Introduction

The dataset has 82 cars and has 5 variables were measured for each of these cars in the dataset. The variables in this dataset are Cubic feet for cab space (VP), Engine Horse Power(HP), SP, Weight(WT) ,and Miles Per gallon(MPG). Based on some preliminary analysis I believe that the dataset has a wide variety of different cars involved. The dataset is not limited to just commercial cars but also high end sports car which can explain many of the outliers that will be later pointed out in the preliminary analysis.

The data describes the effect of the explanatory variables to the response variable.

- VOL (cubic feet of cab space)
- HP engine horsepower (Horse Power)
- SP top speed (MPH)
- WT vehicle weight (Pounds lb)
- MPG (average miles per gallon Miles/gallon)

I am going to investigate the relationship between the explanatory variables HP, SP, and WT to the response variable MPG. I will build a multiple linear regression model to examine the relationship between the explanatory variables (HP, SP,WT) and the response variable(MPG). I will see if HP, weight, and SP are significant in predicting MPG.

The explanatory variables are

- HP engine horsepower (Horse Power)
- SP top speed (MPH)
- WT vehicle weight (Pounds lb)

The response variable is

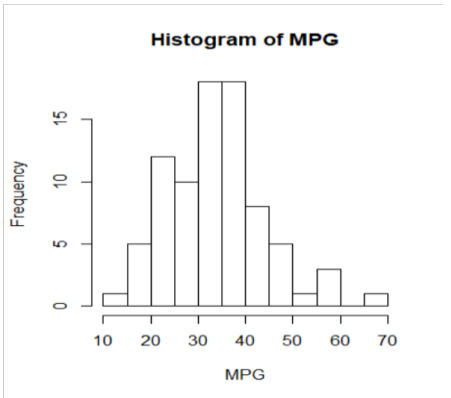
- MPG (average miles per gallon Miles/gallon)

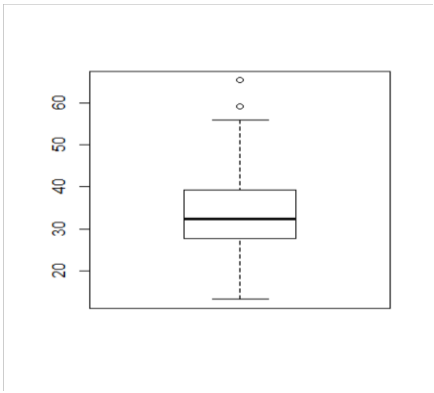
Preliminary Analysis

I will further analyze using multiple linear regression. First, I need a 5-number summary of the dataset variables to better understand the data.

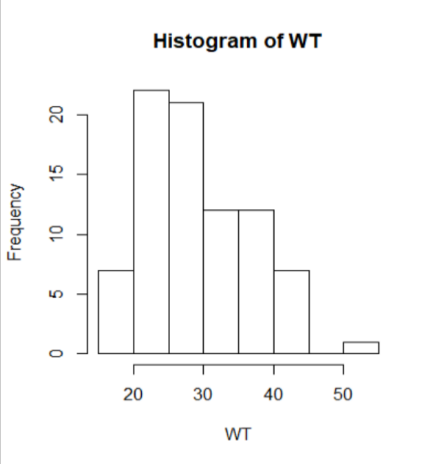
```
> summary(data1)
      VOL              HP              MPG              SP       
 Min.   : 50.0   Min.   : 49.0   Min.   :13.20   Min.   : 90.0  
 1st Qu.: 89.5   1st Qu.: 84.0   1st Qu.:27.77   1st Qu.:105.0  
 Median :101.0   Median : 99.0   Median :32.45   Median :109.0  
 Mean   : 98.8   Mean   :117.1   Mean   :33.78   Mean   :112.4  
 3rd Qu.:113.0   3rd Qu.:140.0   3rd Qu.:39.30   3rd Qu.:114.8  
 Max.   :160.0   Max.   :322.0   Max.   :65.40   Max.   :165.0  
       WT       
 Min.   :17.50  
 1st Qu.:25.00  
 Median :30.00  
 Mean   :30.91  
 3rd Qu.:35.00  
 Max.   :55.00  
> sapply(data1,sd)
      VOL        HP       MPG        SP        WT 
22.166285 56.840857 10.004605 14.037825  8.141422 
```

