COSC 425 Project 1
Ryan Pauly
University of Tennessee Knoxville

## Project Description

The goal of project 1 is to implement a type of linear regression model for multiple independent variables, or multivariate linear regression. With this regression model we will be able to solve for coefficient values which can be used to help more accurately predict or estimate a given dependent variable with its respective independent variables. In this project, we are given a .data file which consists of rows and columns of information related to cars. Miles per gallon, cylinder count, weight, and even model year are just a few of the column types, and our objective is to use these numeric values to estimate miles per gallon. To do this we need to portion off some of the .data rows for training with the goal of calculating a strong coefficient matrix in order to accurately predict miles per gallon. The remaining portion is to be used for analysis of accuracy between the estimate and actual miles per gallon, the test matrices.

## Pre-processing Steps

Upon receiving and looking over the data file, there are two things that immediately stood out. The more obvious of the two is that for some of the rows, the horsepower independent variable had a '?' listed rather than a numeric value. There is a plethora of ways to deal with errors in a dataset. For this project, to deal with this issue and to take advantage of all the rows of data available, I took the average horsepower of the entire horsepower column and filled this average in the place for any invalid or non-numeric horsepower values.

Additionally, the other yet less obvious attribute to the rows in the .data file deals with an uneven distribution of miles per gallon values. Upon closer inspection of the data, one will find that the miles per gallon (mpg) grows larger near the end of the .data file, the value gradually increasing from the beginning to the end of the column. If the training and testing were separated simply by taking the first and second half respectively, there would be a lack of larger mpg values for creating more accurate coefficient values. Essentially, if only the first half of the dataset were used for training, there would be less accurate estimates for larger mpg values. In order to create a more even distribution of samples for training, I implemented a randomization of the rows and output the new shuffled matrix to a .csv file so that the same shuffled dataset could be used for comparisons (i.e. standardized and original independent variables). An issue with shuffling the rows is that now the training dataset is at the mercy of how well the data was shuffled, which could lead to a less accurate estimate over a more thoughtful approach to selecting testing and training rows. Conversely, a shuffled dataset could also potentially yield a very accurate testing result, but again it is not guaranteed.

The data is split into two portions, training and testing. The training dataset is the first 318 of the 398 of the shuffled rows, and the testing dataset consists of the remaining rows. Essentially, the data has a 70/30 split, where 70% of the rows are sampled for training, and the remaining 30% are sampled for testing.

**Description**

In this project we utilize a multivariate linear regression model in order to estimate miles per gallon values. We are given seven numeric independent variables to calculate the dependent variable, i.e. miles per gallon. These seven independent variables are: cylinders, displacement, horsepower, weight, acceleration, model year, and origin.

To solve for miles per gallon given these independent variables, we need the following multivariate linear regression equation:

$$Y_i = \alpha + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \cdots + \beta_n x_i^{(n)} \quad (1)$$

The dependent variable $Y_i$ is in this case the miles per gallon value we wish to estimate for the $i^{th}$ row. $\beta$ represents the coefficient values for the $n^{th}$ attribute of the respective columns (independent variables). The *x* term represents the individual independent element of an attribute column and row, where $i$ and $n$ represent its respective row and column. The constant α is the error term and is further explained with the coefficient and independent variable matrix.

While these variables could be equated utilizing for-loops, it is far more efficient to use them as intended, in a matrix. More specifically, linear/matrix algebra is the key to solving these equations. There are three main matrices we need, one for the dependent, independent, and coefficients. First, the dependent matrix is the following:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \dots \\ Y_m \end{bmatrix} \quad (2)$$

The *Y* matrix is very straightforward, it is a vertical matrix consisting of a mpg value for each row.

$$X = \begin{bmatrix} X_1 \\ \dots \\ X_m \end{bmatrix} \quad (3)$$

The *X* matrix is a little more complex, where each *X* has *m* rows. Each *X* term contains 8 attributes, or *n + 1* attributes, where *n = 7* for our 7 attributes, and the additional column consists of the all 1 column for the α term. To finish creating matrix *X*, we thus need to concatenate a column of 1's with respect to the number of rows in matrix *X* to the original independent variable matrix. This 1 column is required for receiving the constant value α in equation 1 from the coefficient matrix.

$$C = \begin{bmatrix} \alpha \\ \beta_1 \\ \dots \\ \beta_n \end{bmatrix} \quad (4)$$

The final matrix is the coefficient matrix C. Where α is the coefficient of the constant term α in equation 1 above. Of course, $\beta_n$ is a coefficient for each respective attribute.

The first objective is to utilize a training set from our cleaned and shuffled dataset to calculate coefficients in order to estimate Y given a testing set of independent attributes and its real Y for observing accuracy. Firstly then, we need to use matrices 2 and 3 to calculate for the coefficient matrix 4:

$$C = (X^T X)^{-1} X^T Y \tag{5}$$

$X^T$ is the transform of matrix *X*. Further, because we are using matrices, we will be finding the dot product with each multiplication. We will use the dot product again for the training dataset in which we estimate *Y*. To calculate matrix *Y*, we need the following equation:

$$Y = XC \tag{6}$$

Equation 6 is straightforward. As mentioned previously, we simply take the dot product of both matrix *X* and *C* to estimate matrix *Y*, which is just a single column of the estimated mpg values based off the independent variables *X* and the previously calculated coefficients.

