STAT 6021 Project 1

Bobby Andris (rta8y), Hemani Choksi (hc8nd), Farhan Kanani (fk3ak), Camille Leonard (cvl7qu)

Section 1 - Exploratory Data Analysis

Initial exploration of the mileage data set was conducted by producing a scatterplot matrix and correlation matrix of the response and potential regressors (shown below in Figures 1-2). Variable x11 was excluded from the matrices because it was categorical. Linear relationships in the scatterplot matrix indicate correlation between the variables. The correlation matrix quantifies the correlations.

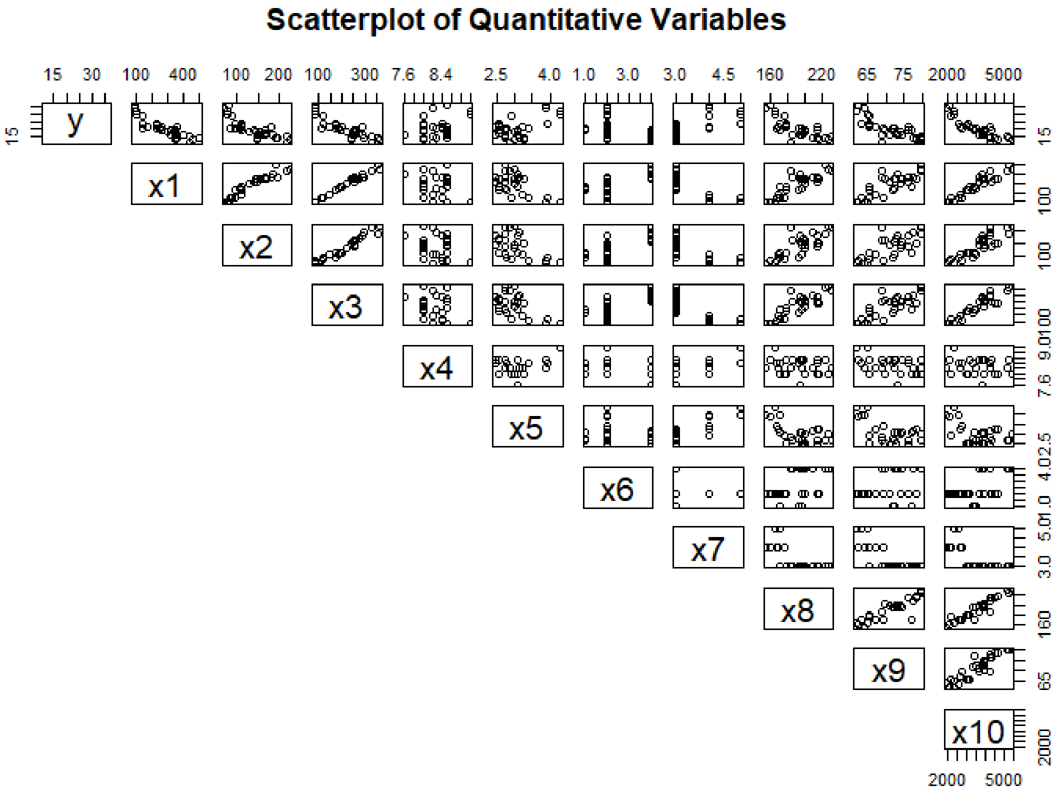


Figure 1 - A scatterplot matrix of the response variable, y, and possible regressor variables, x1-x10.

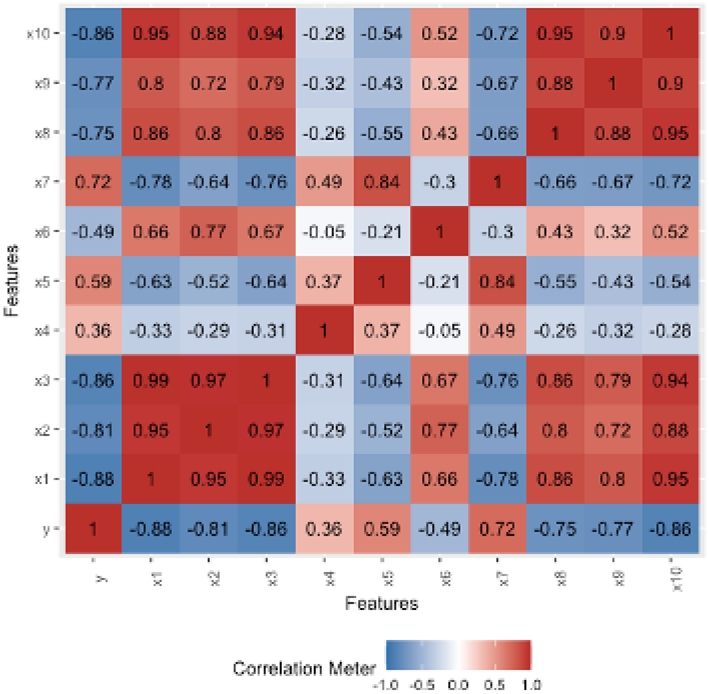


Figure 2 – Correlation chart that corresponds with Figure 1. Note, this matrix is rotated 90˚ counterclockwise compared to the scatterplot matrix.

