Predicting Hourly Bike Rentals

Dhaval Delvadia

DePaul University

*Abstract:*

Bike-sharing systems are available in many major cities for their citizens to commute and visitors to sightseeing without having to rely on other modes of transportation. In this article, we try to predict total amount of bike rentals at certain hours during a given day at a bike-sharing system in a Washington, D.C. Figuring this out, the bike rental owner can determine ahead of time how many bikes to stock for a given day or their revenue for a given day or hour. We took Multiple Linear Regression analysis approach to predict bike-sharing demand at a specific hour from different independent variables. In order to find the best model under multiple Linear Regression analysis, the dataset was split into two, training and test validation, respectively. On the training set, the stepwise and mallows’, cp, methods were employed. The best model was found from these methods and it was tested on test dataset.

1. **Introduction:**

Heavy street traffic in busy cities with high cost of traditional modes of transportation makes rental bike share program an attractive alternative for travel or sightsee a city [1]. Bike-sharing program is an environmentally friendly form of travel and sightsee the city. Also, bike-sharing services have gained considerable traction in U.S. in the past decade [2]. Therefore, it’s really important for the owners of the bike-sharing program to determine the demand per hour level for a given day based on all the environmental factors. In this study, we try to predict the demand for bike per hour for Capital Bikeshare Program implemented in Washington, D.C.

1. **Methodology:**

The data was obtained from an online website called UCI Repository Machine Learning Repository [3]. This dataset contains hourly rental data for the years 2011 and 2012. The dataset contained total 17,379 observations. The original data contained 9 features and 3 labels classified as follows:

|  |  |
| --- | --- |
| **Table 1: Features and Labels for the Bike Sharing Dataset** | |
| **Independent Variables:** | |
| Date and Time | Format is MM/DD/YYYY HH:MM |
| Season | Takes 4 values: 1 = spring, 2 = summer, 3 = fall, 4 = winter |
| Holiday | 1 = yes, 0 = no |
| Working day | 1 = yes, 0 = no |
| Weathersit | 1: Clear, few clouds, partly cloudy  2: Mist + cloudy, mist + broken clouds, mist + few cloud  3: Light snow, light rain + thunderstorm + Scattered clouds, light rain + scattered clouds.  4: Heavy Rain + Ice Pallets+ thunderstorm + mist, snow + fog |
| Temp | Normalized temperature in Celsius. The values are divided to 41 (max) |
| Atemp | Normalized feeling temperature in Celsius. The values are divided to 50 (max) |
| Hum | Normalized humidity. The values are divided to 100 (max) |
| Windspeed | Normalized wind speed. The values are divided to 67 (max) |
| **Dependent Variables:** | |
| Casual | count of casual users |
| Registered | count of registered users |
| Cnt | count of total rental bikes including both casual and registered |

As seen in **Table 1** above, the date and time of each sample was given as one attribute which made the attribute difficult to interpret and incorporate into a model. Therefore, the date and time was split into 4 separate features: year, month, day, and time. This made the total number of features 13 - making it easier to integrate the features into the model.

From here, every member of the team started to come up with their own best model and tested their model on each of their respective test dataset.

The categorical independent variables, season and weather, were transformed into dummy variables. The ‘temp’, ‘atemp’, ‘hum’, and ‘windspeed’ attributes were changed to raw temperature, feel temperature, humidity, and wind speed. Since the casual and the registered dependent variables are combined to provide the total bike rental count as the ‘cnt’ variable, they were not evaluated to build model. Since some hours clearly had more demand than other hours, we found that the weekdays, the peak hours for bike rentals were between 7-9 am and 5-7 pm and on weekends, the peak hours were anywhere between 10 am - 6 pm. They were classified as the ‘week\_peak’ hours and the ‘weekend\_peak’ hours variables [1].

In order to train and test models, the dataset were split into 70% training and 30% test set. The histogram of the output variable, count, was plotted to visually determine normality of the output. The histogram is skewed to the right. In order to normally distribute the output variable, log, square root, inverse, square, and cube transformation were applied. Only the square root transformation came close to the being normally distributed, and thus it was employed as new y-variable. See **Figure 1** in **Appendix A** for before and after transformation.

