Ghoshal_Gourav_HW2

Gourav Ghoshal

October 4, 2018

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?

```r
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 3.4.2
```

```r
data_auto = Auto
head(data_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
```

```r
attach(data_auto)
lm_fit <- lm(mpg~horsepower, data = data_auto)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = 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
```

** ans - ** Yes,there is a statistically significant relationship between predictor and response. The p-value for null hypotesis is very low and hence, we reject the null hypothesis. The relationship is as follow: As the horsepower increases by 1 unit, the mpg reduces by 0.157845 units.

ii. How strong is the relationship between the predictor and the response?

The relationship can be expressed as following: mpg = 39.935861 - 0.157845*horsepower

iii. Is the relationship between the predictor and the response positive or negative?

** ans - ** The relationship is NEGATIVE

iv. How to interpret the estimate of the slope?

** ans ** The interpretation of slope is - As the horsepower increases by 1 unit, the mpg reduces by 0.157845 units.

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

```r
## Prediction @ horsepower = 98
y_hat <- predict(lm_fit, data.frame(horsepower = 98))
print(y_hat)
```

```
##        1
## 24.46708
```

```r
## Confidence and Prediction intervals
print(predict(lm_fit, data.frame(horsepower = 98), interval = c('confidence')))
```

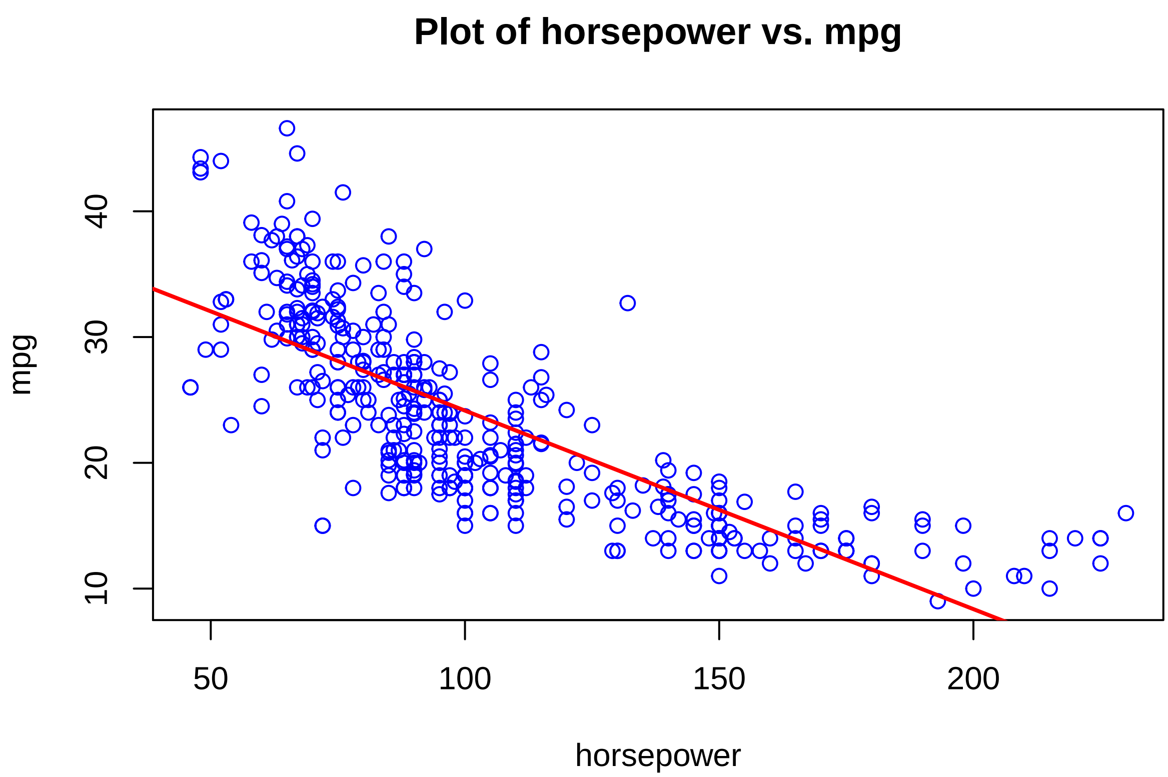
```
##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

```r
print(predict(lm_fit, data.frame(horsepower = 98), interval = c('prediction')))
```

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

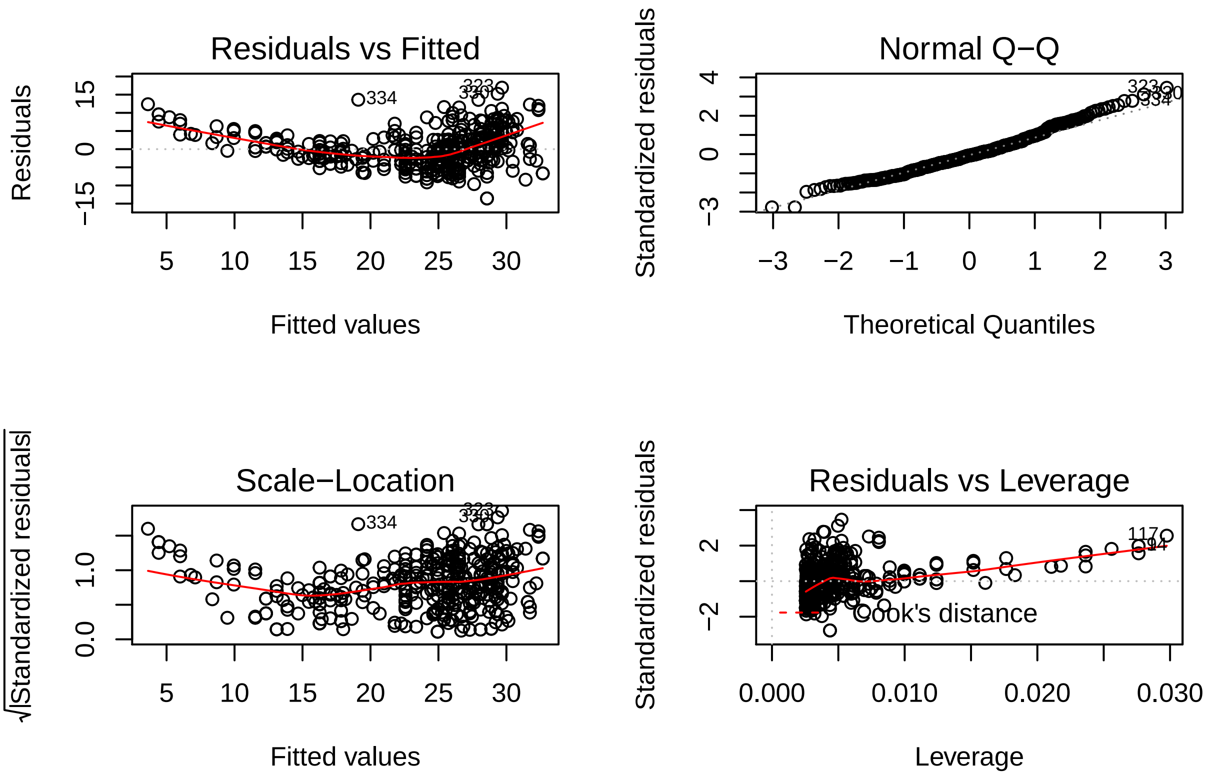
(b) Plot the response and the predictor. Display the least squares regression line in the plot.

