**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.

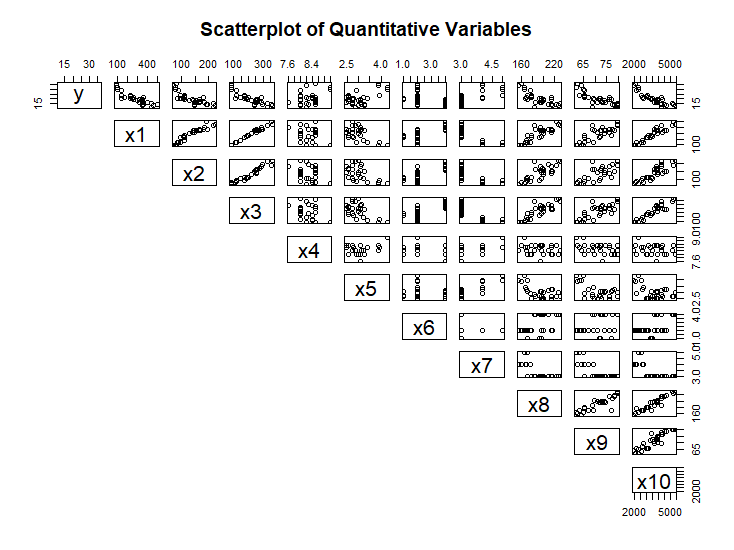


Figure 1 - A scatterplot matrix of the response variable, y, and possible regressor variables, x1-x10.

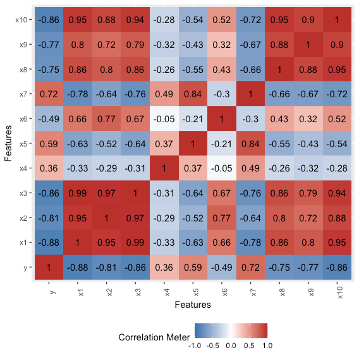


Figure 2 – Correlation chart that corresponds with Figure 1. Note, this matrix is rotated 90˚ counterclockwise compared to the scatterplot matrix.

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.

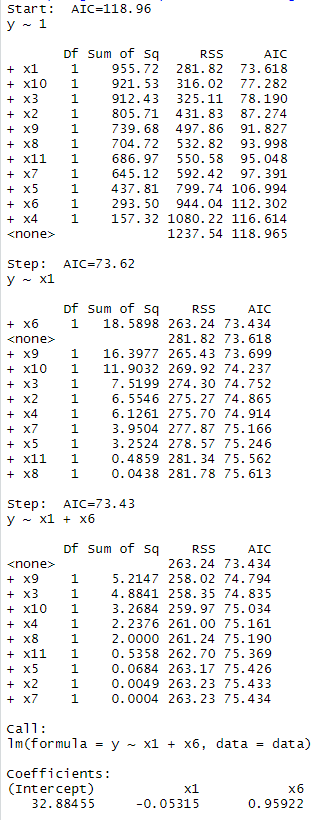
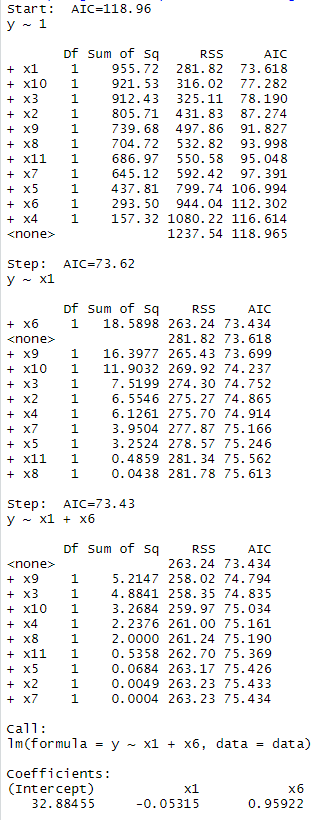
 

Figure 3 – Forward selection R output for mileage data set.

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).

