STAT 6021: Project 1 Report

Claire Setser (cas3hp)
Natalie Zimmer (naz6hd)
Mike Wetklow (mw8hu)

March 23, 2020

1. Initial Model Considered

First, we decided that it was important to inform ourselves of the meanings of all the predictor variables being considered, so we did some research (see Exhibit 1 below) to familiarize ourselves with the terminology. We took note of which variables we thought were most likely to be significant, and which ones we thought might cause multicollinearity problems if they were all included in the model.

Exhibit 1: Predictor Variables

● **X1 Displacement (cubic in.):** Engine size. ● **X2 Horsepower (ft-lb):** Engine power. ● **X3 Torque (ft-lb):** Engine strength. ● **X4 Compression Ratio:** Engine combustion. ● **X5 Rear Axle Ratio:** Rotation rate of axle. ● **X6 Carburetor (barrels):** Device that regulates air flow.	● **X7 No. of Transmission Speeds:** Number of car gears. ● **X8 Overall Length (in.)** ● **X9 Width (in.)** ● **X10 Weight (lb.)** ● **X11 Transmission (Automatic/Manual)**

While we thought that it was important to be informed on a high level of what would potentially be going into the model, we decided to begin the model building process using various statistical methods in R.

After loading and attaching the data, we decided to first use forward addition, backward elimination, and stepwise regression to build our regression model and compare the outputs from the functions in R. Using the full model with all of the variables included as the upper bound and an intercept-only model as the lower bound, forward addition and stepwise regression both gave the result of $y = \beta_0 + \beta_1 x_1 + \beta_6 x_6$, while backward elimination yielded $y = \beta_0 + \beta_5 x_5 + \beta_8 x_8 + \beta_{10} x_{10}$. Because stepwise regression takes more into consideration by evaluating the partial F-statistic of the model both before and after adding each variable, we decided that $y = \beta_0 + \beta_1 x_1 + \beta_6 x_6$ would be the first model we would consider using. The fact that

this equation was also arrived at by building the model using forward addition also helped us to make this decision. Exhibit 2 below provides the relevant R output for each of the stepwise regression functions.

Exhibit 2: R Output Initial Model Considered

```
> summary(forward)

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

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
```

```
> summary(backward)

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

Residuals:
   Min     1Q Median     3Q    Max
-4.512 -1.945 -0.631  1.931  6.003

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.010946  11.275042   0.444   0.6601
x5           2.625031   1.202720   2.183   0.0376 *
x8           0.211874   0.078850   2.687   0.0120 *
x10         -0.009334   0.001702  -5.485 7.37e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.859 on 28 degrees of freedom
Multiple R-squared:  0.8151,    Adjusted R-squared:  0.7953
F-statistic: 41.14 on 3 and 28 DF,  p-value: 2.156e-10
```

```
> summary(stepwise)

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

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
```

2. Other Models Considered

In order to begin evaluating the adequacy of the model, we first simply regressed y against x_1 and x_6 and looked at the summary from the R output (see Exhibit 3 below). We immediately noticed that while the t-statistic for x_1 was highly significant, the t-statistic for x_6 was not. Based off of this, we decided to revise the model that we were considering to simply $y = \beta_0 + \beta_1 x_1$.

Exhibit 3: R Output Other Models Considered

```
> result.stepwise<-lm(y~x1+x6)
> summary(result.stepwise)

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
```

At this point, we decided to also consider the model $y = \beta_0 + \beta_5 x_5 + \beta_8 x_8 + \beta_{10} x_{10}$ that backward elimination provided us with. The R^2_{Adj} value for this model was 0.7953, while the R^2 value for the model $y = \beta_0 + \beta_1 x_1$ was 0.7723, so choosing between these models meant either choosing a model that was more complicated with a slightly better fit, or a model that was extremely simple with a slightly worse fit, respectively.

To see if there were any models we were overlooking by using only the stepwise methods, we decided to evaluate all possible first order models using the *regsubsets* function from the *leaps* library. We inserted the output into a data frame along with the statistical quantities R^2, R^2_{Adj}, *MSE,* C_P, and *BIC*. R^2 and R^2_{Adj} values are used to compare different models, whereas *MSE,* C_P, and *BIC* were leveraged to assess overall model adequacy (see Exhibit 4 below). The model that provided both the best *MSE* and C_P was $y = \beta_0 + \beta_5 x_5 + \beta_8 x_8 + \beta_{10} x_{10}$, while the model that yielded the best *BIC* was $y = \beta_0 + \beta_1 x_1$. This confirmed to us that the two models we were considering were both probably good choices, and that we should do some further analysis to determine which would be the best for the client's needs.

