Assignment #2: Linear Regression

SANDEEP DASARI **UIN - 829002252**

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.

```
> library("ISLR")
> attach(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
```

i. **Is there a relationship between the predictor and the response?**
Yes, there is a relationship between the predictor and the response. We can answer this question by testing the null hypothesis(βi=0). But the p-value corresponding to the F-statistic is 2.2e-16, which is very small. This shows that there is a relationship between mpg and horsepower. If the p-value is small, then we reject the null and conclude that βi≠0.

ii. **How strong is the relationship between the predictor and the response?**
Because the R^2 is 0.6059, it means approximately 60.59% of the variability in mpg can be explained using horsepower. Mean of mpg is 23.4459184. RSE of lm.fit is 4.906 which shows that there is error of 20.9237141%.

iii. **Is the relationship between the predictor and the response positive or negative?**
The relation between predictor and response is negative because the slope or (horsepower coefficient) is negative.

iv. **How to interpret the estimate of the slope?**
The slope is -0.1578, it is negative. Therefore, it means if horsepower increases mpg will decrease.
It basically means that for every 100 units increase in horsepower the mpg decreases by 15 units.

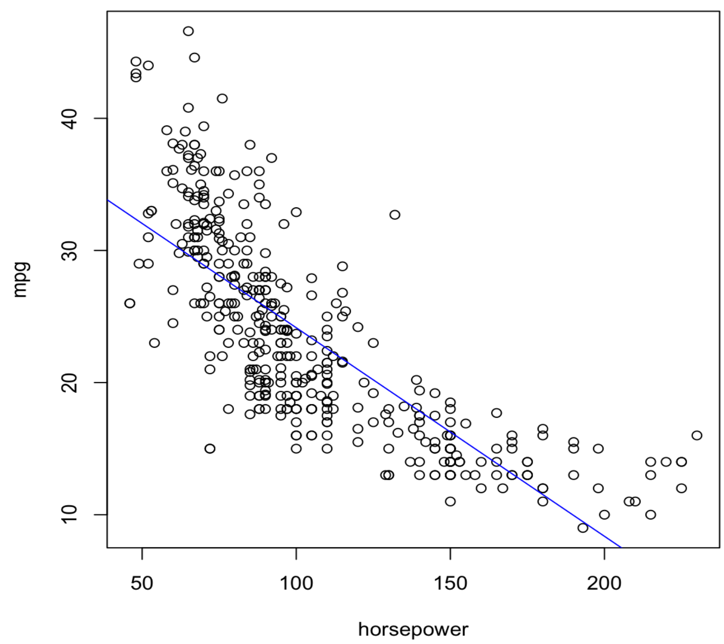
v. **What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?**

*The fit can be described as: <u>mpg= 39.9358-0.1578*hp</u>*
The predicted value of mpg for horsepower of 98 is 24.46. Lower limit of confidence interval us 23.97, upper limit is 24.96. Lower limit for prediction interval is 14.80 and upper limit is 34.12.

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

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

```
> plot(horsepower,mpg)
> abline(lm.fit,col='blue')
```