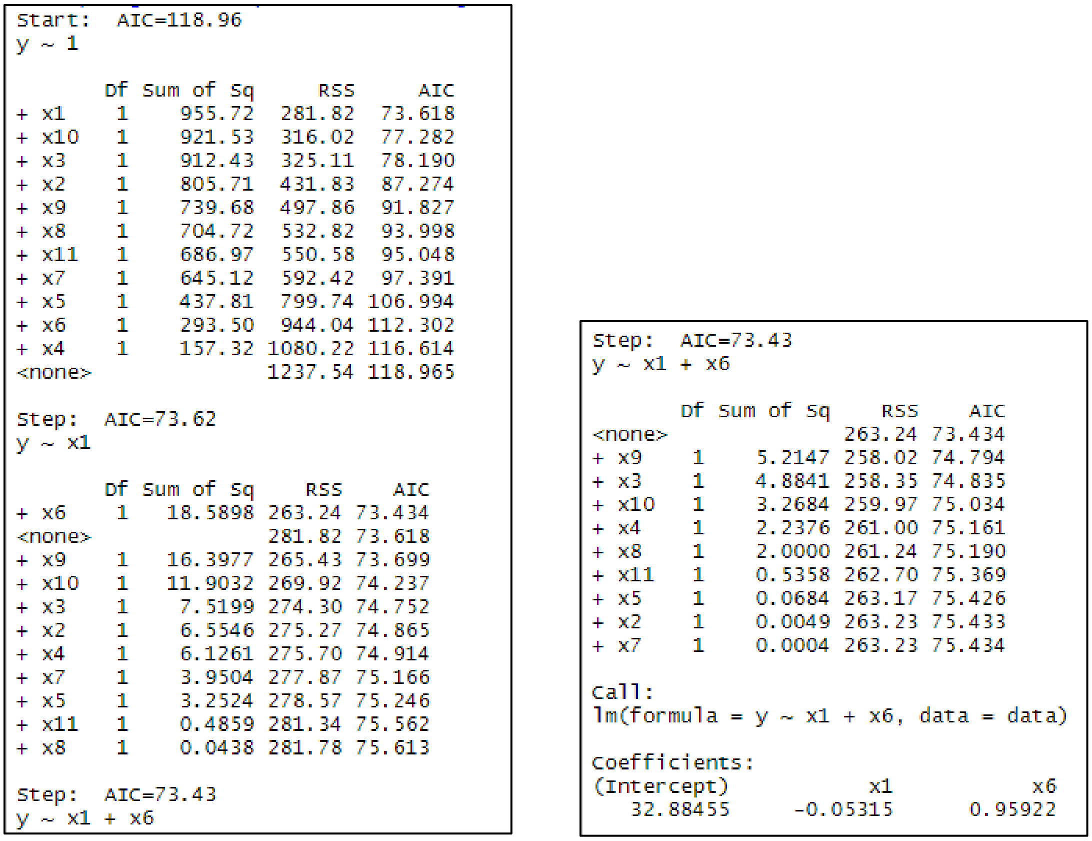
Based on Figures 1 and 2, variables x1, x2, x3, x8, x9, and x10 exhibit correlations with other data set variables. Variables x4, x5, x6, and x7 do not exhibit a significant correlation with other data set variables. We may be able to remove multiple variables from consideration in the models due to their correlation. Elimination of variables from consideration in the model due to multicollinearity was conducted via model selection.

Section 2 - Initial Model Considered

Section 2.1 - Forward Selection

We started by performing forward selection on an intercept only model against all the predictors from the mileage data set. Forward selection function returns the model with the lowest AIC.

The forward selection function returned a “best fit” model of gas mileage = intercept + x1(displacement) + x6(carburetor). The R output is provided in Figure 3.