Exhibit 4: R Regsubsets Outputs

```
> best[order(best$r2,decreasing=TRUE),]
          x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11manual p        r2     adjr2      mse         cp       bic
8  ( 1 )   *  *  *     *     *  *  *  *           9 0.8417263 0.7866746 8.516116  6.2850767 -27.79812
8  ( 2 )   *     *  *  *     *  *  *  *           9 0.8377880 0.7813664 8.728024  6.7898342 -27.01161
8  ( 3 )   *  *  *     *        *  *  *         * 9 0.8368785 0.7801406 8.776959  6.9063957 -26.83269
```

```
> best[order(best$adjr2,decreasing=TRUE),]
          x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11manual p        r2     adjr2      mse         cp       bic
3  ( 1 )              *        *     *           4 0.8150777 0.7952646 8.173197 -0.2995134 -40.14728
4  ( 1 )           *  *        *     *           5 0.8202994 0.7936771 8.236572  1.0312518 -37.59814
4  ( 2 )              *        *  *  *           5 0.8196137 0.7928898 8.268001  1.1191330 -37.47626
```

```
> best[order(best$mse,decreasing=FALSE),] #(want smallest mse)
          x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11manual p        r2     adjr2      mse         cp       bic
3  ( 1 )              *        *     *           4 0.8150777 0.7952646 8.173197 -0.2995134 -40.14728
4  ( 1 )           *  *        *     *           5 0.8202994 0.7936771 8.236572  1.0312518 -37.59814
4  ( 2 )              *        *  *  *           5 0.8196137 0.7928898 8.268001  1.1191330 -37.47626
```

```
> best[order(best$cp,decreasing=FALSE),]  #(want smallest cp)
          x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11manual p        r2     adjr2      mse         cp       bic
3  ( 1 )              *        *     *           4 0.8150777 0.7952646  8.173197 -0.2995134 -40.14728
4  ( 1 )           *  *        *     *           5 0.8202994 0.7936771  8.236572  1.0312518 -37.59814
4  ( 2 )              *        *  *  *           5 0.8196137 0.7928898  8.268001  1.1191330 -37.47626
```

