**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.**

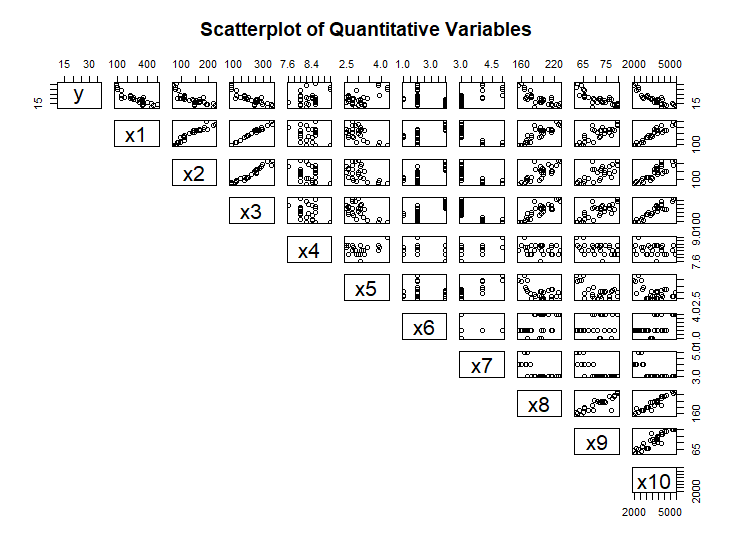
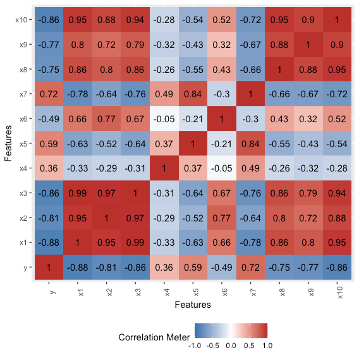
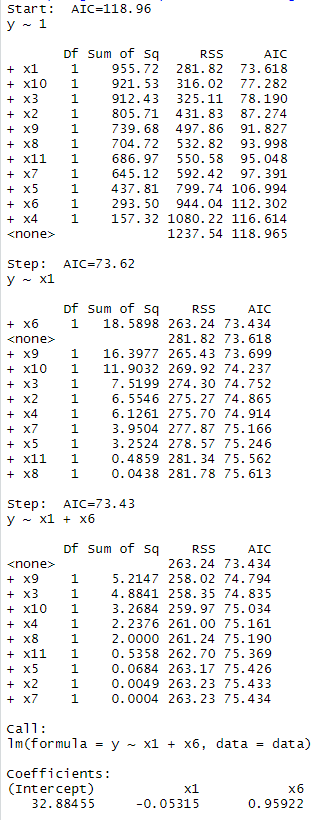


Figure 1 - A scatterplot matrix of the response variable, y, and possible regressor variables, x1-x10.



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

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.

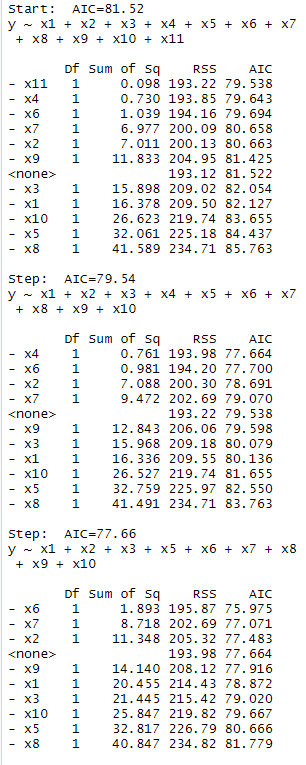
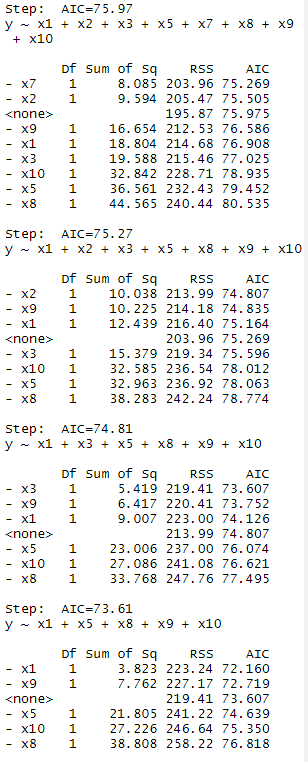
 

Figure # - Output from backward elimination.

