Performance analysis of Students

Team-8

15/10/2021

Team Members
1. Ketan Bassi (30146366)
2. Nanditha Sriram(30149248)
3. Shubham Bansal(30153948)
4. Oham Ugochukwu(30160663)

Motivation:
Earning a good grade is not only a measure of subject matter knowledge or intelligence. Instead, it's a composite of knowledge, skills, and personality traits. For example, a student with a good work ethic and discipline could help their grades because they turn in homework assignments on time and have good class attendance. Similarly, a student who is driven would be willing to do additional research for assignments or to seek out learning resources if they were struggling. Because grades are a composite measurement of student performance, they can be a better predictor of success than other narrow measures.

The aim of the project is to analyze the performance of student throughout their academic life and gauge if and how past performance affects future performance.

Data Collection and wrangling
The data set we have is in CSV format. It has 15 columns and 215 rows. The data set has columns like Gender, it contains scores from Secondary, higher secondary, degree and MBA along with the specialization of higher secondary, degree and MBA. The data also consists of employment test percentage as well salary of the placed students.
The data is pretty clean so we do not need to do any data wrangling on the

data set.

```
data =  read.csv("project.csv")
head(data)
```

```
##   sl_no gender ssc_p   ssc_b hsc_p   hsc_b    hsc_s degree_p  degree_t we
## 1     1      M 67.00  Others 91.00  Others Commerce    58.00  Sci&Tech
## 2     2      M 79.33 Central 78.33  Others  Science    77.48  Sci&Tech
## 3     3      M 65.00 Central 68.00 Central     Arts    64.00 Comm&Mgmt
## 4     4      M 56.00 Central 52.00 Central  Science    52.00  Sci&Tech
## 5     5      M 85.80 Central 73.60 Central Commerce    73.30 Comm&Mgmt
## 6     6      M 55.00  Others 49.80  Others  Science    67.25  Sci&Tech
##   etest_p specialisation mba_p     status salary
## 1    55.0         Mkt&HR 58.80     Placed 270000
## 2    86.5        Mkt&Fin 66.28     Placed 200000
## 3    75.0        Mkt&Fin 57.80     Placed 250000
## 4    66.0         Mkt&HR 59.43 Not Placed     NA
## 5    96.8        Mkt&Fin 55.50     Placed 425000
## 6    55.0        Mkt&Fin 51.58 Not Placed     NA
```

```
str(data)
```

```
## 'data.frame':    215 obs. of  15 variables:
##  $ sl_no         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ gender        : chr  "M" "M" "M" "M" ...
##  $ ssc_p         : num  67 79.3 65 56 85.8 ...
##  $ ssc_b         : chr  "Others" "Central" "Central" "Central" ...
##  $ hsc_p         : num  91 78.3 68 52 73.6 ...
##  $ hsc_b         : chr  "Others" "Others" "Central" "Central" ...
##  $ hsc_s         : chr  "Commerce" "Science" "Arts" "Science" ...
##  $ degree_p      : num  58 77.5 64 52 73.3 ...
##  $ degree_t      : chr  "Sci&Tech" "Sci&Tech" "Comm&Mgmt" "Sci&Tech" ...
##  $ workex        : chr  "No" "Yes" "No" "No" ...
##  $ etest_p       : num  55 86.5 75 66 96.8 ...
##  $ specialisation: chr  "Mkt&HR" "Mkt&Fin" "Mkt&Fin" "Mkt&HR" ...
##  $ mba_p         : num  58.8 66.3 57.8 59.4 55.5 ...
##  $ status        : chr  "Placed" "Placed" "Placed" "Not Placed" ...
##  $ salary        : int  270000 200000 250000 NA 425000 NA NA 252000 2310
```

