Homework1 Regression Analysis on Bike Share

Xuanhui Liao, Yunchan Sun, Becca Wernick, Taowan Yang, Xiaowen Zhang

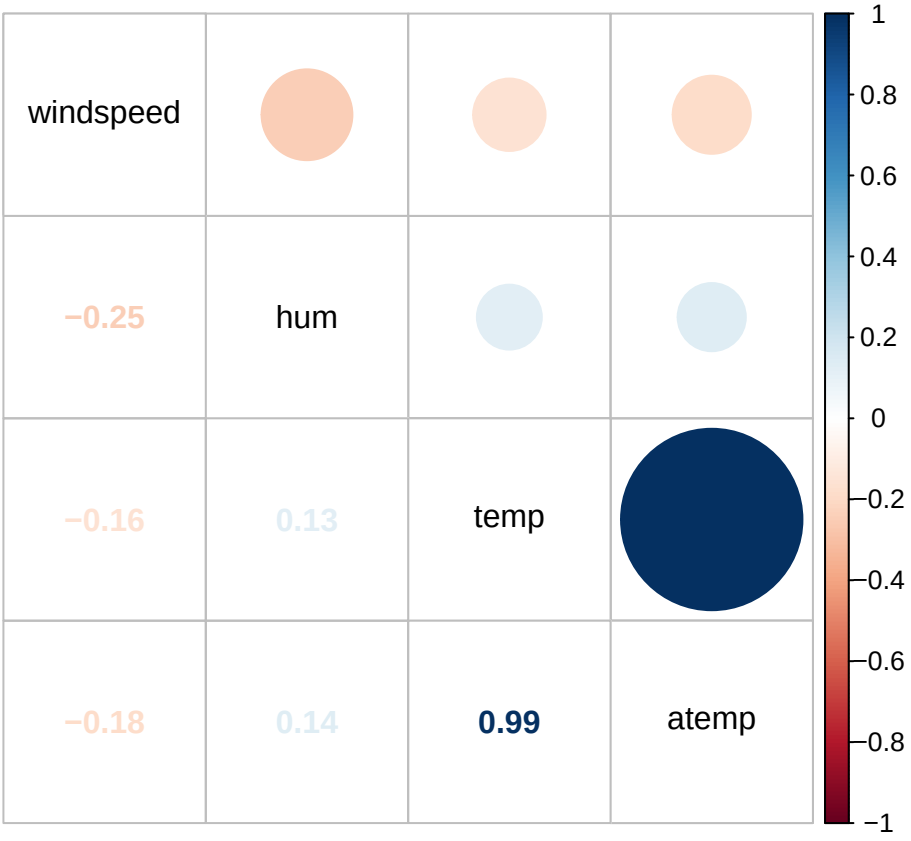
Oct. 26 2018

Project Overview

In this project, we performed regression analysis on the bike-sharing data to predict the total count of daily bike users. Based on our "best" model, wind speed and normalized feeling temperature are the two most influencial factors to the total counts of daily bike users. In the end, we provided useful suggestions for maximizing the profits of the company.

Exploration of the Raw Data

The graph below shows the correlation matrix of the numeric independent variables. It is shown that the variables "temp" and "atemp" have the highest positive correltion, while "windspeed" and "humidity" have the lowest negative correlation.

