Multiple Linear Regression
On
Car Data

The Issues:

To find a multiple linear regression of the given data auto which has 4 predictor and response variable we have to address the following questions:

1. Is at least one of the predictors useful in predicting the response?
2. Do all the predictors help to explain the response, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Findings:

After performing all the statistical operations on the data I come to the conclusion,That all the predictors(i.e displacement,weight,acceleration,horsepower) in the given dataset have significance in predicting the response(i.e mpg).Though,after calculation using correlation we can see that is horsepower is less significant but still after further assessing the data the fit of the model drops hence,we cannot eliminate accelration. As we have 4 predictor we have the chance of getting 16 subset(i.e subsets = 2^n,n=number of predictors) by using backward selection i have come to conclusion that the fit of the model is highest when all predictor are included and it sums up to 0.674 the score should usually be near to 0.9 which means that it is a moderate fit and choosing some other model might be recommended.When I passed a set of predictor values on data with 95% confidence I would got a accurate prediction of 45 which means that the accuracy is average. Hence, I would like to conclude that multiple linear regression on the given data has low accuracy and has a moderate fit.

Discussion

Appendix A: Method

The data was downloaded from the excel sheet form and was imported into R studio. There are 4 factors/predictors(i.e horsepower.displacement,acceleration and weight) which are present and one response variable i.e mpg were extracted and all the blank data was omitted from the raw data.

We apply all the descriptive statistics operations on the data to find all the basic summary of the data. After which we try to find the correlation among data and from that we look at the p values and then backward selection I have removed that which was a less significant variable/predictor.After which I have created a subset and all the required predictors we perform an F-test on the subset to see the significance of it and then to find the we use the r*2 method check the value according try to hit and trail with the predictors until you achieve the highest possible value in r*2 and at the create new predictor value and try to find the prediction accuracy.We can find the confidence of interval and remove all the residuals in the data.Then come to a conclusion that how well does the model find the data and also the accuracy of it.

