DADM Finals

Akhil Patil

December 4, 2018

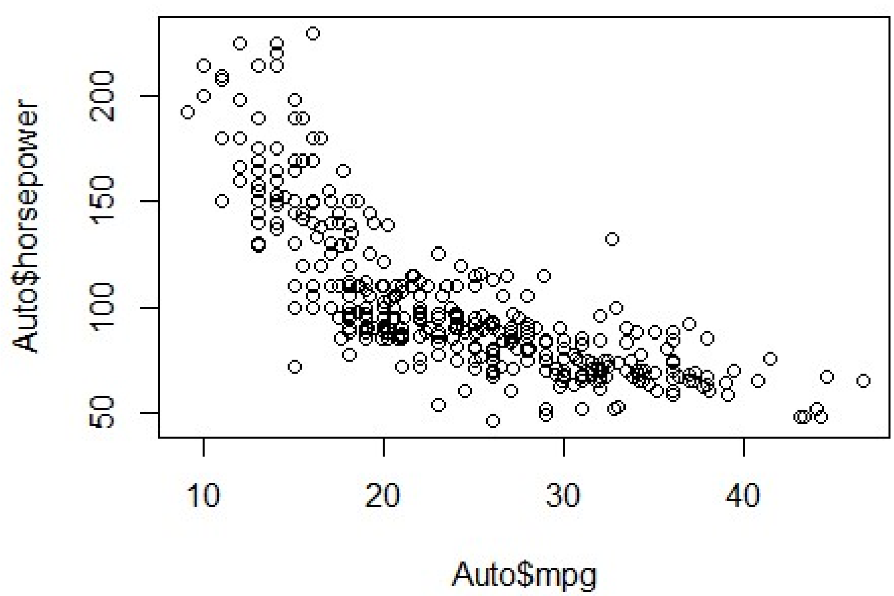
```
################## Q1 - > a ################

#(a) Describe the null hypotheses to which the p-values given in the
following table correspond. Explain what conclusions you can draw based on
these p-values. Your explanation should be phrased in terms of sales, TV,
radio, and newspaper, rather than in terms of the coefficients of the linear
model.

#=>

#As per the table the null hypothesis indicates that the advertising budgets
of "TV", "Radio" , "Newspaper" do not have any effect on sales.
#H(1)0:??1=0
#H(2)0:??2=0 and
#H(3)0:??3=0
# The pvalues for "TV" and "Radio" are highly significant whereas pvalues are
not significant for "Newspaper",
#so we reject H(1)0 and H(2)0 and accept H(3)0.
#In conclusion we can say that Newspaper advertisiong budget does not affect
Sales.

################## Q1 - > c ################

##
#  Included in hand written attachment
##

################## Q1 - > c ################

# Spliting Dataset in Train and Test Data
set.seed(123)
dim(iris)

## [1] 150   5

trainidx = sample(nrow(iris),nrow(iris)*0.80)
train <- iris[trainidx, ]
dim(train)

## [1] 120   5
```

```r
test <- iris[-trainidx, ]
dim(test)

## [1] 30  5

colnames(iris)

## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"
## [5] "Species"

str(iris)

## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1
1 1 1 1 ...

fit.species <- lm(as.numeric(Species) ~ .,data = train)
summary(fit.species)

##
## Call:
## lm(formula = as.numeric(Species) ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57618 -0.15628  0.01322  0.12722  0.55173
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.01972    0.24857   4.102 7.67e-05 ***
## Sepal.Length -0.07769    0.06893  -1.127  0.26206
## Sepal.Width  -0.03597    0.06904  -0.521  0.60340
## Petal.Length  0.22160    0.06635   3.340  0.00113 **
## Petal.Width   0.59234    0.10981   5.394 3.73e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2322 on 115 degrees of freedom
## Multiple R-squared:  0.9213, Adjusted R-squared:  0.9186
## F-statistic: 336.6 on 4 and 115 DF,  p-value: < 2.2e-16

# Predicting some sample species

#Assume you have obtained samples from three plants, with measurements as
listed below. Predict the likelihood that each of these plants belongs to the
species .
```

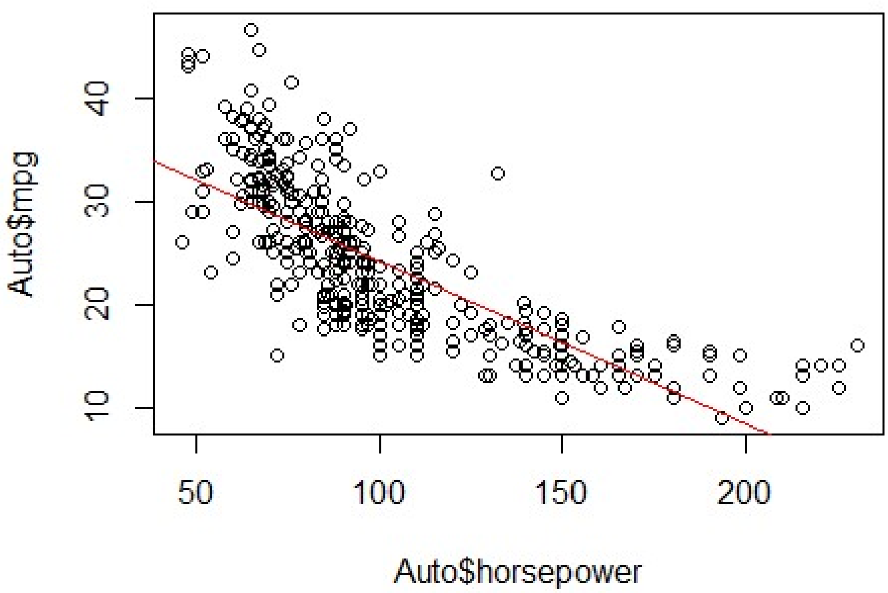
```
plant1 <- data.frame(Sepal.Length=0.4, Sepal.Width=0.8, Petal.Length=4.6,
Petal.Width=1.8)
plant2 <- data.frame(Sepal.Length=6.3, Sepal.Width=2.5, Petal.Length=4.1,
Petal.Width=1.7)
plant3 <- data.frame(Sepal.Length=6.7, Sepal.Width=3.3, Petal.Length=5.2,
Petal.Width=2.3)

predict(fit.species, plant1, type="response")

##        1
## 3.045458

predict(fit.species, plant2, type="response")

##       1
## 2.35592

predict(fit.species, plant3, type="response")

##        1
## 2.895239

################## Q2 - > a ################

library(ISLR)

str(Auto)

## 'data.frame':    392 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : num  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : num  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : num  3504 3693 3436 3433 3449 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year        : num  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ name        : Factor w/ 304 levels "amc ambassador brougham",..: 49 36
231 14 161 141 54 223 241 2 ...

fit.auto = lm(mpg ~ horsepower,data = Auto)

summary(fit.auto)

##
## 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
```

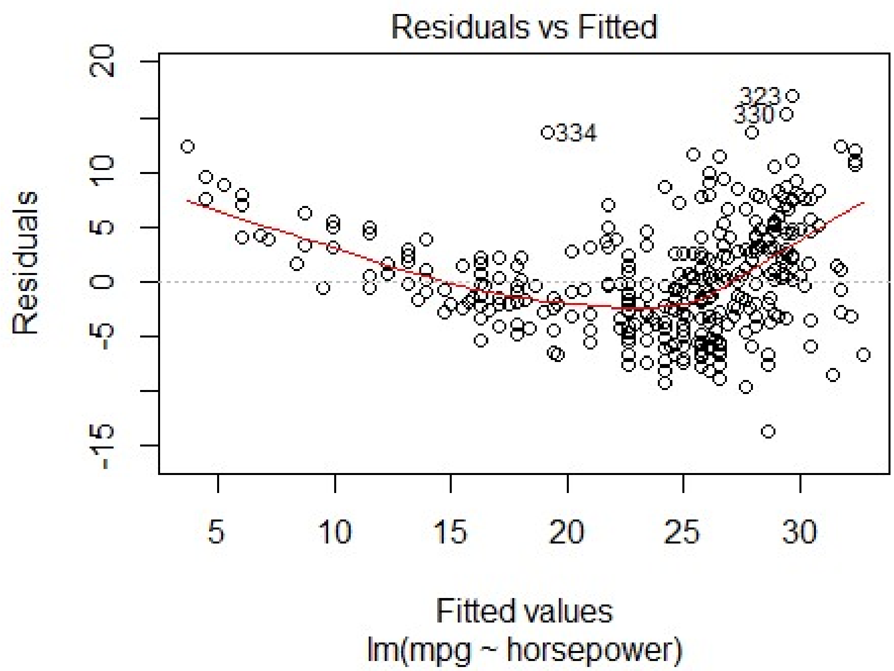
```r
# i. Is there a relationship between the predictor and response?

# => Yes, there is a relationship between mpg(Response) and
horsepower(Predictor) as p-value(2e-16) is less than 0.05 significance level

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

# => From Summary of fit.auto we can say that the R-squared value is around
60% which shows that there is 60% variation in mpg(response variable) because
of horsepower(predictor Variable)

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

plot(Auto$mpg, Auto$horsepower)
```

