1. For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(a) The sample size n is extremely large, and the number of predictors p is small

Better: As the sample size is extremely large and number of predictors is small we have low variance, so to minimize error we just need to minimize bias which could be done by a flexible model.

(b) The number of predictors p is extremely large, and the number of observations n is small.

Worse: As the number of observations are small by default we have high variance as in adding in more value will significantly change the estimator thus using a flexible method would end up over fitting the data and in turn will increase MSE.

(c) The relationship between the predictors and response is highly non-linear.

Better: As the relationship is highly non-linear a flexible model will be able to capture the relationship better. Here the dominating error part in MSE is bias error so to minimize it we should use more flexible model.

(d) The variance of the error terms, i.e. σ2 =Var(ε), is extremely high.

Worse: As the variance is high a flexible model would end up higher MSE. Thus, here to minimize variance error it is better to use inflexible models.

2. We now revisit the bias-variance decomposition.

(a) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

Please see Fig.1 on next page

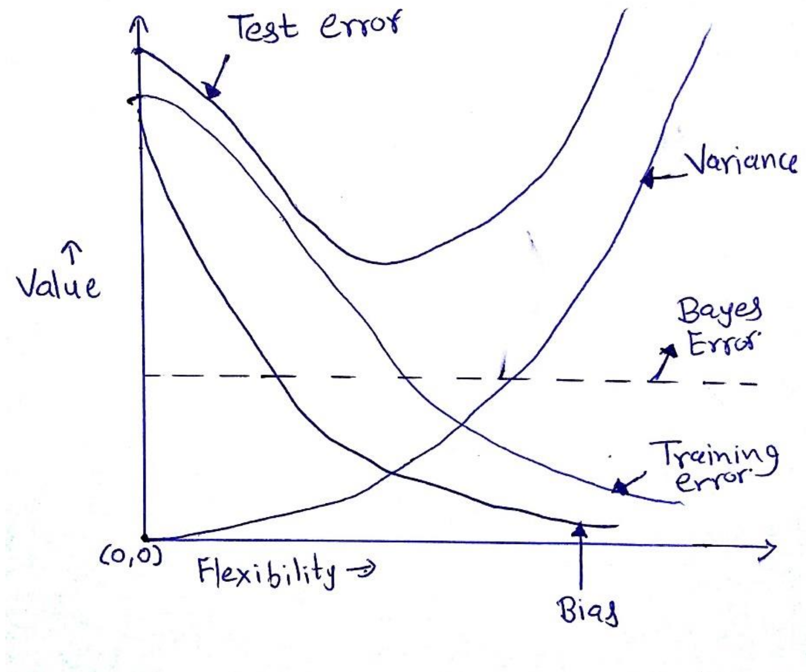


Fig.1 Sketch for various errors

(b) Explain why each of the five curves has the shape displayed in part (a).

Bayes Error (irreducible error): Irreducible error is the lower limit of the error as in we cannot get an error less than this. Variance of this error is one part of the test error which comes due to incapability to get perfect estimator. When the training error goes less than this error we can say the method has overfitted. The Bayes error is same as irreducible error but for classification problem belongs to [0,1).

Variance: As the flexibility of the estimator increases the variance increases monotonically, this happens because even slight change in training data ends up in significant change in the estimator.

Bias (Squared): In the start when estimator is inflexible it is biased towards finding linear relationship between predictors and estimator, but as the flexibility increases it monotonically decreases the bias towards linear relationship

Training error: As the flexibility keeps on increasing the estimator encompasses more and more training point thus decreasing the training error, hence the graph decreases monotonically.

Testing Error: Testing error depends on Bayes error, variance and bias. The Bayes error is constant with respect to flexibility but the variance and bias changes in opposite directions with change in flexibility thus the testing error generally decreases in when bias error is more dominating (less flexible estimators), then it reaches a point from which variance error is more dominating (more flexible estimators) thus increasing the error. This point is the best error we can practically get from the estimator.

3. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?

Parametric Models:

- A two-step approach process where parameters are randomly assumed first and then they are tuned by using trained data
- Explicitly assume the form of estimator
- Reduce the problem of finding estimators to finding few parameters

Non- Parametric Models:

- Try to find an estimator close to data points
- Do not assume the form of estimator
- A very large number of observation required

Advantage of using parametric models for regression or classification is it is easier to implement and requires only small amount of data compared to non-parametric models. While the disadvantage is often they can't find true estimator.

4. This exercise relates to the College data set, which can be found in the file College.csv. It contains several variables for 777 different universities and colleges in the US.

(a) Use the read.csv() function to read the data into R. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.
Code

```
library(ISLR)
College<-read.csv("College.CSV",header=TRUE,sep=",")
College
```

(b)Look at the data using the fix() function. You should notice that the first column is just the name of each university. We don't really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands:
Code

```
rownames (College)=College [,1]
summary(College)
fix(College)
```

```
rownames(College)
```

Output: first 6 row names

```
  [1] "Abilene Christian University"                  "Adelphi University"
  [3] "Adrian College"                                "Agnes Scott College"
  [5] "Alaska Pacific University"                     "Albertson College"
```

c) i. Use the summary() function to produce a numerical summary of the variables in the data set.

Code

```
summary(College)
```

Output

```
                             X       Private        Apps           Accept
Enroll       Top10perc
 Abilene Christian University:  1   No :212   Min.   :   81   Min.   :   72
Min.   :  35   Min.   : 1.00
 Adelphi University          :  1   Yes:565   1st Qu.:  776   1st Qu.:  604
1st Qu.: 242   1st Qu.:15.00
 Adrian College              :  1             Median : 1558   Median : 1110
Median : 434   Median :23.00
 Agnes Scott College         :  1             Mean   : 3002   Mean   : 2019
Mean   : 780   Mean   :27.56
 Alaska Pacific University   :  1             3rd Qu.: 3624   3rd Qu.: 2424
3rd Qu.: 902   3rd Qu.:35.00
 Albertson College           :  1             Max.   :48094   Max.   :26330
Max.   :6392   Max.   :96.00
 (Other)                     :771
   Top25perc      F.Undergrad     P.Undergrad         Outstate       Room.B
oard       Books            Personal
 Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :
1780   Min.   :  96.0   Min.   : 250
 1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:
3597   1st Qu.: 470.0   1st Qu.: 850
 Median : 54.0   Median : 1707   Median :  353.0   Median : 9990   Median :
4200   Median : 500.0   Median :1200
 Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :
4358   Mean   : 549.4   Mean   :1341
 3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:
5050   3rd Qu.: 600.0   3rd Qu.:1700
 Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :
8124   Max.   :2340.0   Max.   :6800

      PhD            Terminal       S.F.Ratio      perc.alumni        Expen
d        Grad.Rate
 Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   :
3186   Min.   : 10.00
 1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.:
6751   1st Qu.: 53.00
```