```r
{plot(horsepower, mpg, main = 'Plot of horsepower vs. mpg', col = 4)
abline(lm_fit, col = 2, lwd = 2)}
```



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

```r
par(mfrow=c(2,2))
plot(lm_fit)
```



** ans- **

Residual Vs. Fitted plot: There is pattern evident in the plot, suggesting that errors are not independent and the red line is not horizontal.

Normal Q-Q plot: The q-q plot follows the straight line and hence, it is safe to assume that the normality assumption is valid

Scale-Location plot: The variance increases as the mean value of y increases and hence, the homoskedastic assumption,i.e. constant variance assumption is violated, a weighted least square model can be used for this model. Also, transformation of predictors can be used to alleviate the problem

Residual-Leverage plot: There is no influential observation in the data

(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.

Log transformation:

```
lm_log <- lm(mpg ~ log(horsepower), data = data_auto)
summary(lm_log)
```

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

** Comments **

i) Yes, there is statistically significant relationship between log(horsepower) and mpg

ii) mpg = 108.6997 - 18.5822*log(horsepower)

iii) Negative

iv) With increase in unit of log(horsepower), the mpg reduces by 18.5822 units

v)

```r
## Confidence and Prediction intervals
print(predict(lm_log, data.frame(horsepower = 98)))
```

```
##        1
## 23.50099
```

```r
print(predict(lm_log, data.frame(horsepower = 98), interval = c('confidence')))
```

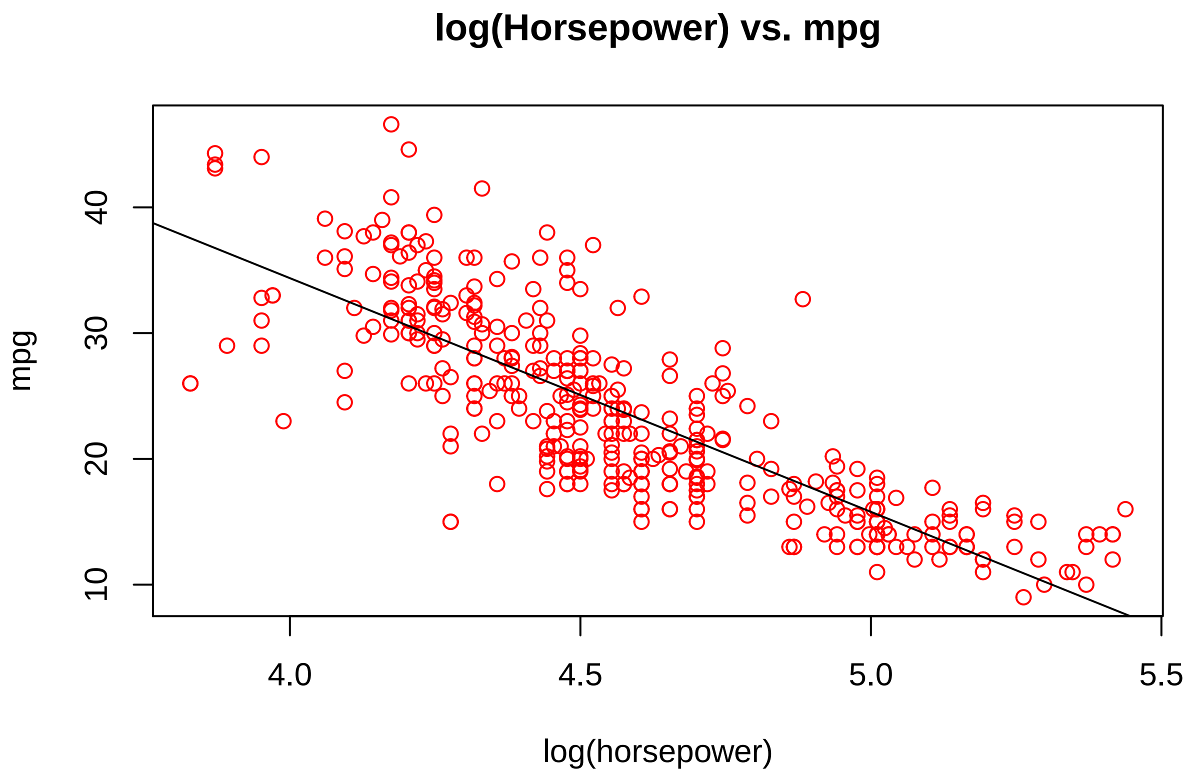
```
##        fit      lwr      upr
## 1 23.50099 23.05405 23.94794
```