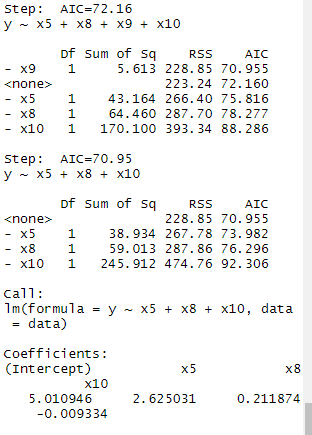
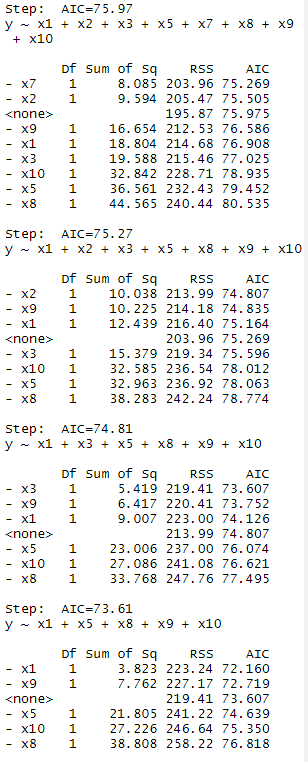
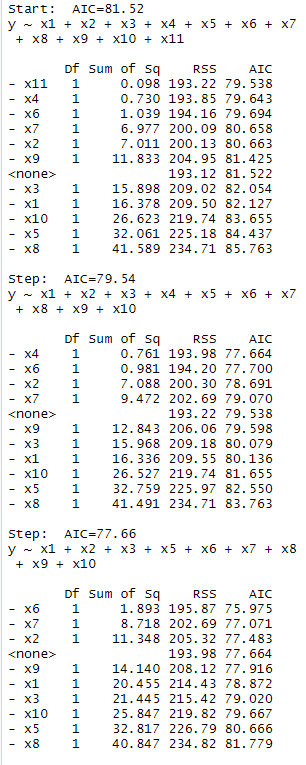


Figure 4 – R output from backward elimination of the mileage data set.

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).

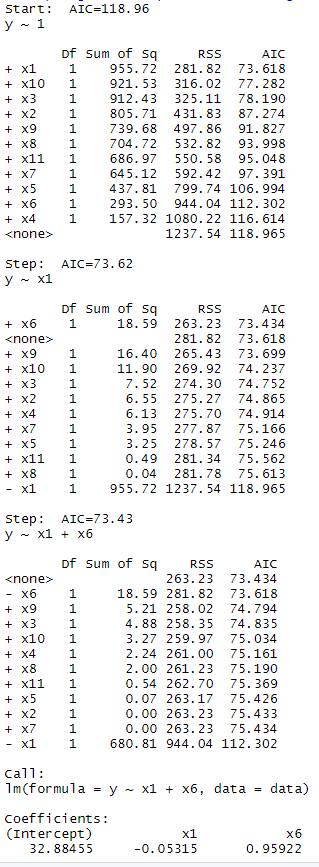
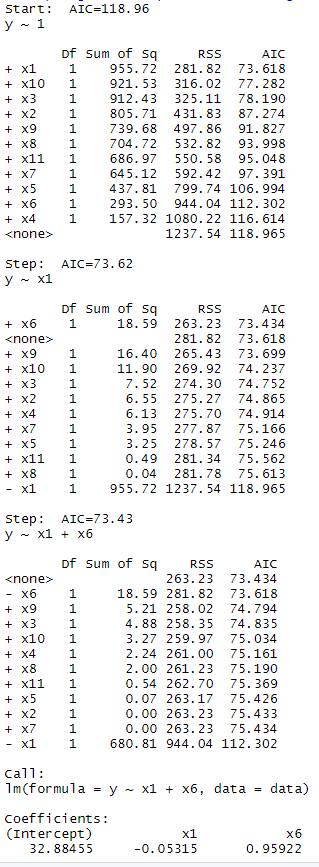
 

Figure 5 – R output from stepwise regression of the mileage data set.

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. 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 with variables x1 and x6 which produced the following function: y = 32.884551 - 0.053148(x1) + 0.959223(x6). This model is, henceforth, referred to as model\_1. The regression results indicated 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.

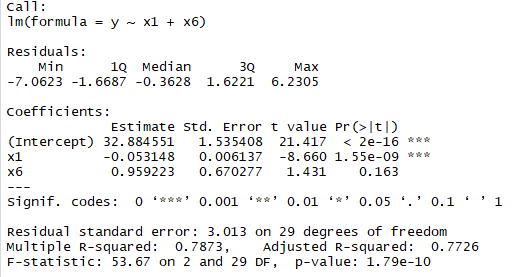


Figure 6 - Regression results for model 1.

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: H0:β6=0. The alternative hypothesis was Ha: β6 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.