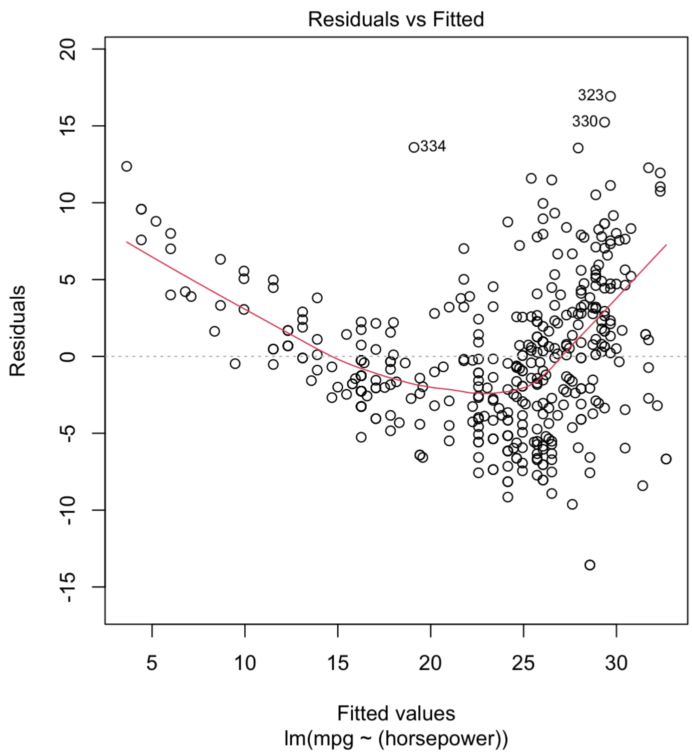


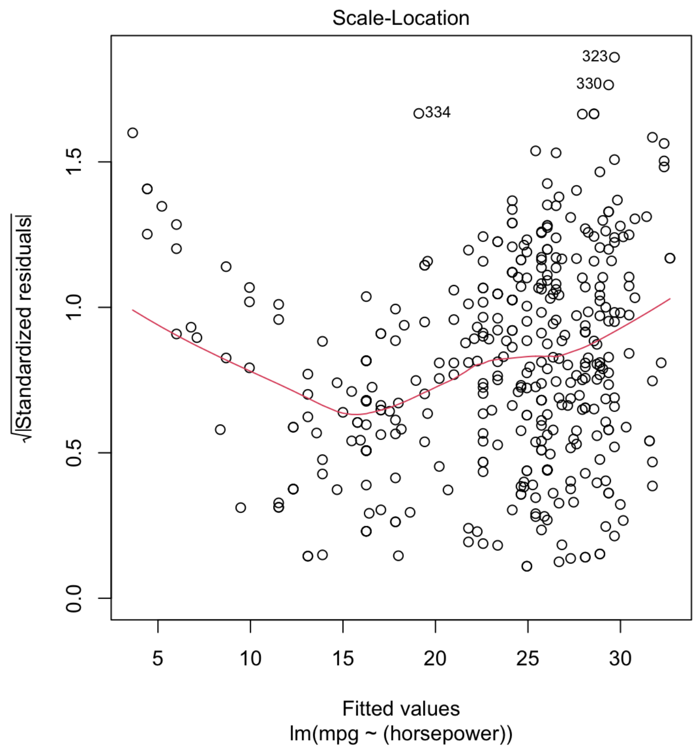
(b) Produce the diagnostic plots of the least squares regression fit. Comment on each plot.

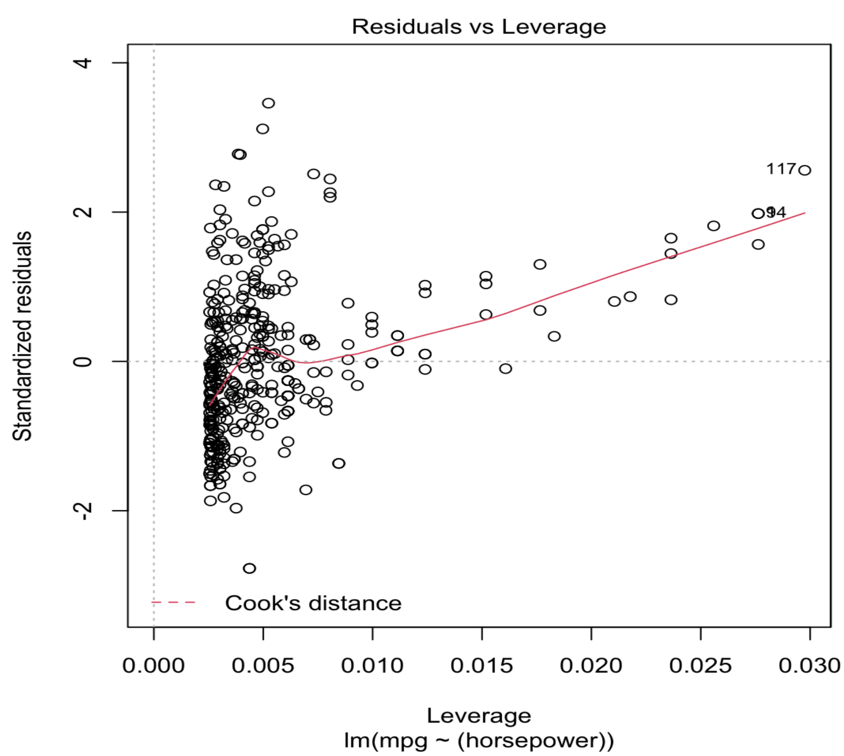
```
> plot(lm.fit, which=1)
> plot(lm.fit, which=3)
> plot(lm.fit, which=5)
```

