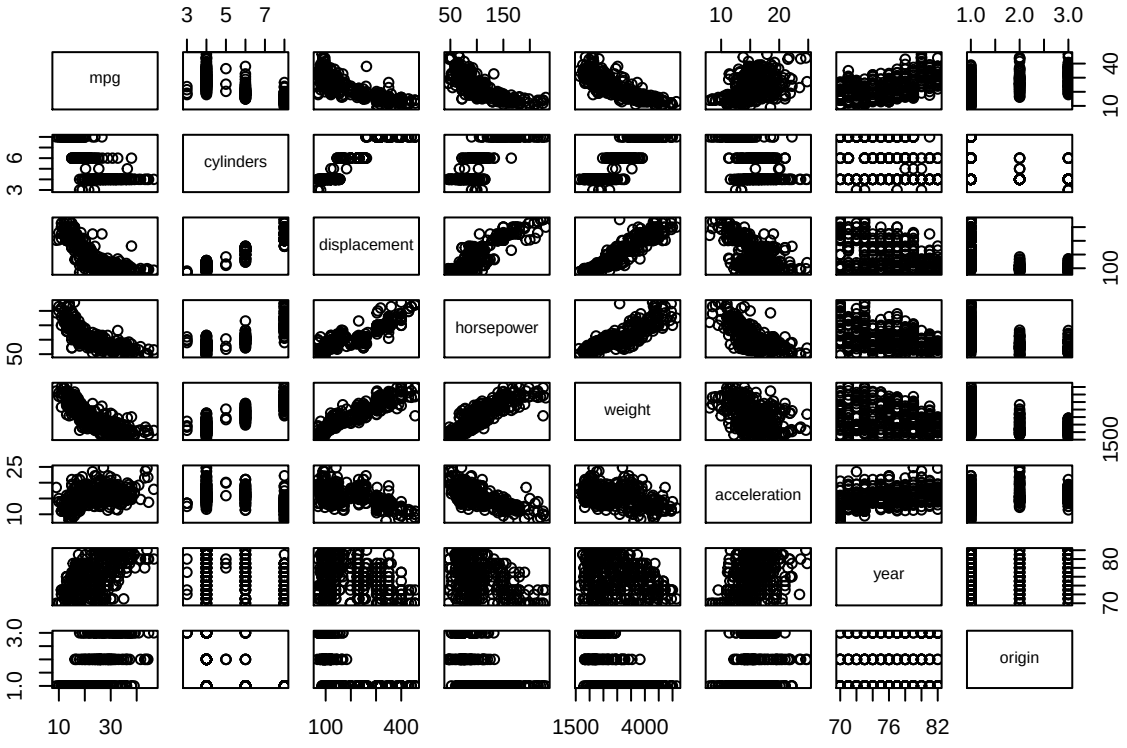
A1Q7

Undergraduate Student

(a)

```
auto <- read.csv('auto.csv', header = TRUE, na.strings = "?")
auto$horsepower <- as.numeric(auto$horsepower)
newAuto <- subset(auto[!is.na(auto$horsepower),], select  = -c(name))
pairs(newAuto)
```



(b)

```
cor(newAuto)
```

```
##                     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)

