Assignment-Week2

Mohit Mehndiratta

Question 1

Part (a) - Upload the Advertising dataset and explore it.

```
Advertising <- read.csv("../datasets/advertising.csv")
head(Advertising)

##      TV Radio Newspaper Sales
## 1 230.1  37.8      69.2  22.1
## 2  44.5  39.3      45.1  10.4
## 3  17.2  45.9      69.3   9.3
## 4 151.5  41.3      58.5  18.5
## 5 180.8  10.8      58.4  12.9
## 6   8.7  48.9      75.0   7.2

attach(Advertising)
dim(Advertising)

## [1] 200   4

sapply(Advertising, mean)

##        TV     Radio Newspaper     Sales
##  147.0425   23.2640   30.5540   14.0225

sapply(Advertising, var)

##         TV      Radio  Newspaper      Sales
## 7370.94989  220.42774  474.30833   27.22185

sapply(Advertising, class)

##        TV     Radio Newspaper     Sales
## "numeric" "numeric" "numeric" "numeric"
```

From here, we can observe that Advertising Data set has 200 Observations of 4 variables and all the variables are numeric.

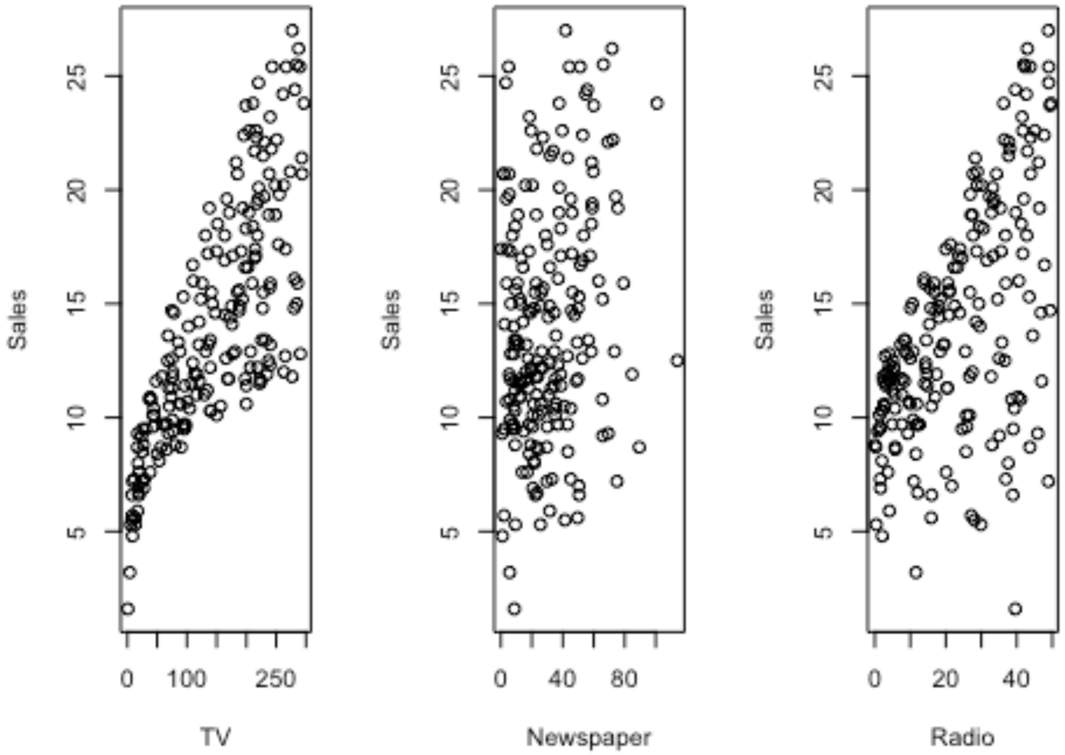
Part(b) - Construct scatter plots to visualize the relationship between following variables:

- Sales and TV

- Sales and Radio

- Sales and Newspaper

```
par(mfrow=c(1,3))
plot(Sales~TV)
plot(Sales~Newspaper)
plot(Sales~Radio)
```



1. For Sales vs TV plot, we can see that there is a linear relationship between Sales and TV as the we can see the constant increase in number of Sales. But at the end we can observe that plot is a bit more scattered thus there is a chance of high errors.

