STA302 FINAL PROJECT

Group members: Collin Downs, Jaekang Lee, Jonathan Tillmann, Jiayan Li, Aly Abdel Baki

Due date: August 19, 2020

Introduction and data description:

The goal of the model is to examine which variables in the manufacturing of a car correlate to better fuel efficiency in the form of miles per gallon of fuel. The question that we wish to answer is which variables have the highest effect on miles per gallon and how to maximize the mileage of the car. This model is being developed because examining all the variables that go into miles per gallon will allow us to design more efficient cars that will help with climate change. The model will also allow us to make more economical vehicles that will help increase the sales of the car, as many people are looking for cars with good mileage. Furthermore, car developers also have to meet a set of standard regulations for mileage and if they do not meet them they often have to pay a tax. This model will help developers design a vehicle that meets the industry standard so that they can save money by avoiding the regulation tax. The model will be a multiple linear regression done in R with variables taken from the dataset mtcars that has data from 32 different cars from 1973 - 1974. Not all the variables will be used as they will first be examined for multicollinearity and the unnecessary ones will be removed before the final model is created. Variables will also be transformed to better fit the model in the case of non-linear interactions with the response or other predictor variables. The way the model will serve its purpose is that it will find the variables correlated with miles per gallon. This will allow developers to see which variables they should change in the design of the car in order for them to meet industry standards so they can save both money and reduce negative environmental impact.

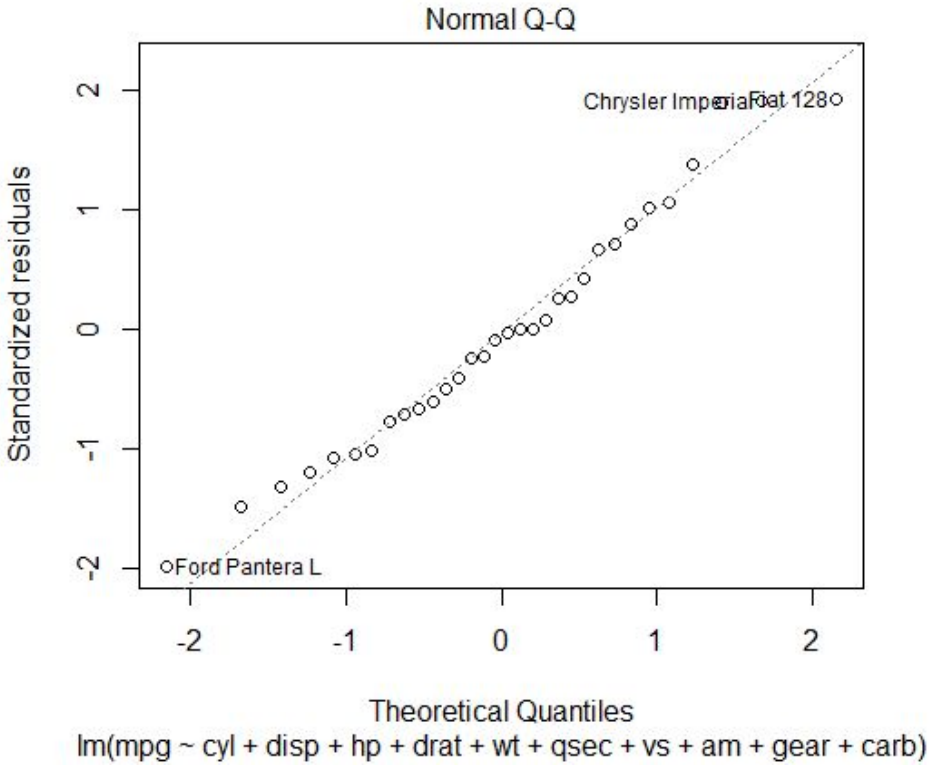
Exploratory Section:

The data set has 11 variables with 32 different observations. The following table lists the variables in the dataset and gives their description.

Variables	**Description**
mpg	Miles per gallon
cyl	Number of cylinders
disp	Displacement
hp	Horsepower
drat	Rear axle ratio
wt	Weight
qsec	¼ mile time
vs	V engine or straight engine
am	Automatic or manual (0 or 1)
gear	Number of forward gears
carb	Number of carburetors

