Assignment 2_Mayank_Jaggi

Mayank Jaggi

February 15, 2018

Problem 1

(a)

```
library(ISLR)
attach(Auto)
lm.fit=lm(mpg~horsepower)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66   <2e-16 ***
## horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

(i) As the p-value is very small for horespower,we reject the null hypothesis, thus the coefficient of horespower is not zero. Therefore, there is a relationship between predictor and repsonse.

(ii) Since the value of adjusted and R^2 statistic is >0.5, there is a strong relationship between predictor and response.

(iii) Negative relationship between predictor and response as the slope is negative.

(iv) With an increase in horespower by 1000 units, it will decrease mpg by 157.845 units

```
#part (v)
predict(lm.fit,data.frame(horsepower=98),interval = "confidence")  #confidence interval
```

```
##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

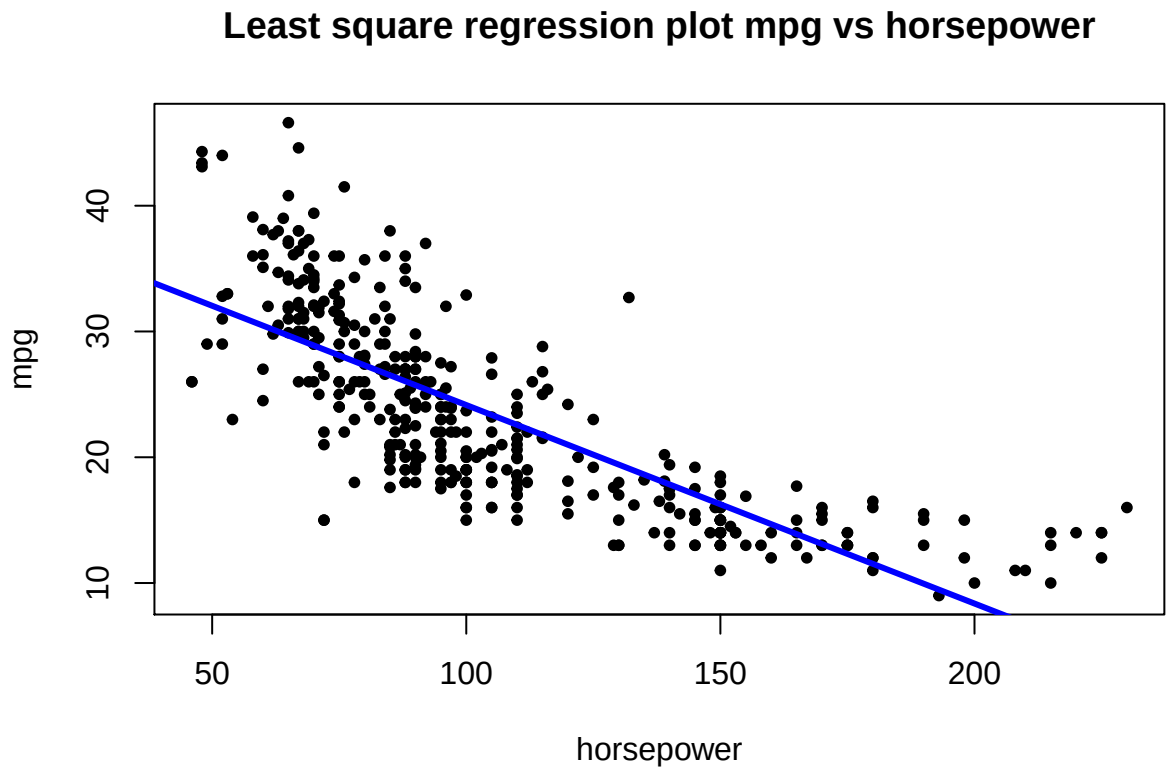
```
predict(lm.fit,data.frame(horsepower=98),interval = "prediction") #prediction interval
```

```
##        fit     lwr      upr
## 1 24.46708 14.8094 34.12476
```

(v) Predicted value of mpg is 24.46708. The values mentioned above are associated with 95% confidence and prediction intervals.

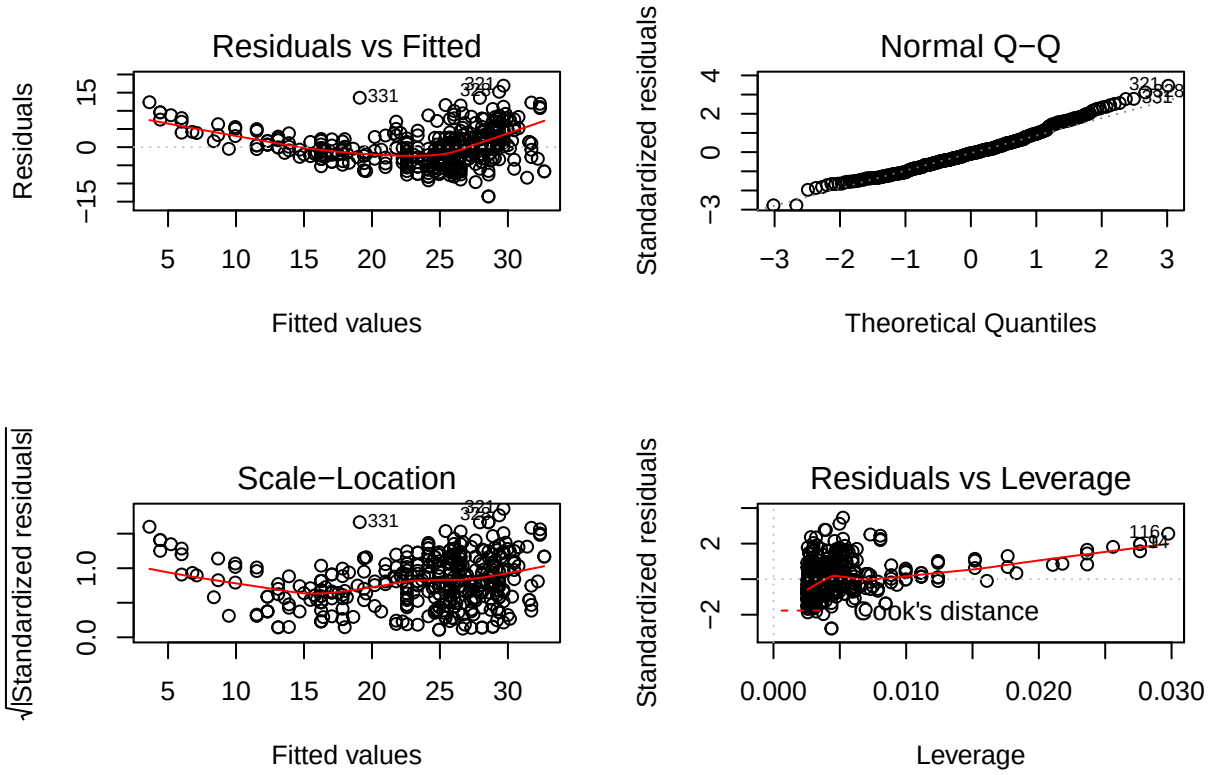
(b)

```
plot(horsepower,mpg,pch=20,col="black",main="Least square regression plot mpg vs horsepower")
abline(lm.fit,lwd=3,col="blue")
```



