Drake Song

Project 2 Technical Report

Project Description

This project aims to utilize Linear Regression on a given set of data to convince the city of Los Angeles to implement a bike sharing program. In order to make the linear regression model work effectively, two methods were used: subset selection and shrinkage.

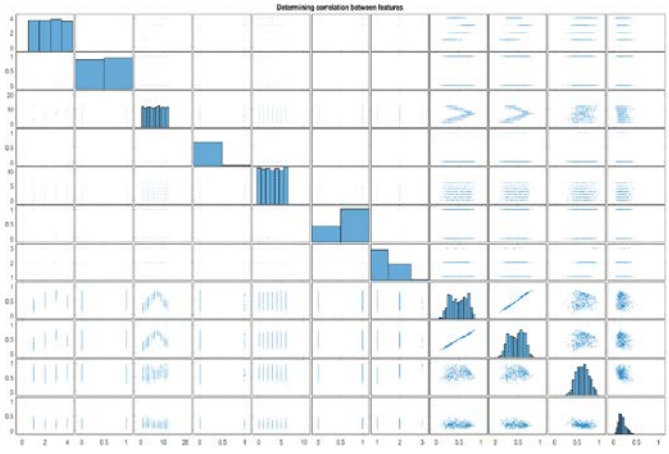
Problem Statement

Given the data consisting of numerous features that affect the number of bikers on a given day, will the environment and climate of the city of Los Angeles be suitable for a bike sharing program?

Feature Selection Rationale

Before anything is discussed, 10% of the data is randomly selected every time the program is run. This creates different results every single run of the program. Although different results are produced, the most occurring and the most reasonable results of the model was used in analysis. Many of the calculated numbers will vary. The total number of users were used as the Y value.

First, correlation between the features of the data was checked. Because having correlated features causes faulty predictions and almost all models of supervised learning to fail, there must be no correlation between features that are used to train the model. After plotting the features against each other (Figure on the left), it is clear to see that features 8 and 9 are correlated. Feature 8 represents normalized temperature in Celsius and feature 9 represents normalized feeling temperature in Celsius. Because both features are measuring the temperature on the same day, it is clear that there exists a correlation between these two features. In order to determine which feature to remove, linear regression models, one with feature 8 removed and the other with feature 9 removed, utilized to calculate the adjusted-R^2. Comparing the adjusted-R^2 (instead of R^2 since R^2 increases whenever you add more data into the model no matter the data you added is helpful or not; adjusted-R^2 accounts for this and does not increase the number if the data is not helpful), when feature 8 is removed, the adjusted-R^2 is 0.8053 and when feature 9 is removed, the adjusted-R^2 is 0.8036. Since 0.8053 is higher, feature 8 is removed.



Using backward step selection, adjusted-R^2 were compared and none of the new models were outputting higher adjusted-R^2 than 0.8053. Then the confidence interval of each features were calculated and feature 6 returned [-31.72, 264.59]. Because this confidence interval contains 0 and only decreases the adjusted-R^2 by 0.005, feature 6 is removed. After doing so, backward step selection was run once more and again no higher adjusted-R^2 was produced.

Model Description

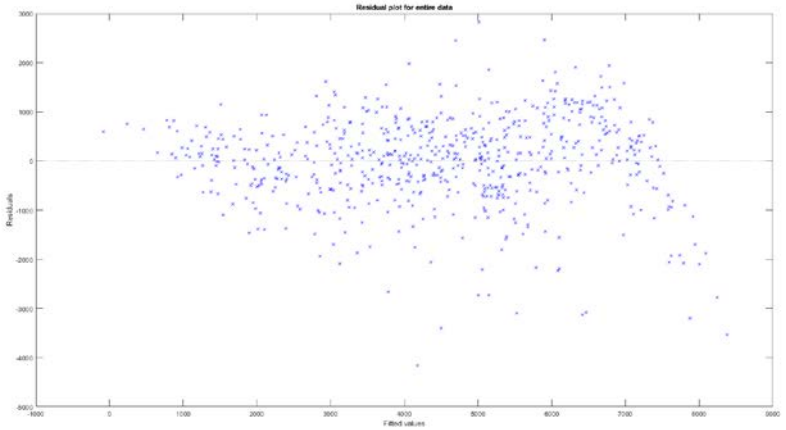
Linear regression utilizes independent features to define a linear relationship between them. Using the given data, linear regression model can predict outcomes of potential data with the model that has been trained with the given data through least squares method. In this project, multivariate linear regression is utilized in order to accompany the numerous features of data that have been provided to work with.

