Header

Names: Avanish Subbiah

Purdue Usernames: subbiah

Path: 1

Dataset

I will be working with the `NYC_Bicycle_Counts_2016_Corrected` dataset, which contains information about the weather, date, and number of cyclists across four bridges. The count's of cyclists is given as strings, along with precipitation. The precipitation value 'T' was assumed to be equivalent to 0, and the precipitation value '(S)' as assumed to be an error and removed (value to left was kept).

Analysis

- 1. I chose to find the mean number of cyclists for each of the four bridges to determine which three bridges to add sensors to. Idealy the dataset should include as many samples as possible, which is why the three bridges out of the four with the highest mean number of cyclists are the most ideal for sensor additions.
- 2. I chose to run a ridge regression with lambda values from 0.1 to 100, using the X values as the normalized high temp, low temp, and precipitation values, and the total cyclists as the Y target values. I expect this to provide me with a linear model that takes normalized value for high temps, low temps, and precipitation values to predict the total number of cyclists.
- 3. I chose to run another ridge regression with lambda values of 0.1 to 100, using X values of the total number of cyclists, and the target Y values as precipitation values. I expect this to provide me with a linear model that can predict precipitation level based on cyclist count.

Results

1.

The output of the mean analysis of the four bridges resulted in the following values:

```
['Brooklyn Bridge', 'Manhattan Bridge', 'Williamsburg Bridge', 'Queensboro
Bridge']
[3030.700934579439, 5052.2336448598135, 6160.873831775701,
4300.72429906542]
```

From these values I conclude that the sensors should be placed on Queensboro, Williamsburg, Manhattan bridges because these bridges have the top three average cyclist counts of the four bridges.

2.

The model found using the analysis is following coefficients.

```
Best lambda tested is 0.1, which yields an MSE of 13975003.010753594
[ 4355.15125561 -1553.88596244 -1938.75893002]
18412.600000000006
```

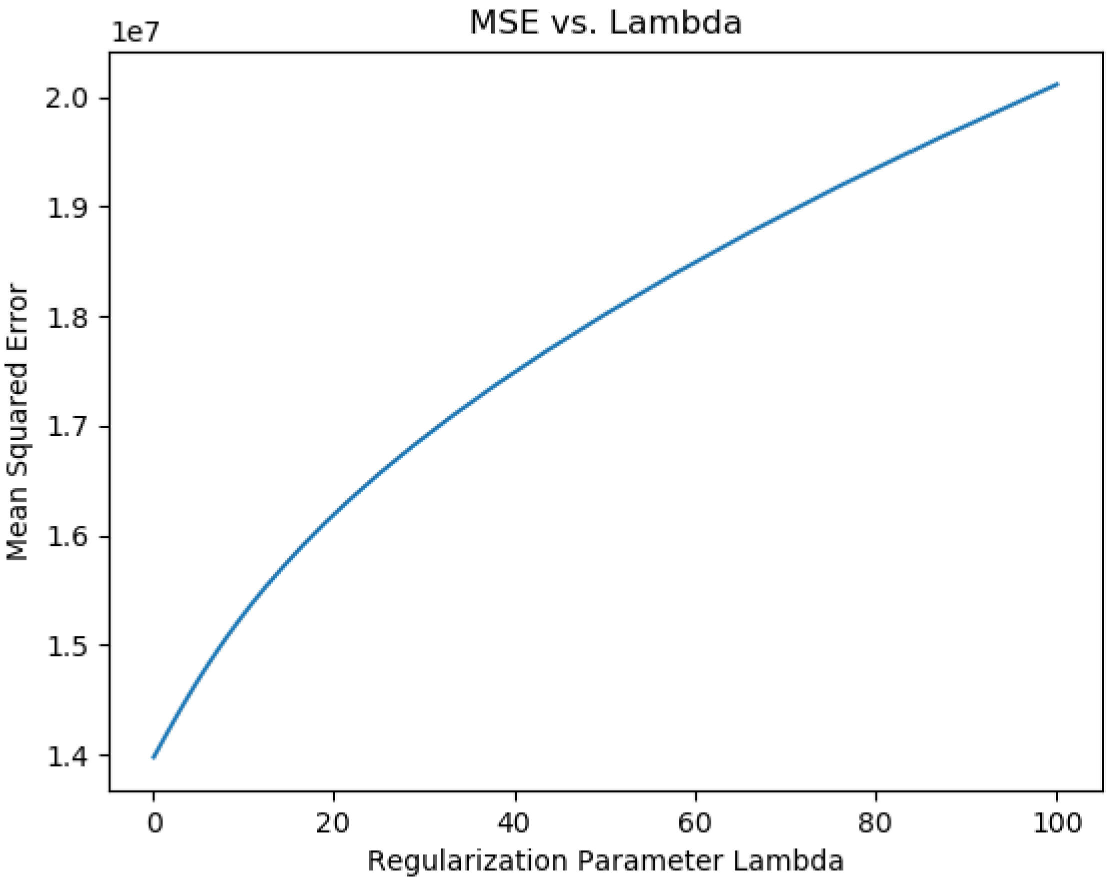
This yields the prediction equation:

y = (4355.15125561 * x_1) + (-1553.88596244 * x_2) + (-1938.75893002 * x_3) + 18412.600000000006