2. For Sales vs Newspaper plot, we can see that plot is heavily spread and have mixed values. It’s hard to depict any relationship between Sales and Newspaper as of now.

3. For Sales vs Radio plot, we can see that the Sales are going up as Radio advertisements are increasing. There is some spread as well which can possible generate errors as well but as of now it is safe to assume that there exists a positive linear relationship.

Part(c) - Find the Correlation Coefficient to measure the strength of the linear relationship of Sales and TV

```
cor(Sales, TV)

## [1] 0.7822244
```

As we can see, Correlation Coefficient between Sales and TV is 0.7822244. As per the correlation coefficient if it is closer to one that means there is a strong positive relationship between the variables.

Since the value is closer to 1 therefore from that we can say that the relationship between Sales and TV is somewhat strong but not very strong. We can safely assume that if TV advertisement go up then there is 78% chance that Sales would go up too.

Part(d) - Find the least square estimates of the linear model of Sales in terms of TV and give the resulting model

```
model1 <- lm(Sales~TV)
summary(model1)

## 
## Call:
## lm(formula = Sales ~ TV)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -8.3860 -1.9545 -0.1913  2.0671  7.2124 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## TV          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
```

```
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

From the above summary, we can see that the least square estimates for the linear model between Sales and TV are 7.032594 and 0.047537 with the standard error as 0.457843 and 0.002691 .

As per our model equation -> E(Y) = b0 + b1(X), where Y is Sales and X is TV. We can find the value for B0^ and B1^ as 7.032594 and 0.047537 respectively.

Part(e) - Assess the accuracy of the parameter estimates

To Assess the accuracy of parameters we can observe the standard error with respect to coefficients. Since the smaller the error the better the coefficients, we can see that errors are quite small for the estimated values of intercept and slope. Since the standard errors are quite small as compared to estimated values we can say that they are somewhat accurate.

```
confint(model1)

##                  2.5 %     97.5 %
## (Intercept) 6.12971927 7.93546783
## TV          0.04223072 0.05284256
```

we can also check the accuracy from confidence interval as well. As we can see 0 is not in between the lower and upper limit therefore we can reject the null hypothesis and can safely say that there is a relationship between Sales and TV.

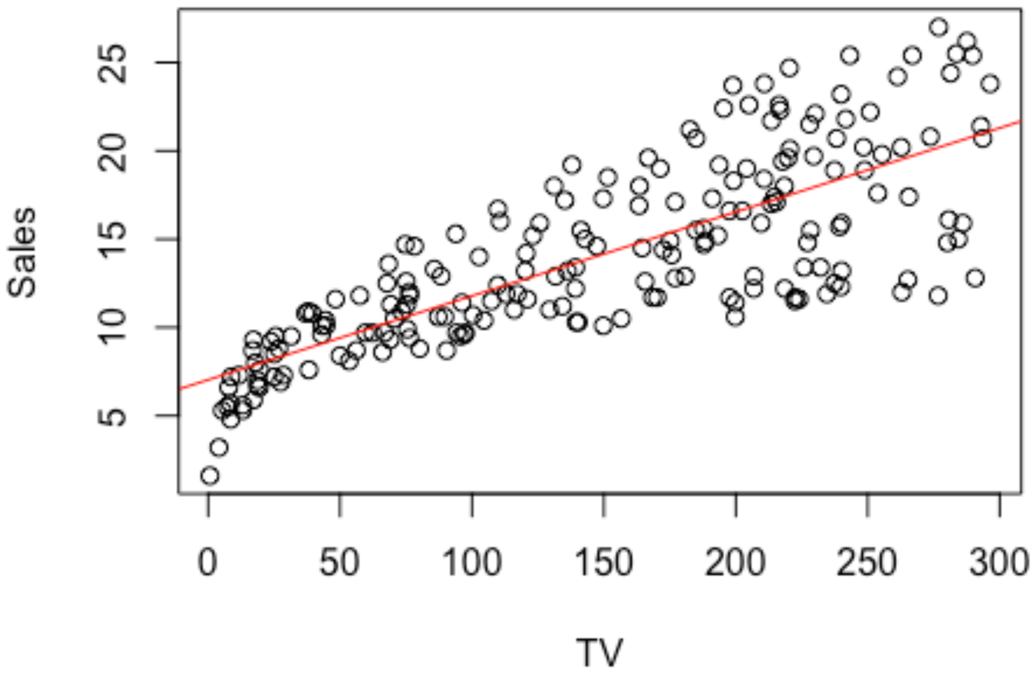
part(f) - Test the significance of the slope of the linear model

For this we can use hypothesis testing. While looking at confidence interval none of the ranges have 0 in them. Therefore we can reject the null hypothesis.

Also, by looking at the p-values for 5% significance, we can see that p-value is much closer to zero and less than 0.05. Therefore we have a strong evidence to reject the null hypothesis and thus it states that there is a significant linear relationship between Sales and TV.

part(g) - Plot the straight line within the scatter plot and comment