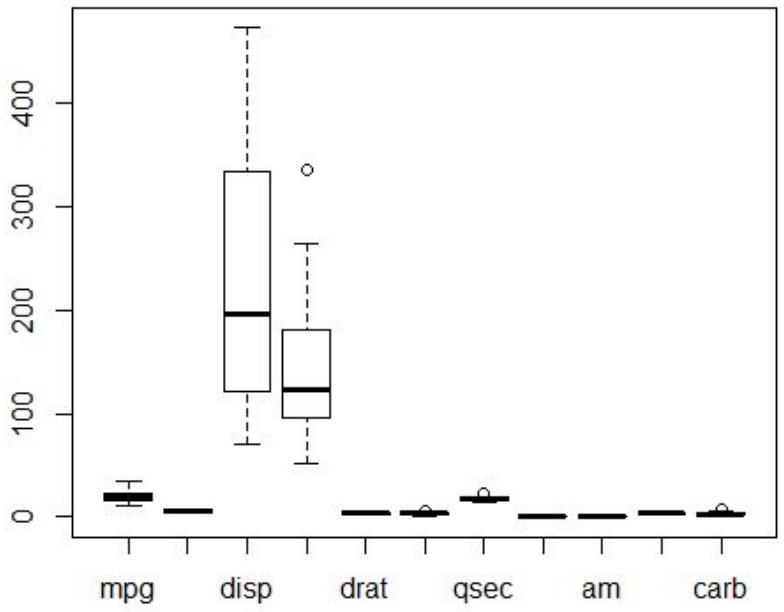
Checking normality and outliers:

The QQ-plot for our dataset is shown below.



From this QQ-plot we note that no values are incredibly far removed from our normal line either above or below, with the majority of values being clustered around the middle of our plot. These traits indicate that our data could be normally distributed. Because of this assumption of normality, it was decided to proceed without any log transformation on our dataset and to see if removing variables was necessary to result in an acceptable final model.

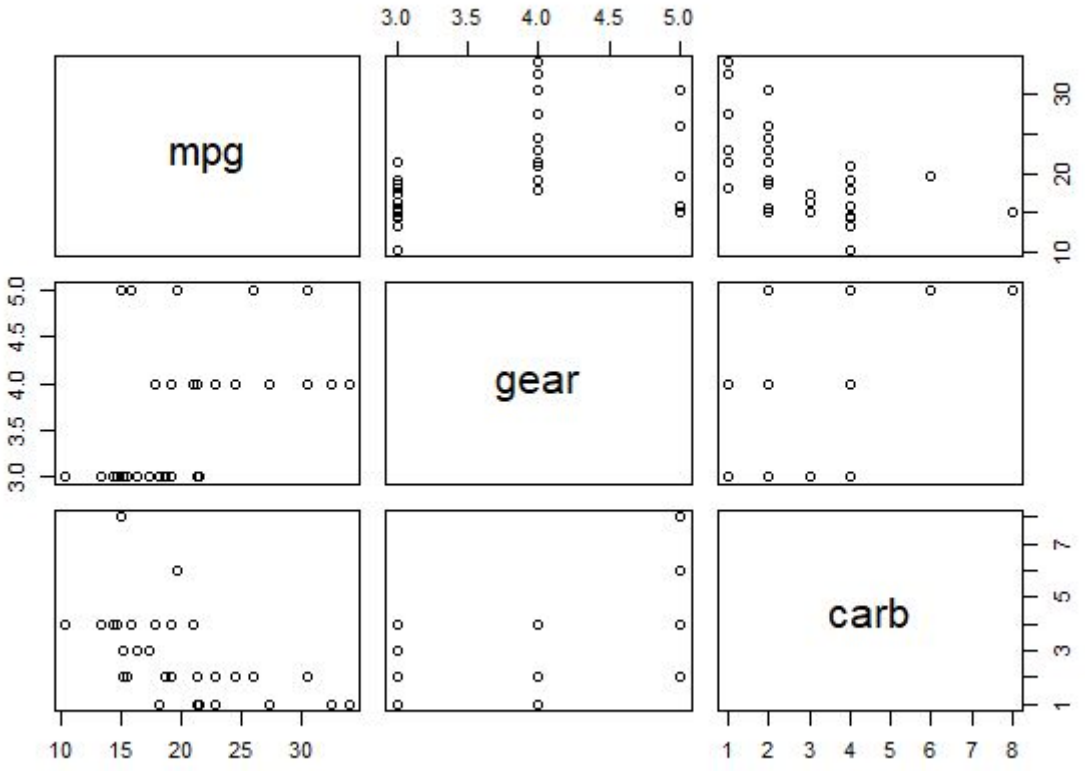
A boxplot of our dataset was used to check for outliers and is shown below.



With the exception of one value in hp, our dataset does not appear to comprise any notable outliers. It was thus decided to leave our dataset as is and move on to the model development process, noting that hp would likely have to be removed due to this aforementioned outlier.

Model development section:

In order to select the variables that will be used in the final model all the variables were examined for both correlation to miles per gallon and for multicollinearity. We decided to use backward elimination to create our model.



There was an assumption that gear and carb had a correlation so they were plotted but they did not seem correlated. At the same time, gear also does not seem correlated to mpg so we checked its correlation and it had one of the weakest of any predictor variables at around 0.43 so we decided to test removing it.