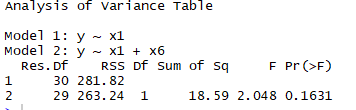
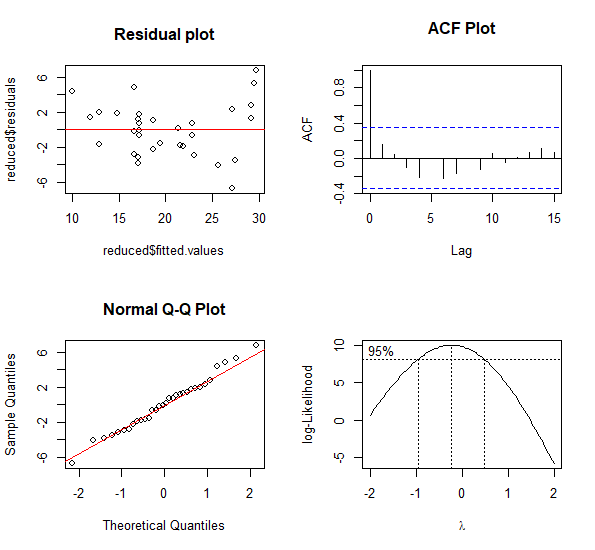


Figure 7 - Results of the partial F test for model 1.

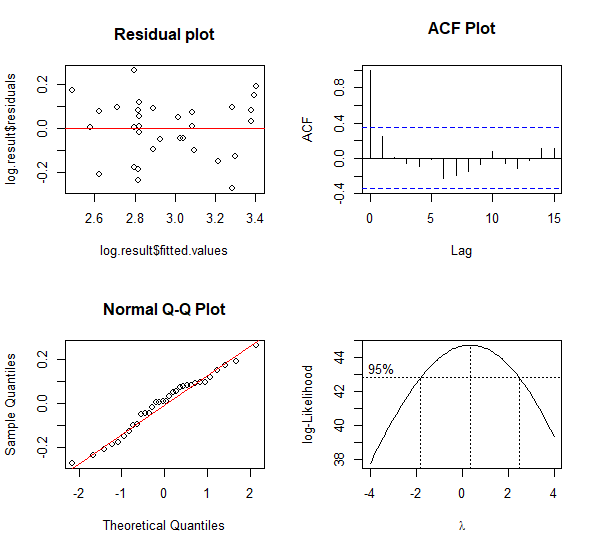
A new reduced model\_1 was produced, y = 33.722677 – 0.047360x1, and is referred to as model\_1r. Our next step was to assess whether the regression assumptions of the reduced model 1 were met. We produced a residual plot, ACF plot, QQ plot and a boxcox plot (shown in Figure 8). The residual plot exhibited 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.



**BoxCox Plot**

Figure 8 – Model\_1r regression assumption plots.

We performed the log transformation on model\_1r response variable. The new regression equation produced after this transformation was model\_1rt: y = 3.5933710 – 0.0022069(x1). 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.



**BoxCox Plot**

Figure 9 – Regression assumption plots of model\_1rt 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. We performed a linear regression with variables x5, x8 and x10 which produced the following function: y = 5.010946 + 2.625031(x5) + 0.211874(x8) – 0.009334(x10). This model is referred to as model\_2. The results of the linear regression and ANOVA indicated all the predictors were significant (Figure 10).

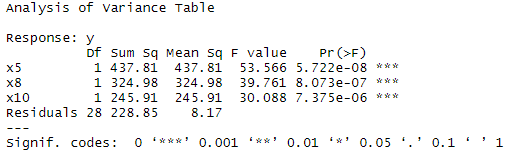
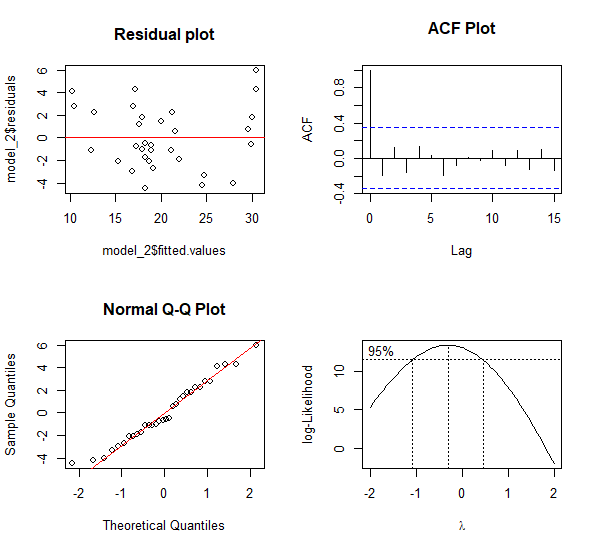


Figure 10 – Model\_2 ANOVA results.

