MIDDLE EAST TECHNICAL UNIVERSITY

DEPARTMENT OF STATISTICS

STAT 364 – LINEAR MODELS II

ANALYSIS OF BIKE SHARING DATASET

Prepared by

Orçun Oltulu

İbrahim Hakkı Erduran

2017-2018

TABLE OF CONTENT

ABSTRACT

The bike sharing data is collected from bike sharing system which name is Captial Bike Sharing (CBS) at Washington, D.C., USA in 2011. It includes number of rental bike and external factors such as weather situation and temperature. Thanks to bike sharing system, members can rent a bike from one point and return it another point. The aim of the project, finding model which can best predict number of rental bikes and examining the relationship between some variables in the data. Poisson regression method is used to find the best model since count data has poisson distribution, but over-dispersion problem occurs. To solve this, negative binomial regression method is used. To make sure that we have the best model for prediction number of rental bikes, other terms such as square of humidity and square of temperature are added to model, but after comparing the two models with the k-fold cross validation, first model is founded the best model for prediction number of rental bikes.

Keywords: Bike sharing data, poisson regression, negative binomial regression, linear model project

1. Background/Introduction

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return has become automatic. Through these systems, user can easily rent a bike from a position and return at another position. Currently, there are more than 400 bike-sharing programs around the world which is composed of over 500 thousand bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real-world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data.

1.1. Data description

Data set is taken from https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset and it is called Bike Sharing Dataset. Variable description can be found below.

Variable Description:

- season: Season variable has 4 categories. 1 refers to Spring, 2 refers to Summer, 3 refers to Fall, 4 refers to Winter.
- hr: Hour is between (0 and 23). It records to time of renting bike.
- holiday: It shows that day is holiday day or not.
- workingday: If day is not weekend or holiday, it represents by 1. If day is weekend or holiday it represents by 0.

- weathersit:
  - 1 refers to Clear, Few clouds, Partly cloudy, Partly cloudy weather.
  - 2 refers to Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist weather.
  - 3 refers to Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds weather.
- temp: Temperature is converted into Celsius.
- hum: Humadity is normalized by diveding it 100.
- windspeed: Wind speed is normalized by diving it 67.
- cnt: Total number of rental bikes.

Name and type of independent variables:

1. **Independent variable:** season, Categorical
2. **Independent variable:** hr, Discreate
3. **Independent variable:** holiday, Categorical
4. **Independent variable:** workingday, Categorical
5. **Independent variable:** weathersit, Categorical
6. **Independent variable:** temp, Continuous
7. **Independent variable:** hum, Continuous
8. **Independent variable:** windspeed, Continuous

Name and type of dependent variable:

cnt: Total number of rental bikes.

1.2. Research questions

1. Which variables are likely to best predict the number of rental bikes(count)?
2. Is there any significance difference between the average number of rental bikes for different weather situation?
3. Are there any relationship between humidity and temperature?
4. Does wind speed decrease humidity?
5. Are there any relationship between temperature and number of rental bikes(count)?

1.3. Aim of the study

In this project, our aim is to estimate the total number of rental bikes by getting help from the past data which contains several explanatory variables.

2. Methodology/Analysis

2.1. Data Preparation:

Our data has 8645 observations and we select a random sample size of 800 observations. At first, we had 15 independent variables; however, we eliminate some of the explanatory variables like date, year, month etc. As it is provided above in data description part, we have 8 explanatory variables which are numerical, continuous and categorical variables.

For season variable we create 3 dummy variables and we take spring as reference group. In addition, for weathersit variable we have 3 levels and we create 2 dummies and we take "Clear, Few clouds, Partly cloudy, Partly cloudy" weathers as reference group.

2.2. Proposed Methods:

Proposed method for the first research question is Poisson Regression method since we try to estimate the total number of rental bikes. The Poisson distribution for a random variable Y has the following probability mass function:

$$\Pr(Y = y \mid \mu) = \frac{e^{-\mu}\mu^{y}}{y!} \quad (y = 0,1,2,\ldots)$$

Furthermore, the Poisson Regression model for each observation is given by:

$$\Pr(Y_i = y_i \mid \mu_i, t_i) = \frac{e^{-\mu_i t_i}(\mu_i t_i)^{y_i}}{y_i!}$$

Poisson Regression method has 3 assumptions:

- Data should be a count data.
- The distribution of counts should follow a Poisson distribution.
- The mean and the variance of the model should be identical.

As you can find in the Results and Findings section we discuss and check the assumptions of Poisson Regression and find out the fact that the equality of mean and the variance of the model is not satisfied.

So, we use Negative Binomial Regression. The Negative Binomial distribution for a random variable Y has the following probability mass function which includes a gamma noise variable whose mean is 1 and scale parameter is denoted as v is:

$$\Pr(Y = y_i|\mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}} \left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{y_i}$$

where

$$\mu_i = t_i\mu$$

$$\alpha = \frac{1}{v}$$

Thus, The Negative Binomial Regression Model for each observation is given by:

$$\Pr(Y = y_i|\mu_i, \alpha) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i + 1)}\left(\frac{1}{1 + \alpha\mu_i}\right)^{\alpha^{-1}}\left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^{y_i}$$

Proposed method for the second research question is checking correlation between total number of rental bikes for each weather type.

Proposed method for the third research question is checking correlation between humidity and temperature.

Proposed method for the fourth research question is checking correlation between wind speed and humidity.

Proposed method for the fourth research question is checking correlation between temperature and number of rental bikes(count).

2.3. K-Fold Cross Validation:

We use K-fold Cross Validation to choose the model which provides smaller MSE values. We choose k = 10; in other words, we divided our data into 10 folds and for each fold we train the model for 720 observations and test the model with remaining 80 observations. You can see Cross Validation part in detail in the Results and Findings section.

3. Results and Findings

Findings of Research Question 1: *Which variables are likely to best predict the number of rental bikes(count)?*

3.1 Assumption Checking:

Assumption 1: Data should be a count data.

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   33.75  112.00  144.24  214.00  601.00
```

Result 1: *summary of the response variable (cnt).*