Further, we wish to compare how standardizing the data compares to non-standardized data. To standardize the dataset, we perform a calculation which scales the independent variables down significantly. The goal with standardization is to scale the data down to a mean which equals 0 and a standard deviation equal to 1. To achieve this, we use the following formula:

$$z = \frac{X - \mu}{\sigma} \tag{7}$$

*X* represents the value we wish to standardize, and *Z* is the result of the standardization. We subtract from *X* the mean of *X* or $\mu$, and lastly, we divide it by the standard deviation of *X* or $\sigma$.

Finally, to calculate and show error between the actual and the estimation of the miles per gallon values, I decided to use the root mean squared error (rmse) formula. First, I calculated for the mean squared error (mse):

$$mse = \frac{1}{n} \sum_{i=1}^{n} (\mathrm{Y_i} - Y_i')^2 \tag{8}$$

In equation 8, $Y_i$ represents an element of the Y vector for n rows. This is simply computing the mean of squared errors. The rmse is thus simply the square root of the mse:

$$rmse = \sqrt{mse} \tag{9}$$

The reason we use equation 9 rather than just equation 8 is because rmse is more meaningful for describing the results it provides. Where if one used mse for computing the error between actual and estimated mpg, then the results would be squared mpg differences rather than simply the differences of mpg, making it somewhat easier to analyze. With equation 8 and 9, values close to zero are desirable, indicating a smaller difference between the estimate and actual.

**Analysis**

To best display the results found by using a multivariate linear regression model I chose to scatter plot both the actual and estimate for the miles per gallon versus each independent variable to show how well the predictors (coefficients) preformed on the test dataset. The first set of graphs are the original graphs with no standardization preformed.

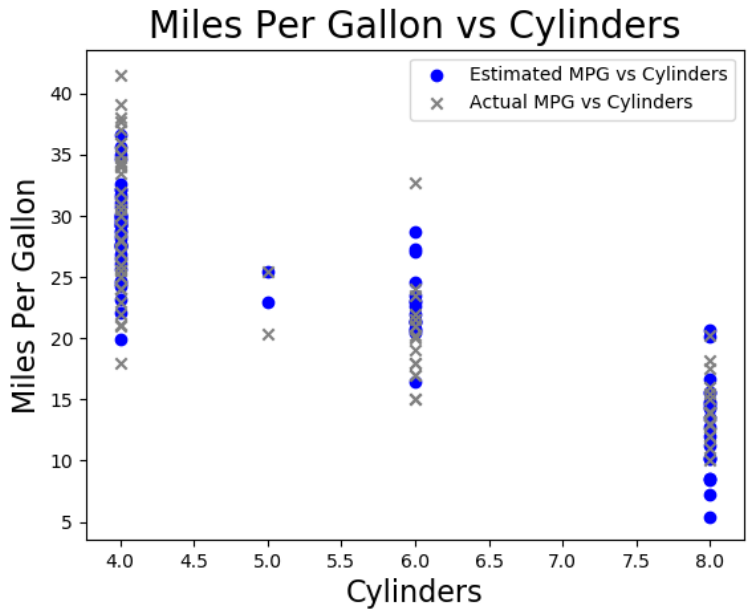


Figure 1. A scatter plot of the mpg dependent variable versus the multi-valued discrete independent variable of Cylinders.

With the mpg versus discrete cylinders graph, the scatter illustrates that for lower cylinder count the estimate was somewhat accurate, but for the ‘extreme’ low and high mpg values with a cylinder count of 4, the estimate struggles to make its mark. At 6 cylinders the estimate becomes less correct with respect to the actual mpg, but at 8 cylinders the estimate improves. There were likely too few training samples with a 6-cylinder attribute, thus this estimate was more askew.

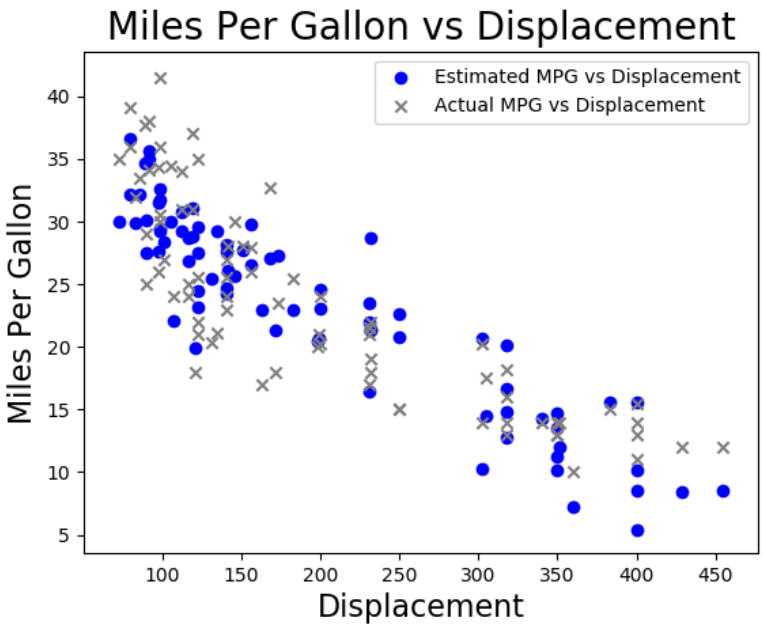


*Figure 2. A scatter plot of the MPG versus Displacement.*

One aspect of the displacement .data column I noticed was that high displacement seemed to be linked to lower mpg values and lower displacement values had some correlation to higher mpg values. Moreover, the mpg versus displacement scatter plot shows a low variance with lower displacement values and high variance with larger displacement values. The bias is high for lower displacement, but larger displacement values appear to have a slightly lower bias, most prominently between a displacement of 350 and 450. This would indicate that there are issues with both the testing and training error.

