Bike Share Prediction

Fangfang Lee

New York University

**Overview**

Unlike cars and buses, the use of rental bikes has become automatic. The process of renting, returning and using of the bikes can be recorded and collected through the bike sharing system. Therefore, analyzing and collecting the data on when bike uses occur have implications on policies in transportation and public health. In this paper, I will make use of the Bike Sharing Dataset, acquired through UCI Machine Learning Repository, to predict the amount of bike usages in a given hour during year 2011 to 2012 in Washington D.C. I will make use of quantitative methods in the context of both regression and multivariate statistics. The methods I will use is linear regression with categorical predictors, recursive partitioning, as well as an change test.

**Variables and Model**

In the Bike Share dataset, we will use independent variables on weather condition of a given day to predict the amount of bike share at a given hour. The hypothesis is that the amount of bike uses will be high when weather is warm, nice and mild. There are a total of 16 predictors and 1 outcome. The predictors include:

* + instant: record index
  + dteday: date
  + Season: season of the year (1: spring, 2: summer, 3: fall, 4: winter)
  + Yr: year (0: 2011, 1:2012)
  + Mnth: month (1 to 12, January to December)
  + Hr: hour (0 to 23)
  + Holiday: whether it’s holiday or not (1 if holiday, 0 if not)
  + Weekday: day of the week
  + Workingday: whether the day is a work day or not
  + Weathersit: weather situation
  + Temp: Normalized temperature in Celsius
  + Atemp: Normalized feeling temperature in Celsius
  + Hum: Normalized humidity.
  + windspeed: Normalized wind speed. The values are divided to 67
  + casual: count of casual users
  + registered: count of registered users
  + Cnt(outcome): count of total rental bikes including both casual and registered

The outcome, *Cnt*, is the amount of bike usage at a given hour. “Casual” and “registered” are the amount of casual and registered users, and the two variables compose of the outcome. Of the predictors, I predict instant to have no predictive power since it is only the index. *Season*, *year*, *Month*, *Holiday*, *Weekda*y, *Workingday* and *Weathersit* are categorical variables, and some of them are already dummy coded. The categorical variables that are not yet dummy coded need to be transformed in order to perform analysis. The variables *casual* and *registered* will also be excluded from the analysis since the two account for 100% of the variability in the outcome.

My hypothesis is that *season, weathersit, temp, atemp, hum* and *windspeed* account for most variability in the outcome. Therefore, my hypotheses is broken down into the following: 1) For season, winter will have the least amount of bike rentals and spring will have the most amount of bike rentals; 2) Of all the predictors, temperature, humidity and windspeed will account for the most amount of variability in the outcome. To test for such hypothesis, I will first run a linear regression with categorical variables on weathersit and season. I will then build a change test to examine the importance of the predictors. After that, I will utilize a nonparametric model to examine the importance of all predictors if assumptions in a parametric model are violated. I chose linear regression with dummy coded categorical predictors because the resulting models are easily interpretable. I chose a change test to examine the importance of variables because it allows me to assess each variable stepwise. Finally, I chose decision tree for the overall model because decision tree is a nonparametric model which does not require testing for assumptions.

**Descriptive Statistics**

The most interesting variables to examine for descriptive are the outcome *cnt, registered* and *casual*. Based on the descriptive statistics output, we can see that the outcome, hourly count of bike rental, has a mean of 189.46 and a standard deviation of 181.38. This result indicates the high variability of bike rental in the dataset. Furthermore, the mean for registered users is 153.79 with a standard deviation of 151.35; the mean for casual user is 35.67 with a standard deviation of 49.35. Again, both the registered and casual users manifest high variability. There are significantly more registered users than casual users in bike rental.

After examining a correlation matrix, we discovered that temp and atemp are highly correlated. Therefore, I will only use temp as a predictor in the analysis later to avoid colinearity.