```r
print(predict(lm_log, data.frame(horsepower = 98), interval = c('prediction')))
```

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

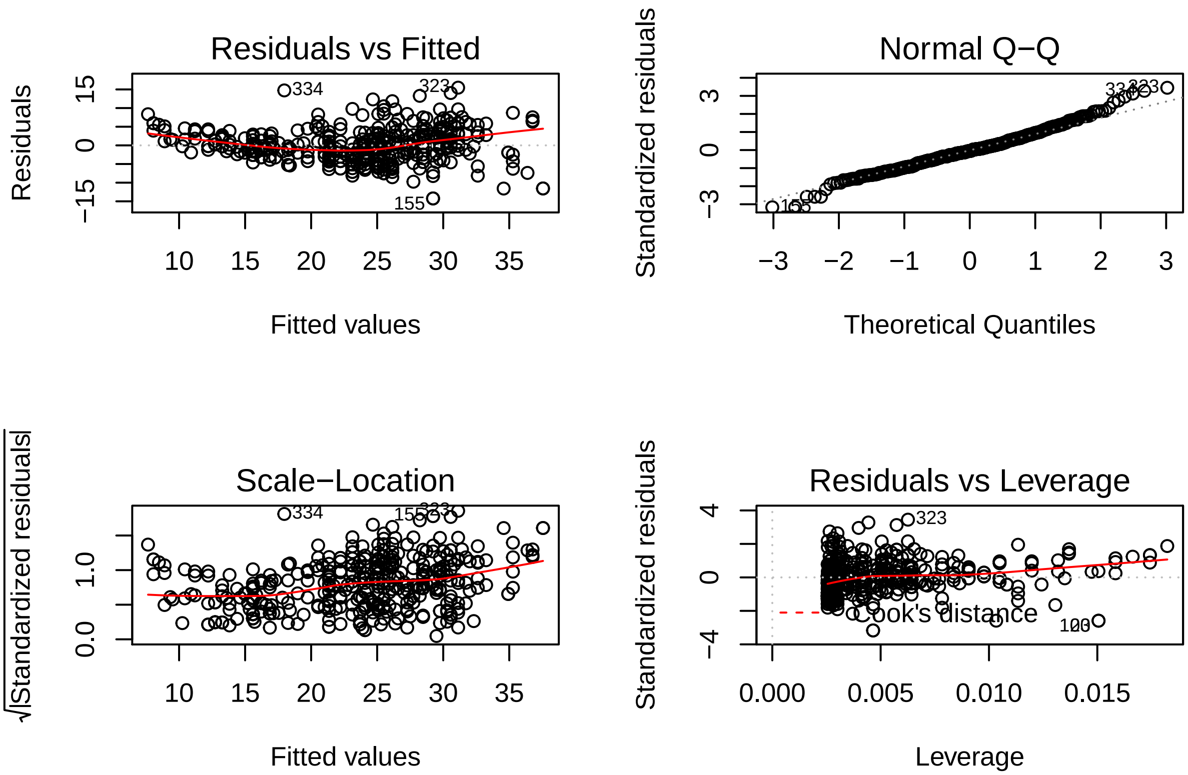
Plot

```r
{plot(log(horsepower), mpg, main = 'log(Horsepower) vs. mpg', col = 2)
abline(lm_log)}
```



Diagnostic plot

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



Residual-Fitted plot: This plot looks great and the assumption that residuals are independent is satisfied as there is no obvious pattern and points are randomly distributed around mean 0 red line

Normal Q-Q plot: The normality assumption looks okay, although there is more points on the top and bottom of the curve, leaving the straight line, heavy tailed distribution

Scale-location: The homoskedastic assumption is looks doubtful as it increase with increasing mean values of y_hat, but I will accept it

Residual - Leverage: The plot is okay with no influential leverage point in the data

sqrt transformation:

```r
lm_sqrt <- lm(mpg ~ sqrt(horsepower), data = data_auto)
summary(lm_sqrt)
```

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

** Comments **

i) Yes, there is statistically significant relationship between sqrt(horsepower) and mpg
ii) mpg = 58.705 - 3.503*(horsepower)
iii) Negative
iv) With increase in unit of sqrt(horsepower), the mpg reduces by 3.503 units
v)

```r
## Confidence and Prediction intervals
print(predict(lm_sqrt, data.frame(horsepower = 98)))
```