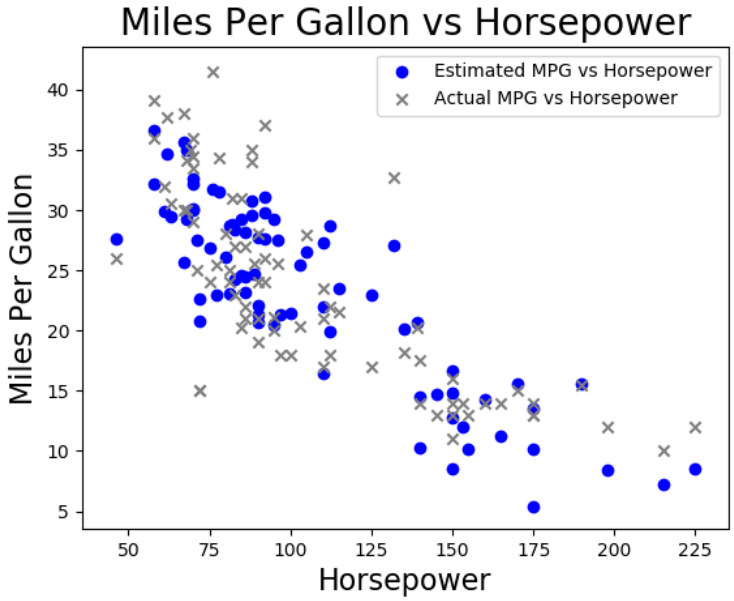


*Figure 3. MPG vs Horsepower scatter plot showing the estimated MPG and the actual MPG values.*

Overall, the estimated mpg scatter is relatively close to its respective actual. However, there certainly are portions that are better than others. There are two clusters between horsepower values of 75 and 100 where the variance is low and the bias questionable. More interestingly, despite the actual and estimate difference towards the end of the scatter plot, the shape of the estimate seems to mirror that of the actual. This may be due to having fewer training samples with a horsepower attribute of 200-225 with a miles per gallon value between 10 and 15. Here we have an estimate which is consistent but inaccurate, indicating that we have high bias but low variance.

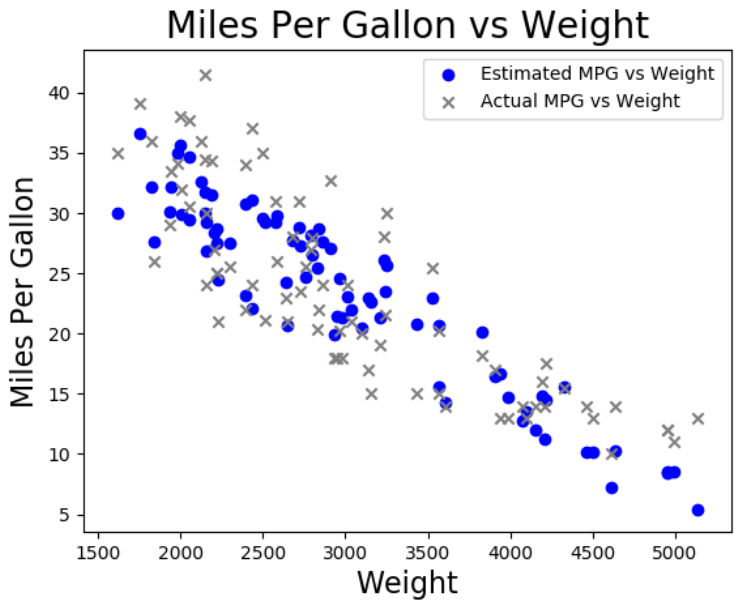


*Figure 4. MPG versus Weight scatter plot showing the estimated and actual MPG values.*

Again, at the beginning of the scatter plot there is a smaller variance in the estimated scatter, but as the weight increases the variance begins to increase in the middle portion of the weight, between 3000 and 4000. Then, interestingly at 5000 the estimate reflects the actual. This is likely due to having too few training samples with larger weight values and a miles per gallon value above 10. This scatter shows a high bias and low variance result. On average the estimate is consistent but is inaccurate.