```
 [1]  54 240  28  18  84  11  10   8 188 258 202   6  22 287   2
[16]   5 290 146 228 429 137 576  37 221 390 260  62  63 349 370
[31] 174  33 154 109  51  10 286 121   1  47 401 219 113   2 104
[46]  15   3   1  72  39
```

Result 2: *Randomly selected sample from response variable (cnt).*

As it can be seen in the *Result 1 and Result 2*, our response variable is a count of the rental bikes. Thus, first assumption is satisfied.

Assumption 2: The distribution of counts should follow a Poisson distribution.

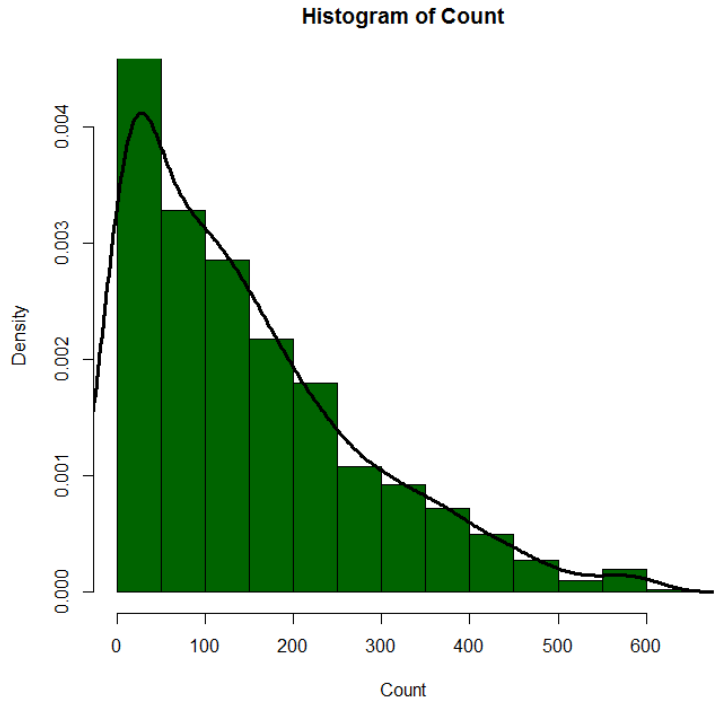

Figure 1: *Histogram of response variable (cnt).*

By looking at the *Figure 1* it is easy to say that response variable follows Poisson Distribution.

Assumption 3: The mean and the variance of the model should be identical.

Mean	144.2413
Variance	17206.93

Result 3: *Mean and Variance of response variable (cnt).*

As it can be seen in *Result 3* the variance is much greater than the mean, this will probably lead to over-dispersion.

Multicollinearity Checking:

	Variables	VIF
1	season	1.213441
2	hr	1.105397
3	holiday	1.069226
4	workingday	1.072580
5	temp	1.182742
6	hum	1.163398
7	windspeed	1.138040

***Table 1*:** VIF table of the data

According to *Table 1*, there is no multicollinearity in between variables since all VIF values are less than 10.

3.2 Model Selection:

Now we are going to build the Poisson regression model. The model with all predictors can be seen in the below.

```
Call:
glm(formula = cnt ~ ., family = poisson(link = "log"), data = new.data.2011)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-18.973   -7.288   -2.314    3.419   29.883

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      4.2080465  0.0209558 200.806   <2e-16 ***
hr               0.0495378  0.0004971  99.644   <2e-16 ***
holiday1        -0.1848652  0.0196488  -9.408   <2e-16 ***
workingday1      0.0066488  0.0066710   0.997   0.3189
temp             0.0386482  0.0006099  63.372   <2e-16 ***
hum             -0.0084815  0.0001773 -47.830   <2e-16 ***
windspeed       -0.0006862  0.0004026  -1.705   0.0883 .
season.dummy11   0.2909661  0.0124969  23.283   <2e-16 ***
season.dummy21   0.2881714  0.0146404  19.683   <2e-16 ***
season.dummy31  -0.4868733  0.0109946 -44.283   <2e-16 ***
weather.dummy11  0.0945743  0.0072466  13.051   <2e-16 ***
weather.dummy21 -0.4644542  0.0168165 -27.619   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 94188  on 799  degrees of freedom
Residual deviance: 54948  on 788  degrees of freedom
AIC: 59896

Number of Fisher Scoring iterations: 5
```

Result 4: *Summarizing the fitted Poisson model with all explanatory variables.*

```
> sqrt(54948 / 788)
[1] 8.350507
```

Result 5: *Dispersion.*

Result 5 implies an over-dispersion problem since the result is greater than 1. The non-equality of mean and variance may be caused by over-dispersion.

Since there is an over-dispersion problem, in order to eliminate this problem, we use negative binomial regression method by "glm.nb" function which is in "MASS" package in R.

```
Call:
glm.nb(formula = cnt ~ ., data = new.data.2011, init.theta = 1.34022175,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8135  -0.9460  -0.2707   0.3631   2.9316

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.6754573  0.2083125  17.644  < 2e-16 ***
hr               0.0758170  0.0048992  15.475  < 2e-16 ***
holiday1        -0.0953453  0.1961401  -0.486 0.626890
workingday1      0.1000469  0.0689011   1.452 0.146491
temp             0.0406477  0.0063103   6.441 1.18e-10 ***
hum             -0.0064000  0.0018782  -3.407 0.000656 ***
windspeed       -0.0006035  0.0042209  -0.143 0.886298
season.dummy11   0.2464986  0.1134261   2.173 0.029765 *
season.dummy21   0.2466268  0.1441337   1.711 0.087063 .
season.dummy31  -0.4948788  0.0992970  -4.984 6.23e-07 ***
weather.dummy11  0.0521972  0.0742852   0.703 0.482268
weather.dummy21 -0.6451706  0.1334031  -4.836 1.32e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.3402) family taken to be 1)

    Null deviance: 1364.86  on 799  degrees of freedom
Residual deviance:  895.73  on 788  degrees of freedom
AIC: 9198.4

Number of Fisher Scoring iterations: 1


              Theta:  1.3402
          Std. Err.:  0.0633

 2 x log-likelihood:  -9172.3660
```

Result 6: *Summarizing the fitted Negative Binomial model with all explanatory variables.*

In *Result 6*, model has several insignificant variables which have to be eliminated.

```
> sqrt(895.73 / 788)
[1] 1.066168
```

Result 7: *Dispersion.*

By looking at *Result 7* we can say that there is no over-dispersion problem.

After eliminating the insignificant variables, we fit the model and print summary of it.

```
Call:
glm.nb(formula = cnt ~ hr + temp + hum + season.dummy3 + weather.dummy2,
    data = new.data.2011, link = "log", init.theta = 1.329141308)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7424  -0.9593  -0.2811   0.3289   3.0084

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       3.521976   0.165444  21.288  < 2e-16 ***
hr                0.075222   0.004801  15.667  < 2e-16 ***
temp              0.049407   0.003866  12.781  < 2e-16 ***
hum              -0.005201   0.001709  -3.044  0.00233 **
season.dummy31   -0.352361   0.072379  -4.868 1.13e-06 ***
weather.dummy21 -0.657984   0.126508  -5.201 1.98e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.3291) family taken to be 1)

    Null deviance: 1353.98  on 799  degrees of freedom
Residual deviance:  896.27  on 794  degrees of freedom
AIC: 9194

Number of Fisher Scoring iterations: 1


              Theta:  1.3291
          Std. Err.:  0.0627

 2 x log-likelihood:  -9180.0020
```

Result 8: *Summarizing the fitted Negative Binomial model with only significant explanatory variables.*