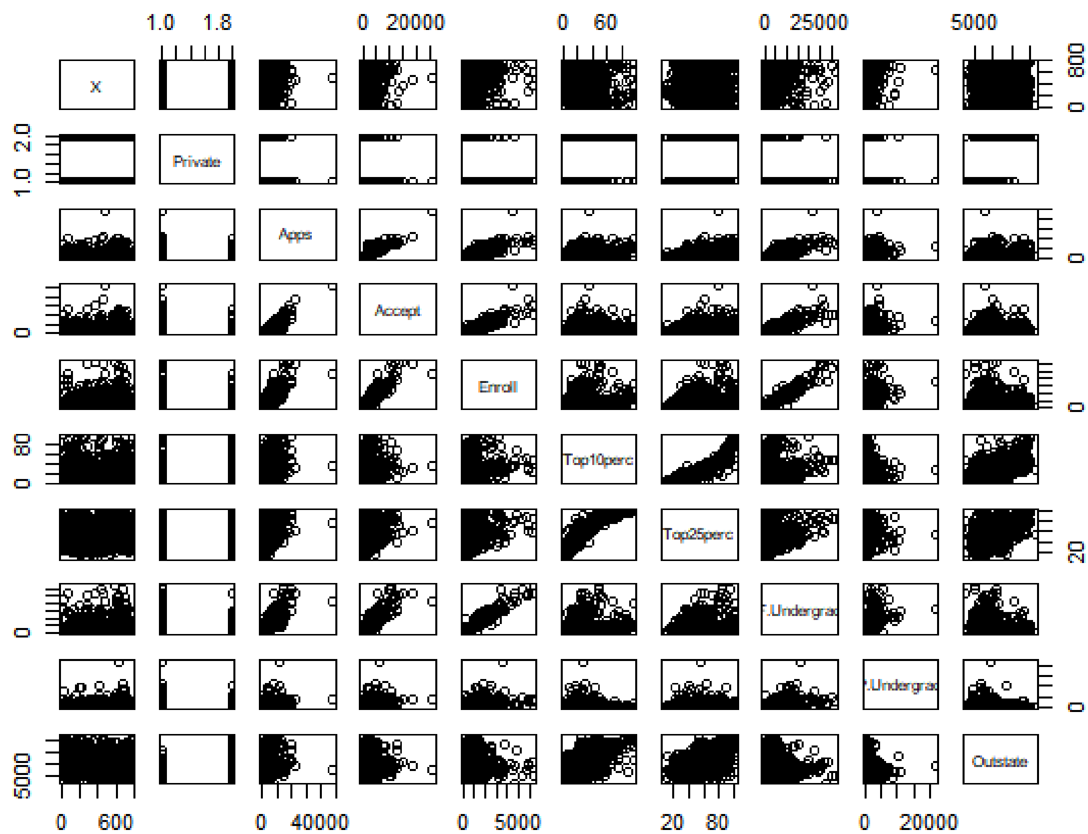
```
 Median : 75.00   Median : 82.0   Median :13.60   Median :21.00   Median :
8377   Median : 65.00
 Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   :
9660   Mean   : 65.46
 3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:1
0830   3rd Qu.: 78.00
 Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :5
6233   Max.   :118.00
```

ii. Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].

Code

```
Pairs(College[,1:10])
```

Output

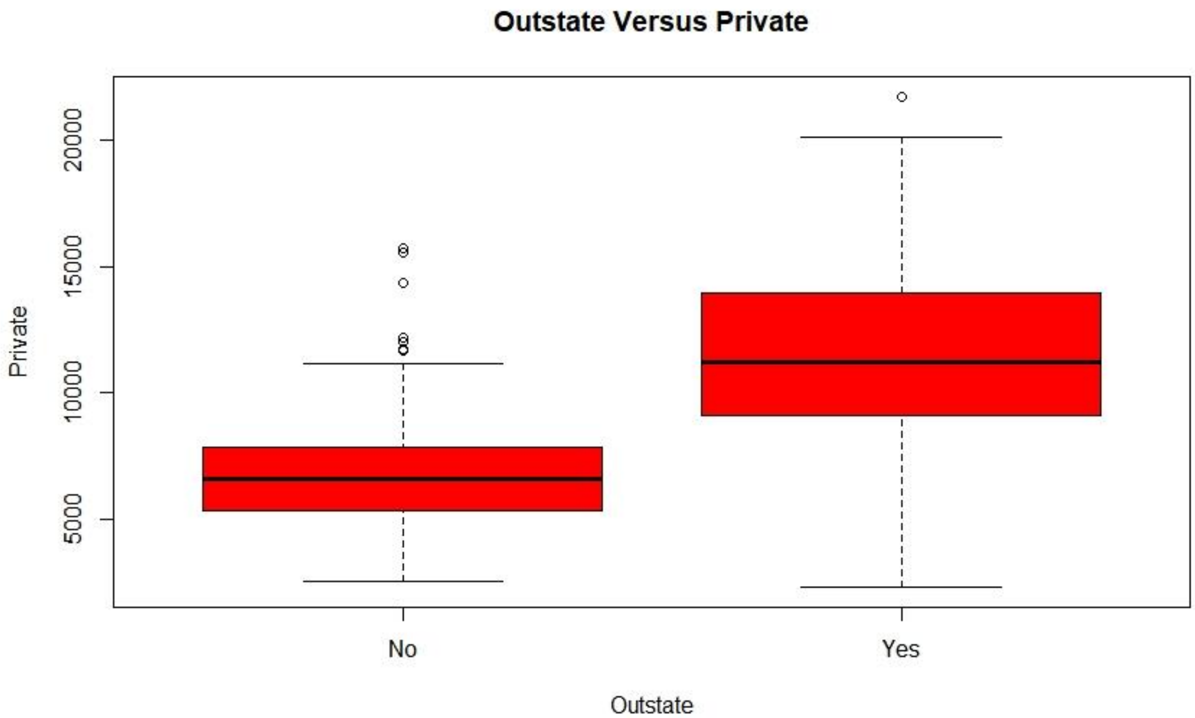


iii. Use the plot() function to produce side-by-side boxplots of Outstate versus Private.

