3.7 Exercises Problem 9

Section 23 - Group 6 Project Groups (Bosan Hsu, Fan Liu, Jimeng Yin, Michael Liu, Richard Wang, Zhuoqian Zhang

This question involves the use of multiple linear regression on the Auto data set.

```
library(ISLR)
attach(Auto)
Auto = na.omit(Auto)
View(Auto)
summary(Auto)
```

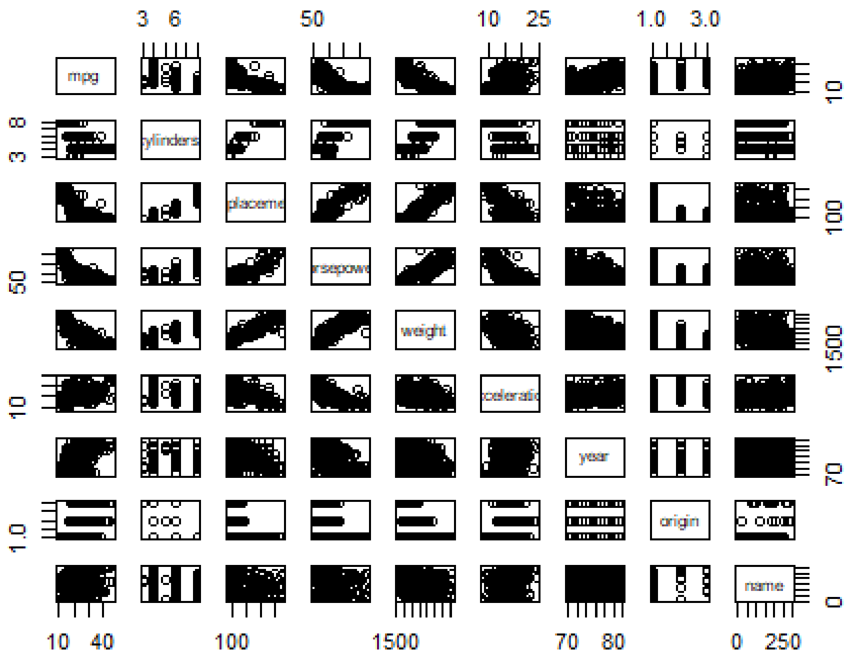
```
##       mpg          cylinders      displacement     horsepower
 weight
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0   Min.
   :1613
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0   1st
 Qu.:2225
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5   Med
ian :2804
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5   Mea
n   :2978
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0   3rd
 Qu.:3615
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0   Max.
   :5140
##

##   acceleration        year           origin                       nam
e
##  Min.   : 8.00   Min.   :70.00   Min.   :1.000   amc matador        :
  5
##  1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000   ford pinto         :
  5
##  Median :15.50   Median :76.00   Median :1.000   toyota corolla     :
  5
##  Mean   :15.54   Mean   :75.98   Mean   :1.577   amc gremlin        :
  4
##  3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000   amc hornet         :
  4
##  Max.   :24.80   Max.   :82.00   Max.   :3.000   chevrolet chevette:
  4
##                                                  (Other)            :
365
```

```
names(Auto)
```

```
## [1] "mpg"          "cylinders"    "displacement" "horsepower"    "wei
ght"
## [6] "acceleration" "year"         "origin"       "name"

dim(Auto)

## [1] 392   9
```

(a) Produce a scatterplot matrix which includes all of the variables in the data set.

```
pairs(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.

```
quantitative_predictor = Auto[ ,1:8]
cor(quantitative_predictor)

##                     mpg  cylinders displacement horsepower     weigh
t
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.832244
2
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.897527
3
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.932994
4
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.864537
7
```

```
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.000000
0
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.416839
2
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.309119
9
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.585005
4
##              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) Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results. Comment on the output. For instance:

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

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

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

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

```
## 
## 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. Is there a relationship between the predictors and the response?
ii. Which predictors appear to have a statistically significant relationship to the response?

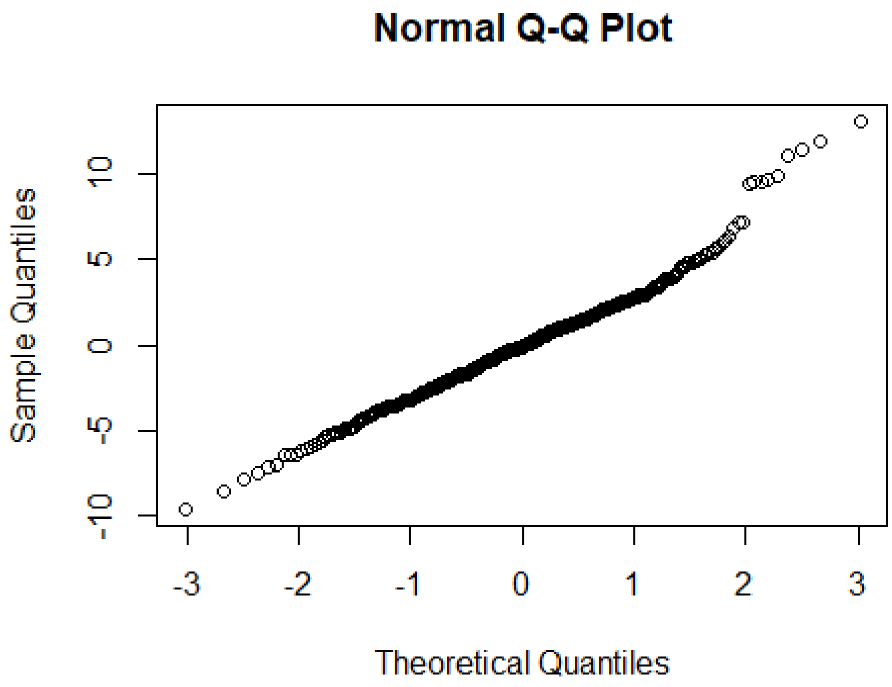
mpg has a statistically significant positive relationship with displacement, weight (negative), year, and origin; while does not has a clear relationship with cylinders, horsepower, and acceleration. Adjusted R-squared of 0.8182 is high.

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

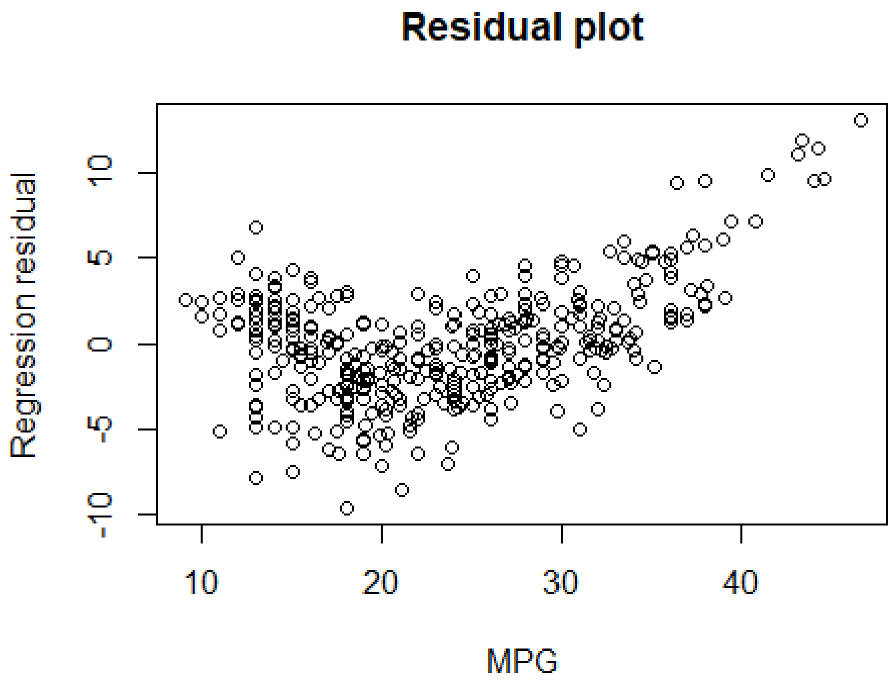
the coefficient for the year variable is 0.750773. When the year variable increase by 1 unit, mpg will increase by 0.750773 on average