```
> sqrt(896.27 / 794)
[1] 1.062452
```

Result 9: *Dispersion*

The model in *Result 8* consists only significant explanatory variables and it seems in *Result 9*, there is no over-dispersion problem since the ratio is close to 1.

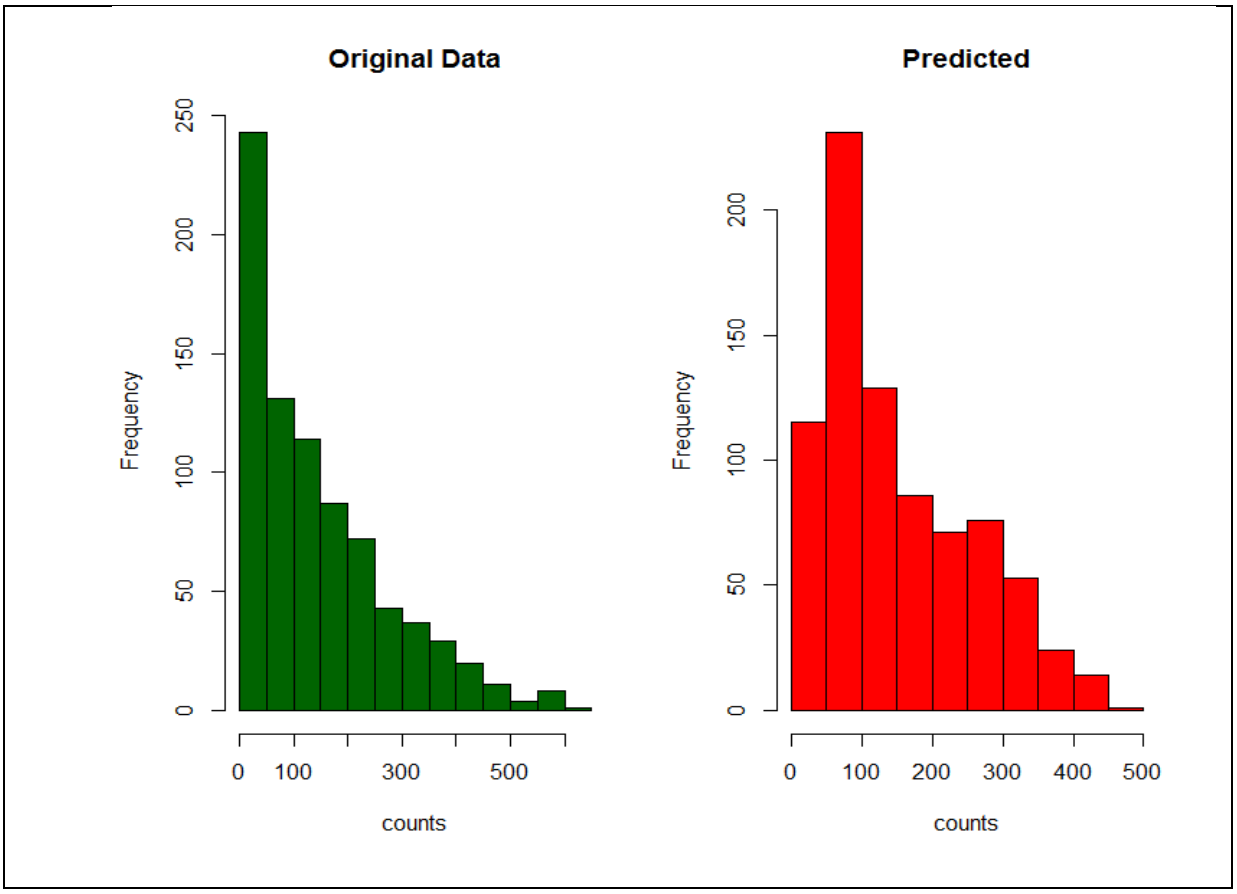

Figure 2: *Comparison of Original data and Prediction.*

Although we successfully eliminate the over-dispersion problem, still our model is not working well for predicting counts under 100.

In order to solve this problem, we fit the model again with hum^2 and temp^2.

```
Call:
glm.nb(formula = cnt ~ hr + I(temp^2) + I(hum^2) + season.dummy3 +
    weather.dummy2, data = new.data.2011, link = "log", init.theta = 1.312064576)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6882  -0.9729  -0.2848   0.3140   2.9089

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.807e+00  1.192e-01  31.930  < 2e-16 ***
hr               7.674e-02  4.807e-03  15.966  < 2e-16 ***
I(temp^2)        1.164e-03  9.566e-05  12.163  < 2e-16 ***
I(hum^2)        -3.523e-05  1.367e-05  -2.578  0.00995 **
season.dummy31  -3.888e-01  7.350e-02  -5.290 1.22e-07 ***
weather.dummy21 -6.567e-01  1.293e-01  -5.079 3.80e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.3121) family taken to be 1)

    Null deviance: 1337.21  on 799  degrees of freedom
Residual deviance:  897.03  on 794  degrees of freedom
AIC: 9205.8

Number of Fisher Scoring iterations: 1


              Theta:  1.3121
          Std. Err.:  0.0617

 2 x log-likelihood:  -9191.8220
```

Result 10: *Summarizing the fitted Negative Binomial model with only significant explanatory variables.*

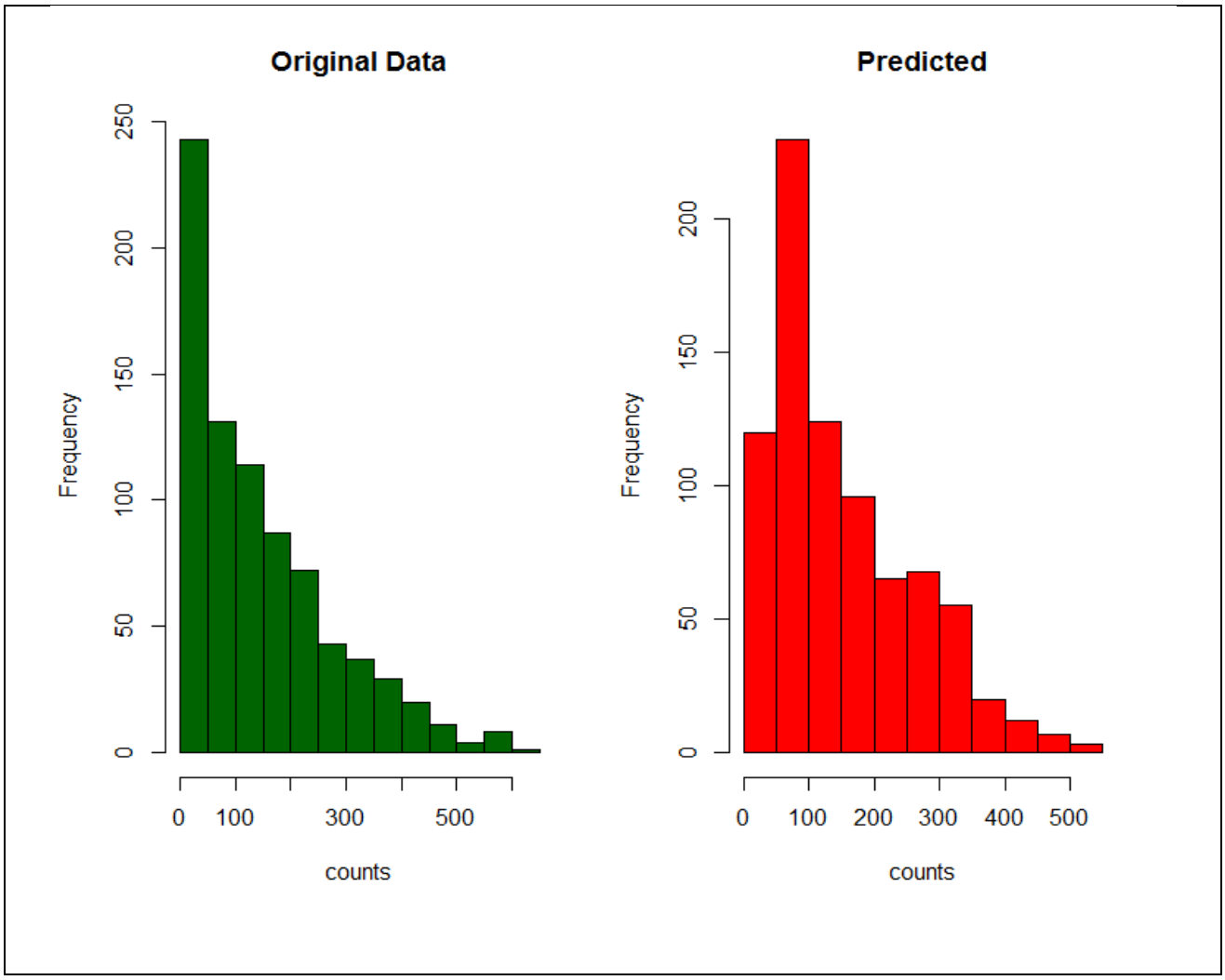

Figure 3: *Comparison of Original data and Prediction.*

Unfortunately, as it can be seen in Figure 3, still our model is not working well for predicting counts under 100.

3.3. K-Fold Cross Validation:

Although both models have problems for predicting counts under 100, we want to compare their Mean Squared Errors to select the best model which has the smallest MSE. In order to compare their MSE results, we used k-fold cross validation technique. We divided our data to 10 folds, each has 720 observations for train and 80 for test set, and calculated MSE for each fold, for both the Negative Binomial models obtained in *Result 8* and *Result 10.*

Model	MSE
hr + temp+ hum + season.dummy3 + weather.dummy2	13437.06
hr + I(temp^2) + I(hum^2) + season.dummy3 + weather.dummy2	15189.87

Result 11: *Average MSE results for 2 different models calculated by cross validation.*

Findings of Research Question 2: *Is thre any significant difference between the average number of rental bikes for different weather situation?*

In this part, firstly we fit the model for checking the ANOVA table to compare average number of rental bikes for different weather situation. Secondly, we look at the box-plot based on categories of weather situation for finding which of the weather situation has different mean.