Code

```
boxplot(College$Outstate ~ College$Private, col="red", main="Outstate Versus Private",
      ylab="Private", xlab="Outstate")
```

Output



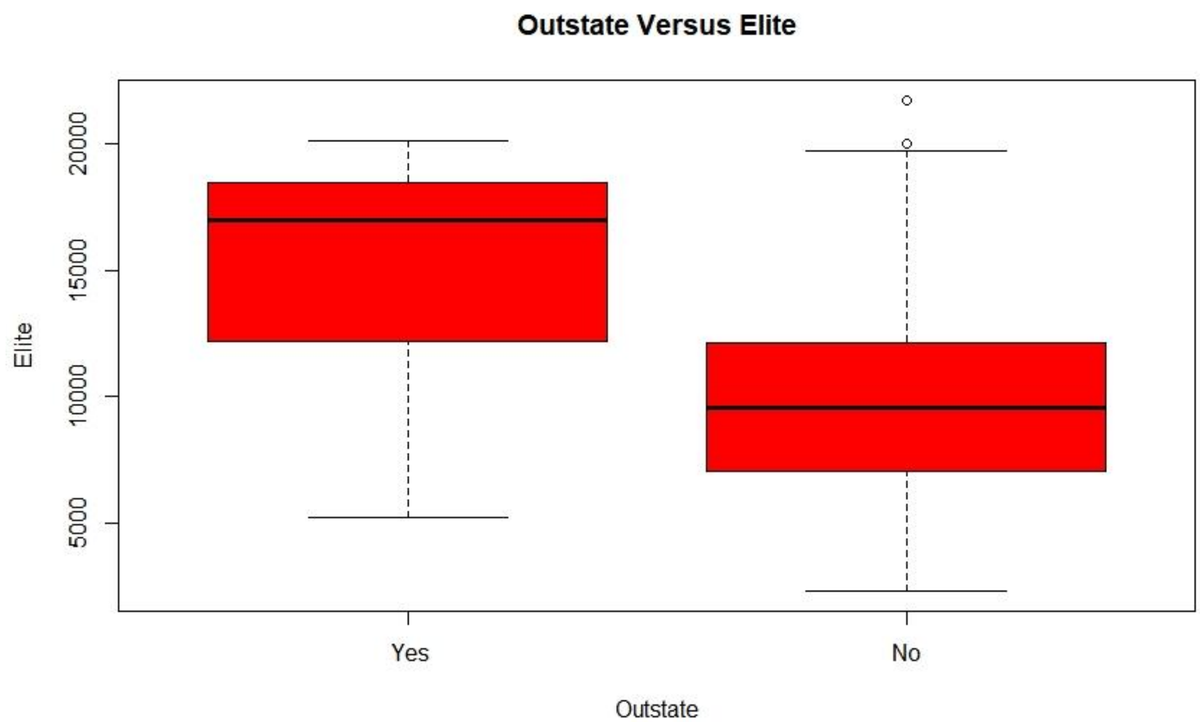
iv. Create a new qualitative variable, called Elite,by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.

Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.

Code

```
Elite=rep("No",nrow(College ))
Elite[College$Top10perc >50]=" Yes"
Elite=as.factor(Elite)
College=data.frame(College ,Elite)
summary(College)
boxplot(College$Outstate ~ College$Elite, col="red", main="Outstate Versus Elite",
      ylab="Elite", xlab="Outstate")
```

Output

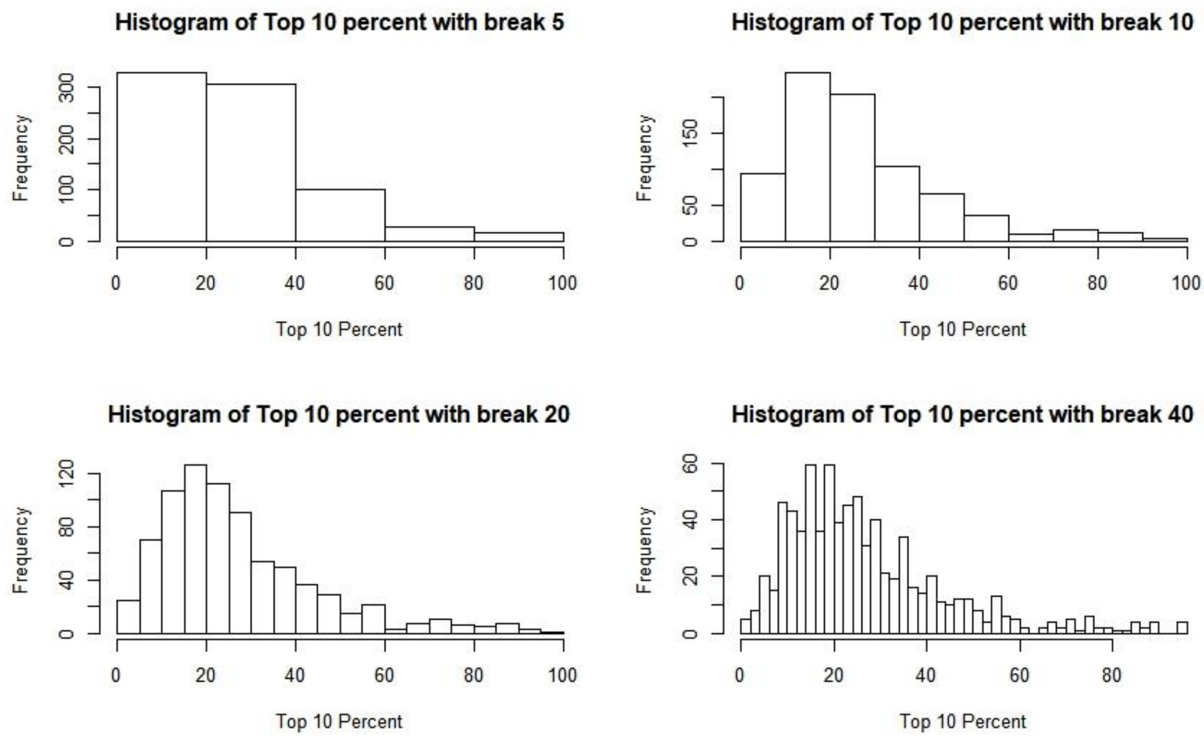


v. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

Code

```
hist(College$Outstate,main="Histogram of Outstate",
    xlab="Outstate",xlim=c(2500,21000),freq=FALSE)
par(mfrow=c(2,2))
hist(College$Top10perc,main="Histogram of Top 10 percent with break 5", xlab="Top 10 Percent", breaks=5)
hist(College$Top10perc,main="Histogram of Top 10 percent with break 10",xlab="Top 10 Percent", breaks=10)
hist(College$Top10perc,main="Histogram of Top 10 percent with break 20",xlab="Top 10 Percent", breaks=20)
hist(College$Top10perc,main="Histogram of Top 10 percent with break 40",xlab="Top 10 Percent", breaks=40)
```