```
plot(Sales~TV)
abline(c(7.032594, 0.047537), col="red")
```



here as per the plot we can see that the line is rising upward which means the sales are increasing with more TV advertisement stating a relationship between them. But it seems like the variance is increasing also and we need to check our assumptions.

part(h) - Assess the overall accuracy of the model

After, seeing the Residual standard error in the summary which is 3.259. We can check the Sales summary and then we can compare the error.

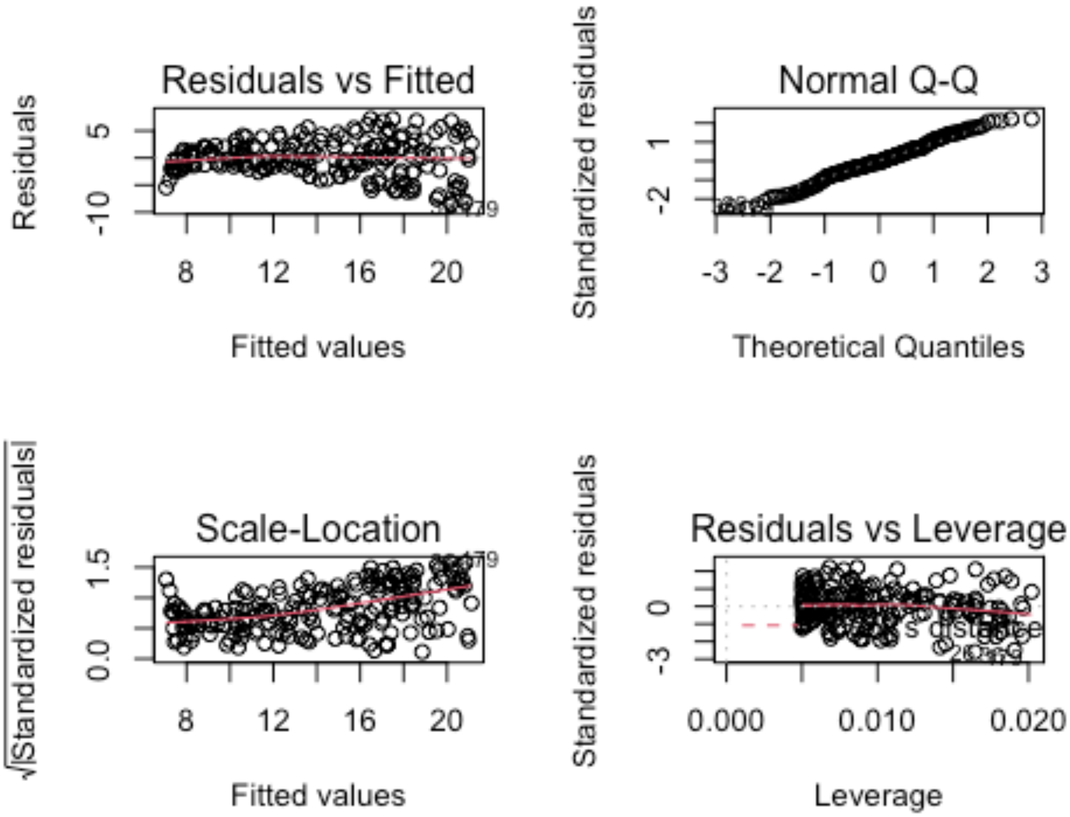
```
summary(Sales)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.60   10.38   12.90   14.02   17.40   27.00
```

So, for the summary we can observe that the Residual standard error or standard deviation seems to be high, which is not good.

By looking at the summary we can see that value for R squared is 61.19%, which means 61.19 % is the proportion of variable explained by the model. The higher the squared, the better is the model. Therefore, it is somewhat better.

We can also check our basics assumptions which are : 1. LINEARITY 2. NORMALITY (NORMAL DISTRIBUTION) 3. CONSTANT VARIANCE ASSUMPTION (Heteroscedasticity)