(c)

```
par(mfrow=c(2,2))
plot(lm.fit)
```



Residual vs Fitted: The above plot shows a parabola and shows that the model failed to explain the non-linear relationship.

Normal Q-Q: Since its a straight line, the residuals are normally distributed.

Scale- Location: Residuals are not equally spaced along the line and are crowded towards right of the X axis. So variance may not be equal.

Residuals vs Leverage: All the points are within the Cook's distance line, so there are no influential outliers present.

(d)

1.) log(x)

```
x=horsepower                                    #new variable
lm.fit1=lm(mpg~log(x))
summary(lm.fit1)
```

```
##
## Call:
## lm(formula = mpg ~ log(x))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2299  -2.7818  -0.2322   2.6661  15.4695
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 108.6997     3.0496   35.64   <2e-16 ***
## log(x)      -18.5822     0.6629  -28.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.501 on 390 degrees of freedom
## Multiple R-squared:  0.6683, Adjusted R-squared:  0.6675
## F-statistic: 785.9 on 1 and 390 DF,  p-value: < 2.2e-16
```

(i) As the p-value is very small for log(x),we reject the null hypothesis, thus the coefficient of log(x) is not zero. Therefore, there is a relationship between predictor and repsonse.

(ii) Since the value of adjusted and R^2 statistic is >0.5, there is a strong relationship between predictor and response.

(iii) Negative relationship between predictor and response as the slope is negative.

(iv) `With an increase in horsepower by 1000 units, it will decrease mpg by 55.74 units`

```
#part (v)
predict(lm.fit1,data.frame(x=98),interval = "confidence")  #confidence interval
```

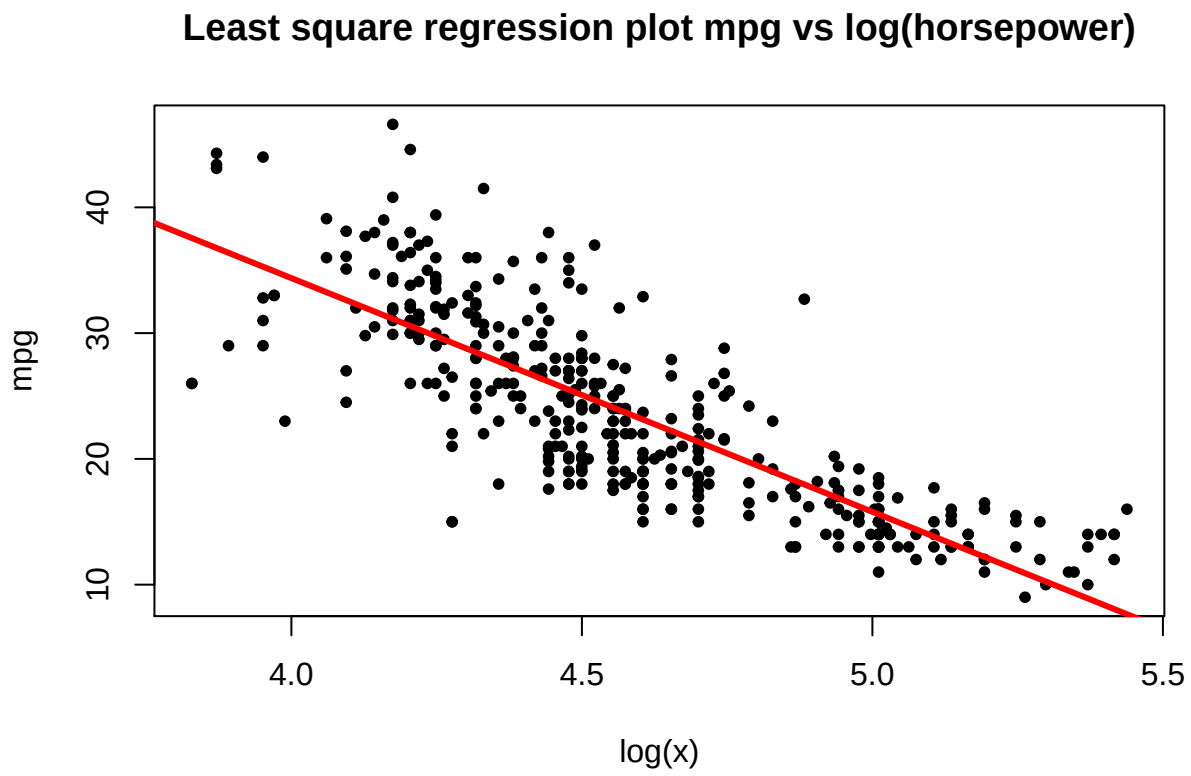
```
##        fit      lwr      upr
## 1 23.50099 23.05405 23.94794
```

```
predict(lm.fit1,data.frame(x=98),interval = "prediction") #prediction interval
```