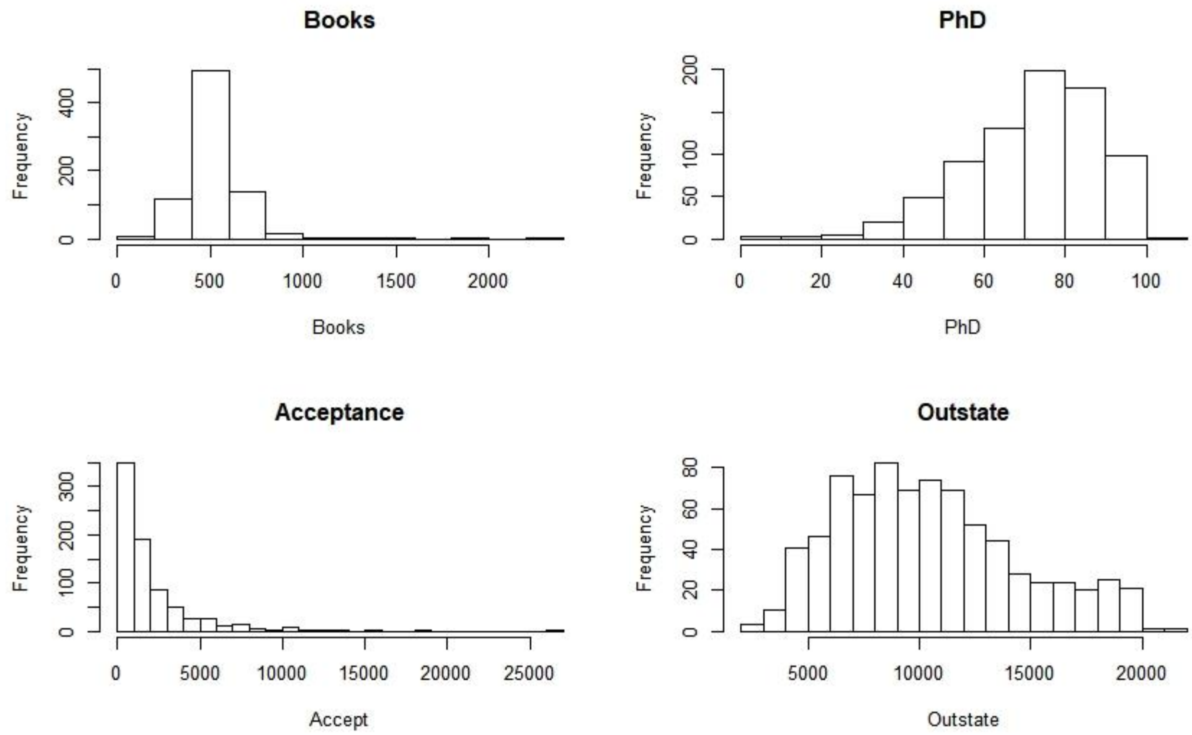
Output



Code

```
par(mfrow=c(2,2))
hist(College$Books,main = "Books",xlab = "Books", breaks=10)
hist(College$PhD,main="PhD",xlab = "PhD",breaks=10)
hist(College$Accept,main="Acceptance",xlab = "Accept",breaks=20)
hist(College$Outstate,main="Outstate", xlab="Outstate",breaks=20)
```

Output



vi. Continue exploring the data and provide a brief summary of what you discover.

Few observations from the data is:

- As the Outstate variable increases Grad.Rate increases
- College acceptance has very low effect on Grad.Rate
- College enrollment does not affect graduation rate
- Private universities get more applications than public

5. This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data.

(a) Which of the predictors are quantitative, and which are qualitative?

Code

```
library(ISLR)
data("Auto")
Auto$origin<-factor(Auto$origin)
Auto$year<-factor(Auto$year)
sapply(Auto,class)
```

Output

```
mpg cylinders displacement horsepower weight acceleration year origin  name
"numeric"    "numeric"    "numeric"    "numeric"    "numeric"    "numeric"
"factor"     "factor"     "factor"
```

Thus, predictors mpg, cylinders, displacement, horsepower, weight and acceleration are quantitative while year, origin, name are qualitative

(b) What is the range of each quantitative predictor? You can answer this using the range() function.

```
 range(Auto$mpg)
[1] 9.0 46.6
 range(Auto$cylinders)
[1] 3 8
 range(Auto$displacement)
[1] 68 455
 range(Auto$horsepower)
[1] 46 230
 range(Auto$weight)
[1] 1613 5140
 range(Auto$acceleration)
[1] 8.0 24.8
```

(c) What is the mean and standard deviation of each quantitative predictor?

Mean of the quantitative variables is as follows

```
sapply(Auto[, quant],mean)
 mpg    cylinders displacement   horsepower      weight     acceleration
 23.445918   5.471939   194.411990   104.469388 2977.584184    15.541327
```

Standard Deviation of the quantitative variables is as follows:

```
sapply(Auto[, quant],sd)
  mpg    cylinders displacement   horsepower      weight acceleration
 7.805007     1.705783   104.644004    38.491160   849.402560     2.758864
```

(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

```
AutoNew<-Auto[-c(10:85),]
sapply (AutoNew[, quant],range)
   mpg     cylinders displacement horsepower weight acceleration
[1,] 11.0         3           68         46   1649          8.5
[2,] 46.6         8          455        230   4997         24.8

sapply (AutoNew[, quant],mean)
   mpg    cylinders displacement   horsepower      weight acceleration
24.404430     5.373418   187.240506   100.721519 2935.971519    15.72689

sapply (AutoNew[, quant],sd)
   mpg    cylinders displacement   horsepower      weight acceleration
```