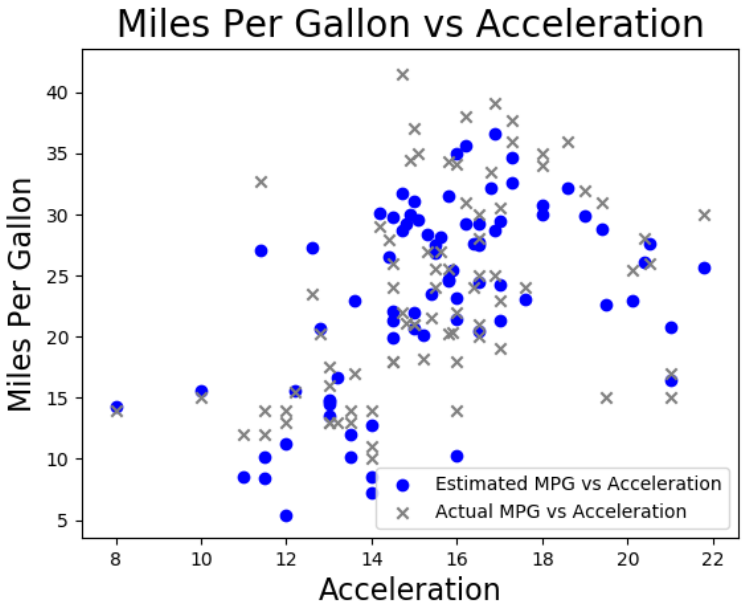


*Figure 5. MPG versus Acceleration, showing again both the estimate and actual MPG values.*

The bulk of the estimate scatter in figure 5 surprisingly does a decent job at capturing the 'shape' of the actual mpg scatter. Between an acceleration of about 11 to 14 and again between about 16 and 22, the actual mpg scatter is mirrored very closely by the estimate. Looking at the graph for the overall shape and 'big-picture,' the graph seems to have a high bias and low variance. It has high bias because the estimate is consistent with the shape of the actual, but low variance because the estimate is on average off mark of the actual.

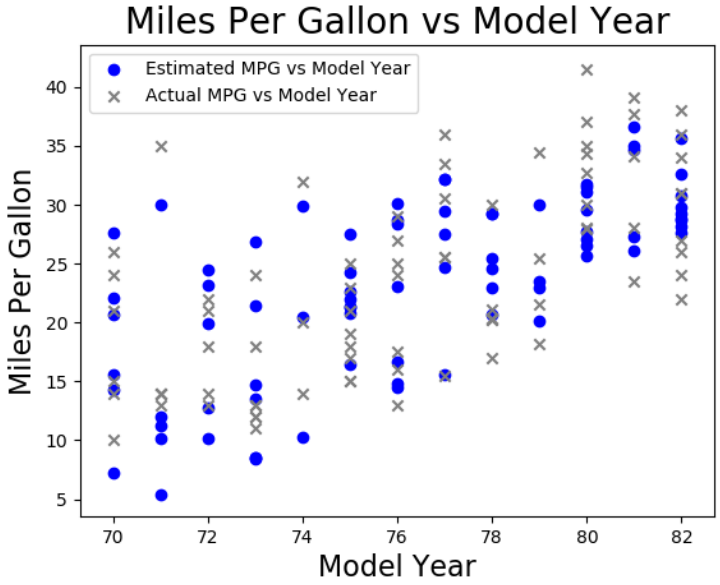


*Figure 6. MPG versus Model Year scatter plot comparing the actual and estimate MPG values.*

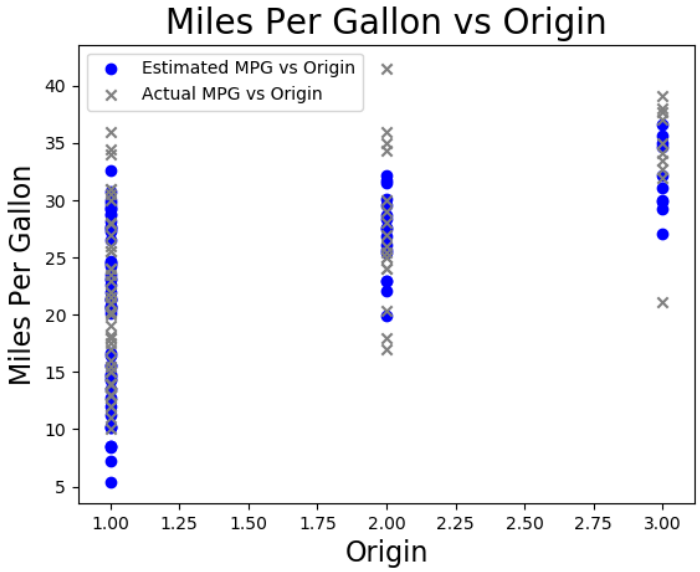


*Figure 7. MPG versus Origin scatter plot comparing the actual and estimate MPG values.*

The origin and model year scatter plots of figure 6 and 7, again show us a common result like that of the previous scatter plots. Most scatter plots have a high or mid-range bias with low variance, meaning the shape is typically consistent but inaccurate to the actual.

The next set of graphs are the root mean squared error for the mpg with respect to the estimate and actual mpg values along each of the independent attributes.