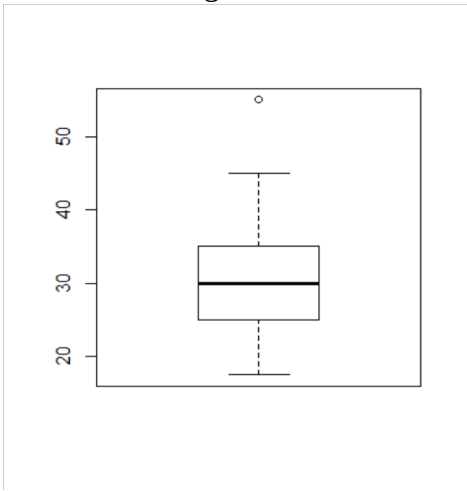
I will also graph the data to visually see the box plots and histograms for the five-number summary to see the distribution of the variables. Each boxplot is next to the histogram related to its variable.



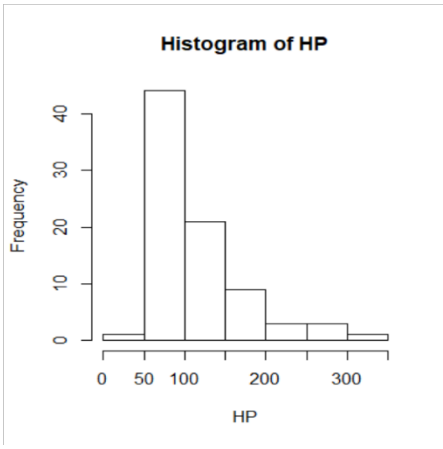


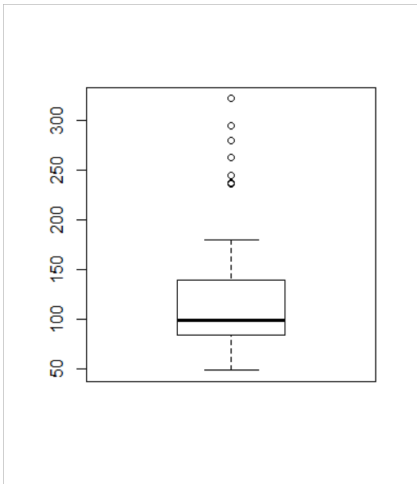
Miles Per Gallon Seems like a normal distribution according to the histogram. The boxplot also shows that there are little outliers while still having a small middle 25-75 percentile



