FUEL EFFICIENCY REGRESSION MODEL

Aditya Jain, Ahmed Omar, Yedukrishnan

Dec 8, 2023

Introduction

The daily commute for many individuals, whether it's to go to work, school, or any other location, requires the use of a motor vehicle. Although the world is transitioning towards greener sources of transportation, the most common method of transport for many people is a gas powered vehicle. On a monthly basis, cars tend to be somewhat expensive for many people due to the cost of insurance, maintenance, as well as gas. If an individual wishes to cut their spending, a possible goal to focus on would be to reduce their spending on gas by purchasing a more fuel-efficient vehicle. The objective of this report is to highlight what features of a car have the most significant effect on its fuel efficiency. To do so, we will be looking at the independent variable, miles per gallon (mpg), to highlight its relation to dependent variables such as the cylinder number of the vehicle, displacement, horsepower, weight, acceleration, and if the origin (1: American, 2: European, 3: Japanese) of the vehicle is also a significant predictor.

Dataset

We will focus on the use of one dataset in this report, a csv labelled as "Fuelcar." This file was obtained from the Carnegie Mellon University website through their StatLib data collection [1]. It is an open dataset and the university has provided free use of this file for educational purposes. The dataset contains 399 rows and 9 columns. About 7 null values exist and therefore, the data was cleaned to remove any missing values using built in excel functions.

Methodology

Again, the data file we have contains 399 rows and 9 columns, where the variables are:

mpg (Quantitative) = fuel efficiency measured in miles per gallon (mpg)

cylinders (Qualitative) = number of cylinders in the engine (4,5,6,8)

displacement (Quantitative) = engine displacement (in cubic inches)

horsepower (Quantitative) = unit of measurement

weight (Quantitative) = vehicle weight (in pounds)

acceleration (Quantitative) = time to accelerate from O to 60 mph (in seconds)

model.year (Qualitative) = model year

origin (Qualitative) = origin of car (1: American, 2: European, 3: Japanese)

car.name (Qualitative) = car name

The main dependent variable is quantitative, and measures miles per gallon (mpg) where the independent variables will include the rest of the listed columns. Please note that the “car.name” and “model.year” variables will not be included in this report only because of the amount of unique values available, and we did not feel that treating both as dummy variables would be ideal.

Below, we will be conducting a multiple linear regression analysis to propose a model to use for predicting our dependent variable. All steps below have been conducted in relative order to test and propose the best fit model, while testing our assumptions for linear regression analysis.

Techniques Used and Their Justification

Interactions:

Use: Introducing interaction terms in a regression model allows you to account for the combined effect of two or more variables, acknowledging that their joint impact may differ from the sum of their individual effects.

Justification: Interactions are useful when the effect of one variable on the dependent variable depends on the level of another variable. It helps capture non-additive relationships.

Individual t-tests:

Use: Conducting individual t-tests for coefficients tests the hypothesis that the coefficient for a specific variable is significantly different from zero.

Justification: Helps determine the significance of individual predictors and assess their contribution to the model.

Global and Partial F-Tests:

Use: Global F-tests assess the overall significance of the regression model, while partial F-tests assess the significance of a subset of variables.

Justification: Global tests ensure the model as a whole is significant, while partial tests help evaluate the importance of specific sets of variables.

Statistical Tests of Residuals:

Use: Diagnostic tests such as residuals analysis, normality tests, and heteroscedasticity tests assess the adequacy of the model assumptions.

Justification: Ensures that the residuals meet the assumptions of the regression model, validating the reliability of the estimated coefficients.

For all statistical tests below, an alpha value of 0.05 will be used. For the report, the work was split as such:

Aditya Jain – Work on producing the models

Ahmed Omar – Work on assumptions

Yedu Krishnan – Write up and data visualization

Overall, all three of us worked together on the project and assisted eachother with each each of our sections. Model creation, assumptions, and visualizations were all written together through a zoom call to make sure our data and assumptions are all clear. GitHub and teams were both used for the presentation and the report.

1
2
3
4
5
6

6 rows | 1-1 of 10 columns

Full Model

In the first step, our objective is to create a fullmodel to test the significance of the independent variables, or predictors, on our dependent variable (mpg). This means that we will test the following hypothesis:

$$H_0: \ \beta 1 = \beta 2 = ... = \beta p = 0$$
$$H_A: \ \text{at least one } \beta i \text{ is NOT EQUAL TO 0 } (i = 1,2,...,p)$$

```
##
## Call:
## lm(formula = mpg ~ factor(cylinders) + displacement + horsepower +
##     weight + acceleration + factor(origin), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0297  -2.1973  -0.5707   1.7730  16.0313
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        35.6607269  3.0730677  11.604  < 2e-16 ***
## factor(cylinders)4  8.5432570  2.0817350   4.104 4.97e-05 ***
## factor(cylinders)5 10.8584486  3.1601616   3.436 0.000655 ***
## factor(cylinders)6  4.5601901  2.3004164   1.982 0.048160 *
## factor(cylinders)8  6.5969826  2.6585973   2.481 0.013518 *
## displacement        0.0059735  0.0090521   0.660 0.509715
## horsepower         -0.0774551  0.0163065  -4.750 2.89e-06 ***
## weight             -0.0038575  0.0007805  -4.942 1.16e-06 ***
## acceleration       -0.0943069  0.1169949  -0.806 0.420701
## factor(origin)2     0.0868317  0.6814804   0.127 0.898678
## factor(origin)3     2.6453931  0.6654358   3.975 8.41e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.91 on 381 degrees of freedom
## Multiple R-squared:  0.7555, Adjusted R-squared:  0.7491
## F-statistic: 117.7 on 10 and 381 DF,  p-value: < 2.2e-16
```

The output result is a p-value of 2.2e-16 for the full model, which is less than an alpha value of 0.05 enabling us to reject the null hypothesis. We can infer the alternative hypothesis that states at least one βi is not zero, and that atleast one of the predictor variables is statistically significant.

From the fullmodel summary, “displacement” and “acceleration” have a p-value of 0.509715 and 0.420701 respectively, which means we FAIL to reject the null hypothesis that states βi = 0. This means that these variables may not have an effect on the dependent variable “mpg” or that they are not statistically significant to the model. We can remove “displacement” and “acceleration” from our model in a reducedmodel.

It is important to note that “cylinders” and “origin” are both qualitative variables, and therefore the factor function has been utilized to account for the differences.

Regression Procedure to Verify fullmodel Results

Stepwise Regression Procedure

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4002  -2.2047  -0.5689   1.7741  16.3854
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        34.0149296  2.5525144  13.326  < 2e-16 ***
## weight             -0.0039465  0.0006271  -6.293 8.52e-10 ***
## factor(cylinders)4  8.6440072  2.0182505   4.283 2.34e-05 ***
## factor(cylinders)5 11.0624033  3.0904967   3.579 0.000389 ***
## factor(cylinders)6  5.0487106  2.1033193   2.400 0.016856 *
## factor(cylinders)8  7.5237315  2.2425536   3.355 0.000873 ***
## horsepower         -0.0652994  0.0117926  -5.537 5.71e-08 ***
## factor(origin)2    -0.0871501  0.6432645  -0.135 0.892303
## factor(origin)3     2.4816343  0.6334064   3.918 0.000106 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.906 on 383 degrees of freedom
## Multiple R-squared:  0.7547, Adjusted R-squared:  0.7496
## F-statistic: 147.3 on 8 and 383 DF,  p-value: < 2.2e-16
```

Backward Elimination Procedure

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4002  -2.2047  -0.5689   1.7741  16.3854
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        34.0149296  2.5525144  13.326  < 2e-16 ***
## factor(cylinders)4  8.6440072  2.0182505   4.283 2.34e-05 ***
## factor(cylinders)5 11.0624033  3.0904967   3.579 0.000389 ***
## factor(cylinders)6  5.0487106  2.1033193   2.400 0.016856 *
## factor(cylinders)8  7.5237315  2.2425536   3.355 0.000873 ***
## horsepower         -0.0652994  0.0117926  -5.537 5.71e-08 ***
## weight             -0.0039465  0.0006271  -6.293 8.52e-10 ***
## factor(origin)2    -0.0871501  0.6432645  -0.135 0.892303
## factor(origin)3     2.4816343  0.6334064   3.918 0.000106 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.906 on 383 degrees of freedom
## Multiple R-squared:  0.7547, Adjusted R-squared:  0.7496
## F-statistic: 147.3 on 8 and 383 DF,  p-value: < 2.2e-16
```