```
model <- lm(mpg~., data = newAuto)
summary(model)
```

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

Comment: Yes, there is a linear relationship between predictors and the response.However, there are a few predictors such as cylinders, horsepower and acceleration that are not significant, they possibly do not have a relationship with mpg. This needs a further exploration.

(ii)

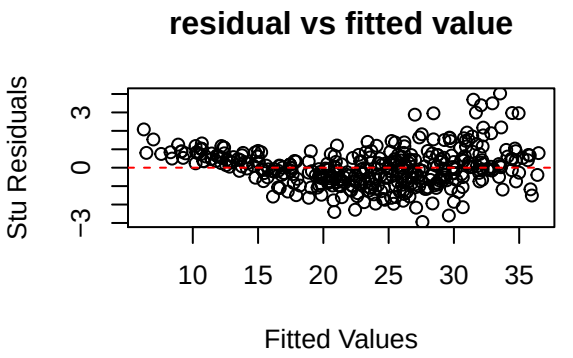
Comment: Displacement, weight, year and origin. They have p-value that are significantly smaller than 0.05.

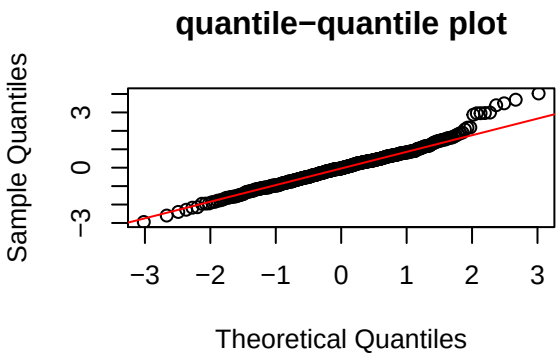
(iii)

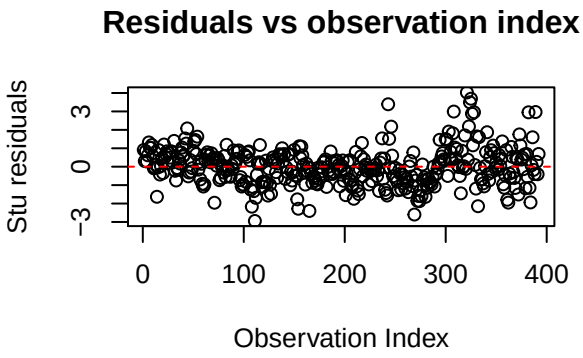
Comment: There will be 0.750773 increase in mpg by increasing year by one while holding other predictors constant.

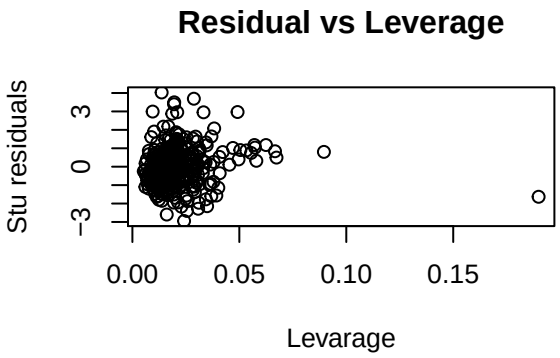
(d)

```
library(MASS)
rownames(newAuto) <- 1:392
par(mfrow=c(2,2))
plot(model$fitted.values, studres(model), xlab = "Fitted Values", ylab = "Stu Residuals",
     main="residual vs fitted value")
abline(h=0, lty=2, col = 'red')

qqnorm(studres(model), main='quantile-quantile plot')
qqline(studres(model), col = 'red')

plot(1:392, studres(model), xlab="Observation Index",ylab="Stu residuals",
     main="Residuals vs observation index")
abline(h=0, lty=2, col = 'red')

hat.model <- lm.influence(model)$hat
plot(hat.model,studres(model), xlab="Levarage", ylab="Stu residuals",
     main = "Residual vs Leverage")
```










```
cbind(newAuto, res=studres(model))[studres(model)>3 | studres(model)< -3,]
```

```
##      mpg cylinders displacement horsepower weight acceleration year origin
## 243 43.1         4           90         48   1985         21.5   78      2
## 321 46.6         4           86         65   2110         17.9   80      3
## 324 44.3         4           90         48   2085         21.7   80      2
## 325 43.4         4           90         48   2335         23.7   80      2
##          res
## 243 3.390068
## 321 4.029537
## 324 3.494823
## 325 3.690246
```

```
cbind(newAuto, lev=hat.model)[hat.model>3/35,]
```

```
##    mpg cylinders displacement horsepower weight acceleration year origin
## 14  14         8          455        225   3086         10.0   70      1
## 29   9         8          304        193   4732         18.5   70      1
##           lev
## 14 0.18991289
## 29 0.08954137
```

Comment: The Residual Plot shows that there is a nonlinear relationship between predictors and the response. There are some potential outliers, as listed above, no.243, 321, 324, 325. The qqplot shows that residuals are following normal distribution with a right tail. The leverage plot shows two points that have higher leverage than others, as listed above, no.9 and 14, but they are within the reasonable range.

(e)

```
summary(model)$cov
```

```
##                (Intercept)     cylinders  displacement    horsepower
## (Intercept)   1.947851e+00 -2.399521e-02  2.238143e-04 -2.557254e-03
## cylinders    -2.399521e-02  9.438004e-03 -1.456053e-04  4.929374e-05
## displacement  2.238143e-04 -1.456053e-04  5.100155e-06 -2.744079e-06
## horsepower   -2.557254e-03  4.929374e-05 -2.744079e-06  1.716520e-05
## weight        4.997596e-05 -2.423003e-06 -1.706933e-07 -3.690479e-07
## acceleration -1.812188e-02  1.457337e-04  6.858721e-06  7.603665e-05
## year         -1.891041e-02  2.265548e-05  2.308244e-06  1.411975e-05
## origin       -1.314375e-02 -8.009994e-04  5.984678e-05 -8.111916e-05
##                      weight  acceleration          year        origin
## (Intercept)   4.997596e-05 -1.812188e-02 -1.891041e-02 -1.314375e-02
## cylinders    -2.423003e-06  1.457337e-04  2.265548e-05 -8.009994e-04
## displacement -1.706933e-07  6.858721e-06  2.308244e-06  5.984678e-05
## horsepower   -3.690479e-07  7.603665e-05  1.411975e-05 -8.111916e-05
## weight        3.839504e-08 -2.800311e-06 -5.035301e-07  1.570354e-06
## acceleration -2.800311e-06  8.823183e-04  3.566167e-05 -2.301055e-05
## year         -5.035301e-07  3.566167e-05  2.346382e-04 -1.270933e-05
## origin        1.570354e-06 -2.301055e-05 -1.270933e-05  6.986038e-03
```