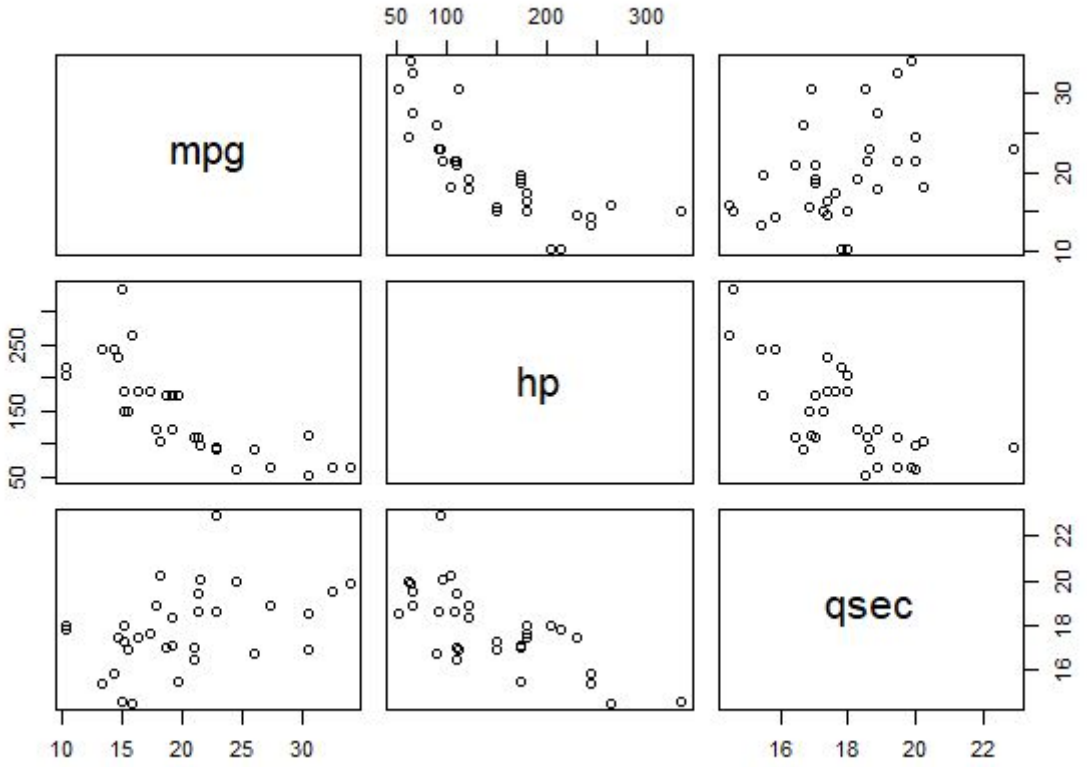
Before removing gear:

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared: 0.869,   Adjusted R-squared: 0.8066
F-statistic: 13.93 on 10 and 21 DF, p-value: 3.793e-07

After removing gear:

Residual standard error: 2.601 on 22 degrees of freedom
Multiple R-squared: 0.8678, Adjusted R-squared: 0.8137
F-statistic: 16.05 on 9 and 22 DF, p-value: 9.885e-08

Removing the gear variable increased our Adjusted R-squared and our F-statistic so we decided to keep it removed from our model.



Next hp was plotted vs qsec because there was an assumption they were correlated (the horsepower of a car should be related to how fast it can travel ¼ mile). From the graph it looked like they are correlated so it was decided to test removing either of the variables.

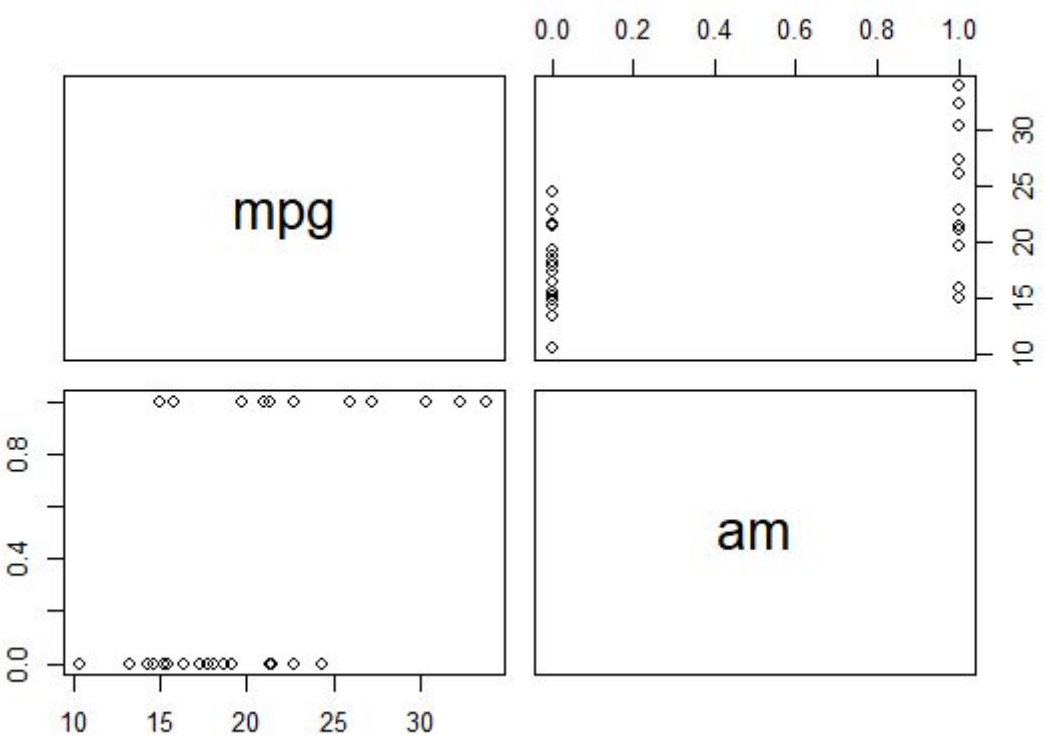
Summary with HP removed:

Residual standard error: 2.598 on 23 degrees of freedom
Multiple R-squared: 0.8622, Adjusted R-squared: 0.8142
F-statistic: 17.98 on 8 and 23 DF, p-value: 3.439e-08

Summary with qsec removed:

Residual standard error: 2.614 on 23 degrees of freedom
Multiple R-squared: 0.8604, Adjusted R-squared: 0.8118
F-statistic: 17.72 on 8 and 23 DF, p-value: 3.965e-08

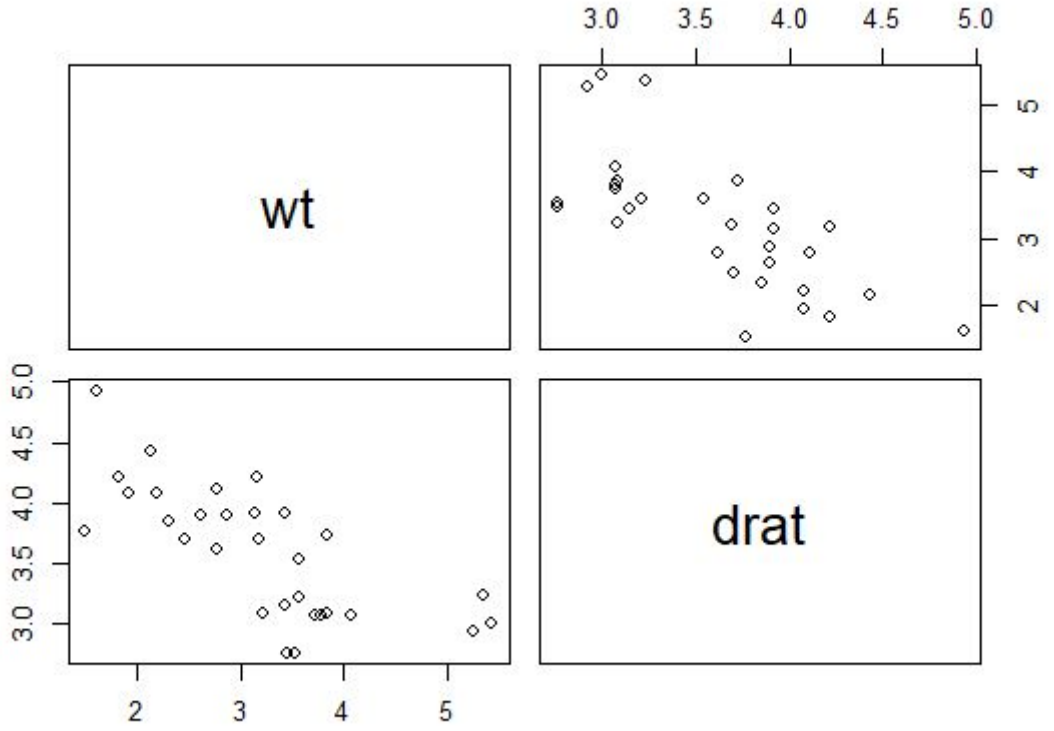
It appeared removing hp had a better effect on the adjusted R-squared and F-statistic marginally so we decided to remove the hp variable for now. We decided we should test which variable would work better with our final model at the end of variable selection.