**Model Checking**

In order to test for a change test of the continuous variables temp, hum and windspeed, we need to examine the following assumptions: 1) normality of the residuals; 2) linearity; 3) homoscedasticity of residuals; 4) independence of observation. We will examine the assumptions using visualization.

1. Normality of residuals

Upon examining the histogram of the distribution of residuals, we can see that the assumption is violated. There is a positive skew in the residuals.

1. Linearity

The assumption of linearity is also violated. The relationship between temperature, windspeed and humidity are not linear but could rather be curvilinear.

1. Homoscedasticity of residuals

The residuals are not equal around the mean. The scatterplot shows a clear uneven spread of residuals around the mean.

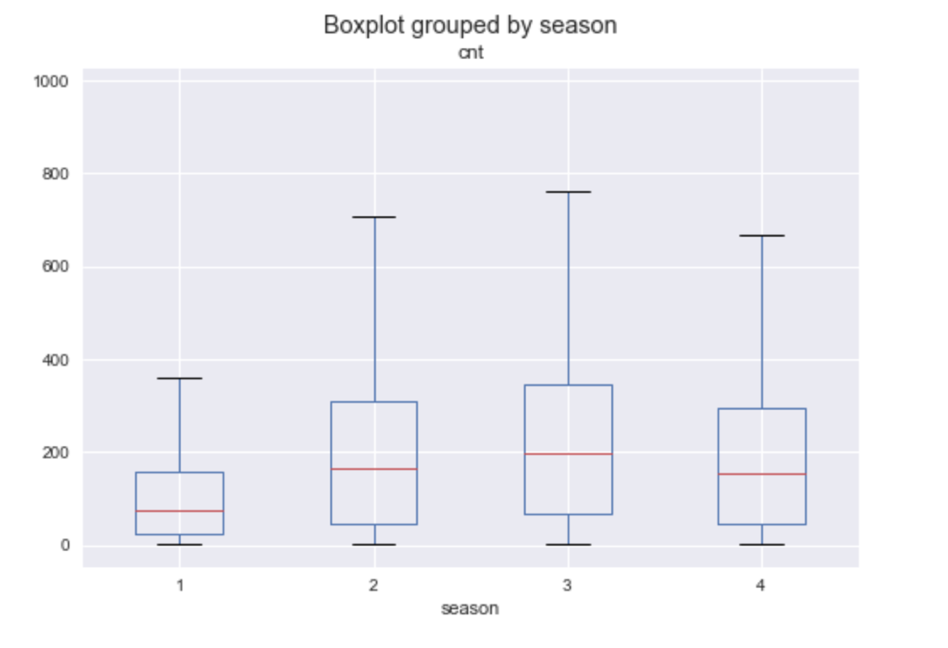
1. Independence of observation

This is the only plausible assumption in this dataset. We can assume that bike rental each hour is not contingent upon the bike rental of other hours.

Since the data do not satisfy most of the assumptions in a linear regression, we only have two remedies: 1) to ignore the assumptions and still use a linear regression, and 2) use a nonparamtric model. In the following sections, I will use both remedies and assess the models.

