Daily Bike Sharing Predictions:

Consultation for Company Decision Makers on Marketing Strategies

**Introduction**

Bike sharing systems are readily available in many cities in the States, especially the areas with heavy transportation burdens like San Francisco bay area, New York, Boston, D.C. et al. This new type of public transportation greatly facilitates people’s ordinary life by making it cheap and easy to rent and return the bikes. Our project took the daily rental data of Capital bikeshare system from the D.C. area between 2011 and 2012, hoping to identify important factors that affect the daily rental bike counts and construct a model that works fairly well to predict the daily rental count in the future. With such predictors and model by hand, the company decision makers can adjust their marketing strategies accordingly. For example, more maintenance staff and custom services should be assigned on the potential busy days.

**Data Description**

Dataset is found on the UCI Machine Learning Repository website, with 731 observations (rows) and 16 variables (columns). The target response variable the total daily count, including both casual user count and registered user count. Boxplot of the total count shows a huge spread, with minimal 2 count per day and maximal 8714 count per day (see Appendix A for the boxplot). Other variables include the date (year, season, month, weekday) and weather conditions (temperature, humidity, weather situations, wind speed). Except season and weather situations are categorical, all other variables are numeric. The weather situation is divided into three categories: 1 for clear or partly cloudy, but no precipitation; 2 for mist with clouds; 3 for light precipitations; and 4 for sever weathers like heavy rains, ice pallets etc. However the severe weather was not observed in our dataset, cuz D.C. area didn’t have severe weather conditions from 2011 to 2012.

Among all the predictors, we choose 6 variables of our interests: season, holiday/weekend, weather situation, temperature, humidity, and wind speed to construct our regression model.

**Statistical Analysis and Result**

After plotting each individual predictors with the total daily count (Appendix B), we found that temperature and humidity showed some sort of curved relationships to the response variable (total count), suggesting that a quadratic model might be more appropriate. This confirms our common sense as people might choose to ride a bike between a more comfortable temperature and humidity range. Any extreme conditions beyond that range, either above or below, keeps people from using this bike sharing service.

With the thoughts above, we first built a polynomial model that include the second order of temperature and humidity, plus all the other variables like season, weather situation, wind speed and holiday/weekend. It turned out that the holiday/weekend variable has a large p value of 0.2469, and it was not significant in our model (Appendix C). With this information, we updated our model by dropping the holiday/weekend variable while keeping all other terms of our initial model. This time, our model shows very small p values for all the predictors in our model (Appendix C), meaning that all the variables we included in this model are important predictors for our daily total bike counts.

Our final model is total count = 4738.9 – 15107.3\*temperature^2 + 27359.2\*temperature – 6501.2\*humidity^2 – 14248.3\*humidity – 3969.3\*wind speed + 595.7\*season2 + 768.5\*season3 + 1013.7\*season4 – 115.9\*weather situation2 – 1174.1\*weather situation3, among which season 2 to 4 and weather situation 2 to 3 are dummy variables for season and weather situation which can only has the value of 0 or 1. The second order terms of temperature and humidity both have negative coefficients, indicating that the curve opens downwards, thus the highest count happens only for optimal ranges. This again confirms our common sense and the curves presented in Appendix B. Wind speed has a negative slope coefficient, which means the rental count will decrease with increase of wind speed. Weather situations 2 and 3 both have negative slopes, suggesting reduced rental counts when the weather situations are bad. It seems that season 4 has the highest bike count over other three seasons, which could help company decision makers to adjust the marketing strategies seasonally.

To validate our regression model, a few assumptions should be check: data normality, constant variances, linearity of response and predictors, and outliers (results seen Appendix D). Shapiro-Wilk normality test showed a W value of 0.975, which is quite large, thus we claim our data is normally distributed even the p value is very small. Our model failed the non-constant variance test though (p < 0.05), meaning that our data has large variability or spread. Residual vs each predictor plots showed straight lines for quadratic temperature, quadratic humidity, and wind speed, all of which are what we want. The medians and spreads are similar in the boxplots for season variable, but weather situation 3 has a significant smaller spread which again shows the large variability of our dataset. The residual vs fitted value plot didn’t give a straight line because of the non-constant variances of the dataset, which is the only unmet model assumptions in our case. To check if any outliers exist in our dataset, we ran the outlier test which no significant outliers; we also check the influence points, hoping that the residual vs fitted value plot might look better after dropping the most influence point (69th row in our data). The results are shown in Appendix D, no obvious changes after dropping the influential points. This result is not surprising as our problem comes from the large variability, and dropping a few points is unlikely to make the variability smaller.