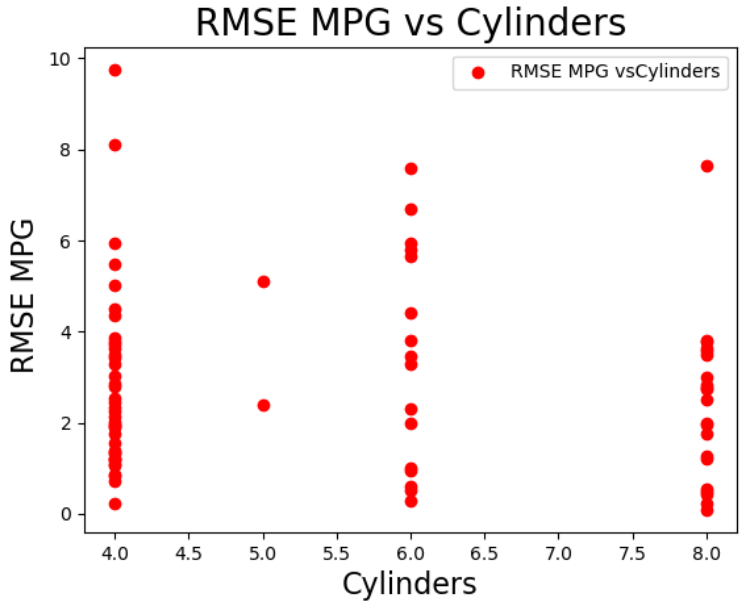


*Figure 8. A scatter plot of the root mean squared error of the estimate MPG versus cylinders.*

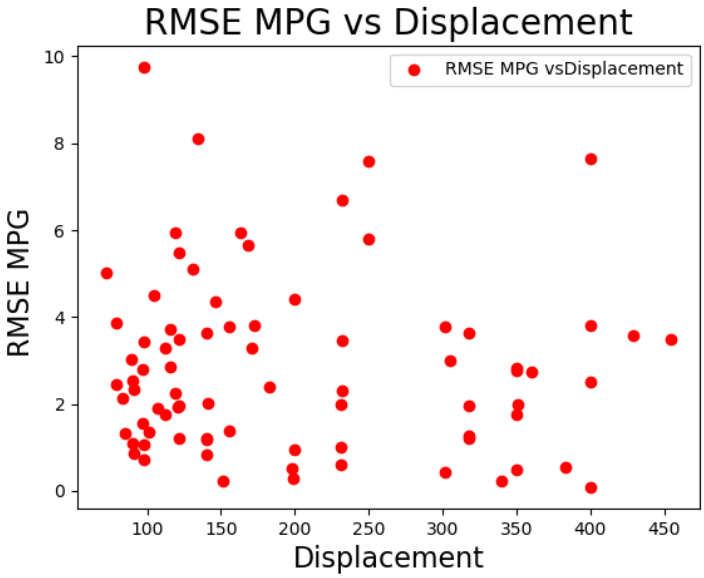


*Figure 9. A scatter plot of the root mean squared error of the estimate MPG versus displacement.*

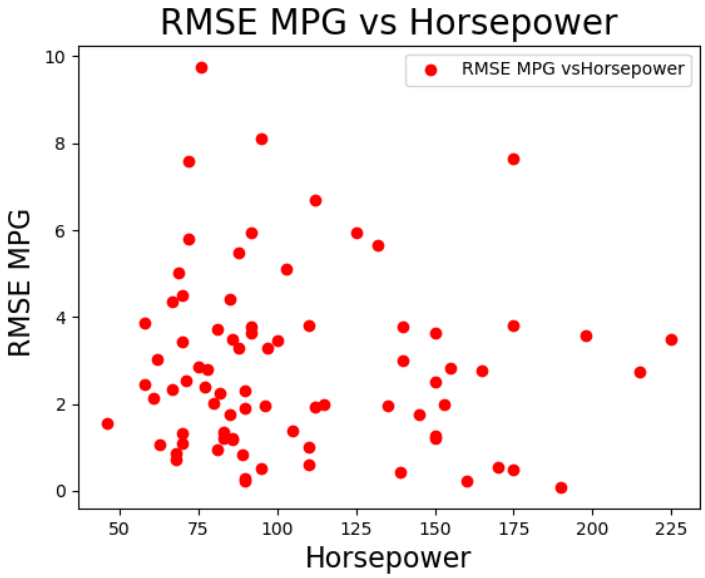


*Figure 10. A scatter plot of the root mean squared error of the estimate MPG versus horsepower.*

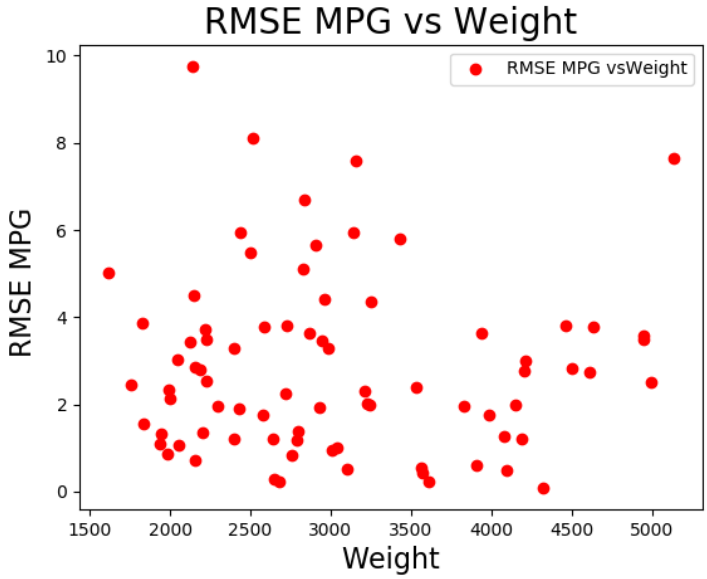


*Figure 11. A scatter plot of the root mean squared error of the estimate MPG versus weight.*