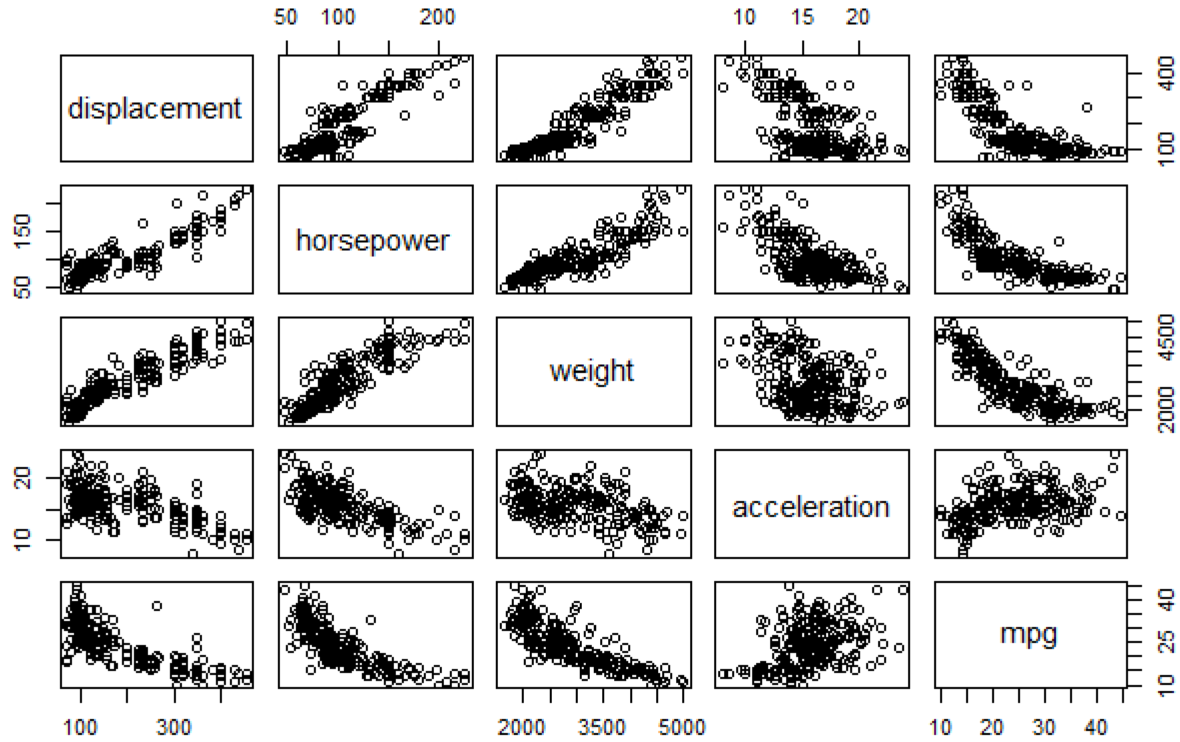
Appendix B: Results

From the dataset there are 405 data points containing 4 predictor factors and 1 response.

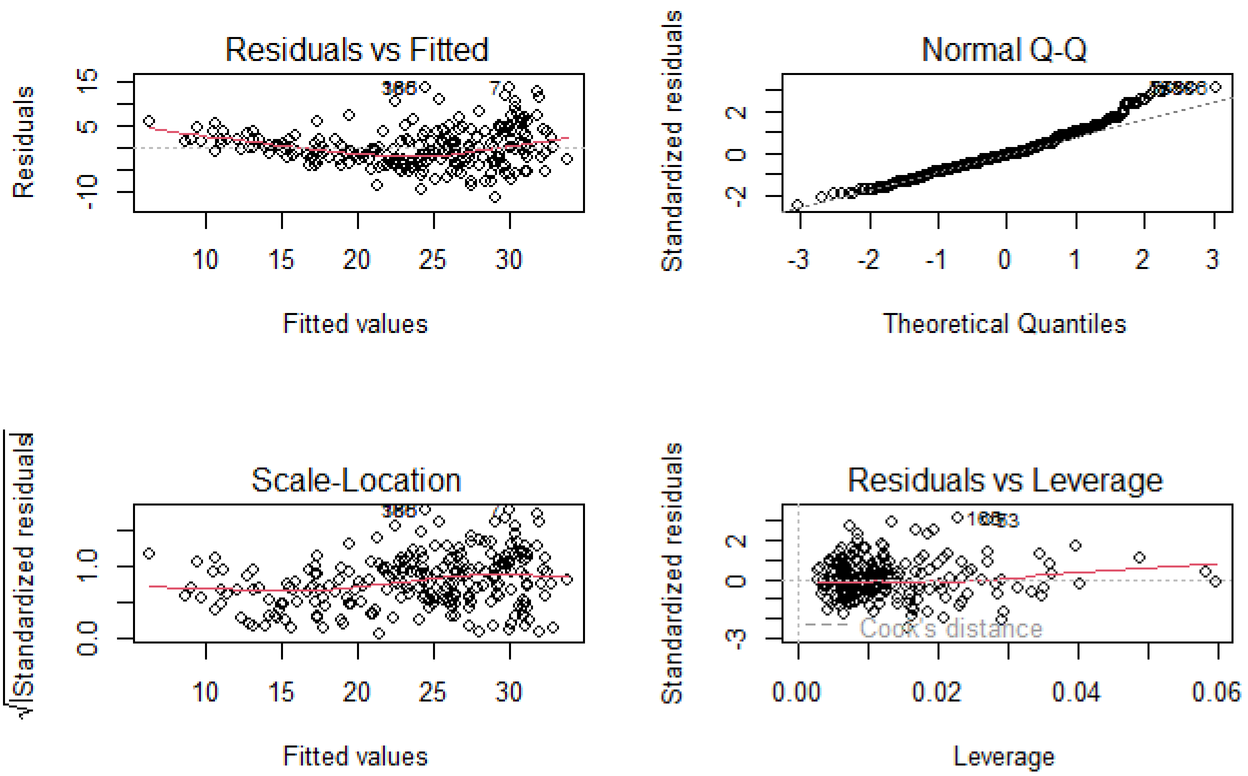
On applying the descriptive statistics on the both variable,I found the different stats

```
 displacement     horsepower        weight      acceleration        mpg       
 Min.   : 70.0   Min.   : 48.0   Min.   :1649   Min.   : 8.00   Min.   :10.00  
 1st Qu.: 98.0   1st Qu.: 75.0   1st Qu.:2215   1st Qu.:14.00   1st Qu.:18.00  
 Median :151.0   Median : 90.0   Median :2774   Median :15.50   Median :23.00  
 Mean   :190.4   Mean   :103.1   Mean   :2921   Mean   :15.59   Mean   :23.92  
#th 3rd Qu.:250.0   3rd Qu.:120.0   3rd Qu.:3520   3rd Qu.:17.30   3rd Qu.:30.00  
 Max.   :455.0   Max.   :225.0   Max.   :4997   Max.   :23.70   Max.   :44.60  
```