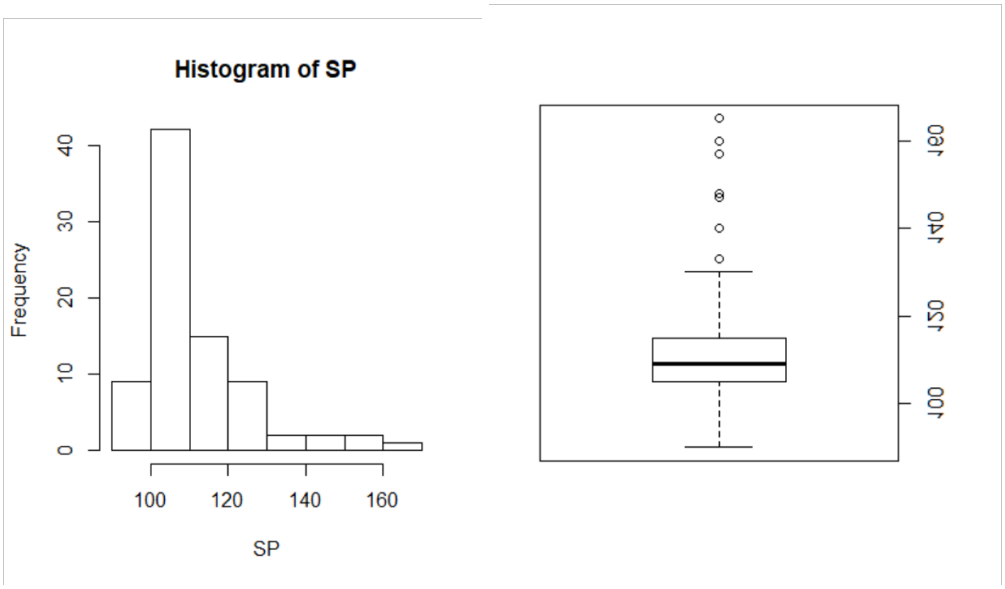


The Historam of WT is skewed to right which demonstrates that many of the cars are on the lower end of theweight spectrum. The boxplot also shows little outliers with only 1 car being over 50.





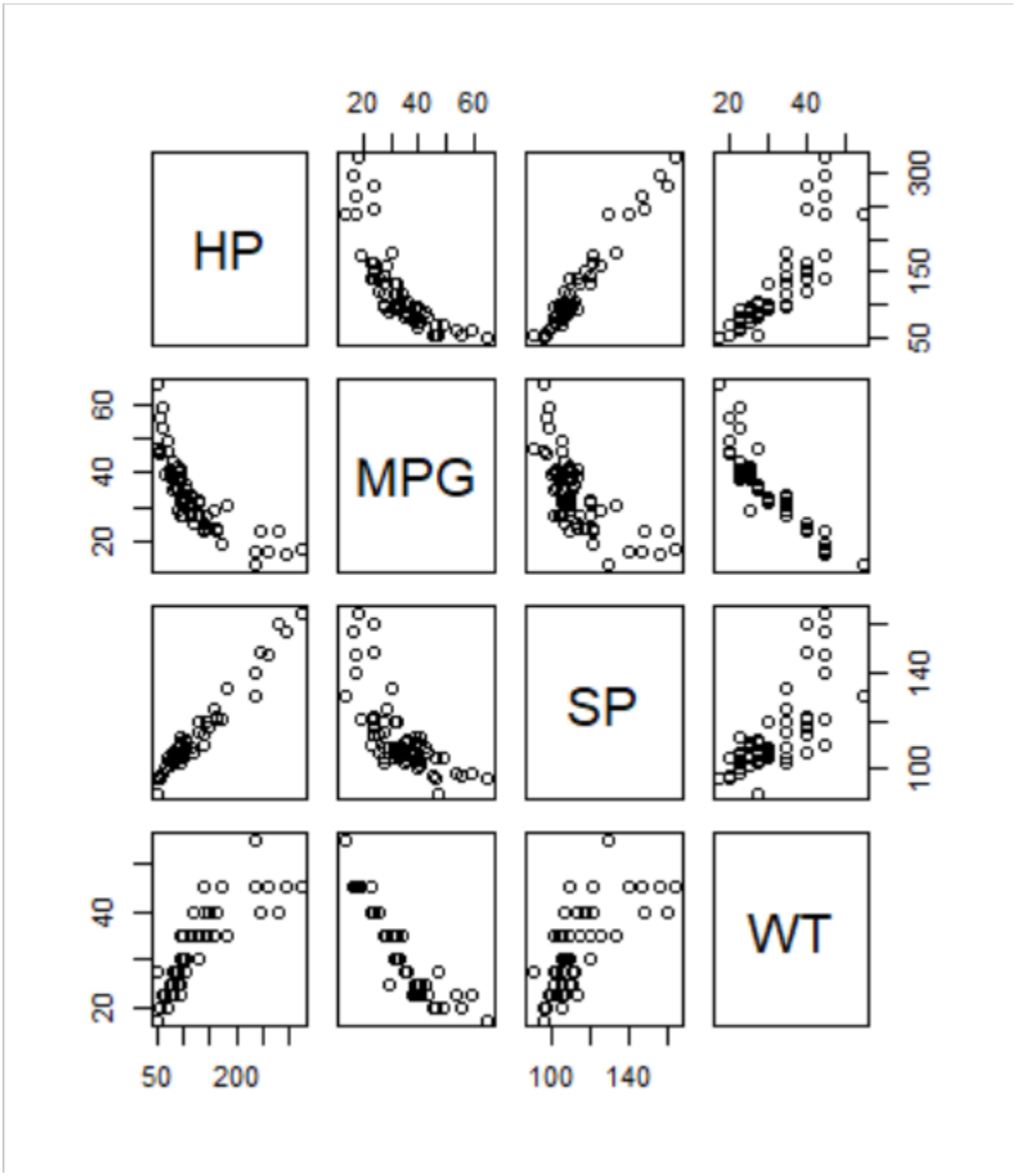
The histogram of HP shows that the data is skewed to the right making us understand how horse power and weight are related given skewness and basic understanding between the relationship of weight and horse power. There are many outliers with horsepower which might be a byproduct of sports cars being the outliers in the dataset.