```
par(mfrow=c(2,2))
plot(model1)
```



For Linearity: if this assumption has to be valid then the plot for Residual vs fitted should be scattered/random and there should not be any pattern. Since we can see pattern here then we can say that Linearity assumption is not valid and same thing is reflecting in scale-location vs Fitted values plot.

For Heteroscedasticity: if this assumption has to be valid then the plot for Residual vs fitted should spread constantly or the variance should be constant throughout. But here also we can see that the the spread is not constant, therefore this assumption is also not valid.

For Normality: if this assumption has to be valid then the plot for Normal Q-Q vs Theoretical Quantities has to have all the plots aligned to the straight line but we do not have all points lied on the straight line therefore, this assumption is also not valid.

Since, our assumptions are not valid this means that there is more room to improve this model.

Part(i) - Use the model to make predictions

```
predict(model1, list(TV=400))

##        1 
## 26.04725
```

For predictions our model is predicting the Sales of 26.04725 when TV advertisement reaches 400.

Question 2

Part (a) - Upload the Auto Dataset and explore it.

```
Auto <- read.csv("../datasets/auto.csv")
head(Auto)

##   mpg cylinders displacement horsepower weight acceleration year
origin
## 1  18         8          307        130   3504         12.0   70
1
## 2  15         8          350        165   3693         11.5   70
1
## 3  18         8          318        150   3436         11.0   70
1
## 4  16         8          304        150   3433         12.0   70
1
## 5  17         8          302        140   3449         10.5   70
1
## 6  15         8          429        198   4341         10.0   70
1
##                         name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

```
attach(Auto)
dim(Auto)
```

```
## [1] 397   9
```

```
sapply(Auto, mean)
```

```
## Warning in mean.default(X[[i]], ...): argument is not numeric or
logical:
## returning NA

## Warning in mean.default(X[[i]], ...): argument is not numeric or
logical:
## returning NA
```

```
##          mpg    cylinders displacement   horsepower       weight
acceleration
##    23.515869     5.458438   193.532746           NA  2970.261965
15.555668
##         year       origin         name
##    75.994962     1.574307           NA
```

```
sapply(Auto, var)
```

```
## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion
```

```
##          mpg    cylinders displacement   horsepower       weight
acceleration
## 6.124321e+01 2.895364e+00 1.089510e+04           NA 7.189414e+05
7.562474e+00
##         year       origin         name
## 1.361614e+01 6.440857e-01           NA
```

```
sapply(Auto, class)
```

```
##          mpg    cylinders displacement   horsepower       weight
acceleration
##    "numeric"    "integer"    "numeric"  "character"    "integer"
```