```
##        1
## 24.02206
```

```r
print(predict(lm_sqrt, data.frame(horsepower = 98), interval = c('confidence')))
```

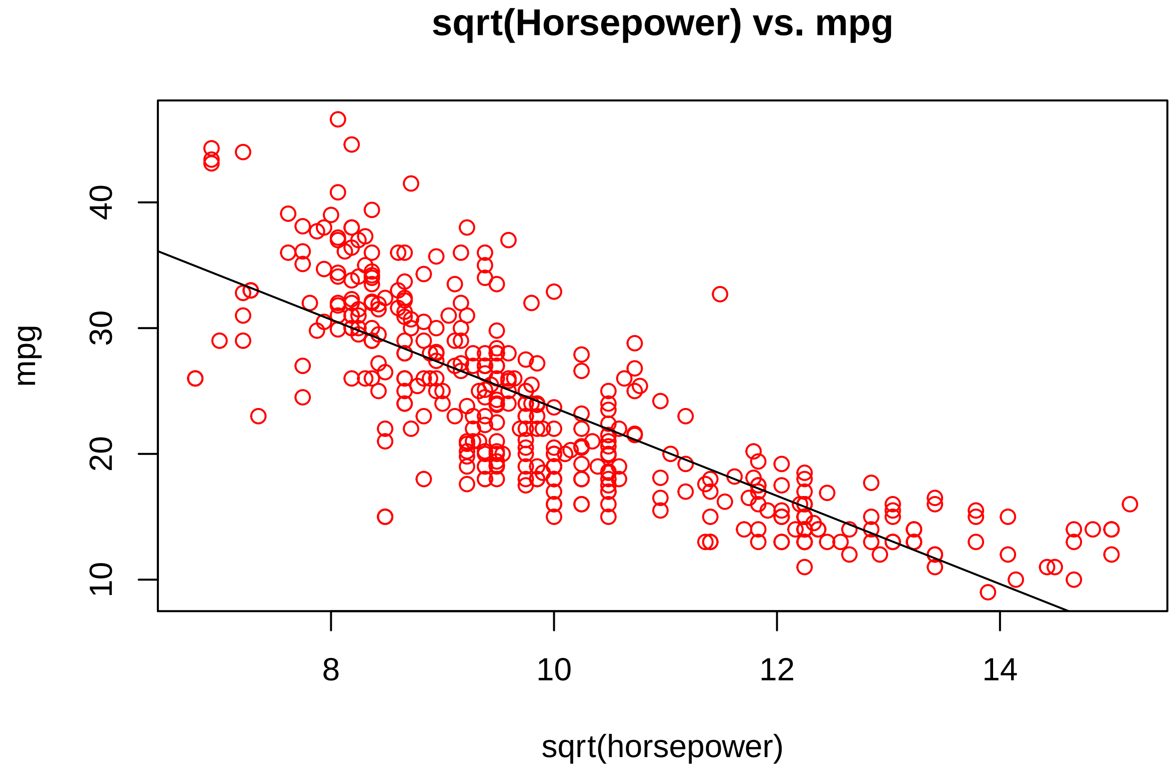
```
##        fit      lwr      upr
## 1 24.02206 23.55687 24.48724
```

```r
print(predict(lm_sqrt, data.frame(horsepower = 98), interval = c('prediction')))
```

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

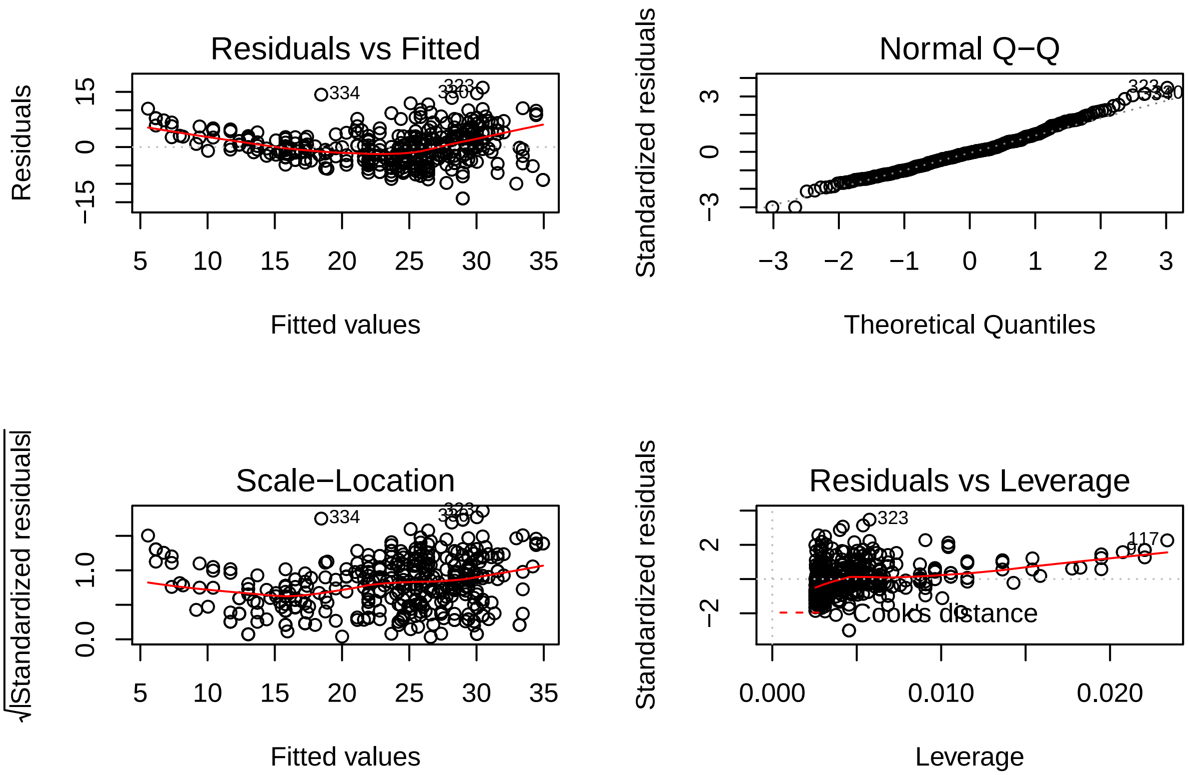
Plot

```r
{plot(sqrt(horsepower), mpg, main = 'sqrt(Horsepower) vs. mpg', col = 2)
abline(lm_sqrt)}
```



Diagnostic plot

```r
par(mfrow=c(2,2))
plot(lm_sqrt)
```



Residual-Fitted plot: This plot looks very good and the assumption that residuals are independent is satisfied as there is no obvious pattern and points are randomly distributed around mean 0 red line

Normal Q-Q plot: The normality assumption looks okay, although there is more points on the top and bottom of the curve, leaving the straight line, suggesting heavy tailed distribution

Scale-location: The homoskedastic assumption is okay

Residual - Leverage: The plot is okay with no influential leverage point in the data

square transformation:

```
lm_square <- lm(mpg ~ I(horsepower^2), data = data_auto)
summary(lm_square)
```

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

** Comments **

i) Yes, there is statistically significant relationship between (horsepower)^2 and mpg

ii) mpg = 30.47 - 0.0005665*(horsepower)^2

iii) Negative

iv) With increase in unit of (horsepower)^2, the mpg reduces by 0.0005665 units

v)

```r
## Confidence and Prediction intervals
print(predict(lm_square, data.frame(horsepower = 98)))
```