Calculate the coefficients of the data



Plotting linear regression between



Calculating summary of the model 1

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 45.9940079  2.5513318  18.027  < 2e-16 ***
disp         0.0020013  0.0070823   0.283  0.77764
hp          -0.0508593  0.0183578  -2.770  0.00586 **
acc          0.0006884  0.1324467   0.005  0.99586
weg         -0.0058949  0.0009346  -6.308 7.51e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By is anova we are calculate the F value

```
Response: mpg
           Df  Sum Sq Mean Sq F value    Pr(>F)
disp        1 14402.1 14402.1 745.659 < 2.2e-16 ***
hp          1   565.1   565.1  29.258 1.093e-07 ***
acc         1   267.8   267.8  13.866 0.0002245 ***
weg         1   768.5   768.5  39.786 7.512e-10 ***
Residuals 400  7725.8    19.3
---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 4.395 on 400 degrees of freedom
Multiple R-squared:  0.6744,    Adjusted R-squared:  0.671
```

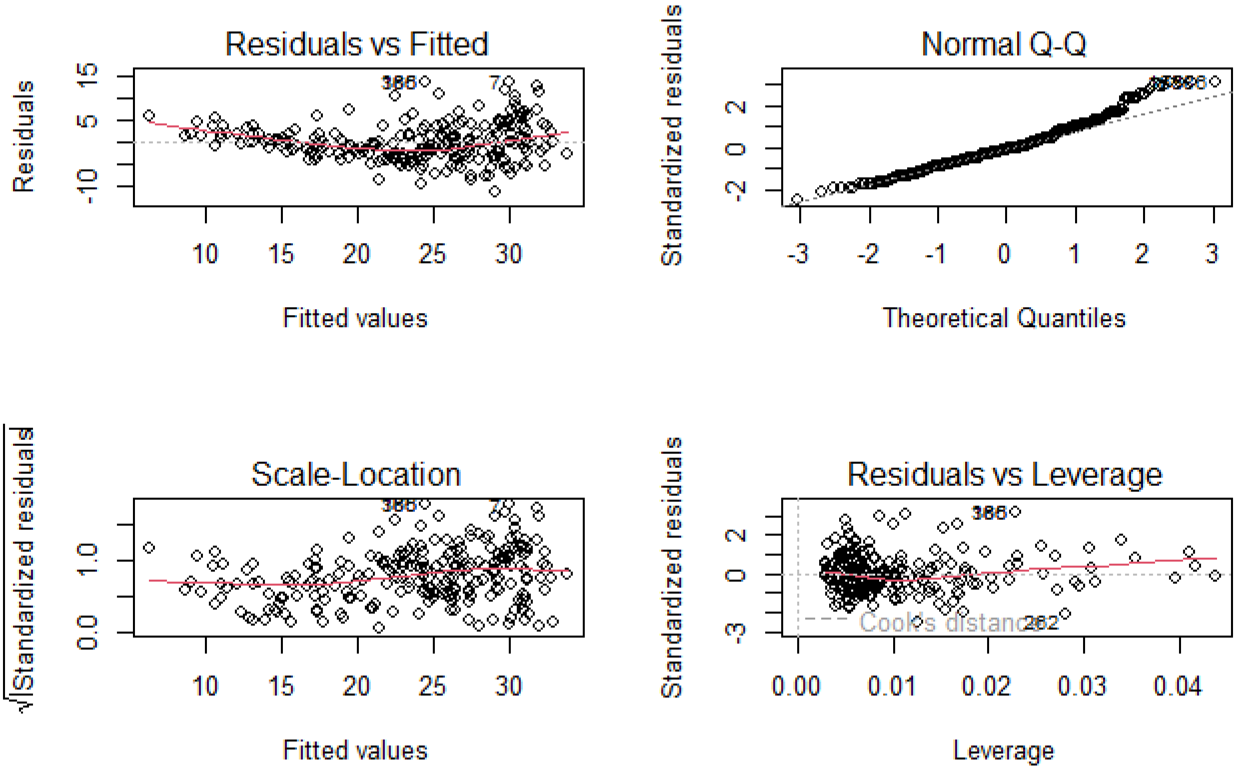
Checking the r*2 value

```
summary(model)$r.squared
[1] 0.6744181
```