Similar to horse power top speed follows a very similar distrubution which helps us understand that in the dataset there are cars that are for spprts but the vats majority of cars have a top speed on the lower end of spectrum. The histogram is skewed to the right given that most commercial cars are slower.

I will run a correlation test to see the linear relationship between all the variables.

```
> pairs(data1)
> cor(data1)
           HP        MPG         SP         WT
HP   1.0000000 -0.7898564  0.9665452  0.8322202
MPG -0.7898564  1.0000000 -0.6884462 -0.9050849
SP   0.9665452 -0.6884462  1.0000000  0.6785339
WT   0.8322202 -0.9050849  0.6785339  1.0000000
```



Based on the graph and R output I believe that there is a negative correlation between all 3 explanatory variables to the MPG. Also based on the R output I believe that al 3 explanatory

variables have a positive correlation with each other. There is a strong correlation between explanatory variables and each other while all 3 variables ultimately have negative correlation to the response variable.

Multiple Linear Regression Results

```
> carmile.lm=lm(MPG~HP+WT+SP,data=data1)
> summary(carmile.lm)

Call:
lm(formula = MPG ~ HP + WT + SP, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1633 -2.8387  0.2464  1.7889 12.5566

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 194.12962   23.32213   8.324 2.22e-12 ***
HP            0.40518    0.07891   5.135 2.03e-06 ***
WT           -1.92210    0.19238  -9.991 1.31e-15 ***
SP           -1.32000    0.24118  -5.473 5.19e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.64 on 78 degrees of freedom
Multiple R-squared:  0.8725,   Adjusted R-squared:  0.8676
F-statistic: 177.9 on 3 and 78 DF,  p-value: < 2.2e-16
```

Linear Model Equation

194.12962+0.40518HP-1.92210WT-1.32000SP

Model Coefficient analysis:

HP: With all thing being equal each 1 increase in one unit of HP the MPG increases by 0.40518

WT: With all thing being each 1 increase in WT the MPG decreases by -1.922210

SP: With all thing being each 1 increase in SP the MPG decreases by -1.320.

Y intercept: With no HP, WT,OR SP the starting point for MPG consumption starts at 194.12962. However this does not make sense since no car starts with 0 HP,WT, and SP.

R2 Interpretation

The R^2 for the Multiple regression model is 0.8676. This means there is 86.67% variation in fuel economy (MPG) can be explained by the SP, HP, and WT in this data set.

