STATISTICAL ANALYSIS OF BIKE SHARING DATA

İbrahim Hakkı ERDURAN*

Orçun OLTULU*

Middle East Technical University, *Ankara, Turkey

INTRODUCTION

The bike sharing data is collected from bike sharing system which name is Capital Bike Sharing (CBS) at Washington, D.C., USA in 2011. Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues. The number of major cities that are becoming bike-friendly is growing day-by-day. It is expected that in a near future, most major cities provide this service along their other public transport services. 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.

METHODOLOGY

Proposed method for the main research question is Poisson Regression method since we try to estimate the total number of rental bikes. Furthermore, the Poisson Regression model for each observation is given by:

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

Since our data does not satisfy assumptions of Poisson Regression, we use Negative Binomial Regression. The Negative Binomial Regression Model for each observation is given by:

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

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.

RESULTS & FINDINGS

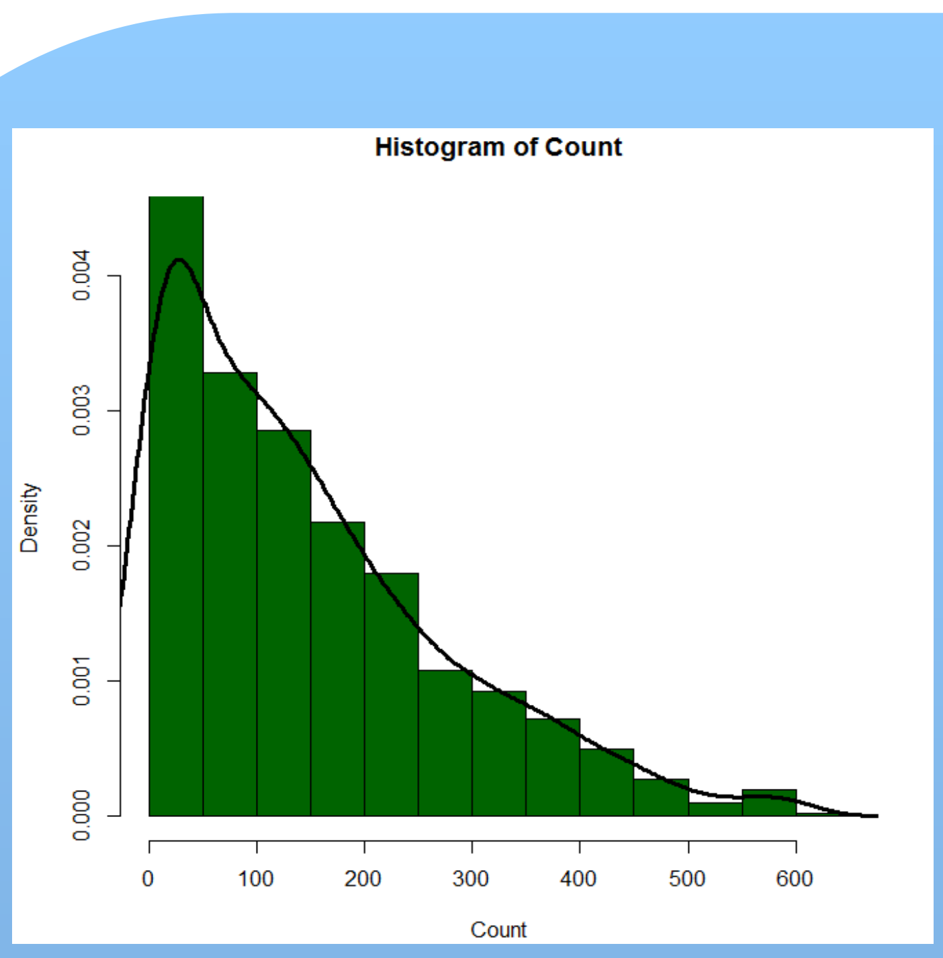
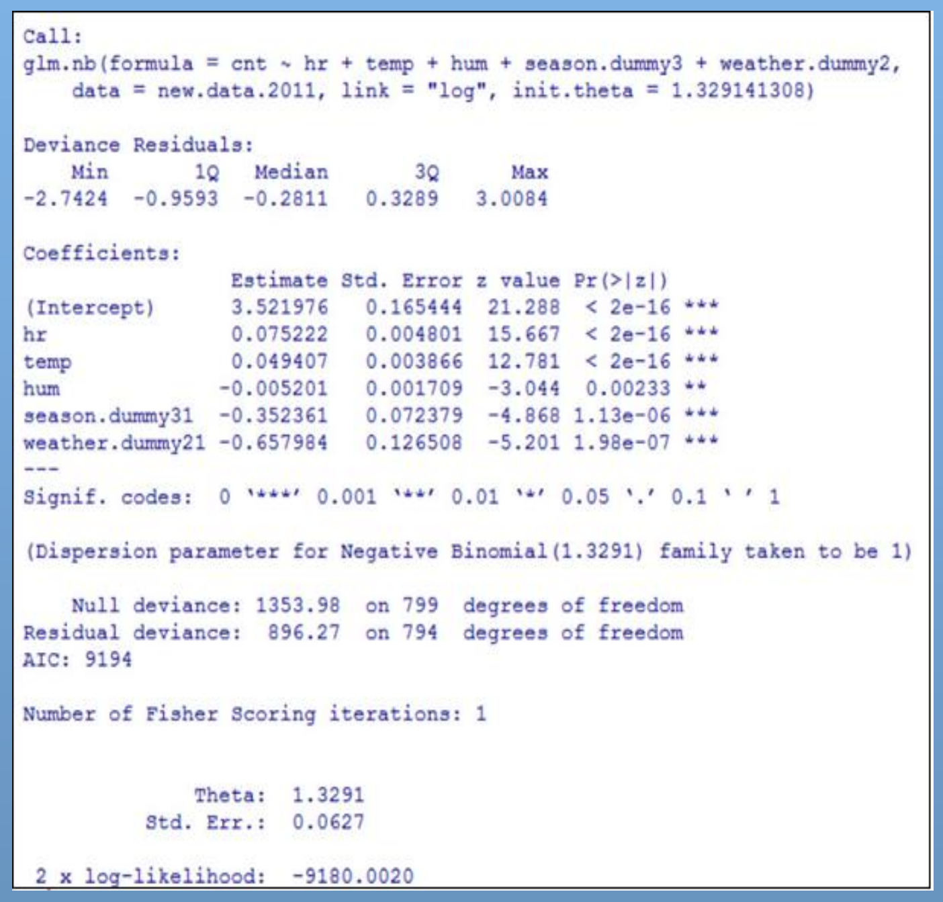