```
summary(lm(mpg~.-acceleration-horsepower, data = newAuto))
```

```
##
## Call:
## lm(formula = mpg ~ . - acceleration - horsepower, data = newAuto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0622  -2.0922  -0.0593   1.8165  13.2758
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.781e+01  4.070e+00  -4.375 1.57e-05 ***
## cylinders    -4.240e-01  3.221e-01  -1.316   0.1889
## displacement  1.176e-02  6.685e-03   1.759   0.0793 .
## weight       -6.506e-03  5.591e-04 -11.637  < 2e-16 ***
## year          7.724e-01  4.977e-02  15.518  < 2e-16 ***
## origin        1.250e+00  2.673e-01   4.676 4.05e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.343 on 386 degrees of freedom
## Multiple R-squared:  0.8189, Adjusted R-squared:  0.8166
## F-statistic: 349.1 on 5 and 386 DF,  p-value: < 2.2e-16
```

```
summary(lm(mpg~.-acceleration-cylinders, data = newAuto))
```

```
##
```

```
## Call:
## lm(formula = mpg ~ . - acceleration - cylinders, data = newAuto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4882 -2.1157 -0.1645  1.8650 13.0544
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.669e+01  4.120e+00  -4.051 6.16e-05 ***
## displacement  1.137e-02  5.536e-03   2.054   0.0406 *
## horsepower   -2.192e-02  1.078e-02  -2.033   0.0428 *
## weight       -6.324e-03  5.685e-04 -11.124  < 2e-16 ***
## year          7.484e-01  5.089e-02  14.707  < 2e-16 ***
## origin        1.385e+00  2.772e-01   4.998 8.80e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.333 on 386 degrees of freedom
## Multiple R-squared:   0.82,  Adjusted R-squared:  0.8177
## F-statistic: 351.7 on 5 and 386 DF,  p-value: < 2.2e-16
```

```
summary(lm(mpg~.-acceleration-cylinders+displacement:horsepower, data = newAuto))
```

```
##
## Call:
## lm(formula = mpg ~ . - acceleration - cylinders + displacement:horsepower,
##     data = newAuto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7264 -1.7408 -0.1554  1.4346 11.8969
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -6.248e+00  3.769e+00  -1.658  0.09823 .
## displacement            -5.331e-02  7.853e-03  -6.788 4.30e-11 ***
## horsepower              -1.617e-01  1.634e-02  -9.894  < 2e-16 ***
## weight                  -4.013e-03  5.477e-04  -7.328 1.39e-12 ***
## year                     7.457e-01  4.491e-02  16.604  < 2e-16 ***
## origin                   8.221e-01  2.504e-01   3.283  0.00112 **
## displacement:horsepower  4.579e-04  4.353e-05  10.520  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.941 on 385 degrees of freedom
## Multiple R-squared:  0.8602, Adjusted R-squared:  0.858
## F-statistic: 394.8 on 6 and 385 DF,  p-value: < 2.2e-16
```

```
summary(lm(mpg~.-acceleration-cylinders+displacement:horsepower+
             displacement:year+displacement:weight, data = newAuto))
```

```
##
```

```
## Call:
## lm(formula = mpg ~ . - acceleration - cylinders + displacement:horsepower +
##     displacement:year + displacement:weight, data = newAuto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8385 -1.5885 -0.0483  1.2275 12.9544
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -3.055e+01  7.517e+00  -4.064 5.86e-05 ***
## displacement              9.165e-02  3.990e-02   2.297   0.0222 *
## horsepower               -9.159e-02  2.160e-02  -4.240 2.81e-05 ***
## weight                   -7.569e-03  1.034e-03  -7.323 1.45e-12 ***
## year                      1.107e+00  9.432e-02  11.737  < 2e-16 ***
## origin                    6.037e-01  2.488e-01   2.426   0.0157 *
## displacement:horsepower   1.573e-04  7.352e-05   2.139   0.0331 *
## displacement:year        -2.033e-03  4.956e-04  -4.103 4.98e-05 ***
## displacement:weight       1.392e-05  3.442e-06   4.043 6.38e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.844 on 383 degrees of freedom
## Multiple R-squared:  0.8699, Adjusted R-squared:  0.8672
## F-statistic: 320.2 on 8 and 383 DF,  p-value: < 2.2e-16
```