- *The plot between residuals and fitted values clearly shows there is non-linearity in relationship between response and predictor. There is also a funneling pattern in the plot, which shows the plot also exhibits heteroscedasticity..*
- *The 2nd plot shows there is non-constant variance of errors also known as heteroscedasticity in the model.*
- *As we do not have any points beyond the boundary, there are not many high influential points.*

Point 117 looks like a high leverage point since it has very high leverage value.







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

Log(X) has the best adj R^2 value of 0.6675. Even $\sqrt{\boldsymbol{X}}$ has a better R^2 *value of 0.6428. Meaning they both fit the data better than X. On the other side X^2 has the least R^2 value of 0.5061.*
From the diagnostic plots we can see that there is very high non-linearity while using X^2 as predictor. The non-linearity is better for $\sqrt{X}$, and best for log(X).

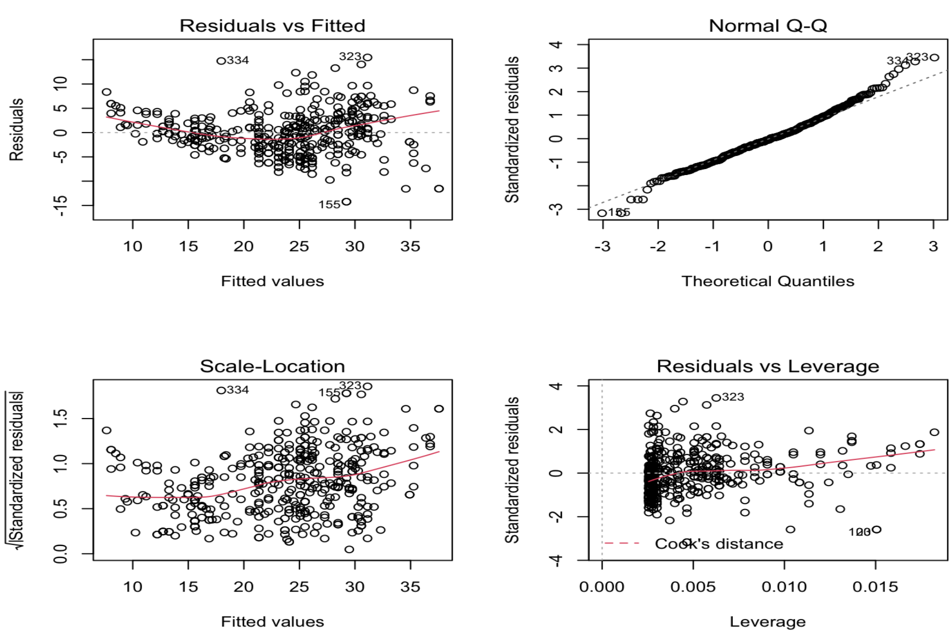
Log(X)

```
> lm.fit = lm(mpg~log(horsepower),data= Auto)
> summary(lm.fit)
> par(mfrow = c(2, 2))
> plot(lm.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-1
```