Forward Selection Procedure

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4002  -2.2047  -0.5689   1.7741  16.3854
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        34.0149296  2.5525144  13.326  < 2e-16 ***
## weight             -0.0039465  0.0006271  -6.293 8.52e-10 ***
## factor(cylinders)4  8.6440072  2.0182505   4.283 2.34e-05 ***
## factor(cylinders)5 11.0624033  3.0904967   3.579 0.000389 ***
## factor(cylinders)6  5.0487106  2.1033193   2.400 0.016856 *
## factor(cylinders)8  7.5237315  2.2425536   3.355 0.000873 ***
## horsepower         -0.0652994  0.0117926  -5.537 5.71e-08 ***
## factor(origin)2    -0.0871501  0.6432645  -0.135 0.892303
## factor(origin)3     2.4816343  0.6334064   3.918 0.000106 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.906 on 383 degrees of freedom
## Multiple R-squared:  0.7547, Adjusted R-squared:  0.7496
## F-statistic: 147.3 on 8 and 383 DF,  p-value: < 2.2e-16
```

All three Regression Procedures (Stepwise Regression, Backward Elimination, and Forward Selection) have the same predictors that are significant compared to the manual fullmodel. Therefore, only “weight,” “cylinders,” “horsepower,” and “origin” will will be used for the possible reducedmodel, but the multicollinearity assumption must be checked beforehand.

All-possible-Regressions-Selection

```
##        rsquare AdjustedR         cp      AIC
## [1,] 0.6926304 0.6918423 90.9753171 2265.939
## [2,] 0.7249044 0.7213409 42.6827232 2230.453
## [3,] 0.7420802 0.7380607 17.9174680 2207.181
## [4,] 0.7547294 0.7496062  0.2063045 2191.469
## [5,] 0.7552240 0.7494570  1.4354723 2192.677
## [6,] 0.7555035 0.7490862  3.0000000 2194.229
```

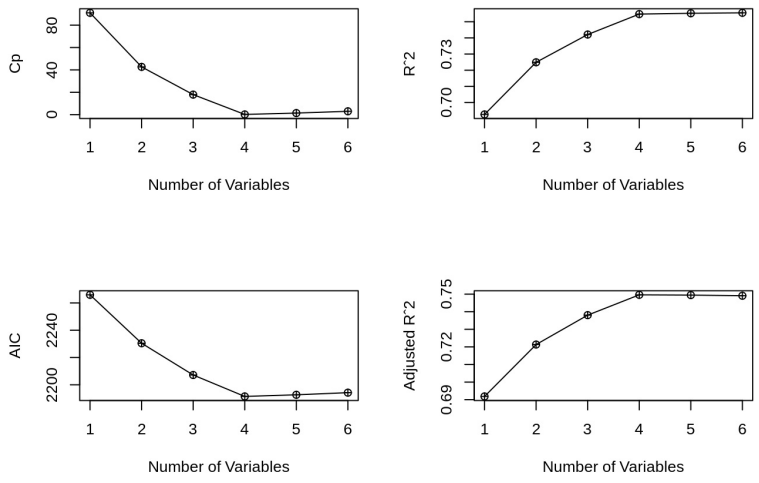

Figure 1

Cp (Mallows’ Cp): A small Cp value (small total mean square error - MSE) means that the model is relatively precise. Lower Cp values indicate better models.

AIC (Akaike Information Criterion): AIC is another measure of model fit that balances goodness of fit and model complexity. Lower AIC values indicate better models.

Based on the predictor selection, we wish to get a balance of the r-square, cp, and AIC. In this case, a model with 4 variables appears to have the lowest cp and AIC values, and the adjusted r-square is relatively high.

Multicollinearity

Before keeping all these variables, multicollinearity will be tested before reducing the model using the following hypothesis:

$$\mathrm{H}_0: \text{ There is NO multicollinearity that exists between the independent variables}$$
$$\mathrm{H}_A: \text{ There is multicollinearity that exists between the independent variables}$$

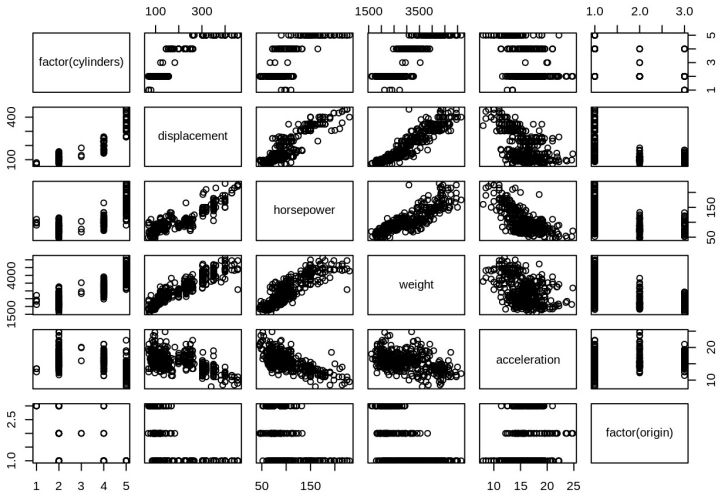

Figure 2

```
##
## Call:
## imcdiag(mod = fullmodel, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                         VIF detection
## factor(cylinders)4 27.7782         1
## factor(cylinders)5  1.9451         0
## factor(cylinders)6 22.6513         1
## factor(cylinders)8 35.1142         1
## displacement       22.9527         1
## horsepower         10.0774         1
## weight             11.2440         1
## acceleration        2.6650         0
## factor(origin)2     1.7077         0
## factor(origin)3     1.8274         0
##
## Multicollinearity may be due to factor(cylinders)4 factor(cylinders)6 factor(cylinders)8 displacement horsepow
er weight regressors
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## ==================================
```

VIF values of 1 indicate NO collinearity

VIF values of 1 - 5 indicate MODERATE collinearity

VIF values of > 5 indicate CRITICAL values of collinearity (p-value becomes questionable)

Based on the VIF test, most of the variables indicate collinearity due to the high VIF values. The independent variables, “cylinders,” “displacement,” “horsepower,” and “weight,” all have VIF values that are greater than 10, and therefore this indicates critical values of collinearity. Since “cylinders” has the highest VIF value, we will remove it from the model to see if anything changes.

Multicollinearity (No Cylinders Variable)

```
##
## Call:
## lm(formula = mpg ~ displacement + horsepower + weight + acceleration +
##     factor(origin), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.3131  -2.9021  -0.3505   2.2157  14.8163
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     43.3905094  2.4526690  17.691  < 2e-16 ***
## displacement     0.0026806  0.0072662   0.369 0.712395
## horsepower      -0.0584607  0.0167637  -3.487 0.000544 ***
## weight          -0.0049546  0.0008035  -6.166 1.77e-09 ***
## acceleration    -0.0234516  0.1232560  -0.190 0.849200
## factor(origin)2  1.0787637  0.7016362   1.537 0.124993
## factor(origin)3  2.8365644  0.6930489   4.093 5.19e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.166 on 385 degrees of freedom
## Multiple R-squared:  0.7194, Adjusted R-squared:  0.7151
## F-statistic: 164.5 on 6 and 385 DF,  p-value: < 2.2e-16
```

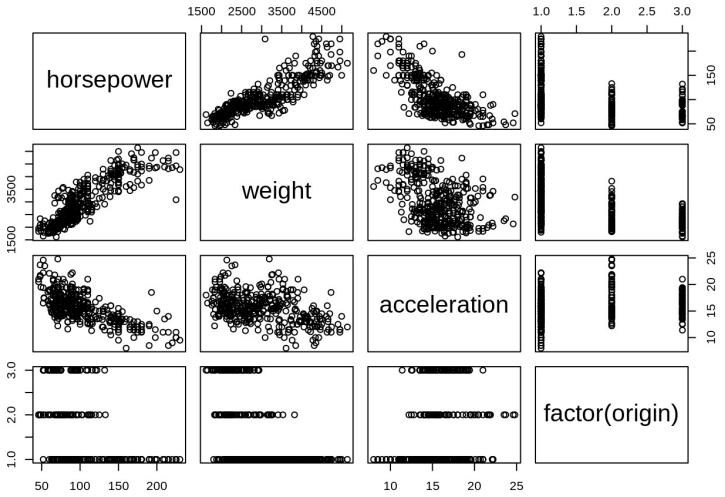

Figure 3

```
##
## Call:
## imcdiag(mod = fullmodel2, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                     VIF detection
## displacement    13.0228         1
## horsepower       9.3783         0
## weight          10.4923         1
## acceleration     2.6046         0
## factor(origin)2  1.5940         0
## factor(origin)3  1.7454         0
##
## Multicollinearity may be due to displacement weight regressors
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

Based on the VIF test without “cylinders”, most of the variables now indicate no multicollinearity, except “displacement” and “weight.” Since “weight” does not have a really high VIF compared to the originally removed “cylinders,” we plan to keep it in our model. Furthermore, “horsepower” now has a reduced VIF and the detection is below 1, which may indicate that “horsepower” and “cylinders” had a strong collinear relation. “displacement” will be removed and two reduced model will be created below, one with “cylinders” and one without to better see, through an ANOVA test, on whether to remove cylinders or not.