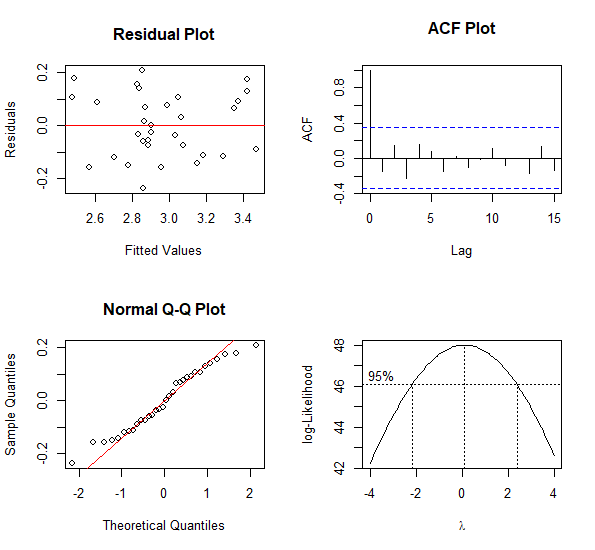
We produced the same regression assumption plots for model\_2 that we did for previous models (Figure 11). The residual plot indicated non-constant variance and, therefore, the need for a transformation. The ACF plot showed no significant lags indicating the errors were uncorrelated. The QQ plot sufficiently satisfied the normality assumption. 0 fell within the confidence interval for the BoxCox plot indicating a log transformation of the response variable was needed.



**BoxCox Plot**

Figure 11 - Model\_2 regression assumption plots.

The new regression equation produced after this transformation was model\_2t: y = 2.374 + 6.052E-02(x5) + 1.102E-02(x8) – 4.767E-04(x10). 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.



**BoxCox Plot**

Figure 12 – Model\_2t regression assumption plots.

#### **Section 4 - Summary of Findings**

Forward selection and stepwise regression suggested a model containing predictors x1 (displacement) and x6(carburator barrels) would best model gas mileage. Backward elimination suggested a model containing predictors x5(rear axle raito), x8(overall length), and x10(weight) would best model gas mileage. The two initial models were created: module\_1 y = 32.884551 - 0.053148(x1) + 0.959223(x6) and model\_2 y = 5.010946 + 2.625031(x5) + 0.211874(x8) – 0.009334(x10).

Hypothesis testing supported dropping x6 as a predictor from model\_1. Testing of the regression assumptions determined both model\_1r and model\_2 did not meet the constant variance assumption and, therefore, needed transformations.

The two final models chosen for comparison. The models were model\_1rt y = 3.5933710 – 0.0022069(x1) and model\_2t: y = 2.374 + 6.052E-02(x5) + 1.102E-02(x8) – 4.767E-04(x10).

The summary statistics, specifically the R2­adj, of the finalized model\_1rt and model\_2t were compared. Model\_1rt had a R2­adj of 0.7873. Model\_2t had an R2­adj of 0.8129. Both models exhibited a good fit. Additionally, the PRESS statistic was calculated for each model. Model\_1rt had a PRESS statistic of 0.625. Model\_2t had a PRESS statistic of 0.610.

Even though model\_2t had the better R2adjand the better (lower) PRESS statistic, we recommend model\_1rt to the client. The difference between the R2adjvalues was small in magnitude, as was the difference between the PRESS statistics. Model\_1rt 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 R2 prediction for both the models. This is calculated by dividing the PRESS statistic by the Total Sum of Squares (SST). The corresponding R2 prediction for model\_1rt and model\_2t were 24% and 23% respectively. These values tell us how much variability in new observations a model might be able to explain. Both R2 prediction values are low however the value for model\_1rt is slightly higher than that of model\_2t indicating that model\_1rt with displacement as its only predictor has a higher predictive ability than model\_2t.

Based on our findings, we concluded that model\_1rt, y = 3.5933710 – 0.0022069(x1), 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 the model y (gas mileage) = 3.5933710 – 0.0022069\*x1(displacement), referred to as model\_1rt in this report 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](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" \l ":~:text=The" \t "_blank) 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](https://en.wikipedia.org/wiki/PRESS_statistic" \t "_blank)