ii.Exercise3.7.9

Kaarthik Sundaramoorthy, Sahil Shah and Vidhi Shah

6/3/2020

This question involves the use of **Multiple Linear Regression** on the **Auto** data set.

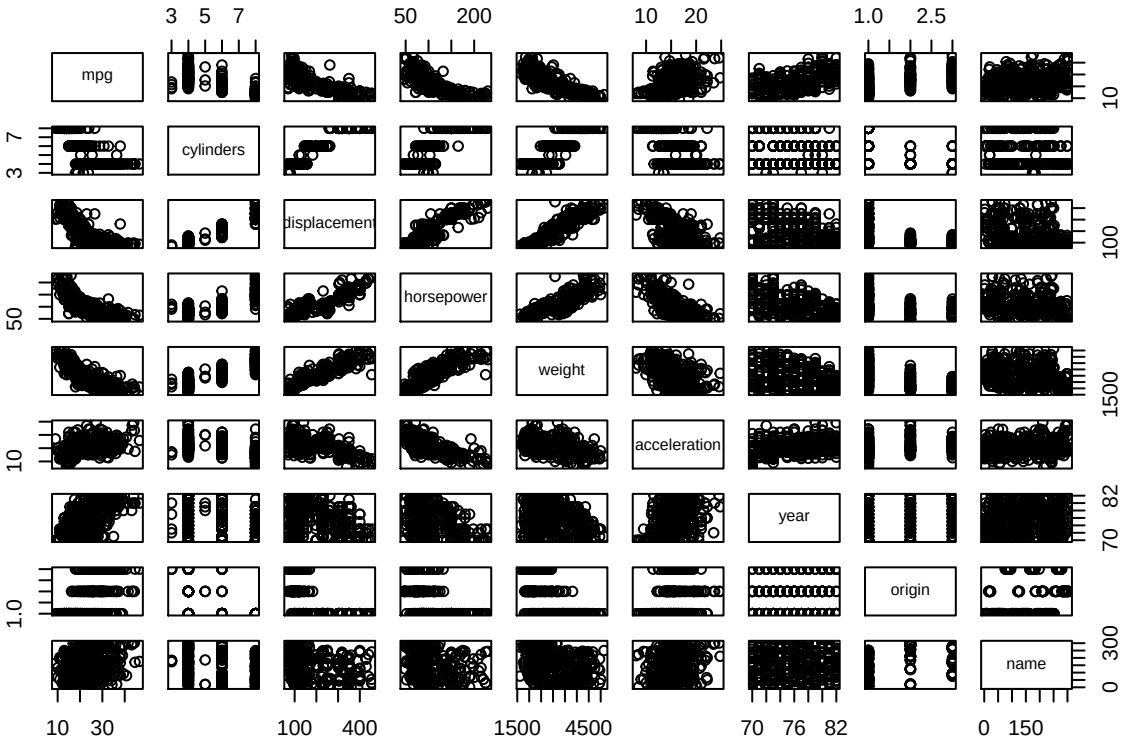
```
require(ISLR)
```

```
## Loading required package: ISLR
```

```
data(Auto)
attach(Auto)
```

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

```
par(mfrow = c(1,1))
pairs(Auto)
```



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

```
head(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
```

```
##help("cor")
cor(subset(Auto, select = -name))
```

```
##                     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) 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:

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