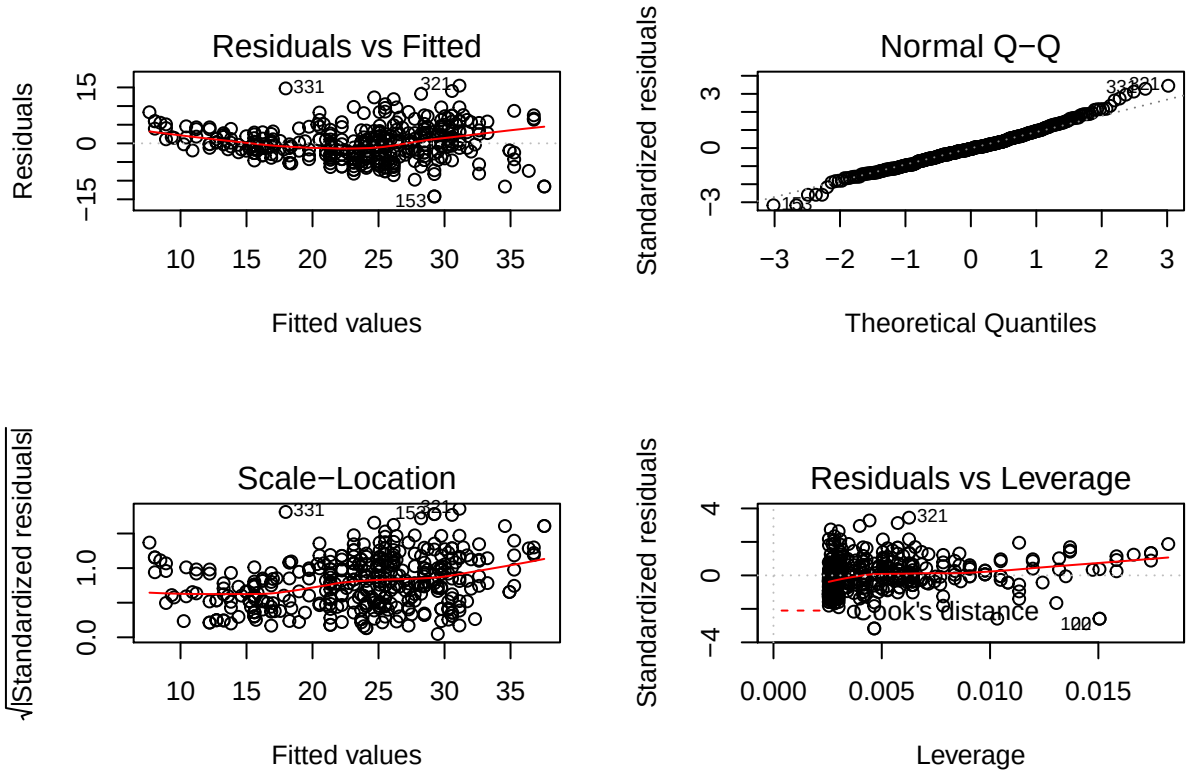
```
##        fit      lwr      upr
## 1 23.50099 14.64106 32.36093
```

(v) Predicted value of mpg is 23.5009. The values mentioned above are associated with 95% confidence and prediction intervals.

```
plot(log(x),mpg,pch=20,col="black",main="Least square regression plot mpg vs log(horsepower)")
abline(lm.fit1,lwd=3,col="red")
```




```
par(mfrow=c(2,2))
plot(lm.fit1)
```



Residual vs Fitted: The above plot shows a parabola and shows that the model failed to explain the non-linear relationship.

Normal Q-Q: Residuals beyond 2 and below -2 (x axis) do not follow normal distribution

Scale- Location: Residuals are equally spaced so the variance is equal for all the observations.

Residuals vs Leverage: All the points are within the Cook's distance line, so there are no influential outliers present.

2.) sqrt(x)

```
lm.fit2=lm(mpg~sqrt(x))
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = mpg ~ sqrt(x))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.9768  -3.2239  -0.2252   2.6881  16.1411
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   58.705      1.349   43.52   <2e-16 ***
## sqrt(x)       -3.503      0.132  -26.54   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.665 on 390 degrees of freedom
## Multiple R-squared:  0.6437, Adjusted R-squared:  0.6428
## F-statistic: 704.6 on 1 and 390 DF,  p-value: < 2.2e-16
```

(i) As the p-value is very small for sqrt(x),we reject the null hypothesis, thus the coefficient of sqrt(x) is not zero. Therefore, there is a relationship between predictor and repsonse.

(ii) Since the value of adjusted and R^2 statistic is >0.5, there is a strong relationship between predictor and response.

(iii) Negative relationship between predictor and response as the slope is negative.