```
Call:  glm(formula = cnt ~ weathersit, family = poisson(link = "log"),
    data = data.2011)

Coefficients:
(Intercept)  weathersit2  weathersit3
     5.0489      -0.1338      -0.8034

Degrees of Freedom: 799 Total (i.e. Null);  797 Residual
Null Deviance:      94190
Residual Deviance: 90810        AIC: 95740
```

***Result 12**: Model for number of rental bikes depends on the weather situation.*

```
             Df   Sum Sq Mean Sq F value   Pr(>F)
weathersit    2   411292  205646   12.29 5.54e-06 ***
Residuals   797 13337042   16734
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

***Result 13:** ANOVA table for comparing means.*

By checking the ANOVA table, we reject null hypothesis, which is average number of rental bikes is equal for each weather type, since p-value than (5.54e-06) is less 0.05. We can conclude that at least one of the weather situation is different from others.

Now let's look the box plot which is on the other page.

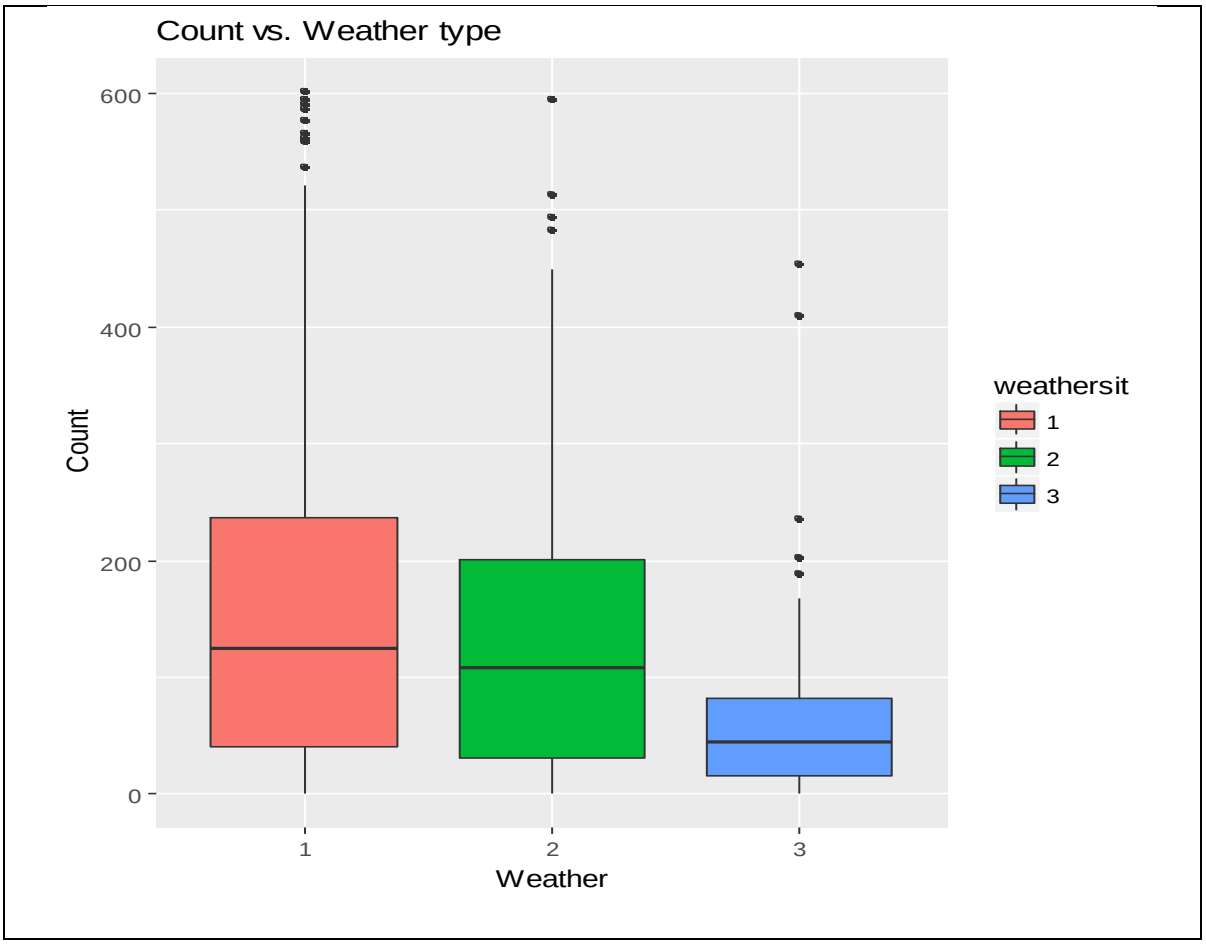

Figure 4: *Box plot of count versus categorical variable weathersit.*

As we explain in the data description section, weather situation 1 represents clear weather, weather situation. 2 represents misty weather and weather situation 3 represents snowy weather. By looking at the *Figure x*, we can say that there are outliers for all three-weather situation. We can easily see that average number of rental bikes is the highest for the clear weather and has the widest variety in weather situation. Average number of rental bikes decreases as weather situation gets worse. So, it seems to be there is a significant difference between the average number of rental bikes for different weather situation and different one is the weather situation 3 which is snowy weather.

