Introduction

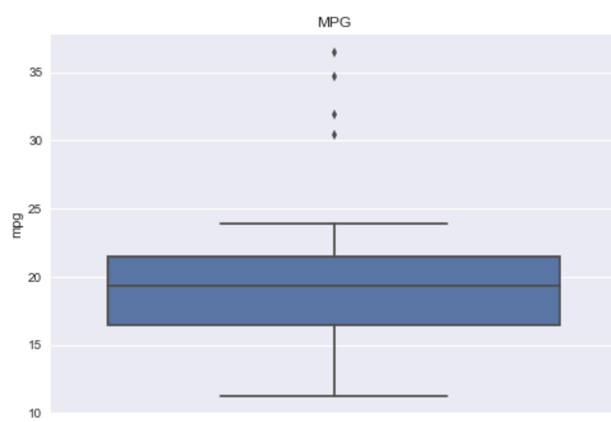
For this assignment, we are looking at fuel efficiency data published in 1975 by Motor Trend magazine. The data report fuel efficiency and 11 other factors for 30 different models of cars. We were asked to create a multiple regression model using all the variables as predictors, and a second model using a subset of the variables as predictors. The variables in the subset are chosen through an automated selection process.

Trying different methods for automation variable selection showed that 3 of the 4 methods attempted produced very similar results. The best model produced only explained ~75% of the variance in the data set. While feature selection is beneficial somewhat in this problem, additional data to improve the model coverage might be advisable before using it for making business level decisions.

Sample Definition

I did not do any reductions to the data. There are a small number of records, and even though a boxplot of MPG (Figure 1) shows a few outliers, I felt that this exercise would be best using all the data provided. A review of the outliers shows the MPG values to be rational based on real-world data. In an article in 2013 from Motor Trend[1], we are told there was a Honda Civic in 1975 which got 40 MPG. Based on this fact, I determined the outliers to be plausibly correct, and retained them in the set.

FIGURE 1

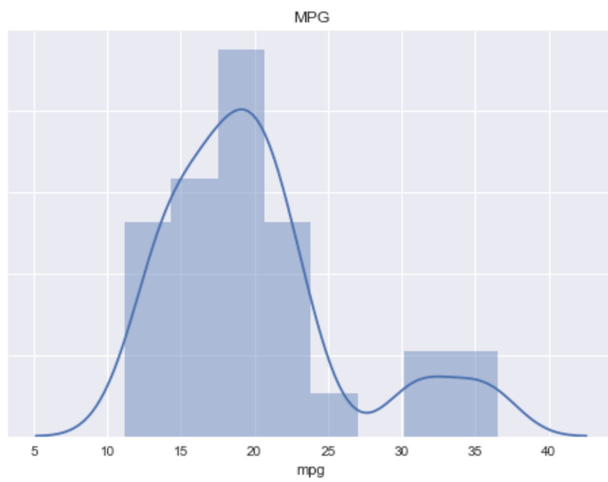


[1] *1975 HONDA CIVIC CVCC AND 1979 HONDA CIVIC CVCC WAGON, Innovative Import Made the Rules by Which All Other Economy Cars Are Judged*, Motor Trend, January 8, 2013. http://www.motortrend.com/cars/honda/civic/1975/12q2-1975-honda-civic-cvcc-1979-wagon/

Exploratory Data Analysis and Simple Linear Regression

I determined that there are no missing values by using the df.describe() method. This is important, since later on we will compare AIC results, and in order to compare models' AICs they must have a complete data set[2]. The distribution of the predictors variables does not form a normal curve. As seen in Figure 2, the distribution is bi-modal. We haven't really covered what to in this case; whether you split the data, or what transformation is appropriate. So, I am noting the departure from a normal distribution and continuing to work with the data set as is.