(d) Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

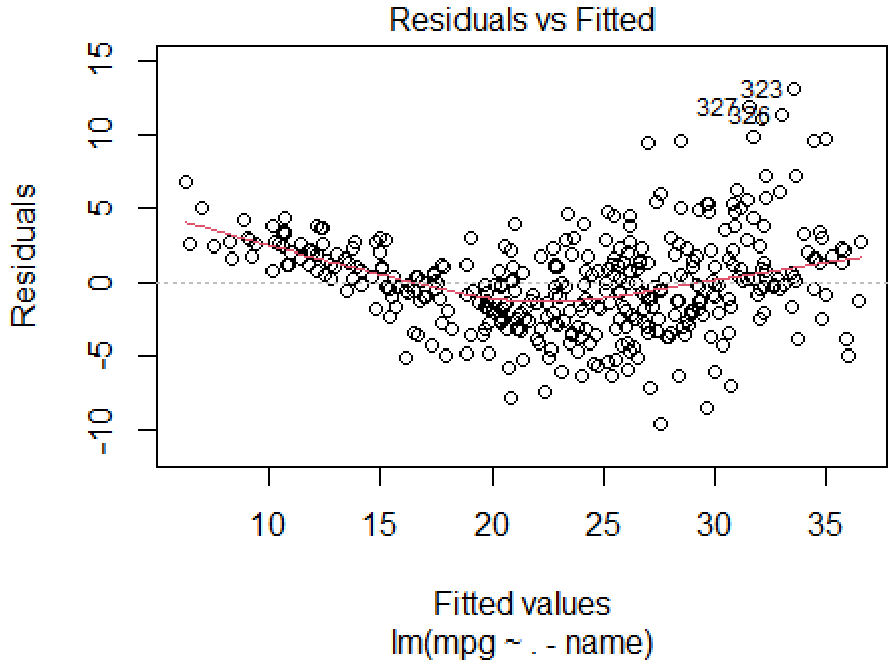
```
normal_qq_plot = qqnorm(lm.fit$residuals)
```

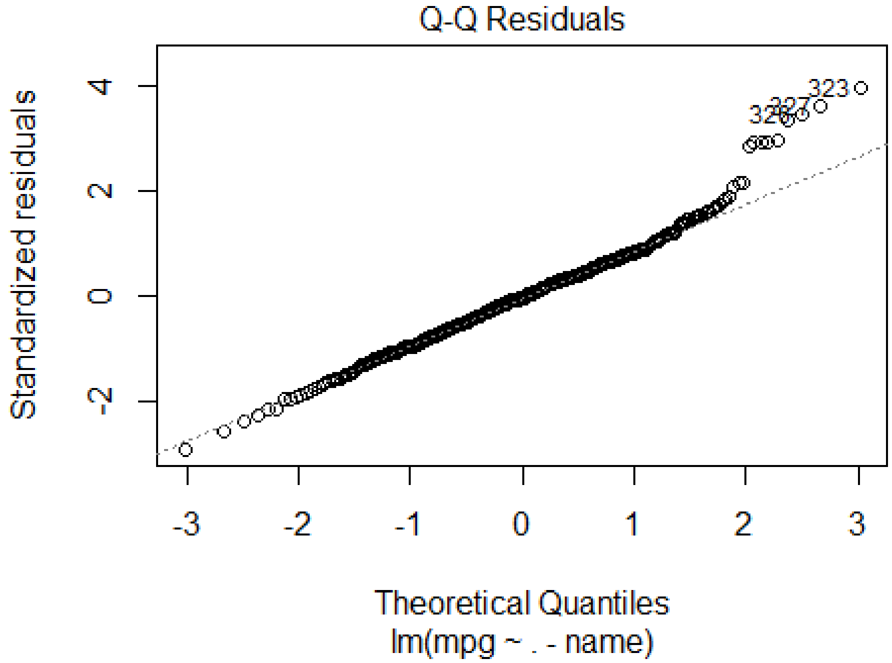


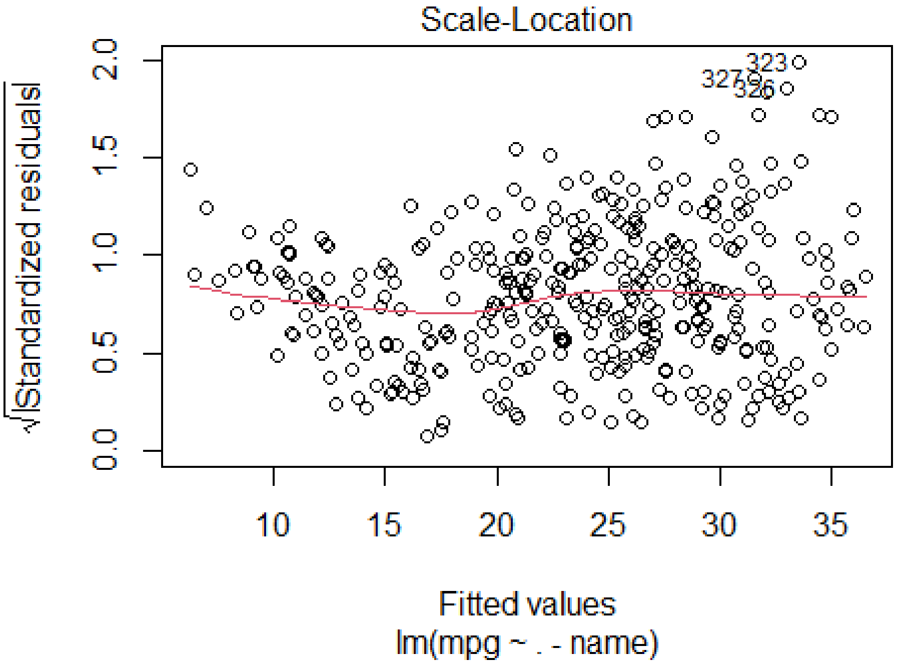
```
residual_plot = plot(Auto[, 1], lm.fit$residuals, xlab = "MPG", ylab =
"Regression residual", main = "Residual plot")
```

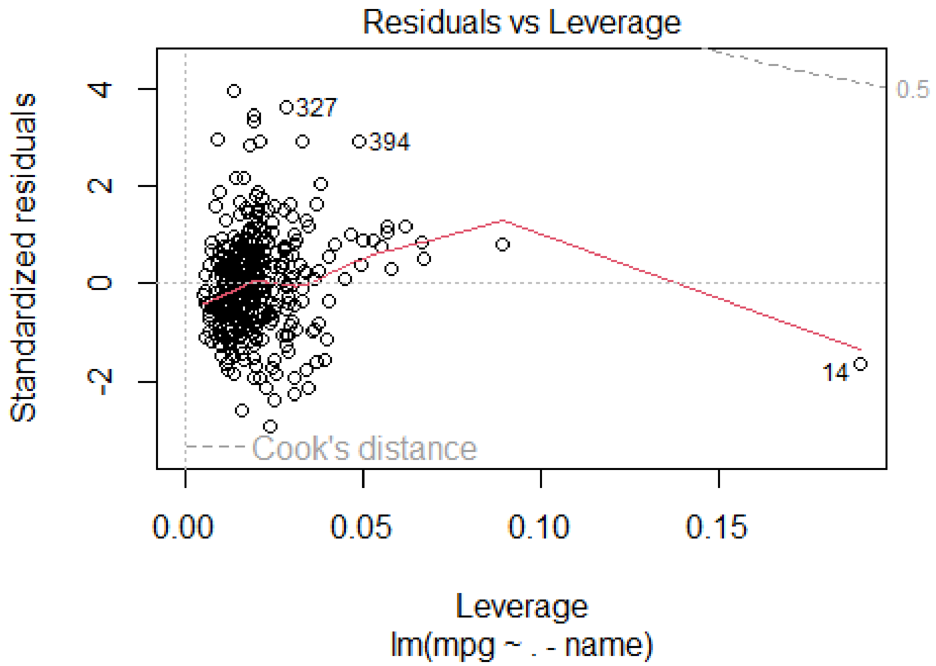