(iv) `With an increase in hosrepower by 1000 units, it will decrease mpg by 110 units`

```
#part (v)
predict(lm.fit2,data.frame(x=98),interval = "confidence")  #confidence interval
```

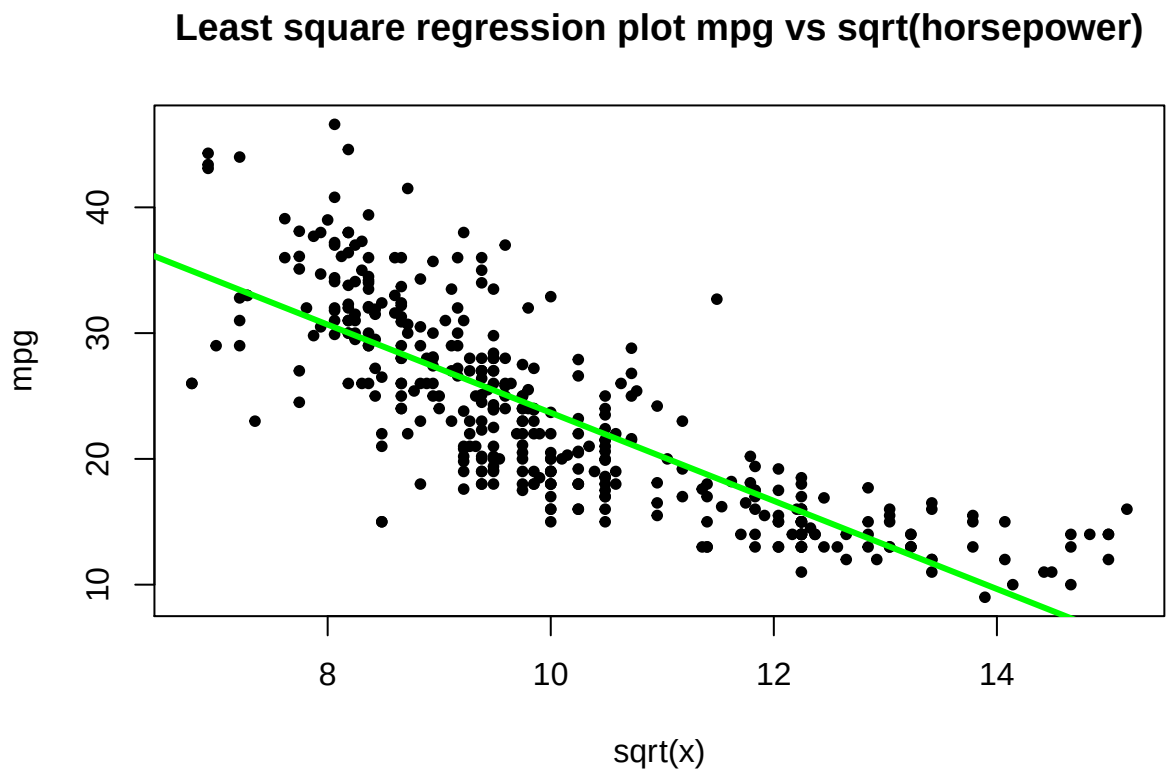
```
##        fit      lwr      upr
## 1 24.02206 23.55687 24.48724
```

```
predict(lm.fit2,data.frame(x=98),interval = "prediction") #prediction interval
```

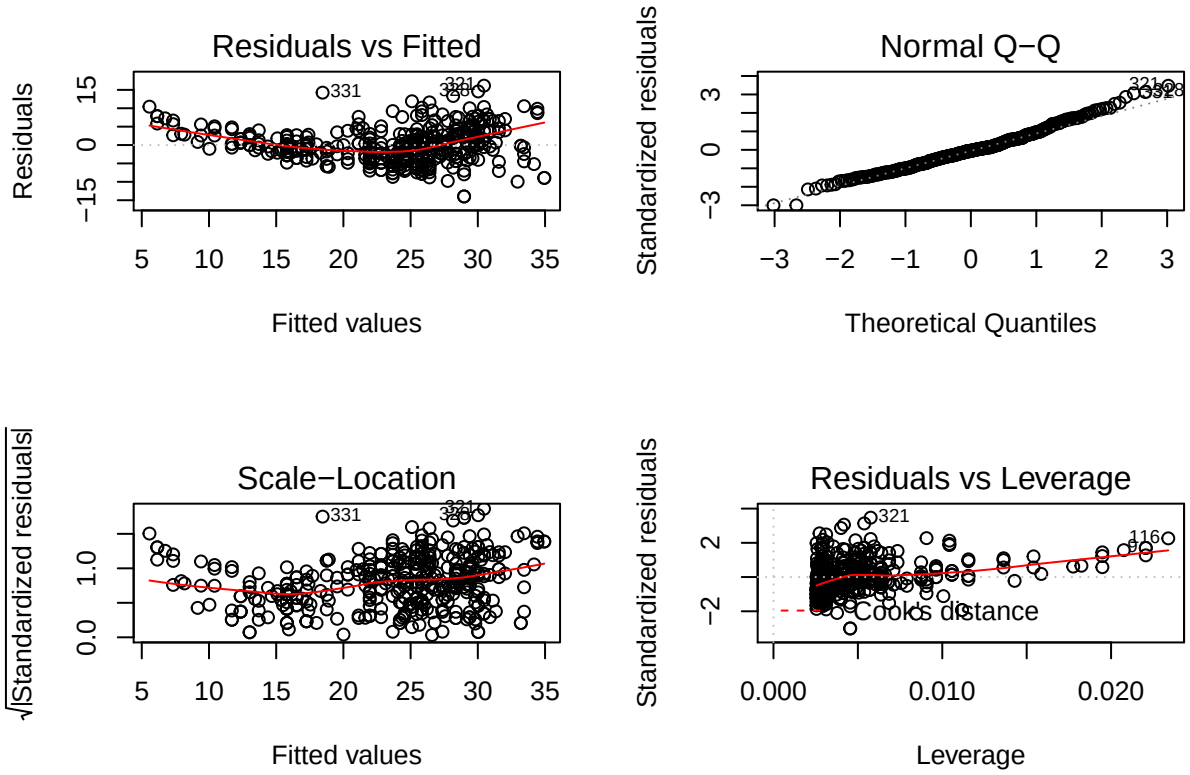
```
##        fit      lwr      upr
## 1 24.02206 14.83892 33.20519
```