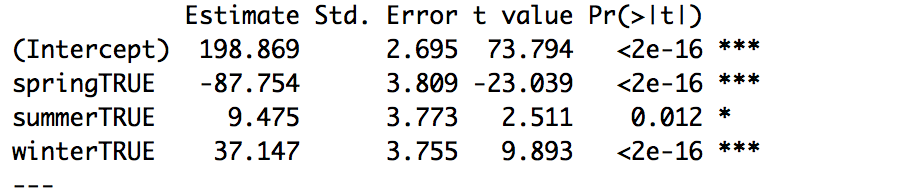
**Model Results**

1. Categorical Variables



*Figure1*. Boxplot of bike rental count grouped by season

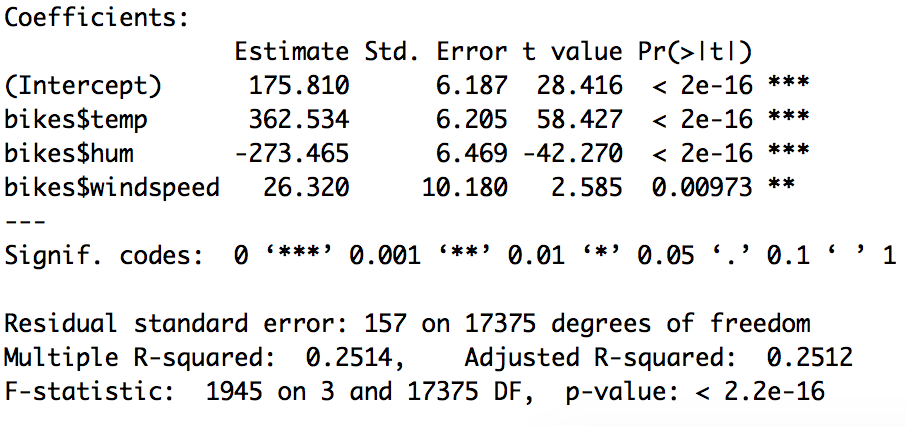
As figure 1 has illustrated, spring shows the least amount of bike rentals and fall shows most amount of bike rentals, followed by summer and winter. The linear regression with dummy variable on season has corroborated the plot: spring shows the least amount of bike share, with a statistically significant coefficient of -87.75. The coefficient represents people rent 87.75 less bikes per hour than the mean in the spring. The season that shows the highest bike rental is the fall, with a significant coefficient of 198.87. In other words, people rent 198.87 more bikes per hour on average in the fall.



*Figure2*. Coefficients for dummy coded variable

1. change model

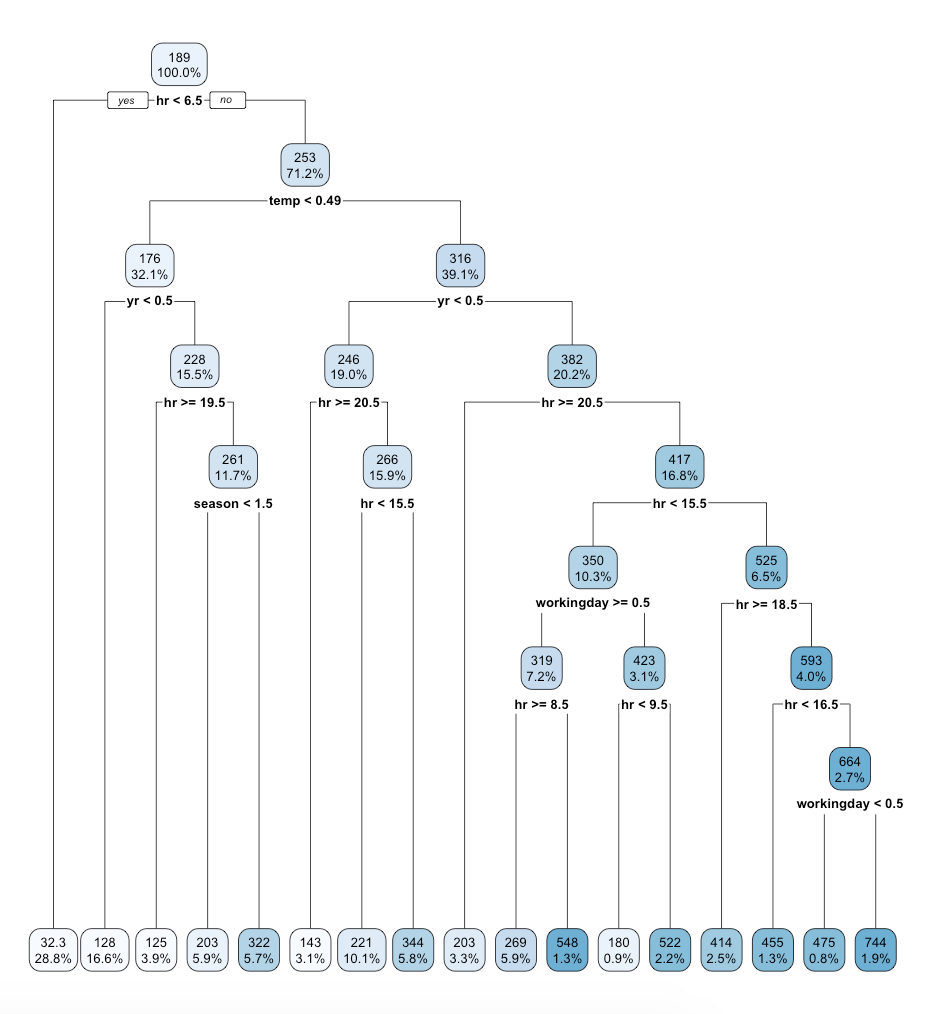
By ignoring the violation of the assumptions, we are able to run a linear regression with windspeed, humidity and temperature as predictors. All of the coefficients are significant, and the three factors account for 25.12% of the variability in bike rental.



