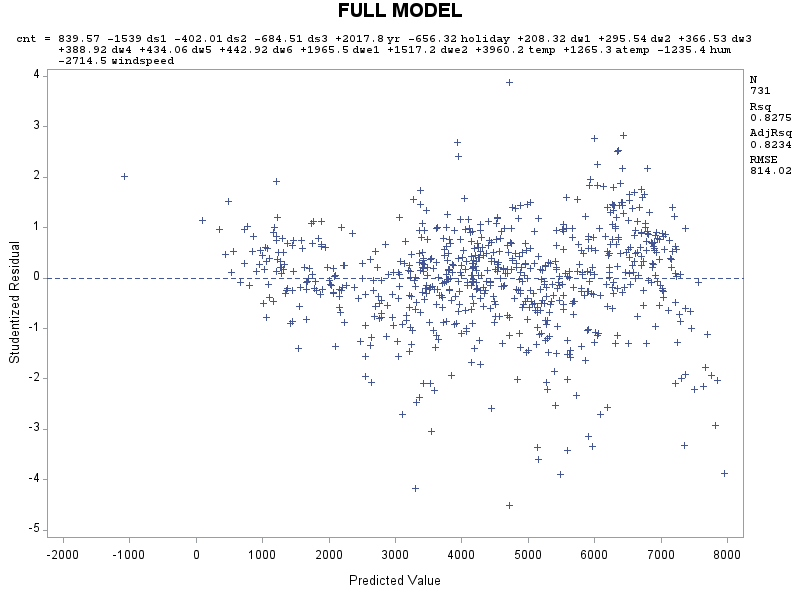
A-9: Boxplot between count and workingday

****

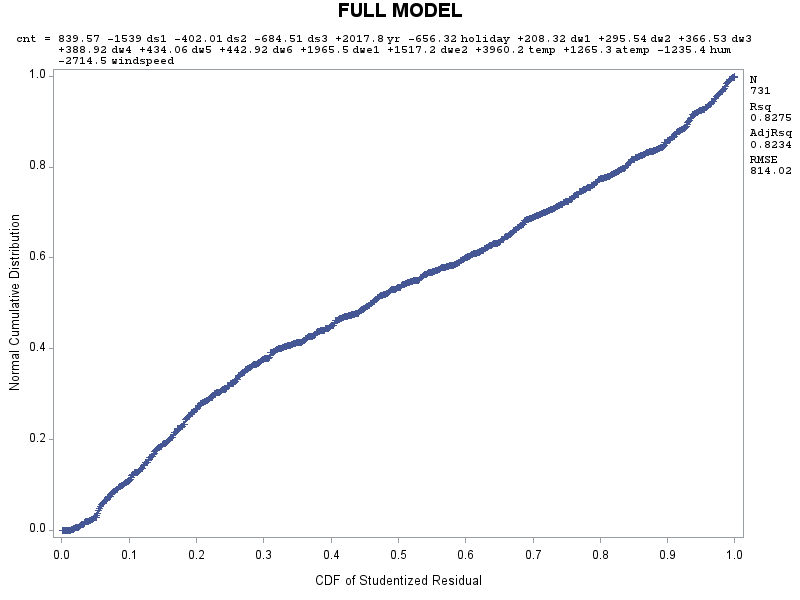
A-10: Full model (1)