(v) Predicted value of mpg is 24.02206. The values mentioned above are associated with 95% confidence and prediction intervals.

```
plot(sqrt(x),mpg,pch=20,col="black",main="Least square regression plot mpg vs sqrt(horsepower)")
abline(lm.fit2,lwd=3,col="green")
```




```
par(mfrow=c(2,2))
plot(lm.fit2)
```



Residual vs Fitted: The above plot shows a parabola and shows that the model failed to explain the non-linear relationship.

Normal Q-Q: Residuals beyond 2 and below -2 (x axis) do not follow normal distribution

Scale- Location: Residuals are equally spaced so the variance is equal for all the observations.

Residuals vs Leverage: All the points are within the Cook's distance line, so there are no influential outliers present.

3.) x^2

```
lm.fit3=lm(mpg~I(x^2))
summary(lm.fit)
```

```
##
## Call:
## lm(formula = mpg ~ I(x^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.529  -3.798  -1.049   3.240  18.528
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.047e+01  4.466e-01   68.22   <2e-16 ***
## I(x^2)      -5.665e-04  2.827e-05  -20.04   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.485 on 390 degrees of freedom
## Multiple R-squared:  0.5074, Adjusted R-squared:  0.5061
## F-statistic: 401.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

(i) As the p-value is very small for x^2,we reject the null hypothesis, thus the coefficient of x^2 is not zero. Therefore, there is a relationship between predictor and repsonse.

(ii) Since the value of adjusted and R^2 statistic is >0.5, there is a moderately strong relationship between predictor and response.

(iii) Negative relationship between predictor and response as the slope is negative.

(iv) `With an increase in hosrepower by 1000 units, it will decrease mpg by 566.5 units`

```
#part (v)
predict(lm.fit3,data.frame(x=98),interval = "confidence")  #confidence interval
```

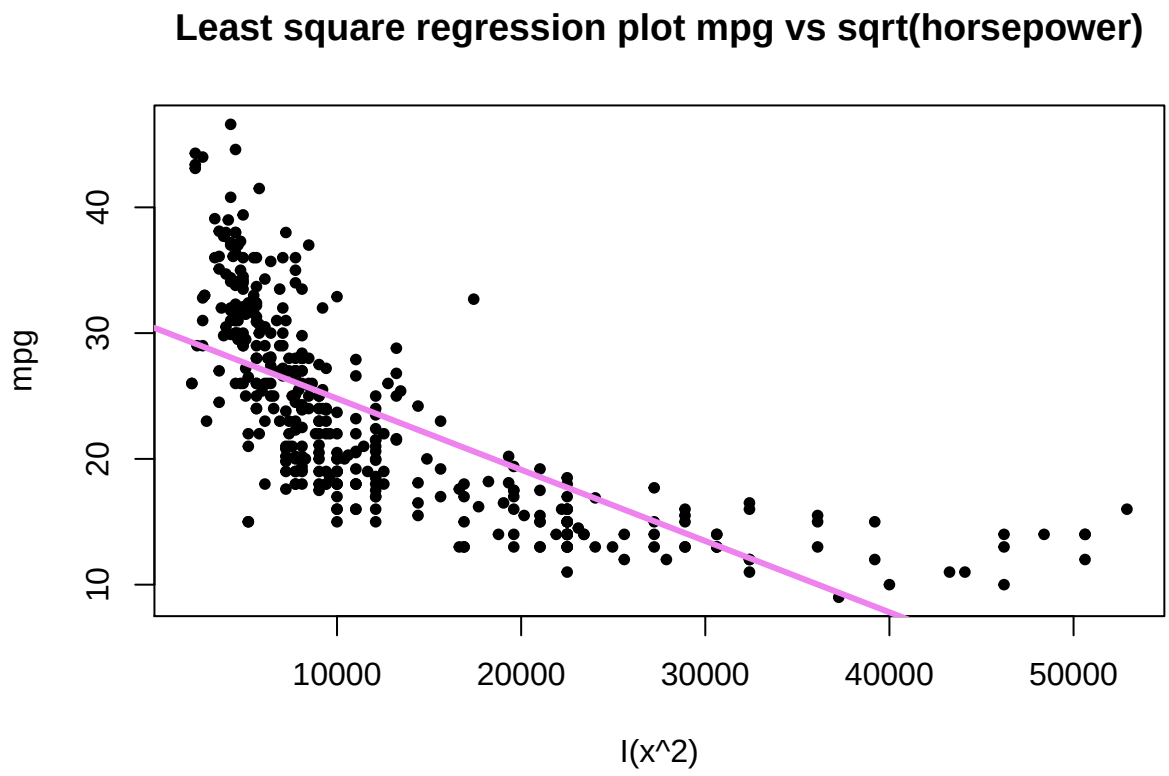
```
##        fit      lwr     upr
## 1 25.02512 24.45883 25.5914
```