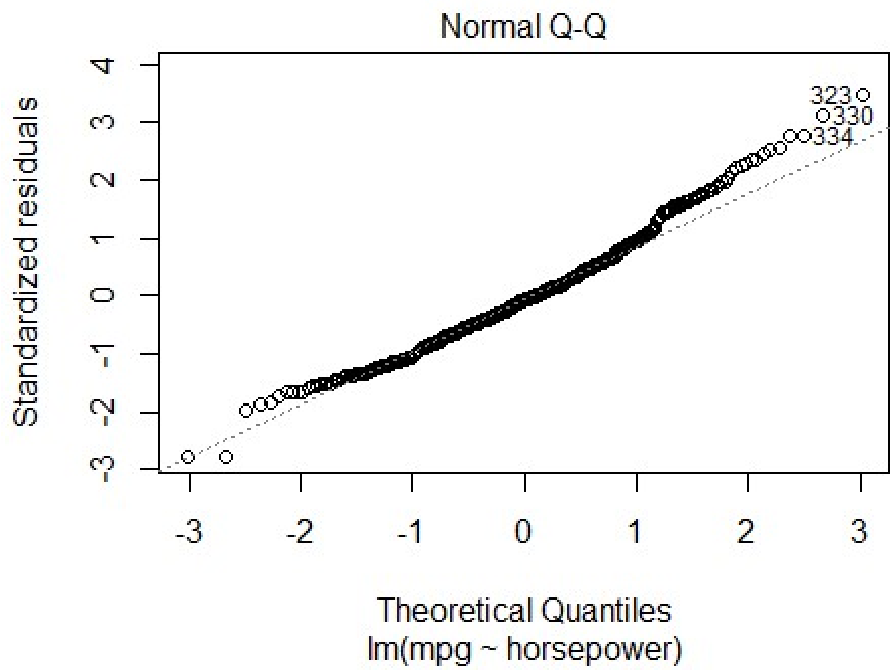


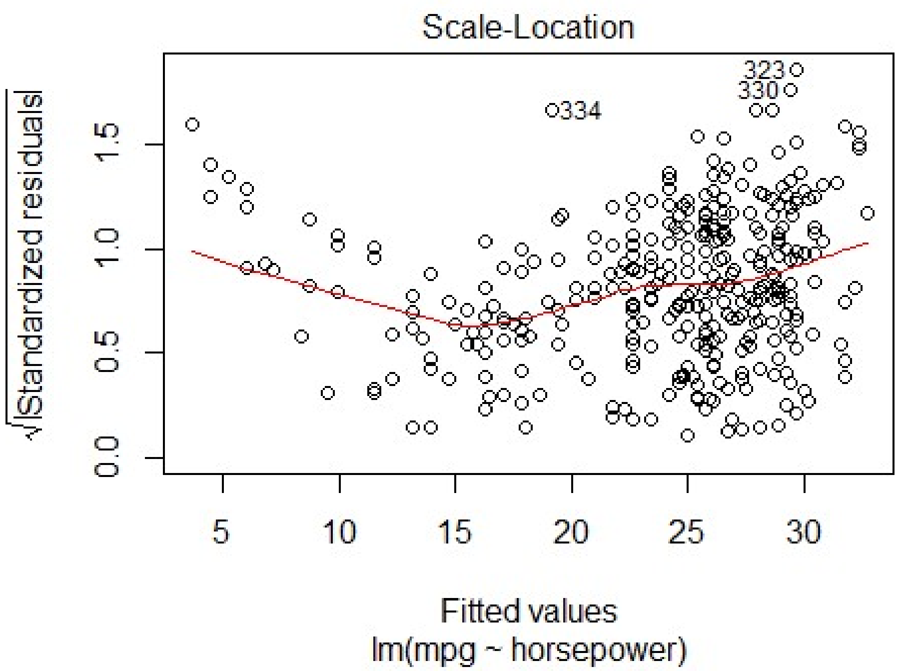
```r
# => From the graph we can say that there is NEGATIVE relationship between mpg(response variable)  & horsepower(predictor Variable) as Horsepower decreases with increase in mpg

#iv. What is predicted mpg associated with horsepower of 98? What is the associated 95% confidence and prediction intervals?

# For Confidence Interval
predict(fit.auto,data.frame(horsepower=c(98)),interval="confidence")

##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108

# For Predicition Interval
predict(fit.auto,data.frame(horsepower=c(98)),interval="predict")

##        fit     lwr      upr
## 1 24.46708 14.8094 34.12476

################## Q2 - > b ################

#(b) Plot the response and the predictor. Display the least square regression line.


plot(Auto$horsepower , Auto$mpg )
abline(fit.auto,col= "red")
```

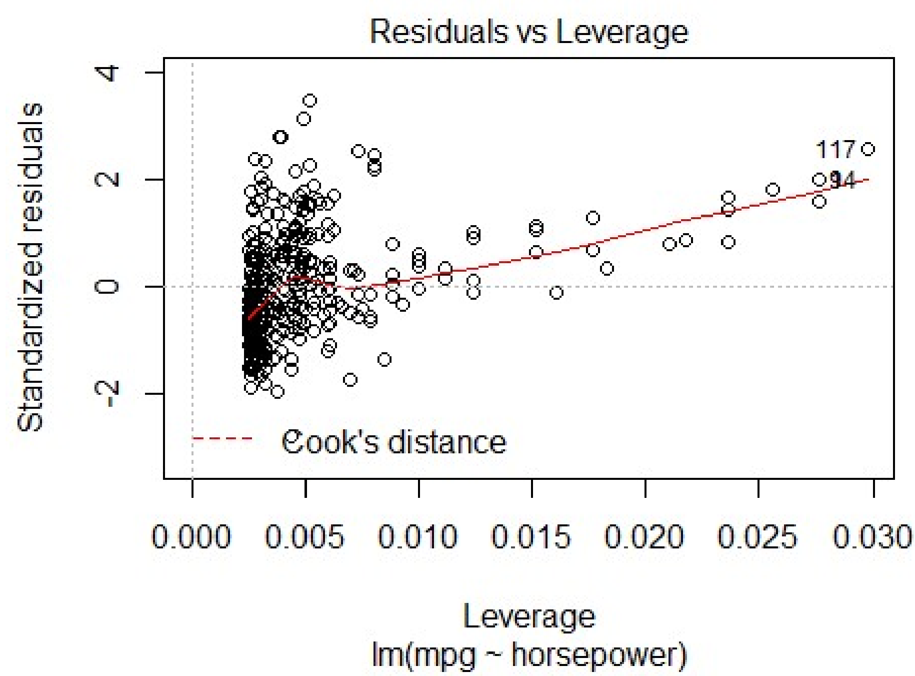


```r
################## Q2 - > c ################

#(c) Use plot () function to produce diagnostic plots of the least square
regression fit. Comment on any problems you see with the fit.

plot(fit.auto)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ horsepower)
334
323
330

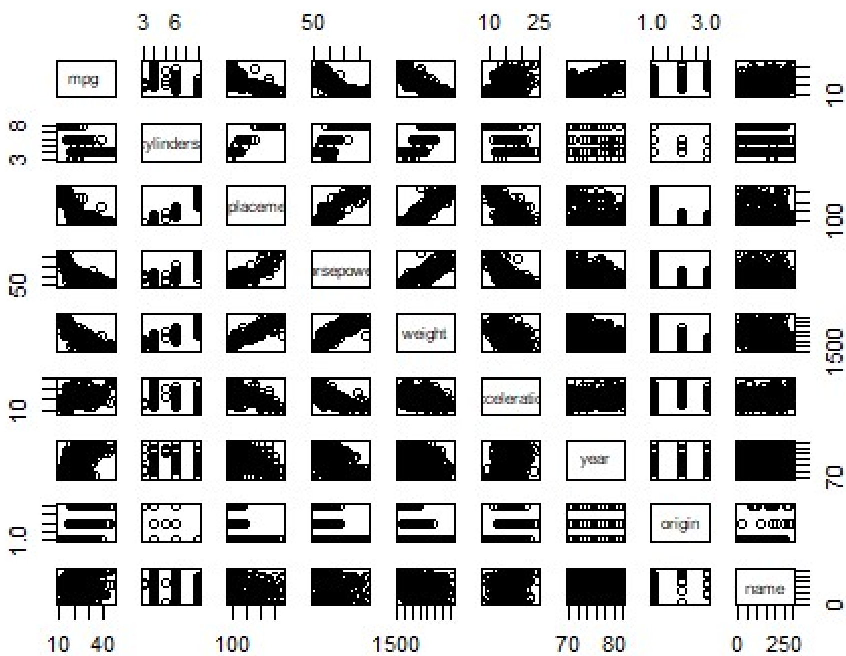

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ horsepower)
323
330
334