```
summary(lm(mpg~.-acceleration-cylinders+displacement:horsepower+
             displacement:weight+year:origin, data = newAuto))
```

```
##
## Call:
## lm(formula = mpg ~ . - acceleration - cylinders + displacement:horsepower +
##     displacement:weight + year:origin, data = newAuto)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.635 -1.711 -0.064  1.401 12.652
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               1.649e+01  7.881e+00   2.093 0.037049 *
## displacement             -7.109e-02  8.886e-03  -8.000 1.49e-14 ***
## horsepower               -1.048e-01  2.134e-02  -4.913 1.33e-06 ***
## weight                   -7.265e-03  1.039e-03  -6.991 1.22e-11 ***
## year                      5.055e-01  9.898e-02   5.107 5.16e-07 ***
## origin                   -1.150e+01  4.161e+00  -2.764 0.005979 **
## displacement:horsepower   2.453e-04  6.907e-05   3.551 0.000431 ***
## displacement:weight       1.278e-05  3.455e-06   3.700 0.000247 ***
## year:origin               1.559e-01  5.338e-02   2.920 0.003703 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.874 on 383 degrees of freedom
## Multiple R-squared:  0.8672, Adjusted R-squared:  0.8644
```

```
## F-statistic: 312.5 on 8 and 383 DF,  p-value: < 2.2e-16
```

Comment: The last model shows that displacement and horsepower, displacement and weight, year and origin interactions are statistically significant.

(f)

```
summary(lm(mpg~log(displacement)+log(horsepower)+log(weight)+log(year)+
             log(origin)+log(cylinders)+log(acceleration), data = newAuto))
```

```
##
## Call:
## lm(formula = mpg ~ log(displacement) + log(horsepower) + log(weight) +
##     log(year) + log(origin) + log(cylinders) + log(acceleration),
##     data = newAuto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5987 -1.8172 -0.0181  1.5906 12.8132
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -66.5643    17.5053  -3.803 0.000167 ***
## log(displacement)  -1.0551     1.5385  -0.686 0.493230
## log(horsepower)    -6.9657     1.5569  -4.474 1.01e-05 ***
## log(weight)       -12.5728     2.2251  -5.650 3.12e-08 ***
## log(year)          54.9857     3.5555  15.465  < 2e-16 ***
## log(origin)         1.5822     0.5083   3.113 0.001991 **
## log(cylinders)      1.4818     1.6589   0.893 0.372273
## log(acceleration)  -4.9831     1.6078  -3.099 0.002082 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.069 on 384 degrees of freedom
## Multiple R-squared:  0.8482, Adjusted R-squared:  0.8454
## F-statistic: 306.5 on 7 and 384 DF,  p-value: < 2.2e-16
```

(g)

```
summary(lm(mpg~sqrt(cylinders)+log(horsepower)+year+weight+origin+
             I(displacement^2)+log(acceleration), data = newAuto))
```

```
##
## Call:
## lm(formula = mpg ~ sqrt(cylinders) + log(horsepower) + year +
##     weight + origin + I(displacement^2) + log(acceleration),
##     data = newAuto)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6605 -1.9139 -0.1281  1.6556 12.3032
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.000e+01  1.052e+01   3.804 0.000166 ***
## sqrt(cylinders)    -3.005e+00  1.151e+00  -2.611 0.009369 **
## log(horsepower)    -9.983e+00  1.477e+00  -6.760 5.15e-11 ***
## year                7.332e-01  4.678e-02  15.674  < 2e-16 ***
## weight             -4.422e-03  6.520e-04  -6.783 4.46e-11 ***
## origin              1.287e+00  2.408e-01   5.344 1.57e-07 ***
## I(displacement^2)   5.321e-05  9.627e-06   5.527 6.03e-08 ***
## log(acceleration) -4.021e+00  1.624e+00  -2.475 0.013747 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.056 on 384 degrees of freedom
## Multiple R-squared:  0.8494, Adjusted R-squared:  0.8467
## F-statistic: 309.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

Comments: transformation can reduce the p-value of those predictors which have high p-value before, making them become statistically significant. We see that after transformation, the R-squared value increases.