Findings of Research Question 3: *Are there any relationship between humidity and temperature?*

In this part, firstly we check the correlation between humidity and temperature. Secondly, we draw scatter plot of Temperature vs Humidity with %95 confidence region for the best-fit regression line and try to understand relationship between humidity and temperature.

```
[1] -0.06930486
```

Result 14: *Correlation between humidity and temperature.*

By looking at the correlation *Result 14,* we can see that there is almost no relationship between humidity and temperature since correlation is too weak.

Now, let's look at the scatter plot of temperature vs humidity with %95 confidence region for the best-fit regression line.

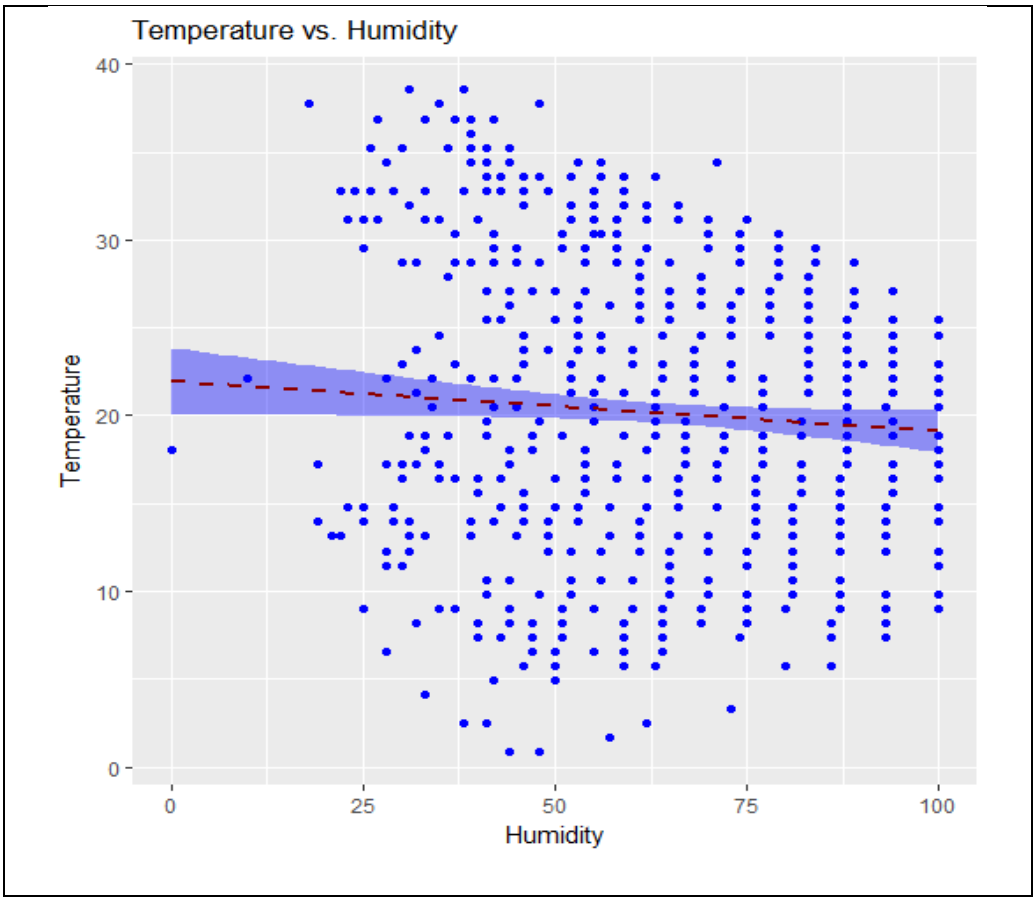

Figure 5: *Scatter plot of Temperature vs Humidity with %95 confidence region for the best-fit regression line.*

According to *Figure 5*, there is no relationship between temperature and humidity. The red line represents to best-fist regression line and shows the predicted values according the regression. The blue region around the red line shows the %95 confidence region for the best-fit regression line. So, we can see that actual values are far away from the predicted values and most of the actual values stay outside of the regression line.

So, we conclude that there is no relationship temperature and humidity.

Findings of Research Question 4: *Does wind speed decrease humidity?*

In this part, firstly we check the correlation between wind speed and humidity. Secondly, we draw scatter plot of Wind Speed vs Humidity with %95 confidence region for the best-fit regression line and try to understand relationship between wind speed and humidity.

```
[1] -0.2797574
```

Result 15: *Correlation between wind speed and humidity.*

By looking at the correlation *Result 15*, we can see that there is a weak negative relationship between wind speed and humidity.

Now, let's look at the scatter plot of wind speed and humidity with %95 confidence region for the best-fit regression line.