**Discussion**

One nagging question about our model is the holiday/weekend term. Our initial thoughts and intuition tell us that this predictor should be very important to the bike count prediction. However, with a p value of 0.25, this predictor is not significant at all. As a result, it was dropped from our final model. To explain this contradiction, we changed our response variable (total count) into the casual user count and register user count respectively. The boxplots of these two separate counts by holiday/weekend factor are attached in the Appendix E. From the boxplot we can see that more casual usage happens on holidays/weekends, whereas more registered usage is on working days. This suggests that the users from these two groups might use the bike sharing service for different purposes: casual users are probably visitors and use the bike to travel around the city on holidays and weekends, and registered users are probably using bikes to commute to work or subway stations on working days. To further prove this usage difference is the reason that holiday/weekend is not significant in our total model, we ran our regression models on casual user count and registered user count respectively on all the predictors including holiday/weekend (results shown in Appendix E). This time, both slopes for holiday/weekend are significant. The comparison can be found in the table below.

|  |  |  |
| --- | --- | --- |
|  | Slope Coefficient for holiday/weekend | p value |
| Casual user model | 812.24 | <2e-16 |
| Registered user model | -923.92 | <2e-26 |

Casual user model has a positive slope for holiday/weekend, and registered user model has a negative slope for holiday/weekend, both of which are highly significant. This is consistent with the boxplot shown in Appendix E, but further confirms that the different usage purposes are the reason that holiday/weekend was not significant in the total model. Because two subsets of the total count have opposite preference in terms of holiday/weekend, their effects cancelled out the significance of the holiday/weekend term in the total model.

**Conclusion**

Our model successfully identified all the important predictors for the daily total rental count (all predictors highly significant in the final model), and brought up a detailed mathematical way to predict the future daily rental count as temperature, humidity, wind speed, and weather situations are easily obtained in any weather forecasts. This can help the bike company adjust their resources and staff based on different daily bike usage.

The downside of our project is that our dataset didn’t really meet the regression model assumptions, especially the equal variance assumption. This will somewhat affect our prediction’s accuracy. A possible way to solve this non-constant variance problem of our dataset is to try a few transformations of our response variable. However it would be very hard to explain the results after transformation, therefore we leave the total model as it is.

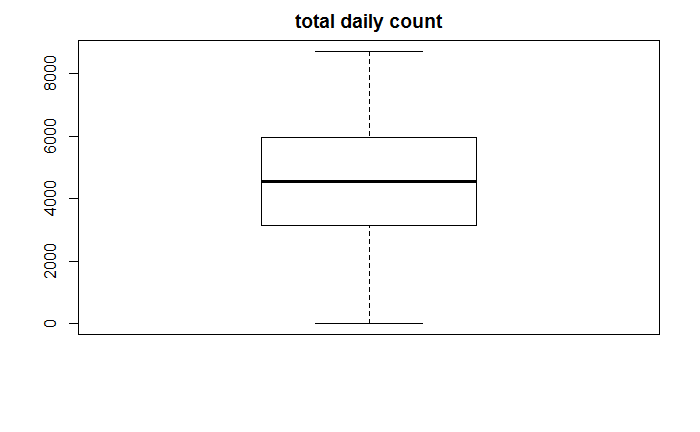
To evaluate how well this model is doing, we need to test our model on future data. If more data in the following years is available, we would be more than happy to test our model accuracy on these data which has nothing to do with the dataset from which we built model. It would be fairly reasonable performance to report how well our model is doing.

References

Fanaee-T, H., & Gama, J. (2013). Event labeling combining ensemble detectors and background knowledge. *Prog Artif Intell, DOI 10.1007/s13748-013-0040-3.*

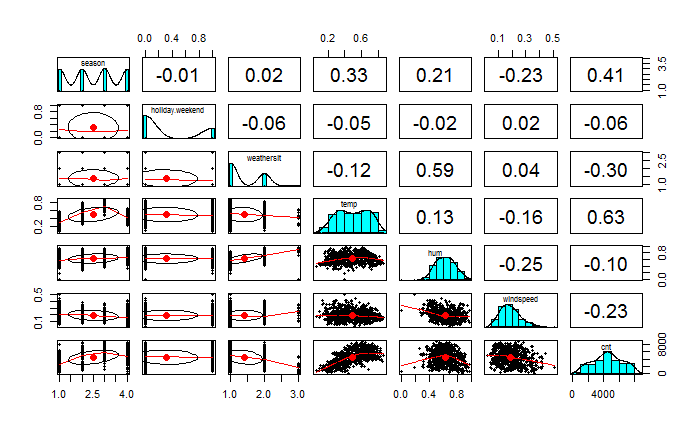
**APPENDIX A**

The boxplot of the daily total count. The data has a large spread, with minimal 2 counts and maximal 8714 counts.