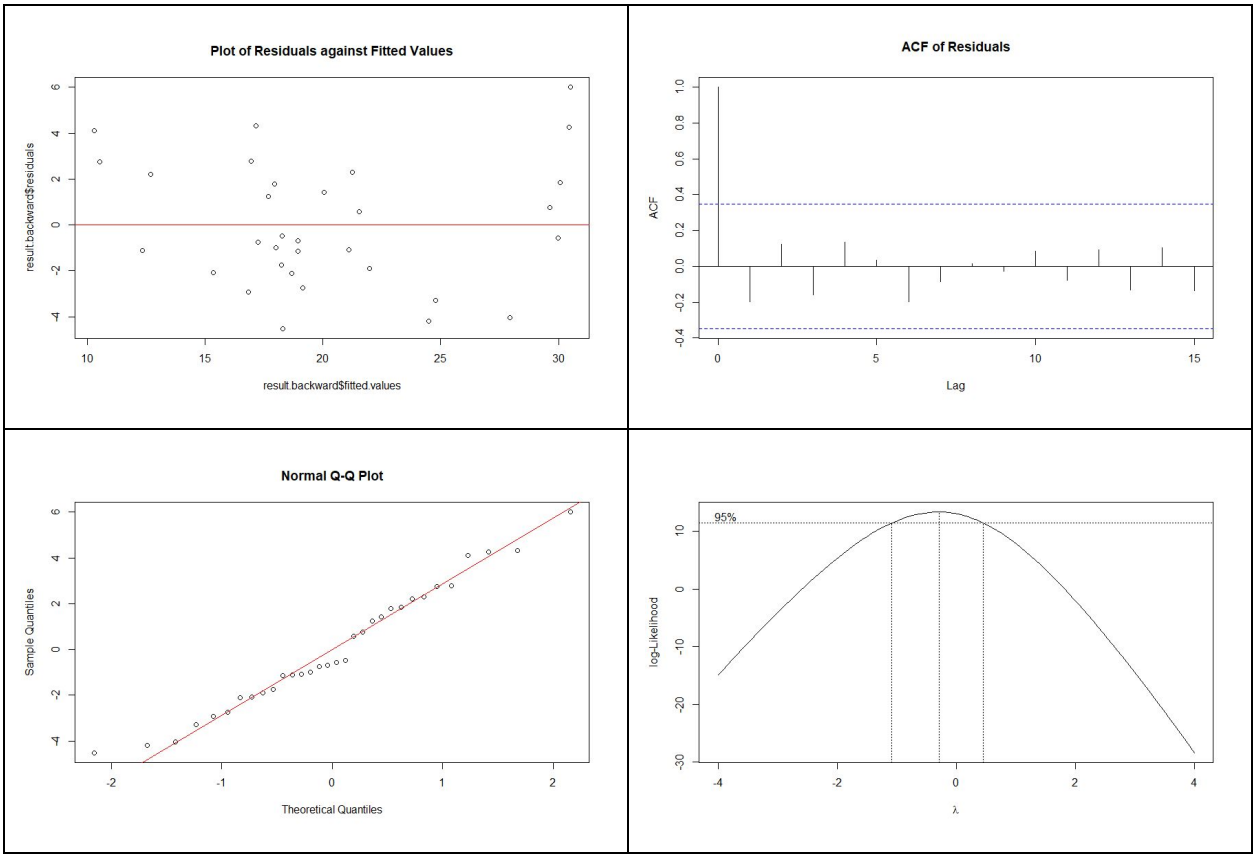
```
> best[order(best$bic,decreasing=FALSE),]
          x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11manual p        r2     adjr2      mse         cp       bic
1  ( 1 )   *                                      2 0.7722712 0.7646803  9.394146  1.1867551 -40.41573
3  ( 1 )              *        *     *           4 0.8150777 0.7952646  8.173197 -0.2995134 -40.14728
2  ( 1 )   *              *                      3 0.7872928 0.7726233  9.077053  1.2615248 -39.13363
```

3. Summary of Findings

We recommend $y = \beta_0 + \beta_5 x_5 + \beta_8 x_8 + \beta_{10} x_{10}$ as our final model, with the regression equation of $y = 5.011 + 2.625\, x_5 + 0.212 x_8 - 0.009 x_{10}$. To decide between the $y = \beta_0 + \beta_1 x_1$ and $y = \beta_0 + \beta_5 x_5 + \beta_8 x_8 + \beta_{10} x_{10}$ models we conducted partial F tests compared to the full model and concluded that $y = \beta_0 + \beta_5 x_5 + \beta_8 x_8 + \beta_{10} x_{10}$ is the preferred model based on the fact that it has the highest F-statistic, meaning that the predictor variables not included in the model are jointly insignificant and can be dropped from the model. While $y = \beta_0 + \beta_1 x_1$ also had a high F-statistic, it was lower than the one for $y = \beta_0 + \beta_5 x_5 + \beta_8 x_8 + \beta_{10} x_{10}$, meaning that the variables being dropped from that model were more significant, which is less desirable.

In addition, we compared PRESS Statistics for our final two models, and concluded $y = \beta_0 + \beta_5 x_5 + \beta_8 x_8 + \beta_{10} x_{10}$ had the lowest PRESS Statistic, meaning that it is the best at predicting new data. This model also has a reasonably strong R^2_{Adj} value of 0.7953, meaning that about 80% of the variation in the response variable can be explained by the model. This model also had the lowest *MSE* and C_p, and the second lowest *BIC*, which again are all statistical quantities that are useful in determining a model's adequacy. As a final step in concluding $y = \beta_0 + \beta_5 x_5 + \beta_8 x_8 + \beta_{10} x_{10}$ as our recommended model, we assessed whether the major regression assumptions are met. The results of our first assumptions check are displayed in Exhibit 5 below.

