Project ECE 20875: Python for Data Science

Spring 2022

1. Project team information

Mini-Project Spring 2022

ECE 20875

Liangcheng Jiang- jiang786 - jiang786@purdue.edu

Ruihang Ni – ni102 – ni102@purdue.edu

Path chosen: Path 1

2. Descriptive Statistics

The dataset we will analyze in this project, which names "NYC_Bicycle_Counts_2016_Corrected.csv", is a CSV file that records the data of the weather conditions and number of riders pass through the four bridges in New York City from April 1 to Oct. 31. In the weather condition, the data includes the high and low temperature in Fahrenheit scale (°F), and the precipitation in inch scale of each day. The data of the riders every day for Brooklyn Bridge, Manhattan Bridge, Williamsburg Bridge, and Queensboro Bridge are recorded separately and there is also a total number of riders every day at the final column. We aim to find out the relationship between the weather condition and the number of riders on the four bridges through analysis of the data recorded in these days and predict more cases based on the model we get in the analysis.

3. Approach

- Problem 1:

  For problem 1, we decided to use linear regression to build a model based on the normalized data to reflect the effect of number of bikes passing on each bridge to the total amount of the people. The model should have the form as following:

  $$y = ax_1 + bx_2 + cx_3 + dx_4 + e$$

  In this model, x_1, x_2, x_3, and x_4 are normalized data of Brooklyn Bridge, Manhattan Bridge, Williamsburg Bridge, and Queensboro Bridge respectively. The *y* is the normalized total numbers of riders. We will write the code applying least square equation: $\beta = (X^T X)^{-1} X^T y$, where $\beta = [a, b, c, d, e]$.

  In this model, the parameter with the least coefficient is least important to the total numbers of the riders passing the bridges. The

bridge corresponding to that parameter is the one we choose to not install the monitor on it.

- Problem 2:

  For this question, we decide to perform ridge regression with different values of regularization parameters and find the regularization parameter that can generate a ridge model with the least mean square error. With this ridge regression model, we can predict the total number of riders passing through the four bridges more precisely based on the weather condition.

  In this model, we will have three exploratory variables: x_1 as the high temperature, x_2 as the low temperature, and x_3 as precipitation. The target variable is *y*, which is the total numbers of riders passing the bridges on each day. These data will all be normalized before regression to reduce the effects of difference in scale. The model we will generate here will include a list of coefficients, [*a, b, c*], that corresponding to the three exploratory variables and an intercept *d*.

  We swap the regularization parameter λ from 10^{-5} to 10^{5} to generate a list of models and calculate the mean squared errors of these models respectively. Then plot the graph of mean square error vs. regularization parameters to find the model with the regularization parameter that causes the least mean square error.

- Problem 3:

  In this problem, the data we will involve is precipitation and the number of riders on the four bridges. Since we need to predict whether it's raining based on the riders, precipitation is converted to a binary data called "*if_rain*", which equals to 0 when the precipitation is 0 and 1 otherwise.

  To deal with the binary data, we decide to use logistic regression The model we generate here has the following form:

$$P(\text{if\_rain} = 1 \mid x) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4))}$$

  The x_1, x_2, x_3, and x_4 represent for the numbers of riders on Brooklyn Bridge, Manhattan Bridge, Williamsburg Bridge, and Queensboro Bridge respectively each day. The function P depends on the four independent variables and represent for the probability of raining on that day. With the coding in Python, we will generate the logistic regression model and get a list of coefficients in the following form:

$$[\beta_0, \beta_1, \beta_2, \beta_3, \beta_4]$$

  After we get the coefficients of the model, we will use accuracy score to estimate the goodness of our logistic regression model. We will use the

model to get predicted logics of whether raining based on the riders on the different bridges, compare them to the real data, and get the accuracy score.

4. Analysis

- Problem 1:

After running the code of regression, we get the model coefficient as the following:

$$a = 0.19888252576089743$$

$$b = 0.306113602111023$$

$$c = 0.22114472049754003$$

$$d = 0.33507804824624693$$

$$e = -1.9081958235744878 \times 10^{-16}$$

From the data above, we can see that *a* is about 0.199, which is the smallest among the four coefficients multiplied by the parameters. The parameter corresponding to *a* is x_1, which represents for Brooklyn Bridge. Here we can conclude that we can choose to not install the monitor on Brooklyn Bridge since it has the least importance to the total number of riders passing the bridges on each day.