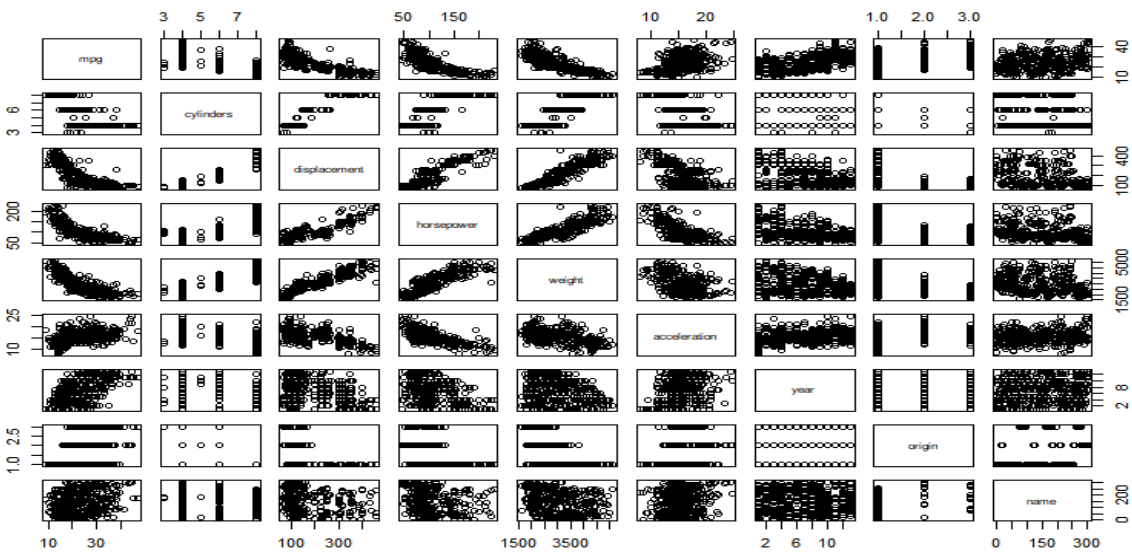
```
7.867283      1.654179     99.678367     35.708853    811.300208
2.693721
```

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

Code

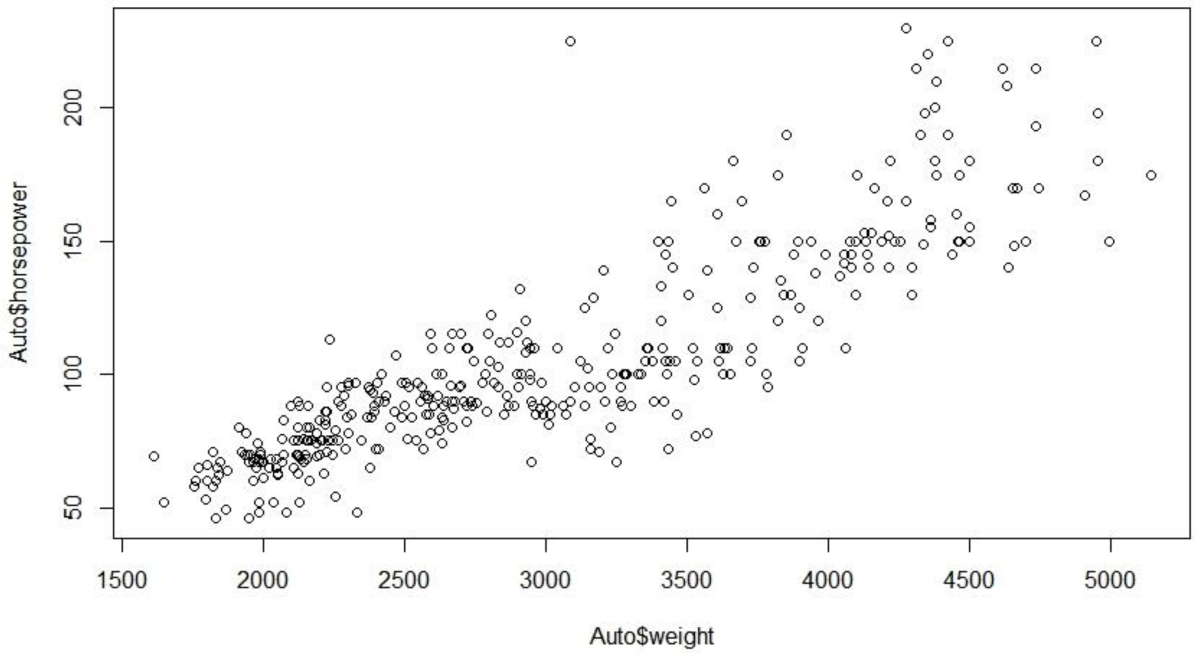
```
pairs(Auto)
```

Output



Code

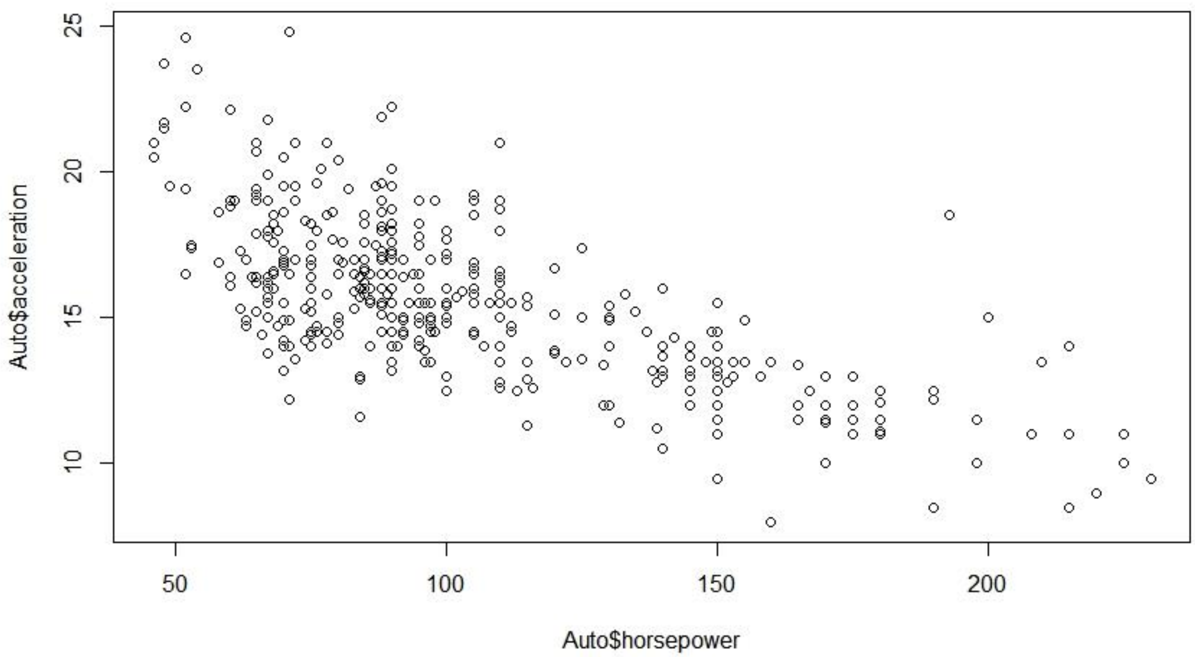
```
plot(Auto$weight,Auto$horsepower,type='p')
```



