SAMBANDH BHUSAN DHAL

ISEN 613

2ND ASSIGNMENT

Problem 1

Use the Auto data set to answer the following questions:

(a) Perform a simple linear regression with mpg as the response and horsepower as the predictor. Comment on the output. For example

i. Is there a relationship between the predictor and the response?
ii. How strong is the relationship between the predictor and the response?
iii. Is the relationship between the predictor and the response positive or negative?
iv. How to interpret the estimate of the slope?
v. What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?

(b) Plot the response and the predictor. Display the least squares regression line in the plot.

(c) Produce the diagnostic plots of the least squares regression fit. Comment on each plot.

(d) Try a few different transformations of the predictor, such as $\log(X)$, $\sqrt{X}$, X^2, and repeat (a)-(c). Comment on your findings.

Ans 1.

```
> fix(Auto)
> lm.fit=lm(mpg~horsepower,data=Auto)
> summary(lm.fit)

Call:
lm(formula = mpg ~ horsepower, data = Auto)

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
```

(a). i.As the value of B1 is not zero, there exists a relationship between the predictor and the response.

ii.As we can see that the RSE is 4.906, this means that even if the model is correct and the values of B0 and B1 are known to us, but any prediction of mpg based on horsepower would still differ by 4906 units on average.

We can also compute it based on the R square value= 0.6049 which means that 60.49% of the total variability of mpg can be explained by a simple linear regression on horsepower.

iii.The value of the slope is -0.158 which means that the relationship between mpg and horsepower is negative. As the value of horsepower goes on to increase, the value of mpg decreases.

iv.The estimate of the slope is -0.158 which means that the value of mpg goes on to decrease by 0.158 units every year if the value of all the predictors stay constant.

v.

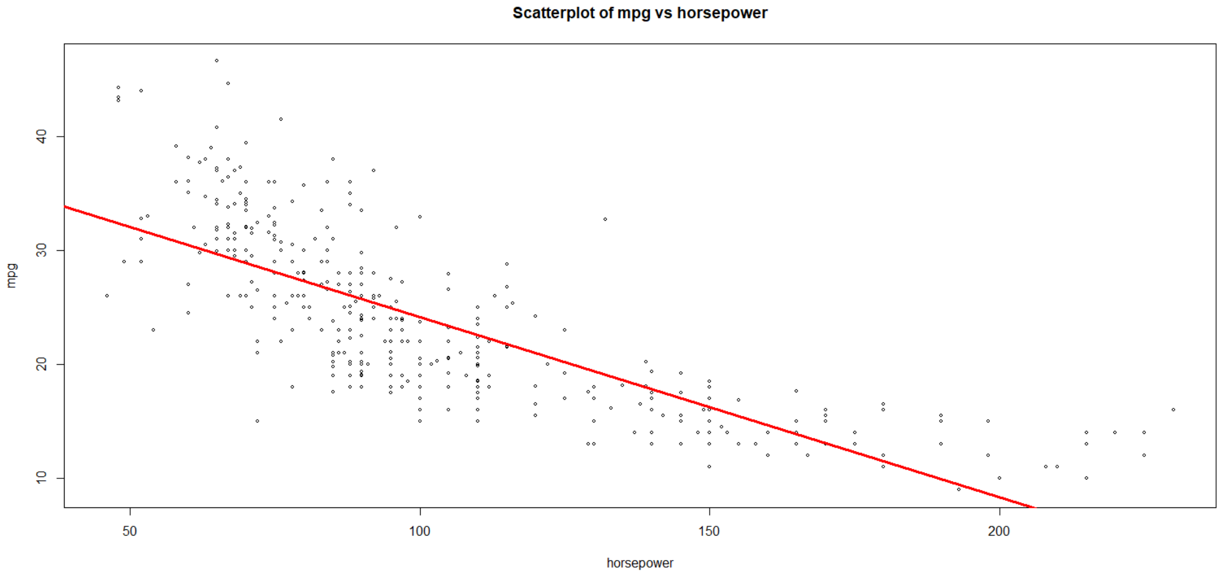
```
> predict(lm.fit,data.frame(horsepower=98),interval="confidence")
       fit      lwr      upr
1 24.46708 23.97308 24.96108
> predict(lm.fit,data.frame(horsepower=98),interval="prediction")
       fit     lwr      upr
1 24.46708 14.8094 34.12476
```

The predicted mpg with a horsepower of 98 was 24.46708

The 95% confidence interval has the lower limit of 23.97 and the upper limit of 24.96 where as the 95% prediction interval has the lower limit of 14.81 and the upper limit of 34.12

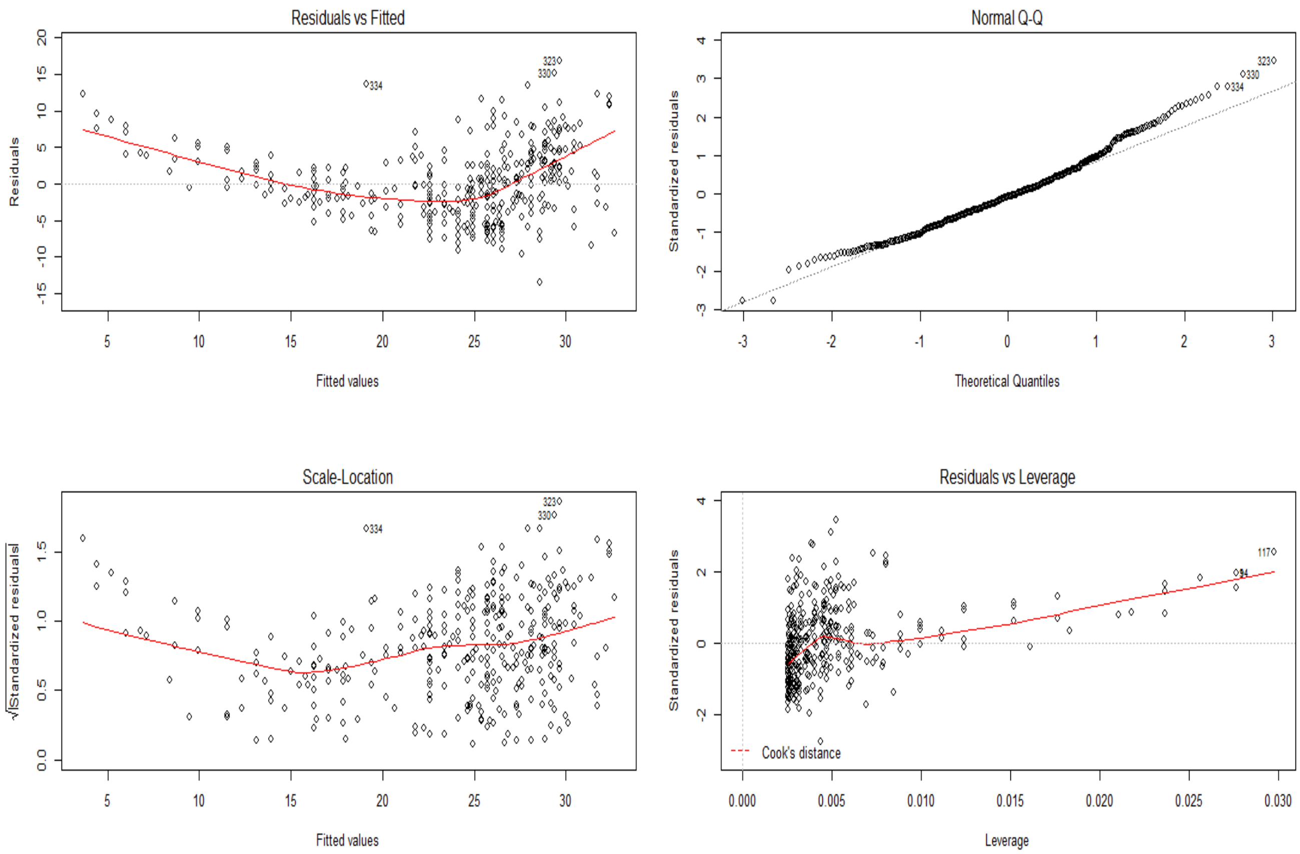
(b)

```
> plot(Auto$horsepower, Auto$mpg, main = "Scatterplot of mpg vs. horsepower", xl
ab = "horsepower", ylab = "mpg",cex=0.6)
> abline(lm.fit, col='red',lwd=3)
```



(c).