Reduced Model (no cylinders)

```
##
## Call:
## lm(formula = mpg ~ horsepower + weight + factor(origin), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.3772  -2.9241  -0.3991   2.2345  14.8164
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     42.7399583  1.1011112  38.815  < 2e-16 ***
## horsepower      -0.0535442  0.0109835  -4.875 1.59e-06 ***
## weight          -0.0048427  0.0005455  -8.878  < 2e-16 ***
## factor(origin)2  0.9611172  0.6402552   1.501    0.134
## factor(origin)3  2.7422385  0.6517055   4.208 3.21e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.157 on 387 degrees of freedom
## Multiple R-squared:  0.7193, Adjusted R-squared:  0.7164
## F-statistic: 247.9 on 4 and 387 DF,  p-value: < 2.2e-16
```

Reduced Model (with cylinders)

```
##
## Call:
## lm(formula = mpg ~ factor(cylinders) + horsepower + weight +
##     factor(origin), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4002  -2.2047  -0.5689   1.7741  16.3854
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        34.0149296  2.5525144  13.326  < 2e-16 ***
## factor(cylinders)4  8.6440072  2.0182505   4.283 2.34e-05 ***
## factor(cylinders)5 11.0624033  3.0904967   3.579 0.000389 ***
## factor(cylinders)6  5.0487106  2.1033193   2.400 0.016856 *
## factor(cylinders)8  7.5237315  2.2425536   3.355 0.000873 ***
## horsepower         -0.0652994  0.0117926  -5.537 5.71e-08 ***
## weight             -0.0039465  0.0006271  -6.293 8.52e-10 ***
## factor(origin)2    -0.0871501  0.6432645  -0.135 0.892303
## factor(origin)3     2.4816343  0.6334064   3.918 0.000106 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.906 on 383 degrees of freedom
## Multiple R-squared:  0.7547, Adjusted R-squared:  0.7496
## F-statistic: 147.3 on 8 and 383 DF,  p-value: < 2.2e-16
```

The reducedmodel with cylinders has a slight increase in the adjusted R-squared and a reduction in the Residual standard error. By using a Partial F test, we can confirm that the independent variable (removed from fullmodel) should be out of the model at significance level of 0.05.

ANOVA Partial F-test

By using a Partial F test, we can confirm that the independent variables (removed from) should be out of the model at significance level of 0.05. We will test the following hypothesis:

$$H_0: \text{ "displacement," "acceleration," and "cylinders" = 0 in the model Y = B0 + B1X1 + B2X2 + B3X3}$$
$$H_A: \text{ "displacement," "acceleration," and "cylinders" NOT EQUAL TO 0 in the model Y = B0 + B1X1 + B2X2 + B3X3}$$

ANOVA for fullmodel and reducedmodel (No cylinder present)

1
2

2 rows | 1-1 of 7 columns

The p-value for the ANOVA test between the full model and reduced model(no cylinder) is 1.223e-09, which is less than the alpha value of 0.05. Therefore, we REJECT the null hypothesis and infer the alternative. This suggests that one of those variables has a statistical significance on the model, and that their coefficient is NOT equal to 0.

$$H_0: \text{ "displacement" and "acceleration" = 0 in the model Y = B0 + B1X1 + B2X2 + B3X3}$$
$$H_A: \text{ "displacement" and "acceleration" NOT EQUAL TO 0 in the model Y = B0 + B1X1 + B2X2 + B3X3}$$

ANOVA for fullmodel and reducedmodel (cylinder present)

1
2

2 rows | 1-1 of 7 columns

The p-value for the ANOVA test between the full model and reduced model is 0.5476, which is greater than the alpha value of 0.05. Therefore, we FAIL to reject the null hypothesis. This suggests that there is no statistical significance to conclude that at least one of the coefficients for ‘displacement’ or ‘acceleration’ is not equal to 0 (as stated in the alternative hypothesis). In other words, we do not have enough statistical evidence to suggest that these predictors have an effect on the model.

Variable Selection

Based on the All-possible-Regressions-Selection values and the significance of cylinders on mpg found by the fullmodel,Stepwise Regression, Backward Elimination, and Forward Selection models, we decided to keep the cylinder variable in our reduced model. Even though the multicollinearity assumption has been broken, cylinders is important for fuel efficiency as shown in the boxplot below (Figure 1). Furthermore, according to the All-possible-Regressions-Selection, the best model would be one with four variables like in our reducedmodel.

Figure 1 - Boxplot of Mileage for Different Cylinder Size

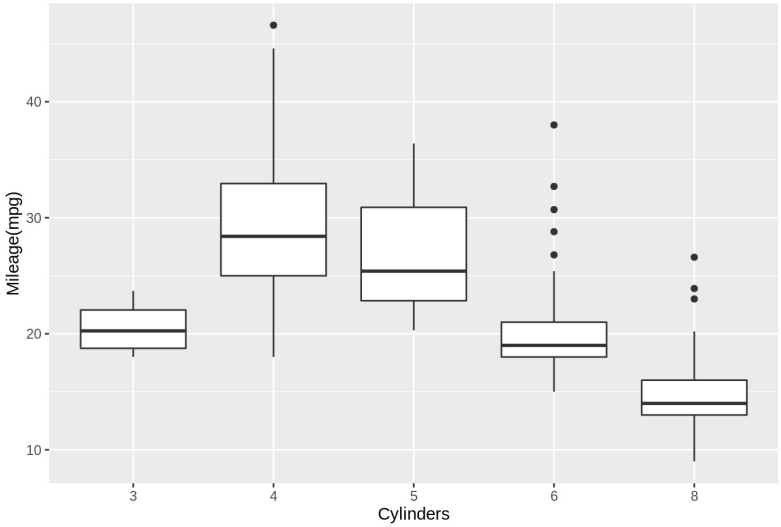

Figure 4

Interaction Model

To see if any interactions exist between the predictor variables, we can create an interacmodel and test the following hypothesis:

H_0: There is NO interaction effect between the independent variables in the model

H_A: There is an interaction effect between the independent variables in the model

```
##
## Call:
## lm(formula = mpg ~ (factor(cylinders) + horsepower + weight +
##     factor(origin))^2, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9334 -2.0661 -0.1924  1.6054 17.5299
##
## Coefficients: (6 not defined because of singularities)
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -3.339e+02  8.916e+02  -0.375   0.7082
## factor(cylinders)4               3.849e+02  8.916e+02   0.432   0.6663
## factor(cylinders)5               4.337e+02  8.922e+02   0.486   0.6272
## factor(cylinders)6               3.733e+02  8.917e+02   0.419   0.6757
## factor(cylinders)8               3.756e+02  8.919e+02   0.421   0.6739
## horsepower                       1.809e+01  4.736e+01   0.382   0.7027
## weight                          -6.033e-01  1.588e+00  -0.380   0.7043
## factor(origin)2                  4.040e+00  6.696e+00   0.603   0.5466
## factor(origin)3                  4.603e+00  7.066e+00   0.651   0.5152
## factor(cylinders)4:horsepower   -1.820e+01  4.736e+01  -0.384   0.7010
## factor(cylinders)5:horsepower   -1.850e+01  4.736e+01  -0.391   0.6963
## factor(cylinders)6:horsepower   -1.812e+01  4.736e+01  -0.383   0.7022
## factor(cylinders)8:horsepower   -1.820e+01  4.736e+01  -0.384   0.7010
## factor(cylinders)4:weight        5.963e-01  1.588e+00   0.375   0.7075
## factor(cylinders)5:weight        5.884e-01  1.588e+00   0.371   0.7112
## factor(cylinders)6:weight        5.966e-01  1.588e+00   0.376   0.7074
## factor(cylinders)8:weight        5.983e-01  1.588e+00   0.377   0.7066
## factor(cylinders)4:factor(origin)2 -2.599e+00  2.755e+00  -0.943   0.3462
## factor(cylinders)5:factor(origin)2         NA         NA      NA       NA
## factor(cylinders)6:factor(origin)2         NA         NA      NA       NA
## factor(cylinders)8:factor(origin)2         NA         NA      NA       NA
## factor(cylinders)4:factor(origin)3 -2.972e+00  2.531e+00  -1.174   0.2411
## factor(cylinders)5:factor(origin)3         NA         NA      NA       NA
## factor(cylinders)6:factor(origin)3         NA         NA      NA       NA
## factor(cylinders)8:factor(origin)3         NA         NA      NA       NA
## horsepower:weight                1.631e-05  2.368e-05   0.689   0.4914
## horsepower:factor(origin)2      -1.151e-01  5.299e-02  -2.171   0.0305 *
## horsepower:factor(origin)3      -1.474e-01  6.657e-02  -2.214   0.0274 *
## weight:factor(origin)2           2.960e-03  2.280e-03   1.298   0.1951
## weight:factor(origin)3           5.155e-03  3.278e-03   1.572   0.1167
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.682 on 368 degrees of freedom
## Multiple R-squared:  0.7905, Adjusted R-squared:  0.7774
## F-statistic: 60.38 on 23 and 368 DF,  p-value: < 2.2e-16
```

There appears to be one interaction term between “horsepower” and “origin.” This is because the p-values for both horsepower:factor(origin)2 and horsepower:factor(origin)3 are 0.0305 and 0.0274 respectively. These values are less than an alpha value of 0.05, which means we can REJECT the null hypothesis and infer the alternative that states that there is an interaction effect between the independent variables in the model.