Where the variables correspond to the following fields:

- y | Number of Cyclists
- x_1 | High Temp (°F) (Normalized)
- x_2 | Low Temp (°F) (Normalized)
- x_3 | Precipitation (Normalized)

This selected a very low lambda value of 0.1 based on the relationship between lambda and MSE shown below:



With the tendency to select extremely low lambda values and high MSE of 13975003, I believe the model is heavily overfitting to the data, dispite the 25% split of training and testing data, which is likely due to there not being enough samples to provide a more generalizable model for the data. As such, I do not believe there is enough samples for the city to use the weather to predict cyclist counts at least with ridge regression modeling.

3.

The model found using the ridge regression analysis are the following parameters:

```
Best lambda tested is 0.1, which yields an MSE of 0.04334012059014044
[[-0.10522649]]
[0.1064375]
```

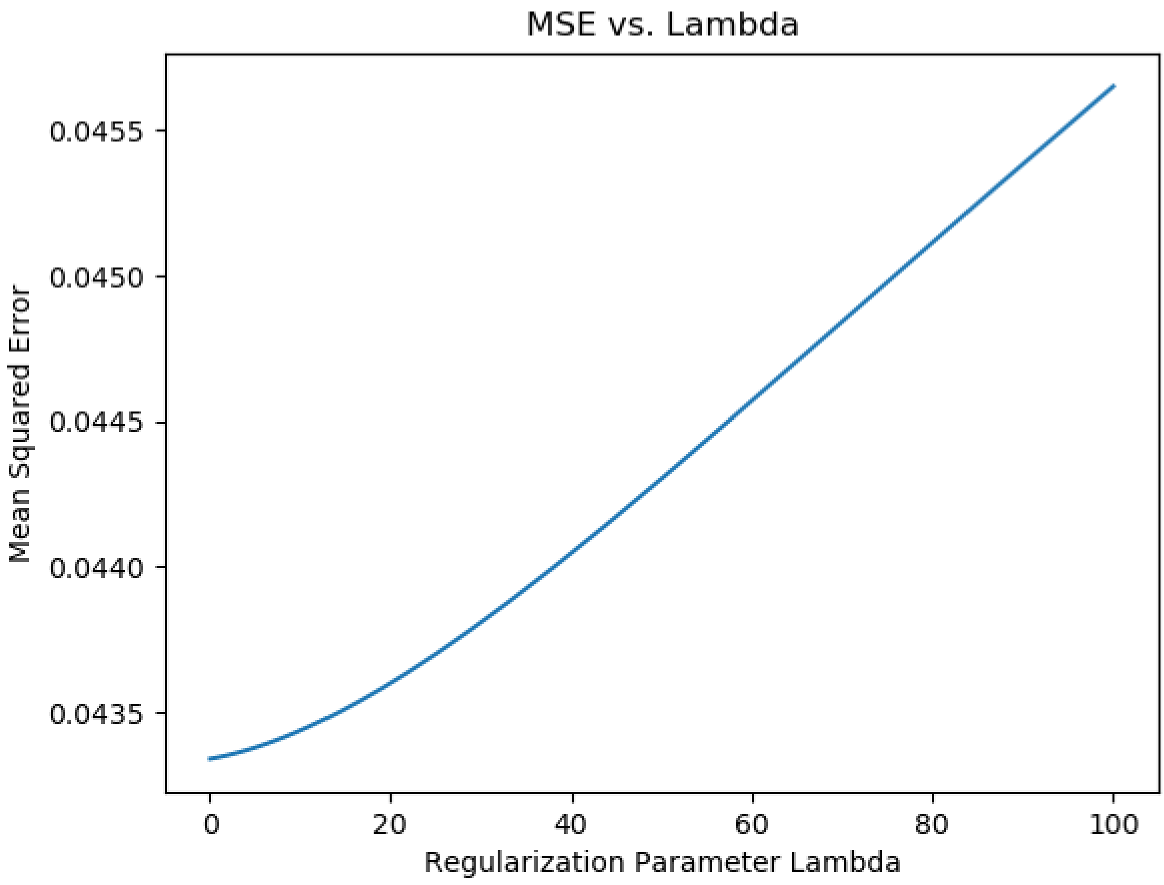
This yields a model equation of the following structure:

y = (-0.10522649 * x) + 0.1064375

Where the values y and x correspond to the following fields:

- y | Precipitation
- x | Total number of cyclists (Normalized)

This also selected a very low lambda (or smoothing) value of 0.1, but also yielded a very low MSE of 0.04 on the testing dataset (75% training 25% testing dataset split). The relationship between MSE and lambda is shown below:



From this analysis, using ridge regression, I can conclude that the data can be used to determine whether it is raining based on cyclist count with a reasonable error (due to the relitively low MSE of 0.4), but there is a chance of model overfit with the still low value of lambda of 0.1. Idealy even for this prediction more samples should be taken for a ridge regression analysis.