As per the plot, as the weight increases the horse power increases linearly

Code

```
plot(Auto$horsepower,Auto$acceleration,type='p')
```



As per the plot as the horsepower increases the acceleration decreases non-linearly

(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

According to sapply(auto) graphs we can observe that mpg decreases with increase in cylinder, displacement, horsepower and weight. The mpg also increases with increase in acceleration and year**.**

If we take p-values with each predictor we will find that origin, year and weight are have significant influence over it while horsepower, displacement and cylinders do not have that much of an influence.

6. Describe the null hypotheses to which the p-values given in Table 3.4 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.

According to table 3.4 it can be concluded that there is very less effect of newspaper advertising on sales. The p-value for newspaper is 0.86 which is very high thus even if we spend money on newspaper advertising the sales won’t increase much.

On the contrary the p-value for both TV and radio in less than 0.0001, thus spending more advertising money on them will result in increase in sales. Especially according the table spending 1000$ on radio advertising will end up increasing sales by 189$ while it would be 46$ for the same spending on TV advertising.

7. Consider the fitted values that result from performing linear regression without an intercept. In this setting, the ith fitted value takes the form

$\hat{y}_i = x_i\beta$, where $\hat{\beta} = \hat{\beta} = (\Sigma x_{i'} y_{i'}) / (\Sigma x_{i'}^2)$.

Show that we can write $\hat{y}_i = \Sigma_i a_{i'} y_{i'}$ .What is a_i' ?

Note: We interpret this result by saying that the fitted values from linear regression are linear combinations of the response values.

$\hat{y}_i = x_i\hat{\beta}$ $\qquad\qquad\qquad$ $\hat{\beta} = (\Sigma x_{i'} y_{i'}) / (\Sigma x_{i'}^2)$

$\hat{y}_i = x_i * (\Sigma_i x_{i'} y_{i'}) / (\Sigma_k x_k^2)$

$\hat{y}_i = \Sigma_i((x_{i'} y_{i'}) / (\Sigma_k x_k^2)) * y_{i'}$

$a_{i'} = (x_{i'} x_i) / (\Sigma_k x_k^2)$

Variable name doesn't matter inside of Σ

8. Using (3.4), argue that in the case of simple linear regression, the least squares line always passes through the point (¯x, ¯y).

A point is on the line if substituting it in the line equation yields zero. Considering 2-dimensional setting let the line equation be:
y = B0 + B1*x ----- eq. (1)

From 3.4 we know that B0 = ¯y – B1*¯x ------ eq. (2), where ¯y and ¯x are average of y and x respectively.

Thus, putting point (¯x, ¯y) in eq. (1) and putting value of B0 from eq. (2) into eq. (1) we get

¯y = ¯y – B1*¯x + B1*¯x. Thus LHS = RHS and we can conclude that the average point lies on the line.

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

a. **Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summary() function to print the results. Comment on the output.**

Code:

```
library(ISLR)
```

```
data(Auto)
fit <- lm(mpg ~ horsepower, Auto)
head(fit)
summary(fit)
```

OUTPUT:

```
> summary(fit)

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
```

I. Is there a relationship between the predictor and the response?

Yes, the predictor p has very low value which is <2e-16, whereas F static is larger than 1. This concludes the null hypothesis as well as suggest a significant relationship between the predictor and the response.

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

The R^2 is almost 60.59 percent of the variance in mpg.

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

Negative. The coefficient has negative value.

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

Code:

```
predict(fit, data.frame(horsepower = 98), interval = "confidence")
predict(fit, data.frame(horsepower = 98), interval = "prediction")
```

Output:

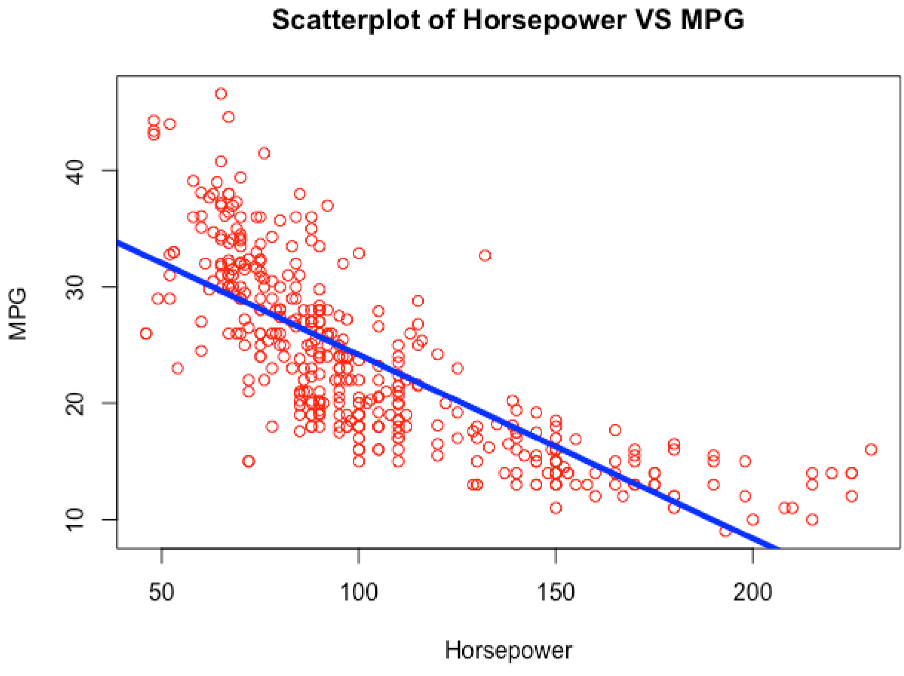
```
> predict(fit, data.frame(horsepower = 98), interval = "confidence")
       fit      lwr      upr
1 24.46708 23.97308 24.96108
> predict(fit, data.frame(horsepower = 98), interval = "prediction")
       fit     lwr      upr
1 24.46708 14.8094 34.12476
```

b. Plot the response and the predictor. Use the abline() function to display the least squares regression line.