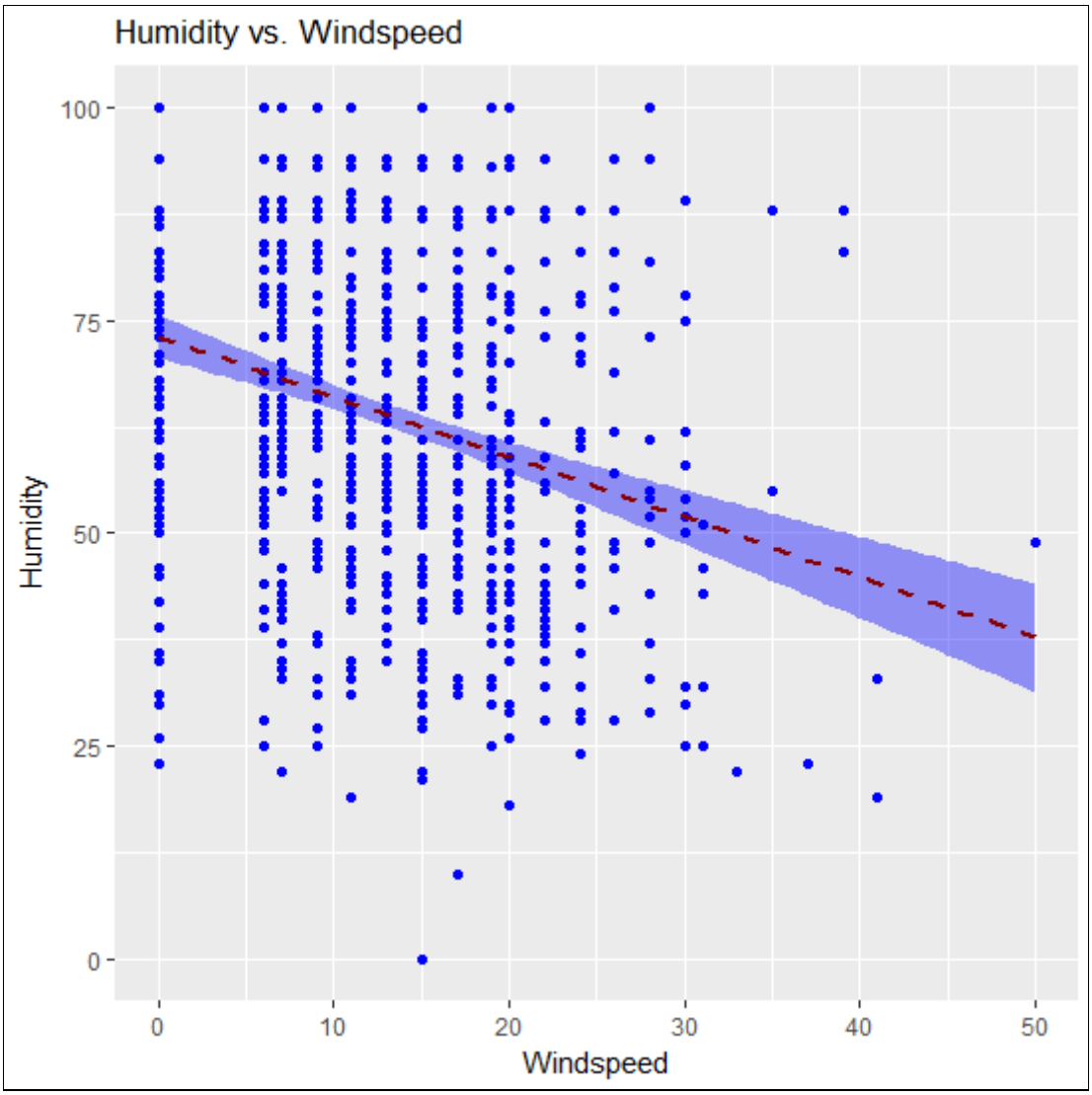

Figure 6: *Scatter plot of Humidity vs Wind Speed with %95 confidence region for the best-fit regression line.*

According to scatter plot, there is a weak relationship between wind speed and humidity. The red line represents to best-fist regression line and shows the predicted values according the regression. The blue region around the red line shows the %95 confidence region for the best-fit regression line. So, we can see that actual values are far away from the predicted values and most of the actual values stay outside of the regression line.

So, we conclude that there is a weak negative relationship between wind speed and humidity.

Findings of Research Question 5: *Are there any relationship between temperature and number of rental bikes(count)?*

In this part, firstly we check the correlation between temperature and number of rental bikes. Secondly, we draw scatter plot of Count vs Temperature with %95 confidence region for the best-fit regression line and try to understand relationship between temperature and number of rental bikes.

```
[1] 0.4531231
```

Result 16: *Correlation between temperature and number of rental bikes (count)*

By looking at the correlation *Result 16,* we can see that there is a positive relationship between temperature and number of rental bikes (count).

Now, let's look at the scatter plot of temperature and number of rental bikes (count) with %95 confidence region for the best-fit regression line.

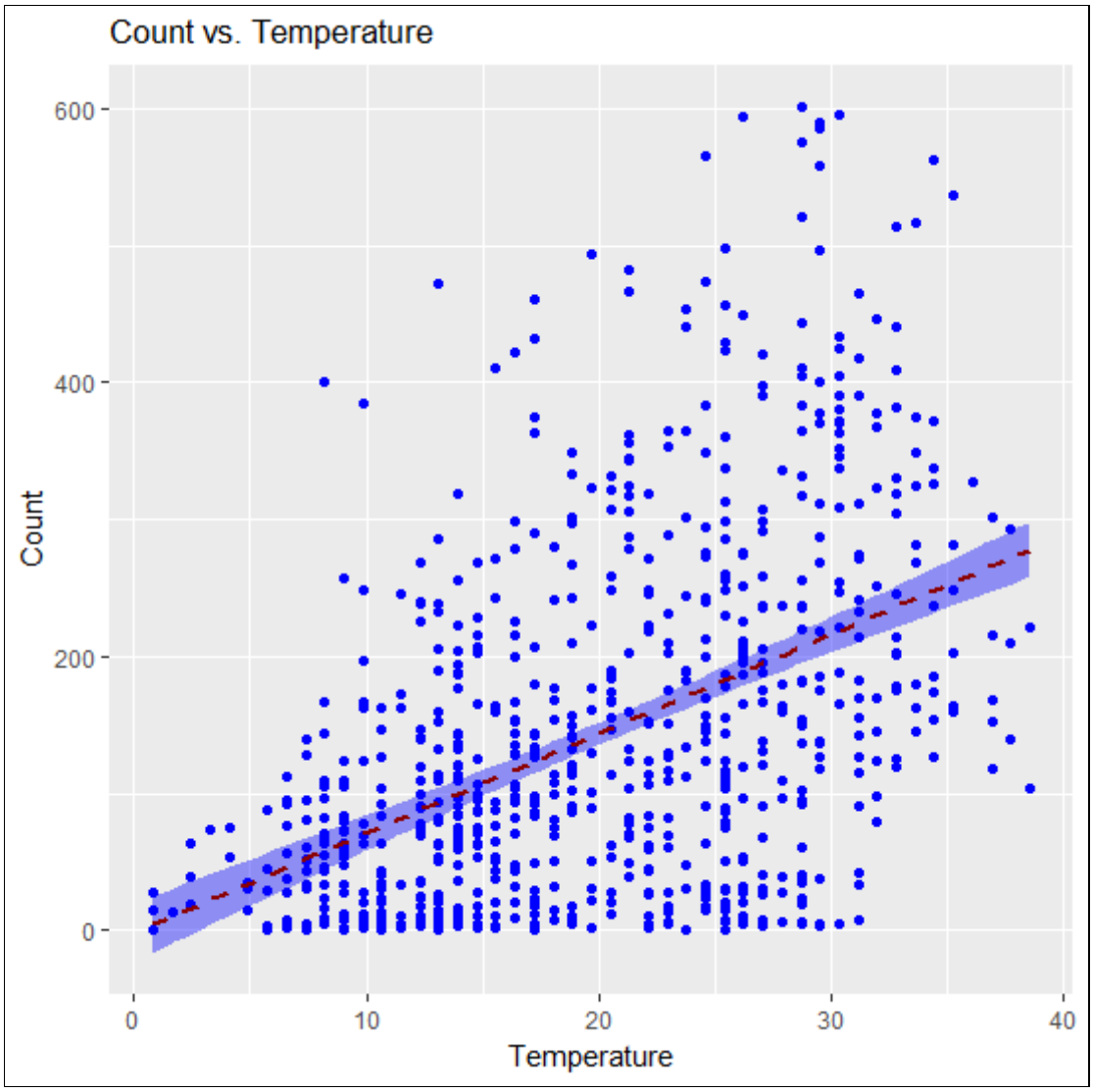

Figure 7: *Scatter plot of Count vs Temperature with %95 confidence region for the best-fit regression line.*

According to scatter plot, there is a positive relationship between temperature and number of rental bikes (count). The red line represents to best-fist regression line and shows the predicted values according the regression. The blue region around the red line shows the %95 confidence region for the best-fit regression line. So, we can see that actual values are far away from the predicted values and most of the actual values stay outside of the regression line.