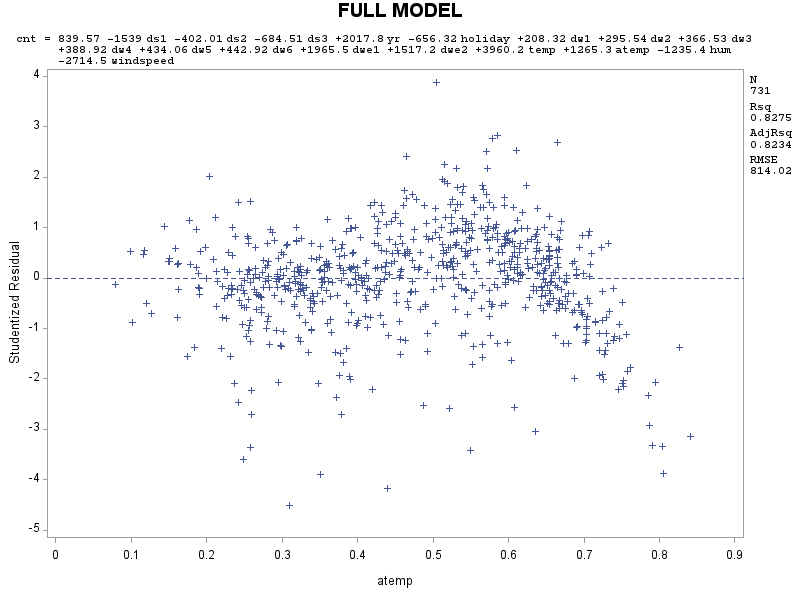
A-11:  Studentized Residual vs Predicted Value

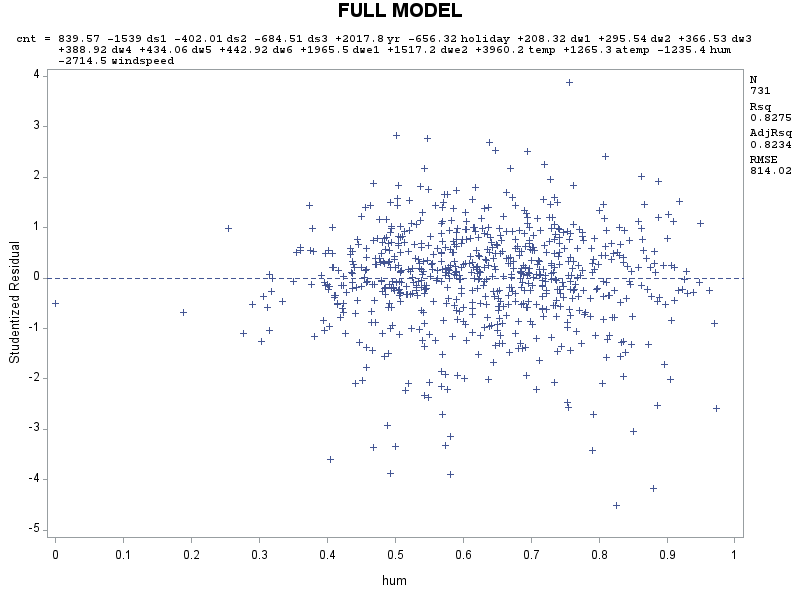


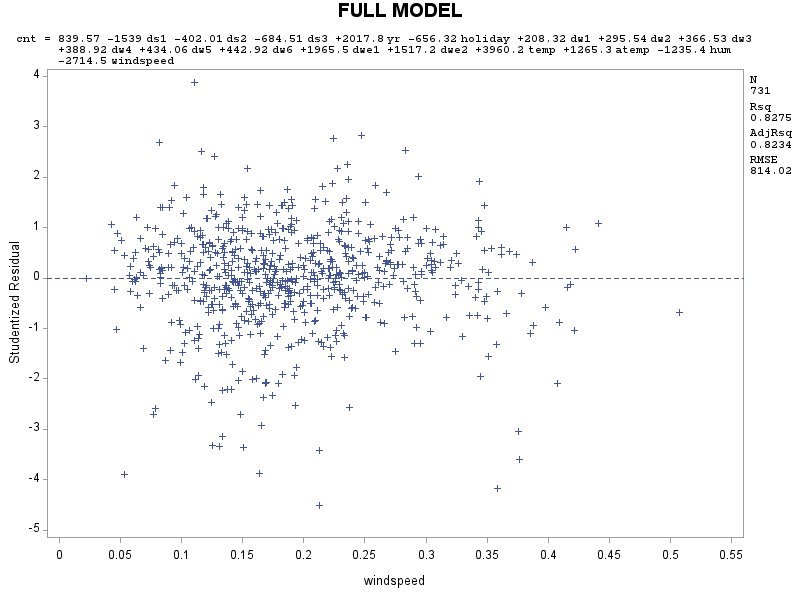
A-12: Normal Probability Plot



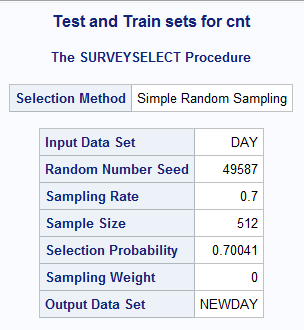
A-13: Studentized Residual Plots for atemp, hum, and windspeed