$\sqrt{\boldsymbol{X}}$

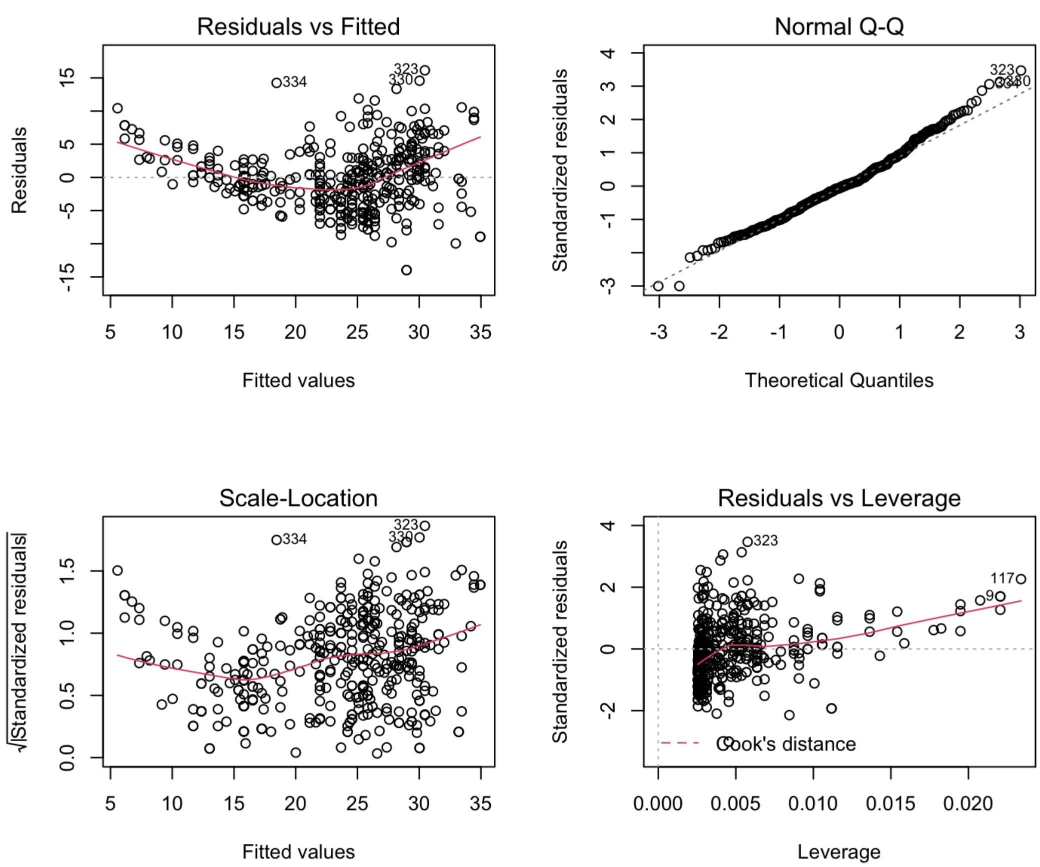
```
> lm.fit = lm(mpg~sqrt(horsepower),data= Auto)
```

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



X^2

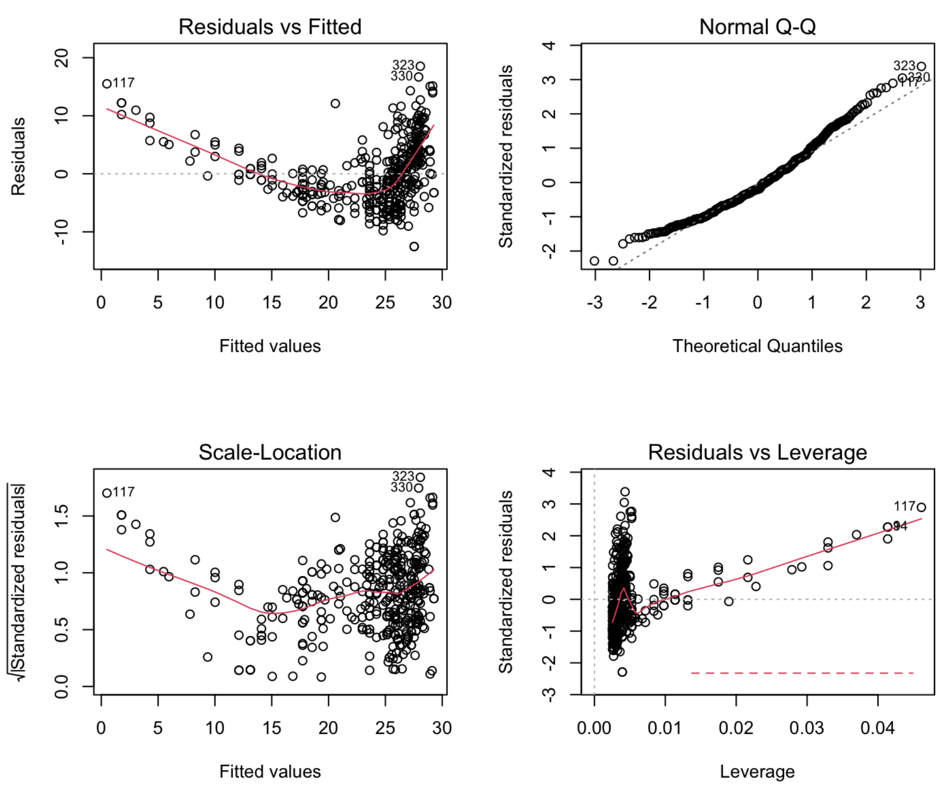
```
> horse_sq =  horsepower^2
```

```
> lm.fit2 = lm(mpg~(horse_sq),data= Auto)
> summary(lm.fit2)
> par(mfrow = c(2, 2))
> plot(lm.fit2)
Call:
lm(formula = mpg ~ (horse_sq), 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 ***
horse_sq    -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
```