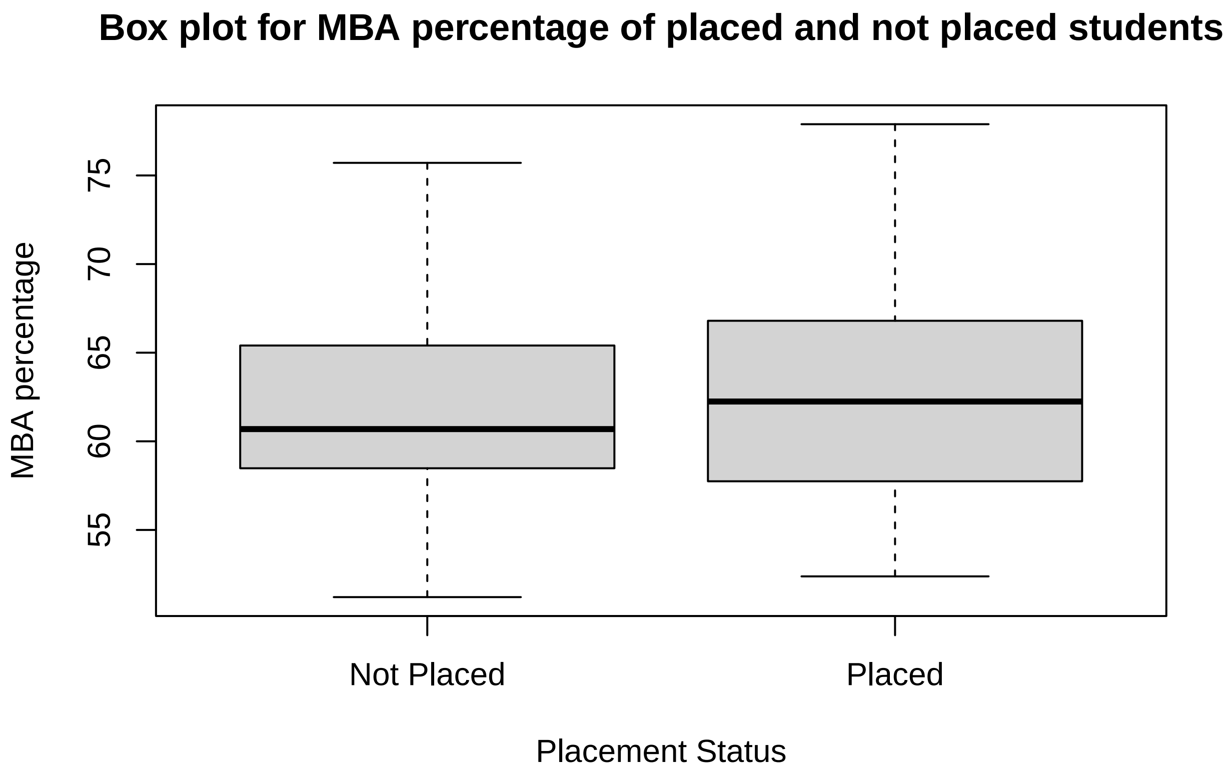
```
data1 = data.frame(a=rep(1,215),data)
```

Question 1 (Hypothesis Testing)

(a) H_0: MBA percentage does not affect their chance of placement
H_a: MBA percentage affects their chance of placement

Solution First we create a Box plot for placed and not placed students along with their mba percentage.

```
boxplot(data$mba_p~data$status,data = data,main= "Box plot for
 ↪  MBA percentage of placed and not placed
 ↪  students",xlab="Placement Status",ylab="MBA percentage")
```



Analysis The average MBA percentage for the placed students is higher than the students that are not placed.

```
t.test(data$mba_p~data$status,data = data,alternative=
 ↪  "two.sided")
```