```
predict(lm.fit3,data.frame(x=98),interval = "prediction") #prediction interval
```

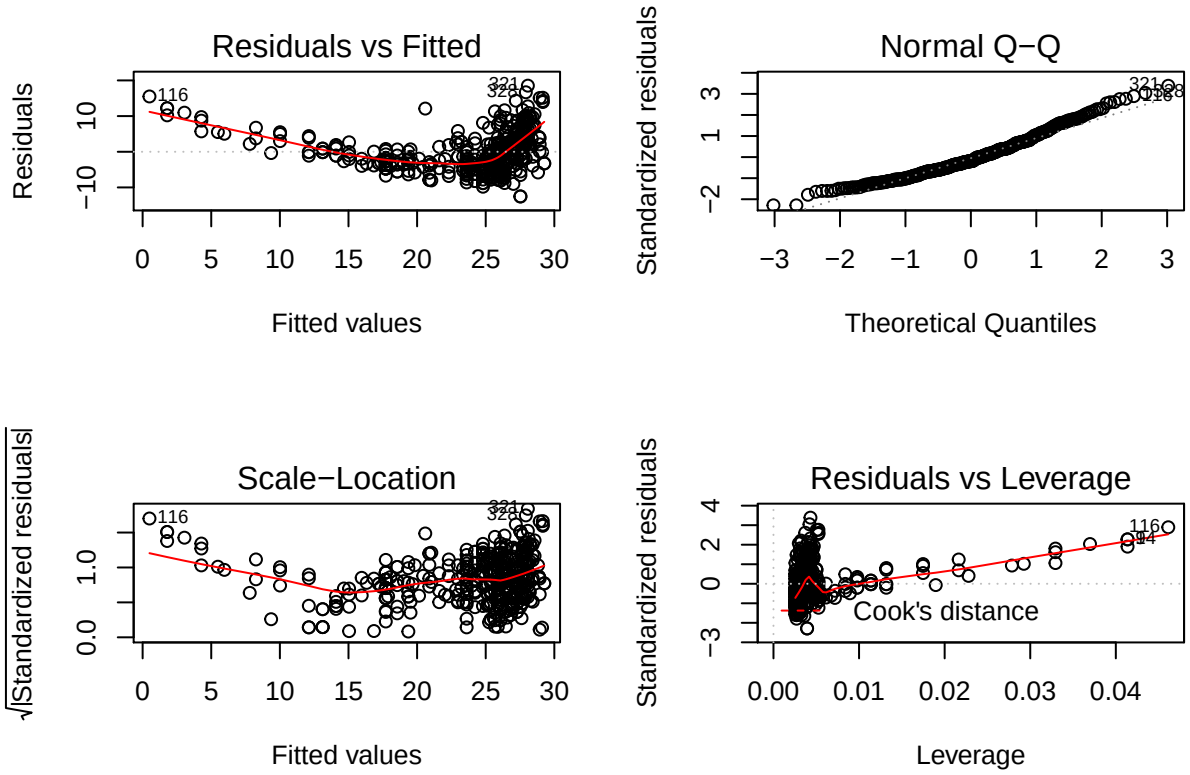
```
##        fit      lwr     upr
## 1 25.02512 14.22603 35.8242
```

(v) Predicted value of mpg is 25.02512. The values mentioned above are associated with 95% confidence and prediction intervals.

```
plot(I(x^2),mpg,pch=20,col="black",main="Least square regression plot mpg vs sqrt(horsepower)")
abline(lm.fit3,lwd=3,col="violet")
```




```
par(mfrow=c(2,2))
plot(lm.fit3)
```



Residual vs Fitted: The above plot shows a parabola and shows that the model failed to explain the non-linear relationship.

Normal Q-Q: Residuals beyond 2 and below -2 (x axis) do not follow normal distribution.

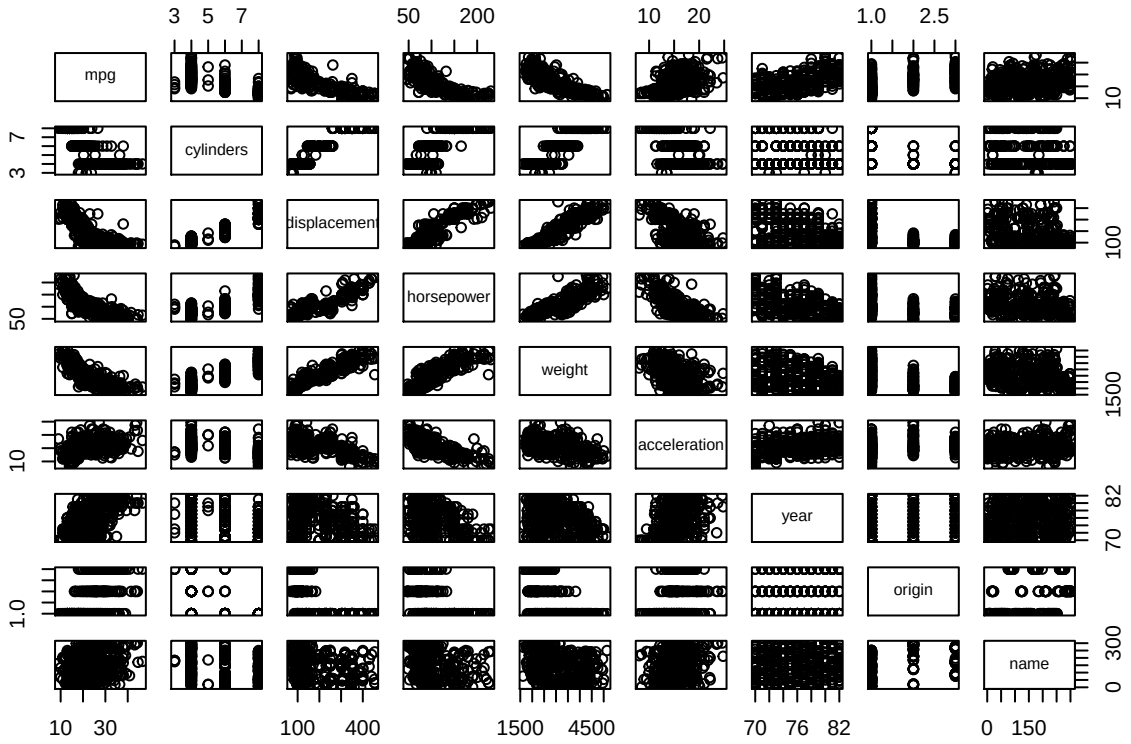
Scale- Location: Residuals are not equally spaced, they tend to spread below 20 units in x-axis so the variance may not be equal.

Residuals vs Leverage: All the points are within the Cook's distance line, so there are no influential outliers present.

Problem 2

(a)