There are several methods used in this project that help to make linear regression to work better and provide a more accurate model: subset selection and shrinkage. Subset selection allows for the best subsets of features to be used in determining the model. Because using all of the features is going to take up a lot of resources and time especially when working with large amount of data, it is best practice to discard some features that are not adding much to the model. On the other hand, shrinkage is an alternative to subset selection in which all features are utilized but the weight of the features are increased and lessened depending on the importance of the feature. More commonly known as regularization, shrinkage methods utilize something called shrinkage penalty in which lambda is tuned accordingly to increase or decrease the penalty.

Assumptions

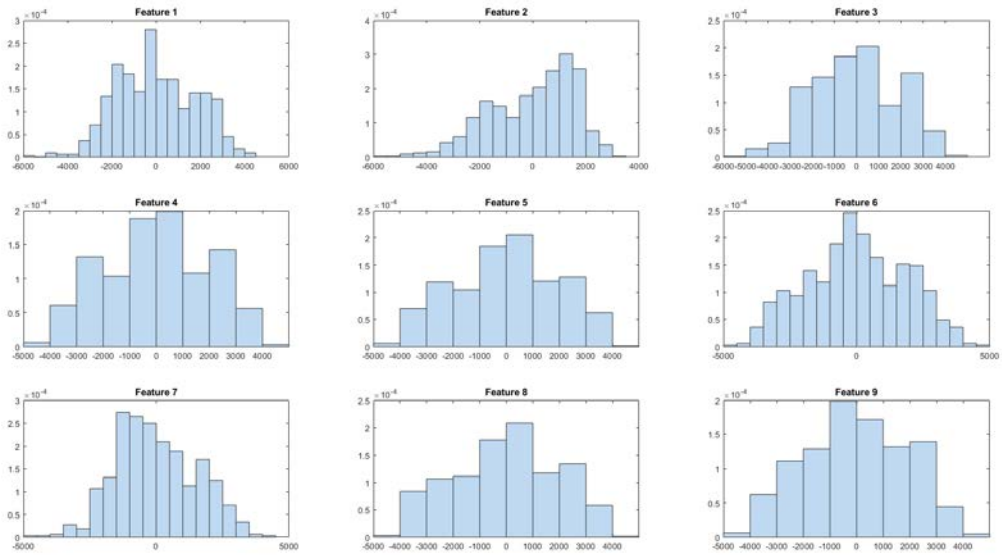
There are two main assumptions when using linear regression model: independent features and somewhat linear trend of the data. Although some linear regression models such as the Bayesian model can handle correlations between features, if there are correlated features, linear regression fails miserably because the model would be heavily biased towards those correlated data. Also, the data has to show some linearity; if the data is all over the place with no observable or probable linear relationship between the features, then the model is not going to be accurate in even predicting the training data.