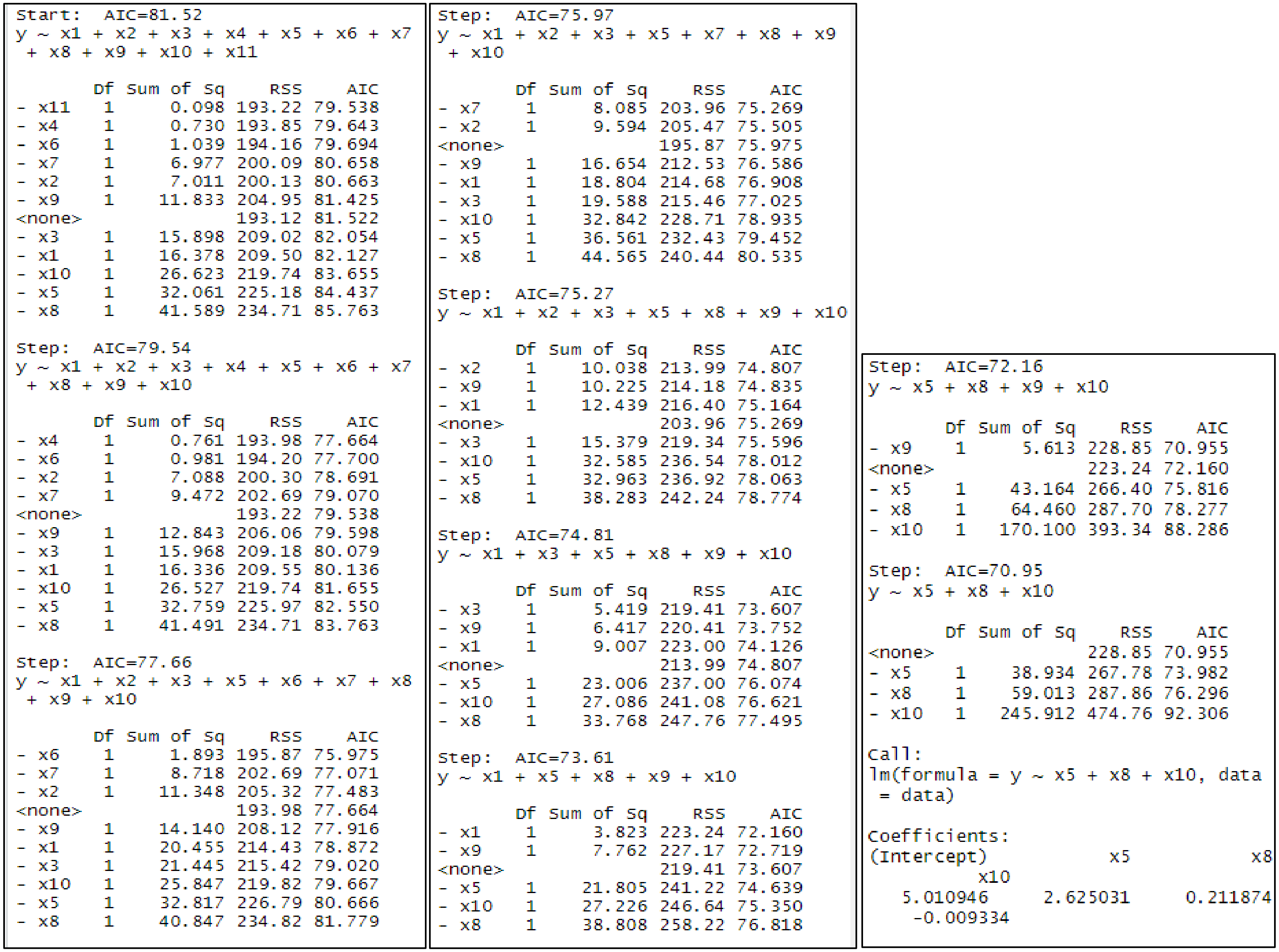
```
Start:  AIC=118.96
y ~ 1

       Df Sum of Sq     RSS     AIC
+ x1    1    955.72  281.82  73.618
+ x10   1    921.53  316.02  77.282
+ x3    1    912.43  325.11  78.190
+ x2    1    805.71  431.83  87.274
+ x9    1    739.68  497.86  91.827
+ x8    1    704.72  532.82  93.998
+ x11   1    686.97  550.58  95.048
+ x7    1    645.12  592.42  97.391
+ x5    1    437.81  799.74 106.994
+ x6    1    293.50  944.04 112.302
+ x4    1    157.32 1080.22 116.614
<none>              1237.54 118.965

Step:  AIC=73.62
y ~ x1

       Df Sum of Sq    RSS    AIC
+ x6    1   18.5898 263.24 73.434
<none>              281.82 73.618
+ x9    1   16.3977 265.43 73.699
+ x10   1   11.9032 269.92 74.237
+ x3    1    7.5199 274.30 74.752
+ x2    1    6.5546 275.27 74.865
+ x4    1    6.1261 275.70 74.914
+ x7    1    3.9504 277.87 75.166
+ x5    1    3.2524 278.57 75.246
+ x11   1    0.4859 281.34 75.562
+ x8    1    0.0438 281.78 75.613

Step:  AIC=73.43
y ~ x1 + x6
```

```
Step:  AIC=73.43
y ~ x1 + x6

       Df Sum of Sq    RSS    AIC
<none>              263.24 73.434
+ x9    1    5.2147 258.02 74.794
+ x3    1    4.8841 258.35 74.835
+ x10   1    3.2684 259.97 75.034
+ x4    1    2.2376 261.00 75.161
+ x8    1    2.0000 261.24 75.190
+ x11   1    0.5358 262.70 75.369
+ x5    1    0.0684 263.17 75.426
+ x2    1    0.0049 263.23 75.433
+ x7    1    0.0004 263.23 75.434

Call:
lm(formula = y ~ x1 + x6, data = data)

Coefficients:
(Intercept)           x1           x6
   32.88455     -0.05315      0.95922
```

Figure 3 – Forward selection R output for mileage data set.

Section 2.2 - Backward Elimination

Next, backward elimination was performed on a full model that contained all of the regressors from the mileage data set. Relevant R output is provided in Figure 4. The backward elimination suggested a “best fit” model of gas mileage = intercept +x5(rear axle ratio) + x8(overall length) + x10(weight).



```
Start:  AIC=81.52
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7
 + x8 + x9 + x10 + x11

       Df Sum of Sq    RSS    AIC
- x11   1     0.098 193.22 79.538
- x4    1     0.730 193.85 79.643
- x6    1     1.039 194.16 79.694
- x7    1     6.977 200.09 80.658
- x2    1     7.011 200.13 80.663
- x9    1    11.833 204.95 81.425
<none>              193.12 81.522
- x3    1    15.898 209.02 82.054
- x1    1    16.378 209.50 82.127
- x10   1    26.623 219.74 83.655
- x5    1    32.061 225.18 84.437
- x8    1    41.589 234.71 85.763

Step:  AIC=79.54
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7
 + x8 + x9 + x10

       Df Sum of Sq    RSS    AIC
- x4    1     0.761 193.98 77.664
- x6    1     0.981 194.20 77.700
- x2    1     7.088 200.30 78.691
- x7    1     9.472 202.69 79.070
<none>              193.22 79.538
- x9    1    12.843 206.06 79.598
- x3    1    15.968 209.18 80.079
- x1    1    16.336 209.55 80.136
- x10   1    26.527 219.74 81.655
- x5    1    32.759 225.97 82.550
- x8    1    41.491 234.71 83.763

Step:  AIC=77.66
y ~ x1 + x2 + x3 + x5 + x6 + x7 + x8
 + x9 + x10

       Df Sum of Sq    RSS    AIC
- x6    1     1.893 195.87 75.975
- x7    1     8.718 202.69 77.071
- x2    1    11.348 205.32 77.483
<none>              193.98 77.664
- x9    1    14.140 208.12 77.916
- x1    1    20.455 214.43 78.872
- x3    1    21.445 215.42 79.020
- x10   1    25.847 219.82 79.667
- x5    1    32.817 226.79 80.666
- x8    1    40.847 234.82 81.779
```