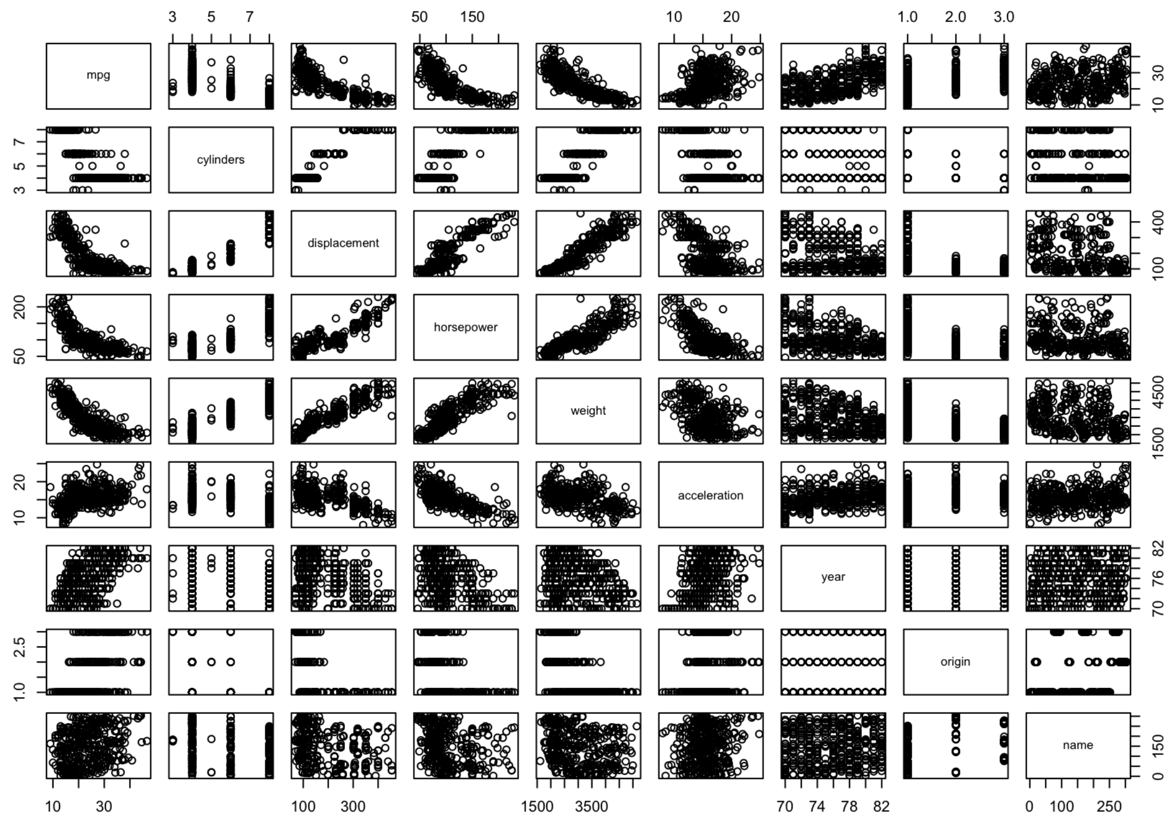


Problem 2

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?