FIGURE 2

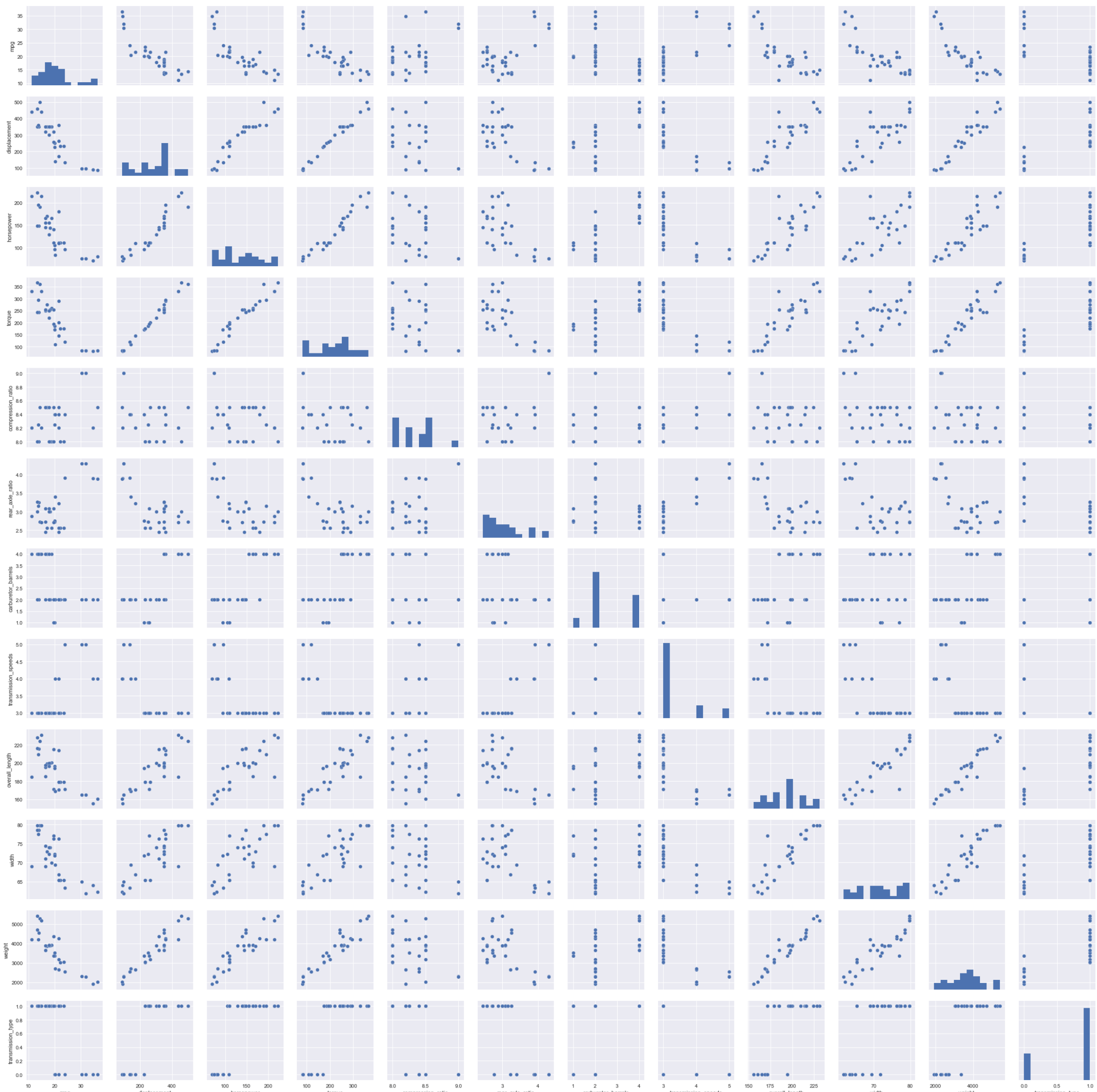


Using a scatterplot matrix to look at the relationship between variables gives us:

---

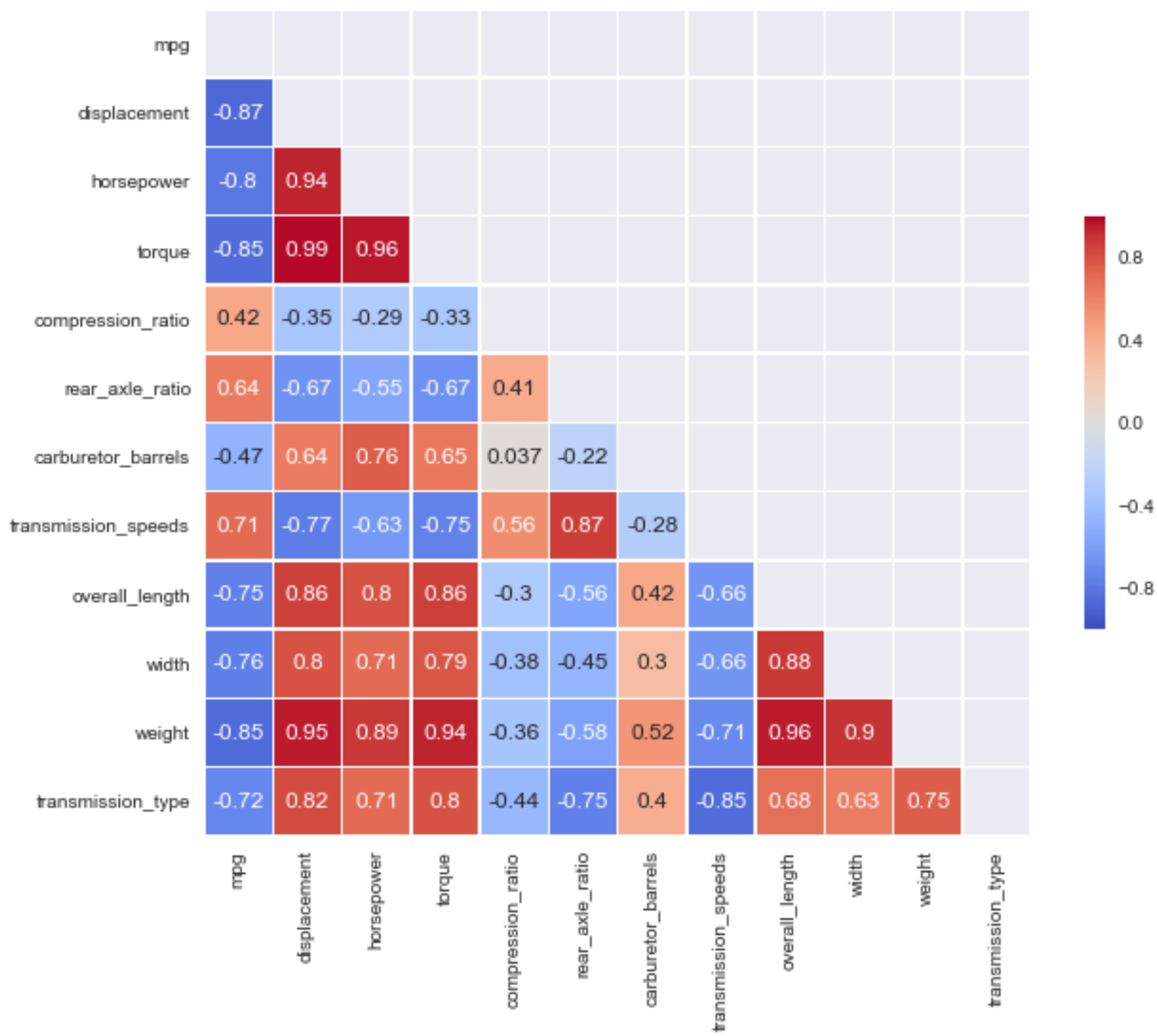
[2] *Regression Analysis by Example*, 5th Ed., Samprit Chatterjee and Ali S. Hadi, page 305

FIGURE 3



The scatterplot matrix shows a number of linear looking relationships between MPG and various predictor variables. However, for more than about 5 variables, I find these hard to read/interpret. As an alternate, we can use a correlation matrix to visualize the relationships. Figure 4 also shows the correlation values for the heatmap. It is faster to compute than the scatterplot matrix, and I find it easier to interpret and work with.

FIGURE 4



From Figure 4 we see that most of the correlations to MPG are negative. Displacement, Horsepower, Torque and Weight having the strongest correlations. Transmission Speeds is the strongest of the few positive correlations. Overall, the data have strong correlations between multiple variables; for example, “overall length” is strongly correlated with “displacement”, “horsepower” and “torque”, as well as strongly negatively correlated with “mpg”.