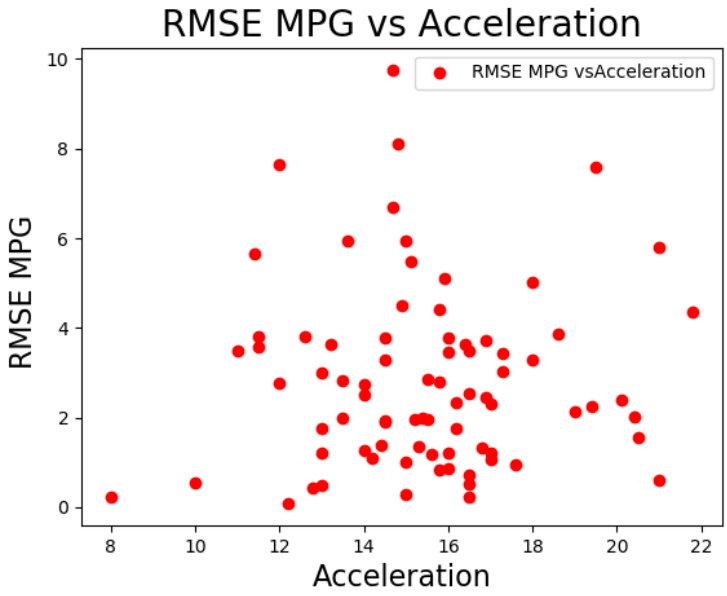


*Figure 12. A scatter plot of the root mean squared error of the estimate MPG versus acceleration.*

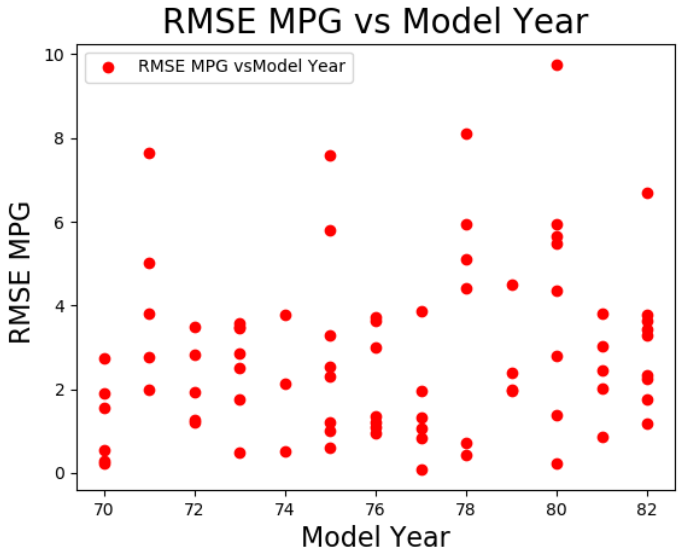


*Figure 13. A scatter plot of the root mean squared error of the estimate MPG versus model year.*

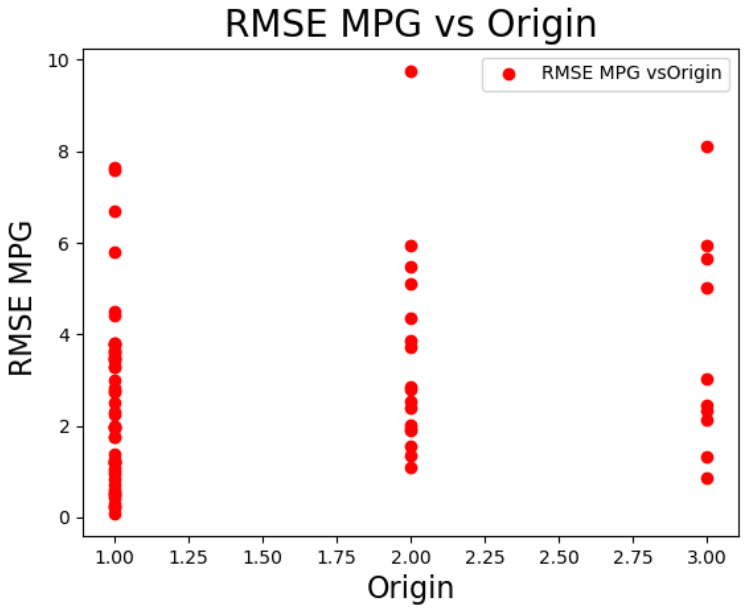


*Figure 14. A scatter plot of the root mean squared error of the estimate MPG versus origin.*

The root mean squared error scatter plots all show an okay result. The majority of the rmse results are below a 5-mpg difference between actual and estimate, and only a couple to a handful above 6 in most of the plots. There certainly is room for improvement.

```
Non-Standardized: Avg_rms =  2.834319496219041
```