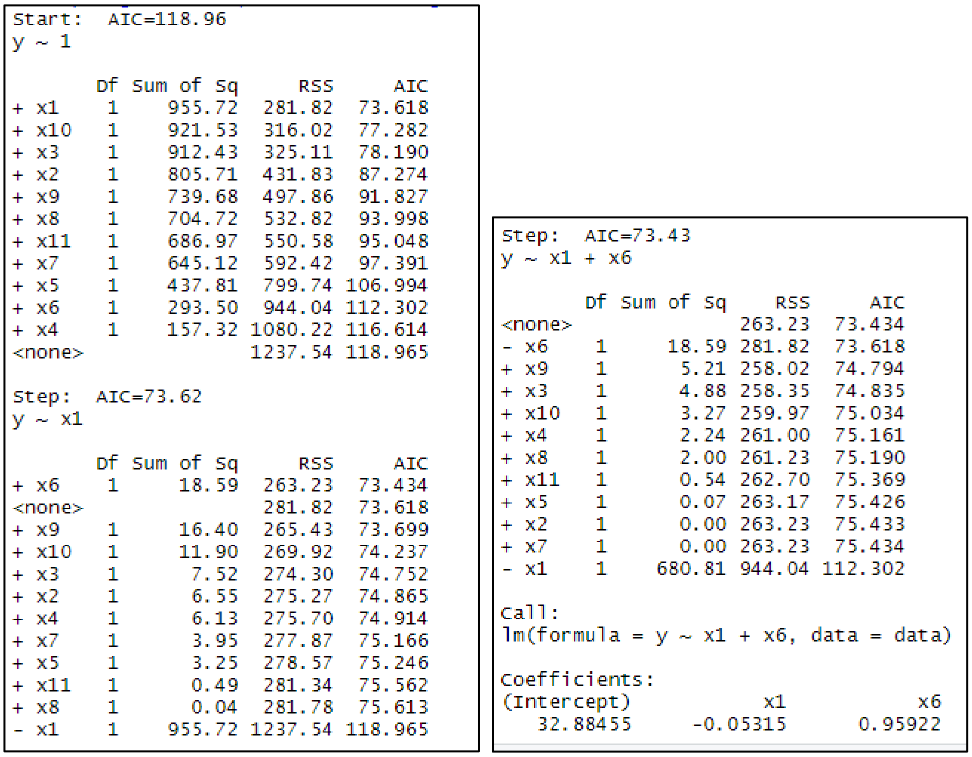
```
Step:  AIC=75.97
y ~ x1 + x2 + x3 + x5 + x7 + x8 + x9
 + x10

       Df Sum of Sq    RSS    AIC
- x7    1     8.085 203.96 75.269
- x2    1     9.594 205.47 75.505
<none>              195.87 75.975
- x9    1    16.654 212.53 76.586
- x1    1    18.804 214.68 76.908
- x3    1    19.588 215.46 77.025
- x10   1    32.842 228.71 78.935
- x5    1    36.561 232.43 79.452
- x8    1    44.565 240.44 80.535

Step:  AIC=75.27
y ~ x1 + x2 + x3 + x5 + x8 + x9 + x10

       Df Sum of Sq    RSS    AIC
- x2    1    10.038 213.99 74.807
- x9    1    10.225 214.18 74.835
- x1    1    12.439 216.40 75.164
<none>              203.96 75.269
- x3    1    15.379 219.34 75.596
- x10   1    32.585 236.54 78.012
- x5    1    32.963 236.92 78.063
- x8    1    38.283 242.24 78.774

Step:  AIC=74.81
y ~ x1 + x3 + x5 + x8 + x9 + x10

       Df Sum of Sq    RSS    AIC
- x3    1     5.419 219.41 73.607
- x9    1     6.417 220.41 73.752
- x1    1     9.007 223.00 74.126
<none>              213.99 74.807
- x5    1    23.006 237.00 76.074
- x10   1    27.086 241.08 76.621
- x8    1    33.768 247.76 77.495

Step:  AIC=73.61
y ~ x1 + x5 + x8 + x9 + x10

       Df Sum of Sq    RSS    AIC
- x1    1     3.823 223.24 72.160
- x9    1     7.762 227.17 72.719
<none>              219.41 73.607
- x5    1    21.805 241.22 74.639
- x10   1    27.226 246.64 75.350
- x8    1    38.808 258.22 76.818
```

```
Step:  AIC=72.16
y ~ x5 + x8 + x9 + x10

       Df Sum of Sq    RSS    AIC
- x9    1     5.613 228.85 70.955
<none>              223.24 72.160
- x5    1    43.164 266.40 75.816
- x8    1    64.460 287.70 78.277
- x10   1   170.100 393.34 88.286

Step:  AIC=70.95
y ~ x5 + x8 + x10

       Df Sum of Sq    RSS    AIC
<none>              228.85 70.955
- x5    1    38.934 267.78 73.982
- x8    1    59.013 287.86 76.296
- x10   1   245.912 474.76 92.306

Call:
lm(formula = y ~ x5 + x8 + x10, data
 = data)

Coefficients:
(Intercept)           x5            x8
          x10
   5.010946     2.625031      0.211874
    -0.009334
```

Figure 4 – R output from backward elimination of the mileage data set.

