**Regression Analysis of Bike Sharing Problems**

Group Members: Tianlong Xu, Lina Hu and

# Introduction

Some cities provide public bikes on street for short-term rent. The daily rentals can be affected by several factors, such as the weather condition, temperature, humidity and wind speed, etc. The primary task of this project is to conduct a statistical analysis, constructing proper relationships of bike rentals vs. possible influential factors. A prediction of future rentals based on this statistical analysis is also conducted at the same time. Several statistical methods like linear regression; time series analysis and tree-based analysis are implemented and compared in this project.

# Data Structure

Our data for this project is obtained from *Kaggle* ([*https://www.kaggle.com/c/bike-sharing-demand/data*](https://www.kaggle.com/c/bike-sharing-demand/data)).

We are provided hourly rental data spanning two years. Training data set is comprised of the first 19 days of each month. Since variables in provided testing data does not match with training data, we decided to divide existing training data into new training and testing data sets. We retained data of the first 17 days to be training data and the rest to be testing data.

Our variables are listed here:

**Datetime** - hourly date + timestamp

**Season** - 1 = spring, 2 = summer, 3 = fall, 4 = winter

**Holiday** - whether the day is considered a holiday

**Workingday** - whether the day is neither a weekend nor holiday

**Weather** -

1: Clear, Few clouds, Partly cloudy, Partly cloudy

2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

**Temp** - temperature in Celsius

**Atemp** - "feels like" temperature in Celsius

**Humidity** - relative humidity

**Windspeed** - wind speed

**Casual** - number of non-registered user rentals initiated

**Registered** - number of registered user rentals initiated

**Count** - number of total rentals

# Models

* 1. **Linear models**

Several linear models were created based on correlations of variables. We treat datetime, season, holiday, workingday and weather as factors.

Attempts are listed here:

Model 1:

count~holiday+season+weather+workingday+temp+atemp+humidity+windspeed

Model 2:

count~datetime+holiday+season+weather+workingday+temp+atemp+humidity+windspeed

Model 3:

count~hour+holiday+season+weather+workingday+temp+atemp+humidity+windspeed

Model 4:

count~hour+holiday+season+weather+workingday+temp+atemp+humidity+windspeed+holiday\*season+workingday\*season+weather\*holiday

Model 5:

count~hour+holiday+season+weather+workingday+atemp+humidity+windspeed+sqrt(windspeed\*humidity)

Model 6:

count~hour+holiday+season+weather+workingday+windspeed\*humidity\*atemp

Model 7:

count~hour+holiday+season+windspeed\*humidity\*atemp

* 1. **Time Series Analysis**
  2. **Tree Based Methods**
  3. **Other Models Tested**

# Results

* 1. **Linear models**

Adjusted R squares and testing errors are summarized in the following table.

|  |  |  |
| --- | --- | --- |
| **Model #** | **Adjusted R square** | **Testing Errors** |
| 1 | 0.2865 | 2.0902559 |
| 2 | / | / |
| 3 | 0.6323 | 1.0678582 |
| 4 | 0.6326 | 1.0736497 |
| 5 | 0.6319 | 1.0509096 |
| 6 | 0.6361 | 1.0011115 |
| 7 | 0.6288 | 0.9687209 |

From the summary above, we can see that Model #6 is preferable over others, despite the fact that its R square is relatively low based on common sense. We can also conclude that linear model does not behave perfectly in this data set.

Several plots for Model #6 are:

* 1. **Time Series Analysis**
  2. **Tree Based Methods**
  3. **Other Models Tested**

# Conclusions

To be summed, we have applied several methods to this bike-sharing problem and our key notes are:

1. Linear regression cannot provide a perfect fitted model for bike rentals and its variables;
2. Inner correlations among variables should be particularly considered;

**Appendix**