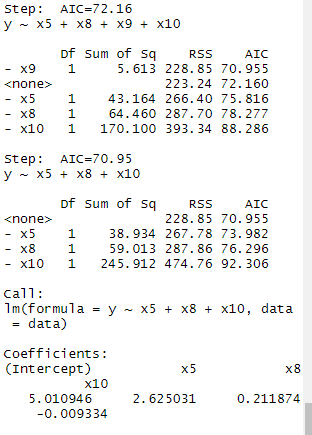
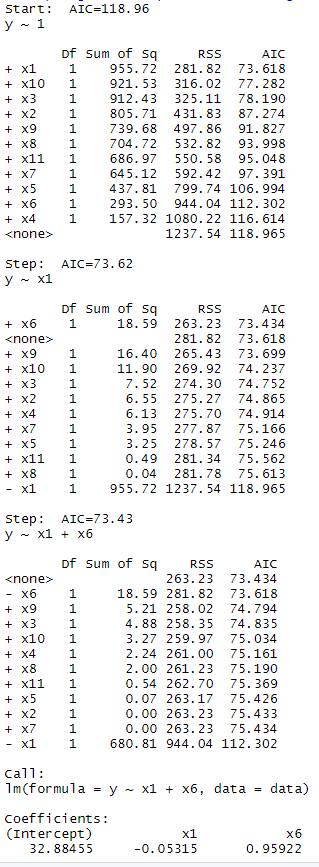
The backward elimination suggested a “best fit” model of: gas mileage = intercept +x5-rear axle ratio + x8-overall length + x10-weight.

Figure # - Output from backward elimination contd.

Section 2.3 Stepwise Regression

Finally, stepwise regression was performed on an intercept only model against all the predictors from the mileage data set. 

The stepwise regression returned a “best fit” model of: gas mileage = intercept + x1-displacement (in^3cubic inches) +and x6-carburetor.

Section 2.4 - Selection and Analysis of Model 1

Each of these models was favored as they yielded the lowest AIC value.

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 two of the methods suggesting this model and because it had fewer predictors than the model suggested by backward elimination.

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

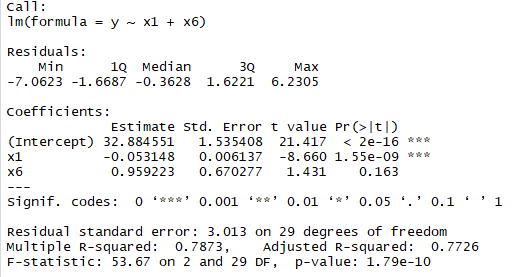


Figure # - Regression results for model 1.

In order to determine whether we could drop x6 as a predictor from the model we conducted a partial F test of the full model against the reduced model. Our hypotheses were: H0: Beta6 = 0 and Ha: Beta6 does not = 0. The partial F test failed to reject the null hypothesis and, therefore, supported dropping x6 from the model.