If *we check for plots of mpg vs others, cylinders, displacement, horsepower and weight seem to have better relation with the response. Rest of the variables; acceleration, year and origin do not seem to have a strong 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.

> cor(Auto[,c(1,2,3,4,5,6,7,8)])

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

(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,

```
> lm2.fit = lm(mpg~ cylinders+displacement+horsepower+weight+acceleration+year+origin, data= Auto)
> summary(lm2.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. Is there a relationship between the predictors and the response?

There is a relationship between the predictors and the response(mpg). We can test the null hypothesis of whether all the regression coefficients are zero. The F-statistic (252.4) is far from 1, with very small p-value, indicating evidence against the null hypothesis. Hence, we can reject the null hypothesis and be sure that there is a relationship between predictor and response.

ii. Which predictors have a statistically significant relationship to the response?

The p-values of displacement, weight, year and origin are very small and also coefficients are large compared to their standard errors. Hence, these four predictors have a statistically significant relationship with the response. Whereas cylinder, horsepower and acceleration are not significant.

iii. What does the coefficient for the year variable suggest?

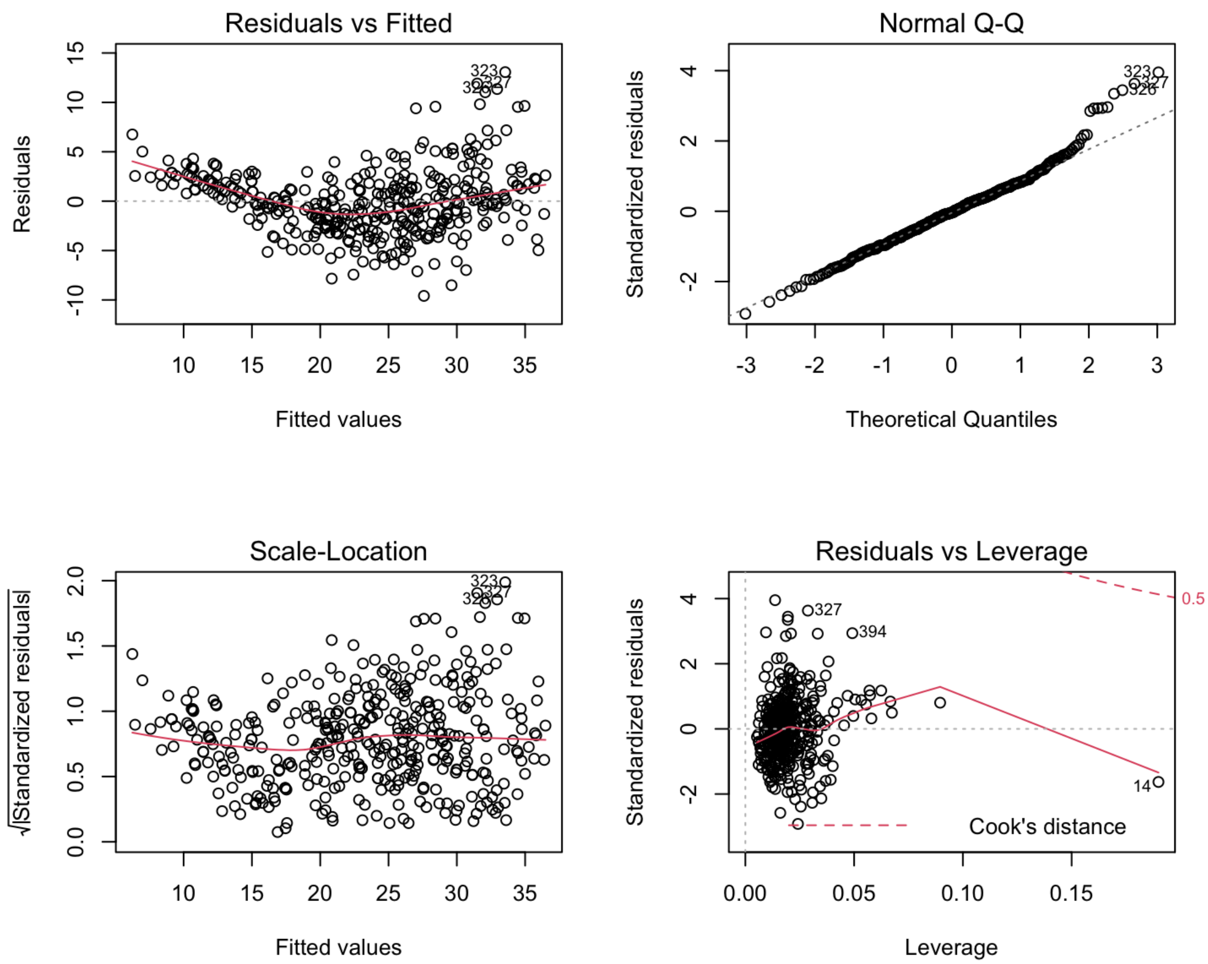
The coefficient associated with year is +0.75077. It is positive, showing positive relationship between year and mpg. It can be inferred as; every year mpg increases by 0.75.

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