Model Analysis

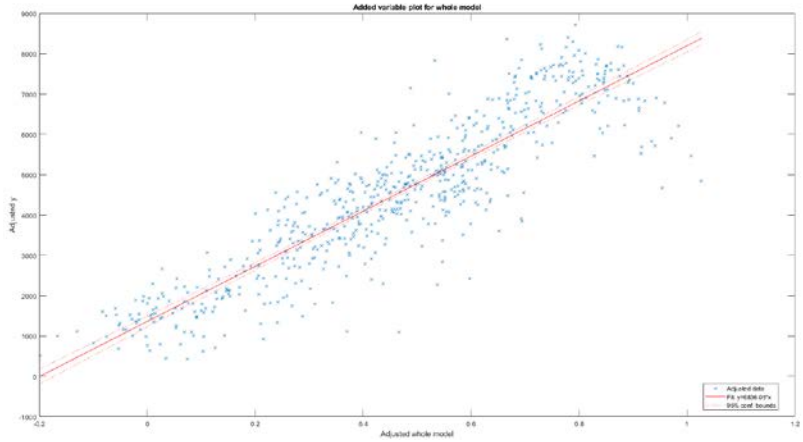


Looking at the residual plot of the entire data set on the left, it is clear to see that the amount of data on the top half of the plot and the bottom half seem to be very similar. Although, the numerical value of some of the residuals on the bottom half seem to reach towards extremities to the point of where they can be considered

outliers. However, it is just not a single point that is an outlier but rather a couple of points, the data points are left in the analysis.



Looking at the features individually, the figure on the right show all of the features demonstrating normality. With the except of feature 2 which seems to be slightly skewed to the left, These error distribution plots reassures us that the features that were used in creating a linear regression model is fairly accurate or to be more exact, has a good chance that the model will likely predict the outcome accurately.



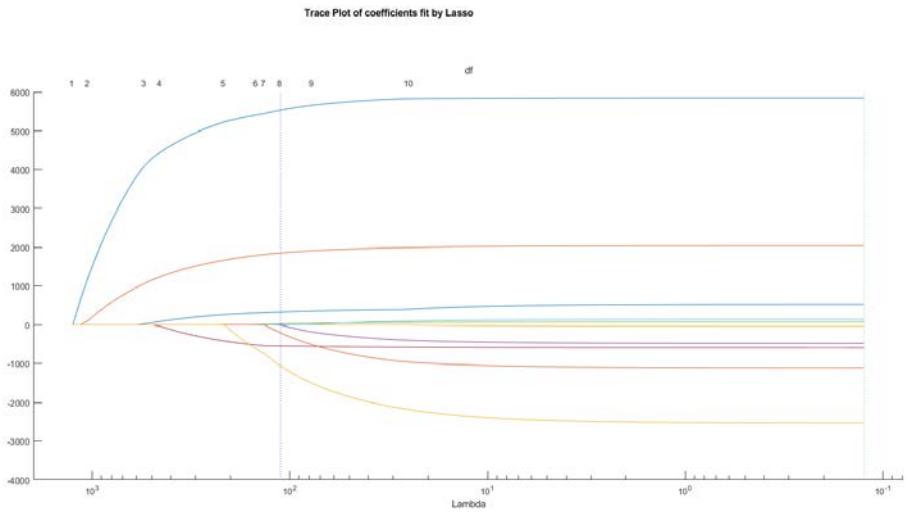
When the data is plotted with the linear regression model along it, it is clear to see that the data do in fact follow the model in a linear fashion. However, the majority of the data points seem to be outside the 95% confidence boundary.

These graphs demonstrate that the linear regression model have in fact been trained somewhat accurately with the data. The error, however, seems to be quite large to the point where the model can be questioned for its accuracy. There is an explanation for this particular outcome: the nature of the given data. Although the data generally follow a linear model, because the Y value is the number of people who biked on the given day, it is highly unlikely that the number of people biking on a given day will be constant no matter how similar the weather conditions and other data features are kept. There are other variables that also contribute to this: the behavior of humans. Some variables to keep in mind when analyzing this data is the type of people biking. Athletes will more likely bike on a regular basis no matter the weather condition while average people will take weather more into consideration. Other factors such as the number of tourists, the efficiency and establishment of other transportation methods in the area, etc. also contribute to the large variance in the number of people using bikes.