After transforming the output variable, the scatterplot of the dependent variables and independent variable was plotted in SAS. **Figure 2** in **Appendix A** shows limited matrix scatter plot of square root of count as independent variables. From this, we can see multicollinearity among ‘raw\_temp’ and ‘raw\_atemp’. **Figure 3 in Appendix A** shows scatter plot of the other variables. Only the hour variable is linear with some high picks during some periods of the day. Other variables are not linear. They are binary. We also confirmed all these by reviewing the Pearson correlation among all variables.

From above the Model 1 was developed based on all independent variables versus square root of count as y-variable except for ‘week\_peak’ and ‘weekend\_peak’. See SAS output for **Model 1 under** **Appendix A**. Based on the output, we can see the multicollinearity or Variance Inflation Factor (VIF) is higher than 10 between ‘raw\_temp’ and ‘raw\_atemp’. We can also see that the residual plot does have constant variance or independence. Also, the normal probability plot is not completely linear. The ends are curved. The F-value was noted as 671, RMSE was 4.805, R2 = 0.4844, and Adj- R2 = 0.4837. Based on this information, the highest VIF value variable, ‘raw\_temp’, was removed. Also, the alternate hypothesis did not meet due to p-value higher than 0.05 for ‘mnth’, ‘d\_heavyRain’, ‘day’, and ‘workingday’ variables. Therefore, highest p-value variables were removed one at a time and model was re-run until all remaining variables p-values were less than 0.05. The final version of re-run model was noted as **model 2** in **Appendix A.**

Model 2 had no issue with multicollinearity or the null hypothesis. The F-value increased to 2618.1, RMSE lowered to 3.340. We added the ‘week\_peak’ and ‘weekend\_peak’ independent variables and the R2 increased to 0.751 and adj- R2 also increased to 0.751. Model 3 was run with stepwise, adj-r2 and mallows’ Cp selection methods to find out the best variables from Model 2 variables. All methods provided same output variables as Method 2. Also, according to Standardized Estimate value, the best predictors are ‘Week\_peak’, ‘weekend\_peak’, ‘raw\_atemp’, and ‘hr’.

The final model was new\_y = bike count per hour = 0.9986 + 0.3126\*hr + 2.65\*yr + 1.1\*d\_summer + 0.39\*d\_fall + 2.3\*d\_winter – 0.9\*holiday + 0.1\*weekday + 0.2\*workingday + 0.24\*d\_mist – 1.8\*d\_LightSnow + 0.25\*raw\_atemp – 0.06\* raw\_hum + 8.7\*week\_peak + 6.4\*weekend\_peak.

This model was cross validated on the test data. After validating on the 30% of the test data, we received Mean Absolute Error, MAE equal to 2.745. The Root Mean Square Error, RMSE is equal to 3.37. The correlation, R2 is 0.753 and CV-R2 is 0.002. Since MAE, RMSE, and R2 are equal to the training set and CV-R2 is less than 0.3, this model seems to be adequate.

After getting this model and conducting cross validation, this model can be improved since the best prodictors are ‘hour’, ‘raw\_atemp’, ‘week\_peak’ hours and ‘weekend\_peak’ hours. Maybe fitting the model only with these predictors may lead to same or better RMSE, R2, adj-R2, and CV-R2 outputs.

1. **Analysis Results and Findings:**

The final model contains following MAE, RMSE, R2, and CV-R2.

|  |  |
| --- | --- |
| # of Ind. Var. | 14 |
| MAE | 2.745 |
| RMSE | 3.37 |
| R2 | 0.753 |
| CV-R2 | 0.002 |

1. **Future Work:**

Using different combinations of models from each team members, we found hour to be the most predictive feature for the model and weather variable contributed little bit. We wish to analyze this model even further using advance algorithms such as Time Series Analysis, Random Forest (RF) and Decision Tree Model specifically Conditional Inference Trees (CTree). May be there is another model even better at predicting the bike rental per hour with far fewer predictors.