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



1. The residuals vs Fitted curve shows non-linearity in the data as the plot is not centred around zero and there is a chance of over-fitting the data. Therefore, this line is not a good fit.
2. The normal Q-Q curve shows a 45 degrees line between the theoretical quantities and the standardized responses except for a few outliers like 323, 330 and 334 but overall, this seems to be a better fit than the rest.
3. The scale-location graph should be straight to ensure that it is a good fit but here, that is not the case. Therefore, it is a bad fit.
4. Here point 117 is an outlier point. It can be removed to achieve a better fit.

Log(x):

```
> lm2.fit=lm(mpg~log(horsepower),data=Auto)
> summary(lm2.fit)

Call:
lm(formula = mpg ~ log(horsepower), data = Auto)

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(horsepower) -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

> 
```

(a). i.As the value of B1 is not zero, there exists a relationship between the predictor and the response.

ii.As we can see that the RSE is 4.501, this means that even if the model is correct and the values of B0 and B1 are known to us, but any prediction of mpg based on log(horsepower) would still differ by 4501 units on average.

We can also compute it based on the R square value= 0.6675 which means that 66.75% of the total variability of mpg can be explained by a simple linear regression on log(horsepower).

iii.The value of the slope is -18.58 which means that the relationship between mpg and horsepower is negative. As the value of log(horsepower) goes on to increase, the value of mpg decreases.

iv.The estimate of the slope is -18.58 which means that the value of mpg goes on to decrease by 18.58 units every year if the value of all the predictors stay constant.

v.

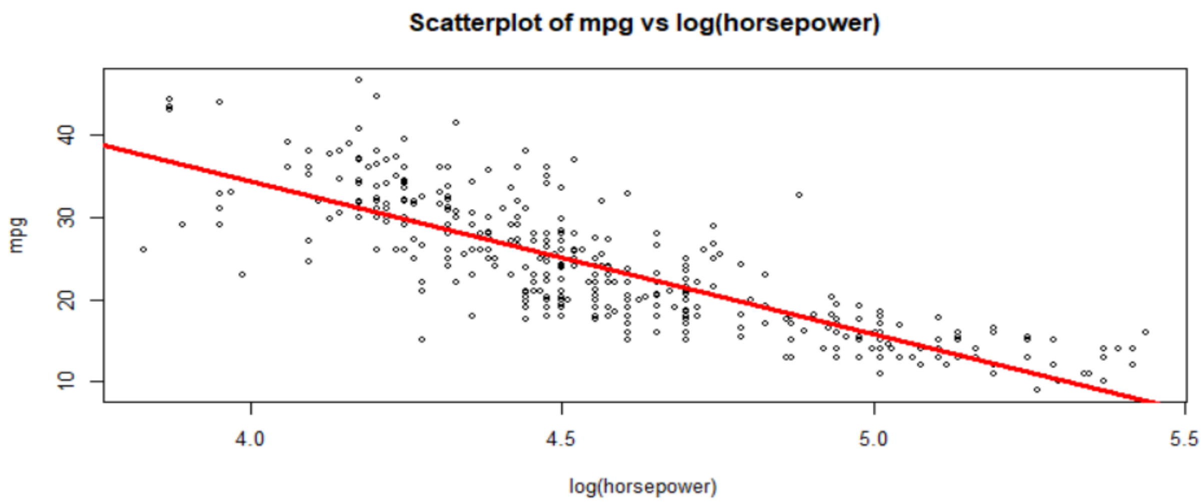
```
> predict(lm2.fit,data.frame(horsepower=98),interval="prediction")
       fit      lwr      upr
1 23.50099 14.64106 32.36093
> predict(lm2.fit,data.frame(horsepower=98),interval="confidence")
       fit      lwr      upr
1 23.50099 23.05405 23.94794
```

The predicted mpg with a horsepower of 98 was 23.50099

The 95% confidence interval has the lower limit of 23.05 and the upper limit of 23.94 where as the 95% prediction interval has the lower limit of 14.64 and the upper limit of 32.361

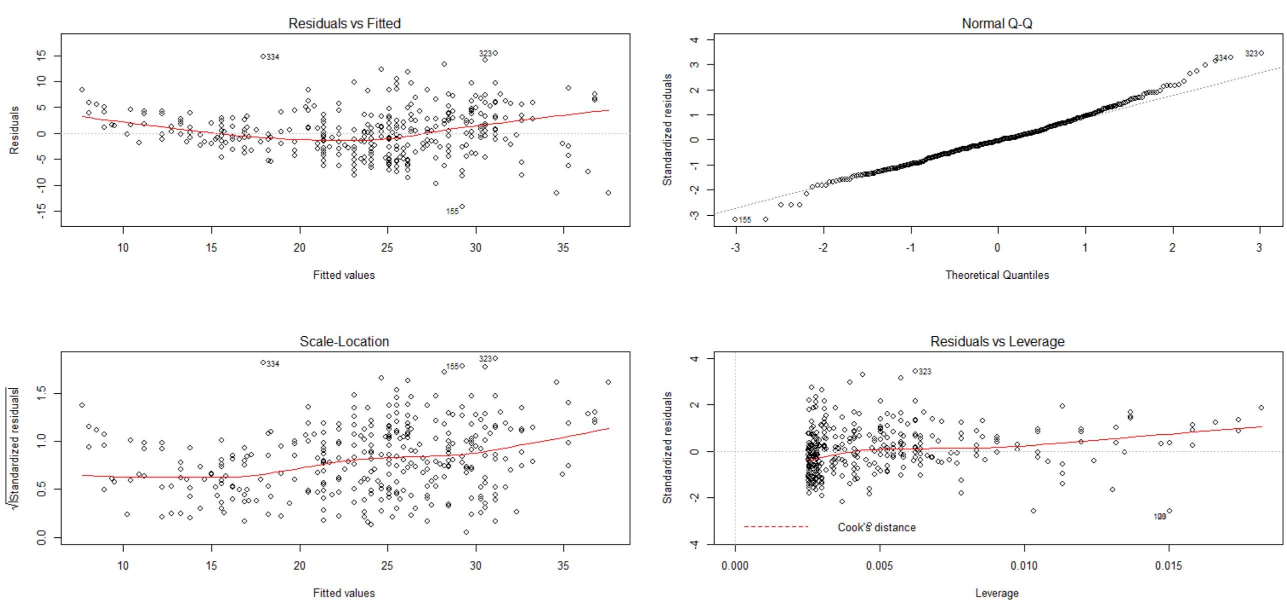
(b).

```
> plot(log(Auto$horsepower),Auto$mpg,main="Scatterplot of mpg vs log(horsepower)",xlab="log(horsepower)",ylab="mpg",cex=0.6)
> abline(lm2.fit,col="red",lwd=3)
```




```
> par(mfrow=c(2,2))
> plot(lm2.fit)
>
```

(c).