Although large error is the case in this particular data set, the numbers produced on the statistical side of the model is not too terrible. Looking at the figure on the right where the linear regression model is shown, it can be observed that all of the p-values are less than 0.05 or 5%. The adjusted R^2 is 0.805 which basically means that 80.5% of the variance in data can be explained by this model. The root mean squared error, however, is definitely larger than what would be acceptable but the reasons were given earlier that explains this phenomenon.

```
mdl =


Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9

Estimated Coefficients:
                   Estimate      SE       tStat       pValue
                   ________    _____    _______    ___________

    (Intercept)      1295.5    235.18     5.5085    5.2241e-08
    x1               530.24    54.683     9.6966    7.5005e-21
    x2               2023.5    67.415     30.016    9.8157e-125
    x3              -42.98     16.955    -2.5349     0.011481
    x4              -540.38    205.95    -2.6238     0.0088997
    x5               62.856    16.774     3.7472     0.0001948
    x6              -543.78     80.83    -6.7274     3.8e-11
    x7               6054.8    227.15     26.655    2.985e-106
    x8              -1094.3    317.89    -3.4425     0.00061345
    x9              -2293.5    463.04    -4.9532     9.3243e-07


Number of observations: 658, Error degrees of freedom: 648
Root Mean Squared Error: 854
R-squared: 0.808,  Adjusted R-Squared 0.805
F-statistic vs. constant model: 303, p-value = 1.32e-225
```

In order to make sure that backward step selection was able to produce an acceptable model, lasso shrinkage method was used at the end to compare and contrast the models produced by the two different methods on the same type of data. As seen on the left figure, the lasso behaves similarly: the shrinkage method zeros out the same features that were discarded out during the process of backward step selection. Comparing the Root MSE produced by these two models, backward step selection produced about 854 while lass produced about 861.29 which are very close numbers. Thus, it is safe to say that the backward step selection method used earlier was able to produce an acceptable model.



Conclusion

```
The average number of users in Janurary for Los Angeles is: 4831.400502 vs. 2225.642857
The average number of users in February for Los Angeles is: 4968.725715 vs. 2664.666667
The average number of users in March for Los Angeles is: 3616.396299 vs. 3630.127273
The average number of users in April for Los Angeles is: 3917.960703 vs. 4515.910714
The average number of users in May for Los Angeles is: 4182.246629 vs. 5339.363636
The average number of users in June for Los Angeles is: 4886.803963 vs. 5674.339623
The average number of users in July for Los Angeles is: 5130.398179 vs. 5655.909091
The average number of users in August for Los Angeles is: 5195.458178 vs. 5641.732143
The average number of users in September for Los Angeles is: 5298.696152 vs. 5861.226415
The average number of users in October for Los Angeles is: 4522.127760 vs. 5369.175439
The average number of users in November for Los Angeles is: 4408.786642 vs. 4221.313725
The average number of users in December for Los Angeles is: 4788.579167 vs. 3362.150000
```

Los Angeles is a vast area with tightly packed attractions and wonders to explore. Although Los Angeles does have a public transit system, more and more people are using bikes in the area as can be seen from the large coefficient (2023.5) on feature x2 (year) [the figure with linear regression model output]. On top of that, Los Angeles is a tourist attraction with a countless number of people coming in and out of the area to visit as many places as possible. The cheapest option to explore around the city of Los Angeles, other than walking, is by using bikes. By implementing a bike sharing program, all different kinds of people such as workers, athletes, students, tourists, and others will be able to enjoy the quick, cheap, and efficient way of going around the city.

Some might worry that weather conditions greatly impact the number of bike users. That is true; however, it is slightly different for the city of Los Angeles. Because Southern California climate has less articulated seasons other than for summer, there isn't much decline in usage of bikes other than during the summer when the weather can be too hot to bike around. Throughout the entire year of Los Angeles climate (data collected from multiple sources that provided the average temperature, humidity, wind speed, etc. for every month), the number of bike users stayed relatively stable around the upper 4000 users. This can only represent that bikes will always be in demand no matter what time of the year it is in Los Angeles.

Not only will having a bike sharing program offer an efficient method of transportation for everyone who are wanting to get around the city of Los Angeles, the program will definitely bring a stable income due to the fact that there will always be a demand for bikes no matter what season it is. In addition to these, bikes are also a step closer in saving our environment and having a less polluted air which the majority of major cities around the world seem to have. With all of these reasons, I highly encourage that you implement a bike sharing program for the city of Los Angeles.