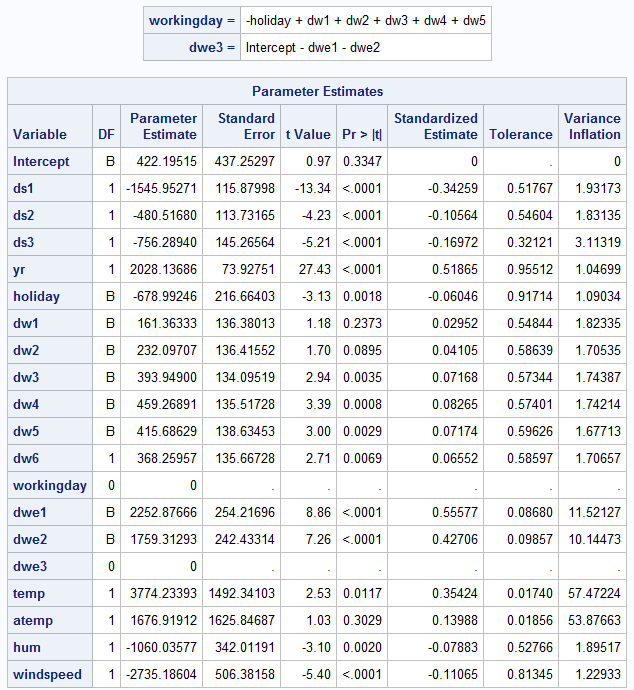
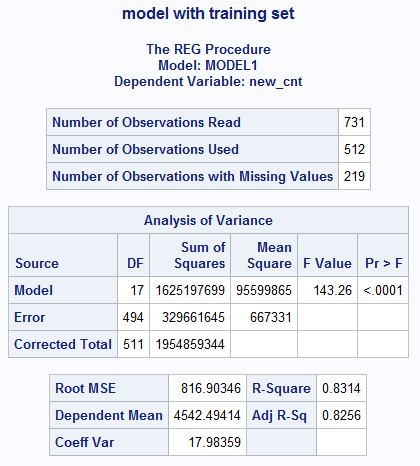