Entire Model

Hypothesis Testing

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$

The explanatory variables are the same and are not significant to the response variable. All the coefficients of the explanatory variables are equal to zero.

H_a: At least one of the β_j is not 0

At least one of the explanatory variables is significant to the response variable (MPG).

F statistic: 177.9

P value: 2.2e-16

Conclusion: Since P<0.05 then we reject null hypothesis which means that at least 1 explanatory variable is significant in determining the miles per gallon of a car. We can draw on this conclusion with 95% confidence.

HP

Hypothesis Testing

H_0: $\beta_1 = 0$

The explanatory variable not significant in determining the response variable.

H_a: β_1 is not 0

The explanatory variable is significant in determining the response variable.

T value: 5.135

P value: 2.03e-06

Conclusion: Since P<0.05 then we reject null hypothesis which means that this explanatory variable (HP) is significant in determining the miles per gallon of a car with the other explanatory variables being involved in the model. We can draw on this conclusion with 95% confidence.

WT

Hypothesis Testing

H_0: $\beta_2 = 0$

The explanatory variable not significant in determining the response variable.

H_a: β_2 is not 0

The explanatory variable is significant to the response variable.

F statistic: -9.991

P value: 1.31e-15

Conclusion: Since P<0.05 then we reject null hypothesis which means that this explanatory variable (WT) is significant in determining the miles per gallon of a car with the other explanatory variables being involved in the model. We can draw on this conclusion with 95% confidence.

.

SP

Hypothesis Testing

H_0: $\beta_3 = 0$

The explanatory variable not significant in determining the response variable.

H_a: β_3 is not 0

The explanatory variable is significant to the response variable.

F statistic: -5.473

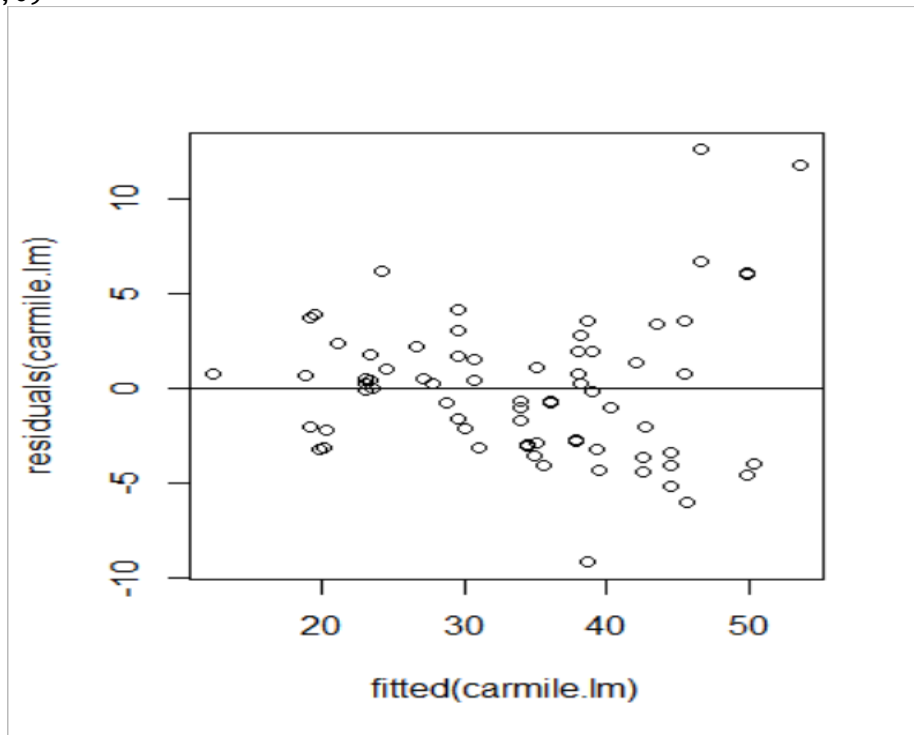
P value: 5.19e-07

Conclusion: Since P<0.05 then we reject null hypothesis which means that this explanatory variable (SP) is significant in determining the miles per gallon of a car with the other explanatory variables being involved in the model. We can draw on this conclusion with 95% confidence.