1. **References:**

[1] Jimmy Du, Rolland He, Zhivko Zhechev, “Forecasting Bike Rental Demands”, <http://cs229.stanford.edu/proj2014/Jimmy%20Du,%20Rolland%20He,%20Zhivko%20Zhechev,%20Forecasting%20Bike%20Rental%20Demand.pdf>

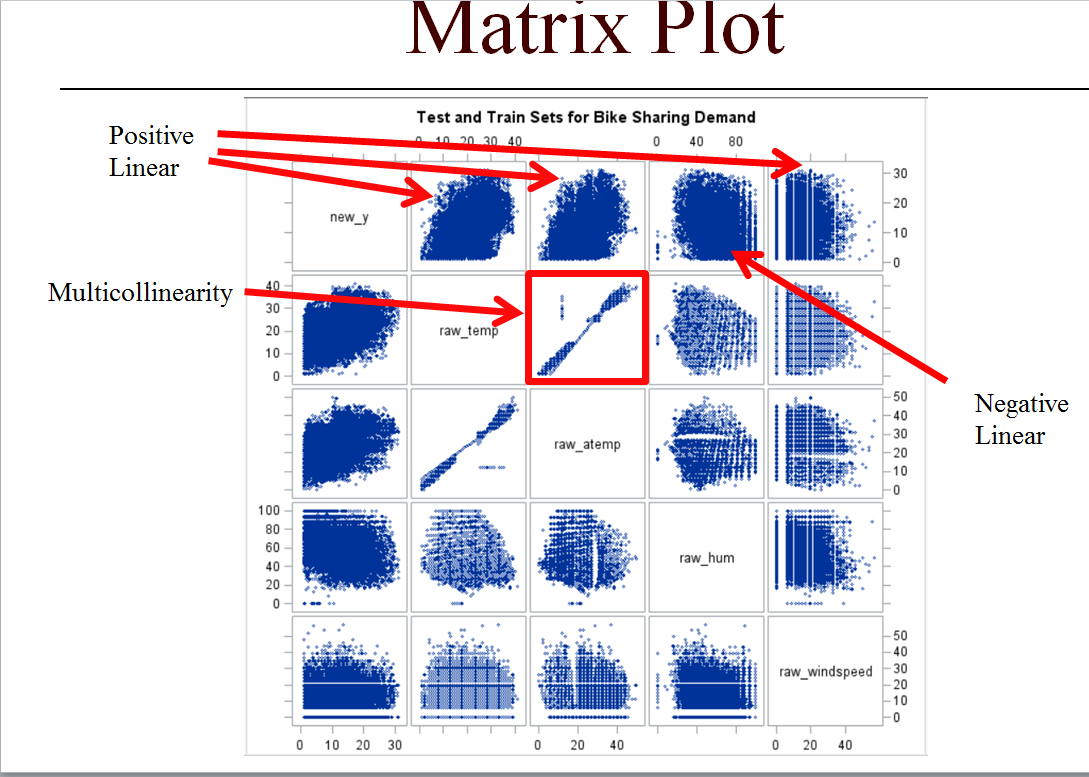
[2] Bill Chappell. Firm buys big bike-share service; expansion and higher rates seen, 2014.

[3] Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi:10.1007/s13748-013-0040-3.