1. The residuals vs Fitted curve shows non-linearity in the data as the plot is not centred around zero and there is a chance of over-fitting the data. Therefore, this line is not a good fit.
2. The normal Q-Q curve shows a 45 degrees line between the theoretical quantities and the standardized responses except for a few outliers like 323, 155 and 334 but overall, this seems to be a better fit than the rest.
3. The scale-location graph should be straight to ensure that it is a good fit but here, that is not the case. Therefore, it is a bad fit.
4. Here point 323 is an outlier point. It can be removed to achieve a better fit.

Sqrt (X):

```
> lm3.fit=lm(mpg~sqrt(horsepower),data=Auto)
> plot(lm3.fit)
> summary(lm3.fit)

Call:
lm(formula = mpg ~ sqrt(horsepower), data = Auto)

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(horsepower)   -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
```

(a). i.As the value of B1 is not zero, there exists a relationship between the predictor and the response.

ii.As we can see that the RSE is 4.665, this means that even if the model is correct and the values of B0 and B1 are known to us, but any prediction of mpg based on log(horsepower) would still differ by 4665 units on average.

We can also compute it based on the R square value= 0.6428 which means that 66.75% of the total variability of mpg can be explained by a simple linear regression on log(horsepower).

iii.The value of the slope is -3.503 which means that the relationship between mpg and horsepower is negative. As the value of log(horsepower) goes on to increase, the value of mpg decreases.

iv.The estimate of the slope is -3.503 which means that the value of mpg goes on to decrease by 3.503 units every year if the value of all the predictors stay constant.

v.

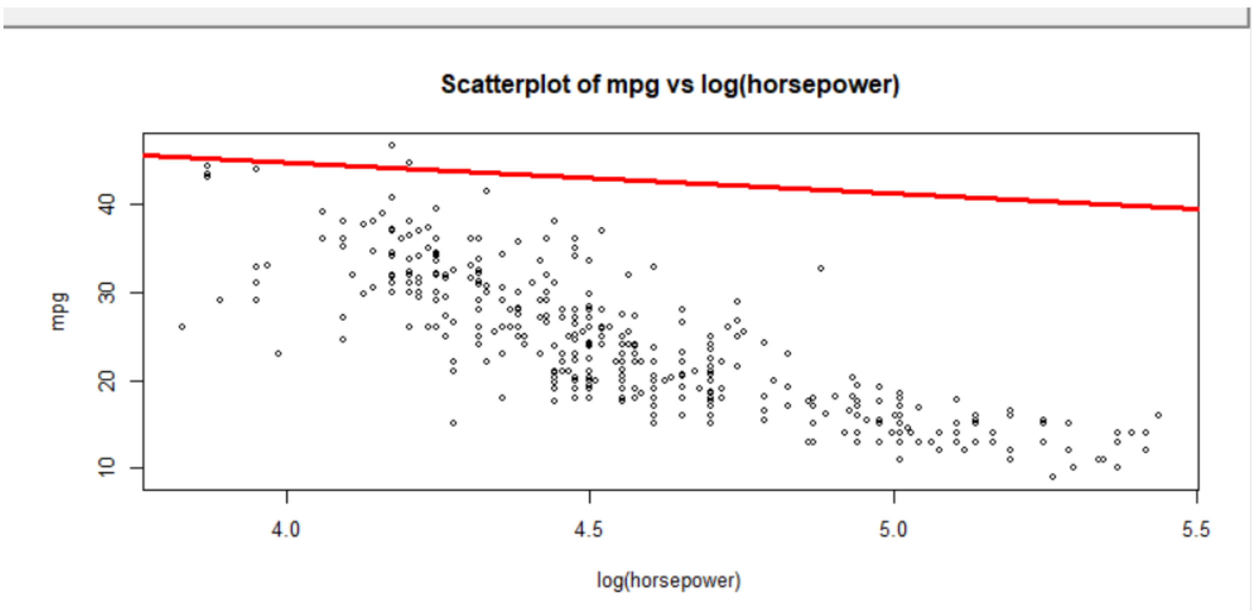
```
> predict(lm3.fit,data.frame(horsepower=98),interval="confidence")
       fit      lwr      upr
1 24.02206 23.55687 24.48724
> predict(lm3.fit,data.frame(horsepower=98),interval="prediction")
       fit      lwr      upr
1 24.02206 14.83892 33.20519
>
```

The predicted value of mpg for the model when horsepower is 98 is 24.022

The 95% confidence interval has the lower limit of 23.557 and the upper limit of 24.49 where as the 95% prediction interval has the lower limit of 14.84 and the upper limit of 33.20

(b).

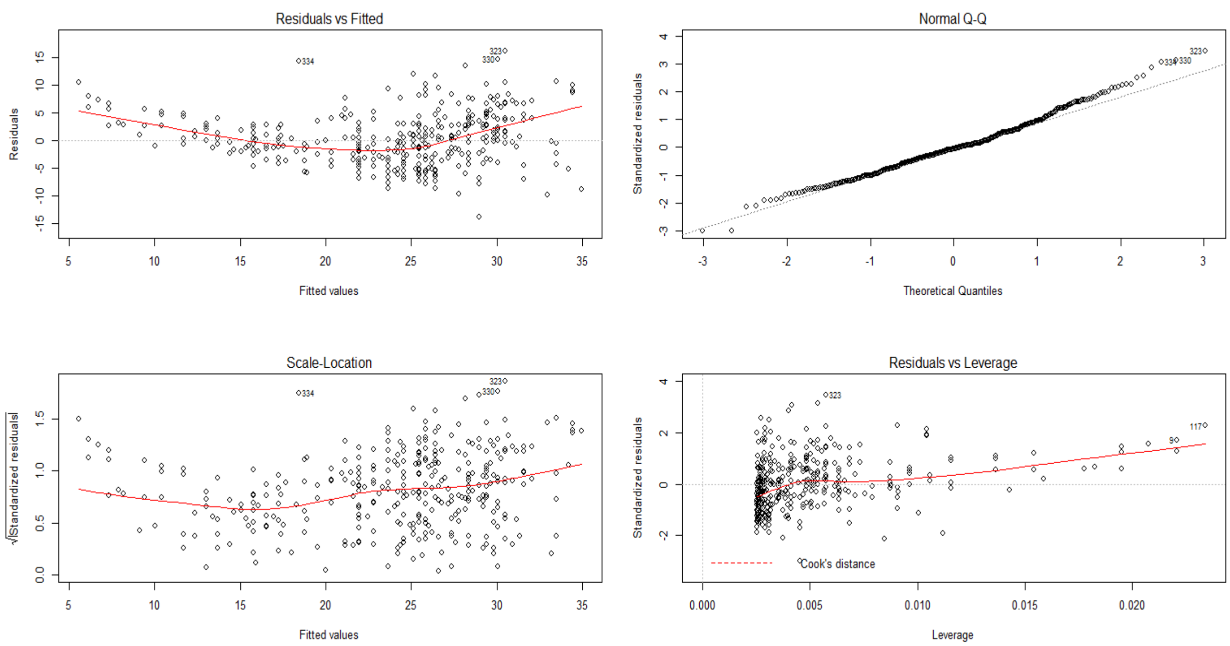
```
> plot(log(Auto$horsepower),Auto$mpg,main="Scatterplot of mpg vs log(horsepower)",xlab="log(horsepower)",ylab="mpg",cex=0.6)
> abline(lm3.fit,col="red",lwd=3)
```



The linear regression line between log(horsepower) and mpg does not encompass most of the points in the graph. Hence, it is not a good fit.