Multiple Linear Regression - Full Model

Fitting a model using all of the predictors gives us the following:

TABLE 1

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.835
Model:                            OLS   Adj. R-squared:                  0.735
Method:                 Least Squares   F-statistic:                     8.297
Date:                Thu, 20 Jul 2017   Prob (F-statistic):           5.29e-05
Time:                        15:46:48   Log-Likelihood:                -70.046
No. Observations:                  30   AIC:                             164.1
Df Residuals:                      18   BIC:                             180.9
Df Model:                          11
Covariance Type:            nonrobust
=========================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
Intercept              17.7732     30.509      0.583      0.567     -46.323      81.870
displacement           -0.0779      0.059     -1.330      0.200      -0.201       0.045
horsepower             -0.0734      0.089     -0.825      0.420      -0.260       0.113
torque                  0.1211      0.091      1.326      0.201      -0.071       0.313
compression_ratio       1.3290      3.100      0.429      0.673      -5.183       7.841
rear_axle_ratio         5.9760      3.159      1.892      0.075      -0.660      12.612
carburetor_barrels      0.3042      1.289      0.236      0.816      -2.404       3.012
transmission_speeds    -3.1986      3.105     -1.030      0.317      -9.723       3.326
overall_length          0.1854      0.129      1.434      0.169      -0.086       0.457
width                  -0.3991      0.324     -1.233      0.234      -1.079       0.281
weight                 -0.0052      0.006     -0.881      0.390      -0.018       0.007
transmission_type       0.5987      3.021      0.198      0.845      -5.748       6.945
==============================================================================
Omnibus:                        0.612   Durbin-Watson:                   1.890
Prob(Omnibus):                  0.736   Jarque-Bera (JB):                0.619
Skew:                           0.302   Prob(JB):                        0.734
Kurtosis:                       2.638   Cond. No.                     1.96e+05
==============================================================================
```

FIGURE 5

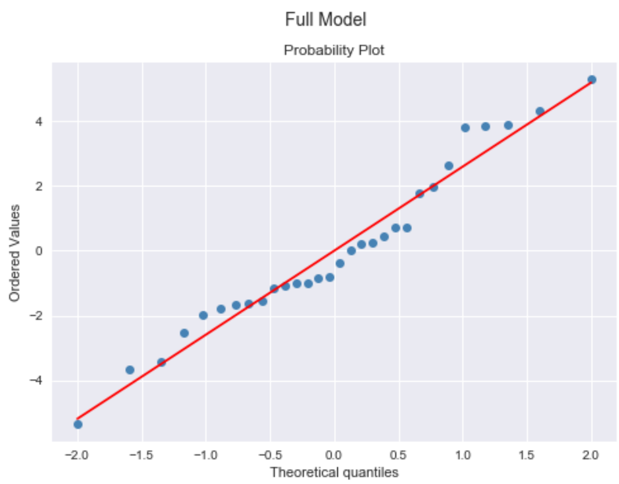
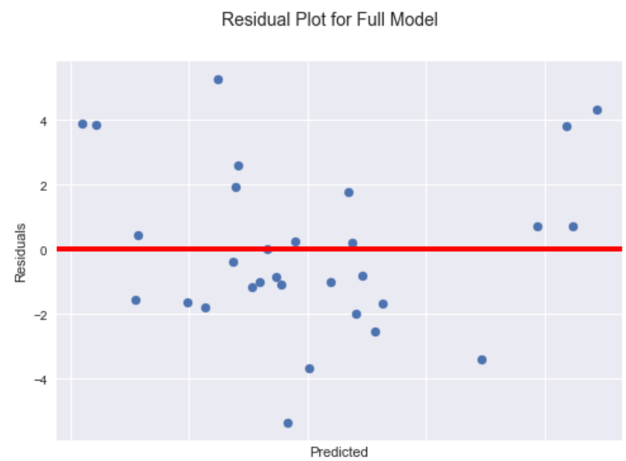


FIGURE 6



We can see from the information above, that the full model accounts for roughly 84% of the variance in the data set, the data have a slight skew and are leptokurtic. The F-statistic for the model is ~8.3 with a p-value of 5.29-e5, which is significant for alpha = .01; so, we reject the null hypothesis that there is no significant relationship between the predictors and the response variable. There is 10% gap between R^2 and adjusted-R^2, indicating there may be non-significant predictors in the model. The QQ plot shows a close-to-normal distribution, and the plot of residuals is decently close to normal.

Multiple Linear Regression - Subset Model

To select from the options for automated feature reduction, I experimented with a number of methods. Several are listed in the SKLearn API documentation as Automated Feature selection methods, but we know from class that Ridge can also be used to select features. While not covered in the text, Lasso is very similar to Ridge, with the main difference being it can reduce features by driving some coefficients to 0. The code for my experimentation is included in the .py file for the assignment. The output for 4 of the 5 models were very close, with RFE being decidedly less effective. Of the models, the Lasso results were the best, so I elected to use that as my Subset model. Details on the other models are shown in the appendix.

I ran the Lasso model, found the variables it selected, and used those in an OL regression to permit comparison to the Full model. The fitted results for the reduced model are:

TABLE 2

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.787
Model:                            OLS   Adj. R-squared:                  0.743
Method:                 Least Squares   F-statistic:                     17.74
Date:                Thu, 20 Jul 2017   Prob (F-statistic):           2.27e-07
Time:                        16:00:06   Log-Likelihood:                -73.899
No. Observations:                  30   AIC:                             159.8
Df Residuals:                      24   BIC:                             168.2
Df Model:                           5
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept            22.0313     25.656      0.859      0.399     -30.920      74.982
rear_axle_ratio       1.0229      1.634      0.626      0.537      -2.351       4.396
width                -0.1857      0.267     -0.694      0.494      -0.738       0.366
displacement         -0.0341      0.021     -1.631      0.116      -0.077       0.009
weight            -9.607e-05      0.003     -0.030      0.976      -0.007       0.007
compression_ratio     2.1929      2.419      0.907      0.374      -2.799       7.185
==============================================================================
Omnibus:                        1.451   Durbin-Watson:                   2.105
Prob(Omnibus):                  0.484   Jarque-Bera (JB):                0.985
Skew:                           0.442   Prob(JB):                        0.611
Kurtosis:                       2.931   Cond. No.                     1.67e+05
==============================================================================
```