```
################## Q3 - > a ################
#This question involves the use of multiple linear regression on the Auto
dataset.
```

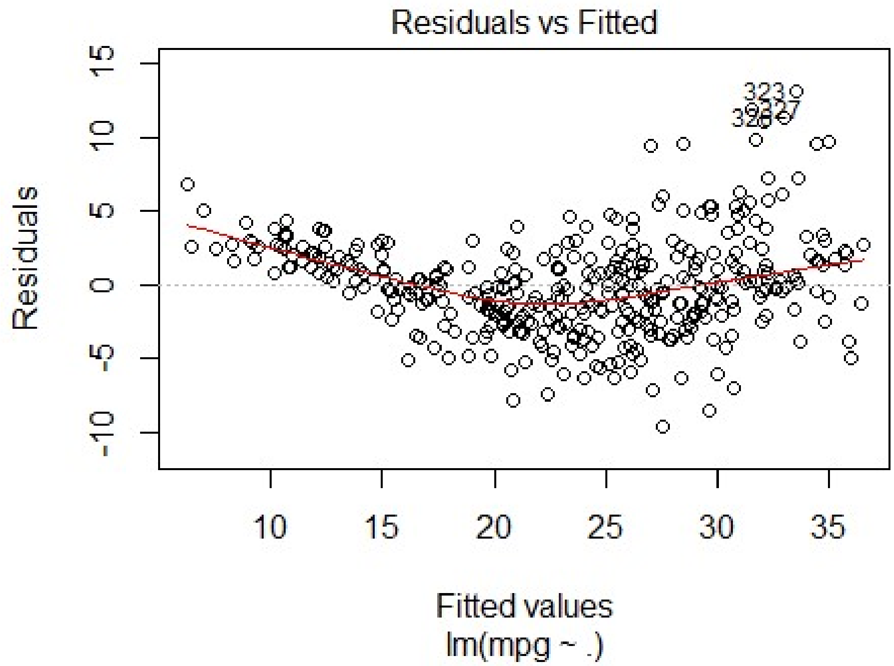
```
#(a) Produce a scatterplot matrix which includes all of the variables in the
dataset.

pairs(Auto)
```

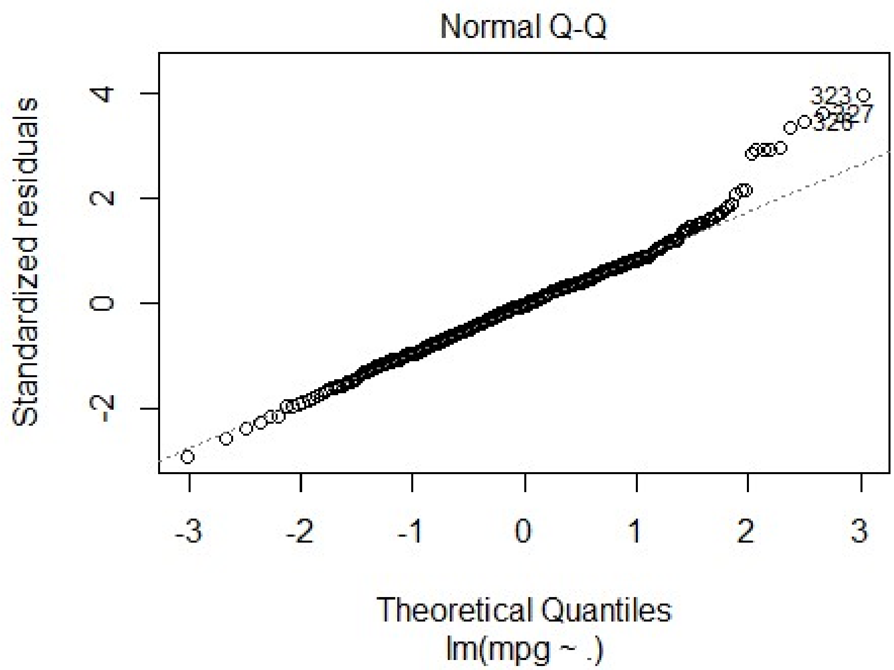


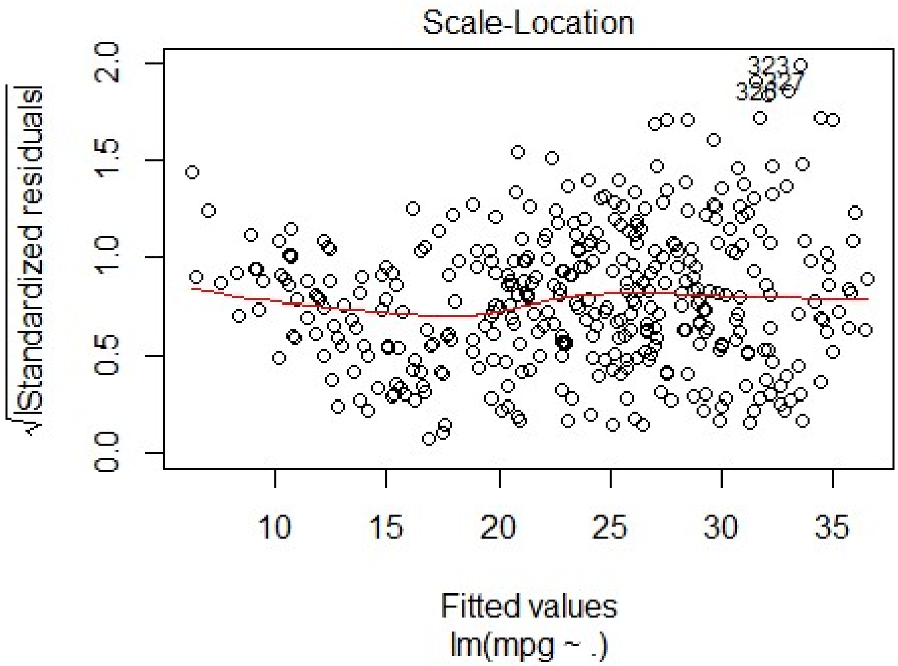

```
################## Q3 - > b ################

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

#Excluding the Name column
Auto$name<-NULL

# Default method is "Pearson"
cor(Auto)

##                     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
```

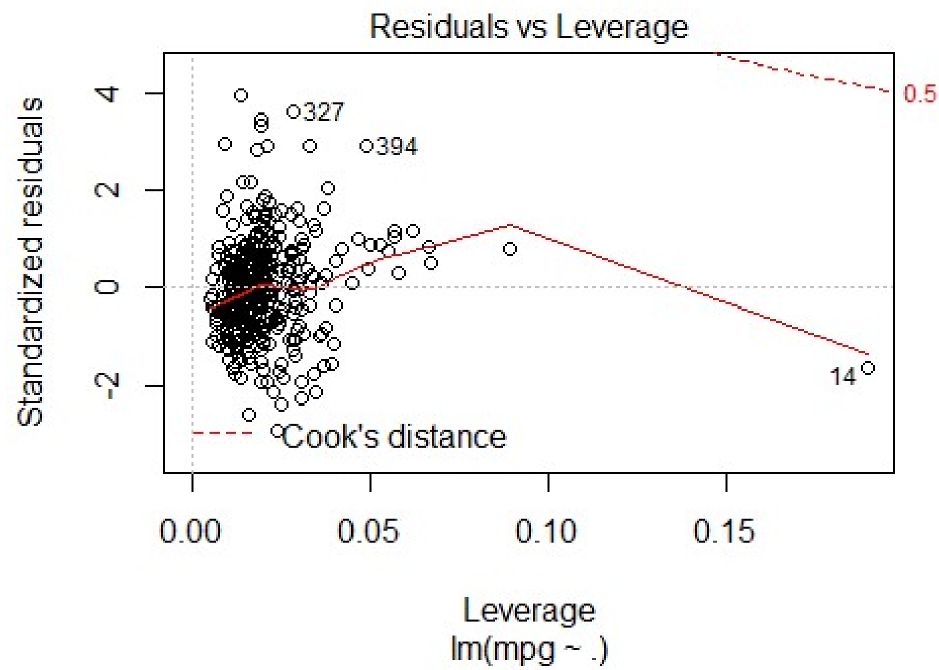
```
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000

################## Q3 - > c ################

#(c) Perform multiple linear regression with mpg as the response and all
other variables except name as the predictors.

# since we have eliminated name valiable earlier we can use same Auto dataset
here.

fit.mul = lm(mpg~.,Auto)

summary(fit.mul)

##
## Call:
## lm(formula = mpg ~ ., 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 predictors and the response?
#=>
#Yes, There is Relationship between the Mpg(response variable) and the
predictors.We can obtain itby testing hypothesis H0:??i=0 ???i . the p-value
corresponding to F-statistic is less than 2.2e-16 , this proves that there is
```

```r
a relationship between "mpg" and other predictors.

#ii. Which predictors appear to have a statistically significant relationship
to the response?
#=>
#From the summary t-statistic we can get that only four predictors are
statically significant to "mpg" they are "displacement" , "weight","year" &
"origin".