(c).

```
> par(mfrow=c(2,2))
> plot(lm3.fit)
> 
```



1. The residuals vs Fitted curve shows non-linearity in the data as the plot is not centred around zero and there is a chance of over-fitting the data. Therefore, this line is not a good fit.
2. The normal Q-Q curve shows a 45 degrees line between the theoretical quantities and the standardized responses except for a few outliers like 323, 330 and 334 but overall, this seems to be a better fit than the rest.
3. The scale-location graph should be straight to ensure that it is a good fit but here, that is not the case. Therefore, it is a bad fit.
4. Here point 323,117 and 90 are outlier points. It can be removed to achieve a better fit.

X^2 :

```
> lm4.fit=lm(mpg~I(horsepower^2),data=Auto)
> summary(lm4.fit)

Call:
lm(formula = mpg ~ I(horsepower^2), data = Auto)

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(horsepower^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
```

(a). i.As the value of B1 is not zero, there exists a relationship between the predictor and the response.

ii.As we can see that the RSE is 5.485, this means that even if the model is correct and the values of B0 and B1 are known to us, but any prediction of mpg based on log(horsepower) would still differ by 5485 units on average.

We can also compute it based on the R square value= 0.5061 which means that 50.61% of the total variability of mpg can be explained by a simple linear regression on log(horsepower).

iii.The value of the slope is -5.66e-04 which means that the relationship between mpg and horsepower is negative. As the value of log(horsepower) goes on to increase, the value of mpg decreases.

iv.The estimate of the slope is -5.66e-04 which means that the value of mpg goes on to decrease by 3.503 units every year if the value of all the predictors stay constant.

v.

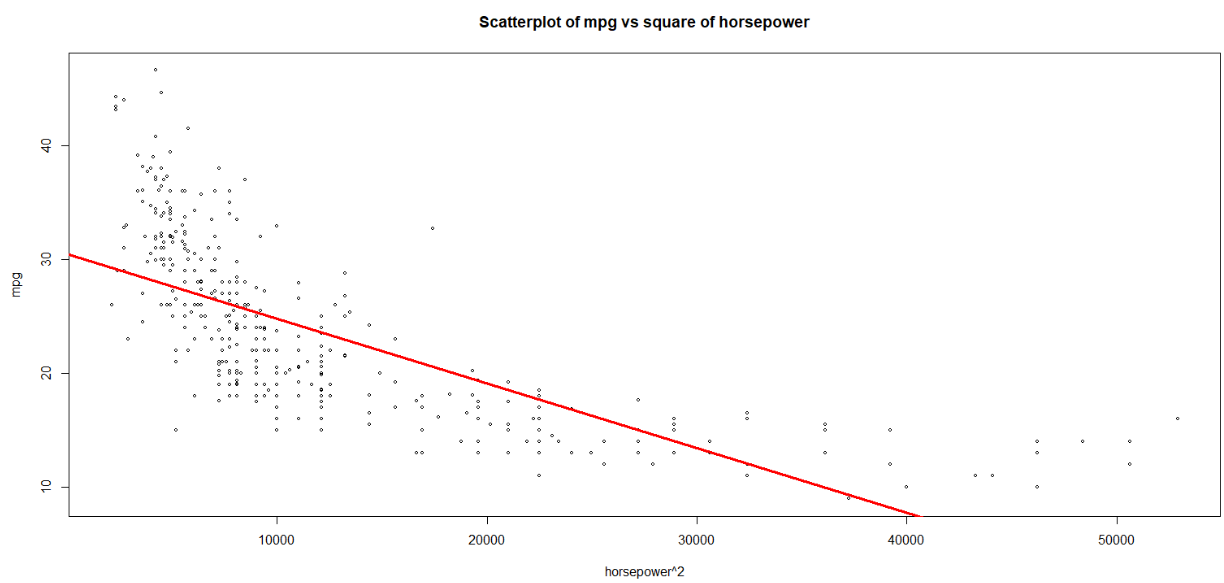
```
> predict(lm4.fit,data.frame(horsepower=98),interval="prediction")
       fit      lwr     upr
1 25.02512 14.22603 35.8242
> predict(lm4.fit,data.frame(horsepower=98),interval="confidence")
       fit      lwr     upr
1 25.02512 24.45883 25.5914
>
```

The predicted value of mpg for the model when horsepower is 98 is 25.02512

The 95% confidence interval has the lower limit of 24.459 and the upper limit of 25.591 where as the 95% prediction interval has the lower limit of 14.226 and the upper limit of 35.824

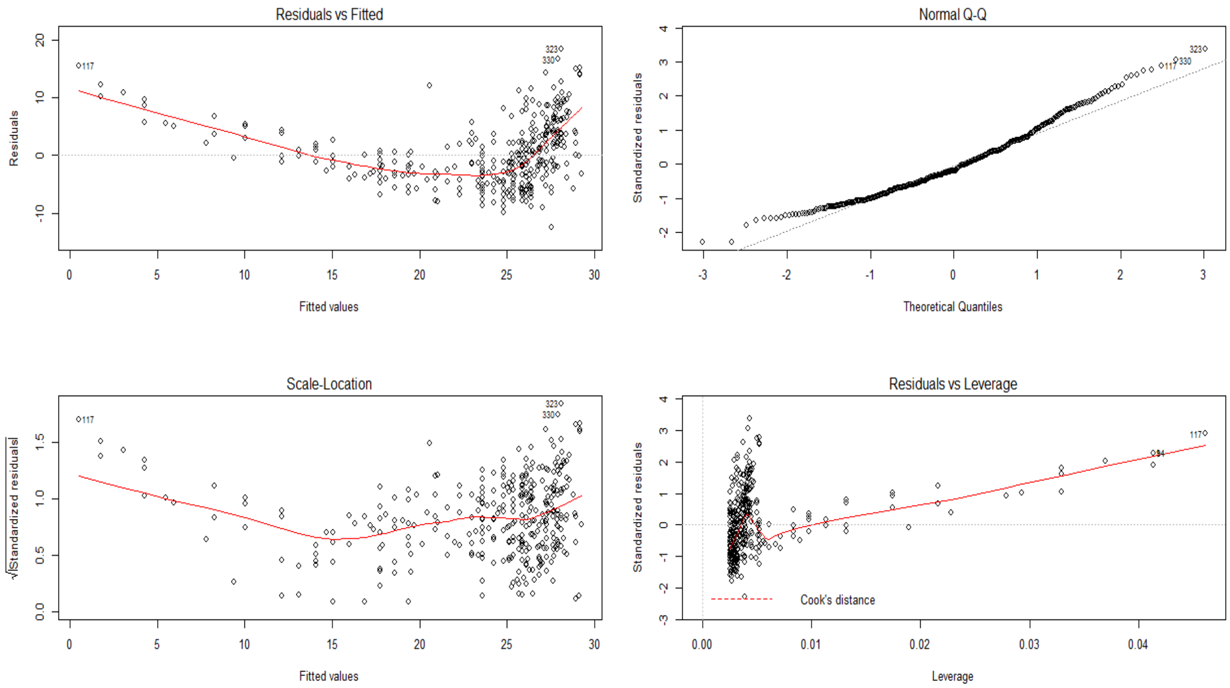
(b).

```
> plot((Auto$horsepower)^2,Auto$mpg,main="Scatterplot of mpg vs square of horsepower",xlab="horsepower^2",ylab="mpg",cex=0.6)
> abline(lm4.fit,col="red",lwd=3)
>
```



(c).

```
> par(mfrow=c(2,2))
> plot(lm4.fit)
>
```



1. The residuals vs Fitted curve shows non-linearity in the data as the plot is not centred around zero and there is a chance of over-fitting the data. Therefore, this line is not a good fit.
2. The normal Q-Q curve shows a 45 degrees line between the theoretical quantities and the standardized responses except for a few outliers at the end points but overall, this seems to be a better fit than the rest.
3. The scale-location graph should be straight to ensure that it is a good fit but here, that is not the case. Therefore, it is a bad fit.
4. Here point 117 is an outlier point. It can be removed to achieve a better fit.