Code:

```
plot(Auto$horsepower,Auto$mpg, xlab = 'Horsepower', ylab = 'MPG', col = 'red', main = 'Scatterplot of Horsepower VS MPG')
abline(fit, col='blue', lwd = '4')
```

Output:

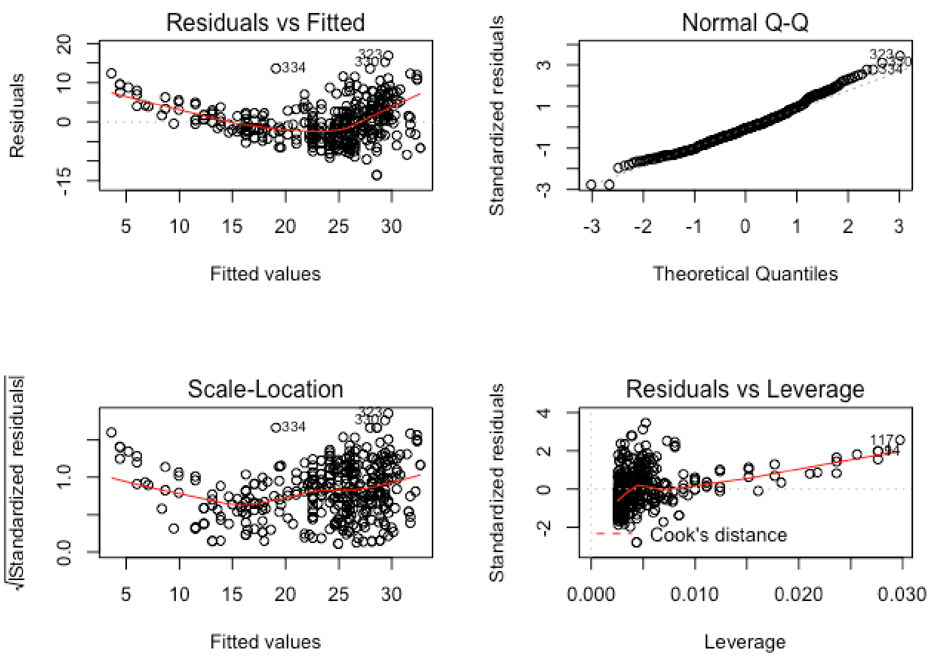


c. **Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.**

Code:

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

Output:



- The plot of residual versus fitted values depicts a soft U-shaped tendency, this infers presence of non-linearity in data.
- The plot of standardized residuals versus leverage indicates the presence of a few outliers (higher than 2 or lower than -2) and a few high leverage points.
- The residuals suggest a non-linear relationship as they don't seem to be independent of outcome.

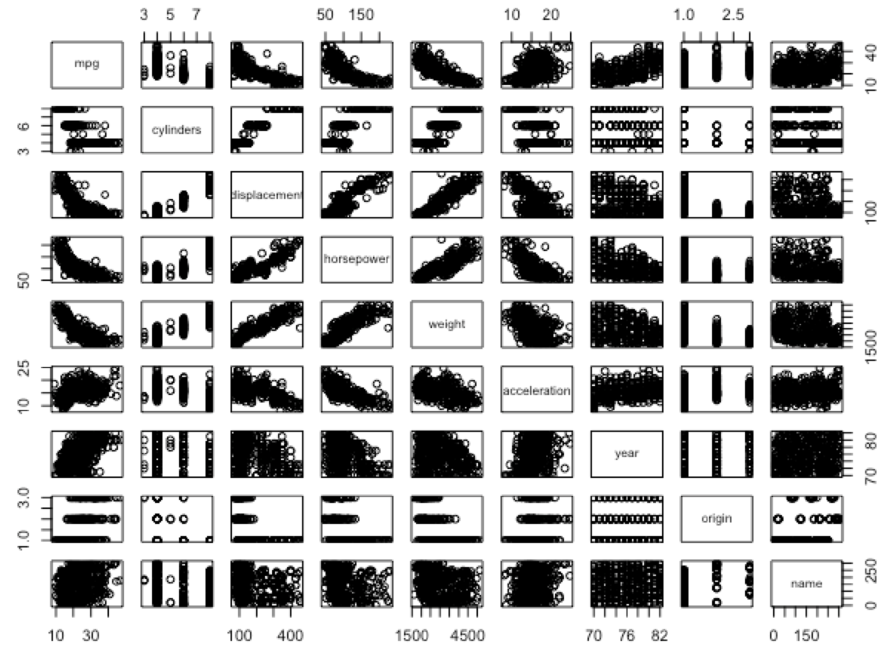
10. This question involves the use of multiple linear regression on the Auto data set.

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

Code:

```
pairs(Auto)
```

Output:



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.

Code:

```
cor(Auto[, names(Auto) !="name"])
```

Output:

```
> cor(Auto[, names(Auto) !="name"])
                    mpg  cylinders displacement horsepower     weight acceleration       year     origin
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442    0.4233285  0.5805410  0.5652088
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273   -0.5046834 -0.3456474 -0.5689316
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944   -0.5438005 -0.3698552 -0.6145351
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377   -0.6891955 -0.4163615 -0.4551715
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000   -0.4168392 -0.3091199 -0.5850054
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392    1.0000000  0.2903161  0.2127458
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199    0.2903161  1.0000000  0.1815277
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054    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 thesummary() function to print the results. Comment on the output. For instance:

Code:

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

Output:

```
> fit2 <- lm(mpg ~ .-name, data=Auto)
> summary(fit2)

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?

   Yes, there exists a relationship as the p value is very small which is <2.2e-16.

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

   Year, weight, origin, displacement has a significant relationship with the response.