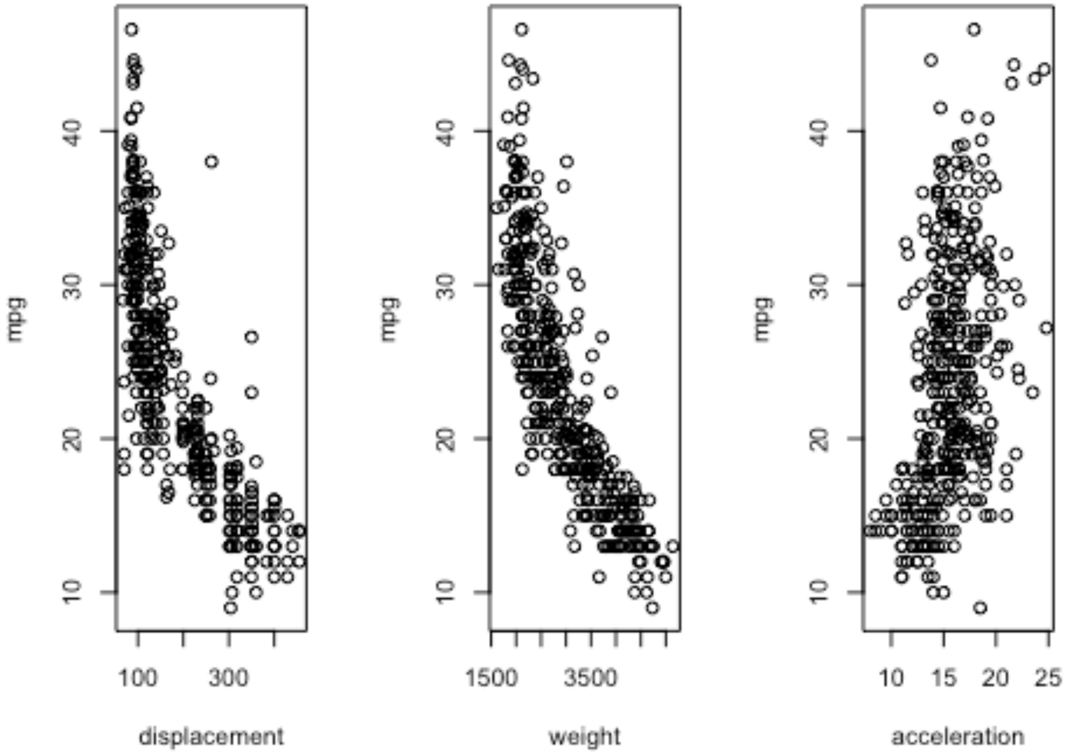
```
"numeric"
##         year       origin         name
##    "integer"    "integer"  "character"
```

From here, we can observe that Auto Data set has 397 Observations of 9 variables. The data set is about the various specifications of the different cars.

Part(b) - Construct scatter plots to visualize the relationship between following variables:

- **mpg and displacement**
- **mpg and weight**
- **mpg and acceleration**

```
par(mfrow=c(1,3))
plot(mpg~displacement)
plot(mpg~weight)
plot(mpg~acceleration)
```



1. For mpg vs displacement plot, we can see that there is a mix of mpg values when displacement is low for various cars. But when displacement is large, mpg is going down. Therefore, it looks like we have a negative relationship between them.

2. For mpg vs weight plot, we can see that heavy cars have low mpg as compared to light weight cars. The plot is depicting a negative linear relationship between mpg and weight.

3. For mpg vs acceleration plot, the scatter plot is heavily spread and have mixed values. It's hard to depict any relationship between mpg and acceleration as of now.

Part(c) - Find the Correlation Coefficient to measure the strength of the linear relationship of mpg and weight.

```
cor(mpg, weight)

## [1] -0.8317389
```

As we can see, Correlation Coefficient between mpg and weight is -0.8317389. As per the correlation coefficient if it is closer to -1 that means there is a strong negative relationship between the variables.

Since the value is closer to -1 therefore, from that we can say that the relationship between mpg and weight is somewhat strong but not very strong. We can safely assume that the light weight cars should have high mpg.

Part(d) - Find the least square estimates of the linear model of mpg in terms of weight. and give the resulting model

```
model2 <- lm(mpg~weight)
summary(model2)

##
## Call:
## lm(formula = mpg ~ weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0123  -2.8076  -0.3541   2.1145  16.4802
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.3173992  0.7962915   58.17   <2e-16 ***
## weight      -0.0076766  0.0002578  -29.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.35 on 395 degrees of freedom
## Multiple R-squared:  0.6918, Adjusted R-squared:  0.691
## F-statistic: 886.6 on 1 and 395 DF,  p-value: < 2.2e-16
```

From the above summary, we can see that the least square estimates for the linear model between mpg and weights are 46.3173992 and -0.0076766 with the standard error as 0.7962915 and 0.0002578 .

As per our model equation -> E(Y) = b0 + b1(X), where Y is mpg and X is weight. We can find the value for B0^ and B1^ as 46.3173992 and -0.0076766 respectively.