*Figure 15. The Non-Standardized root mean squared error average value.*

The average root mean squared error value for the non-standardized data is roughly 2.83.

```
Matrix_C =

[[-1.66754058e+01]
 [-2.76682611e-01]
 [ 2.04802569e-02]
 [-1.44265241e-02]
 [-7.26822759e-03]
 [ 1.17226044e-01]
 [ 7.51085281e-01]
 [ 1.19102605e+00]]
```

*Figure 16. A matrix of the coefficient values calculated with the training dataset.*

The coefficient matrix of figure 15 is very straightforward. With equation 5, the C matrix can be solved with a dedicated training dataset. The first value is the α value, the constant coefficient found in equation 1. Then the following coefficients are multiplied to each independent variable, as shown in equation 1. The values determined for each attribute tells us how each attribute is 'weighted,' but this will be further discussed in the discussion section.

The scatter plot graphs for the standardized dataset all yield very similar results to the previous scatter plots for the non-standardized dataset. They are however slightly more accurate than the non-standardized set. Here are a few of these graphs which have the most noticeable improvements:

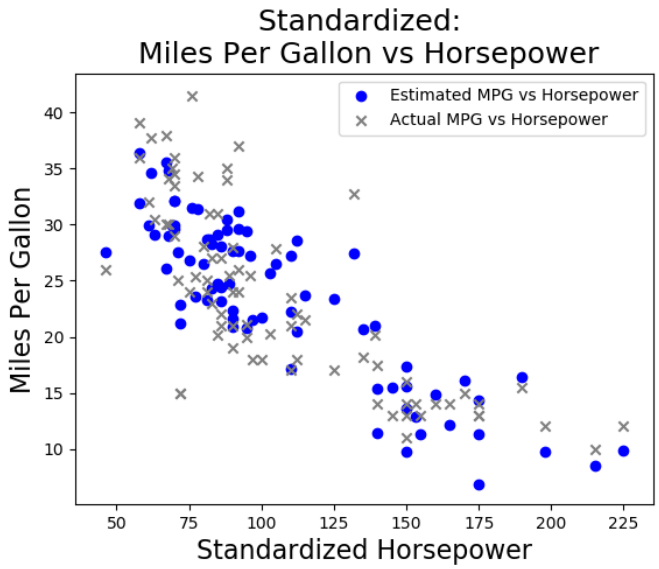


*Figure 17. Standardized: A scatter plot of MPG versus horsepower. Upon closer inspection one may notice a slightly lower variance than its non-standardized counterpart.*

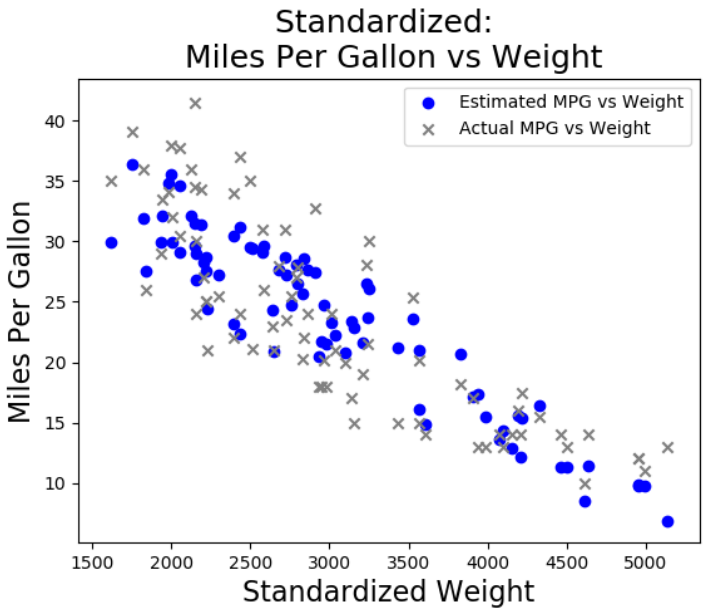


*Figure 18. Standardized: A scatter plot of MPG versus weight. Again, upon closer inspection one may notice a slightly lower variance than its non-standardized counterpart.*

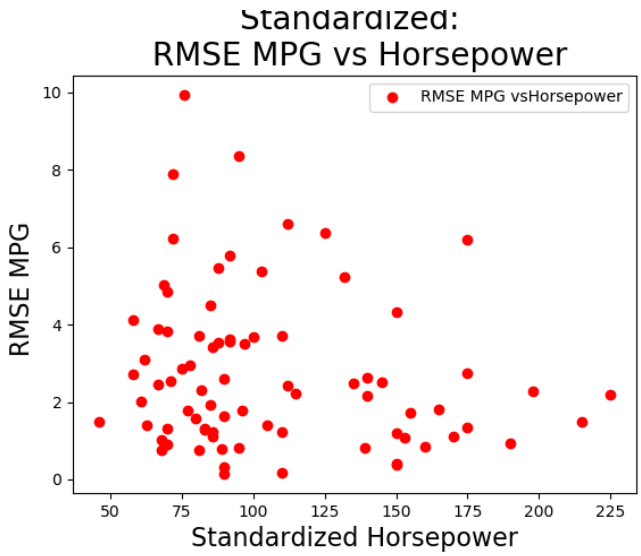


*Figure 19. Standardized: A scatter of the root mean squared error of the mpg estimate.*