```
##
## 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
```

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

Yes, there is a relationship between the predictors and the response variables. The F-Statistics is really far from 1, which results into small p-value. Indicating that there is an evidence against the null hypothesis.

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

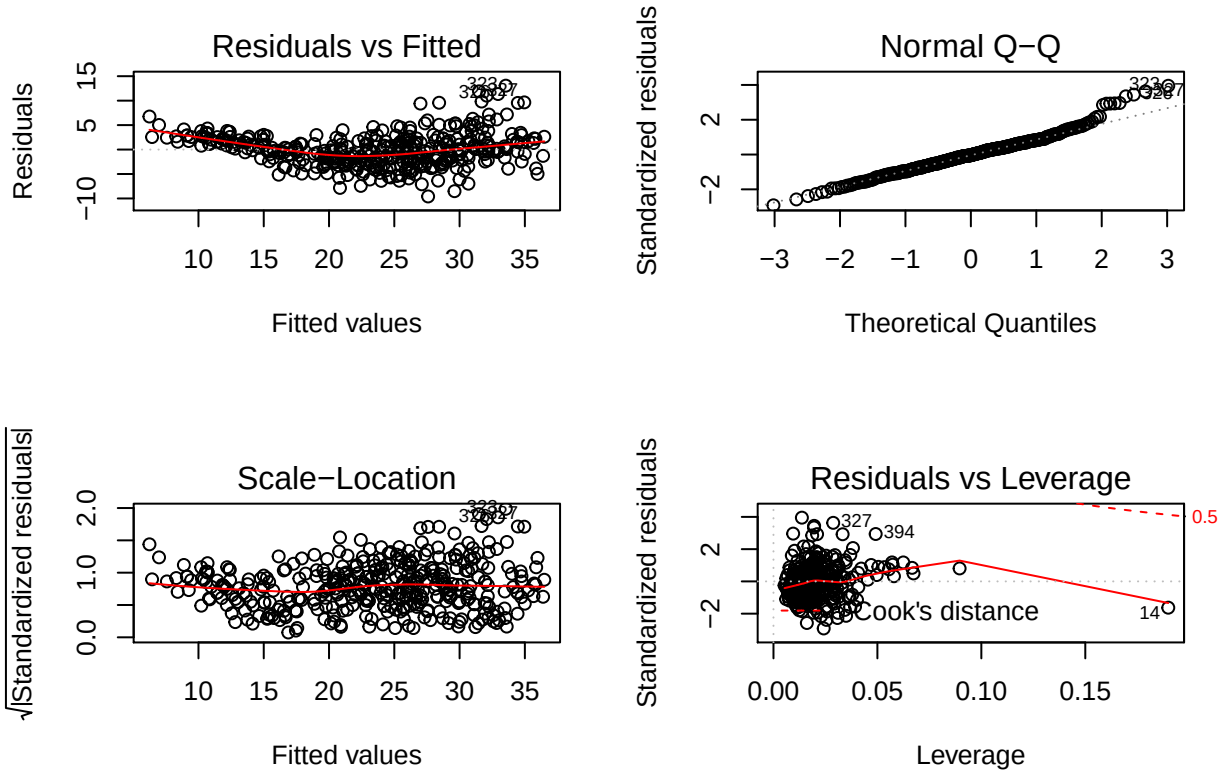
Predictors like **origin**, **year**, **weight** and **displacement** have the statistically significant relationship with the response variable.

(c) iii. What does the coefficient for the year variable suggest?

Ans. The coefficient for the **year** i.e. 0.75 with the additional year of the car's age, the **mpg** will on average increase by 0.75. In other words it suggests that later model year car will have better **mpg**.

(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?

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



There is an evidence of the non-linearity. The plot of Residuals vs Fitted shows that variance of the error terms increase with the values of the response. The Residuals vs Leverage plot shows no outliers, while the high leverage point exists(i.e. 14).

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

```
lm.fit1 <- lm(mpg~cylinders*displacement+displacement*weight, data = Auto)
summary(lm.fit1)
```

```
##
## 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
```

```
lm.fit2 <- lm(mpg ~.-name+displacement:weight, data = Auto)
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = mpg ~ . - name + displacement:weight, data = Auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9027 -1.8092 -0.0946  1.5549 12.1687
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -5.389e+00  4.301e+00  -1.253   0.2109
## cylinders            1.175e-01  2.943e-01   0.399   0.6899
## displacement        -6.837e-02  1.104e-02  -6.193 1.52e-09 ***
## horsepower          -3.280e-02  1.238e-02  -2.649   0.0084 **
## weight              -1.064e-02  7.136e-04 -14.915  < 2e-16 ***
## acceleration         6.724e-02  8.805e-02   0.764   0.4455
## year                 7.852e-01  4.553e-02  17.246  < 2e-16 ***
## origin               5.610e-01  2.622e-01   2.139   0.0331 *
## displacement:weight  2.269e-05  2.257e-06  10.054  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.964 on 383 degrees of freedom
## Multiple R-squared:  0.8588, Adjusted R-squared:  0.8558
## F-statistic: 291.1 on 8 and 383 DF,  p-value: < 2.2e-16
```