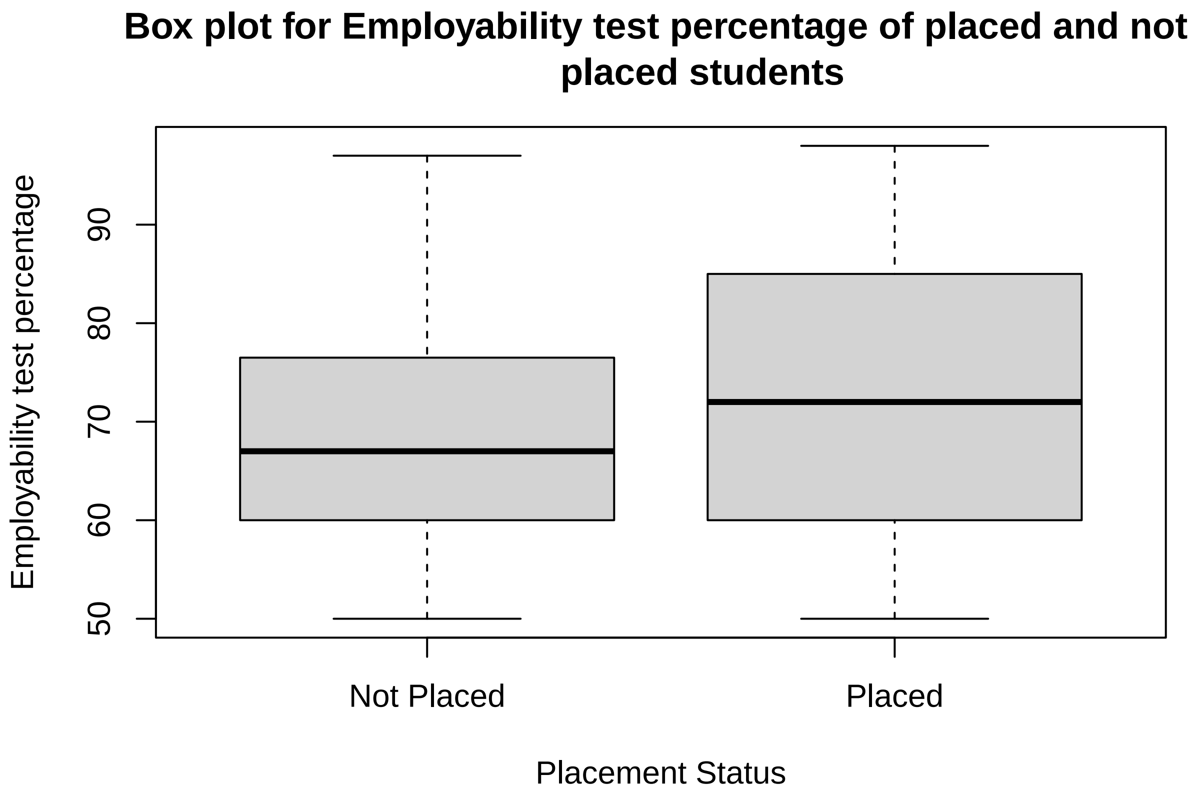
```
##
##  Welch Two Sample t-test
##
## data:  data$mba_p by data$status
## t = -1.1392, df = 131.21, p-value = 0.2567
## alternative hypothesis: true difference in means between group Not Place
## 95 percent confidence interval:
##  -2.6449696  0.7118575
## sample estimates:
## mean in group Not Placed     mean in group Placed
##                 61.61284                 62.57939
```

Since the p-value is greater than alpha we fail to reject the null hypothesis. From this we can conclude that the percentage of the students during their MBA does not affect their placements.

(b)

H_0:Employability test scores does not affect their chance of placement
H_a:Employability test scores affects their chance of placement

```
boxplot(data$etest_p~data$status,data = data, main="Box plot for
 ↪  Employability test percentage of placed and not
        placed students",xlab="Placement
 ↪  Status",ylab="Employability test percentage")
```



Analysis The average Employability test percentage for the placed students is higher than the students that are not placed.

```
t.test(data$etest_p~data$status,data = data,alternative=
 ↪  "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  data$etest_p by data$status
## t = -1.9801, df = 145.39, p-value = 0.04958
## alternative hypothesis: true difference in means between group Not Place
## 95 percent confidence interval:
##  -7.293445062 -0.006815124
## sample estimates:
## mean in group Not Placed     mean in group Placed
##                 69.58791                 73.23804
```