Q2.

Problem 2

Use the Auto data set to answer the following questions:

(a) Produce a scatterplot matrix which includes all of the variables in the data set. Which predictors appear to have an association with the response?

(b) Compute the matrix of correlations between the variables (using the function cor()). You will need to exclude the name variable, which is qualitative.

(c) Perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Comment on the output. For example,

i. Is there a relationship between the predictors and the response?
ii. Which predictors have a statistically significant relationship to the response?
iii. What does the coefficient for the year variable suggest?

(d) Produce diagnostic plots of the linear regression fit. Comment on each plot.

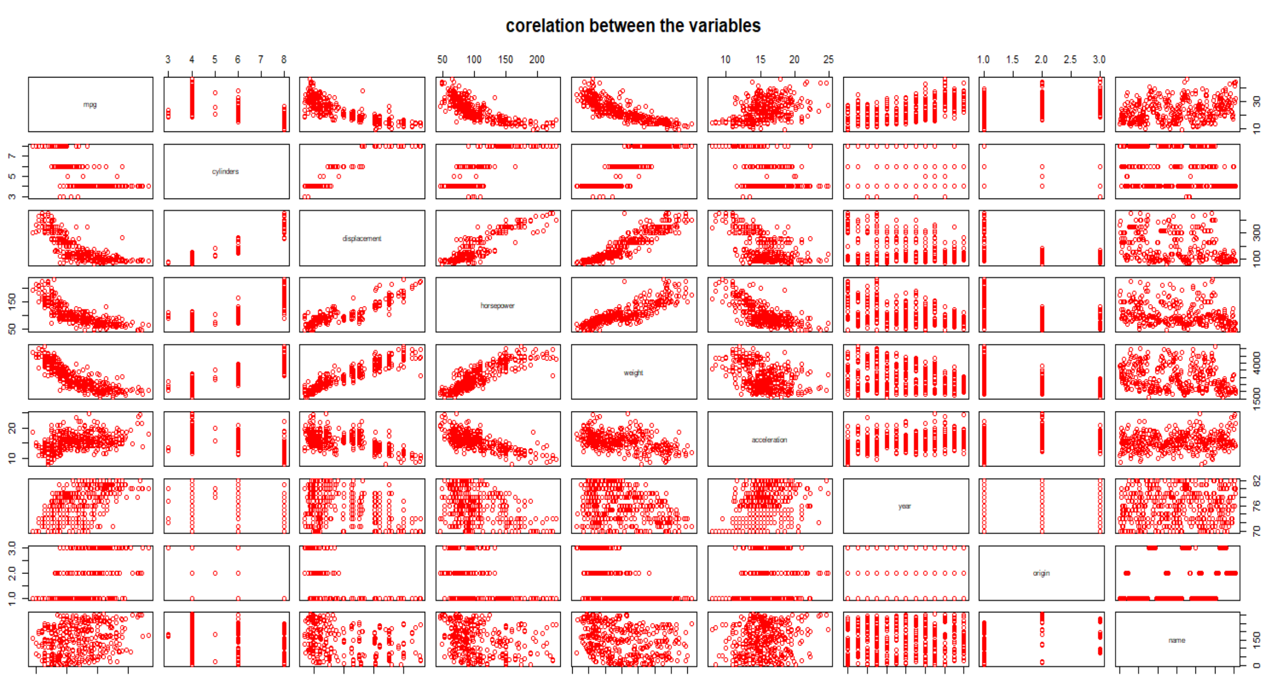
(e) Is there serious collinearity problem in the model? Which predictors are collinear?

(f) Fit linear regression models with interactions. Are any interactions statistically significant?

Ans 2.

(a).

```
> pairs(Auto,main="corelation between the variables",col="red")
>
```



Considering mpg as the predictor, the relation between the parameters can be stated below:

mpg and cylinders do not seem to have any correlation between them.

Mpg and displacement have negative correlation among them.

Mpg and horsepower have negative correlation among them.

Mpg and weight have negative correlation among them.

Mpg-acceleration, Mpg-year and Mpg-origin have positive correlation among them.

(b).

```
> cor(Auto[1:8])
                    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).

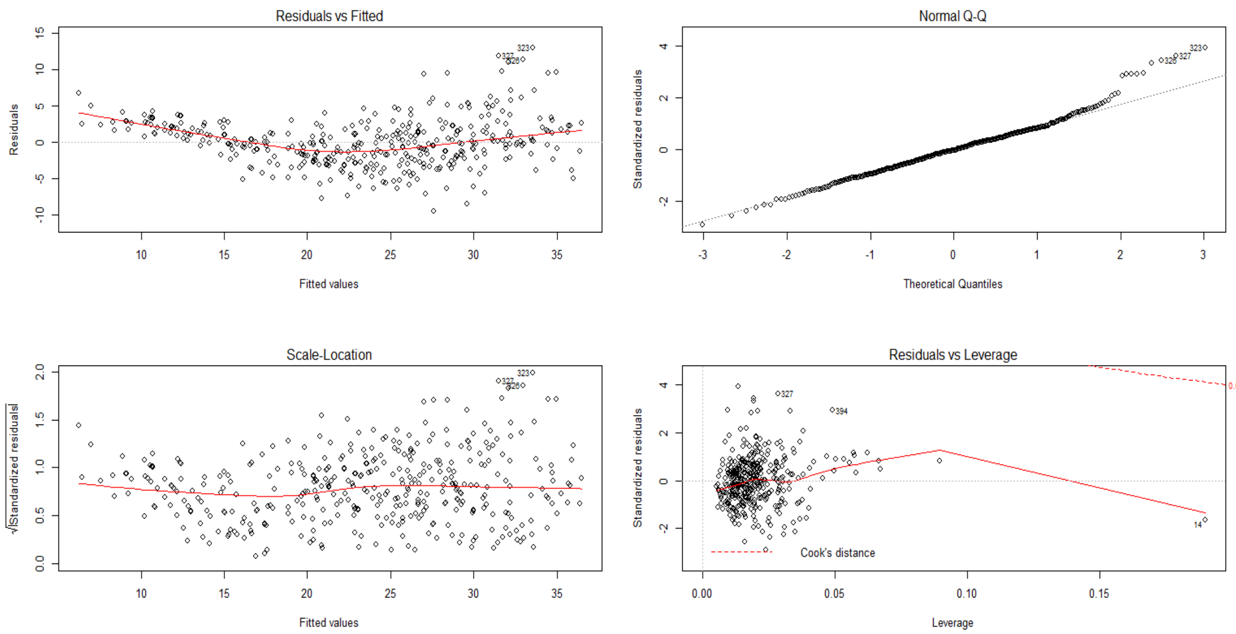
```
> lm.fit=lm(mpg~cylinders+displacement+horsepower+weight+acceleration+year+origin,data=Auto)
> summary(lm.fit)

Call:
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
    acceleration + year + origin, 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),(ii). mpg is negatively correlated with cylinders, horsepower and weight as we can see that the value of their slope is negative which we can see from the table we have generated. Both horsepower and weight have slope coefficients which even though being negative are not significant which means that although there is a negative correlation between them but still, it may not have a large impact on the predictor mpg.