FIGURE 7

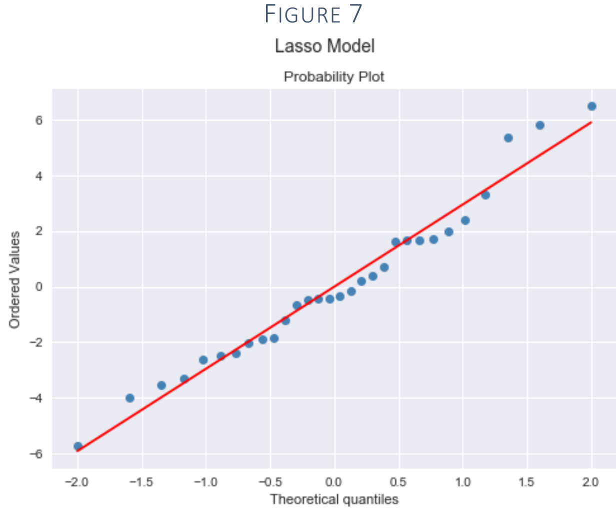
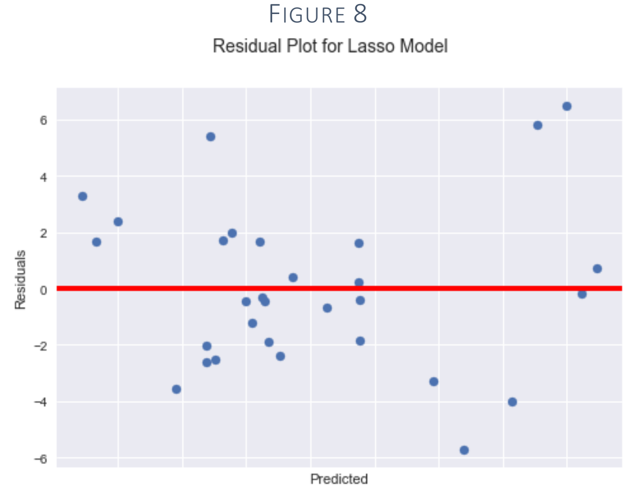


FIGURE 8



In this model, we account for ~79% of the variance in the data set. The F-statistic of 17.74 has a p-value of 2.27e-7, so again we'd reject the null hypothesis for the model. The R^2 versus adjusted-R^2 comparison shows a drop of 4.4%, indicating there may still be some non-significant variables in the model. Both the QQ plot and scatterplot of residuals appear sufficiently close to normal to allow us to work with the models in confidence we will not be "bitten" by a departure from normalcy.