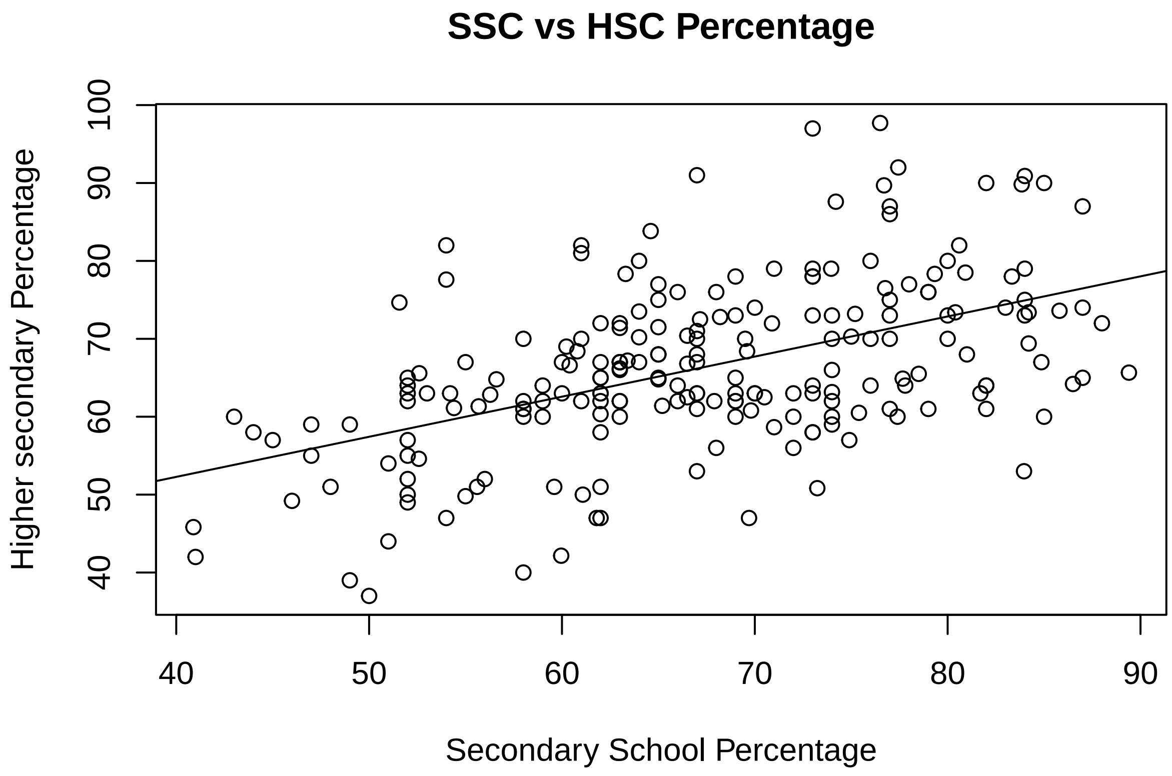
Since the P-value is very close to the alpha(0.05), so we do not have enough evidence to reject or accept the null hypothesis.

Question 2

(a)

Performing simple linear regression to estimate higher secondary score based on secondary score

```
plot(data$hsc_p~data$ssc_p,xlab="Secondary School Percentage",ylab="Higher secondary Percentage",main= "SSC vs HSC Percentage")
reg = lm(data$hsc_p~data$ssc_p,data = data)
abline(reg)
```