Interaction model with the interaction term

```
##
## Call:
## lm(formula = mpg ~ factor(cylinders) + horsepower + weight +
##     factor(origin) + horsepower:origin, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6525 -2.2614 -0.4183  1.5222 16.4393
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         33.2860094  2.5128789  13.246  < 2e-16 ***
## factor(cylinders)4   6.9937917  2.0257183   3.452 0.000618 ***
## factor(cylinders)5   9.3422078  3.0657524   3.047 0.002470 **
## factor(cylinders)6   3.4493793  2.1049267   1.639 0.102096
## factor(cylinders)8   4.2967744  2.3506473   1.828 0.068343 .
## horsepower           0.0060120  0.0215692   0.279 0.780601
## weight              -0.0037599  0.0006175  -6.089 2.77e-09 ***
## factor(origin)2      4.3378334  1.2938478   3.353 0.000880 ***
## factor(origin)3     10.9214441  2.2418011   4.872 1.62e-06 ***
## horsepower:origin   -0.0511702  0.0130586  -3.919 0.000106 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.834 on 382 degrees of freedom
## Multiple R-squared:  0.7642, Adjusted R-squared:  0.7587
## F-statistic: 137.6 on 9 and 382 DF,  p-value: < 2.2e-16
```

The interacmodel model will be tested for higher-order models below.

Higher Order Model

1
2
3
4
5

5 rows | 1-1 of 6 columns

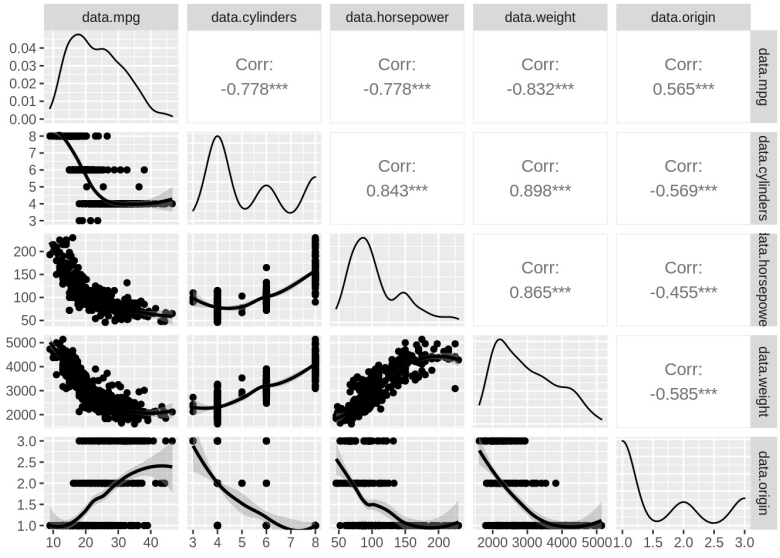

Figure 5

By looking at the higher order model, there appears to be significance for both horsepower and weight. Below, we will check the square model to see the impacts on our model.

Square Model

```
##
## Call:
## lm(formula = mpg ~ factor(cylinders) + horsepower + I(horsepower^2) +
##     weight + I(weight^2) + factor(origin) + horsepower:origin,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4517 -2.2092 -0.3657  1.7551 15.7949
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.601e+01  4.742e+00   9.703  < 2e-16 ***
## factor(cylinders)4  7.048e+00  1.990e+00   3.542 0.000446 ***
## factor(cylinders)5  9.645e+00  3.037e+00   3.176 0.001615 **
## factor(cylinders)6  4.396e+00  2.090e+00   2.103 0.036117 *
## factor(cylinders)8  5.980e+00  2.347e+00   2.548 0.011242 *
## horsepower         -2.040e-01  6.125e-02  -3.330 0.000953 ***
## I(horsepower^2)     5.385e-04  1.607e-04   3.351 0.000886 ***
## weight             -5.822e-03  2.691e-03  -2.163 0.031138 *
## I(weight^2)         3.787e-07  3.932e-07   0.963 0.336113
## factor(origin)2     3.593e-01  1.614e+00   0.223 0.823897
## factor(origin)3     3.467e+00  2.889e+00   1.200 0.230852
## horsepower:origin  -7.542e-03  1.681e-02  -0.449 0.653944
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.765 on 380 degrees of freedom
## Multiple R-squared:  0.7738, Adjusted R-squared:  0.7673
## F-statistic: 118.2 on 11 and 380 DF,  p-value: < 2.2e-16
```

From the square model, the p-value for the horsepower variable squared is 0.000886 whereas the weight variable is 0.336113. Therefore, horsepower^2 will be added to the model since it's significant, whereas the weight^2 is not.

Cube Model

```
##
## Call:
## lm(formula = mpg ~ factor(cylinders) + horsepower + I(horsepower^2) +
##     I(horsepower^3) + weight + factor(origin) + horsepower:origin,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7565 -2.3238 -0.2804  1.7637 15.8582
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.934e+01  5.905e+00   8.355 1.24e-15 ***
## factor(cylinders)4  7.045e+00  1.987e+00   3.545 0.000441 ***
## factor(cylinders)5  9.415e+00  3.009e+00   3.129 0.001891 **
## factor(cylinders)6  4.122e+00  2.073e+00   1.989 0.047411 *
## factor(cylinders)8  5.210e+00  2.409e+00   2.163 0.031139 *
## horsepower         -3.932e-01  1.457e-01  -2.698 0.007293 **
## I(horsepower^2)     2.054e-03  1.096e-03   1.874 0.061766 .
## I(horsepower^3)    -3.640e-06  2.701e-06  -1.348 0.178595
## weight             -3.320e-03  6.172e-04  -5.378 1.32e-07 ***
## factor(origin)2     9.288e-02  1.636e+00   0.057 0.954758
## factor(origin)3     3.205e+00  2.886e+00   1.111 0.267342
## horsepower:origin  -5.816e-03  1.687e-02  -0.345 0.730422
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.761 on 380 degrees of freedom
## Multiple R-squared:  0.7744, Adjusted R-squared:  0.7678
## F-statistic: 118.5 on 11 and 380 DF,  p-value: < 2.2e-16
```

By keeping the squared horsepower value and cubing it, the p-value is 0.178595. This means that horsepower^3 is not significant for the model, and therefore will be removed. Instead, through higher order model checks, only horsepower^2 will be kept in the proposed model.

Proposed Model

```
##
## Call:
## lm(formula = mpg ~ factor(cylinders) + horsepower + I(horsepower^2) +
##     weight + factor(origin) + horsepower:origin, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4463 -2.2982 -0.3434  1.7760 15.7298
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        42.9163816  3.4893884  12.299  < 2e-16 ***
## factor(cylinders)4  7.0076868  1.9890044   3.523 0.000478 ***
## factor(cylinders)5  9.2606126  3.0102569   3.076 0.002247 **
## factor(cylinders)6  4.1528498  2.0746180   2.002 0.046021 *
## factor(cylinders)8  5.9547754  2.3468076   2.537 0.011566 *
## horsepower         -0.2143289  0.0602955  -3.555 0.000426 ***
## I(horsepower^2)     0.0005906  0.0001513   3.903 0.000112 ***
## weight             -0.0032992  0.0006177  -5.341 1.59e-07 ***
## factor(origin)2     0.6807706  1.5785558   0.431 0.666522
## factor(origin)3     4.1272148  2.8063095   1.471 0.142200
## horsepower:origin  -0.0106022  0.0165056  -0.642 0.521039
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.765 on 381 degrees of freedom
## Multiple R-squared:  0.7733, Adjusted R-squared:  0.7673
## F-statistic: 129.9 on 10 and 381 DF,  p-value: < 2.2e-16
```

Compare the adjusted R-squared and Residual standard error for reducedmodel and proposedmodel.

```
## Adjusted R-squared for Reduced Model: 0.7496062
```

```
## Adjusted R-squared for Proposed Model: 0.7673217
```

```
## Residual Standard Error for Reduced Model: 3.905576
```

```
## Residual Standard Error for Proposed Model: 3.764881
```

The adjusted R-squared increased from the reducedmodel(0.7496062) to the proposedmodel(0.7673217).

The standard error decreased from the reducedmodel(3.905576) to the proposedmodel (3.764881).