Displacement,acceleration,year and origin are positively correlated with mpg as we can notice from their slope coefficients. Year and origin do have a slope estimate which is considerably large which shows that there is a high positive correlation between them and mpg.

(iii). The coefficient of slope for the year variable is +0.75 which means that the value of mpg increases by 0.75 units per year if all other predictors are kept constant.

(d).



For the Residual vs Fitted graph, most of the points seem to be centred around zero except a few points which makes it a good fit.

For the normal Q-Q graph, all the points must lie on the 45 degrees line and in our case, it is the case except a few outliers like 323 and 327 but overall, it seems to be a decent fit.

For the scale-location graph, most of the points must lie on the horizontal line and in this case too, it is the case. There by, we can conclude that it is a good fit.

For the residuals vs leverage graph, they show a high leverage point above except for a few outlier points above +2.

(e).

```
> library(car)
> library(ISLR)
> fix(Auto)
> multiplelr=lm(mpg~.-name,data=Auto)
> vif(multiplelr)
   cylinders displacement   horsepower       weight acceleration         year
   10.737535    21.836792     9.943693    10.831260     2.625806     1.244952
      origin
    1.772386
> multiplelr1=lm(mpg~.-name-displacement,data=Auto)
> vif(multiplelr1)
   cylinders   horsepower       weight acceleration         year       origin
    6.008253     9.088413     9.219674     2.598356     1.239409     1.594220
> multiplelr2=lm(mpg~.-name-displacement-weight,data=Auto)
> vif(multiplelr2)
   cylinders   horsepower acceleration         year       origin
    4.155143     5.323311     1.996560     1.209909     1.495100
> multiplelr3=lm(mpg~.-name-displacement-weight-horsepower,data=Auto)
> vif(multiplelr3)
   cylinders acceleration         year       origin
    1.999959     1.384478     1.159429     1.495041
> 
```

For the first case, the variance inflation factor of cylinders,displacement,horsepower and weight is more than 5. It states that all of these predictors are seriously collinear and these predictors can be dropped to better the model.

For the second case, the variance inflation factor of cylinders, horsepower and weight is more than 5.It suggests that we can still better our model by dropping a few of them.

For the third case, the variance inflation factor of horsepower is more than 5. So, we can try some more permutations with our model to try avoiding serious collinearity in the model.

For the fourth case, we can see that all the predictors have VIF more than 1 but very much within 5. Thereby, we can use this model.

(f).Trying to fit different linear regression models to select the best model, we try out different models by modelling the predictors:

1st case:

```
> lm2.fit=lm(mpg~log(cylinders)+log(weight)+acceleration+origin+year,data=Auto)
> summary(lm2.fit)

Call:
lm(formula = mpg ~ log(cylinders) + log(weight) + acceleration +
    origin + year, data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4638 -1.9754 -0.0177  1.6941 12.9668

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    116.42995    9.29955  12.520  < 2e-16 ***
log(cylinders)   0.80572    1.18663   0.679  0.49755
log(weight)    -19.55917    1.23806 -15.798  < 2e-16 ***
acceleration     0.09086    0.06630   1.370  0.17135
origin           0.78764    0.24984   3.153  0.00174 **
year             0.77264    0.04612  16.755  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.124 on 386 degrees of freedom
Multiple R-squared:  0.8419,    Adjusted R-squared:  0.8398
F-statistic: 411.1 on 5 and 386 DF,  p-value: < 2.2e-16

> vif(lm2.fit)
log(cylinders)    log(weight)   acceleration         origin           year
      5.155083       4.857121       1.340837       1.623089       1.156489
```

In this case, even though the adjusted R-square value is high which shows that 83.9% variability in mpg is dependent on the predictors but since here the VIF of log(cylinders) is more than 5 , we can reject this model since it shows high collinearity.

2nd case:

```
> lm1.fit=lm(mpg~I(cylinders^2)+log(weight)+acceleration+origin+year,data=Auto)
> summary(lm1.fit)

Call:
lm(formula = mpg ~ I(cylinders^2) + log(weight) + acceleration +
    origin + year, data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4492 -1.9987 -0.0298  1.6959 12.8766

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    122.07612   10.01324  12.191  < 2e-16 ***
I(cylinders^2)   0.02600    0.01795   1.448  0.14838
log(weight)    -20.31303    1.23976 -16.385  < 2e-16 ***
acceleration     0.11551    0.06860   1.684  0.09301 .
origin           0.78885    0.24754   3.187  0.00156 **
year             0.77853    0.04613  16.878  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.117 on 386 degrees of freedom
Multiple R-squared:  0.8426,    Adjusted R-squared:  0.8405
F-statistic: 413.1 on 5 and 386 DF,  p-value: < 2.2e-16

> vif(lm1.fit)
I(cylinders^2)    log(weight)   acceleration         origin           year
      5.299736       4.891138       1.441339       1.600125       1.161998
```

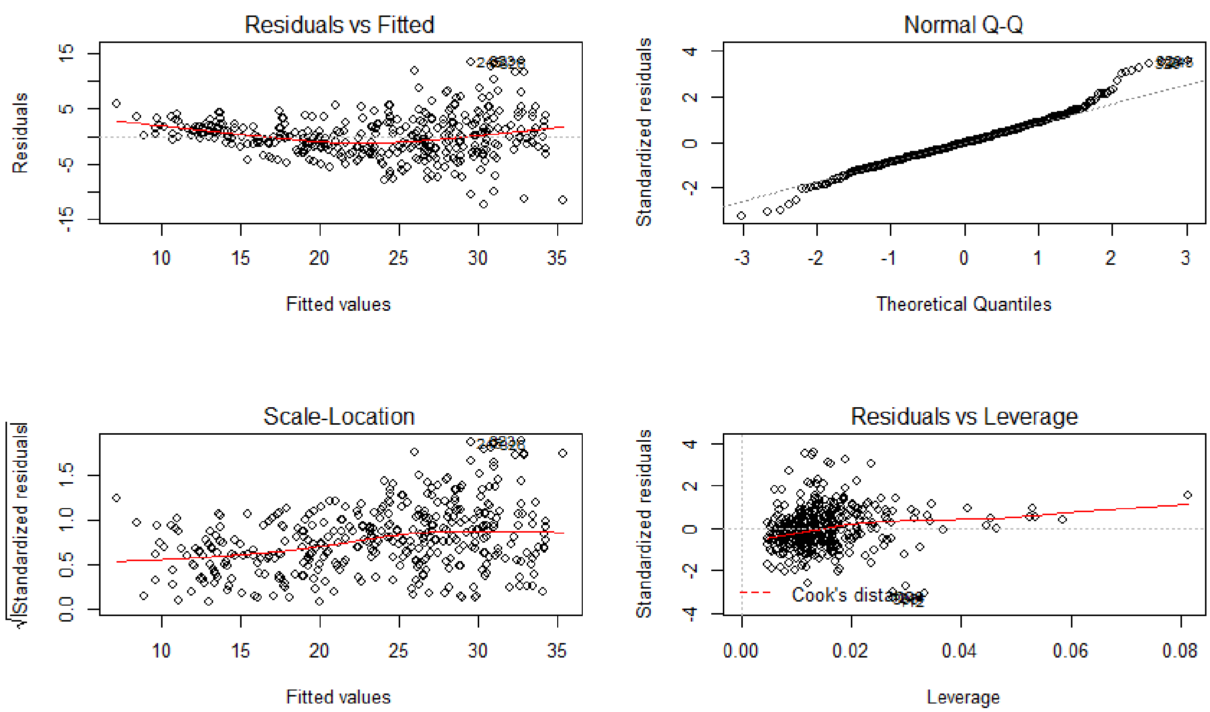
In the second case, the adjusted R square value was found to be 0.84 which goes on to explain that 84% variability in mpg can be explained by linear regression on these predictors but as the VIF of I(cylinders^2) and log(weight) is close to 5, it shows that the predictors are highly collinear so we tend to ignore this model and we try for the third case.