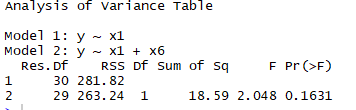
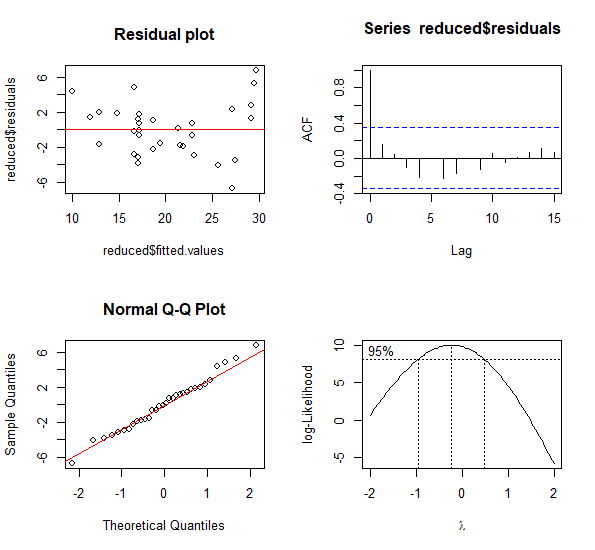
We conducted a hypothesis test to determine if we could remove predictor x6 (carburetor) from the model. To do this, we performed a partial F-test between the full model and the reduced model where x6 was removed. The null hypothesis was: H0:β6=0. The alternative hypothesis was Ha=β6 is not equal to 0. The partial F test returned a p value of 0.1631 which is greater than 0.05. Therefore,we failed to reject the null hypothesis and were able to drop predictor x6 from the model. 

Figure X - 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 X). The residual plot below shows a non-constant variance of residuals violating the homoscedasticity assumptions. Indicating a transformation is needed.

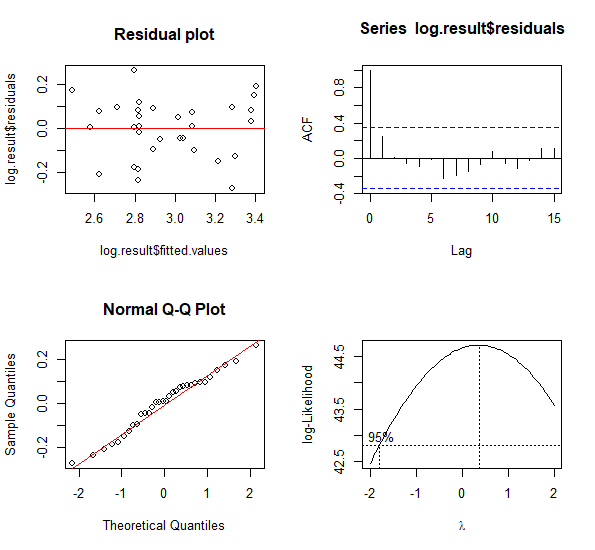


**BoxCox Plot**

Figure X – Reduced model 1 regression assumption plots.

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.

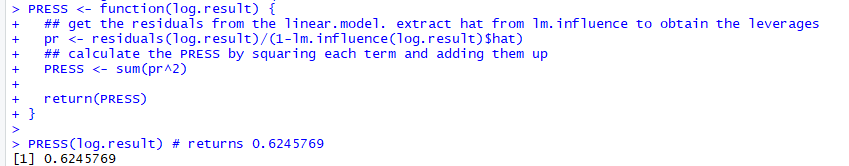
A boxcox plot was created to determine what transformation was needed. Since 1 was not in the reduced model 1 boxcox confidence interval (CI) but 0 was we determined that a log transformation of the response variable needed to be performed. We performed the log transformation resulting in boxcox plot that contained 1 in the CI . We checked the regression assumptions again after the transformation. Based on the plots, we determined no further transformations were necessary.



**BoxCox Plot**

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

Lastly, we calculated the press statistic for reduced model 1. We obtained a value of approximately 0.625. See the output below. This statistic can be thought of as cross-validation performed in regression analysis to provide a summary measure of the fit of a model to a sample of observations that were not themselves used to estimate the model. Ideally, the lower the PRESS statistic the better the model.