```
##        1
## 25.02512
```

```r
print(predict(lm_square, data.frame(horsepower = 98), interval = c('confidence')))
```

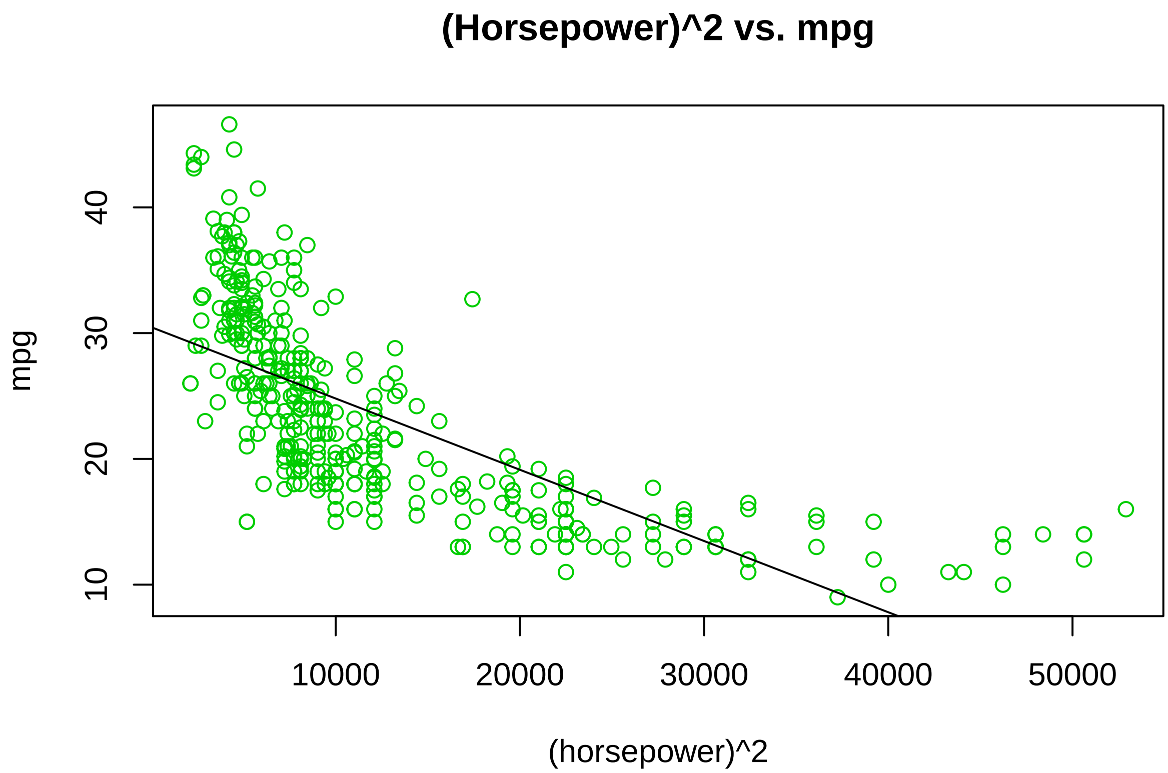
```
##        fit      lwr     upr
## 1 25.02512 24.45883 25.5914
```

```r
print(predict(lm_square, data.frame(horsepower = 98), interval = c('prediction')))
```

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

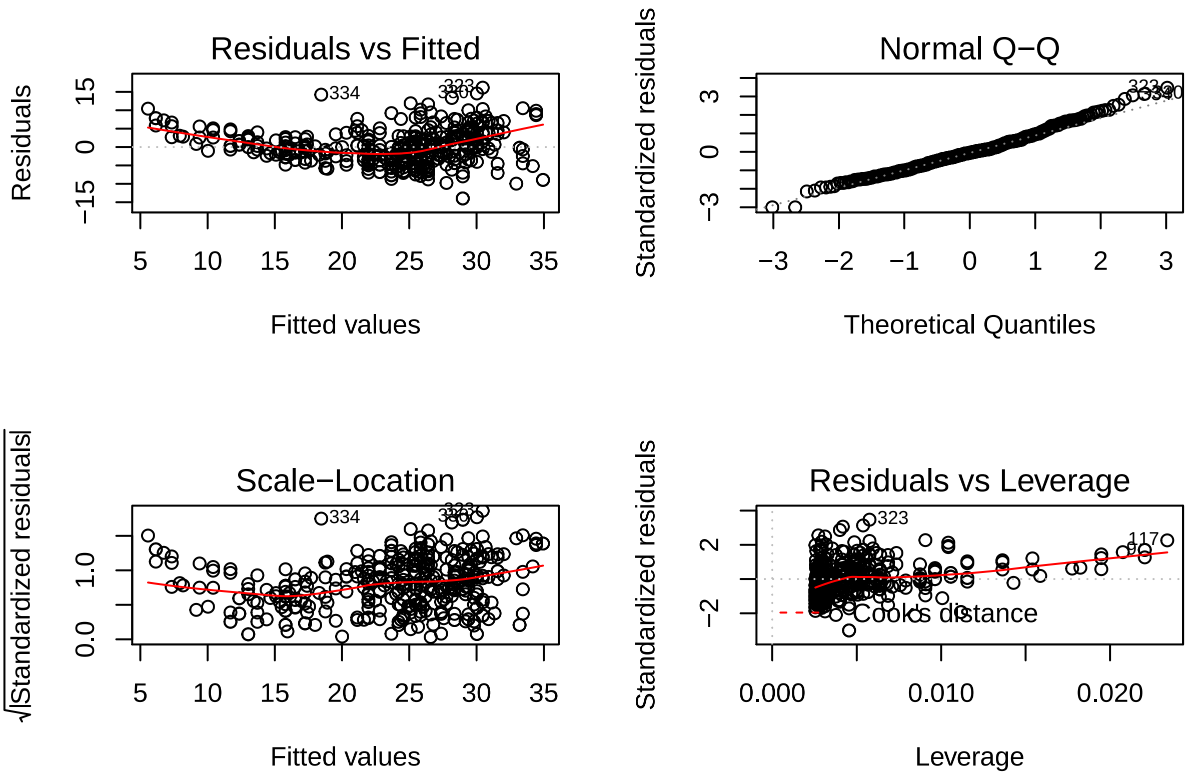
Plot