Figure 1: *Histogram of response variable (cnt)*

By looking at the *Figure 1*, it is easy to say that response variable count has Poisson distribution. For this reason, we try Poisson regression method to estimate number of rental bikes after checking assumptions. Unfortunately, this result have over-dispersion problem since square root of residual deviance/df, sqrt(5498/788)=8.350507, is greater than 1. To solve this problem, we use Negative Binomial regression method. After applying Negative Binomial regression method and eliminating insignificant variables, we find the model whose R output in the below.



```
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 1: *Summarizing the fitted Negative Binomial model with only significant explanatory variables.*

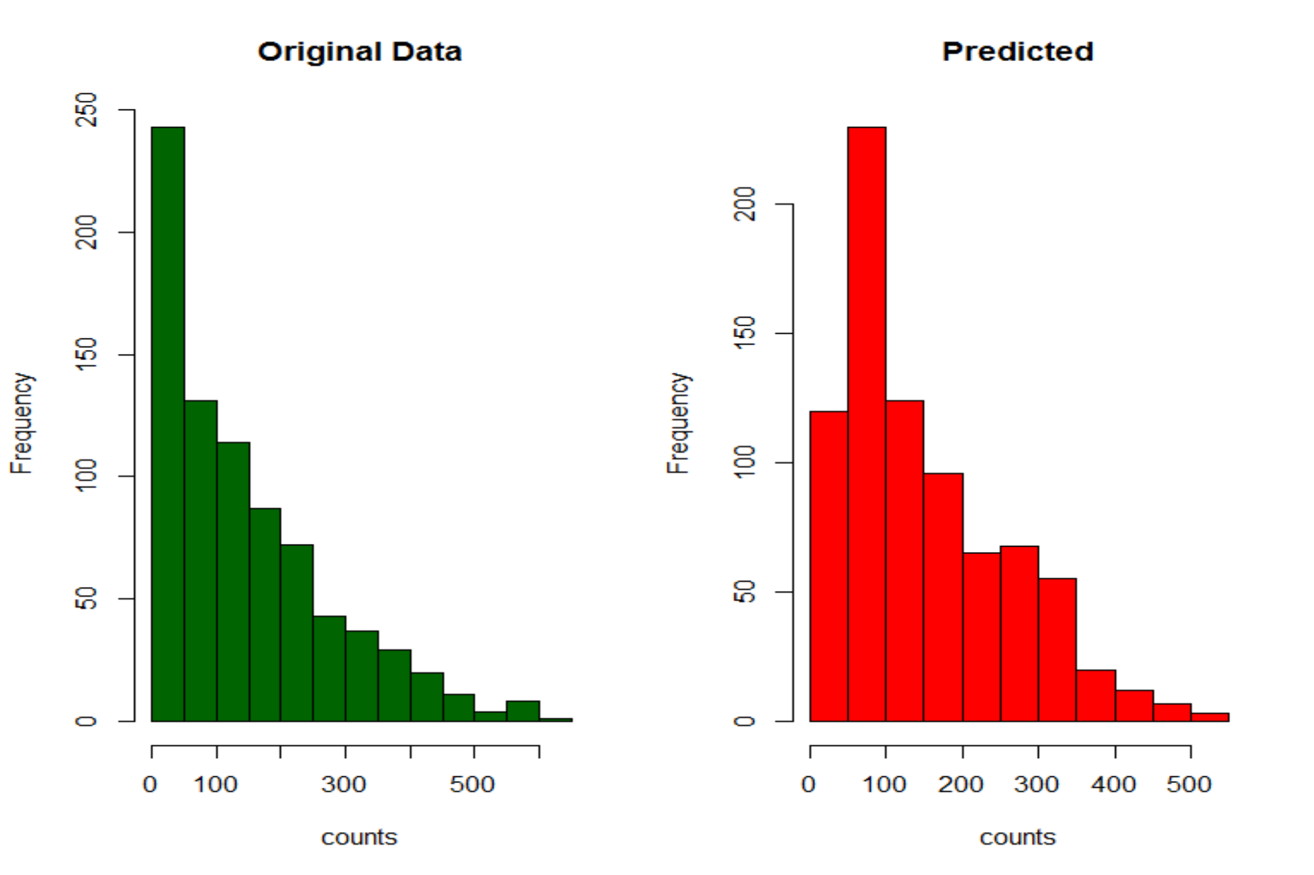


Figure 2: *Comparison of Original data and Prediction.*

Unfortunately, as it can be seen in Figure 2, still our model is not working well for predicting counts under 100. For this reason, we try to find different model with interaction terms and other terms such as square of humidity and square of temperature, but after comparing models with k-fold cross validation, the model which is found by Negative Binomial regression method is chosen the best predicted model.

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

CONCLUSION

To sum up, 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.

$$\hat{Y} = \exp(3.521976 + 0.07522*hr + 0.049407*temp – 0.005201 * hum – 0.352361*season.dummy3 – 0.657987*weather.dummy2)$$

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. 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

2. Negative Binomial Regression | R Data Analysis Examples. (n.d.). Retrieved from https://stats.idre.ucla.edu/r/dae/negative-binomial-regression

3. Poisson Models for Count Data. [online] Available at: http://data.princeton.edu/wws509/notes/c4.pdf.

4. 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