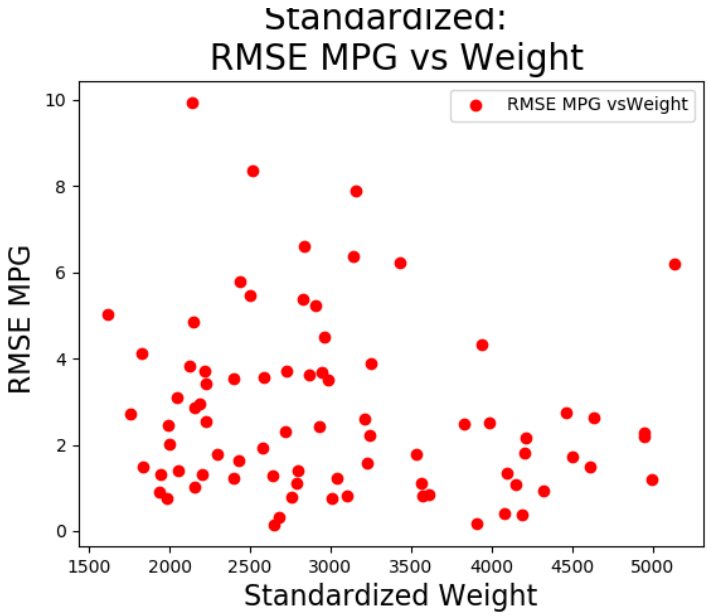


*Figure 20. Standardized: A root mean squared error scatter of the mpg estimate.*

The standardization of the dataset changes the results very little, but it did bring down the average root mean squared error compared to the non-standardized dataset. In other words, the standardized dataset did indeed improve upon the non-standardized predictions.

```
Standardized: Avg_rms =  2.7498395463904166
```

*Figure 21. The standardized average of the root mean squared error.*

```
Matrix_C =

[[-0.46309791]
 [ 0.02247869]
 [-0.03528406]
 [-0.00690737]
 [-0.0198661 ]
 [ 0.58534942]
 [ 1.06776834]]
```

*Figure 22. A matrix of the coefficient values calculated with a standardized training dataset.*

Because the estimate mpg values are now much closer to the actual mpg values, the average root mean squared error decreased slightly, but not overwhelmingly so. Additionally, the coefficients as seen in figure 22 are scaled due to the standardization of the dataset. Interestingly, the signs on some of the coefficients in figure 22 changed from that of figure 16.

**Discussion**

When first approaching a machine learning or statistical problem, understanding the data is likely one of the most crucial steps to reaching a sound conclusion. Upon investigation there were two main issues with the data we received in this project. That being that there were both missing horsepower values for some of the rows, and that the data was distributed in such a way that lower mpg values were ‘clustered’ at the beginning of the .data file and larger mpg values towards the end of the .data file. To solve this problem, the average horsepower of the valid horsepower values was calculated and inserted into any invalid horsepower elements. Then, to help give a fairer distribution of data for training and testing, the data rows were sampled and shuffled. There are likely better ways to both fix invalid horsepower values and select a fair testing and training set of the provided data, but these solutions certainly worked.

Multivariate linear regression is one of many statistical tools for prediction. There are many ways to go about using multivariate linear regression, from varying sample sizes for training and testing, to standardization or non-standardization, even normalization. With a 70/30 split of the data for training and testing respectively, the multivariate linear regression received an average root mean squared error of about 2.834 for non-standardized and 2.749 for standardized. These results are achieved by receiving a well distributed and shuffled training and testing dataset, but with an unfavorably shuffled dataset these results could yield higher error. Standardization is clearly a helpful step in decreasing error and is the superior method between non-standardization and standardization.

The final question is, which attribute(s) is the most determinative of the mpg value. My assumption is that this question can be solved by observing the max/min of the coefficient matrices. For non-standardized, weight has the greatest negative impact on miles per gallon, and model year had the greatest positive impact on miles per gallon. For standardized, the origin had the greatest positive impact on miles per gallon, and displacement had the greatest negative impact on miles per gallon. For the non-standardized, the max and min are very reasonable attributes to expect, weight certainly could be a detrimental part of a car’s mpg. Additionally, the model year, indicating that it is highly likely a newer more modern vehicle will have better mpg than an older model year, is understandably a strong factor in predicting mpg.

**Sources**


Alpaydin, Ethem. *Introduction to Machine Learning*. third ed., MIT Press, 2014.

"General Linear Model." *Wikipedia*, Wikimedia Foundation, 17 June 2019, https://en.wikipedia.org/wiki/General_linear_model.

"Mean Squared Error." *Wikipedia*, Wikimedia Foundation, 25 Sept. 2019, https://en.wikipedia.org/wiki/Mean_squared_error.

"Multivariate Linear Regression Tutorials & Notes: Machine Learning." *HackerEarth*, https://www.hackerearth.com/practice/machine-learning/linear-regression/multivariate-linear-regression-1/tutorial/.

Stephanie. "Standardized Variables: Definition, Examples." Statistics How To, 12 Oct. 2017, www.statisticshowto.datasciencecentral.com/standardized-variables/.