```r
{plot((horsepower)^2, mpg, main = '(Horsepower)^2 vs. mpg', col = 3)
abline(lm_square)}
```



Diagnostic plot

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



Residual-Fitted plot: This plot looks not okay, as there is parabolic shape suggesting non-linearity

Normal Q-Q plot: The normality assumption looks okay

Scale-location: The homoskedastic assumption is looks doubtful as it increase with increasing mean values of y_hat, but again, I will accept it

Residual - Leverage: The plot is okay with no influential leverage point in the data

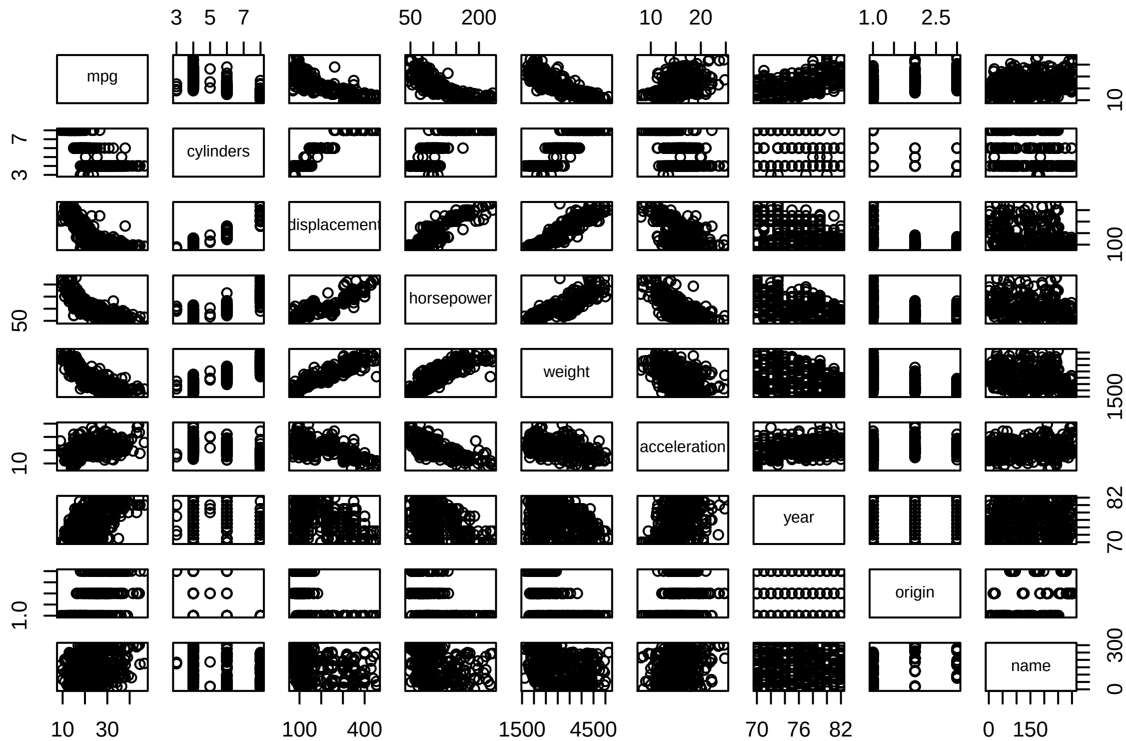
Among all the transformation, the diagnostic plot of lo transformation looks the best and I will go ahead with that model

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?

```
pairs(data_auto)
```



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

```r
colnames(data_auto)
```

```
## [1] "mpg"          "cylinders"    "displacement" "horsepower"  
## [5] "weight"       "acceleration" "year"         "origin"      
## [9] "name"
```

```r
new_data <- data_auto[,-9]
cor(new_data)
```

```
##                     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) 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

```r
attach(new_data)
```

```
## The following objects are masked from data_auto:
##
##     acceleration, cylinders, displacement, horsepower, mpg,
##     origin, weight, year
```

```r
lm_multi <- lm(mpg ~., data = new_data)
summary(lm_multi)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = new_data)
##
## 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
```

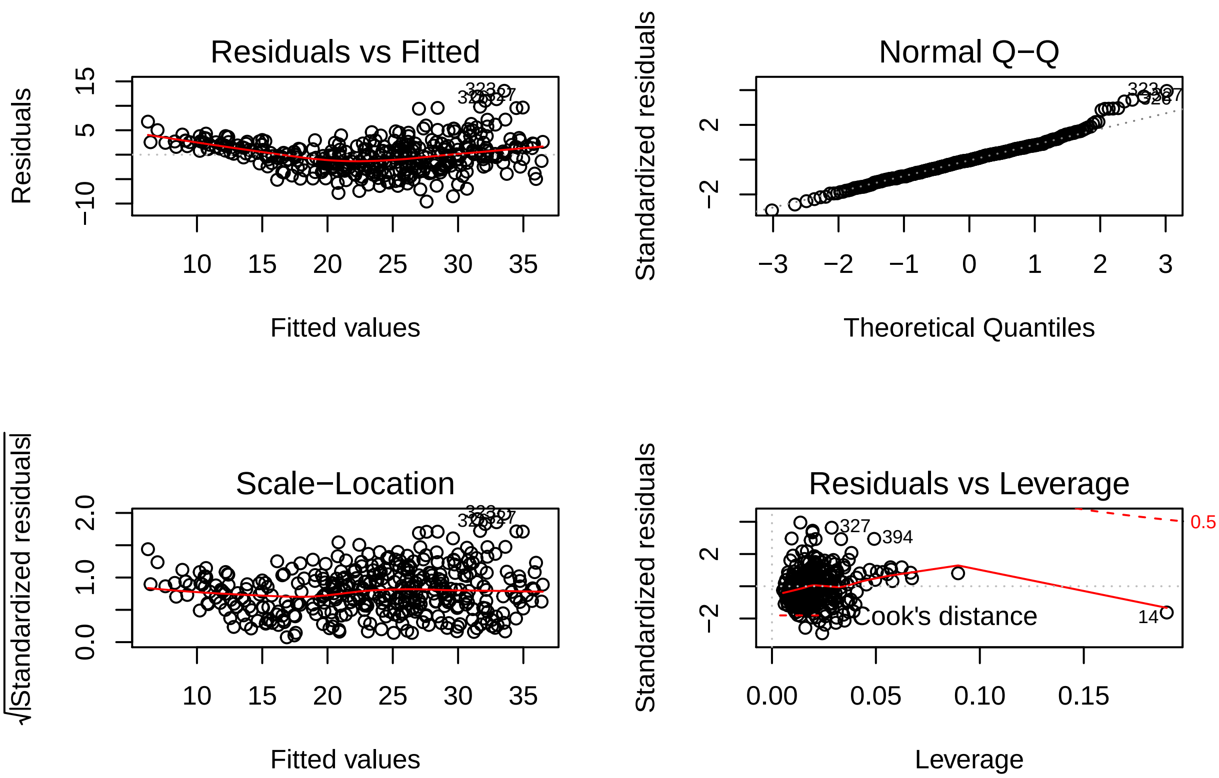
i) Is there a relationship between the predictors and the response?