```
lm.fit3 <- lm(mpg~year+origin+displacement*weight, data = Auto)
summary(lm.fit3)
```

```
##
## Call:
## lm(formula = mpg ~ year + origin + displacement * weight, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.6119  -1.7290  -0.0115   1.5609  12.5584
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -8.007e+00  3.798e+00  -2.108   0.0357 *
## year                 8.194e-01  4.518e-02  18.136  < 2e-16 ***
```

```
## origin               3.567e-01  2.574e-01   1.386   0.1666
## displacement        -7.148e-02  9.176e-03  -7.790 6.27e-14 ***
## weight              -1.054e-02  6.530e-04 -16.146  < 2e-16 ***
## displacement:weight  2.104e-05  2.214e-06   9.506  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.016 on 386 degrees of freedom
## Multiple R-squared:  0.8526, Adjusted R-squared:  0.8507
## F-statistic: 446.5 on 5 and 386 DF,  p-value: < 2.2e-16
```

From all the above models, the 2nd model has more no. of variables being significant. The R-squared statistics estimates that 86% of the changes in the response can be explained by this particular set of predictors.

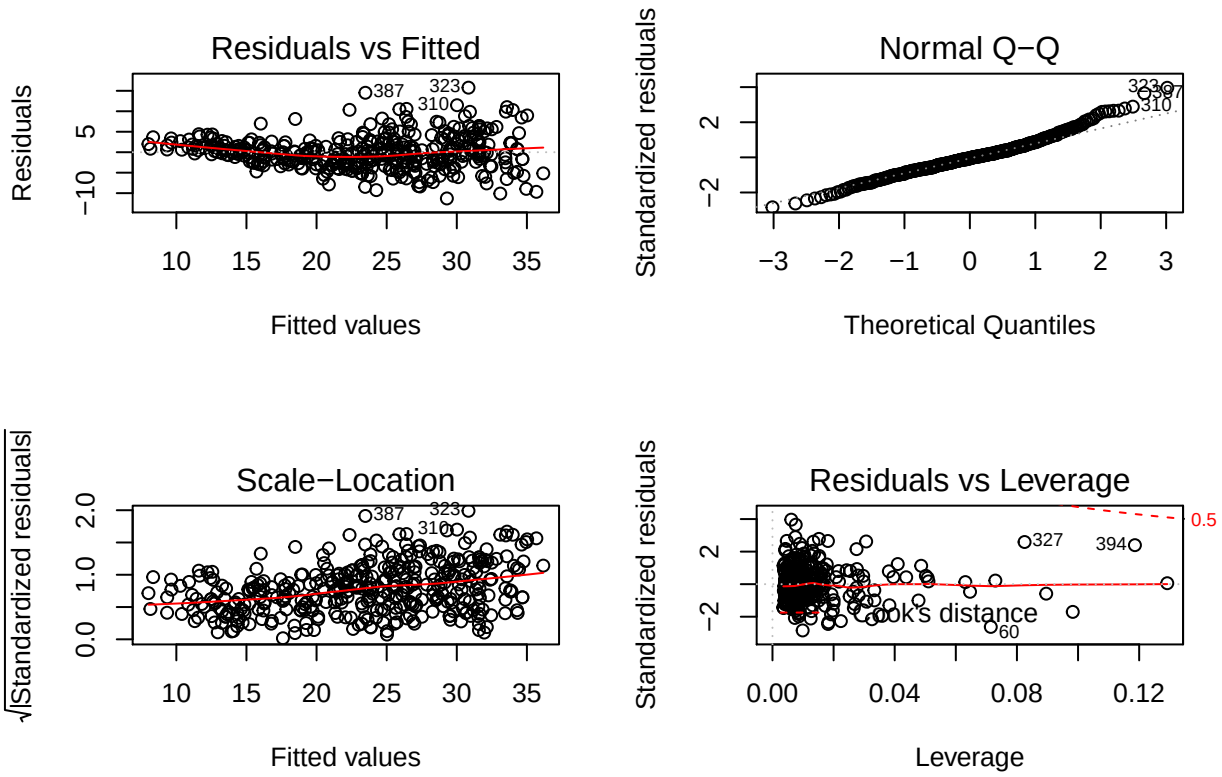
(f) Try a few different transformations of the variables, such as log(X), √X, X^2. Comment on your findings.