Section 2.3 Stepwise Regression

Finally, stepwise regression was performed on an intercept only model against all the predictors from the mileage data set. Relevant R output is provided in Figure 5. The stepwise regression returned a “best fit” model of gas mileage = intercept + x1(displacement) + x6(carburetor).



```
Start:  AIC=118.96
y ~ 1

       Df Sum of Sq     RSS     AIC
+ x1    1    955.72  281.82  73.618
+ x10   1    921.53  316.02  77.282
+ x3    1    912.43  325.11  78.190
+ x2    1    805.71  431.83  87.274
+ x9    1    739.68  497.86  91.827
+ x8    1    704.72  532.82  93.998
+ x11   1    686.97  550.58  95.048
+ x7    1    645.12  592.42  97.391
+ x5    1    437.81  799.74 106.994
+ x6    1    293.50  944.04 112.302
+ x4    1    157.32 1080.22 116.614
<none>              1237.54 118.965

Step:  AIC=73.62
y ~ x1

       Df Sum of Sq     RSS     AIC
+ x6    1     18.59  263.23  73.434
<none>               281.82  73.618
+ x9    1     16.40  265.43  73.699
+ x10   1     11.90  269.92  74.237
+ x3    1      7.52  274.30  74.752
+ x2    1      6.55  275.27  74.865
+ x4    1      6.13  275.70  74.914
+ x7    1      3.95  277.87  75.166
+ x5    1      3.25  278.57  75.246
+ x11   1      0.49  281.34  75.562
+ x8    1      0.04  281.78  75.613
- x1    1    955.72 1237.54 118.965
```

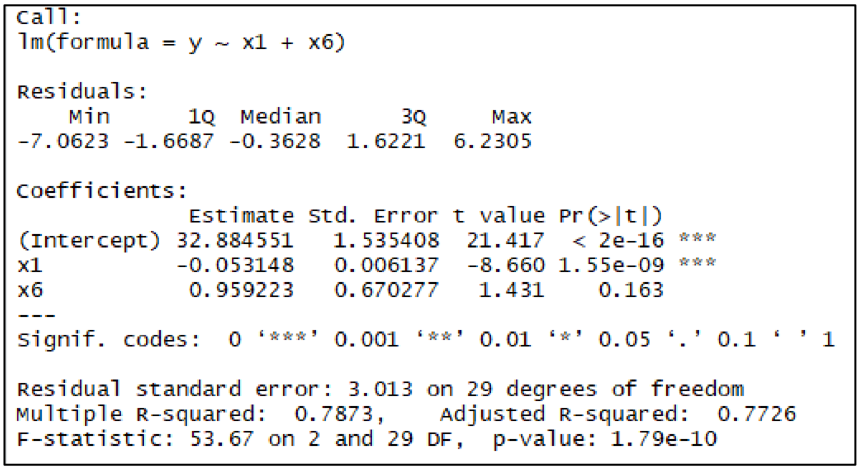
```
Step:  AIC=73.43
y ~ x1 + x6

       Df Sum of Sq     RSS     AIC
<none>               263.23  73.434
- x6    1     18.59  281.82  73.618
+ x9    1      5.21  258.02  74.794
+ x3    1      4.88  258.35  74.835
+ x10   1      3.27  259.97  75.034
+ x4    1      2.24  261.00  75.161
+ x8    1      2.00  261.23  75.190
+ x11   1      0.54  262.70  75.369
+ x5    1      0.07  263.17  75.426
+ x2    1      0.00  263.23  75.433
+ x7    1      0.00  263.23  75.434
- x1    1    680.81  944.04 112.302

Call:
lm(formula = y ~ x1 + x6, data = data)

Coefficients:
(Intercept)           x1           x6
   32.88455     -0.05315      0.95922
```

Figure 5 – R output from stepwise regression of the mileage data set.

Section 2.4 - Selection and Analysis of Model 1

Our choice to explore model selection satisfied the client’s goal of exploring the relationship between the predictors and response variable. Forward selection, backward elimination and stepwise regression all return models that exhibit the lowest AIC of all compared models. After reviewing the results of the three model selection functions, we decided to first consider the model suggested by the forward selection and stepwise regression, y (gas mileage) = x1 (displacement) + x6 (carburetor barrels). Henceforth referred to as model 1. Our decision was motivated by forward selection and stepwise regression both suggesting this model. Additionally, model 1 had fewer predictors than the model suggested by backward elimination, which follows the client's goal of producing a simple model.

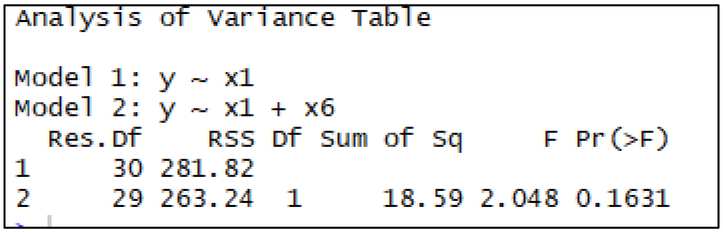
We performed a linear regression on model 1 and observed the carburetor predictor (x6) had a p-value of 0.163 (shown in Figure 6). As this is greater than 0.05, we considered predictor x6 to be a possible candidate for removal from the model.



```
Call:
lm(formula = y ~ x1 + x6)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.0623 -1.6687 -0.3628  1.6221  6.2305 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 32.884551   1.535408  21.417  < 2e-16 ***
x1          -0.053148   0.006137  -8.660 1.55e-09 ***
x6           0.959223   0.670277   1.431    0.163    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.013 on 29 degrees of freedom
Multiple R-squared:  0.7873,	Adjusted R-squared:  0.7726 
F-statistic: 53.67 on 2 and 29 DF,  p-value: 1.79e-10
```

Figure 6 - Regression results for model 1.