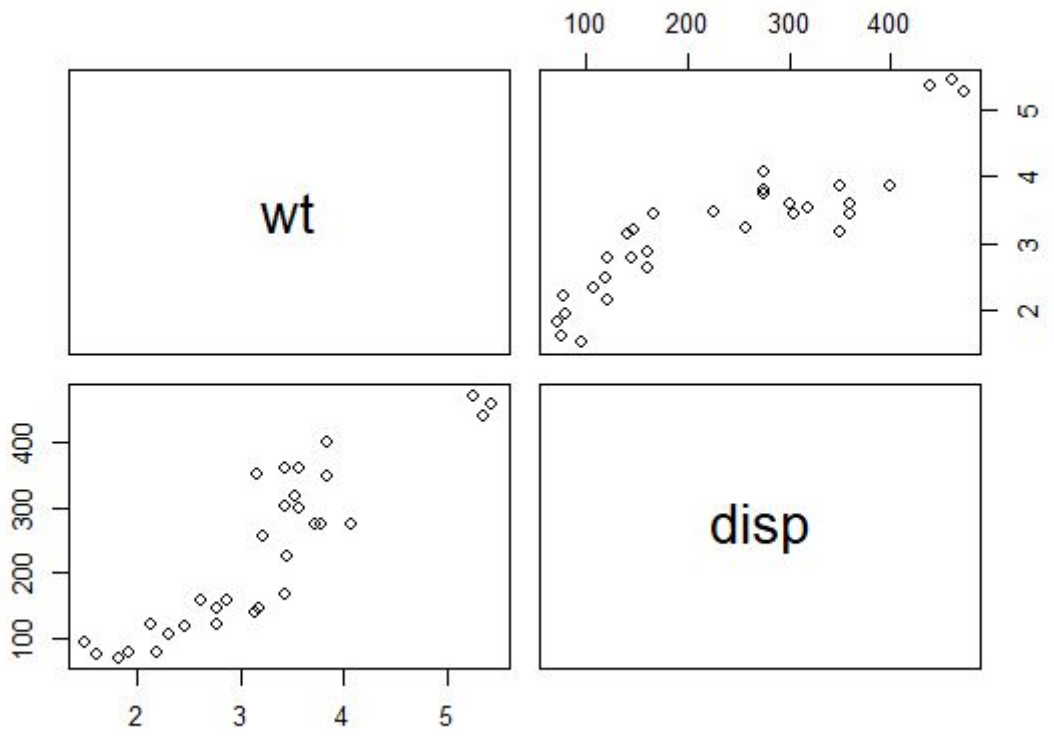
We decided to check the relationship between am and mpg and there seems to be a tenuous correlation between mpg and am. However, we decided not to remove it as it seemed to have one of the lowest p-values of any of our predictor variables in our model.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	13.824216	16.659099	0.830	0.415	
cyl	-0.425169	0.945668	-0.450	0.657	
disp	0.004878	0.014735	0.331	0.744	
qsec	0.875807	0.709154	1.235	0.229	
drat	0.966241	1.593247	0.606	0.550	
wt	-3.401678	1.762300	-1.930	0.066	.
vs	-0.197208	1.979886	-0.100	0.922	
am	2.655035	1.908000	1.392	0.177	
carb	-0.491998	0.573612	-0.858	0.400	



We then saw from plotting some of our remaining variables that wt seemed to be correlated with drat and disp.



It was then decided to see what the change in model would be like if we removed either wt or disp and drat since they are both correlated with wt.

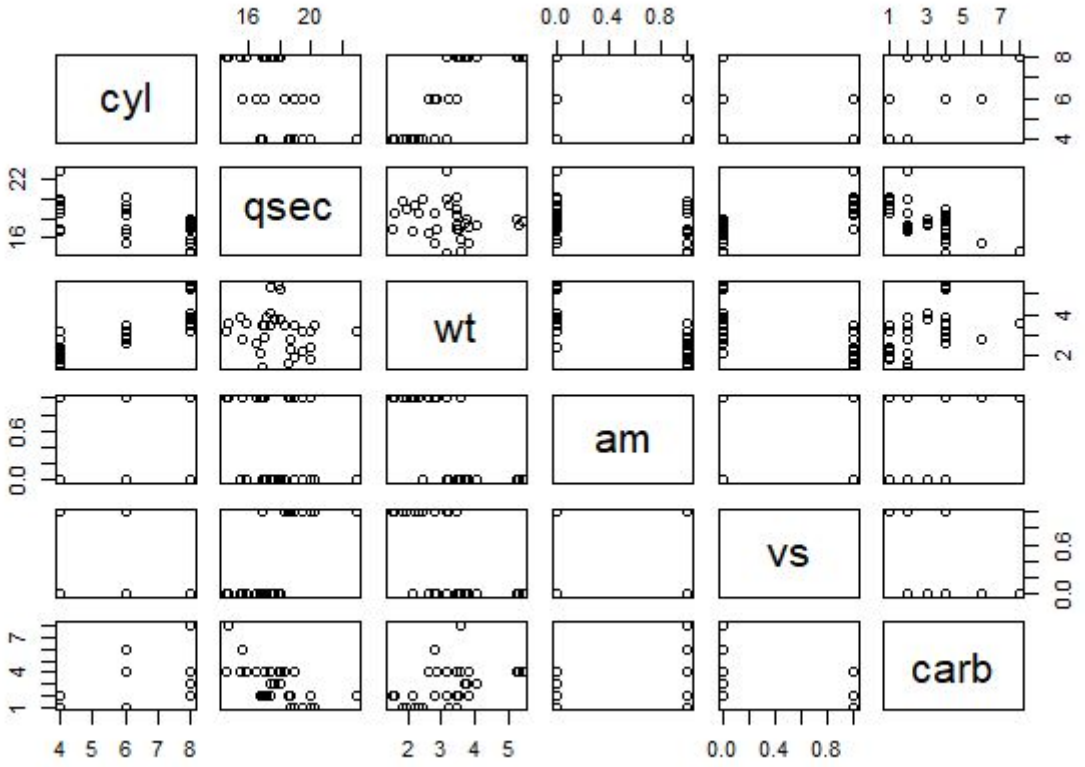
Without wt:

Residual standard error: 2.741 on 24 degrees of freedom
Multiple R-squared: 0.8398, Adjusted R-squared: 0.7931
F-statistic: 17.98 on 7 and 24 DF, p-value: 4.018e-08

Without drat and disp:

Residual standard error: 2.52 on 25 degrees of freedom
Multiple R-squared: 0.8591, Adjusted R-squared: 0.8252
F-statistic: 25.4 on 6 and 25 DF, p-value: 1.708e-09

The model has a much higher adjusted R-squared and F-statistic without disp and drat so we decided to keep them removed.



It was decided to plot the leftover variables and noticed cyl seems to be correlated with qsec and wt so we decided to see what the model would look like without it.

Residual standard error: 2.489 on 26 degrees of freedom
Multiple R-squared: 0.8569, Adjusted R-squared: 0.8294
F-statistic: 31.15 on 5 and 26 DF, p-value: 3.431e-10

The model had a significant increase in the F-statistic and the adjusted R-squared slightly increased so we removed cyl.

There didn't seem to be any variables left that were related to each other so we decided to test removing others that seemed to have a weak relation to mpg.

Model without vs:

Residual standard error: 2.444 on 27 degrees of freedom
Multiple R-squared: 0.8568, Adjusted R-squared: 0.8356
F-statistic: 40.39 on 4 and 27 DF, p-value: 5.064e-11

Model without vs and carb:

Residual standard error: 2.459 on 28 degrees of freedom
Multiple R-squared: 0.8497, Adjusted R-squared: 0.8336
F-statistic: 52.75 on 3 and 28 DF, p-value: 1.21e-11

We decided to remove vs and carb from our model after witnessing their removal significantly increasing our F-statistic and lowering the p-value and increasing our adjusted R-squared.

Finally we compared models that had qsec vs the hp variable to see which variable had a more significant effect on mpg.

Model without hp:

Residual standard error: 2.459 on 28 degrees of freedom
Multiple R-squared: 0.8497, Adjusted R-squared: 0.8336
F-statistic: 52.75 on 3 and 28 DF, p-value: 1.21e-11

Model without qsec:

Residual standard error: 2.538 on 28 degrees of freedom
Multiple R-squared: 0.8399, Adjusted R-squared: 0.8227
F-statistic: 48.96 on 3 and 28 DF, p-value: 2.908e-11

Since the model with qsec had a significantly higher F-statistic and adjusted R-squared we decided to keep it in our final model and discard hp.

Our final model looks like this:

Call:
lm(formula = mpg ~ qsec + wt + am, data = mtcars)

Residuals:
   Min   1Q Median   3Q   Max
-3.4811 -1.5555 -0.7257 1.4110 4.6610

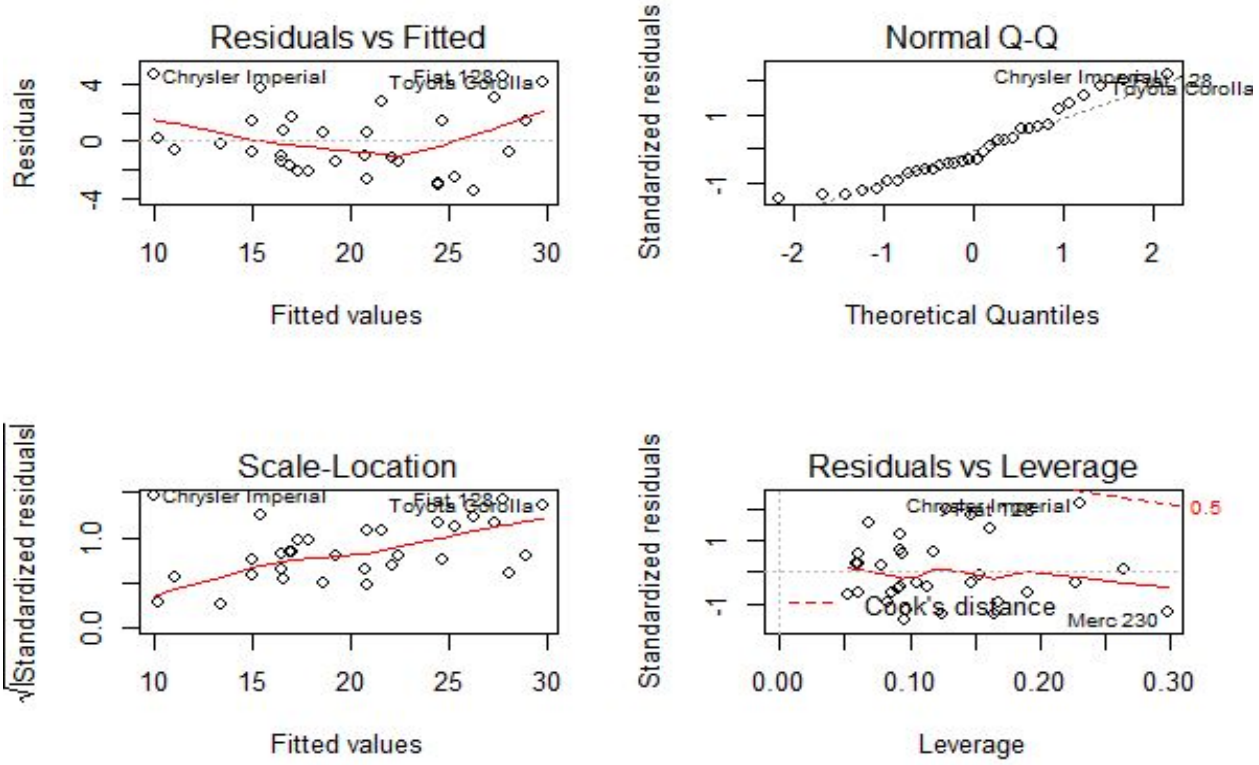
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.6178   6.9596   1.382 0.177915
qsec   1.2259   0.2887   4.247 0.000216 ***
wt   -3.9165   0.7112 -5.507 6.95e-06 ***
am   2.9358   1.4109   2.081 0.046716 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.459 on 28 degrees of freedom
Multiple R-squared: 0.8497, Adjusted R-squared: 0.8336
F-statistic: 52.75 on 3 and 28 DF, p-value: 1.21e-11