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

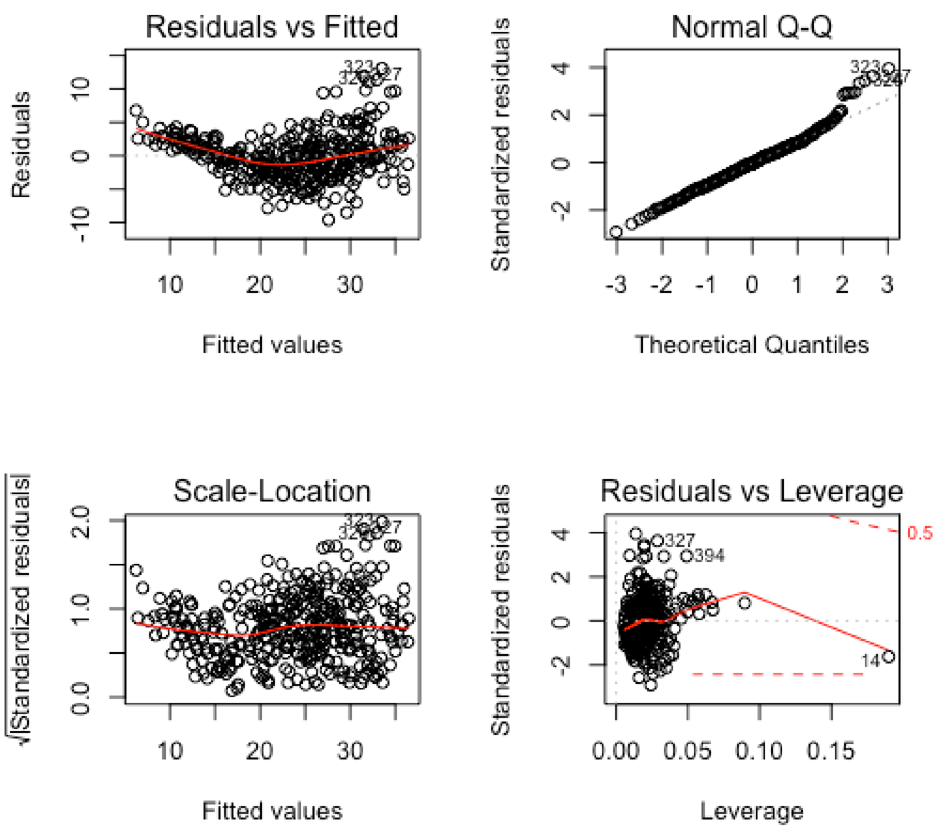
   Coefficient for the year variable suggest that the average effect of an increase in 1 year leads to an increase of 0.750773 in mpg.

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?

Code:

```
par(mfrow = c(2, 2))
plot(fit2)
```

Output:



- The plot of standardized residuals versus leverages indicate the presence of few outlier points in the data greater than 3 and some less than -2 and one high leverage point 14.
- At the Scale-Location graph, some observations are potential outliers, mainly the observation 323.

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

Code:

```
fit3= lm(mpg ~.-
name+cylinders:acceleration+year:origin+displacement:weight+displacement:weight+acceleration:
horsepower+acceleration:weight, data=Auto)
summary(fit3)
```

Output:

```
Call:
lm(formula = mpg ~ . - name + cylinders:acceleration + year:origin +
    displacement:weight + displacement:weight + acceleration:horsepower +
    acceleration:weight, data = Auto)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1712 -1.6213  0.0804  1.4452 13.0961

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.525e+01  9.922e+00   1.537 0.125126
cylinders                9.096e-01  1.126e+00   0.808 0.419558
displacement            -8.753e-02  1.259e-02  -6.953 1.58e-11 ***
horsepower               9.164e-02  3.803e-02   2.409 0.016452 *
weight                  -1.395e-02  3.001e-03  -4.650 4.59e-06 ***
acceleration             1.179e-01  2.830e-01   0.417 0.677174
year                     5.040e-01  9.906e-02   5.088 5.72e-07 ***
origin                  -1.223e+01  4.131e+00  -2.961 0.003260 **
cylinders:acceleration  -3.037e-02  7.190e-02  -0.422 0.672965
year:origin              1.627e-01  5.305e-02   3.067 0.002319 **
displacement:weight      2.304e-05  2.983e-06   7.722 1.03e-13 ***
horsepower:acceleration -9.474e-03  2.616e-03  -3.621 0.000333 ***
weight:acceleration      3.036e-04  1.622e-04   1.872 0.061983 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.862 on 379 degrees of freedom
Multiple R-squared:  0.8696,	Adjusted R-squared:  0.8655
F-statistic: 210.7 on 12 and 379 DF,  p-value: < 2.2e-16
```

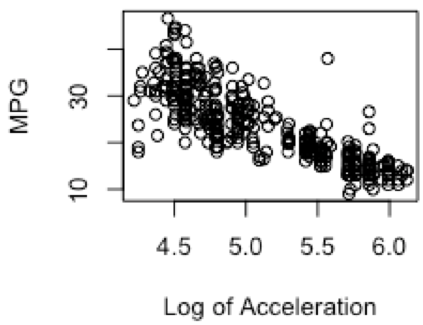
Here the R-squared statistics estimates that 87% of the changes in the response can be explained by this particular set of predictors.

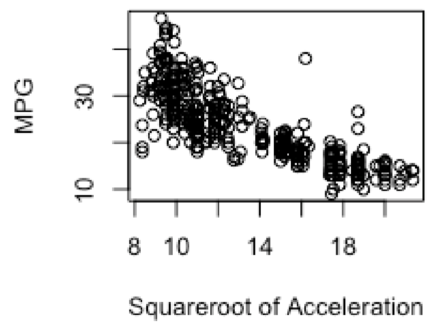
f. Try a few different transformations of the variables, such as log(X), √X, X2. Comment on your findings

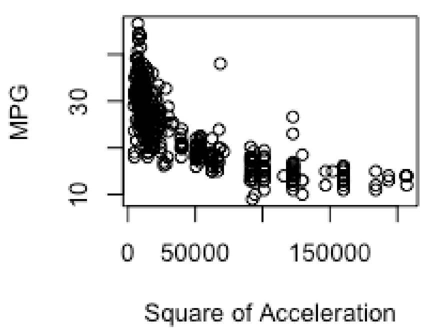
Code:

```
par(mfrow = c(2, 2))
plot(log(Auto$displacement), Auto$mpg, xlab = 'Log of Displacement', ylab = 'MPG')
plot(sqrt(Auto$displacement), Auto$mpg, xlab = 'Squareroot of Displacement', ylab = 'MPG')
plot((Auto$displacement)^2, Auto$mpg, xlab = 'Square of Displacement', ylab = 'MPG')
```

Output:







Plot of squareroot of Acceleration VS MPG is most linear among the other 2.