We conducted a hypothesis test to determine if we could remove predictor x6 (carburetor) from the model. To do this, we performed a partial F-test between the full model and the reduced model where x6 was removed. The null hypothesis was: H_0:β_6=0. The alternative hypothesis was H_a: β_6 is not equal to 0. The partial F test returned a p value of 0.1631, which is greater than 0.05 (shown in Figure 7). Therefore, we failed to reject the null hypothesis and were able to drop predictor x6 from the model.



```
Analysis of Variance Table

Model 1: y ~ x1
Model 2: y ~ x1 + x6
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     30 281.82                          
2     29 263.24  1     18.59 2.048 0.1631
```

Figure 7 - Results of the partial F test for model 1.

Model 1 was now a simple linear regression. Our next step was to assess whether the regression assumptions of the reduced model 1 were met. We produced a residual plot, ACF plot, QQ plot and a boxcox plot (shown in Figure 8). The residual plot exhibited possible curvature and evidence of non-constant variance which violates the homoscedasticity assumptions. Indicating a transformation was needed. The ACF plot showed no lag indicating the error terms are uncorrelated. The QQ Plot showed the errors follow a normal distribution meeting the regression assumption. Zero fell within the confidence interval of the BoxCox plot. Indicating a log transformation of the response variable was needed.

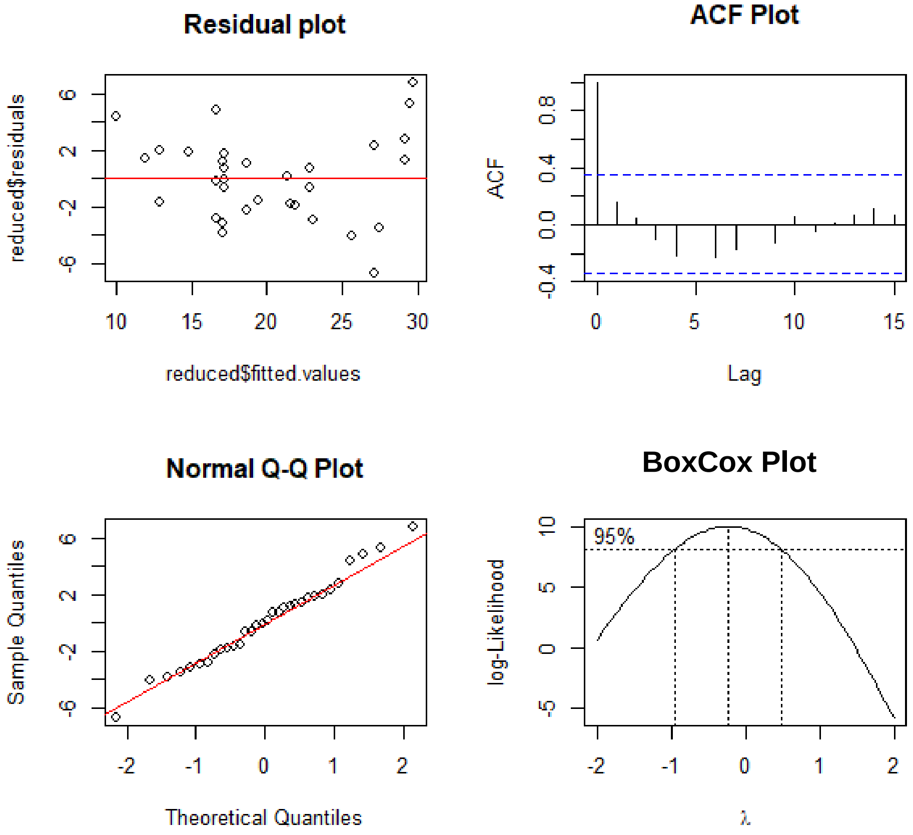


Figure 8 – Reduced model 1 regression assumption plots.

We performed the log transformation on the reduced model 1 response variable. The regression assumption plots were generated after the transformation (Figure 9). The residual plot exhibited constant variance and no curvature. The ACF and QQ plots indicated the errors were uncorrelated, and the normality assumption was reasonably satisfied. The BoxCox plot contained the value 1 in the confidence interval indicating no further transformations were necessary.