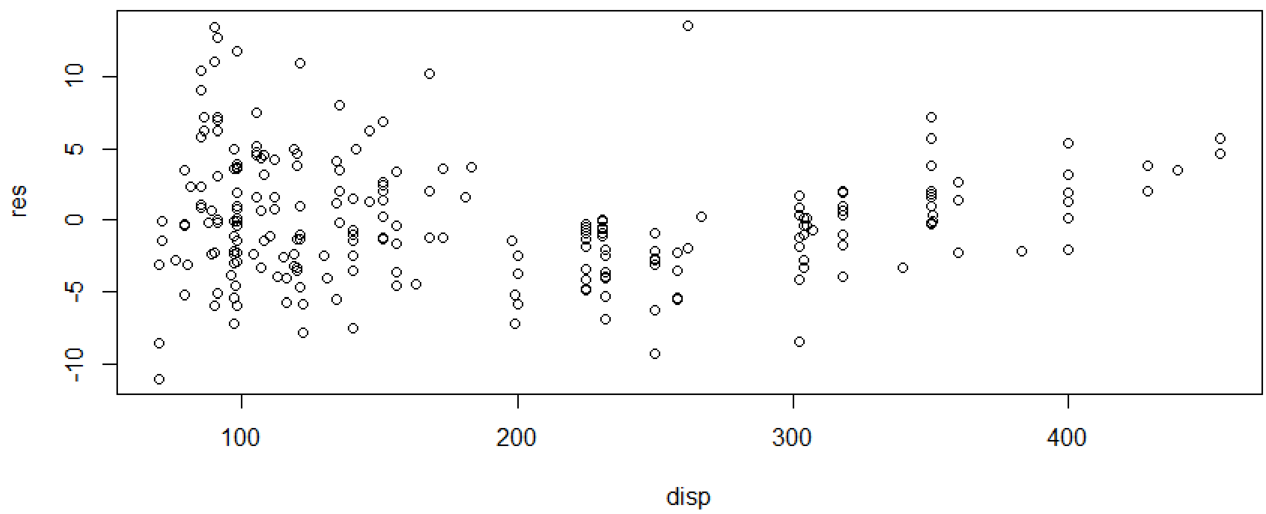
Confidence Interval

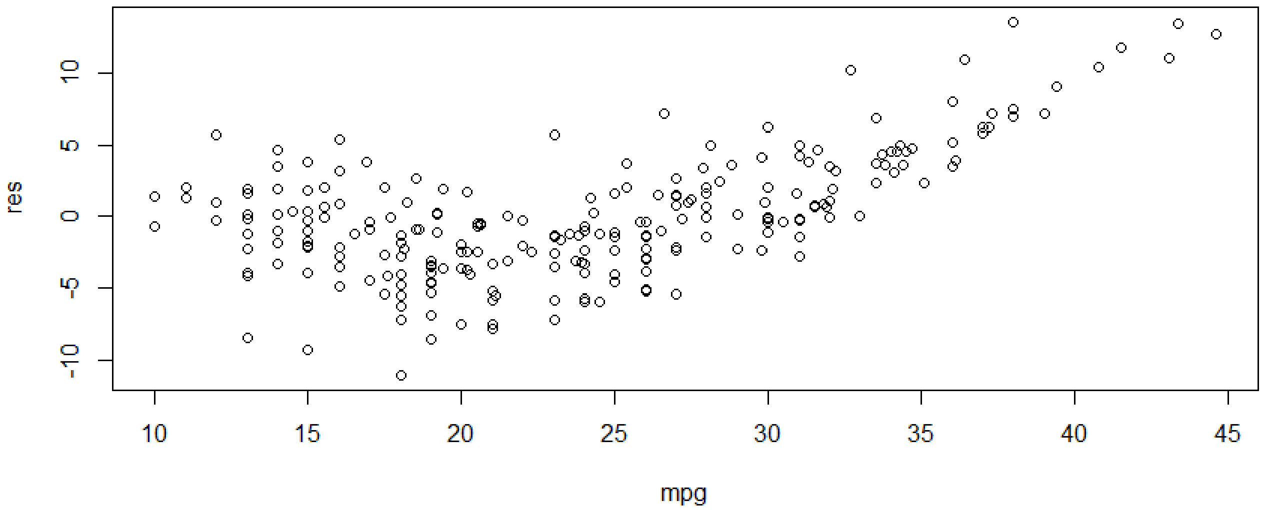
```
> confint(model)
                   2.5 %       97.5 %
(Intercept) 40.978313314 51.009702530
disp        -0.011921797  0.015924482
hp          -0.086949087 -0.014769591
acc         -0.259690085  0.261066984
weg         -0.007732128 -0.004057608
```

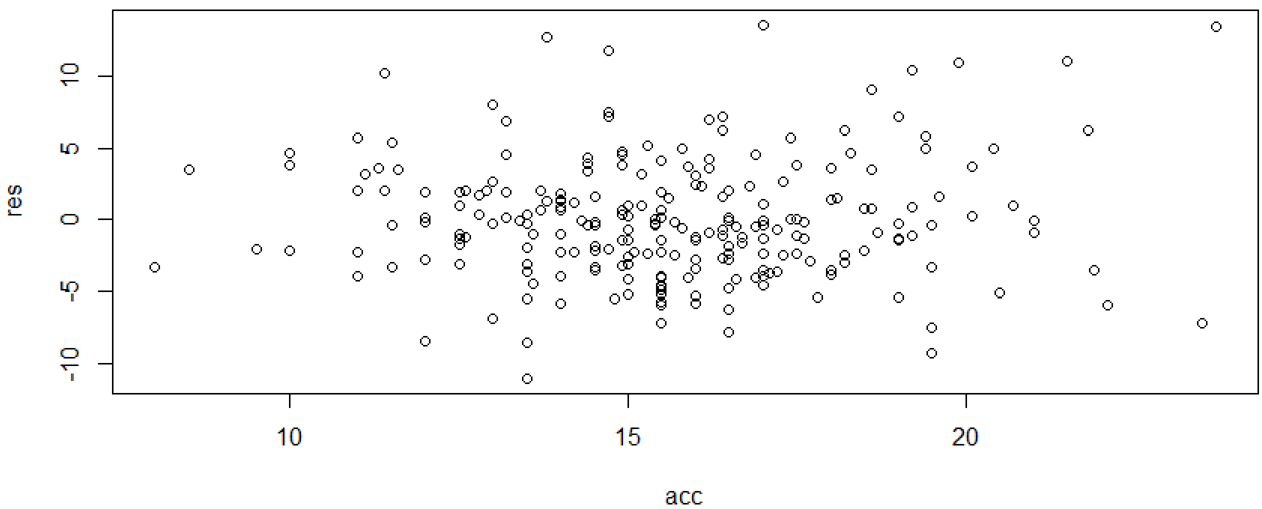
We remove acc as it f value is less than 0.05 has f value less than and make a new subset model.