The final proposedmodel is:

$$Y_{\text{mpg}} = 42.9164 + 7.0076X_{\text{Cy4}} + 9.2606X_{\text{Cy5}} + 4.1529X_{\text{Cy6}} + 5.9548X_{\text{Cy8}} - 0.2143X_{\text{Hp}} + 0.0006X_{\text{Hp}}^2 - 0.0033X_{\text{Wt}} + 0.6808X_{\text{Org2}} + 4.1272X_{\text{Org3}} - 0.0106022X_{\text{Hp:Org}}$$

Assumption Checks

Linearity

Figure 2 - Linearity Assumption on Proposed Linear Regression Model

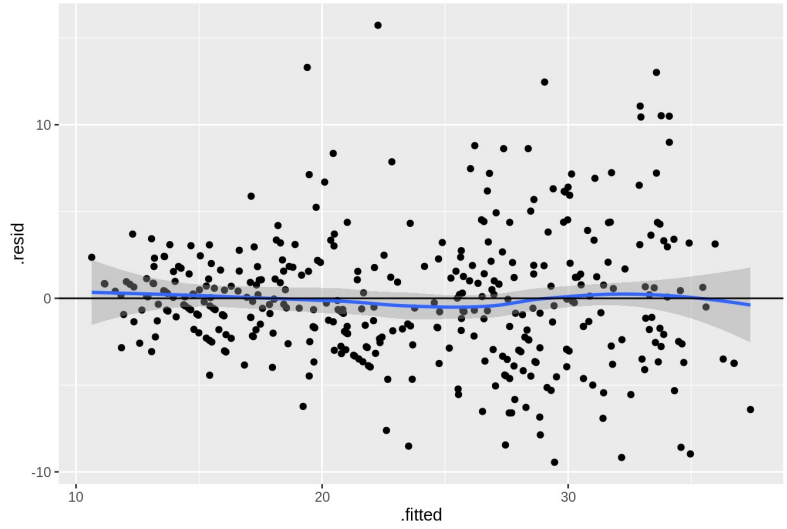

Figure 6

There does appear to be a problem with the linearity assumption. Most of the data points are scattered around the zero-residual line, but there is a slight curve around the right end of the zero line. This indicates that the linear model may not be the best fit for the data and the assumptions of linearity and constant variance of the residuals are not met. These will be checked below.

Homoscedasticity

To statistically test for heteroscedasticity, the following hypothesis will be used using the Breusch-Pagan test:

H_0: Heteroscedasticity is NOT present (homoscedasticity)

H_A: Heteroscedasticity is present

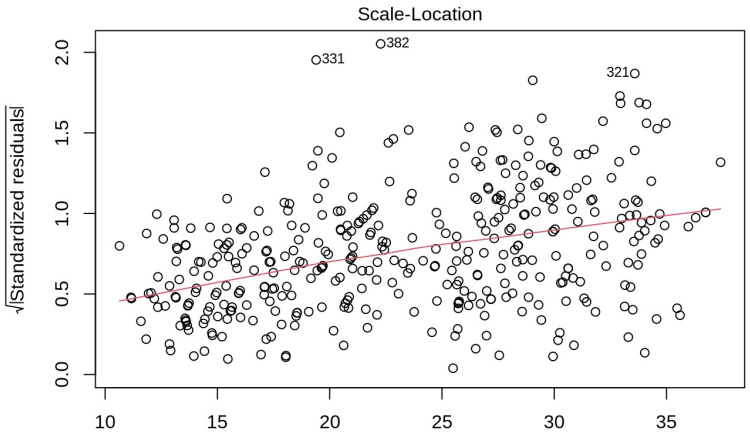

Figure 7

```
##
##  studentized Breusch-Pagan test
##
## data:  proposedmodel
## BP = 60.654, df = 10, p-value = 2.725e-09
```

The scale-location plot does not appear to be straight line meaning that just visually, we won't assume that the data meets the assumption of homoscedasticity.

The p-value for the Breusch-Pagan test is 7.487e-05, which is less than an alpha value of 0.05. Therefore, we REJECT the null hypothesis that heteroscedasticity is NOT present, and infer the alternative that suggests that heteroscedasticity is present in the proposedmodel.

Normality

To statistically test for normality, the following hypothesis will be used using Shapiro Wilk test:

$$H_0: \text{ The sample data are normally distributed}$$
$$H_A: \text{ The sample data are NOT normally distributed}$$

Figure 3 - Histogram for Residuals

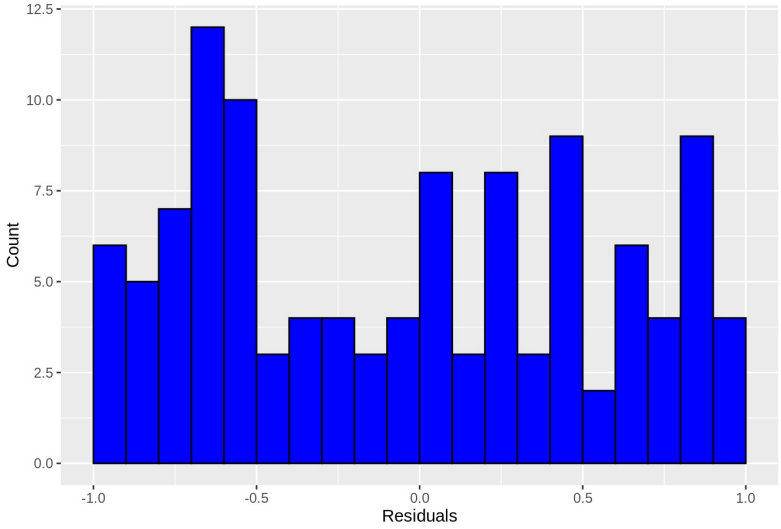

Figure 8

Figure 4 - Normal QQ-Plot

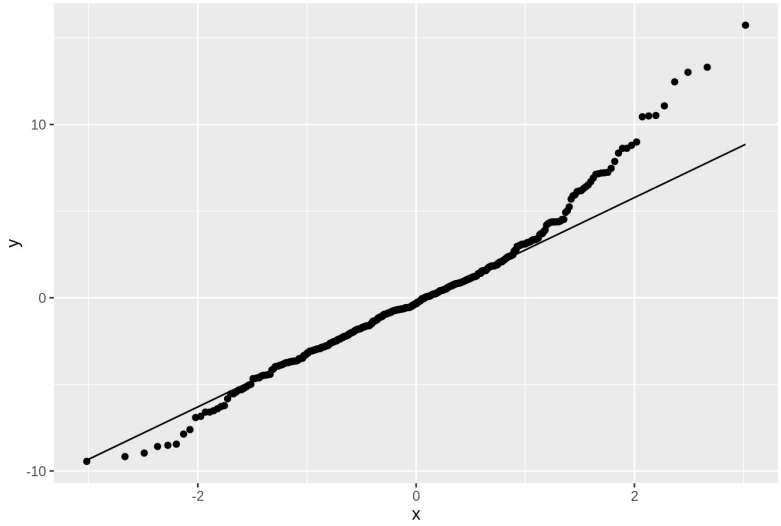

Figure 9

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(proposedmodel)
## W = 0.96412, p-value = 3.3e-08
```

Based on the histogram and normal QQ plot, the sample data appears to NOT be normally distributed due to the shape of the histogram as well as most data points are at the polar ends on the QQ plot (far from the center line on the right hand side). Furthermore, by running a Shapiro Wilk test, the p-value obtained is 3.3e-08 which is less than an alpha value of 0.05. This means that we can reject the null hypothesis and instead infer the alternative that states that the data points are NOT normally distributed.

Outliers

Below, we will check for outliers in the data by plotting leverage vs fitted graphs as well as Cook's distance to get a rough estimate of any influencing outliers.