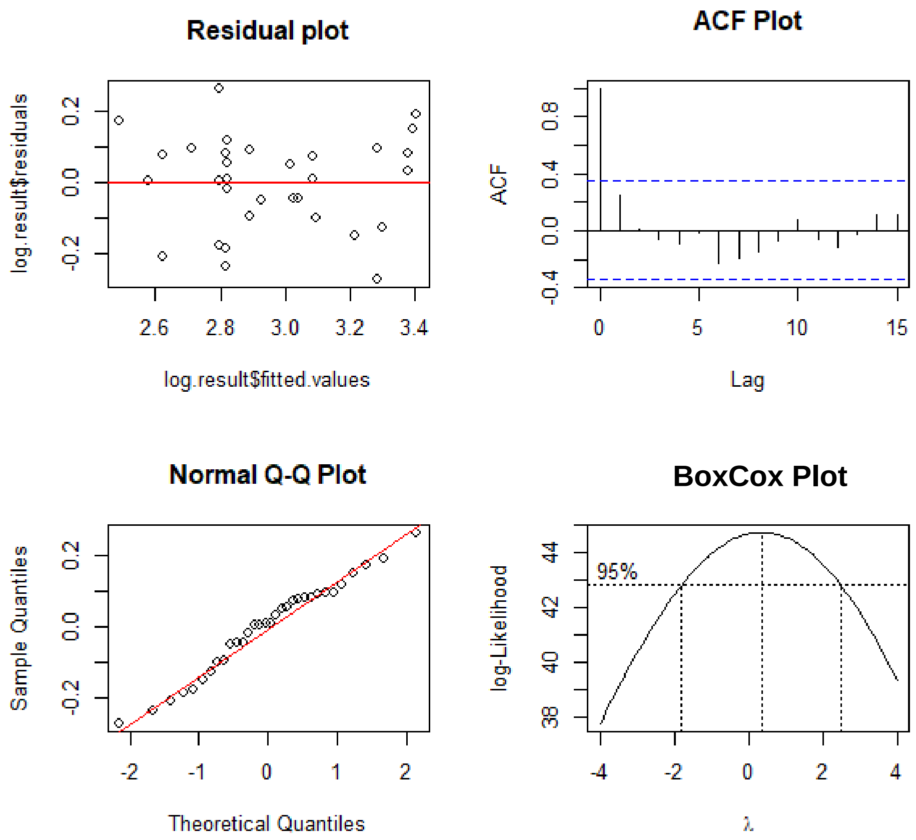
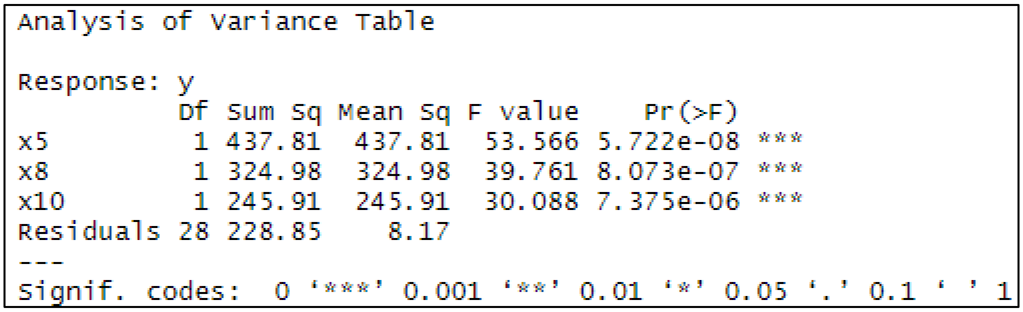


Figure 9 – Regression assumption plots of reduced model 1 after the log transformation.