```
lm.fit5 = lm(mpg~log(weight)+sqrt(horsepower)+acceleration+I(acceleration^2), data = Auto)
summary(lm.fit5)
```

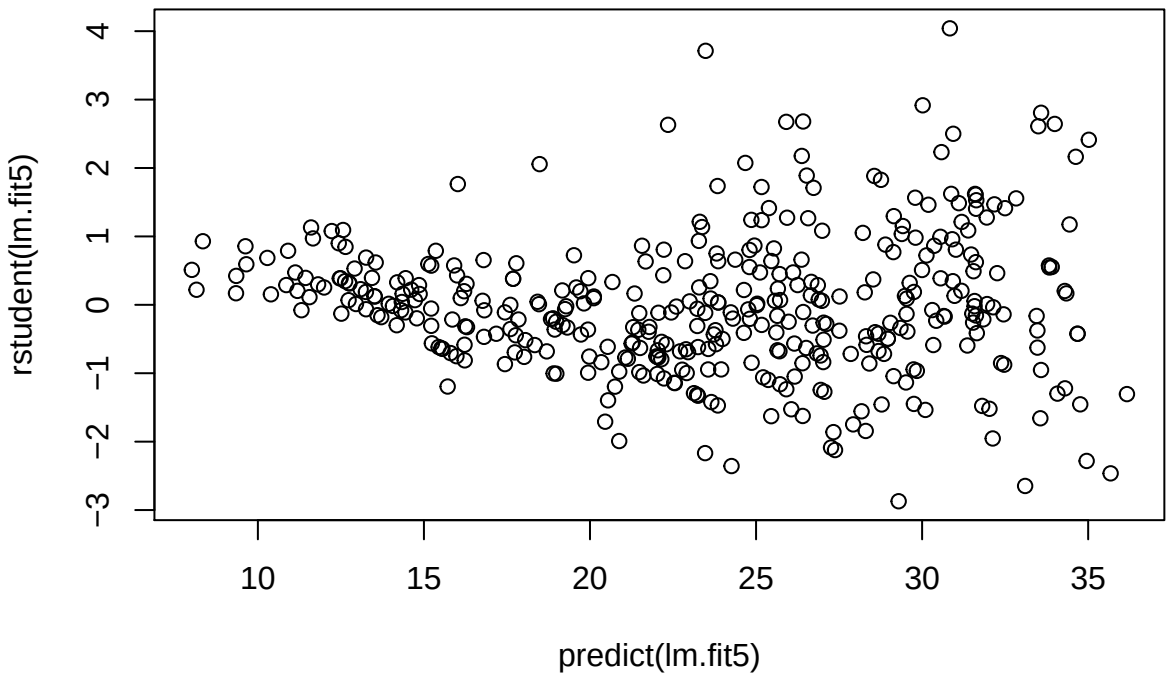
```
##
## Call:
## lm(formula = mpg ~ log(weight) + sqrt(horsepower) + acceleration +
##     I(acceleration^2), data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2932  -2.5082  -0.2237   2.0237  15.7650
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       178.30303   10.80451  16.503  < 2e-16 ***
## log(weight)       -14.74259    1.73994  -8.473 5.06e-16 ***
## sqrt(horsepower)   -1.85192    0.36005  -5.144 4.29e-07 ***
## acceleration       -2.19890    0.63903  -3.441 0.000643 ***
## I(acceleration^2)   0.06139    0.01857   3.305 0.001037 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.99 on 387 degrees of freedom
## Multiple R-squared:  0.7414, Adjusted R-squared:  0.7387
## F-statistic: 277.3 on 4 and 387 DF,  p-value: < 2.2e-16
```

log(weight), sqrt(horsepower) and accelaration^2 have an evidence of statistical significance from the p-values.