Checking Model Assumptions

Entire Model

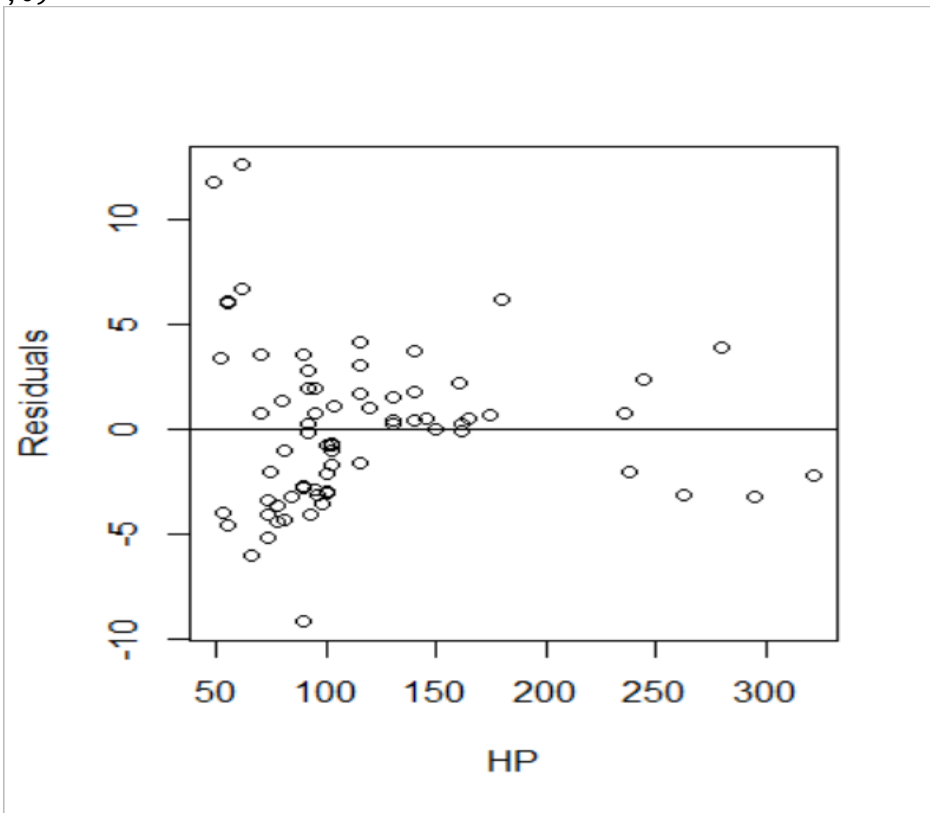
```
> plot(fitted(carmile.lm),residuals(carmile.lm))
> abline(0,0)
```



Analysis: There is no curved trend which justifies that the multiple linear regression is valid and does not have problems. As x increases the variance increases.