Part(e) - Assess the accuracy of the parameter estimates

To Assess the accuracy of parameters we can observe the standard error with respect to coefficients. Since the smaller the error the better the coefficients, we can see that errors are quite small for the estimated values of intercept and slope. Since the standard errors are quite small as compared to estimated values we can say that they are somewhat accurate.

```
confint(model2)

##                     2.5 %       97.5 %
## (Intercept) 44.751899760 47.882898590
## weight      -0.008183466 -0.007169746
```

we can also check the accuracy from confidence interval as well. As we can see 0 is not in between the lower and upper limit therefore we can reject the null hypothesis and can safely say that there is some relationship between mpg and weight.

As, we can see the value for the slope is negative, from that we can evidently say that there exists a negative linear relationship.

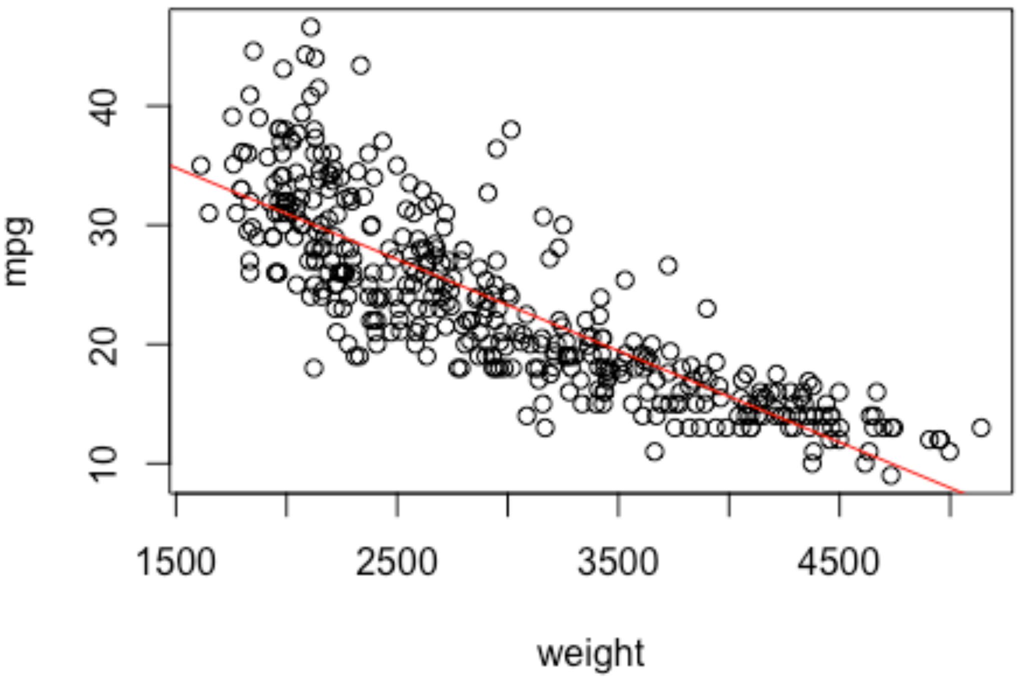
part(f) - Test the significance of the slope of the linear model

For this we can use hypothesis testing. While looking at confidence interval none of the ranges have 0 in them. Therefore we can reject the null hypothesis.

Alternatively, by looking at the p-values for 5% significance, we can see that p-value is much closer to zero and less than 0.05. Therefore we have a strong evidence to reject the null hypothesis and thus it states that there is a significant linear relationship between mpg and displacement.

part(g) - Plot the straight line within the scatter plot and comment

```
plot(mpg~weight)
abline(c(46.3173992, -0.0076766), col="red")
```



here as per the plot we can see that the line is going downward which states that heavy cars have generally less mpg. But we can also see that there aren't any constant spread in the plot. We need to check our assumptions here.

part(h) - Assess the overall accuracy of the model

After, seeing the Residual standard error in the summary which is 4.35. We can check the mpg summary and then we can compare the error.