Yes.

p-value for more than one variable is very low and F-statistic is significant with very low p-value. Hence, null hypothesis is rejected. #### ii) Which predictors have a statistically significant relationship to the response? displacement, weight, year, origin #### iii) What does the coefficient for the year variable suggest? With the increase in year by 1 unit, the mpg increases by 0.750773 units. This indicates that over the years mpg of automobile improved/increased!

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

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



** comments **

Residual-Fitted plot: This plot looks very good and the assumption that residuals are independent is satisfied as there is no obvious pattern and points are randomly distributed around mean 0 red line

Normal Q-Q plot: The normality assumption looks okay, although there is more points on the top of the curve leaving the straight line, suggesting right skewed distribution of the residuals

Scale-location: The homoskedastic assumption is justified and this plot is very good as the variance remains almost constant as the mean value of y_hat increases

Residual - Leverage: The plot is okay with no influential leverage point in the data

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

Yes.

From the pairwise scatter plot it is clear. However the VIF helps us identifying the most collinear predictors:

```
#install.packages('car')
library(car)
```

```
## Warning: package 'car' was built under R version 3.4.4
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 3.4.4
```

```r
print(vif(lm_multi))
```

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

VIF > 5 variables are : cylinders, displacement, horsepower, weight; Among these displacement is most correlated with other variable (VIF = 21.83)

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

```r
lm_inter <- lm(mpg~(cylinders+displacement+horsepower+weight+acceleration+year+origin)^2, data = data_au
summary(lm_inter)
```

```
##
## Call:
## lm(formula = mpg ~ (cylinders + displacement + horsepower + weight +
##     acceleration + year + origin)^2, data = data_auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6303 -1.4481  0.0596  1.2739 11.1386
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                3.548e+01  5.314e+01   0.668  0.50475
## cylinders                  6.989e+00  8.248e+00   0.847  0.39738
## displacement              -4.785e-01  1.894e-01  -2.527  0.01192 *
## horsepower                 5.034e-01  3.470e-01   1.451  0.14769
## weight                     4.133e-03  1.759e-02   0.235  0.81442
## acceleration              -5.859e+00  2.174e+00  -2.696  0.00735 **
## year                       6.974e-01  6.097e-01   1.144  0.25340
## origin                    -2.090e+01  7.097e+00  -2.944  0.00345 **
## cylinders:displacement    -3.383e-03  6.455e-03  -0.524  0.60051
## cylinders:horsepower       1.161e-02  2.420e-02   0.480  0.63157
## cylinders:weight           3.575e-04  8.955e-04   0.399  0.69000
## cylinders:acceleration     2.779e-01  1.664e-01   1.670  0.09584 .
## cylinders:year            -1.741e-01  9.714e-02  -1.793  0.07389 .
## cylinders:origin           4.022e-01  4.926e-01   0.816  0.41482
## displacement:horsepower   -8.491e-05  2.885e-04  -0.294  0.76867
## displacement:weight        2.472e-05  1.470e-05   1.682  0.09342 .
## displacement:acceleration -3.479e-03  3.342e-03  -1.041  0.29853
## displacement:year          5.934e-03  2.391e-03   2.482  0.01352 *
## displacement:origin        2.398e-02  1.947e-02   1.232  0.21875
## horsepower:weight         -1.968e-05  2.924e-05  -0.673  0.50124
## horsepower:acceleration   -7.213e-03  3.719e-03  -1.939  0.05325 .
## horsepower:year           -5.838e-03  3.938e-03  -1.482  0.13916
## horsepower:origin          2.233e-03  2.930e-02   0.076  0.93931
```

```
## weight:acceleration        2.346e-04  2.289e-04   1.025  0.30596
## weight:year               -2.245e-04  2.127e-04  -1.056  0.29182
## weight:origin             -5.789e-04  1.591e-03  -0.364  0.71623
## acceleration:year          5.562e-02  2.558e-02   2.174  0.03033 *
## acceleration:origin        4.583e-01  1.567e-01   2.926  0.00365 **
## year:origin                1.393e-01  7.399e-02   1.882  0.06062 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.695 on 363 degrees of freedom
## Multiple R-squared:  0.8893, Adjusted R-squared:  0.8808
## F-statistic: 104.2 on 28 and 363 DF,  p-value: < 2.2e-16
```

At alpha = 0.05, only following pairs are statistically significant: (acceleration:origin), (acceleration:year), (displacement:year)