HP

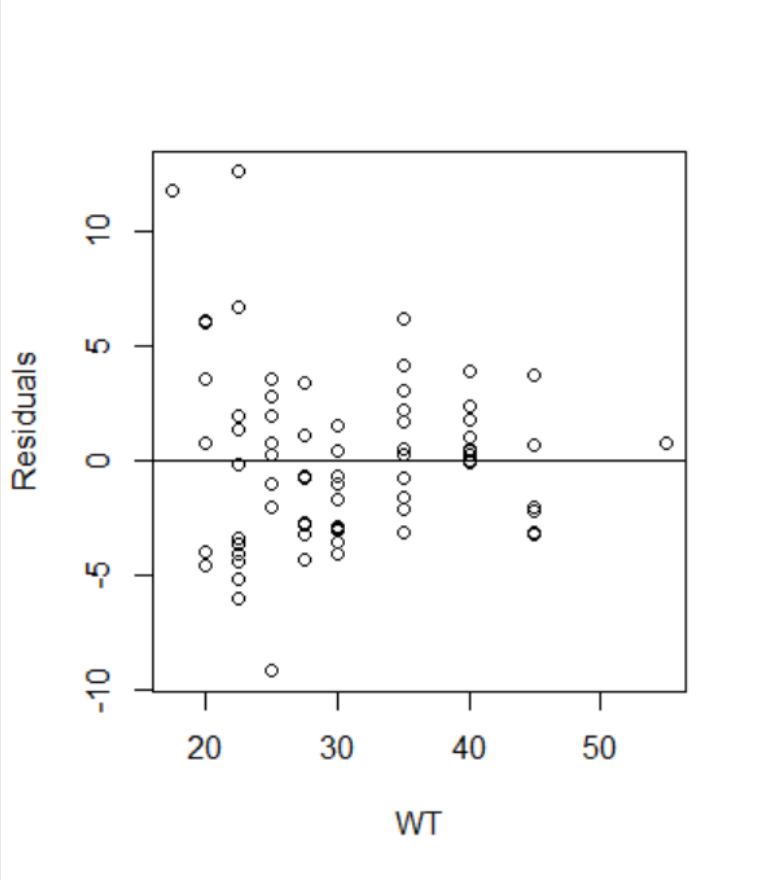
```
> plot(HP,residuals(carmile.lm),xlab="HP",ylab="Residuals")
> abline(0,0)
```



Analysis: The variance decreases as HP increases, however when the horsepower increases beyond 150 the variance starts to increase a little.

WT

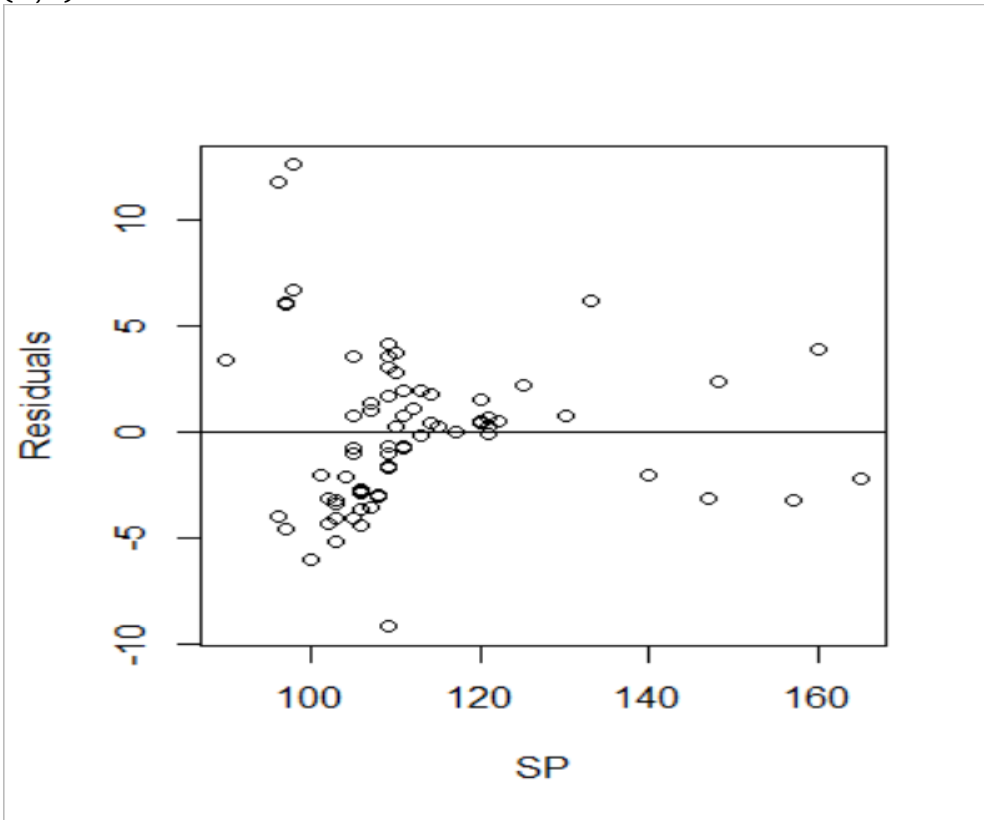
```
> plot(WT,residuals(carmile.lm),xlab="WT",ylab="Residuals")
> abline(0,0)
```