So, we conclude that there is a positive relationship between temperature and number of rental bikes (count).

4. Discussion/Conclusion

In this project we try to find answers of 5 research questions. We used different techniques to answer our questions. The first question was "Which variables are likely to best predict the number of rental bikes(count)?". After checking assumptions, we used Poisson regression analysis for answering the first question since it suits well with count data. Firstly, we build a Poisson regression model, but we faced with over-dispersion problem. To eliminate this problem, we used negative binomial regression model. Thanks to negative binomial regression model we overcome over-dispersion problem. Although overcoming the over-dispersion problem, there were insignificant variables. We eliminated the insignificant variables and test our model. Unfortunately, our new model was not working well for predicting number of rental bikes under 100. In order to solve this problem, we fitted the model again with square of humidity and square of temperature (hum^2 and temp^2). Again, the model was not working well for predicting number of rental bikes under 100. After having second model, we used k-fold cross validation technique to choose best model. As a result of k-fold cross validation, the first model is chosen the best model. The first model is

$$\hat{Y} = e^{3.521976 + 0.07522*hr + 0.049407*temp - 0.005201*hum - 0.352361 *season.dummy3 - 0.657987*weather.dummy2}$$

The second question was "Is there any significance difference between the average number of rental bikes for different weather situation?" To answer this question, we fitted a poisson model with the average number of rental bikes (count) as a response variable and weather situation as explanatory variable. After building the model, we checked anova table for deciding whether there is a significant difference between mean of weather situations or not. As a result, we decided at least one of the weather situation has different mean. To find which one is different, we draw a box-plot based on categories of weather situations, which are clear, misty and snowy weather. The box-plot showed us that there is difference between the average number of rental bikes for different weather situation and different one is snowy weather.

The third question was "Are there any relationship between humidity and temperature?". To give an answer to this question, we checked correlation between humidity and temperature. The result showed that there is no relationship between humidity and temperature since correlation between variables was too weak. In order to be sure, we drew a scatter plot of

temperature vs humidity with the best-fit regression line. According to this result, the answer was same, there is no relationship between humidity and temperature.

The fourth question was "Does wind speed decrease humidity?". We used the same techniques we used in third question. Firstly, we find correlation between wind speed and humidity. There was weak negative relationship between these two variables. It seemed to be wind speed decreases humidity weakly. To get clear answer, we drew a scatter plot of wind speed vs humidity with the best-fit regression line. The scatter plot showed that there is weak negative relationship between wind speed and humidity.

The fifth question was "Are there any relationship between temperature and number of rental bikes(count)?". Again, same techniques, which are used in previous questions, can be applied in that question. The result of correlation showed that there is positive relationship between temperature and number of rental bikes(count). That means number of rental bikes increases as temperature increases. To be sure that our finding is true, we drew a scatter plot temperature and number of rental bikes(count). The scatter plot showed same finding, there is a positive relationship between temperature and number of rental bikes (count).

To sum up, we make our best prediction of the number of rental bikes with the model below,

$$\hat{Y} = e^{3.521976 + 0.07522*hr + 0.049407*temp - 0.005201*hum - 0.352361 *season.dummy3 - 0.657987*weather.dummy2}$$

We conclude that there is difference between the average number of rental bikes for different weather situation and different one is snowy weather, there is no relationship between humidity and temperature, there is weak negative relationship between wind speed and humidity and there is a positive relationship between temperature and number of rental bikes.

References

1. Negative Binomial Regression | R Data Analysis Examples. (n.d.). Retrieved from https://stats.idre.ucla.edu/r/dae/negative-binomial-regression
2. Poisson Models for Count Data. [online] Available at: http://data.princeton.edu/wws509/notes/c4.pdf.
3. Zeileis, A. Regression Models for Count Data in R. Retrieved September 9, 2016 from: https://cran.r-project.org/web/packages/pscl/vignettes/countreg.pdf
4. Ggplot2 scatter plots: Quick start guide - R software and data visualization. (n.d.). Retrieved from http://www.sthda.com/english/wiki/ggplot2-scatter-plots-quick-start-guide-r-software-and-data-visualization