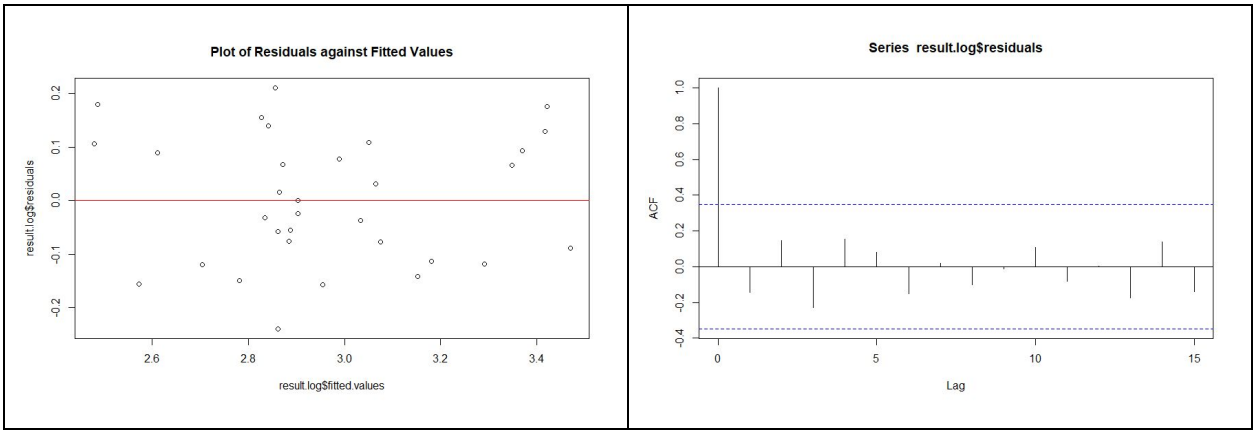
Exhibit 5: First Regression Assumptions Check

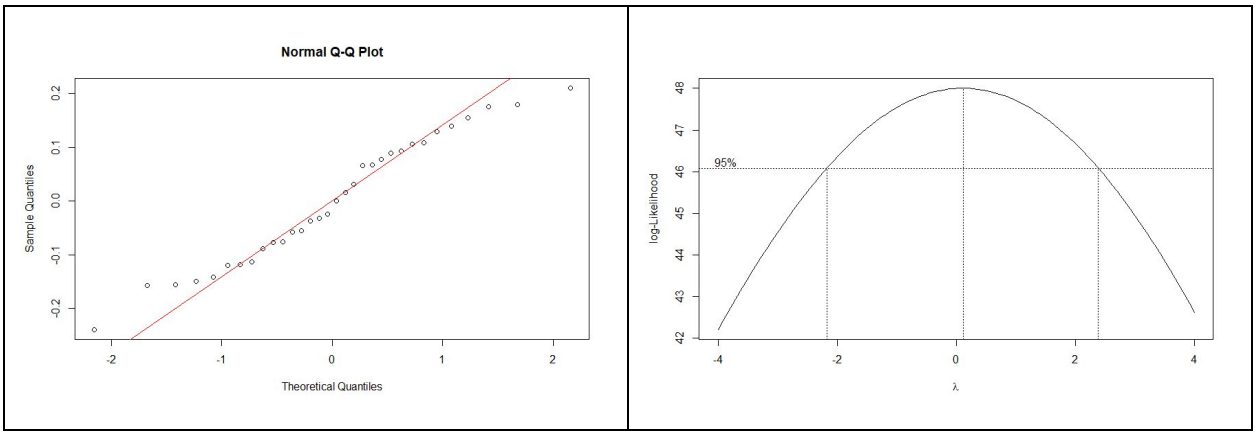


The Plot of Residuals against Fitted Values shows the residuals are somewhat evenly scattered on the horizontal line from the far left to the middle of the plot with some imbalance of negative values to positive towards the far right of the plot indicating a potential violation of the linearity assumption. The vertical spread of the residuals appears constant from the left to the middle of the plot with a large gap towards the upper right portion of the plot indicating a clear violation of the equal variance assumption.

The ACF Plot, from Exhibit 5 above, confirms the independence of error terms as all lags are insignificant. The Normal QQ-Plot confirms a normal distribution as the central values are approximately close to the straight line. Next, our initial check also included a Box-Cox method to confirm the need to transform the y variable to improve normality and correct the nonconstant variance. Given the non-constant error variance, we focused on stabilizing the variance by transforming the y response variable through a log transformation. The standard practice of a y transformation could also potentially help to improve the linearity. The results of our second assumptions check are displayed in Exhibit 6 below.