#iii. What does the coefficient of the year variable suggest?
#=>
#If the other predictors are constant then there is direct proportion between
"mpg" and "year" i.e with every year cars are more feul efficient by 0.75
mpg/year

################## Q3 - > d ################


# (d) Produce the diagnostic plots of linear regression fit. Comment on any
problems you see with the fit.
plot(fit.mul)
```


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

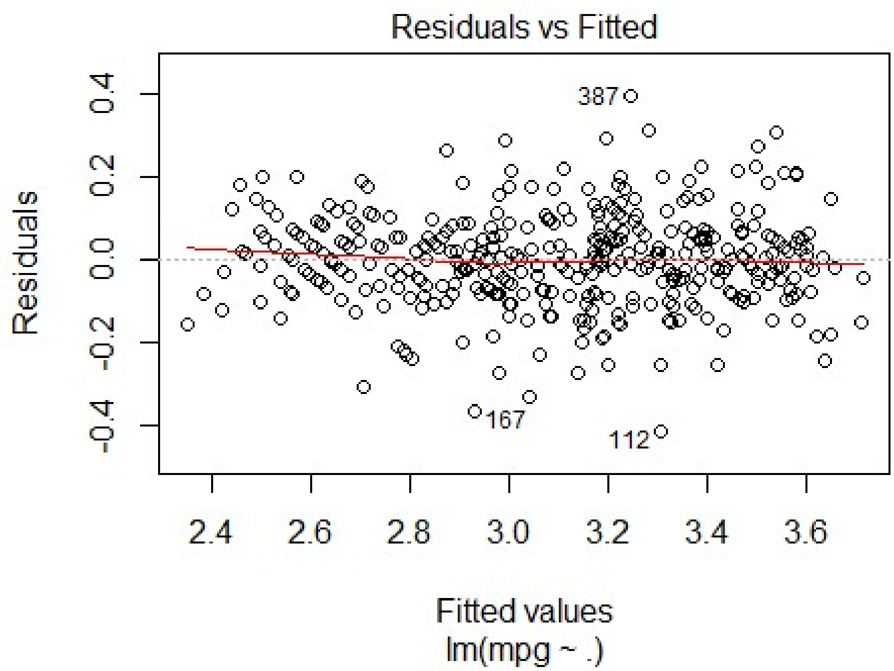

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ .)
323
327
326
4
2
0
-2
-3
-2
-1
0
1
2
3

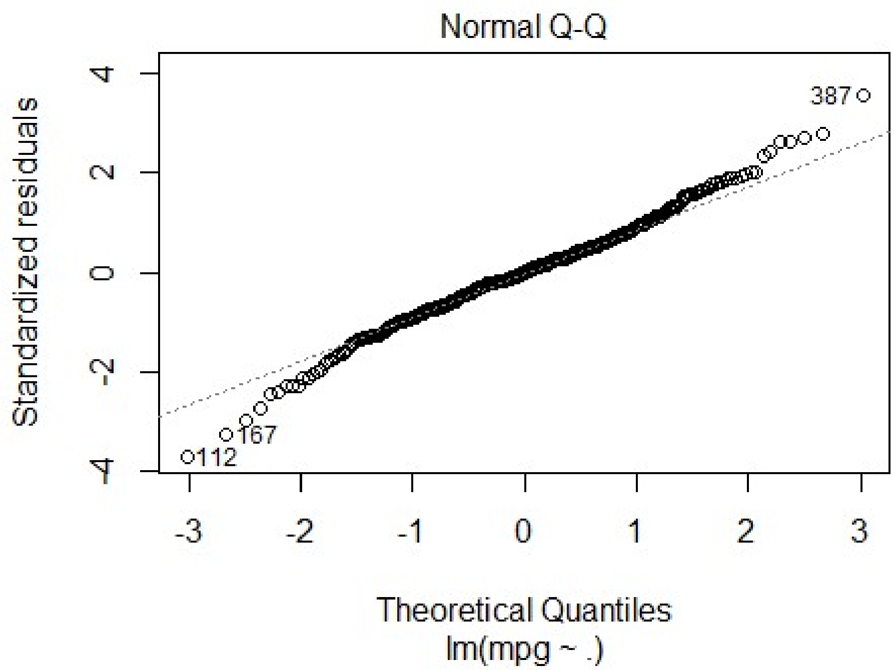


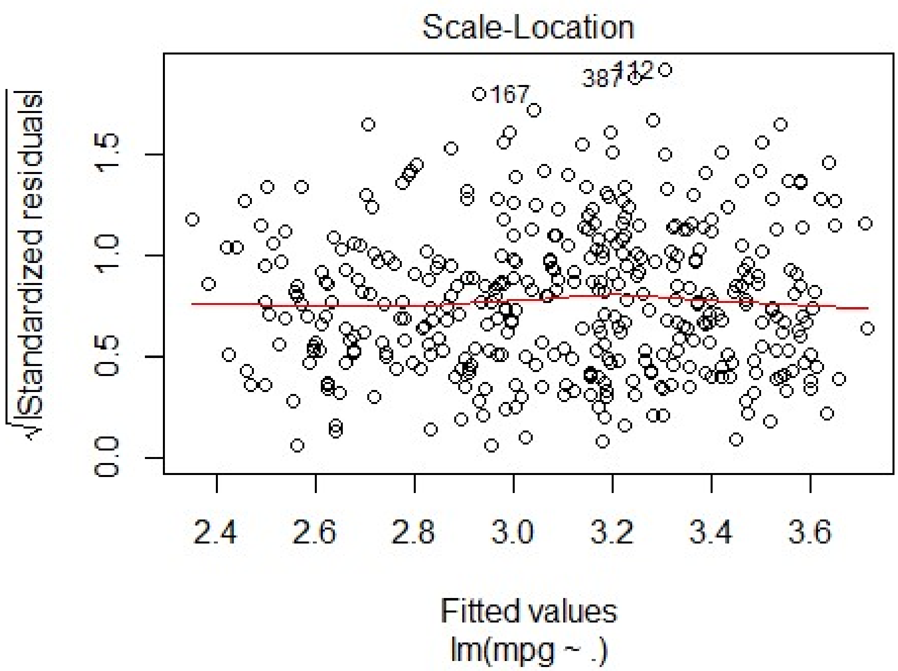


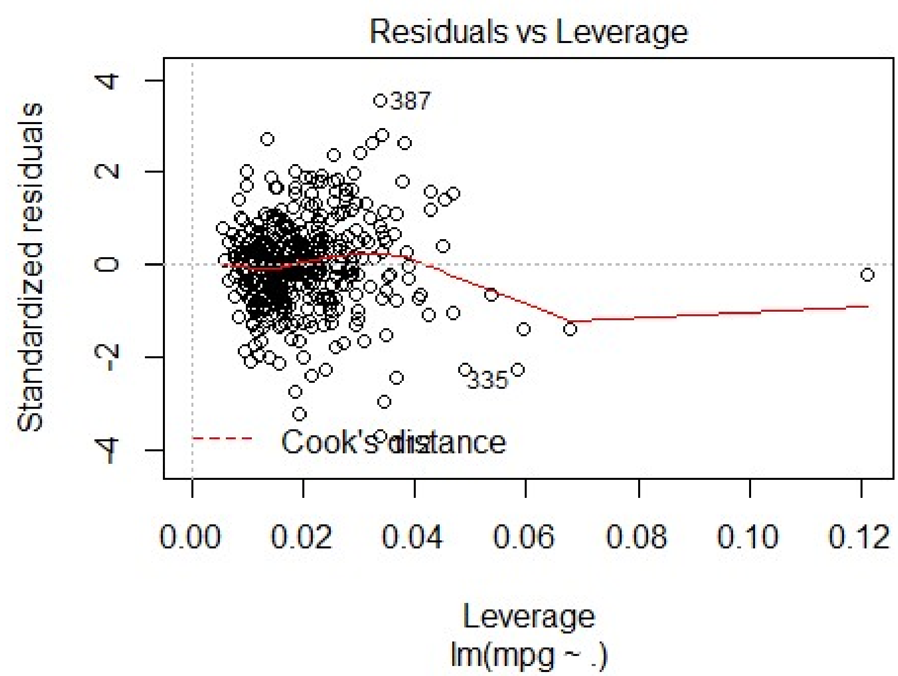

```
#Does the residual plot suggest any unusually large outliers?
#=>
#The residuals vs fitted plot indicates the presence of non-linearity in data
& the residuals versus levegrage plot shows presence of some outliers(>2 or
```

```
<-2) not large outliers.

#Does the leverage plot identify any observations with unusually high
leverage?
#=>
#Yes , the plot of Residuals versus Leverage shows the presence of one high
Leverage point "14".

################## Q3 - > e ################
# (e) Q log(X).sqrt(X),X^2

# Transforming the whole matrix in Log, Sqrt, Sq
Auto.log = log(Auto)
Auto.sqrt = sqrt(Auto)
Auto.sq = (Auto^2)

# log(X)
fit.Auto.log = lm(mpg~.,Auto.log)

summary(fit.Auto.log)

##
## Call:
## lm(formula = mpg ~ ., data = Auto.log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41298 -0.07098  0.00055  0.06150  0.39532
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.155391   0.648230  -0.240  0.81068
## cylinders    -0.082815   0.061429  -1.348  0.17841
## displacement  0.006625   0.056970   0.116  0.90748
## horsepower   -0.294389   0.057652  -5.106 5.18e-07 ***
## weight       -0.569666   0.082397  -6.914 1.98e-11 ***
## acceleration -0.179239   0.059536  -3.011  0.00278 **
## year          2.243989   0.131661  17.044  < 2e-16 ***
## origin        0.044848   0.018821   2.383  0.01767 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1136 on 384 degrees of freedom
## Multiple R-squared:  0.8903, Adjusted R-squared:  0.8883
## F-statistic: 445.3 on 7 and 384 DF,  p-value: < 2.2e-16

plot(fit.Auto.log)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ .)
387
167
112
0.4
0.2
0.0
-0.2
-0.4
2.4
2.6
2.8
3.0
3.2
3.4
3.6


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ .)
387
167
112
4
2
0
-2
-4
-3
-2
-1
0
1
2
3