```
plot(lm.fit)
```


Residuals vs Fitted
Residuals
15
10
5
0
-5
-10
323
327
326
10
15
20
25
30
35
Fitted values
lm(mpg ~ . - name)


Q-Q Residuals
Standardized residuals
4
2
0
-2
323
327
326
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(mpg ~ . - name)





on the first and fourth normal qq plot, almost all the point align along a straight line, which means the most of residuals are normally distributed. But there are a couple of points at the right end that clearly are outliers. on the second and third residual plot, we can find a clear "U" shape, which means non-linearity in the data. on the last plot, we can find a extreme high leverage point

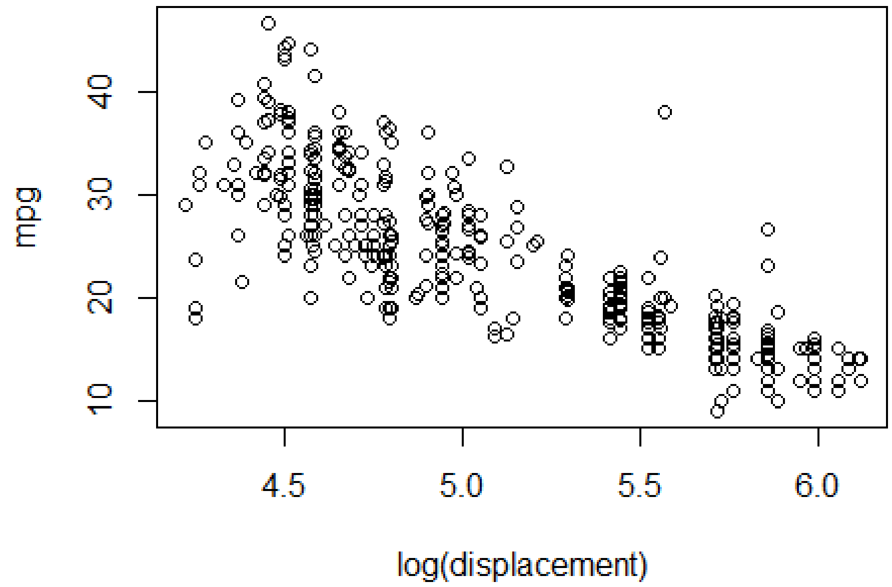
(e) Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