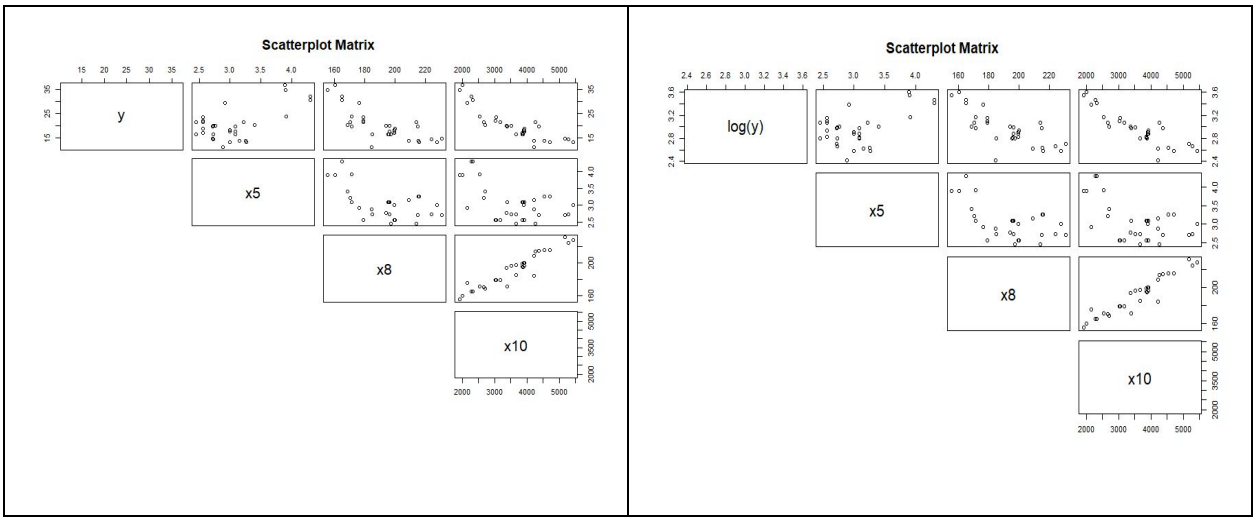
Exhibit 6: Second Regression Assumptions Check





The vertical spread of the residuals appears more constant and the gap from Exhibit 5 appears to have improved to meet the equal variance assumption. The ACF Plot re-confirms the independence of error terms and lags continue to be insignificant. The Normal QQ-Plot's central values appear to be tighter to the straight line. Our second check of the Box-Cox method shows significant improvement as the best lambda value 1 is within the upper bound of the 95 percent confidence interval, which indicates the transformation of the y variable improved normality and corrected the nonconstant variance. Finally, the transformation on y appears to have improved the linearity of the model as the Plot of Residuals against Fitted Values shows the residuals are more evenly scattered on the horizontal line and individual scatter plots were somewhat straightened out (See Exhibit 7 below).

Exhibit 7: Comparison of First and Second Regression Assumption Scatterplot Checks



4. Summary of Findings for the Client

In order to accurately relate the gas mileage of a vehicle, the model we are recommending includes rear axle ratio, overall length in inches, and weight in pounds. We considered three different models and arrived at this model by focusing on your two main goals: the model fits well and the simplicity of the model.

To begin the model selection process, we used automated search procedures which explore the possible subsets of predictors to help determine the most suitable model. Our first model contained only two variables, displacement in cubic inches and carburetor barrels, but further analysis indicated that carburetor barrels were not a significant component of this model. Carburetor barrels were then removed from the model, giving us our second model with only one variable, displacement in cubic inches. Displacement was a significant factor on gas mileage by itself, but we decided to consider another model from the automated search procedures, since mileage can be impacted by more than just one factor. The third model considered included rear axle ratio, overall length in inches, and weight in pounds. This model

proved to have the highest overall model adequacy compared to the others, and each predictor in the model was statistically significant. Statistical adjustments, more technically known as log transformations, were made to ensure we could interpret patterns in the data and meet statistical assumption validity checks. Although three variables are more complex than one, the third model proved to be the best and includes less than half of the predictors in the data. We believe that the models considered provide a reasonable basis for our findings and conclusions.