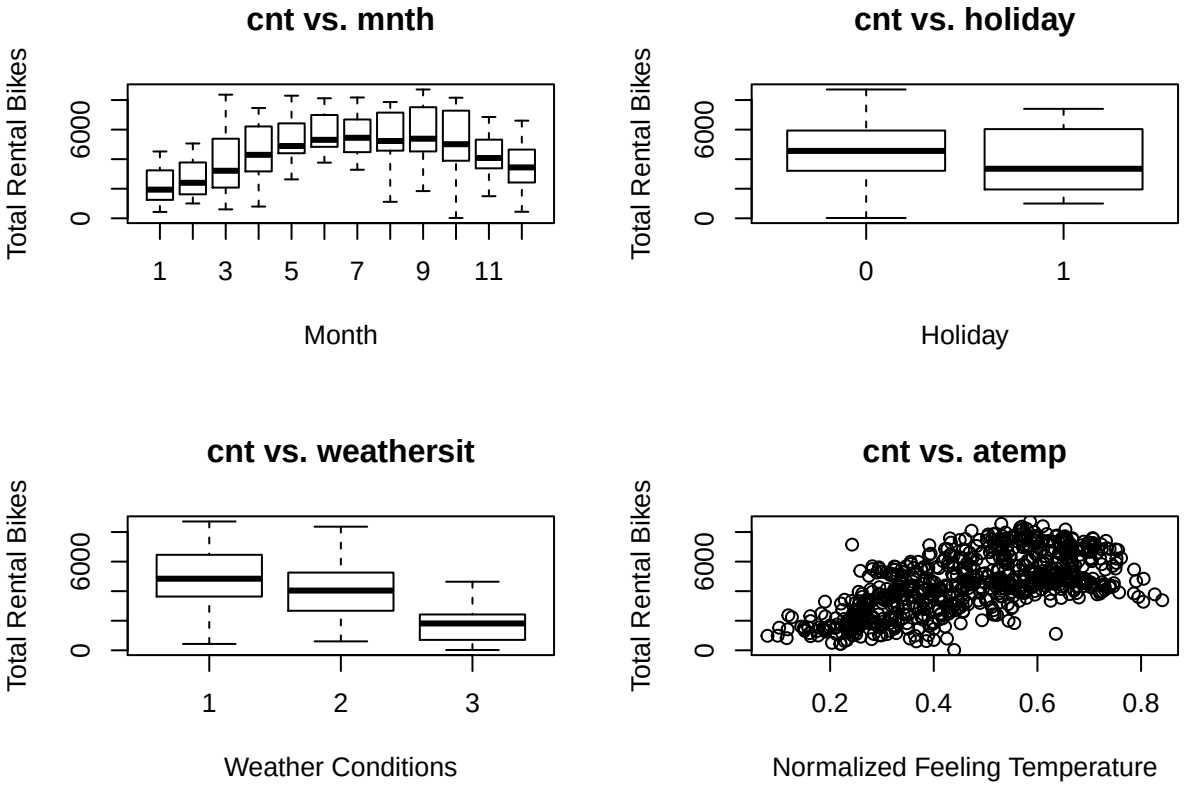


The table below shows some general statistics of the independent numerical variables.

```
##    windspeed           hum             atemp            temp
##  Min.   :0.02239   Min.   :0.0000   Min.   :0.07907   Min.   :0.05913
##  1st Qu.:0.13495   1st Qu.:0.5200   1st Qu.:0.33784   1st Qu.:0.33708
##  Median :0.18097   Median :0.6267   Median :0.48673   Median :0.49833
```

```
##  Mean   :0.19049   Mean   :0.6279   Mean   :0.47435   Mean   :0.49538
##  3rd Qu.:0.23321   3rd Qu.:0.7302   3rd Qu.:0.60860   3rd Qu.:0.65542
##  Max.   :0.50746   Max.   :0.9725   Max.   :0.84090   Max.   :0.86167
```

After exploring all variables, we discovered that the following three categorical variables in box plots have more distingiushed distributions between different levels than others. Also, the scatterplot shows an obvious relationship between "atemp" and the response.



Model Building

We applied t-tests, backward elimination, and ANOVA F-tests to determine the "best" linear regression model to predict the total counts of daily bike users. We then validated our "best" model by performing diagnostic tests. Our final model was able to explained 85.9% of the variability in the response variable.

The variables included in our "best" model, as well as their coefficients, are shown below.

```
## (Intercept)          yr         Feb         Mar         Apr         May
##  1746.35717  2108.37836  -647.42378   -72.53018   451.73201   710.35860
##         Jun         Aug         Sep         Oct     holiday         Mon
##   571.39431   333.62818  1115.18786  1321.12449  -544.98184   206.21638
##         Tue         Wed        Thur         Fri         Sat       atemp
##   353.77188   341.10582   406.27497   505.30719   503.56403  5668.27123
##    weather2    weather3         hum   windspeed
##  -494.75628 -1968.38142 -1315.92517 -2693.90779
```

Based on our "best" model, variable atemp and windspeed have the most influence to total number of bike rent. With one unit increase of normalized feeling temperature in celsius, the total number of bike rent

increases by around 5668 units, holding everything else constant. On the other hand, with one unit increase of normalized wind speed, the total number of bike rent decreases by around 2694 units, holding everything else constant.

Prediction Inerval

Prediction interval tests are conducted for the following three conditions:

1.global mean of all columns as benchmark

2.increase only atemp by 5 Celsius (0.1 after normalized), holding every thing else unchanged (global mean)

3.increase only windspeed by 6.7 (0.1 after normalized), holding every thing else unchanged (global mean)

```
##        fit      lwr      upr
## 1 4560.801 3131.658 5989.945
```

```
##        fit      lwr      upr
## 1 5127.629 3697.731 6557.527
```

```
##        fit      lwr      upr
## 1 4291.411 2860.137 5722.684
```

We can be 95% confident that with a 5 Celsius increase in feeling temperature, the percentage of the total number of bike rent will increase to be in between 3131.658 and 5989.945.

We can be 95% confident that with a 6.7 mph increase in feeling temperature, the percentage of the total number of bike rent will increase to be in between 2860.137 and 5722.684.

Suggestions

CONDITIONAL SUGGESTIONS:

Based on the model and the prediction interval, we recommend the following promotions to boost the demand:

- Rainy Day Promo - running a price promotion on weathersit 3 (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) to increase ridership.
- Holiday Plant-a-Tree Promo - offering bike promotion for eco-friendly concerns - if you bike 10 (or more) miles on a holiday, the company will plant a tree.
- Dining in Winter Promo - collaborating with restaurants in the winter months for a discount on food. (Reciprocal discounts will also be given at partner restaurants.)

GENERAL SUGGESTIONS:

In addition to our suggestions to address specific demand lags, we propose the following to boost overall demand:

- App - seeing other nearby users to make friends / social media meets geotracking
  - mile-ranking among the users
  - providing weather data for next day/week
  - sharing top (safe) bike routes for each city (routes updated in holidays to boost the demand)
  - users are able to decide if geotracking is permitted (for privacy concern)
- Nitty Gritty Bike Event - holding a “Nitty Gritty” biking competition in mud
- Rent-to-Own High-End Bikes - the users are able to purchase the high-end bikes at a discount after renting for a certain period