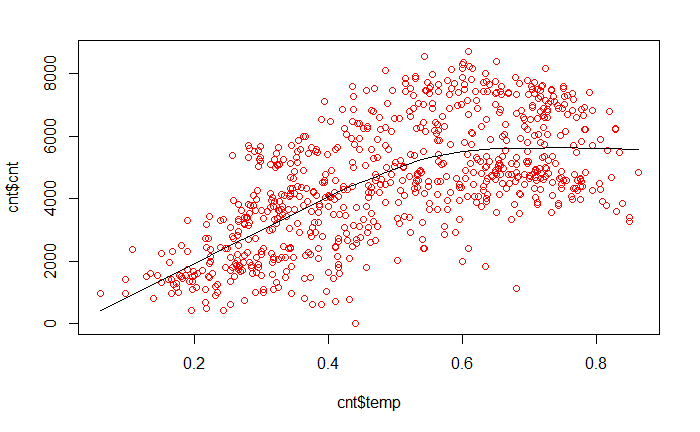


**Appendix B**

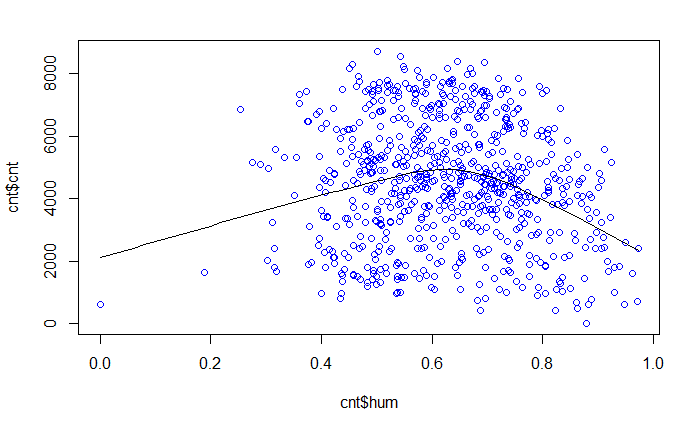
Scatterplot matrix with correlation coefficient to check the relationships between response variable and all potential predictors, and the collinearity between different predictors.



Plot of temperature and response variable with a fitted smooth curve

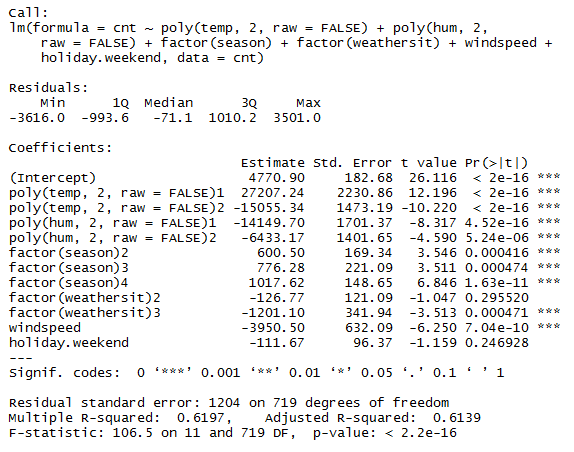


Plot of humidity and response variable with a fitted smooth curve

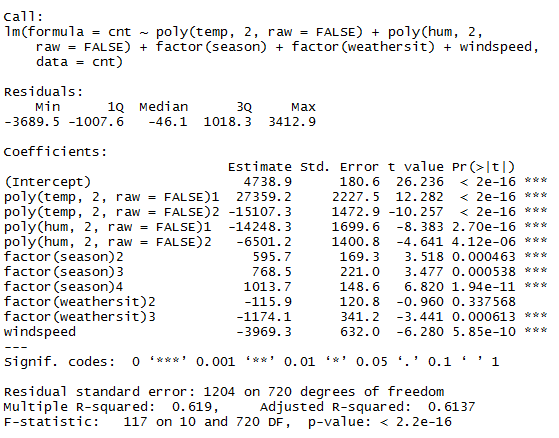


**Appendix C**

Polynomial model with temperature and humidity as quadratic terms, and season, weather situation, wind speed and holiday/weekend as first-order terms.

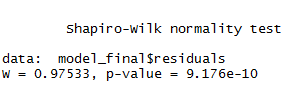


Polynomial model with temperature and humidity as quadratic terms, and season, weather situation, and wind speed as first-order terms.