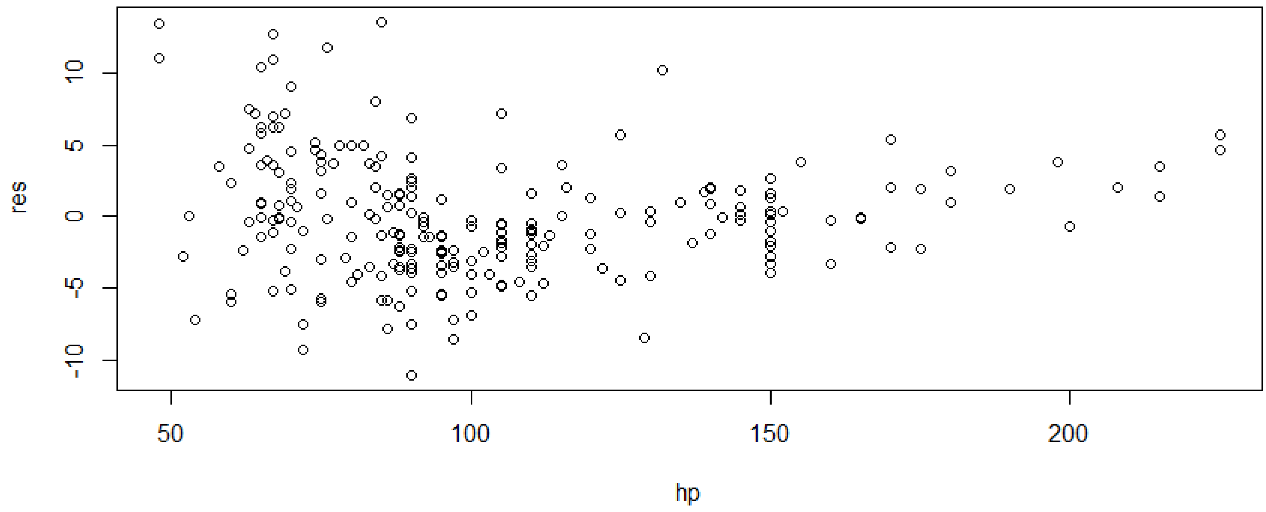


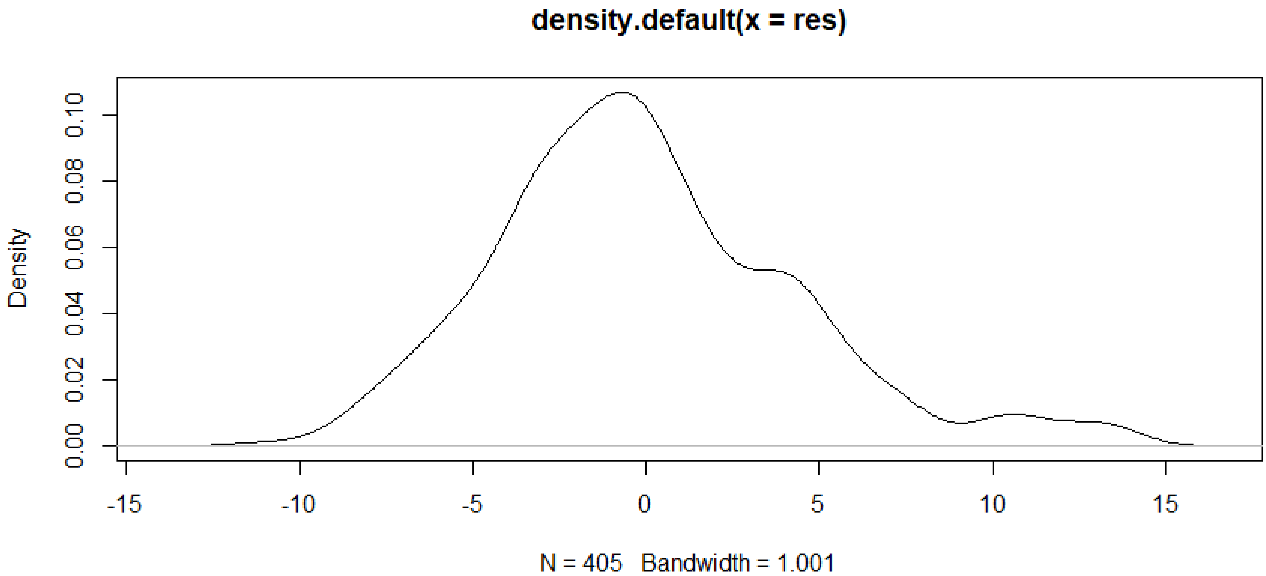
Calculating the residual




res
mpg
10
5
0
-5
-10
10
15
20
25
30
35
40
45


res
acc
10
5
0
-5
-10
10
15
20





Calculating summary of the subset

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.0053540  1.3192105  34.873  < 2e-16 ***
disp         0.0019934  0.0069061   0.289 0.773005
weg         -0.0058924  0.0008034  -7.334 1.25e-12 ***
hp          -0.0509206  0.0140527  -3.624 0.000328 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.389 on 401 degrees of freedom
Multiple R-squared:  0.6744,	Adjusted R-squared:  0.672
F-statistic: 276.9 on 3 and 401 DF,  p-value: < 2.2e-16
```

Using anova calculate the F values

```
> anova(model_subset)
Analysis of Variance Table

Response: mpg
           Df  Sum Sq Mean Sq F value    Pr(>F)
disp        1 14402.1 14402.1 747.523 < 2.2e-16 ***
weg         1  1348.4  1348.4  69.988     1e-15 ***
hp          1   253.0   253.0  13.130 0.0003279 ***
Residuals 401  7725.8    19.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Calculating the R*2

```
> summary(model_subset)$r.squared
[1] 0.6744181
```

Creating the new dataframe or set value to predictor

```
new_data <- data.frame(disp = 5, acc= 13 ,weg= 15)
#predict
prediction <- predict(model, newdata = new_data)
print(prediction)
> print(prediction)
45.41595
```

Calculating the prediction interval interval with 95% confidence

```
prediction_interval <- predict(model, newdata = new_data, interval = "prediction",
level = 0.95)
print(prediction_interval[2:3])
> print(prediction_interval[2:3])
[1] 36.37436 54.45754
```