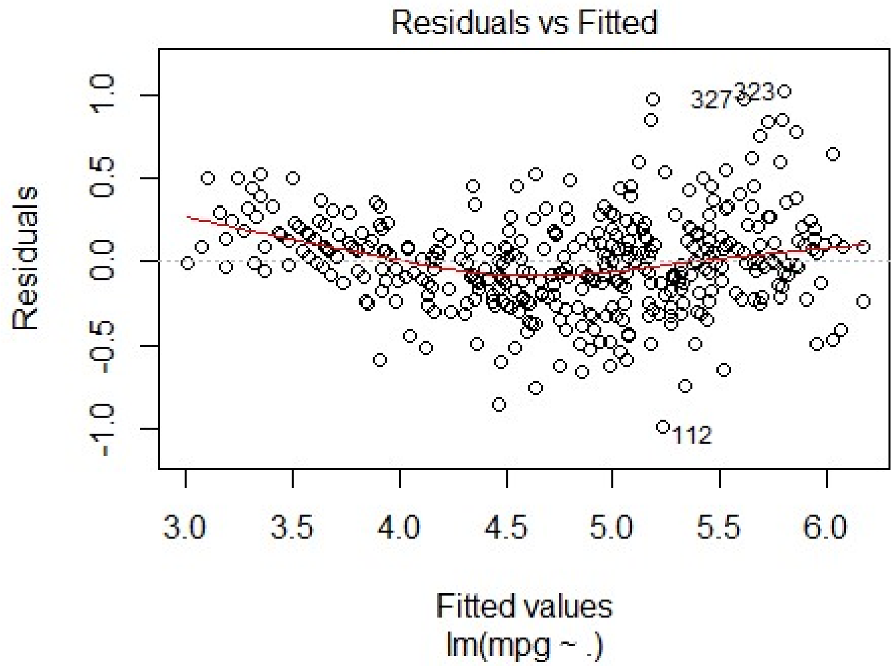


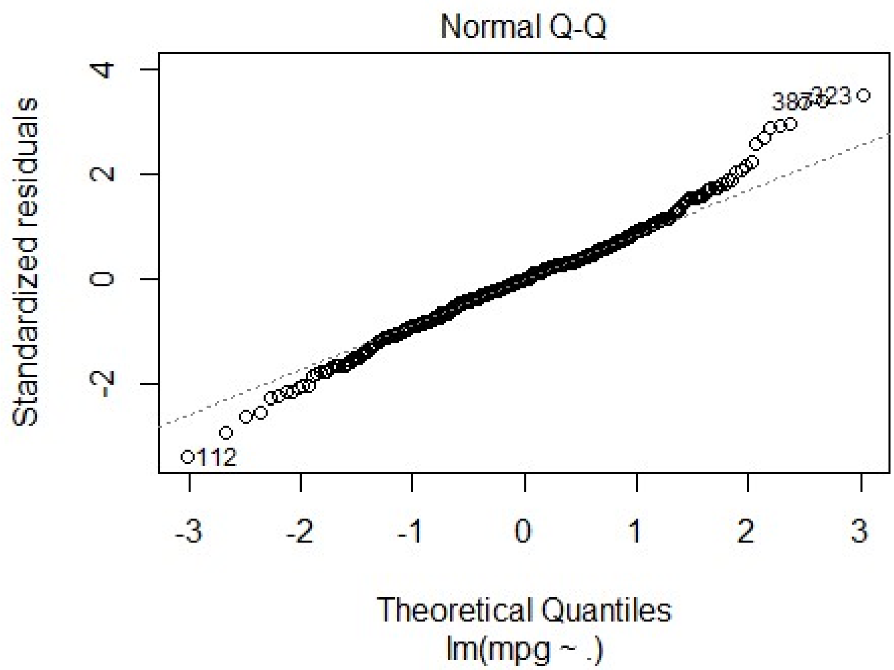
#By Transforming the dataset by Log(X) we can see that the attributes horsepower , weight , acceleration, year, origin appears to be statistically significant with "mpg"

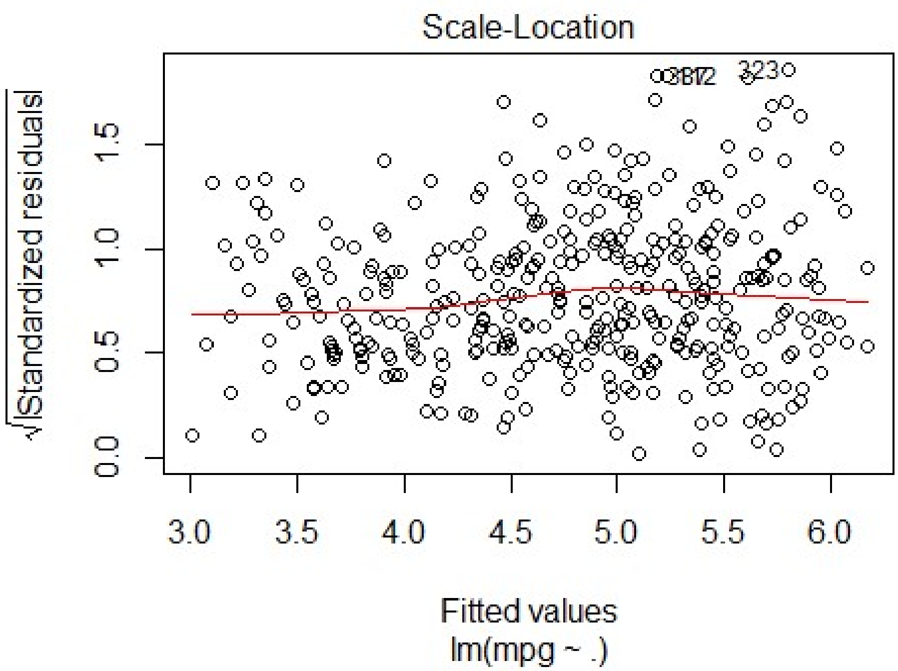
```r
#Residual error is 0.1136
#R^2 = 0.89

#Sqrt(X)
fit.Auto.sqrt = lm(mpg~.,Auto.sqrt)

summary(fit.Auto.sqrt)
```