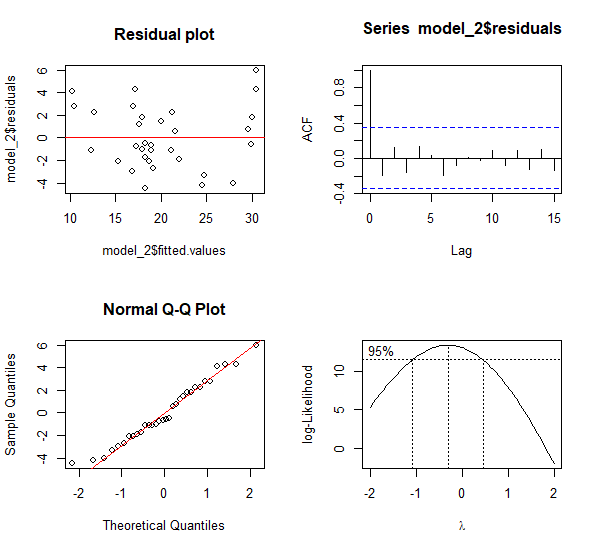


This model shows the client that predictor displacement would be sufficient for them to predict the gas mileage of a given automobile.

#### **Section 3 - Other Models Considered**

The second model considered was, y~ x5+x8+x10, which was the model suggested by backward elimination. The predictors were rear axle ratio (x5), overall length(x8), weight (lb) (x10). The results of the linear regression and ANOVA indicated all the predictors were significant. We set up a hypothesis test to see if at least one of the predictors’ coefficients was not zero in the model, H0:β5=β8=β10=0, Ha=β5=β8=β10 is not equal to zero. After exploring the results and conducting an anova test of the model, the output illustrated that the p value was less than 0.05 for all predictors, so we rejected the null hypothesis.

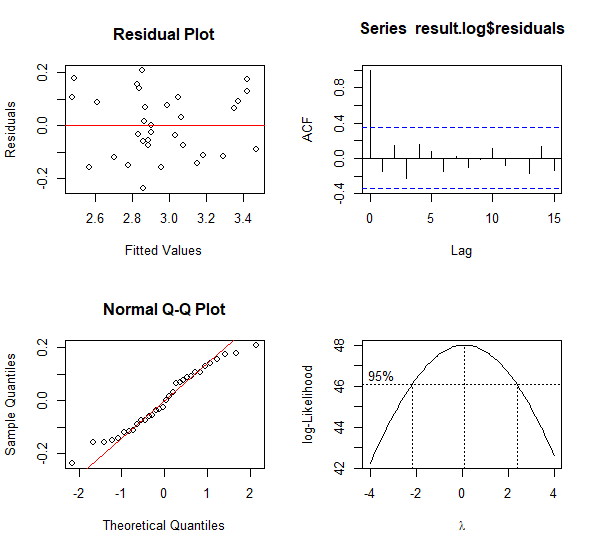
We produced the same regression assumption plots that we did for model 1 (shown in Figure X). The residual plot indicated non-constant variance and, therefore, the need for a transformation. The ACF plot showed no significant lags indicating the errors were uncorrelated. The QQ plot sufficiently satisfied 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 X - Model 2 regression assumption plots.

After the transformation of the response variable the regression assumptions were checked again. Based on the plots, we determined no further transformations were necessary.



**BoxCox Plot**

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

From looking at the final transformed model 2, we came to the conclusion that this was a well fitted model. In fact, this model had a better adj. R2 and the better PRESS statistic than model one. However, though this model had the better adj. R2 and the better PRESS statistic (shown in the outputs), we chose to go with the first model because it was a simpler model and therefore easier to fit and test. The difference between the two adj. R2 values were small in magnitude, as was the difference between the PRESS statistics, and we felt, due to that, x1 was the most important predictor, and that model 1 should be selected. For future experimentation, it would be interesting to see how the models would change if x1 was kept as a required predictor.

#### 

#### **Section 5 - Summary of Findings for the Client**