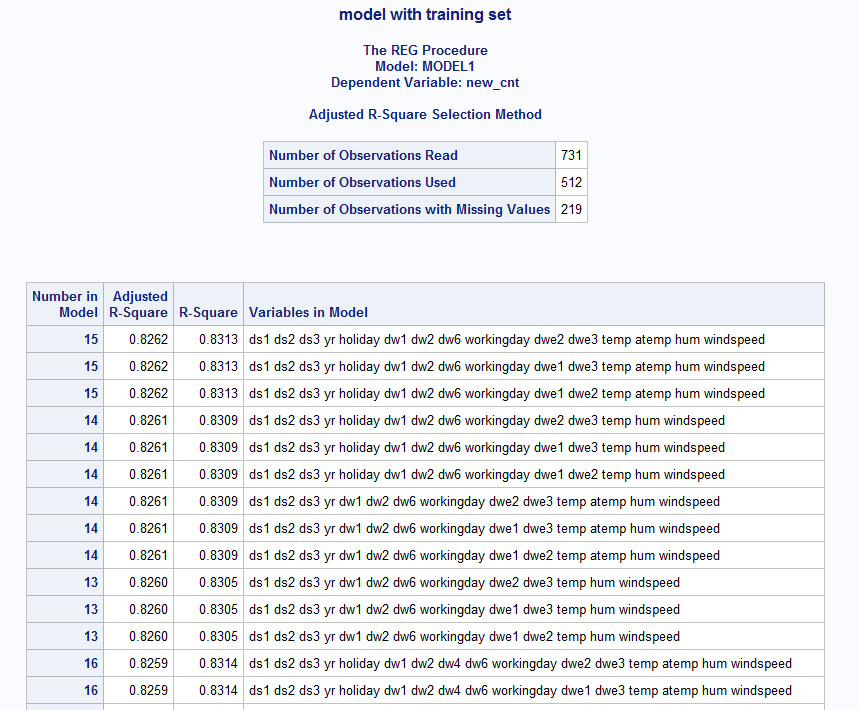
A-14: Test and Train sets



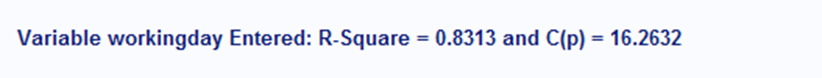
A-15: Model with training set

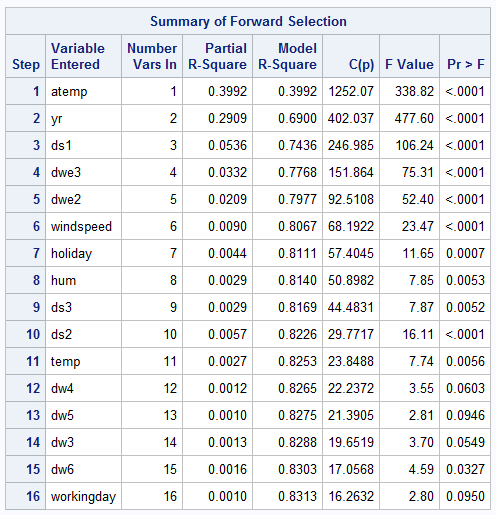


A-16: Model Selection: Adj-R2 (training set)

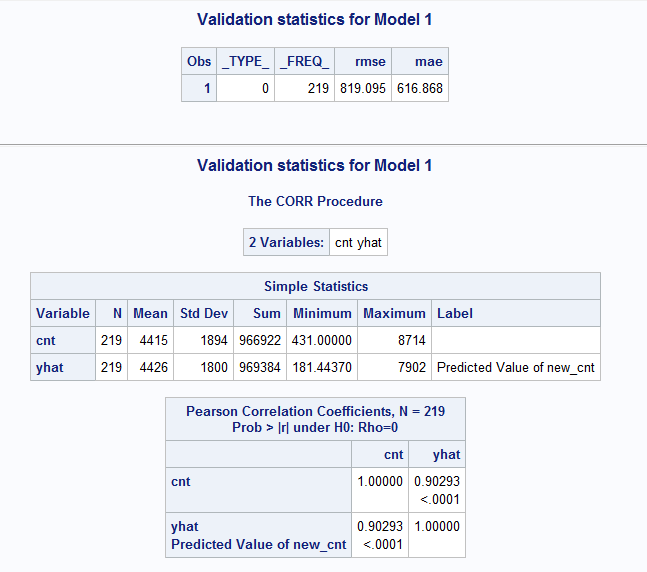


A-17: Model Selection: Forward (training set)