Model Validation

In this section, we will validate our final model with a diagnostic. From our test, since lack of multicollinearity shows that our predictor variables are free from any interference among themselves, we decided that our final model is a good choice to determine miles per gallon when affected by the variables qsec, wt, and am. Since we want developers to see which variables they need to change in order to minimize negative environmental impact and money to build, our final model serves as the perfect model for this purpose. To validate further, consider the following plots.



Residual vs Fitted plot

Here we can notice a slight parabola but not curved significantly enough to say it is non-linear. We see equally spread residuals around the horizontal line without distinct patterns concluding that we have a linear relationship and thus our validation holds.

Scale-location plot

Here we confirm that the assumption of equal variance(or homoscedasticity) holds as we see a horizontal line with equally distributed spread points. (i.e we don't see the graph spreading wider along the fitted values-axis). We also notice that the slope which indicates the spread of residuals for those predictor values is not steep and even less so in the middle.

Normal-QQ Plot

Here we can confirm normal distribution as we see points lined well on the straight dashed line.

Residual vs Leverage

Here we are looking for influential cases(i.e subjects) if any. Note that Cook's distance is labeled with red dashed lines. We can quickly observe that we can barely see Cook's distance lines on the corner of the graph, meaning that all cases are well inside of Cook's distance lines. This is good for our model because there will be no outliers that will impact significantly on our analysis. Meaning that we can proceed to analyze without having to examine noisy cases.

In conclusion, our model is a great representation on determining miles per gallon and we conclude that it won't be necessary to go back to consider adding quadratic terms or log transformation as it might actually remove linearity and normality.

Hypothesis tests for linear relationship between mpg and qsec, wt, and am:

In this section we are testing our final model to determine whether we can conclusively state if a linear relationship exists between our response variable, mpg, and our explanatory variables qsec, wt, and am. Using the summary of our model as given above, our model is shown to be

$$\hat{Y} = 9.6178 + 1.2259\,\beta 1 - 3.9165\,\beta 2 + 2.9358\,\beta 3 ,$$

and our hypotheses are as follows:

$$H_0 : \beta 1 = \beta 2 = \beta 3 = 0$$
$$H_A : \exists\ \beta \text{ such that at least one of } \beta 1, \beta 2, \beta 3 \neq 0$$

Where H_0 states that no linear relationship exists between our response and our explanatory variables, and H_A states that at least one of our explanatory variables is significant to our final model.

From the ANOVA table of our model:

Analysis of Variance Table

Response: mpg

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
qsec	1	197.39	197.39	32.6488	3.963e-06	***

```
wt         1 733.19  733.19 121.2704 1.099e-11 ***
am         1  26.18   26.18   4.3298   0.04672 *
Residuals 28 169.29    6.05
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Therefore, our test statistic $F* = MSReg/MSRes$ is 52.75. These values can again be verified through the summary of our completed model. At significance level 0.05, we show that $F*$ = 52.75 > $F_{0.95, 3, 28}$ = 2.9467, showing that we reject our null hypothesis. This rejection of the null hypothesis indicates that at least one of the variables qsec, wt, or am is significant to our final model.

We can interpret these results further with the given values in our model call and its ANOVA table. Of immediate note is the relatively large p-value for am in comparison to qsec and wt. The p-value of 0.04672 is large when the other two variables are taken into account, and at lower alpha levels such as 0.01, am would be a potential reason as to why we may fail to reject our null hypothesis. This observation was preceded by our noticing earlier in this report that am has at best a tenuous relationship with mpg. While the relationship between am and mpg is not as strong as mpg with qsec and wt, there is still evidence of a relationship. For this reason it was decided that am should remain in our final model, and our final hypothesis test proves that at 5% significance, some relationship does exist between mpg and our three variables.

Through our tests, we have been led to believe that our final model is a good representation of the relationship between the response mpg, and the explanatory variables qsec, wt, and am. Our data is proven to be approximately normal with few outliers, our model validated through various means, and a hypothesis test has conclusively stated that a relationship does exist between mpg and the variables in our chosen final model.

Conclusion and discussion:

Our dataset is all about the fuel consumption and the other 10 variables of automobile design and performance for 32 automobiles samples that were made during 1973–74. Out of these 10 variables, the project studies statistical significance of each variable’s impact on fuel efficiency, proximate by miles per gallon, using a multiple linear regression, after a multicollinearity check. The multicollinearity check helped to choose among highly correlated variables, the existence of which might weaken the statistical power of the multiple linear regression model, thus raising the precision of the estimated coefficients of those variables left. For example, wt, drat and disp have a high correlation, so two of them should be removed to avoid multicollinearity. Since the model looks better without drat and disp, these two variables are taken out and wt is left.

The final model candidate determines to use three variables qsec ¼ mile time), wt (weight), am (0 for automatic or 1 for manual) to predict fuel efficiency. According to the model coefficient estimate, a car with higher ¼ mile time and lower weight would have better fuel efficiency given other things equal. To be more specific, one unit increase in ¼ mile time or one unit decrease in weight would increase miles per gallon by 1.22591 or 3.91652 respectively. Also, cars with manual transmission systems are 2.93583 miles per gallon more efficient in fuel consumption than their automatic counterparts, given other things equal. This might shed a light on automobile developers to design fuel efficient cars, which is to aim for higher ¼ mile time, lower weight, and manual transmission system.

The study has a few limitations. The number of observations, 32, is a relatively small sample size. If sample size was too small, the model would not be able to identify significant relationships within the dataset. Basing the study in a larger sample size could have generated more accurate results. Also, the samples of the dataset are stale because they're related to models manufactured almost 50 years ago, which is definitely not a representative sample that can provide proper guidance to today's automobile designers. Our final model, however, can provide a reference point to manufacturers willing to obtain the best ratio of miles per gallon in a vehicle. By being knowledgeable of which aspects of a car optimize miles per gallon, car manufacturers can design more climate friendly vehicles, which was the original research goal of this report.

Appendix

Exploratory Section:

```
> multi.fit = lm(mpg~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb,
data=mtcars)
> plot(multi.fit, 2)
> boxplot(mtcars)
```

Model Selection:

```
> pairs(~ mpg + gear + carb, data = mtcars)
> cor(mtcars, use="complete.obs", method="kendall")
> multi.fit = lm(mpg~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb,
data=mtcars)
> summary(multi.fit)
> multi.fit_nogear = lm(mpg~ cyl + disp + hp + drat + wt + qsec + vs + am + carb,
data=mtcars)
> summary(multi.fit_nogear)
> pairs(~ mpg + hp + qsec, data = mtcars)
> multi.fit = lm(mpg~ cyl + disp + qsec + drat + wt + vs + am + carb, data=mtcars)
> summary(multi.fit)
> multi.fit = lm(mpg~ cyl + disp + hp + drat + wt + vs + am + carb, data=mtcars)
> summary(multi.fit)
> plot(~ mpg, am, data = mtcars)
> multi.fit = lm(mpg~ cyl + disp + qsec + drat + wt + vs + am + carb, data=mtcars)
> summary(multi.fit)
> pairs(~ wt + drat, data = mtcars)
> pairs(~ wt + disp, data = mtcars)
> multi.fit = lm(mpg~ cyl + disp + qsec + drat + vs + am + carb, data=mtcars)
> summary(multi.fit)
> multi.fit = lm(mpg~ cyl + qsec + wt + vs + am + carb, data=mtcars)
> summary(multi.fit)
> pairs(~cyl + qsec + wt + am + vs + carb, data = mtcars)
> multi.fit = lm(mpg~ qsec + wt + vs + am + carb, data=mtcars)
> summary(multi.fit)
> multi.fit = lm(mpg~ qsec + wt + am + carb, data=mtcars)
> summary(multi.fit)
> multi.fit = lm(mpg~ qsec + wt + carb, data=mtcars)
> summary(multi.fit)
> multi.fit = lm(mpg~ hp + wt + am, data=mtcars)
> summary(multi.fit)
> anova(multi.fit)
```

Model Validation:

```
> multi.fit = lm(mpg~ qsec + wt + am, data=mtcars)
> par(mfrow = c(2, 2))
> plot(multi.fit)
```