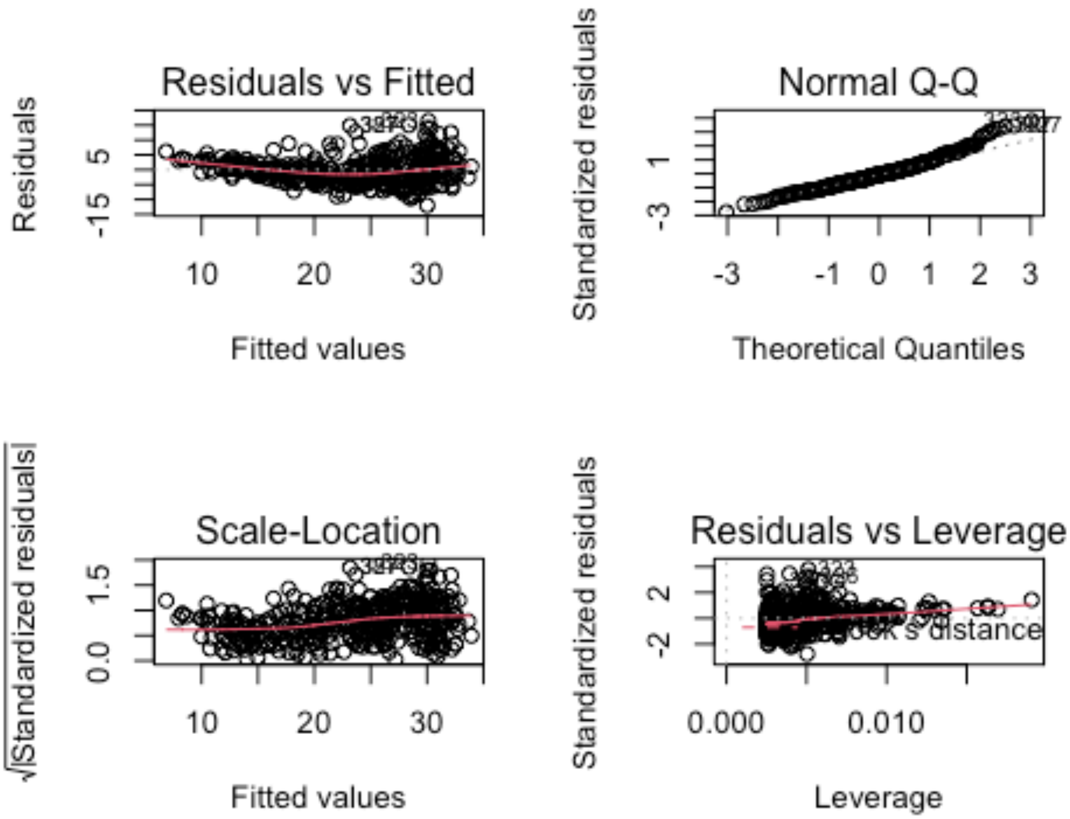
```
summary(mpg)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    9.00   17.50   23.00   23.52   29.00   46.60
```

So, for the summary we can observe that the Residual standard error not seems to be high, which is good.

By looking at the summary we can see that value for R squared is 69.18%, which means 69.18 % is the proportion of variable explained by the model. The higher the R-squared value, the better is the model. Therefore, we can say that it is somewhat better.

We can also check our basics assumptions which are : 1. LINEARITY 2. NORMALITY (NORMAL DISTRIBUTION) 3. CONSTANT VARIANCE ASSUMPTION (Heteroscedasticity)

```
par(mfrow=c(2,2))
plot(model2)
```



For Linearity: if this assumption has to be valid then the plot for Residuals vs fitted should be scattered/random and there should not be any pattern. Since we can see pattern here then we can say that Linearity Assumption is not valid and same thing is reflecting in scale vs Fitted plot.

For Heteroscedasticity: if this assumption has to be valid then the plot for Scale-Location vs Fitted values should spread constantly the variance should be constant throughout. But here also we can see that the the spread is not constant, therefore this assumption is also not valid.

For Normality: if this assumption has to be valid then the plot for Normal Q-Q vs Theoretical Quantities has to have all the plots aligned to the straight line but we do not have all points lied on the straight line therefore, this assumption is also not valid.

Since, our assumptions are not valid this means that there is more room to improve this model.

Part(i) - Use the model to make predictions

```
predict(model2, list(weight=1300))

##        1
## 36.33781
```

For predictions, our model is predicting that if car's weight is reduced to 1300 then the car should expect to have mpg around 36.33781.