Model Comparisons and Recommendation

Comparing the Full and the Subset models, we first see that the adjusted-R2 values are very close. The model created by using the Lasso method to eliminate features with 0 for coefficients has a slightly better adjust-R^2. This means the Subset model accounts for more of the variance than does the Full model.

Per our textbook[3] when comparing AIC values, we want to consider whether the values differ by more than 2 and if so, select the model with the lower AIC. The Full model has an AIC of 164.1 and the Lasso-selected model has an AIC of 159.8, giving us a difference of 4.3. Per guidance from our text, we should select the Subset model. Given the larger adjusted-R^2 value and the lower AIC value, I would recommend using the simpler, Subset model. The final model uses only 5 of the original 11 variables as predictors. In this situation, storage space and compute time are not issues, but for very large data sets, eliminating more than ½ of the variables, while improving over the Full model results can be a big win.

This model can be used to predict the MPG for cars whose rear axle ratios, widths, displacement, weights, and compression ratios fall with the range of the existing data. That is to say, the predictor variables must fall within the ranges in Table 3.

TABLE 3

	Compression ratio	Displacement	Rear axle ratio	Weight	Width
min	8	85.3	2.45	1905	61.8
max	9	500	4.3	5430	79.8

Predictions on variable values outside the range used to create the model are not reliable. Therefore, the model should **not** be used to predict the MPG of say for example, a 6614 pound, 81.3-inch-wide Hummer 2. On a pragmatic note, the model was created using only 30 observations, which are more than 40 years old. Models built with small numbers of observations may not be as robust as models built using more observations. Additionally, state-of-the-art automotive design has progressed, and it is quite possible the data are no longer representative of current car production. While the model may work well for vehicles produced within a few years of 1975, extreme caution should be exercised in trying to use the model to predict MPG for new automotive designs, or for designs of other vehicle-types, like trucks and motorcycles.

---

[3] *Regression Analysis by Example*, 5th Ed., Samprit Chatterjee and Ali S. Hadi, page 305

Conclusion

By reducing the number of predictor variables used in the multiple regression model, via techniques like Lasso Regression, we can improve the model's performance. However, with the data given, there is still roughly 26% of the observed variation in MPG unaccounted for.

If we wanted to use this model for prediction in a business setting, it might be advisable to gather data on additional parameters to see if the overall model could be improved. Having 26% of the variance unaccounted for seems like it would be too high to be useful in a real-world setting. Also. to predict the MPG for new cars, I would suggest that new data be collected. The data set is more than 40 years old, and unlikely to represent actual modern results.