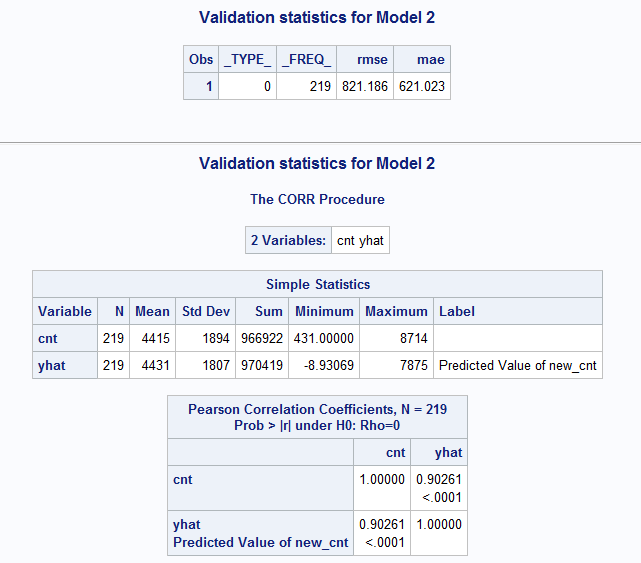
****



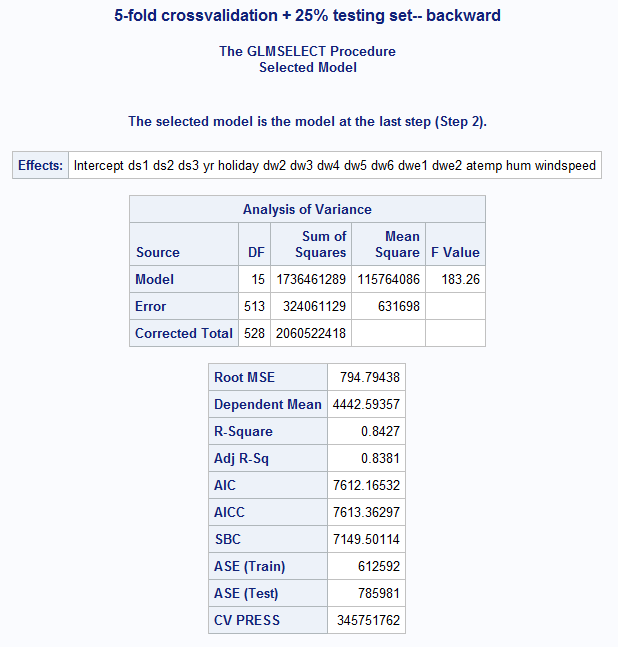
A-18: Validation Statistics for Adj-R2



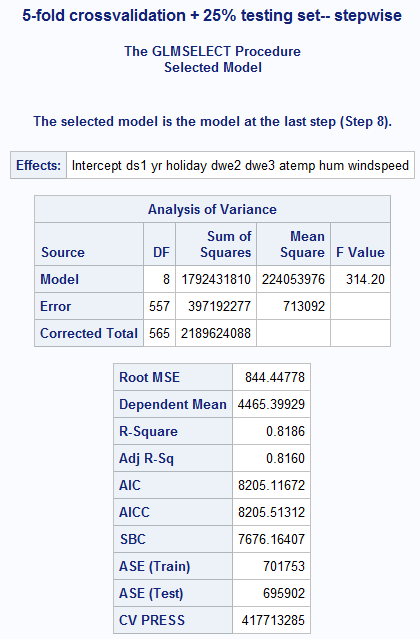
A-19: Validation Statistics for Forward



A-20: 5-fold cross validation with Backward



A-21: 5-fold cross validation with Stepwise



A-22: Final Model

A screenshot of a cell phone

Description automatically generated

A-23: Studentized Residual Plots for atemp, hum, and windspeed

A close up of a map

Description automatically generated

A close up of a map

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A-24: Studentized Residual vs Predicted Value

A screenshot of a social media post

Description automatically generated

A-25: Normal Probability Plot

A close up of a map

Description automatically generated

A-26: Outliers and influential points





A screenshot of a cell phone

Description automatically generated

A-27: After removing outliers and influential points for the first time

A screenshot of a cell phone

Description automatically generated

A-28: Outliers and influential points









A-29: After removing outliers and influential points at second time

A screenshot of a cell phone

Description automatically generated

A-30: Outliers and influential points



A-31: After removing last outliers and influential points

A screenshot of a cell phone

Description automatically generated

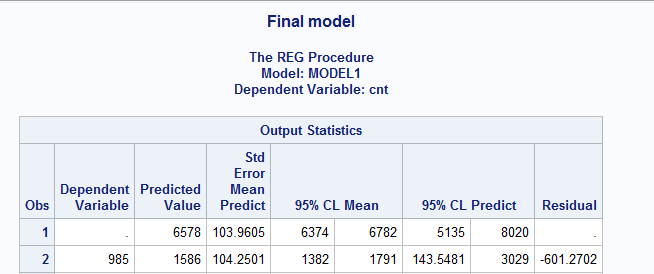
A-32: After removing non-significant variable (dw2):

A screenshot of a cell phone

Description automatically generated

A-33: Prediction - 1





A-34: Prediction - 2