```
> plot(lm2.fit)
```

1) We can see that the points follow non-linearity which is not explained in the model. There seem to be some outliers in the plot.
2) The normal Q-Q plot shows no signs of abnormality except at the ends where some points do not follow a straight line.
3) Standardized residual vs fitted values can be used to prove heteroscedasticity, curve is pretty close to a straight line, indicating that heteroscedasticity is less.

4) We can see in the fourth plot that point 14 appears to be a high leverage point. Points 325 and 389 seem to be outliers.



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

```
> library(car)
> vif(lm2.fit)
  cylinders displacement   horsepower       weight
  10.737535    21.836792     9.943693    10.831260
acceleration         year       origin
   2.625806     1.244952     1.772386
```

From the above results we can see that cylinders, displacement, horsepower, weight all have very high vif, showing that there is serious collinearity.

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

From the below solution we can see that interaction between displacement and weight is significant because of the p-value. Whereas interaction between cylinders and displacement is not statistically significant.

```
> lm.fit2 <-  lm(mpg ~ cylinders * displacement + displacement * weight, data = Auto)
> summary(lm.fit2)

Call:
lm(formula = mpg ~ cylinders * displacement + displacement *
    weight, data = Auto)

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
```

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.

```
> attach(Carseats)
> 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) Provide an interpretation of each coefficient in the model (note: some of the variables are qualitative).

*The p-value for **price** is very small, indicating a relationship with Sales. Also, this relationship is negative. For every 100 units increase in price, the sales reduce by 5.44 units*

*There is not much evidence to suggest a relationship between the location of store (**Urban yes**) and sales.*

*There is again sufficient evidence that a store in US (**USYES**) has an effect on the sales as the p-value is very small. Having a store in US, increases the sales.*

(c) Write out the model in equation form.

Sales = 13.04 + -0.05 Price + -0.02 UrbanYes + 1.20 USYes

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

For price and US, predictors we can reject the null hypothesis. On basis of p-values associated with them.

(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.

```
> lm.fit2 = lm(Sales ~ Price + US, data = Carseats)
> summary(lm.fit2)
Call:
lm(formula = Sales ~ Price + US, data = Carseats)
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) How well do the models in (a) and (e) fit the data?
Both of the models in (a) and (e) have similar R^2 values. Hence, both for the model similarly. R^2 is slightly better for model in (e), hence it fits better. 23% of the variance is explained by both the models.

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

All studentized residuals appear to be bounded by -3 to 3, so no potential outliers can be seen from the model. From the second plot we can see that there is one high leverage point which has leverage value greater than 0.04.