```
pairs(Auto)
```



mpg, displacement, horsepower, weight, accleration appear to have an association with one another.

(b)

```
A=subset(Auto,select=-c(9))
cor(A)
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##              acceleration       year     origin
## mpg             0.4233285  0.5805410  0.5652088
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

(c)

```
lm.fit4=lm(mpg~. -name,data = Auto)
summary(lm.fit4)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

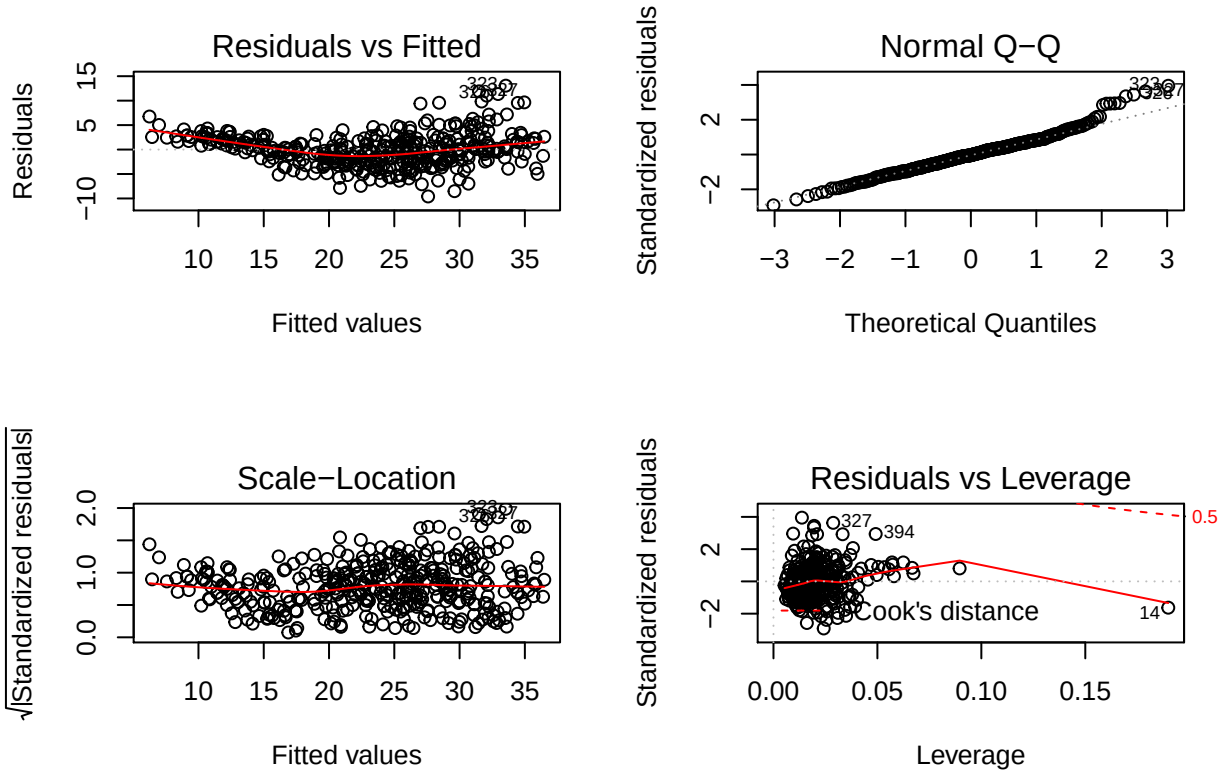
(i) There is a relationship between predictors and response.

(ii) Displacement, weight, year, origin are the predictors having statisically significant relationship with the response mpg.

(iii) With an increase in year by 10 units, it will increase mpg by 7.51 units. There is a positive correlation between the two.

(d)

```
par(mfrow=c(2,2))
plot(lm.fit4)
```



Residual vs Fitted: The above plot shows a parabola and shows that the model failed to explain the non-linear relationship.

Normal Q-Q: Residuals beyond 2 (x axis) do not follow normal distribution.

Scale- Location: Residuals are not equally spaced, they tend to spread above 30 units in x-axis, so the variance may not be equal.

Residuals vs Leverage: All the points are within the Cook's distance line, so there are no influential outliers present.

(e)

```
library(car)
vif(lm.fit4)
```

```
##    cylinders displacement   horsepower       weight acceleration
##    10.737535    21.836792     9.943693    10.831260     2.625806
##         year       origin
##     1.244952     1.772386
```

Yes. There is serious collinearity in the following predictors: Cylinders, displacement, horsepower, weight as their VIF values are >5.

(f)

From the correlation matrix, it is evident that the two highest correlation is between the following predictors: Cylinders: displacement Displacement: weight