The mean squared error we calculate based on the predicted values and the actual value is 0.105305 in this model, which is not very high from my point of view. The coefficient of determination r^2 for this model is 0.894201, which means about 89.4% of variation in the data can be interpreted by this model. In my opinion, the model is reliable enough to explain the relationship between the riders passing the different bridges respectively and total number of riders.

- Problem 2:

After we run the code and generate a list of models with different regularization parameters, we also calculate the mean squared errors corresponding to these different models. Plot the graph of mean squared error vs. regularization parameters and get the graph as following:

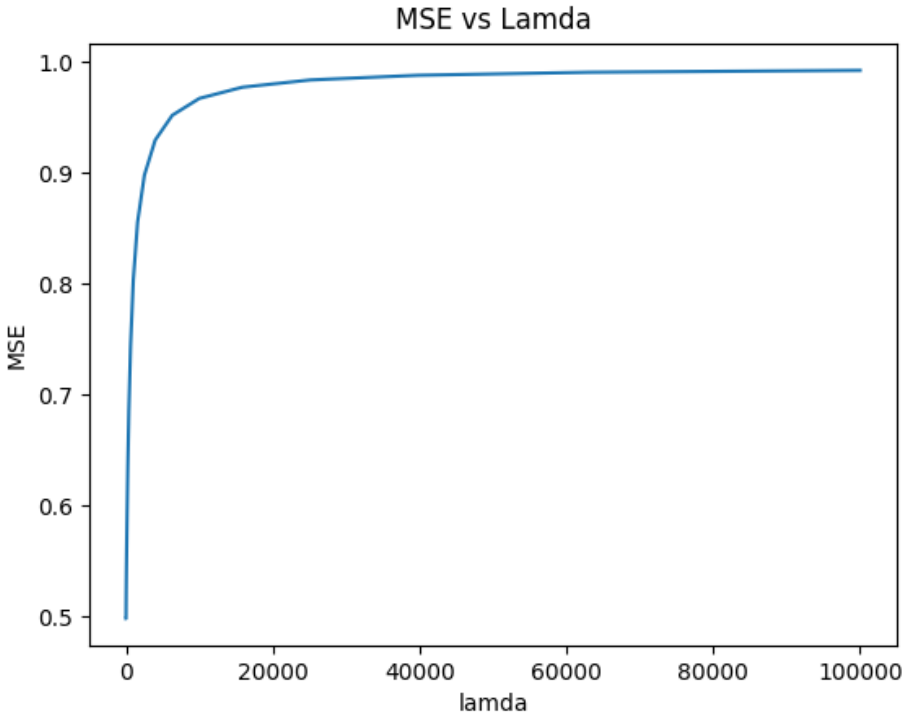

Figure 1. Mean squared error vs. regularization parameter

From the graph, we can see that the least mean squared error is smaller as regularization parameter λ becomes smaller. By index searching in Python coding, we can know that the smallest mean squared error is 0.498204. This occurs at the index 0 of the model list, which corresponding to $\lambda = 10^{-5}$. The model generated in this parameter is as following:

$$a = 0.86007705$$

$$b = -0.33222331$$

$$c = -0.36250909$$

$$d = -6.03848371309352 \times 10^{-17}$$

With this ridge model based on high temperature, low temperature, and precipitation, we can predict the total number of riders passing the bridges each day with some bit of errors. However, the range of date to predict should be restricted from April 1 to Oct. 31 as shown in the dataset. This is because the weather conditions in different seasons can be very different and can have different effects to the number of people. To make more comprehensive cases of prediction, we should have the dataset through the whole year with all different seasons.

- Problem 3:

After we run the code and get the logistic regression model, the output is as following:

$$\beta_0 = 1.35219671$$

$$\beta_1 = -2.71019542 \times 10^{-4}$$

$$\beta_2 = -2.94485328 \times 10^{-4}$$

$$\beta_3 = 8.15896170 \times 10^{-4}$$

$$\beta_4 = -2.71019542 \times 10^{-4}$$

After we get the model, we can make predictions on whether it's rain on a particular day based on the numbers of riders on the four bridges. If the value of function P is greater than or equal to 0.5, we can say it's rain. If the value calculated by P is smaller than 0.5, we can say it's not rain.

The accuracy score we get for this model is 0.775701, which means we have about 77.6% of confidence to make a right prediction of whether it's raining on a particular day based on the numbers of riders on the four bridges.