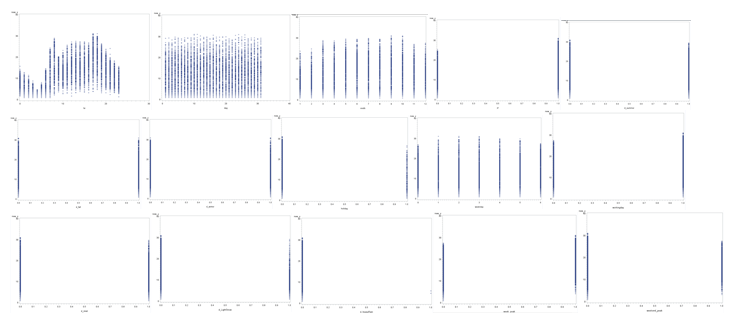
**Appendix A:**



**Figure 1: Histogram of output variable, count before and after transformation**



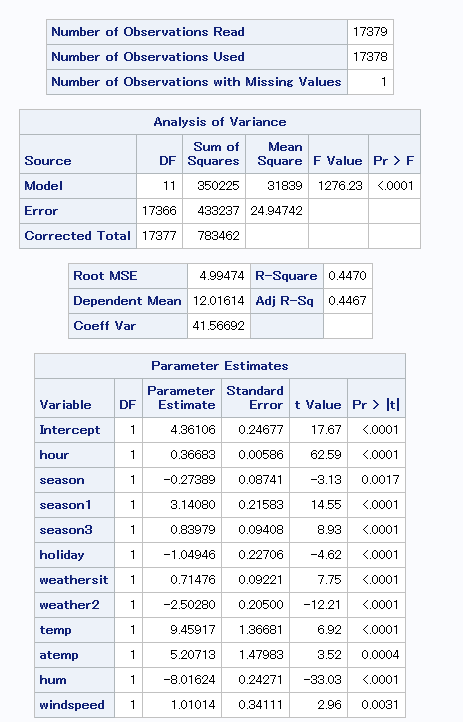
**Figure 2: Limited Matrix scatter plot**



**Figure 3– Scatter plot of new\_y vs each dependent variables**

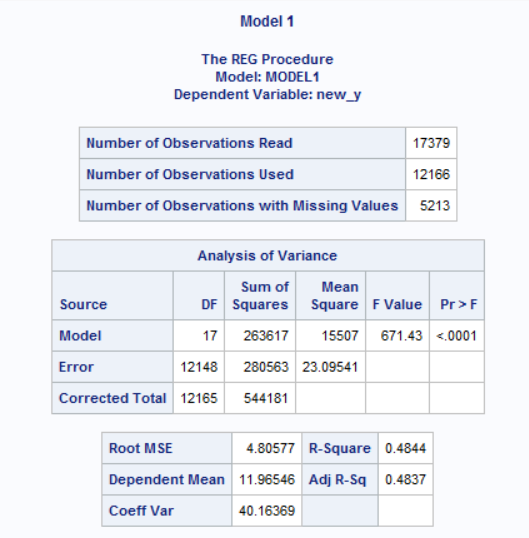
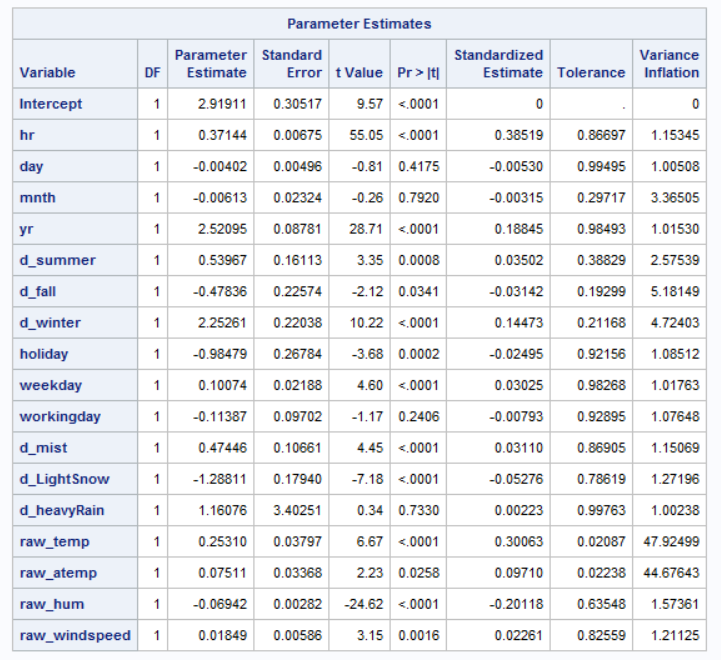
****

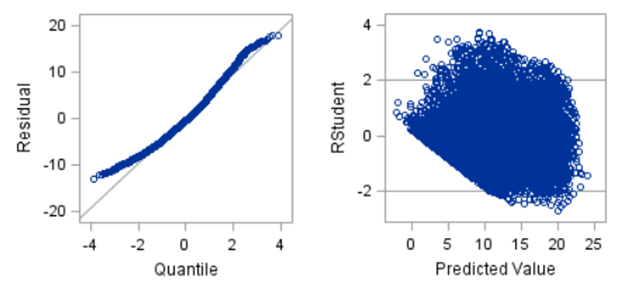
**Cp Selection method results**

****

**Regression Analysis**

**SAS Output Model 1:**



**SAS Output Model 2:**