Analysis: The variance decreases as the weight increases especially on the lower ends of weight.

SP

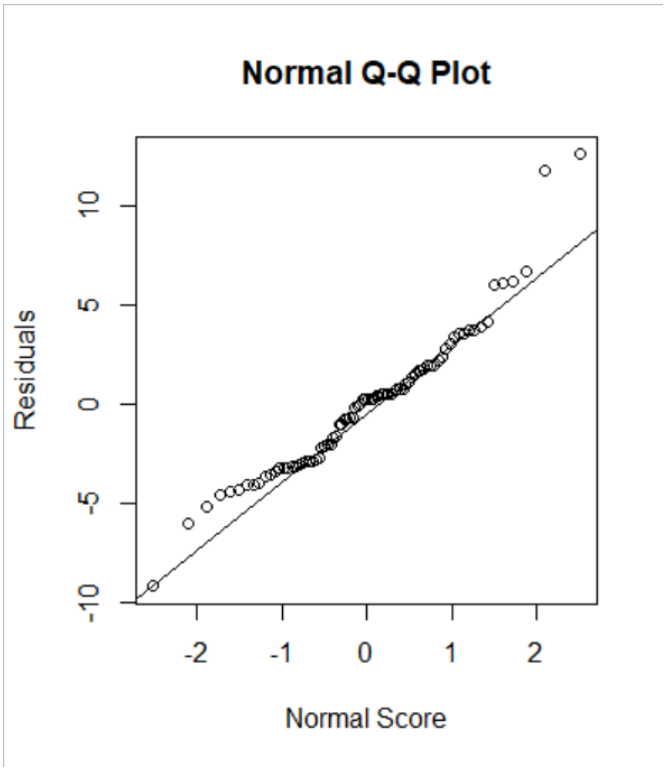
```
> plot(SP,residuals(carmile.lm),xlab="SP",ylab="Residuals")
> abline(0,0)
```



Analysis: Variance decrease as the top speed increases however on the higher ends of top speed the variance starts to increase a little. This also shows similarity to the Horsepower residual.

QQ plot

```
> qqnorm(residuals(carmile.lm),xlab="Normal Score", ylab="Residuals")
> qqline(residuals(carmile.lm))
```



Analysis: The scatterplot follows the lines well therefore we can conclude that the linear regression is appropriate.

Conclusion: While performing the complete regression analysis. I was able to answer my question about predicting the MPG used from the explanatory variables. The relationship between MPG and each variable in my model is negative correlation. The strongest negative linear relationship comes is for HP and WT (-0.7898564, -0.9050849).

Based on analysis in the paper, we can confirm that the variables HP, SP, WT are significant in predicting MPG. I chose to leave out volume as volume does is directly correlated with weight. I also believe based on background information volume is not that significant of a predictor of MPG usage. I obtained a model that concludes that the at least one of the three variables is significant in predicting the MPG. I have also done a hypothesis test for each variable in the model and I came up with the conclusion that each variable in the model is significant in predicting MPG in presence with all other variables in the model. I also tried to confirm that the model is appropriate. I ran the QQ plot and I believe that since the scatterplot fits the line then we can conclude that the linear model is appropriate thus means that the model as is appropriate and helps us predict miles per gallons.