Appendix C: Code

```
library("readxl")
library(tidyverse)

Auto<-read_excel("D:/MSDS/MTH522/assigment1/auto_data_vislavath_lik
hil.xls)
Auto = na.omit(Auto)
disp <-Auto$`displacement`
hp <-Auto$`horsepower`
weg <-Auto$`weight`
acc <-Auto$`acceleration`
mpg <-Auto$`mpg`

plot(disp,mpg)
summary(Auto)
glimpse(Auto)
# Fit a multiple linear regression model
model <- lm(mpg ~ disp + hp + acc + weg , data= Auto );
plot(model)
res <- resid(model)
plot(res)
# Test whether at least one of the predictors is useful
summary(model)

# Perform an F-test for the overall significance of the regression model
anova(model)
# Residual plot against predicted response
plot(mpg , res)
```

```
# Residual plots against each predictor
plot(disp, res)
plot(acc, res)
plot(hp, res)
plot (density(res))

# 3 predictor
model_subset <- lm(mpg ~ disp + weg  + acc  , data= Auto );
plot(model_subset)
# Test the significance of Predictor3
summary(model_subset)

# Compare the models with and without Predictor3
anova(model_subset, model)


summary(model)$r.squared
summary(model_subset)$r.squared
new_data <- data.frame(disp = 5,hp= 10, acc= 13 ,weg= 15)
#predict
prediction <- predict(model, newdata = new_data)
print(prediction)
par(Auto)

# Calculate the prediction interval for the new set of predictor values
prediction_interval <- predict(model, newdata = new_data, interval =
"prediction", level = 0.95)
# Print the lower and upper bounds of the prediction interval
print(prediction_interval[2:4])
plot(hatvalues(model))
plot(cooks.distance(model))


#confidence interval
```

```
confint(model)
cor(Auto)
pairs.panels(Auto,method ="person")
plot(cor(Auto))
```