Final model:

```
> lm.fit=lm(mpg~log(cylinders)+I(weight^4)+log(acceleration)+origin+year,data=Auto)
> summary(lm.fit)

Call:
lm(formula = mpg ~ log(cylinders) + I(weight^4) + log(acceleration) +
    origin + year, data = Auto)

Residuals:
     Min       1Q   Median       3Q      Max
-12.4284  -2.2958  -0.0661   2.0919  13.8620

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.697e+01  6.054e+00  -2.803  0.00531 **
log(cylinders)    -9.686e+00  1.257e+00  -7.707 1.10e-13 ***
I(weight^4)       -1.306e-14  2.513e-15  -5.199 3.25e-07 ***
log(acceleration) -1.923e-01  1.283e+00  -0.150  0.88090
origin             1.723e+00  3.008e-01   5.728 2.04e-08 ***
year               7.348e-01  5.775e-02  12.722  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.877 on 386 degrees of freedom
Multiple R-squared:  0.7564,    Adjusted R-squared:  0.7532
F-statistic: 239.7 on 5 and 386 DF,  p-value: < 2.2e-16

> vif(lm.fit)
   log(cylinders)       I(weight^4) log(acceleration)            origin              year
         3.752406          2.995353          1.401453          1.526557          1.177176
```

In this case, the adjusted R-square value was found to be 0.75 which shows that 75% variability in response mpg can be explained by a linear regression on these predictors which is significant and can make a good model. Baring it, the VIF values of all the predictors all lie between 1 and 5, thus stating that all the predictors are perfect collinear. Therefore, we consider this as our final model.

Producing diagnostic plots of this model, we get the following curves:



For the Residuals vs Fitted graph, most of the points must be centered around zero. In our case, baring point 262 which is an outlier point , all other points are more or less centered around zero which makes it a good fit.

For the Normal Q-Q graph, all the points must lie on the 45 degree line. Here, 70 % of the points lie on the line baring a few points in the end thus making it a decent fit.

For the scale-location graph, all the points must lie on the straight horizontal lie for the model to be considered a perfect fit. In our case, baring a few outliers like point 283 all others lie on the line. Thus, this makes the scale-location graph a good fit.

For the Residuals vs Leverage graph, baring a single leverage point and a few outliers, the model is considerably good.

Problem 3

Use the Carseats data set to answer the following questions:

(a) Fit a multiple regression model to predict Sales using Price, Urban, and US.

(b) Provide an interpretation of each coefficient in the model (note: some of the variables are qualitative).

(c) Write out the model in equation form.

(d) For which of the predictors can you reject the null hypothesis $H_0: \beta_j = 0$?

(e) On the basis of your answer to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the response.

(f) How well do the models in (a) and (e) fit the data?

(g) Is there evidence of outliers or high leverage observations in the model from (e)?

Ans 3:

(a).

```
> library(ISLR)
> data(Carseats)
> head(Carseats)
  Sales CompPrice Income Advertising Population Price ShelveLoc Age Education Urban  US
1  9.50       138     73          11        276   120       Bad  42        17   Yes Yes
2 11.22       111     48          16        260    83      Good  65        10   Yes Yes
3 10.06       113     35          10        269    80    Medium  59        12   Yes Yes
4  7.40       117    100           4        466    97    Medium  55        14   Yes Yes
5  4.15       141     64           3        340   128       Bad  38        13   Yes  No
6 10.81       124    113          13        501    72       Bad  78        16    No Yes
> lm.fit=lm(Sales~Price+Urban+US,data=Carseats)
> summary(lm.fit)

Call:
lm(formula = Sales ~ Price + Urban + US, data = Carseats)

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 intercept value gives the average value of sales which is 13.04 even though the value of all other predictors are zero. The value of the slope for price is -0.05 which shows that although not that significant, but there is a negative correlation between price and sales.

Here, both UrbanYes and USYes are qualitative variables which means that their corresponding slopes are valid if and only if the values of Urban and US are true. For the value of UrbanYes, the value of the slope is -0.02 which means that there is a negative correlation between sales and Urban where as the for the value of USYes, the value of slope is 1.2 which shows that there is a positive linear correlation between sales and US if their values are true.

(c).

$$\hat{y} = \beta_0 + \beta_1 \times price + \beta_2 \times Urban + \beta_3 \times US$$
$$= 13.043 - 0.054 \times price - 0.022 \times Urban + 1.2 \times US$$

Here, both Urban and US are qualitative variables.

If both the values are true, then the above equation holds true.

If Urban=False and US=True, then the equation can be modelled as:

$$\hat{y} = 13.043 - 0.054 \times price + 1.2 \times US$$

If Urban=True and US=False, then the equation can be modelled as:

$$\hat{y} = 13.043 - 0.054 \times price - 0.022 \times Urban$$

If both are False, then the equation is modelled as:

$$\hat{y} = 13.043 - 0.054 \times price$$

(d).

```
> lm.fit=lm(Sales~.,data=Carseats)
> summary(lm.fit)