Appendices

```
#data preparation
getwd()
setwd("C:/Users/orcun.oltulu/Desktop/proje")
data <- read.csv("hour.csv",T,sep=",")
head(data)
data.2011 <- data[which(data$yr==0),]
dim(data.2011)
##2011##
N <- 800
set.seed(6)
random <- sample(1:8645,replace=FALSE,size=800)
data.2011 <- data.2011[random,-c(1,2,4,5,8,12,15,16)]
dim(data.2011)
attach(data.2011)
#creating dummy variables
#dummies for season variable
season.dummy1 <- ifelse(season==2,1,0)
season.dummy2 <- ifelse(season==3,1,0)
season.dummy3 <- ifelse(season==4,0,1)
new.data.2011 <- cbind(data.2011,season.dummy1,season.dummy2,season.dummy3)
new.data.2011 <- new.data.2011[,-1]
head(new.data.2011)
#dummies for weathersit
weather.dummy1 <- ifelse(weathersit==2,1,0)
weather.dummy2 <- ifelse(weathersit==3,1,0)
new.data.2011 <- cbind(new.data.2011,weather.dummy1,weather.dummy2)
new.data.2011 <- new.data.2011[,-4]
head(new.data.2011)
class(new.data.2011)
detach(data.2011)
```

```
attach(new.data.2011)
new.data.2011$holiday <- as.factor(holiday)
new.data.2011$workingday <- as.factor(workingday)
new.data.2011$season.dummy1 <- as.factor(season.dummy1)
new.data.2011$season.dummy2 <- as.factor(season.dummy2)
new.data.2011$season.dummy3 <- as.factor(season.dummy3)
new.data.2011$weather.dummy1 <- as.factor(weather.dummy1)
new.data.2011$weather.dummy2 <- as.factor(weather.dummy2)
new.data.2011$temp <- new.data.2011$temp * 41
new.data.2011$hum <- new.data.2011$hum * 100
new.data.2011$windspeed <- new.data.2011$windspeed * 67
attach(new.data.2011)
head(new.data.2011)

##test_set##
kalan <- data[-random,-c(1,2,4,5,8,12,15,16)]
dim(kalan)
set.seed(7)
random.test <- sample(1:7845,replace=FALSE,size=800)
test <- data[random.test,-c(1,2,4,5,8,12,15,16)]
dim(test)

#creating dummy variables for test set
#dummies for season variable
season.dummy1 <- ifelse(test$season==2,1,0)
season.dummy2 <- ifelse(test$season==3,1,0)
season.dummy3 <- ifelse(test$season==4,0,1)
new.test <- cbind(test,season.dummy1,season.dummy2,season.dummy3)
new.test<- new.test[,-1]
head(new.data.2011)
#dummies for weathersit
weather.dummy1 <- ifelse(test$weathersit==2,1,0)
weather.dummy2 <- ifelse(test$weathersit==3,1,0)
new.test <- cbind(new.test,weather.dummy1,weather.dummy2)
new.test <- new.test[,-4]
head(new.test)
dim(new.test)

new.test$holiday <- as.factor(new.test$holiday)
new.test$workingday <- as.factor(new.test$workingday)
new.test$season.dummy1 <- as.factor(new.test$season.dummy1)
new.test$season.dummy2 <- as.factor(new.test$season.dummy2)
new.test$season.dummy3 <- as.factor(new.test$season.dummy3)
new.test$weather.dummy1 <- as.factor(new.test$weather.dummy1)
new.test$weather.dummy2 <- as.factor(new.test$weather.dummy2)
#they were normilized, we converted them to original.
new.test$temp <- new.test$temp * 41
```

```
new.test$hum <- new.test$hum * 100
new.test$windspeed <- new.test$windspeed * 67

##Assupmtion Checking##
#count data
summary(cnt)
cnt[sample(1:800,replace=FALSE,size=50)]

#poisson distribution
hist(cnt,col="dark green",prob=TRUE,main="Histogram of Count",xlab="Count")
lines(density(cnt),lwd=3)

#mean = variance
mean(cnt);var(cnt)

#multicollinearity checking
install.packages("usdm")
library(usdm)
data.2011 <- data.frame(data.2011)
data.2011<-data.2011[,-5] # omitting categorical variable weathersit
data.2011<-data.2011[,-8] # omitting the response varieble cnt
vif(data.2011)

##models##
model.p1 <- glm(cnt ~ . , data=new.data.2011, family = poisson(link = "log"))
summary(model.p1)

model.p2 <- glm(cnt ~ hr + temp + hum + season.dummy1 + season.dummy2 +
                                        season.dummy3 + weather.dummy2,
                data=new.data.2011, family = poisson(link = "log"))
summary(model.p2)

##glm.nb##
library(MASS)

model.nb <- glm.nb(cnt ~ . , data=new.data.2011)
summary(model.nb)

model.nb2 <- glm.nb(cnt ~ hr + temp + hum + season.dummy1 + season.dummy3 +
weather.dummy2,
            data = new.data.2011, link = "log")
summary(model.nb2)

model.nb3 <- glm.nb(cnt ~ hr + temp + hum + season.dummy3 + weather.dummy2,
            data = new.data.2011, link = "log")
summary(model.nb3)
#after eliminating the insignificant variables.
```

```
model.nb7 <- glm.nb(cnt ~ hr + I(temp^2) + I(hum^2)  + season.dummy3 +
weather.dummy2 ,
                data = new.data.2011, link = "log")
summary(model.nb7)

##predictions##
prediction.1 <- predict(model.nb3, newdata=new.test)
par(mfrow=c(1,2))
hist(cnt,col="dark green", main="Original Data",xlab="counts")
hist(exp(prediction.1),col="red", main="Predicted",xlab="counts")

prediction.2 <- predict(model.nb7, newdata=new.test)
par(mfrow=c(1,2))
hist(cnt,col="dark green", main="Original Data",xlab="counts")
hist(exp(prediction.2),col="red", main="Predicted",xlab="counts")

##Cross Validation##
N <- 800
subset.train <- list()
for(j in 1:10)
  subset.train[[j]] <- setdiff(1:N, ((j-1) * N / 10 + 1):(j * N) / 10)

CV.nb <- c()
for(j in 1:10){
  reg.nb <- glm.nb(cnt ~ hr + temp + hum + season.dummy3 + weather.dummy2 ,
                subset = subset.train[[j]])

  test.nb <- data.frame(hr[-subset.train[[j]]],temp[-subset.train[[j]]],
                    hum[-subset.train[[j]]],season.dummy1[-subset.train[[j]]],
                    season.dummy3[-subset.train[[j]]], weather.dummy2[-subset.train[[j]]])
  colnames(test.nb) <-
c("hr","temp","hum","season.dummy1","season.dummy3","weather.dummy2")
  prediction.nb <- exp(predict(reg.nb, newdata=test.nb))
  CV.nb <- c(CV.nb, mean((prediction.nb - cnt[-subset.train[[j]]])^2))
}
CV.nb
av.mse.nb <- mean(CV.nb);av.mse.nb

CV.nb2 <- c()
for(j in 1:10){
  reg.nb2 <- glm.nb(cnt ~ hr + I(temp^2) + I(hum^2) + season.dummy3 +
weather.dummy2 ,
                subset = subset.train[[j]])

  test.nb2 <- data.frame(hr[-subset.train[[j]]],temp[-subset.train[[j]]],
                    hum[-subset.train[[j]]],season.dummy1[-subset.train[[j]]],
```

```
                season.dummy3[-subset.train[[j]]], weather.dummy2[-subset.train[[j]]])
  colnames(test.nb2) <-
c("hr","temp","hum","season.dummy1","season.dummy3","weather.dummy2")
  prediction.nb2 <- exp(predict(reg.nb2, newdata=test.nb2))
  CV.nb2 <- c(CV.nb2, mean((prediction.nb2 - cnt[-subset.train[[j]]])^2))
}
CV.nb2
av.mse.nb2 <- mean(CV.nb2);av.mse.nb2

###research question-2
#Which type of weather had a most influence on the number of rental bikes?

new.data.2011$weathersit<-as.factor(new.data.2011$weathersit)
names(data.2011)
library(ggplot2)

ggplot(data.2011, aes(x=weathersit, y=cnt, fill=weathersit)) +
  geom_boxplot(outlier.shape=19, outlier.size=1 ) + ggtitle("Count vs. Weather type")
+
  xlab("Weather") + ylab("Count")
names(data.2011)
model.weather <- glm(cnt ~weathersit , data=data.2011, family = poisson(link="log"))
model.weather
summary(aov(model.weather))

###research question-3
#Are there any relationship between humidity and temperature?

cor(new.data.2011$hum, new.data.2011$temp)

ggplot(new.data.2011, aes(x=hum, y=temp))+  geom_point(shape=16, color="blue")+
  geom_smooth(method=lm, linetype="dashed",
          color="darkred", fill="blue") + ggtitle("Temperature vs. Humidity") +
  xlab("Humidity") + ylab("Temperature")

###research question-4
#Does wind speed decrease humidity?

cor(new.data.2011$windspeed, new.data.2011$hum)

ggplot(new.data.2011, aes(x=windspeed, y=hum))+  geom_point(shape=16,
color="blue")+
  geom_smooth(method=lm, linetype="dashed",
          color="darkred", fill="blue") + ggtitle("Humidity vs. Windspeed") +
  xlab("Windspeed") + ylab("Humidity")

###research question-5
```

#Are there any relationship between temperature and number of rental bikes(count)?

```
cor(new.data.2011$temp, new.data.2011$cnt)

ggplot(new.data.2011, aes(x=temp, y=cnt))+  geom_point(shape=16, color="blue")+
  geom_smooth(method=lm, linetype="dashed",
              color="darkred", fill="blue") + ggtitle("Count vs. Temperature") +
  xlab("Temperature") + ylab("Count")
```