**Appendix D**

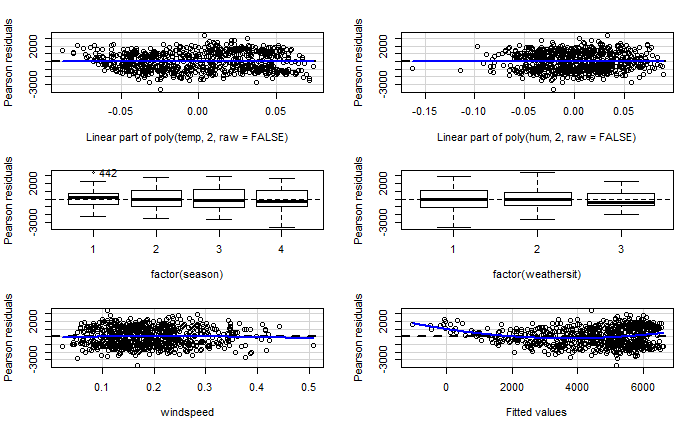
Shapiro-Wilk normality test results



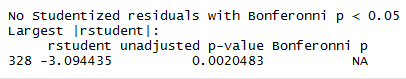
Non-constant variance test



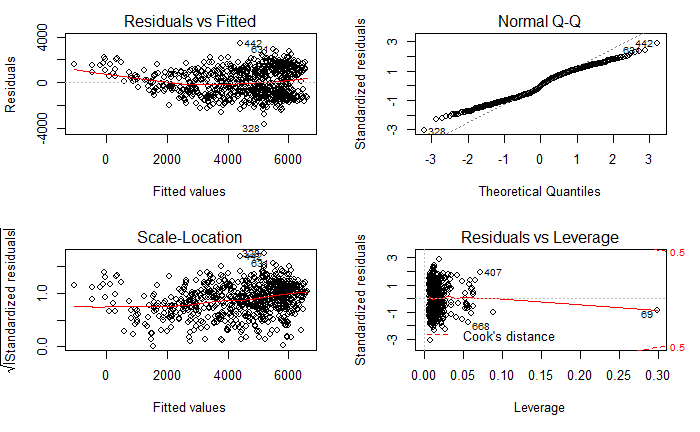
Residual Plots to check linearity



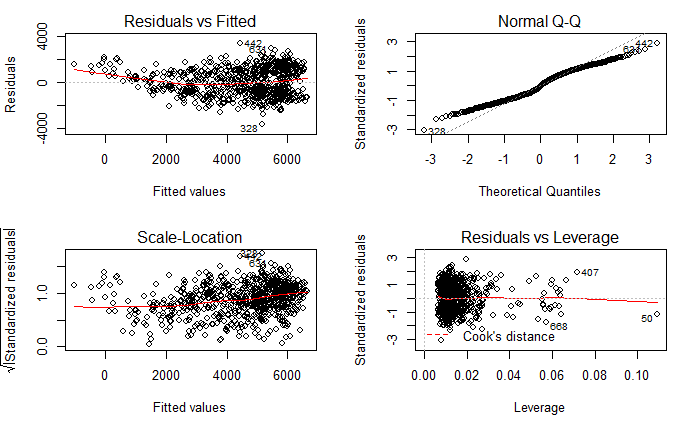
Outlier Test Result



Plots with all observations

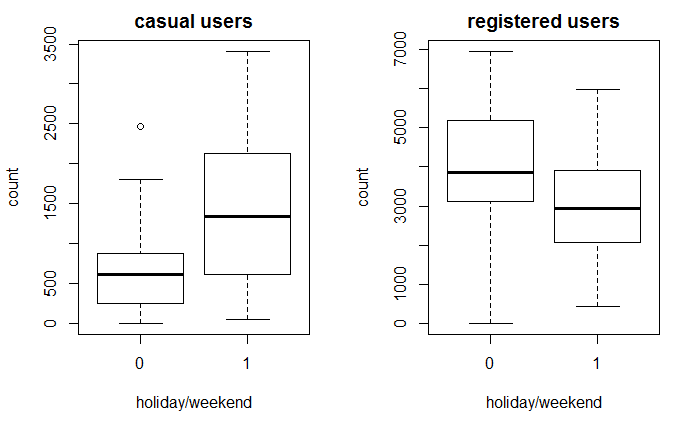


Plots without the most influential 69th observation

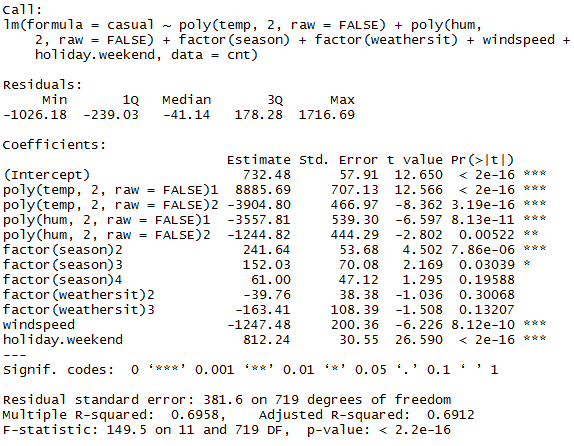


**Appendix E**

Boxplots for casual user count and registered user count by holiday/weekend.



Regression model for casual user count



Regression model for registered user count