```
par(mfrow=c(2,2))
plot(lm.fit5)
```




```
plot(predict(lm.fit5), rstudent(lm.fit5))
```



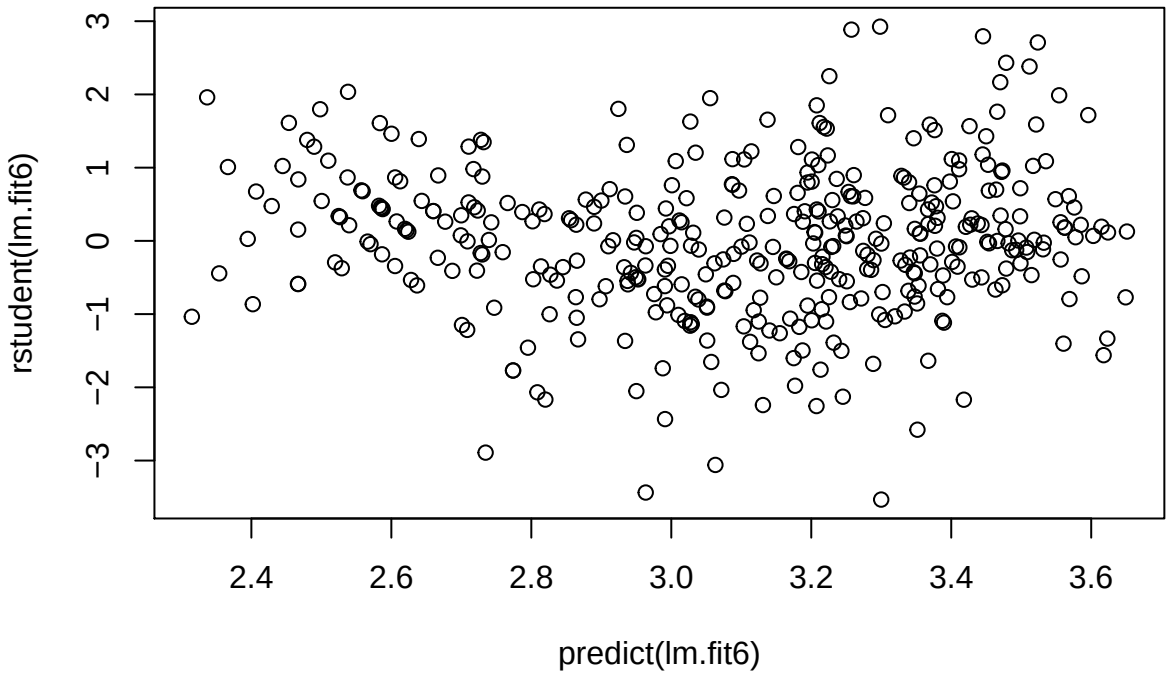
In the residual plot there is less notable pattern than the plot of linear regression terms. The standardized residuals shows the outliers more than 3. In the residuals vs leverage plot indicates there are more than 3 points with the high leverage. However, the Q-Q plot shows there is abnormality of the residuals.

```
lm.fit6<-lm(log(mpg)~cylinders+displacement+horsepower+weight+acceleration+year+origin,
            data=Auto)
summary(lm.fit6)
```

```
##
## Call:
## lm(formula = log(mpg) ~ cylinders + displacement + horsepower +
##     weight + acceleration + year + origin, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40955 -0.06533  0.00079  0.06785  0.33925
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.751e+00  1.662e-01  10.533  < 2e-16 ***
## cylinders    -2.795e-02  1.157e-02  -2.415  0.01619 *
## displacement  6.362e-04  2.690e-04   2.365  0.01852 *
## horsepower   -1.475e-03  4.935e-04  -2.989  0.00298 **
## weight       -2.551e-04  2.334e-05 -10.931  < 2e-16 ***
## acceleration -1.348e-03  3.538e-03  -0.381  0.70339
## year          2.958e-02  1.824e-03  16.211  < 2e-16 ***
```

```
## origin        4.071e-02  9.955e-03   4.089 5.28e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1191 on 384 degrees of freedom
## Multiple R-squared:  0.8795, Adjusted R-squared:  0.8773
## F-statistic: 400.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

```
plot(predict(lm.fit6), rstudent(lm.fit6))
```



Displacement, Horsepower and Weight gives the evidence of non-linear pattern with respect to the mpg. This pattern is nearly close to the log from of mpg. The results shows the better transformation of mpg that graph tends better than model fitting i.e. having good R^2.