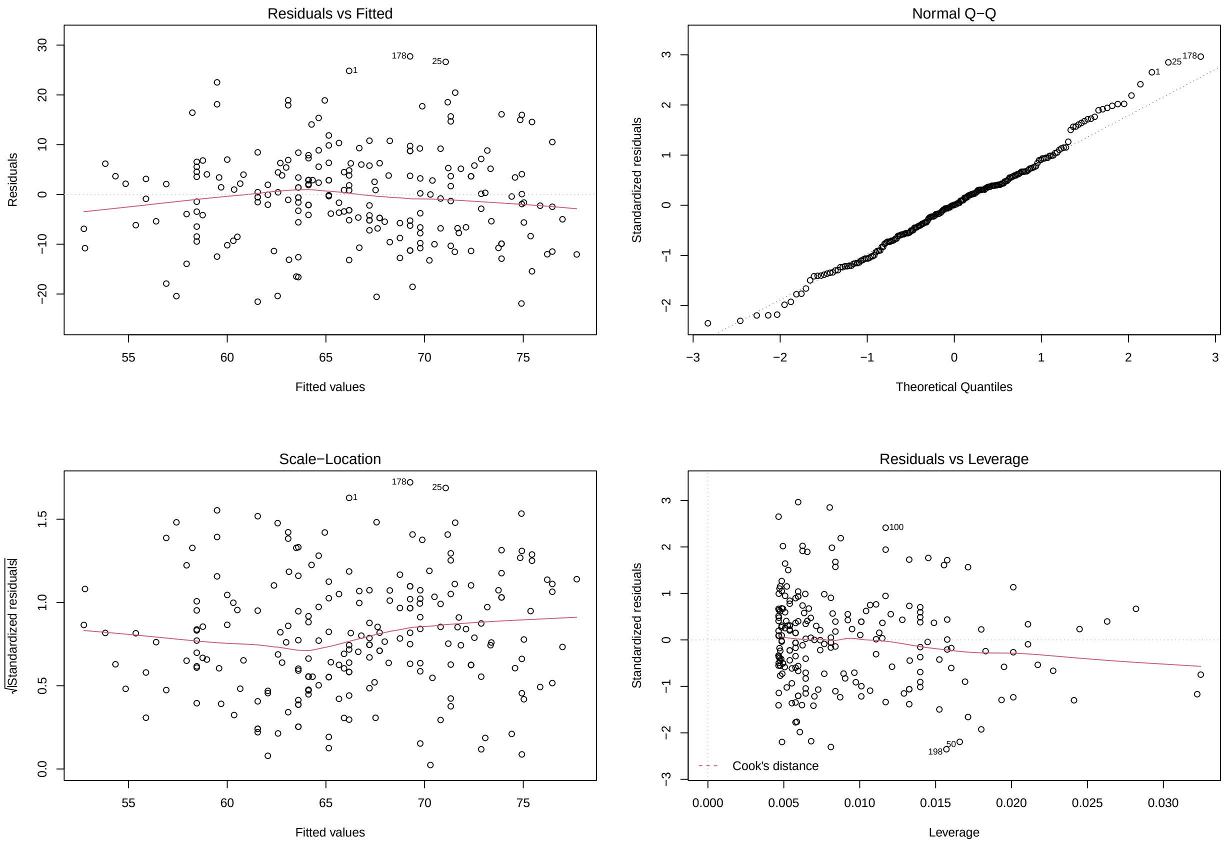


We can see a positive linearity between Secondary School Percentage and Higher Secondary school percentage.

```
summary(reg)
```

```
##
## Call:
## lm(formula = data$hsc_p ~ data$ssc_p, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.9079  -6.2160   0.0715   5.3567  27.7343
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31.68583    4.03948   7.844 2.08e-13 ***
## data$ssc_p   0.51479    0.05926   8.687 9.90e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.386 on 213 degrees of freedom
## Multiple R-squared:  0.2616, Adjusted R-squared:  0.2581
## F-statistic: 75.46 on 1 and 213 DF,  p-value: 9.902e-16
```

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



The plot of the residuals vs fitted, shows that the relationship between secondary school percentage and Higher sceondary percentage is linear since the residuals "bounce randomly" around the 0 line. Secondly, although the red line in the scale-location plot is not perfectly horizontal, the points are randomly spread out. This shows heteroscedasticity, which is a good sign for the model. In addition, the straight line in the normal Q-Q plot shows that the distribution is approximately normal, although with a few departures at the upper and lower tail. Furthermore, since no point falls outside of the Cook's distance in the residuals vs leverage plot, we can assume that there are no influential points in our model.