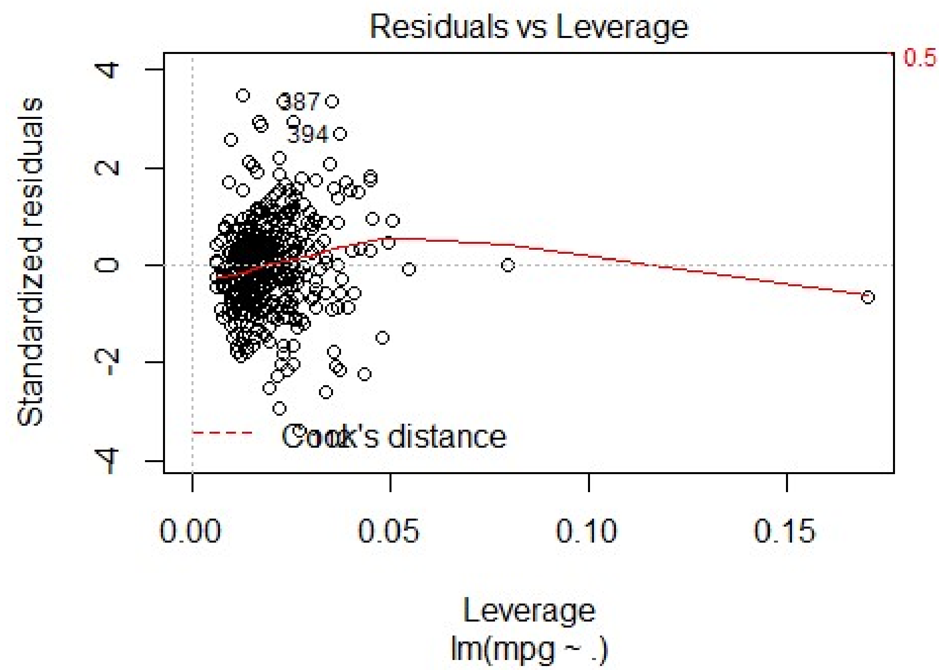
```
##
## Call:
## lm(formula = mpg ~ ., data = Auto.sqrt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98667 -0.17280 -0.00315  0.16145  1.02245
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.949286   0.847481  -2.300 0.021979 *
## cylinders    -0.108552   0.141968  -0.765 0.444964
## displacement  0.019707   0.021182   0.930 0.352752
## horsepower   -0.090896   0.028428  -3.197 0.001502 **
## weight       -0.061414   0.007292  -8.422 7.48e-16 ***
## acceleration -0.107258   0.077048  -1.392 0.164699
## year          1.266015   0.079308  15.963  < 2e-16 ***
## origin        0.272324   0.070883   3.842 0.000143 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2964 on 384 degrees of freedom
## Multiple R-squared:  0.8662, Adjusted R-squared:  0.8638
## F-statistic: 355.1 on 7 and 384 DF,  p-value: < 2.2e-16
```

```r
plot(fit.Auto.sqrt)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ .)
327
323
112
1.0
0.5
0.0
-0.5
-1.0
3.0
3.5
4.0
4.5
5.0
5.5
6.0

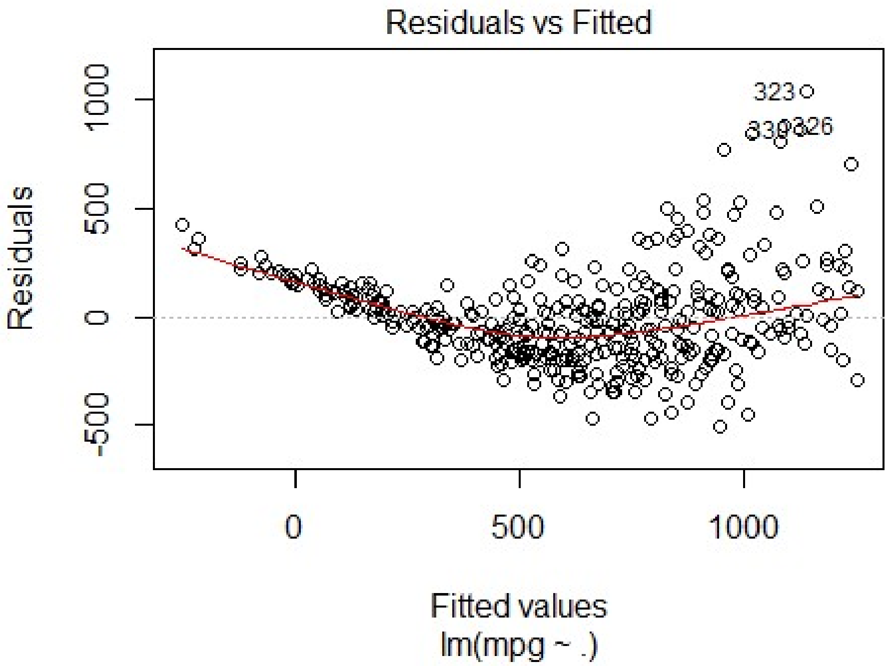

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ .)
387
323
112
4
2
0
-2
-3
-2
-1
0
1
2
3

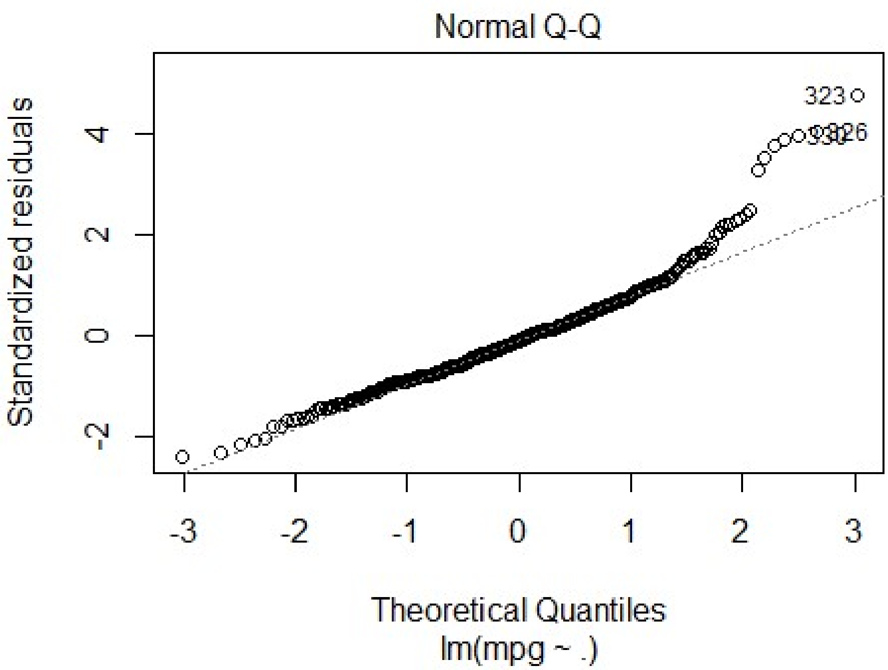


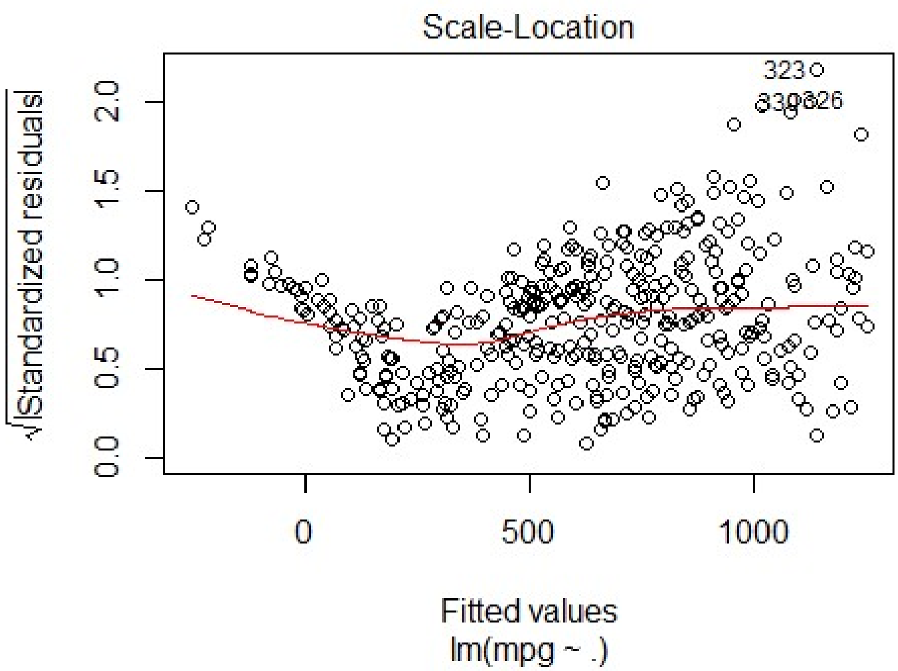


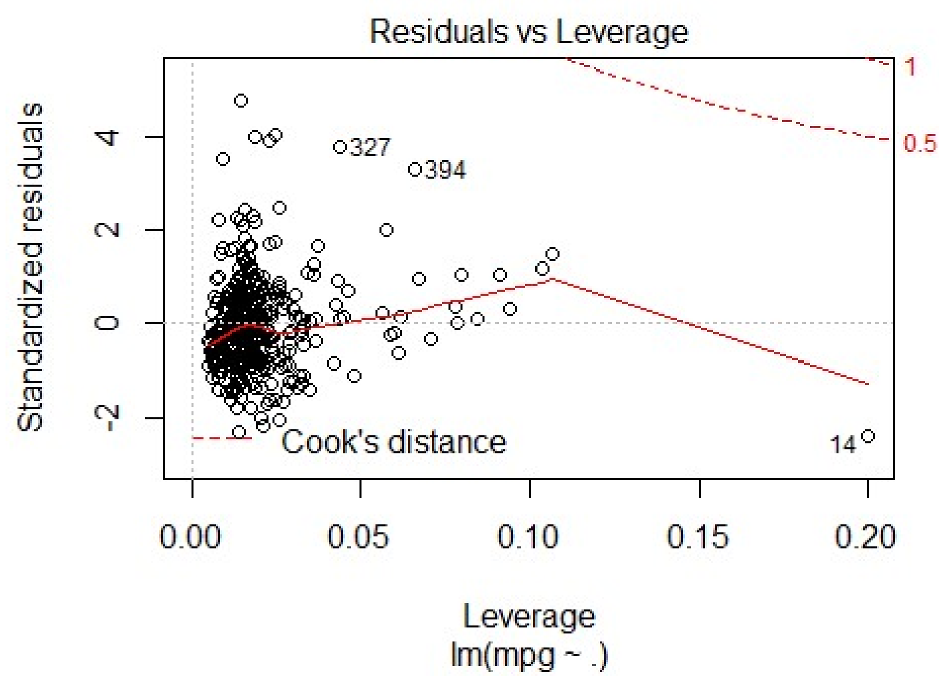
#By Transforming the dataset by Sqrt(X) we can see that the attributes horsepower , weight , year, origin appears to be statistically significant with "mpg" , which is similar to that of X

```r
#Residual error is 0.2964
#R^2 is found to be 0.8662

#(X^2)
fit.Auto.sq = lm(mpg~.,Auto.sq)

summary(fit.Auto.sq)

##
## Call:
## lm(formula = mpg ~ ., data = Auto.sq)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -501.89 -145.36  -18.91  111.41 1034.08
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.523e+02  1.456e+02  -5.165 3.87e-07 ***
## cylinders    -3.746e+00  1.559e+00  -2.403 0.016713 *
## displacement  3.356e-03  8.547e-04   3.926 0.000102 ***
## horsepower    1.279e-04  3.076e-03   0.042 0.966851
## weight       -4.833e-05  5.551e-06  -8.707  < 2e-16 ***
## acceleration  4.892e-01  1.663e-01   2.941 0.003474 **
## year          2.731e-01  2.183e-02  12.513  < 2e-16 ***
## origin        2.608e+01  4.275e+00   6.101 2.57e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 218.8 on 384 degrees of freedom
## Multiple R-squared:  0.7055, Adjusted R-squared:  0.7001
## F-statistic: 131.4 on 7 and 384 DF,  p-value: < 2.2e-16

plot(fit.Auto.sq)
```


Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ .)
323
326
1000
500
0
-500
0
500
1000


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






```
#By Transforming the dataset by Sqrt(X) we can see that the attributes
horsepower , weight , year, origin appears to be statistically significant
with "mpg" , which is similar to that of X
```

```r
#Residual error is 0.2964
#R^2 = 0.8662

################## Q4 - > a ################

#i. Estimate the probability that a student who studies 40 hours and has an
undergrad GPA of 3.5 gets an A in the class.

est.prob <- function(x1,x2){ z <- exp(-6 + 0.05*x1 + 1*x2); return(
round(z/(1+z),2))}

est.prob(40,3.5)

## [1] 0.38

#ii. How many hours would the student in part (i) need to study to have a 50%
chance of getting an A in the class?