Appendix

TABLE 4 – RIDGE SELECTION

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.800
Model:                            OLS   Adj. R-squared:                  0.736
Method:                 Least Squares   F-statistic:                     12.56
Date:                Thu, 20 Jul 2017   Prob (F-statistic):           2.17e-06
Time:                        16:00:06   Log-Likelihood:                -72.969
No. Observations:                  30   AIC:                             161.9
Df Residuals:                      22   BIC:                             173.1
Df Model:                           7
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept            29.3774     27.036      1.087      0.289     -26.691      85.446
weight               -0.0004      0.003     -0.116      0.908      -0.007       0.006
torque                0.0985      0.084      1.171      0.254      -0.076       0.273
displacement         -0.0750      0.043     -1.755      0.093      -0.164       0.014
compression_ratio     1.2764      2.575      0.496      0.625      -4.064       6.617
width                -0.2468      0.285     -0.866      0.396      -0.838       0.344
horsepower           -0.0585      0.072     -0.816      0.423      -0.207       0.090
rear_axle_ratio       2.2967      2.132      1.077      0.293      -2.125       6.718
==============================================================================
Omnibus:                        1.068   Durbin-Watson:                   1.996
Prob(Omnibus):                  0.586   Jarque-Bera (JB):                0.730
Skew:                           0.378   Prob(JB):                        0.694
Kurtosis:                       2.886   Cond. No.                     1.74e+05
==============================================================================
```

TABLE 5 – BACKWARD STEP SELECTION (5)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.786
Model:                            OLS   Adj. R-squared:                  0.741
Method:                 Least Squares   F-statistic:                     17.61
Date:                Sat, 22 Jul 2017   Prob (F-statistic):           2.43e-07
Time:                        10:26:35   Log-Likelihood:                -73.983
No. Observations:                  30   AIC:                             160.0
Df Residuals:                      24   BIC:                             168.4
Df Model:                           5
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept             18.6028     24.553      0.758      0.456     -32.073      69.278
displacement          -0.0497      0.024     -2.097      0.047      -0.099      -0.001
horsepower             0.0230      0.045      0.506      0.618      -0.071       0.117
compression_ratio      2.9762      2.667      1.116      0.275      -2.527       8.480
transmission_speeds   -0.5695      1.824     -0.312      0.758      -4.334       3.195
width                 -0.1448      0.180     -0.803      0.430      -0.517       0.227
==============================================================================
Omnibus:                        1.089   Durbin-Watson:                   2.079
Prob(Omnibus):                  0.580   Jarque-Bera (JB):                0.815
Skew:                           0.394   Prob(JB):                        0.665
Kurtosis:                       2.822   Cond. No.                     1.47e+04
==============================================================================
```

When evaluating RFE and KBest, I set the algorithm to select 5 predictors so I could compare the results easily with the Lasso results. Surprisingly, the Recursive Feature Elimination method produced a very low R^2 compared to the other models tried.

TABLE 6 - RFE (5)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.619
Model:                            OLS   Adj. R-squared:                  0.540
Method:                 Least Squares   F-statistic:                     7.812
Date:                Thu, 20 Jul 2017   Prob (F-statistic):           0.000175
Time:                        16:07:58   Log-Likelihood:                -82.608
No. Observations:                  30   AIC:                             177.2
Df Residuals:                      24   BIC:                             185.6
Df Model:                           5
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept            -11.2324     28.703     -0.391      0.699     -70.472      48.007
rear_axle_ratio        1.5310      3.113      0.492      0.627      -4.894       7.956
transmission_type     -3.8847      3.528     -1.101      0.282     -11.165       3.396
transmission_speeds    1.8749      3.329      0.563      0.578      -4.995       8.745
carburetor_barrels    -1.6589      0.834     -1.990      0.058      -3.379       0.061
compression_ratio      3.3019      3.601      0.917      0.368      -4.131      10.734
==============================================================================
Omnibus:                        3.826   Durbin-Watson:                   1.766
Prob(Omnibus):                  0.148   Jarque-Bera (JB):                2.724
Skew:                           0.732   Prob(JB):                        0.256
Kurtosis:                       3.195   Cond. No.                         369.
==============================================================================
```

TABLE 7 – KBEST

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:                       0.782
Model:                            OLS   Adj. R-squared:                  0.736
Method:                 Least Squares   F-statistic:                     17.18
Date:                Sat, 22 Jul 2017   Prob (F-statistic):           3.05e-07
Time:                        09:57:55   Log-Likelihood:                -74.277
No. Observations:                  30   AIC:                             160.6
Df Residuals:                      24   BIC:                             169.0
Df Model:                           5
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
Intercept             46.2011     17.712      2.609      0.015       9.646      82.756
transmission_speeds   -1.1072      2.447     -0.452      0.655      -6.158       3.944
rear_axle_ratio        2.3569      2.538      0.929      0.362      -2.882       7.596
displacement          -0.0367      0.023     -1.630      0.116      -0.083       0.010
weight                 0.0004      0.003      0.126      0.901      -0.007       0.008
width                 -0.2909      0.303     -0.960      0.347      -0.917       0.335
==============================================================================
Omnibus:                        1.117   Durbin-Watson:                   1.828
Prob(Omnibus):                  0.572   Jarque-Bera (JB):                0.942
Skew:                           0.411   Prob(JB):                        0.624
Kurtosis:                       2.720   Cond. No.                     1.14e+05
==============================================================================
```