```
lm.fit5=lm(mpg~cylinders*displacement+displacement*weight)
summary(lm.fit5)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders * displacement + displacement *
##     weight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.2934  -2.5184  -0.3476   1.8399  17.7723
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             5.262e+01  2.237e+00  23.519  < 2e-16 ***
## cylinders               7.606e-01  7.669e-01   0.992    0.322
## displacement           -7.351e-02  1.669e-02  -4.403 1.38e-05 ***
## weight                 -9.888e-03  1.329e-03  -7.438 6.69e-13 ***
## cylinders:displacement -2.986e-03  3.426e-03  -0.872    0.384
## displacement:weight     2.128e-05  5.002e-06   4.254 2.64e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.103 on 386 degrees of freedom
## Multiple R-squared:  0.7272, Adjusted R-squared:  0.7237
## F-statistic: 205.8 on 5 and 386 DF,  p-value: < 2.2e-16
```

Yes. As we can see the interaction effect between predictors displacement and weight is satistically significant. Whereas, interaction effect bettween cylinders and weight is not satistically significant.

Problem 3

(a)

```
attach(Carseats)
lm.fit6=lm(Sales~Price + Urban + US)
summary(lm.fit6)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Urban + US)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9206 -1.6220 -0.0564  1.5786  7.0581
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
## Price       -0.054459   0.005242 -10.389  < 2e-16 ***
## UrbanYes    -0.021916   0.271650  -0.081    0.936
## USYes        1.200573   0.259042   4.635 4.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.472 on 396 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2335
## F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

b)

The coefficient of the "Price" variable may be interpreted by saying that the effect of a price increase of 1000 dollars is a decrease of 54.4588492 units in sales provided all other predictors remaining fixed.

Since it is Yes for "Urban", it can be interpreted as that urban areas have 21.916 unit sales lower provided all other predictors remain fixed.

The coefficient of the "US" variable may be interpreted by saying that on average the unit sales in a US store are 1.2 units more than in a non US store provided all other predictors remaining fixed.

c)

Model in equation form:

Sales= 13.043469 - 0.054459 x Price + 1.200573 x US + Error

d)

We can reject the null hypothesis for Price and US predictors as its is p-value is very small.

e)

```
lm.fit7=lm(Sales~Price + US)
summary(lm.fit7)
```

```
##
## Call:
## lm(formula = Sales ~ Price + US)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## Price       -0.05448    0.00523 -10.416  < 2e-16 ***
## USYes        1.19964    0.25846   4.641 4.71e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

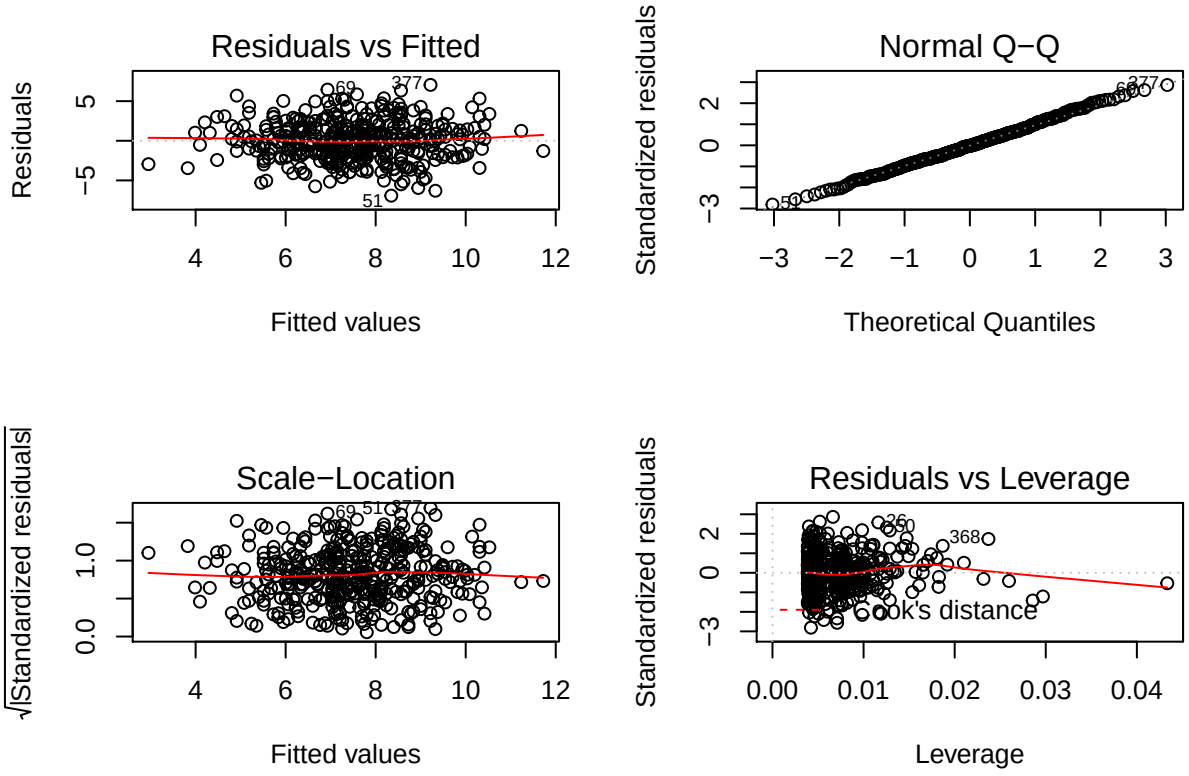
```
##
## Residual standard error: 2.469 on 397 degrees of freedom
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2354
## F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

f)

The R squared values is the same in both models. Whereas, adjusted R squared value is marginally better in model e). Thus, model e) fits the data marginally better. The overall fit is not very well as only 23.54% of the variance can be explained by the model which is lower than the recommended 0.5 value.

g)

```
par(mfrow=c(2,2))
plot(lm.fit7)
```



From the Residual vs Leverage graph, we can say that there are some outliers present in the data. Points above 2 and below -2 along y axis are considered to be outliers.