```
##          9         14         71         95        111        116        123
## 0.08764733 0.13673989 0.25027449 0.09477047 0.25413752 0.10760693 0.10151199
##        242        273        296        326        331        332        358
## 0.25581769 0.34744266 0.34002274 0.33741708 0.14183413 0.25002450 0.09298030
```

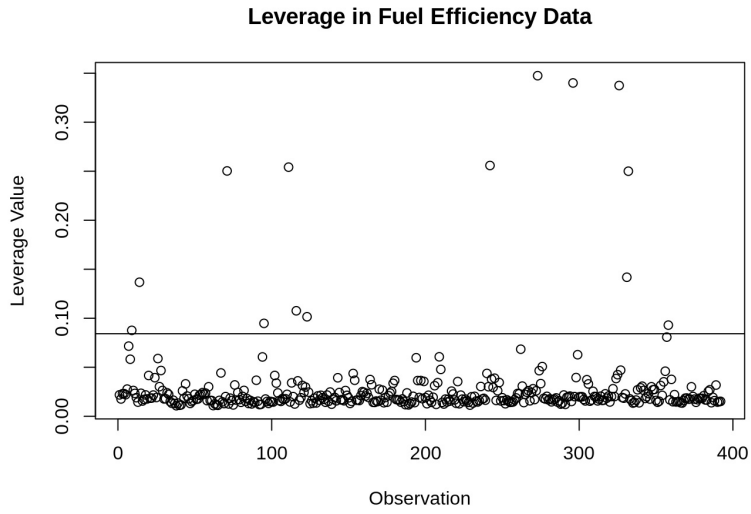

Figure 10

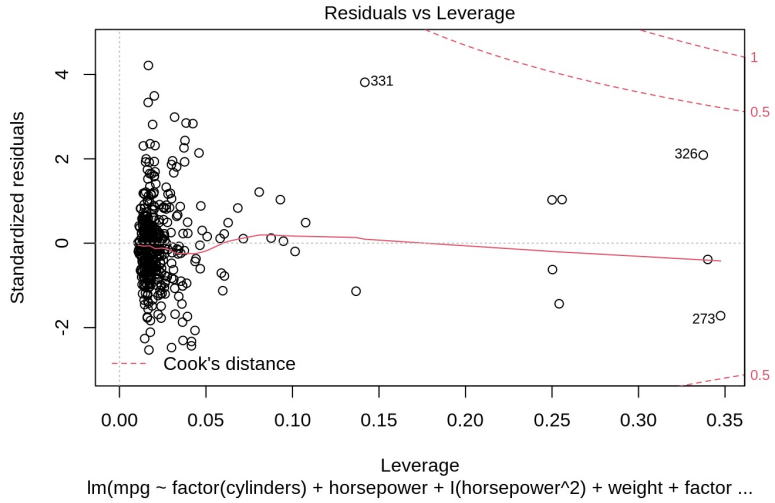

Figure 11

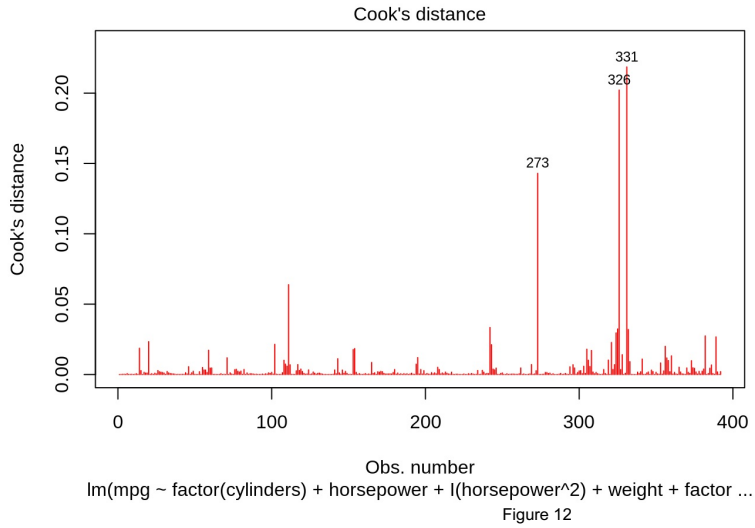

Figure 12

1
2
3
4
5

5 rows | 1-1 of 10 columns

There does appear to be outliers for this dataset, but a general rule for Cook's distance is values greater than 0.5 to have an influence, but since all the data points are not greater than 0.5 in the Residuals vs Leverage plot,we can assume that they don't have significant influence. This is also true for the Cook's Distance plot. There are no influential outliers greater than 0.5 and therefore, we will keep all the outliers.

Assumption Results

Based on the results above, the proposedmodel appears to have failed most assumptions. In this case, our goal is to transform the dataset using a Box-Cox transformation to see if the assumptions can be met through transformation. If so, that will be the final model that will be used for predicting the fuel mileage of a vehicle based on the independent variables.

Proposed Model Transformation: Box-Cox Transformation

In order to transform the data using a Box-Cox transformation, we first have to find the best lambda to transform the data using that value. The best lambda in this case is -0.2323232.

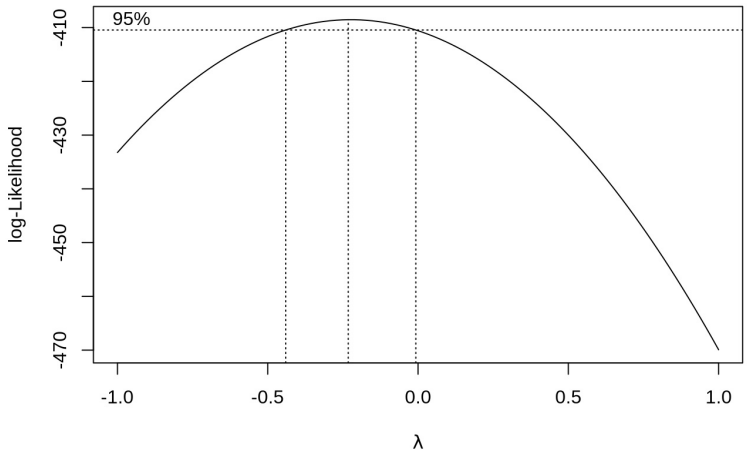

Figure 13

```
## The best lambda is the x-value corresponding to the maximum y-value in the 'bc' data: -0.2323232
```

Create the Box-Cox Model using the best lambda

```
##
## Call:
## lm(formula = (((mpg^bestlambda) - 1)/bestlambda) ~ factor(cylinders) +
##     horsepower + weight + factor(origin) + horsepower:origin,
##     data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.186375 -0.044952 -0.004655  0.045355  0.269866
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.483e+00  4.638e-02  53.531  < 2e-16 ***
## factor(cylinders)4  1.403e-01  3.739e-02   3.752 0.000202 ***
## factor(cylinders)5  1.949e-01  5.659e-02   3.444 0.000637 ***
## factor(cylinders)6  7.742e-02  3.885e-02   1.993 0.047019 *
## factor(cylinders)8  8.147e-02  4.339e-02   1.878 0.061185 .
## horsepower         -1.084e-03  3.981e-04  -2.722 0.006777 **
## weight             -8.736e-05  1.140e-05  -7.664  1.5e-13 ***
## factor(origin)2     1.182e-02  2.388e-02   0.495 0.620810
## factor(origin)3     6.925e-02  4.138e-02   1.674 0.095022 .
## horsepower:origin  -2.243e-04  2.410e-04  -0.931 0.352609
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07077 on 382 degrees of freedom
## Multiple R-squared:  0.8244, Adjusted R-squared:  0.8202
## F-statistic: 199.2 on 9 and 382 DF,  p-value: < 2.2e-16
```

How does the Box-Cox Model Compare to the Proposed Model?

```
## Adjusted R-squared for Proposed Model: 0.7673217
```

```
## Adjusted R-squared for BC Model: 0.8202474
```

```
## Residual Standard Error for Proposed Model: 3.764881
```

```
## Residual Standard Error for BC Model: 0.07077338
```

The adjusted R-squared increased from the proposedmodel(0.7673217) to the bcmodel(0.8202474).

The standard error decreased from the proposedmodel(3.764881) to the bcmodel (0.07077338).

Below, we will check whether the assumptions of the model have been impacted.

Did the Box-Cox Transformation Change the Assumptions?

Linearity

Figure 5 - Linearity Assumption on the Box-Cox Model

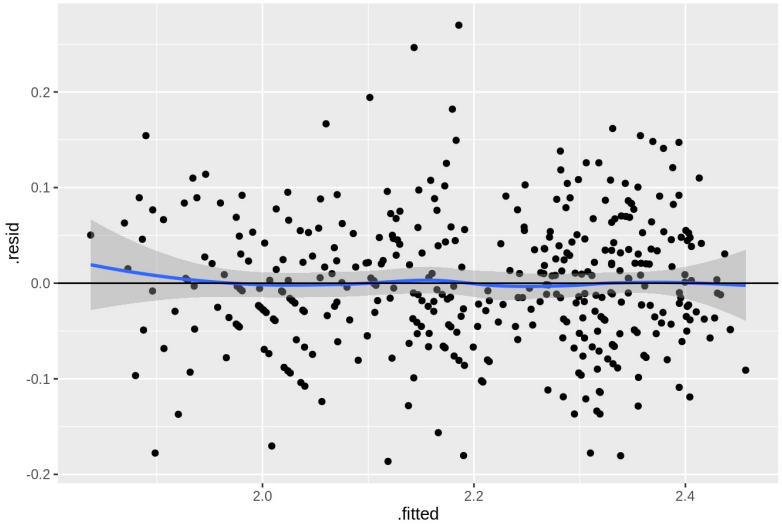

Figure 14

The linearity assumption using the residuals vs. fitted values for the box-cox model appears to be more flattened out in Figure 5, compared to Figure 2 in the proposed model. This indicates that the linearity assumption has improved after the transformation.

Heteroscedasticity

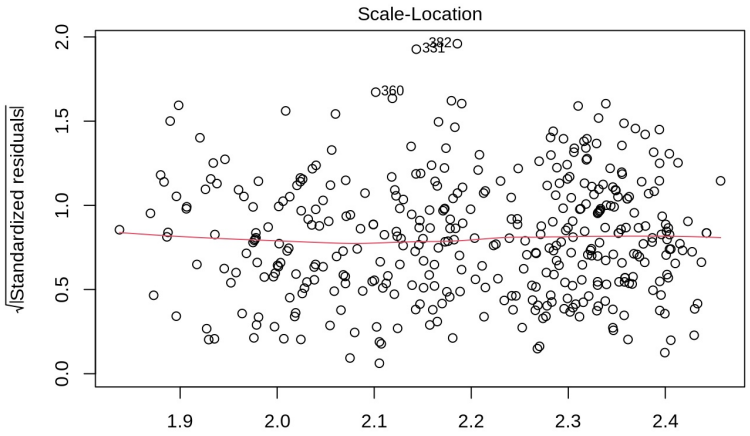

Figure 15