```
lm.fit.e = lm(mpg~displacement * horsepower + weight * acceleration, da
ta = Auto)
summary(lm.fit.e)

##
## Call:
## lm(formula = mpg ~ displacement * horsepower + weight * acceleration,

##     data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5847  -2.2713  -0.2229   1.8801  16.7669
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               6.625e+01  7.319e+00   9.052  < 2e-16 ***
## displacement             -8.441e-02  1.098e-02  -7.684 1.29e-13 ***
## horsepower               -2.509e-01  2.815e-02  -8.913  < 2e-16 ***
## weight                   -2.930e-03  2.201e-03  -1.331    0.184
## acceleration             -6.152e-01  3.881e-01  -1.585    0.114
## displacement:horsepower   5.775e-04  7.526e-05   7.674 1.39e-13 ***
## weight:acceleration       9.477e-05  1.291e-04   0.734    0.463
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.842 on 385 degrees of freedom
## Multiple R-squared:  0.7615, Adjusted R-squared:  0.7577
## F-statistic: 204.8 on 6 and 385 DF,  p-value: < 2.2e-16
```

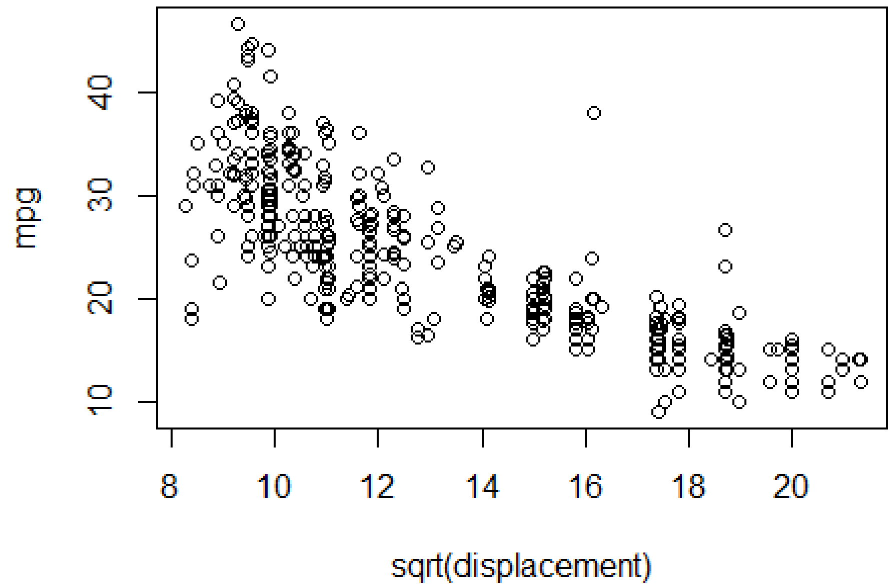
We can see that between displacement and horsepower has statistically significant interaction, while weight and acceleration has no significant interaction.

(f) Try a few different transformations of the variables, such as log(X), √X, X2. Comment on your findings.

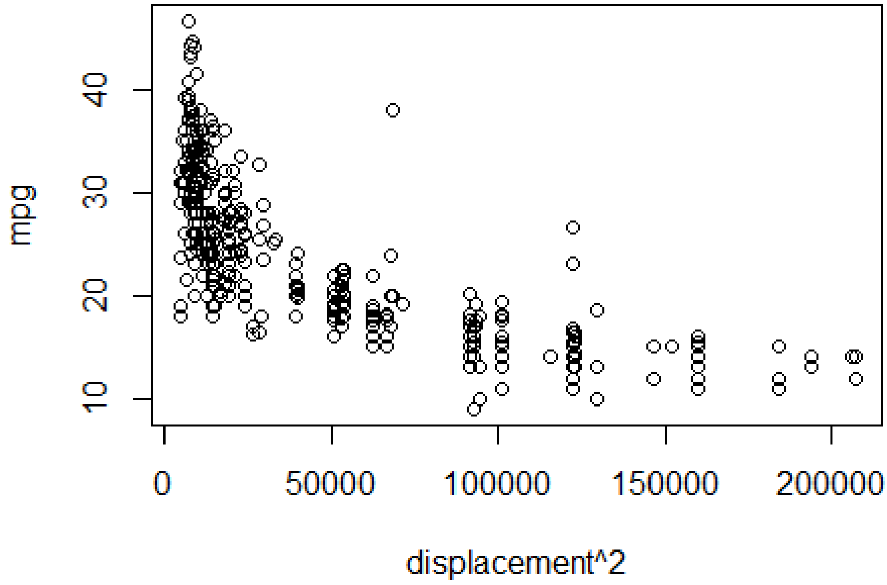
```
plot(log(displacement), mpg)
```



```
plot(sqrt(displacement), mpg)
```



```
plot(displacement^2, mpg)
```



```
lm.fit.f1 = lm(mpg~log(displacement))
lm.fit.f2 = lm(mpg~sqrt(displacement))
lm.fit.f3 = lm(mpg~displacement^2)
summary(lm.fit.f1)
```

```
##
## Call:
## lm(formula = mpg ~ log(displacement))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1204  -2.5843  -0.4217   2.1979  19.9005
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        85.6906     2.1422   40.00   <2e-16 ***
## log(displacement) -12.1385     0.4155  -29.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.377 on 390 degrees of freedom
## Multiple R-squared:  0.6863, Adjusted R-squared:  0.6855
## F-statistic: 853.4 on 1 and 390 DF,  p-value: < 2.2e-16
```

```
summary(lm.fit.f2)
```

```
##
## Call:
## lm(formula = mpg ~ sqrt(displacement))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4034  -2.7367  -0.4956   2.3207  19.3499
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        47.11839    0.86246   54.63   <2e-16 ***
## sqrt(displacement) -1.75878    0.06186  -28.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.458 on 390 degrees of freedom
## Multiple R-squared:  0.6746, Adjusted R-squared:  0.6738
## F-statistic: 808.5 on 1 and 390 DF,  p-value: < 2.2e-16
```

```
summary(lm.fit.f3)
```

```
##
## Call:
## lm(formula = mpg ~ displacement^2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9170  -3.0243  -0.5021   2.3512  18.6128
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.12064    0.49443   71.03   <2e-16 ***
## displacement -0.06005    0.00224  -26.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.635 on 390 degrees of freedom
## Multiple R-squared:  0.6482, Adjusted R-squared:  0.6473
## F-statistic: 718.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

We choose “displacement” as the only predictor and conduct the regression. We can see that plot one (log) and plot two (sqrt) show a clear linear trend. All the three regression lines have a relative high R^2 (>0.6).