Section 3 - Other Models Considered

The model suggested by backward elimination, gas mileage = x5(rear axle ratio) + x8(overall length) + x10(weight), was the second model considered. The results of the linear regression and ANOVA indicated all the predictors were significant (Figure 10).



```
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)    
x5         1 437.81  437.81  53.566 5.722e-08 ***
x8         1 324.98  324.98  39.761 8.073e-07 ***
x10        1 245.91  245.91  30.088 7.375e-06 ***
Residuals 28 228.85    8.17                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 10 – Model 2 ANOVA results.

We produced the same regression assumption plots for model 2 that we did for model 1 (Figure 11). The residual plot indicated non-constant variance and, therefore, the need for a transformation. The ACF plot showed no significant lags indicating the errors were uncorrelated. The QQ plot sufficiently satisfied the normality assumption. 0 fell within the confidence interval for the BoxCox plot indicating a log transformation of the response variable was needed.

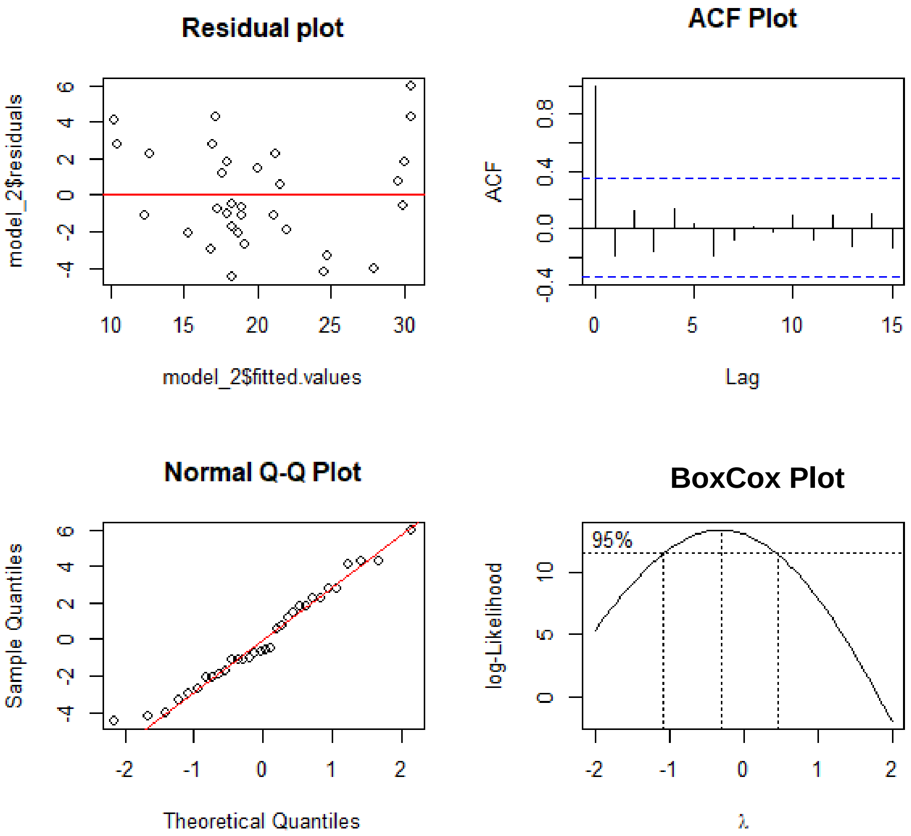


Figure 11 - Model 2 regression assumption plots.

After the transformation of the response variable the regression plots were generated again to check the assumptions (Figure 12). The residual plot exhibited constant variance and no curvature. The ACF and QQ plots indicated the errors were uncorrelated, and the normality assumption was reasonably satisfied. The BoxCox plot contained the value 1 in the confidence interval indicating no further transformations were necessary.

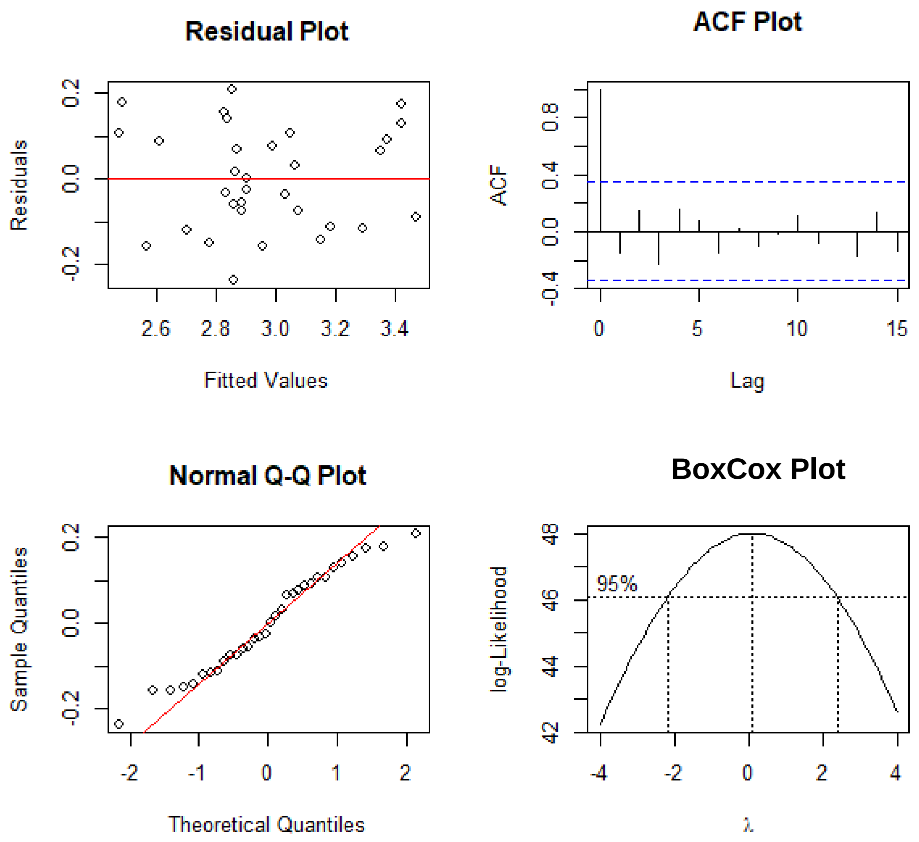


Figure 12 - Model 2 transformed regression assumption plots.

Section 4 - Summary of Findings

Two models were compared, model 1 with predictors x1+x6 and model 2 which had predictors x5+x8+x10. After performing a model summary for each model we found out that in model 1, predictor x6 was insignificant and thus removed it from the model. Predictors in model 2 were all significant. A log transformation was performed on each model.

The summary statistics, specifically the R^2_{adj}, of the finalized model 1 and model 2 were compared. Model 1 had a R^2_{adj} of 0.7873. Model 2 had an R^2_{adj} of 0.8129. Both models exhibited a good fit. Additionally, the PRESS statistic was calculated for each model. Model 1 had a PRESS statistic of 0.625. Model 2 had a PRESS statistic of 0.610.

Even though model 2 had the better R^2_{adj} and the better (lower) PRESS statistic, we recommend model 1 to the client. The difference between the R^2_{adj} values was small in magnitude, as was the difference between the PRESS statistics. Model 1 better satisfies the client's goals of creating a simple model as model one has only one predictor instead of three.

Additionally, we decided to calculate the R^2 prediction for both the models. This is calculated by dividing the PRESS statistic by the Total Sum of Squares (SST). The corresponding R^2 prediction for model 1 and 2 were 24% and 23% respectively. These values tell us how much variability in new observations a model might be able to explain. Both R^2 prediction values are low however the value for model 1 is slightly higher than that of model 2 indicating that model 1 with displacement as its only predictor has a higher predictive ability than model 2.

Based on our findings, we concluded that the simpler model one would best satisfy the clients goals of creating a model that fit well and was simple.

Section 5 - Summary of Findings for the Client

After examining the variables of the full model, we determined that a gas mileage is best related though the variable the displacement (cubic in.) After running several statistical tests to compare different models, we selected a simple model that describes the interaction between gas mileage (miles/gallon) and displacement (cubic in.). Our tests determined that our simple model containing the variable displacement fit the data well. A more complex model with additional predictors such as rear axle ratio, overall length, and weight did not provide a significantly better fit to justify the additional complexity of the model. Therefore, we recommend model 1 gas mileage = intercept + displacement as the simplest, best fit, first-order model.

References:

Editor, M. B. (2013, June 13). Multiple Regression Analysis: Use Adjusted R-Squared and Predicted R-Squared to Include the Correct Number of Variables. Retrieved July 11, 2020, from https://blog.minitab.com/blog/adventures-in-statistics-2/multiple-regession-analysis-use-adjusted-r-squared-and-predicted-r-squared-to-include-the-correct-number-of-variables#:~:text=The adjusted R-squared is,less than expected by chance.PRESS statistic. (2018, February 27). Retrieved July 11, 2020, from https://en.wikipedia.org/wiki/PRESS_statistic