```
##
##  studentized Breusch-Pagan test
##
## data:  bcmodel
## BP = 11.938, df = 9, p-value = 0.2168
```

For the box-cox model, the p-value for the Breusch-Pagan test is 0.2168, which is greater than an alpha value of 0.05. Therefore, we fail to reject the null hypothesis that heteroscedasticity is NOT present and infer that there is statistical evidence for homoscedasticity. By looking at the Scale-Location plot, we can also see a more straight line. In this case, the transformed model has fixed the heteroscedasticity assumption.

Normality

Figure 6 - Histogram for Residuals

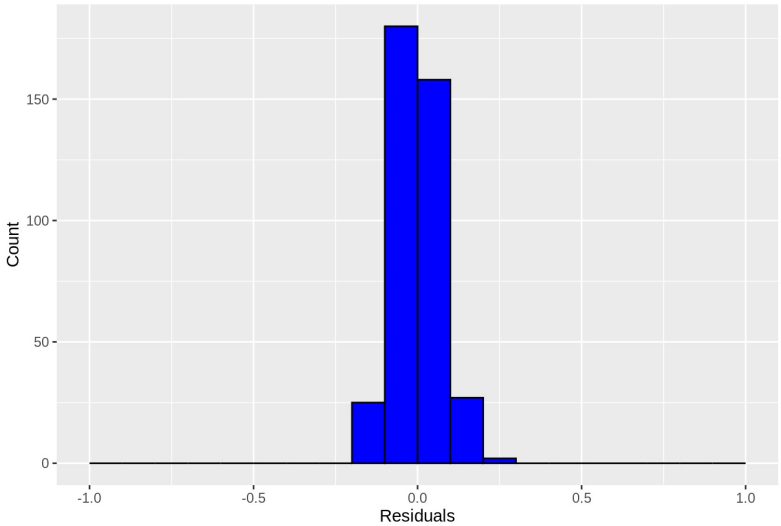

Figure 16

Figure 7 - Normal QQ-Plot For Box-Cox Transformed Model

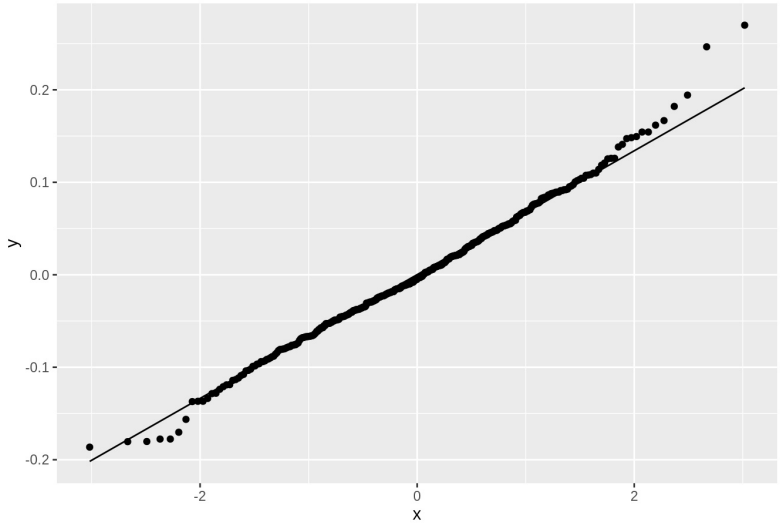

Figure 17

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(bcmodel)
## W = 0.99283, p-value = 0.0578
```

Based on the Figure 6 (Histogram) and Figure 7 (Normal QQ plot), the sample data appears to be normally distributed due to the shape of the histogram as well as most data points are now more so on the straight line. Furthermore, by running a Shapiro Wilk test, the p-value obtained is 0.0578 which is greater than an alpha value of 0.05. This means that we fail to reject the null hypothesis and instead have statistical evidence to suggest that the data points are normally distributed. Again, the transformed model fixed one of our assumptions.

Outliers

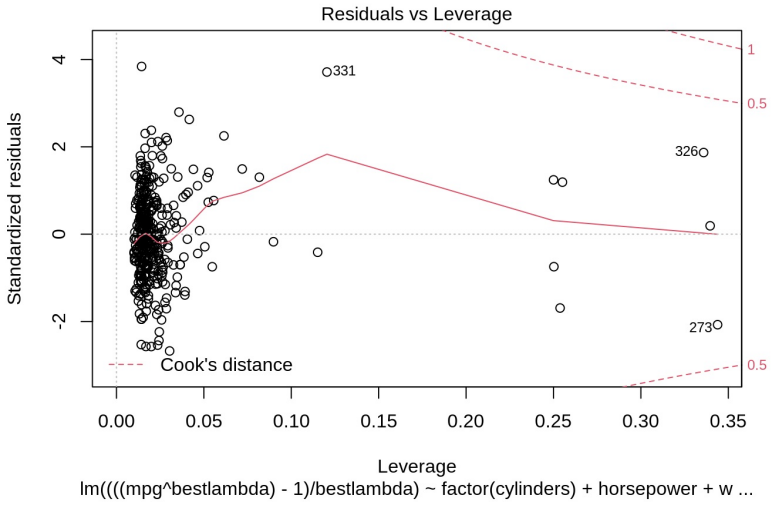

Figure 18

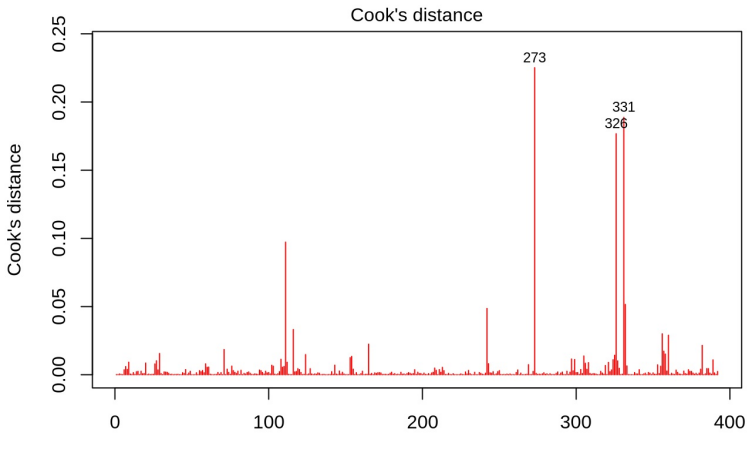

Figure 19

Here, we are taking the extreme points from the Cook's Distance graph for prediction to see how good our prediction and model are.

```
##    mpg cylinders displacement horsepower weight acceleration model.year origin
## 10  15         8          390        190   3850          8.5         70      1
##              car.name
## 10 amc ambassador dpl
```

```
##    mpg cylinders displacement horsepower weight acceleration model.year origin
## 11  15         8          383        170   3563           10         70      1
##               car.name
## 11 dodge challenger se
```

```
##    mpg cylinders displacement horsepower weight acceleration model.year origin
## 12  14         8          340        160   3609            8         70      1
##              car.name
## 12 plymouth 'cuda 340
```

The outliers and Cook's distance after the transformation are still somewhat similar. This is due to us not removing any outliers since Cook's distance was not greater than 0.5.

Model Predictions: Proposed Model Vs. Box-Cox Transformation Model

Below, we will use both the proposed model and the box-cox model to predict fuel mileage based on datapoints provided by the data, and to see the significance of each model's prediction. We are using there data frames new_data1, new_data2 and new_data3 which are the extreme points that we found on the Cook's Distance graph.

Proposed Model

```
## Prediction for point 9: 14.75379 7.254438 22.25314
```

```
## Prediction for point 10: 15.9468 8.45841 23.43519
```

```
## Prediction for point 11: 16.09529 8.622889 23.5677
```

Box-Cox Transformation Model

```
## Prediction for point 9: 14.17187 10.99998 18.55011
```

```
## Prediction for point 10: 15.59851 12.04665 20.53377
```

```
## Prediction for point 11: 15.869 12.25099 20.89864
```

Based on the prediction results, the box-cox model appears better at predicting fuel mileage due to the model being a better fit and therefore, it will be used as our final significant model. Furthermore, the lower and upper intervals are more precise, and therefore provide a better estimate of the actual mpg prediction.