*Figure3*. Coefficients for linear regression

1. Decision Tree

Using a regression tree, we use all predictors in the dataset excluding *registered*, *casual*, and *instant*.



*Figure4.* Decision Tree

**Discussion**

**Interpretation of Results**

1. Dummy variable

Contrary to our prediction where we hypothesize spring to have the highest bike rentals and winter the lowest, spring has the lowest bike rentals and fall has the highest. I believe this result is due to a mistake in the information provided by the dataset. Upon close inspection, the month of January to March is coded as 1, but indicated as spring by the dataset provider. I believe that 1 stands for winter, 2 for spring, 3 for summer and 4 for fall.

1. The F test

The multiple linear regression, or F test, shows significance in the three continuous variables. Confirming to our beliefs, the warmer it is, the more likely that people rent bikes. Humidity is negatively correlated with the outcome, showing a decrease in bike rentals when humidity increases. Wind speed shows a positive correlation with the outcome. I believe the result is intuitive—people are more likely to ride bikes when weather is warm, and not raining. However, the wind speed predictor seems like intuitive. I believe a possible explanation for the significance in wind speed is the sample size. The dataset is composed of a total of 17379 cases, so it is not surprising to find a significant coefficient in dataset of such size.

1. Decision Tree

Contrary to our beliefs, the variable with the most amount of predicting power is hour of the day: people are much more likely to rent bikes when it is after 6:30 am. The second most important predictor is temperature, and then year. There are a lot of interactions between variables.

**Limitation**

There is substantial amount of limitations in the above analysis. First, the dataset itself is flawed. As indicated by the previous section, the season is coded incorrectly by the provider of the dataset. Without the correct coding, the variable will produce misleading predictions. Secondly, the dataset needs extensive recoding on categorical variables. The categorical variables that are not yet dummy coded needed to engineered. Furthermore, if we want to utilize an F test with dummy variables as predictors, we will need to use linear contrast, or contrast coding, which can be highly complicated. Thirdly, the assumptions for a linear regression are violated, so we need to opt for other more advanced algorithms to yield reliable predictions. Lastly, even though the decision tree algorithm does not require examining for assumptions, it creates high variance depending on the specific tree. Therefore, we need to construct multiple tree and ensemble the trees to create a random forest.

**Recommendation**

The possible explanation for the failure of traditional linear regression, or F test, in this dataset is that the dataset is created for machine learning purposes. The magnitude of observations and variables are beyond the scope of traditional linear regression. In order to have more predictive power, we need to try using more complex variable selection algorithms such as lasso regression. Furthermore, in order to better predict bike share, we could collect data on behavioral aspects of the users in addition to weather conditions. Variables such as self-perceived activeness, whether user owns a car, preference for transportation, gender, age, and body mass index etc. could all have an impact on predicting the amount of bike rental.

*References*

Capital Bike Share data*. Retrieved from https://www.capitalbikeshare.com/system-data*