Call:
lm(formula = Sales ~ ., data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8692 -0.6908  0.0211  0.6636  3.4115 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      5.6606231  0.6034487   9.380  < 2e-16 ***
CompPrice        0.0928153  0.0041477  22.378  < 2e-16 ***
Income           0.0158028  0.0018451   8.565 2.58e-16 ***
Advertising      0.1230951  0.0111237  11.066  < 2e-16 ***
Population       0.0002079  0.0003705   0.561    0.575    
Price           -0.0953579  0.0026711 -35.700  < 2e-16 ***
ShelveLocGood    4.8501827  0.1531100  31.678  < 2e-16 ***
ShelveLocMedium  1.9567148  0.1261056  15.516  < 2e-16 ***
Age             -0.0460452  0.0031817 -14.472  < 2e-16 ***
Education       -0.0211018  0.0197205  -1.070    0.285    
UrbanYes         0.1228864  0.1129761   1.088    0.277    
USYes           -0.1840928  0.1498423  -1.229    0.220    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.019 on 388 degrees of freedom
Multiple R-squared:  0.8734,	Adjusted R-squared:  0.8698 
F-statistic: 243.4 on 11 and 388 DF,  p-value: < 2.2e-16
```

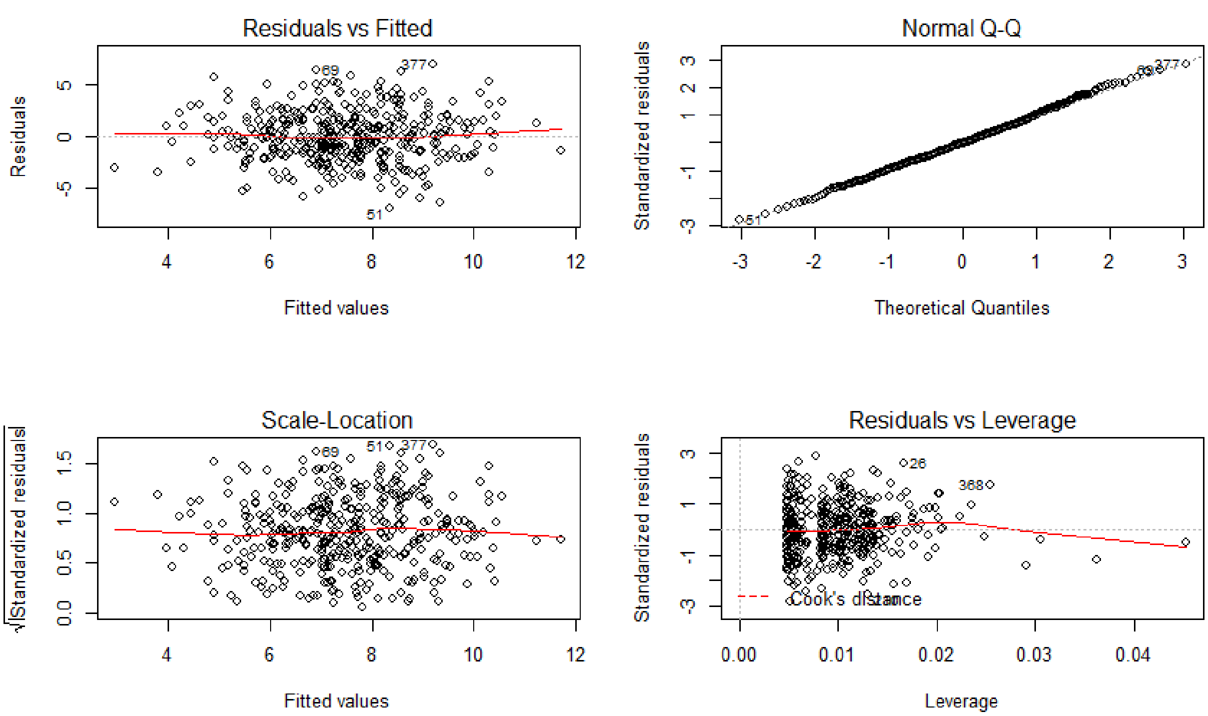
Here, we consider the value of alpha for all the predictors to be 0.05

So, checking for the p-values, the p-values of CompPrice, Income,advertising,price,ShelveLoc and Age are less than 2e-16 which is less than the stipulated value of alpha which we have considered for our model. Therefore, we reject the null hypothesis for these predictors.

(e).

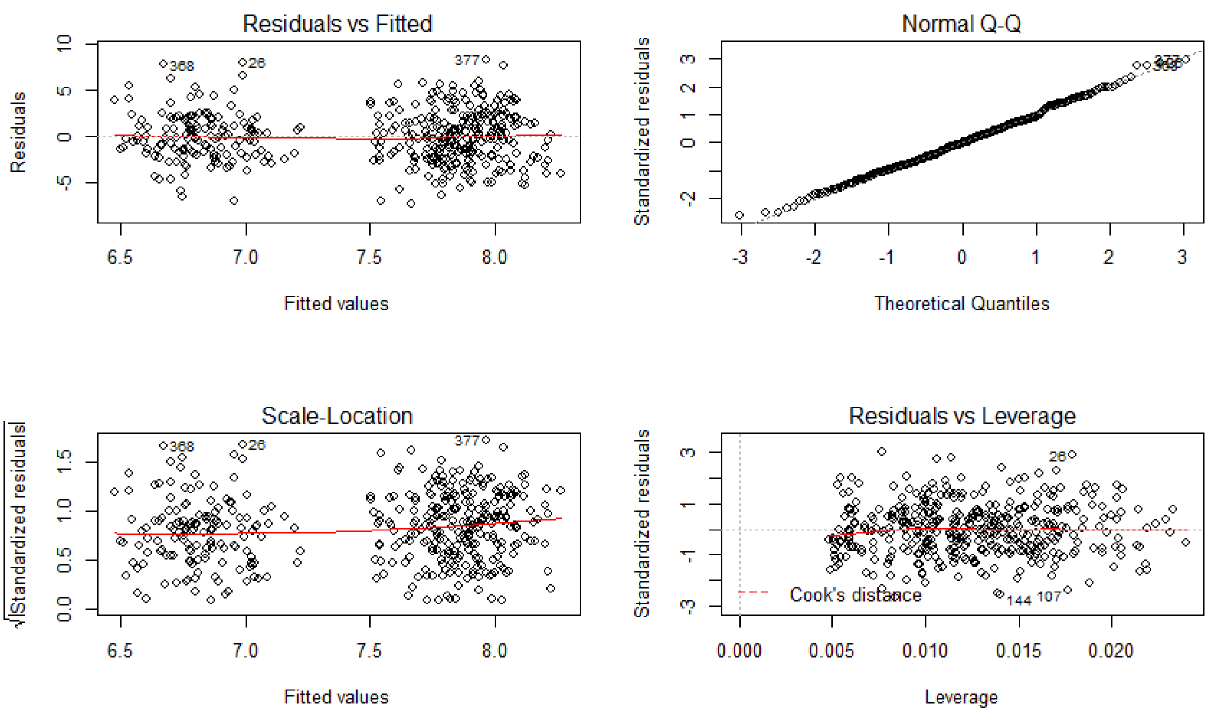
```
> lm1.fit=lm(Sales~Population+Education+Urban+US,data=Carseats)
> summary(lm1.fit)

Call:
lm(formula = Sales ~ Population + Education + Urban + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2945 -1.9766 -0.0256  1.8398  8.3058

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.2905436  0.8900461   8.191 3.63e-15 ***
Population   0.0006710  0.0009557   0.702 0.483040
Education   -0.0381826  0.0537553  -0.710 0.477935
UrbanYes    -0.1418150  0.3067541  -0.462 0.644115
USYes        1.0213930  0.2930368   3.486 0.000546 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.789 on 395 degrees of freedom
Multiple R-squared:  0.03465,	Adjusted R-squared:  0.02487
F-statistic: 3.544 on 4 and 395 DF,  p-value: 0.007411
```

(f).



From the first model, we can see that:

1. The residuals vs Fitted curve shows linearity in the data as the plot is more or less centred around zero. Therefore, this line is a good fit.
2. The normal Q-Q curve shows a 45 degrees line between the theoretical quantities and the standardized responses except point 377 and 51. Thereby, making it a good fit.
3. The scale-location graph should be straight and in our case, all the points are uniformly spread thereby indicating that it is a decent fit.
4. Here, the graph shows a few outlier points like 26 and 368. It can be removed to achieve a good fit.



From the second model, we can see that:

1. The residuals vs Fitted curve shows linearity in the data as the plot is more or less centred around zero except a few outliers like point 368,26 and 377. Therefore, this line is a good fit.
2. The normal Q-Q curve shows a 45 degrees line between the theoretical quantities and the standardized responses. Thereby, making it a good fit.
3. The scale-location graph should be straight and in our case, all the points are not uniformly spread instead, these points are in the form of clusters. Therefore, we can state that the scale-location graph is not a good fit.
4. Here, the graph shows a few outlier points like 26,144 and 107. It can be removed to achieve a good fit.

(g). In (e), we can observe that although there are a few outliers but there is no leverage point as all the points are uniformly spread throughout.