Result

In our first step, we tested all variables in a full model.

The hypothesis for this test to determine significance was:

$$H_0: \ \beta1 = \beta2 = ... = \beta p = 0$$
$$H_A: \ \text{at least one } \beta i \text{ is NOT EQUAL TO 0 } (i = 1,2,...,p)$$

After testing the global t-test, we got a p-value of less than alpha. The full model is shown below.

$$Y_{\text{mpg}} = 35.66 + 8.54\,X_{\text{factor(cylinders)4}} + 10.86\,X_{\text{factor(cylinders)5}} + 4.56\,X_{\text{factor(cylinders)6}} + 6.60\,X_{\text{factor(cylinders)8}} + 0.01\,X_{\text{displacement}} - 0.08\,X_{\text{horsepower}} - 0.00\,X_{\text{weight}} - 0.09\,X_{\text{acceleration}} + 0.09\,X_{\text{factor(origin}}$$

Afterwards, we did the regression procedure to verify the full model results. Every regression procedure provided us with similar results compared to our full model. Then we did the all possible regression selection. Based on the former, we confirmed a model with four variables would be the most ideal due to the having the lowest cp and AIC values and a relatively high adjusted r-square.

For the next step, we did the Multicolinearity test. Based on the VIF test, most of the variables indicate colinearity due to the high VIF values. We removed the Cylinder predictor which was giving the most colinear effect. We then created two reduced model, one with cylinder and without cylinder. The model is shown below.

$$Y_{\text{mpg}} = 42.74 - 0.05\,X_{\text{horsepower}} - 0.00\,X_{\text{weight}} + 0.96\,X_{\text{factor(origin)2}} + 2.74\,X_{\text{factor(origin)3}}$$

$$Y_{\text{mpg}} = 34.01 + 8.64\,X_{\text{factor(cylinders)4}} + 11.06\,X_{\text{factor(cylinders)5}} + 5.05\,X_{\text{factor(cylinders)6}} + 7.52\,X_{\text{factor(cylinders)8}} - 0.07\,X_{\text{horsepower}} - 0.00\,X_{\text{weight}} - 0.09\,X_{\text{factor(origin)2}} + 2.48\,X_{\text{factor(origin)3}}$$

Next we conducted two ANOVA tests to confirm whether the reduced model(with cylinder and without cylinder) are better or not. The Hypothesis for both hypothesis test are down below.

$$H_0: \ \text{"displacement," "acceleration," and "cylinders" = 0 in the model Y = B0 + B1X1 + B2X2 + B3X3}$$
$$H_A: \ \text{"displacement," "acceleration," and "cylinders" NOT EQUAL TO 0 in the model Y = B0 + B1X1 + B2X2 + B3X3}$$

$$H_0: \ \text{"displacement" and "acceleration" = 0 in the model Y = B0 + B1X1 + B2X2 + B3X3}$$
$$H_A: \ \text{"displacement" and "acceleration" NOT EQUAL TO 0 in the model Y = B0 + B1X1 + B2X2 + B3X3}$$

For the first ANOVA, the p-value was less than alpha but for the second test, it was greater. From this, we came to an inference that the cylinder indeed is a significant predictor and the reduced model with cylinder is better than the model without this variable. Therefore, we kept the cylinders variable even though it had an issue with multicolinearity.

We then created an interaction model. The hypothesis are down below.

$$H_0: \ \text{There is NO interaction effect between the independent variables in the model}$$
$$H_A: \ \text{There is an interaction effect between the independent variables in the model}$$

After the test, we came to a conclusion and selected the below model.

$$Y_{\text{mpg}} = 33.29 + 6.99\,X_{\text{factor(cylinders)4}} + 9.34\,X_{\text{factor(cylinders)5}} + 3.45\,X_{\text{factor(cylinders)6}} + 4.30\,X_{\text{factor(cylinders)8}} + 0.01\,X_{\text{horsepower}} - 0.00\,X_{\text{weight}} + 4.34\,X_{\text{factor(origin)2}} + 10.92\,X_{\text{factor(origin)3}} - 0.05\,X_{\text{horsepo}}$$

We also did a test for finding the higher order model, compared all the relevant things from summary such as RMSE, R2Adj etc and we came up with the proposed model. The model is shown below.

$$Y_{\text{mpg}} = 42.92 + 7.01\,X_{\text{factor(cylinders)4}} + 9.26\,X_{\text{factor(cylinders)5}} + 4.15\,X_{\text{factor(cylinders)6}} + 5.95\,X_{\text{factor(cylinders)8}} - 0.21\,X_{\text{horsepower}} + 0.00\,X^2_{\text{I(horsepower)}} - 0.00\,X_{\text{weight}} + 0.68\,X_{\text{factor(origin)2}} + 4.13\,X_{\text{factor(orig}}$$

Afterwards, we did the linear assumptions such as Linearity, Homoscedasticity, Normality and Outliers. The hypothesis for the former are down below.

H_0: Heteroscedasticity is NOT present (homoscedasticity)

H_A: Heteroscedasticity is present

H_0: The sample data are normally distributed

H_A: The sample data are NOT normally distributed

Our model failed most of the assumptions except linearity and any influential outliers. So, in order to fix Normality and Homoscedasticity. a Box-Cox transformation was used to see if the assumption could be met. From that, we got our best lambda as -0.2323232 and used that to propose a transformed model, which is:

$$Y_{\text{mpg}} = 2.48 + 0.14\,X_{\text{factor(cylinders)4}} + 0.19\,X_{\text{factor(cylinders)5}} + 0.08\,X_{\text{factor(cylinders)6}} + 0.08\,X_{\text{factor(cylinders)8}} - 0.00\,X_{\text{horsepower}} + 0.00\,X_{\text{I(horsepower^2)}} - 0.00\,X_{\text{weight}} + 0.01\,X_{\text{factor(origin)2}} + 0.07\,X_{\text{factor(ori}}$$

This fixed our Normality issue and the homoscedasticity. RMSE decreased and R2Adj increased. Then we took 3 extreme points from the Cook's distance graph and predicted with proposed model and transformed model. The transformed model was better at predicting and also had a narrower confidence interval.

Ultimately, we selected the transformed model as our final model. The model is shown below.

$$Y_{\text{mpg}} = 2.48 + 0.14\,X_{\text{factor(cylinders)4}} + 0.19\,X_{\text{factor(cylinders)5}} + 0.08\,X_{\text{factor(cylinders)6}} + 0.08\,X_{\text{factor(cylinders)8}} - 0.00\,X_{\text{horsepower}} + 0.00\,X_{\text{I(horsepower^2)}} - 0.00\,X_{\text{weight}} + 0.01\,X_{\text{factor(origin)2}} + 0.07\,X_{\text{factor(ori}}$$

Discussion

In plain terms, we used statistical modeling to figure out what makes cars get better gas mileage. We tested different combinations of engine features to find the best fitting model. The one that worked best included number of cylinders, horsepower, weight, where the car was made, and an interaction between horsepower and origin.

We did have to transform the gas mileage data to follow normal statistics rules about bell curve distributions and equal scattering. After fixing that through a Box-Cox transformation, our final model did a good job predicting gas mileage, especially for cars that were outside the norm. It was better than our first tries.

Overall, we met our goal of finding the engine things that matter most for efficiency. Step-by-step testing let us shrink down from a big model to a tighter one. We also solved issues about data shape when the numbers didn't follow the usual rules.

To make the model even better, we could add in extra pieces like transmission and rear-wheel/4-wheel drive next time. Checking more interactions might also help. And transforming earlier when the data shape is off could streamline things.

In short, we used stats to pinpoint cylinder number, horsepower, weight, origin, and horsepower-origin interactions as big impacts on MPG. Transforming MPG gave us a model that follows standard data guidelines. These results break down what makes cars sip or gulp fuel based on what we had to work with. That information could guide car companies in designing vehicles that use less gas. We explained the findings in regular language so anyone can understand them.

Conclusion

In conclusion, the objective of this report was to create a multiple linear regression model to predict the fuel efficiency (dependent variable = mpg) of a vehicle given multiple independent variables such as cylinders, displacement, horsepower, weight, acceleration, and it's origin. Once a reduced model was proposed, the model failed multiple assumptions such as multicollinearity, heteroscedasticity, and normality. By creating a transformed model using the Box-Cox transformation, most of the assumptions like heteroscedasticity and normality were met. This model was then used to predict the miles per gallon (mpg) and was more accurate than the proposed model due to a higher adjusted r-square, lower standard error, and a more fitted confidence interval when it came to prediction.

References

[1]'StatLib—Datasets Archive'. Available: http://lib.stat.cmu.edu/datasets/ (http://lib.stat.cmu.edu/datasets/). [Accessed: Nov. 29, 2023]