##To increase the chance of A without alter the GPA, the student have to
increase the number of hours, so i test a sequence of hours and see how the
chances change.

hours <- seq(40,60,1)
probs <- mapply(hours, 3.5, FUN=est.prob)
names(probs) <- paste0(hours,"h")
probs

##  40h  41h  42h  43h  44h  45h  46h  47h  48h  49h  50h  51h  52h  53h  54h
## 0.38 0.39 0.40 0.41 0.43 0.44 0.45 0.46 0.48 0.49 0.50 0.51 0.52 0.54 0.55
##  55h  56h  57h  58h  59h  60h
## 0.56 0.57 0.59 0.60 0.61 0.62

## By seeing the probs output we can interpret that to have 50% cahnge the
sudent needs to study atleast 50 hours.

################## Q4 - > b ################

##
#  Included in hand written attachment
##

################## Q5 - > a ################

# In this problem, you will develop a model to predict whether a given car
gets high or low gas mileage based on the Auto dataset.
# (a) Create a binary variable, mpg01, that contains a 1 if mpg contains a
value above its median, and a 0 if mpg contains a value below its median.

data(Auto)

mpg01 <- rep(0, length(Auto$mpg))
```

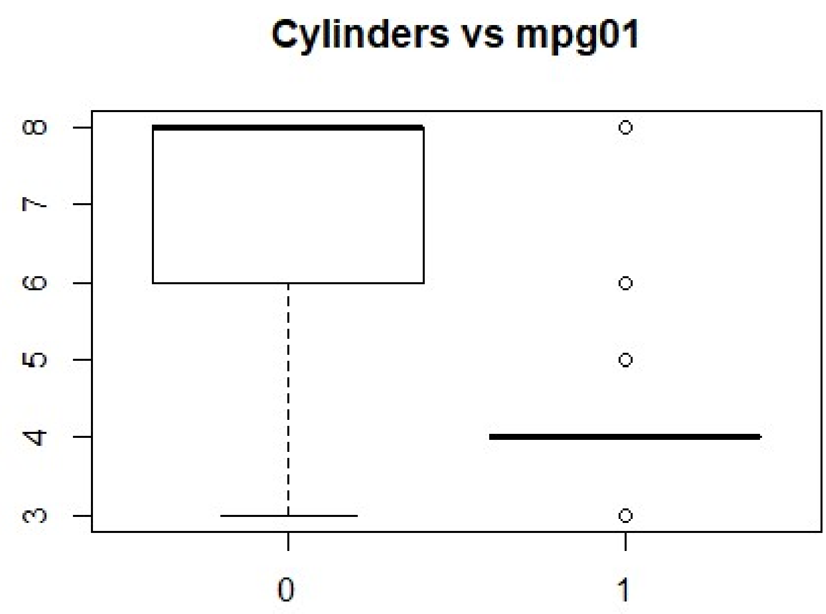
```r
mpg01[Auto$mpg > median(Auto$mpg)] <- 1
Auto <- data.frame(Auto, mpg01)
summary(Auto)
```

```
##       mpg          cylinders      displacement     horsepower
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0
##
##      weight      acceleration        year           origin
##  Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000
##  1st Qu.:2225   1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000
##  Median :2804   Median :15.50   Median :76.00   Median :1.000
##  Mean   :2978   Mean   :15.54   Mean   :75.98   Mean   :1.577
##  3rd Qu.:3615   3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000
##
##                  name         mpg01
##  amc matador       :  5   Min.   :0.0
##  ford pinto        :  5   1st Qu.:0.0
##  toyota corolla    :  5   Median :0.5
##  amc gremlin       :  4   Mean   :0.5
##  amc hornet        :  4   3rd Qu.:1.0
##  chevrolet chevette:  4   Max.   :1.0
##  (Other)           :365
```

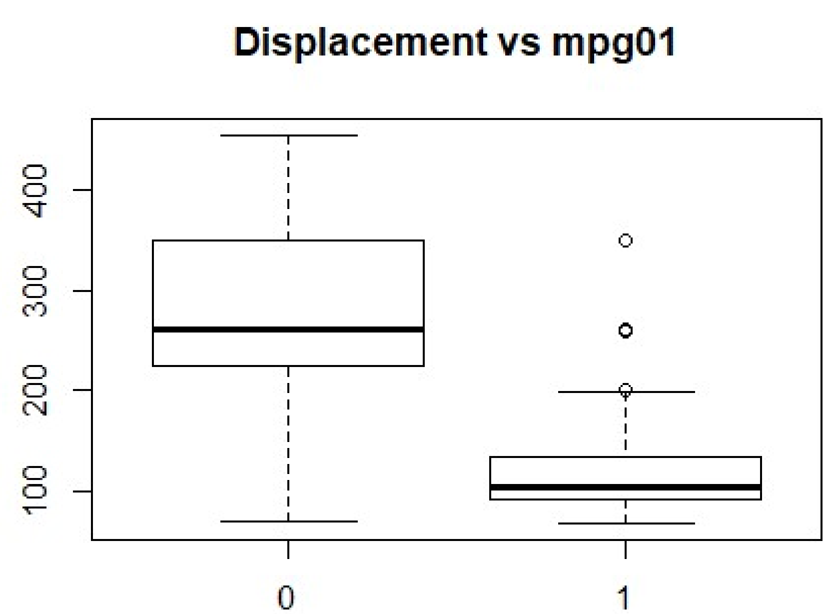
```r
################## Q5 - > b ################

# (b) Explore the data graphically in order to investigate the association
between mpg01 and the other features. Which of the other features seem most
likely to be useful in predicting mpg01? Scatter plots and boxplots may be
useful tools to answer this question.
#Describe your findings.

attach(Auto)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     mpg01
```

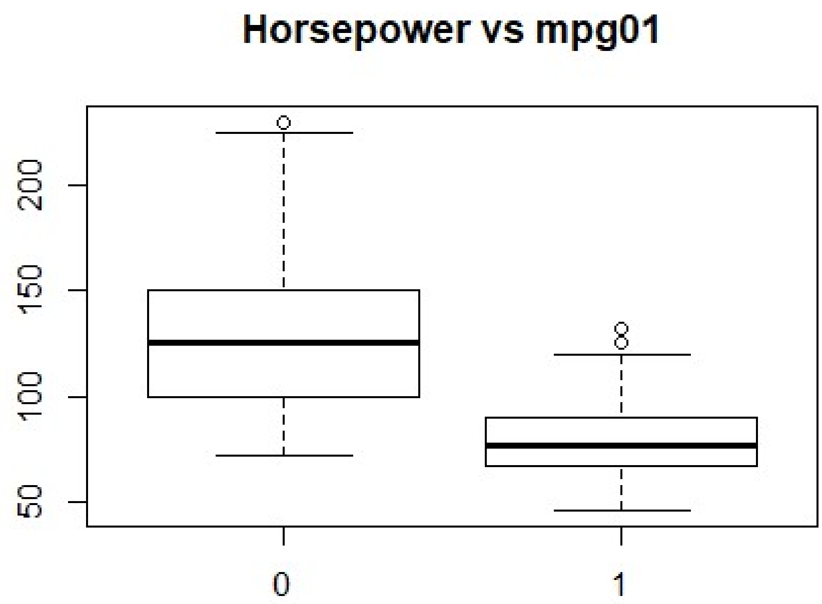
```r
# Boxplots
boxplot(cylinders ~ mpg01, data = Auto, main = "Cylinders vs mpg01")
```



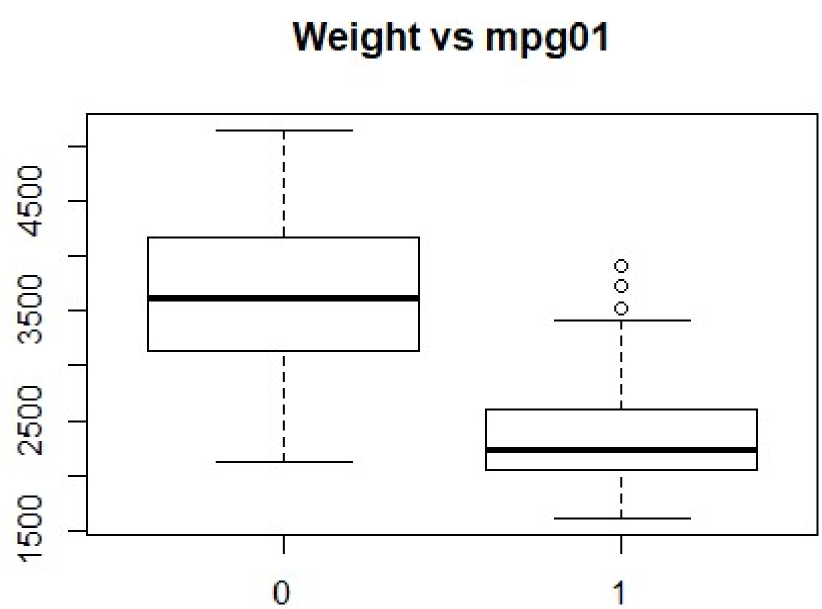