Finally, combining part (e) and (f), keeping model simplicity in mind, only the following predictors are used which generated the final model with RSE = 3.45 and Adjusted R-square = 80.76%

```r
lm_final <- lm(mpg~ (acceleration+year+origin+acceleration:origin+
                      acceleration:year+displacement:year), data=new_data)
summary(lm_final)
```

```
##
## Call:
## lm(formula = mpg ~ (acceleration + year + origin + acceleration:origin +
##     acceleration:year + displacement:year), data = new_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0931  -1.8205  -0.1685   1.7350  15.4803
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          7.405e+01  2.118e+01   3.496 0.000527 ***
## acceleration        -6.840e+00  1.325e+00  -5.161 3.95e-07 ***
## year                -2.624e-01  2.833e-01  -0.926 0.355020
## origin              -1.169e+01  1.729e+00  -6.765 4.98e-11 ***
## acceleration:origin  7.747e-01  1.043e-01   7.430 7.07e-13 ***
## acceleration:year    7.243e-02  1.780e-02   4.070 5.71e-05 ***
## year:displacement   -7.531e-04  3.672e-05 -20.513  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.45 on 385 degrees of freedom
## Multiple R-squared:  0.8076, Adjusted R-squared:  0.8046
## F-statistic: 269.4 on 6 and 385 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

```r
head(Carseats)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age
## 1  9.50       138     73          11        276   120       Bad  42
## 2 11.22       111     48          16        260    83      Good  65
## 3 10.06       113     35          10        269    80    Medium  59
## 4  7.40       117    100           4        466    97    Medium  55
## 5  4.15       141     64           3        340   128       Bad  38
## 6 10.81       124    113          13        501    72       Bad  78
##   Education Urban  US
## 1        17   Yes Yes
## 2        10   Yes Yes
## 3        12   Yes Yes
## 4        14   Yes Yes
## 5        13   Yes  No
## 6        16    No Yes
```

```r
attach(Carseats)
lm_cs <- lm(Sales ~ Price+Urban+US, data = Carseats)
summary(lm_cs)
```

```
##
## 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).

Price: Unit increase in price reduces Sales by 0.054459 units Urban: For Urban (Yes values), Sales reduces by 0.02 units - although the co-eff is not statistically significant and the predictoe can dropped from the model

US: For US (Yes values), Sales increases by 1.200573 units

(c) Write out the model in equation form

Sales = 13.043469 - 0.054459*Price - 0.021916*Urban + 1.200573*US dropping the statistically insignifacnt term, the model becomes: Sales = 13.043469 - 0.054459*Price + 1.200573*US where, US = 1, if yes, 0 otherwise ; Urban = 1, if yes, 0 otherwise

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

Price, US (USYes)

(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.

```r
lm_new <- lm(Sales ~ Price + US, data = Carseats)
summary(lm_new)
```

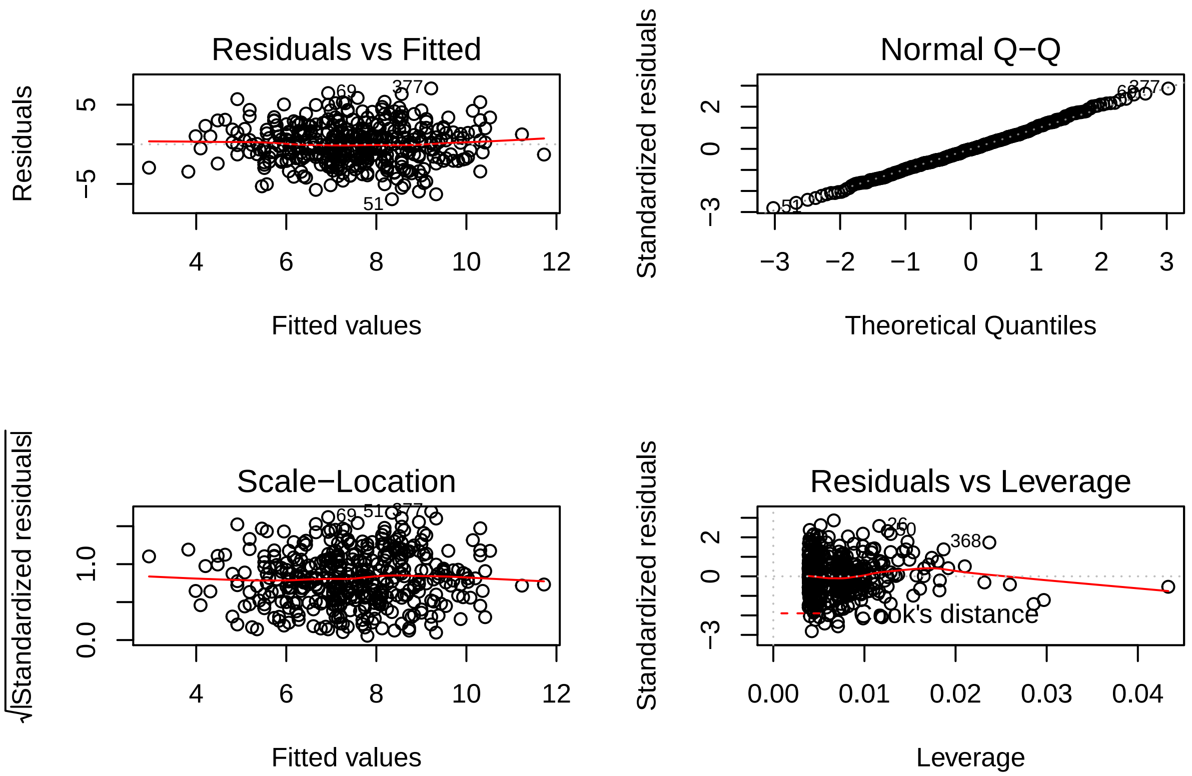
```
##
## Call:
## lm(formula = Sales ~ Price + US, data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9269 -1.6286 -0.0574  1.5766  7.0515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
## 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?

Almost same, RSE reduced in (e) by 0.003

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

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



There are some high leverage points, but none is influential.