```r
boxplot(displacement ~ mpg01, data = Auto, main = "Displacement vs mpg01")
```



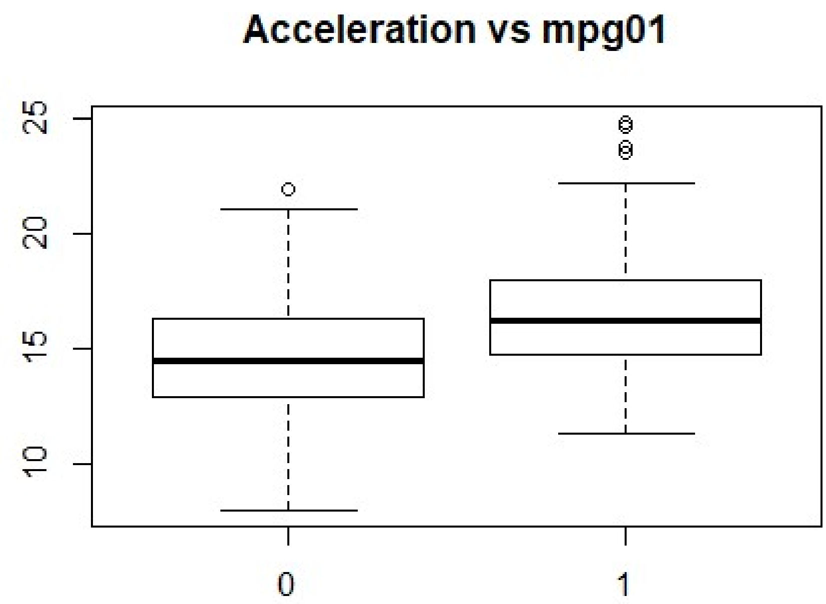

```r
boxplot(horsepower ~ mpg01, data = Auto, main = "Horsepower vs mpg01")
```



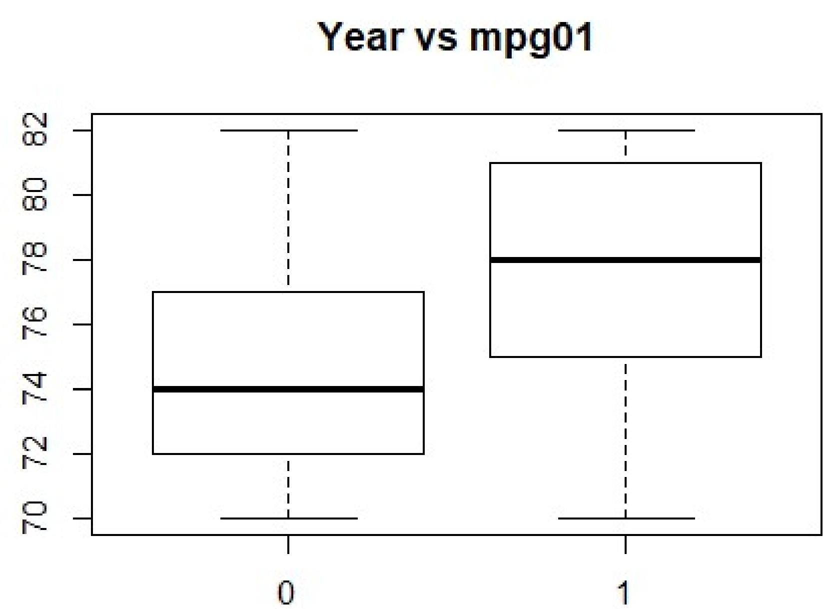

```r
boxplot(weight ~ mpg01, data = Auto, main = "Weight vs mpg01")
```



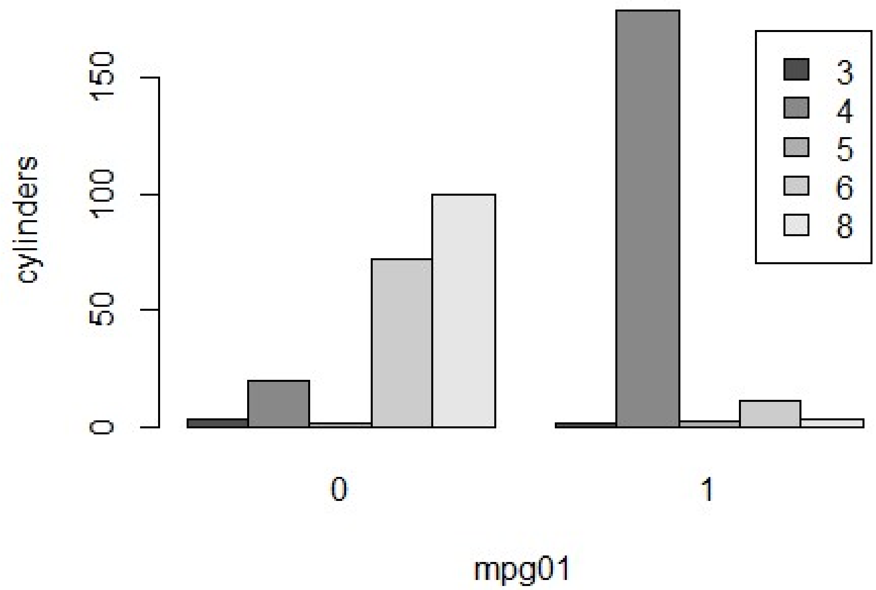

```r
boxplot(acceleration ~ mpg01, data = Auto, main = "Acceleration vs mpg01")
```

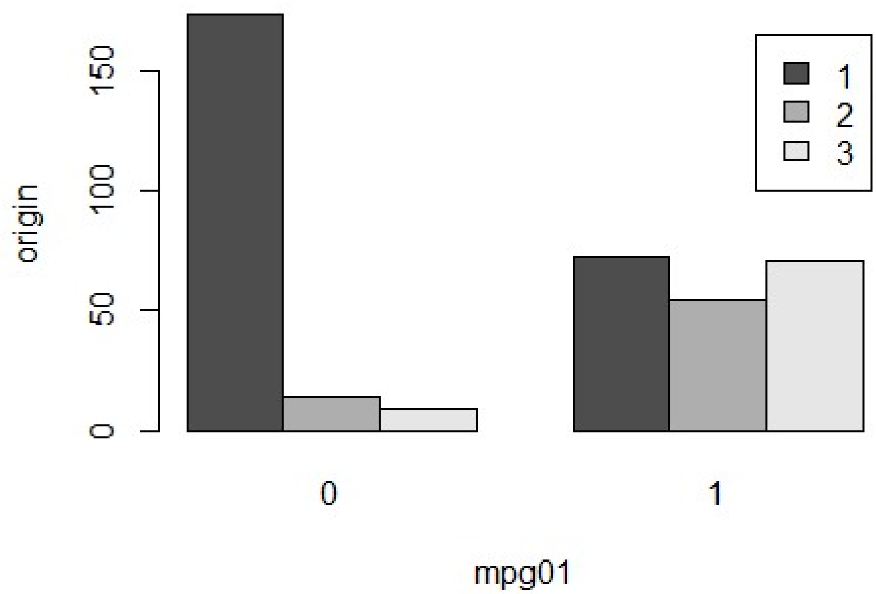



```r
boxplot(year ~ mpg01, data = Auto, main = "Year vs mpg01")
```



```r
# Above Box plots indicates that there exists some relation between "mpg01"
and "cylinders", "weight", "displacement" and "horsepower".

for(i in c("cylinders", "origin")){
  aux <- table(eval(parse(text=i)), mpg01)
  barplot(aux, xlab="mpg01", ylab=i, beside=T,  legend=rownames(aux))
}
```






```
# By the bove Barplots, cylinders and origin also show relation with mpg01.
For instance, on dataset cars of lower mpg are majoraty from origin 1, which
is American.
```

```
################## Q5 - > c ################

#(c) Split the data into training and test set.

# splitting the train and test set into 75% and 25%
set.seed(123)
rows <- sample(x=nrow(Auto), size=.75*nrow(Auto))
trainset <- Auto[rows, ]
testset <- Auto[-rows, ]

dim(trainset)

## [1] 294  10

dim(testset)

## [1] 98 10

################## Q5 - > d ################

#(d) Perform logistic regression on the training data in order to predict
mpg01 using the variables that seemed most associated with mpg01 in
fit.lr <- glm(mpg01 ~ cylinders + weight + displacement + horsepower, data =
trainset, family = binomial)
summary(fit.lr)

##
## Call:
## glm(formula = mpg01 ~ cylinders + weight + displacement + horsepower,
##     family = binomial, data = trainset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4820  -0.1550   0.1131   0.3408   3.2895
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  11.7815748  1.9711595   5.977 2.27e-09 ***
## cylinders     0.1253806  0.3794635   0.330  0.74109
## weight       -0.0023744  0.0008247  -2.879  0.00399 **
## displacement -0.0129040  0.0095681  -1.349  0.17745
## horsepower   -0.0353126  0.0151612  -2.329  0.01985 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 407.35  on 293  degrees of freedom
## Residual deviance: 155.02  on 289  degrees of freedom
## AIC: 165.02
```

```
##
## Number of Fisher Scoring iterations: 7
```

```
probs <- predict(fit.lr, testset, type = "response")
pred.glm <- rep(0,length(probs))
pred.glm[probs > 0.5] <-1
table(pred.glm, testset$mpg01)
```

```
##
## pred.glm  0  1
##        0 46  3
##        1  7 42
```

```
#Q5 -> (d)->  (b). What is the test error of the model obtained?

mean(pred.glm != testset$mpg01)*100
```

```
## [1] 10.20408
```

```
#Test error rate in logistic regression is  10.20%
```