The overall F-test(75.46 on 1 and 213 DF, p-value: 9.902e-16) and the individual t-test(P value for slope is 2.08e-13) all suggest that the model is highly statistically significant. Also looking at the model it can be observed that the model can be used to predict the higher secondary percentage using the secondary percentage.

Regression Equation:

$hsc_p = 31.68583 + 0.51479 * ssc_p$

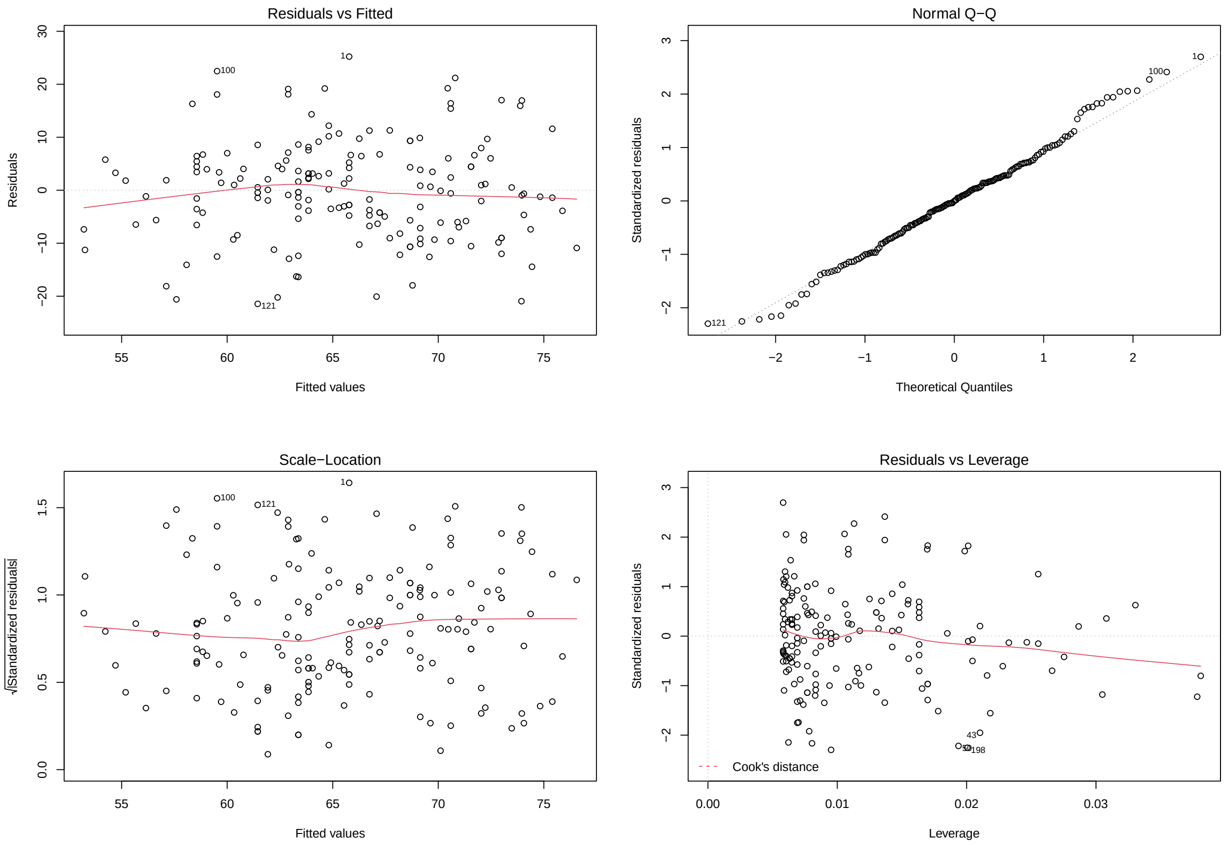
After doing that we train and test our model

```
set.seed(2021)
n = nrow(data)
index =  sample(1:n,round(0.8*n),replace = FALSE)

train80 = data[index, ] # 80% of the data to build a model
test20 = data[-index, ] #20% of the data to test the model

train80_fit = lm(hsc_p~ssc_p,data=train80)

summary(train80_fit)
```

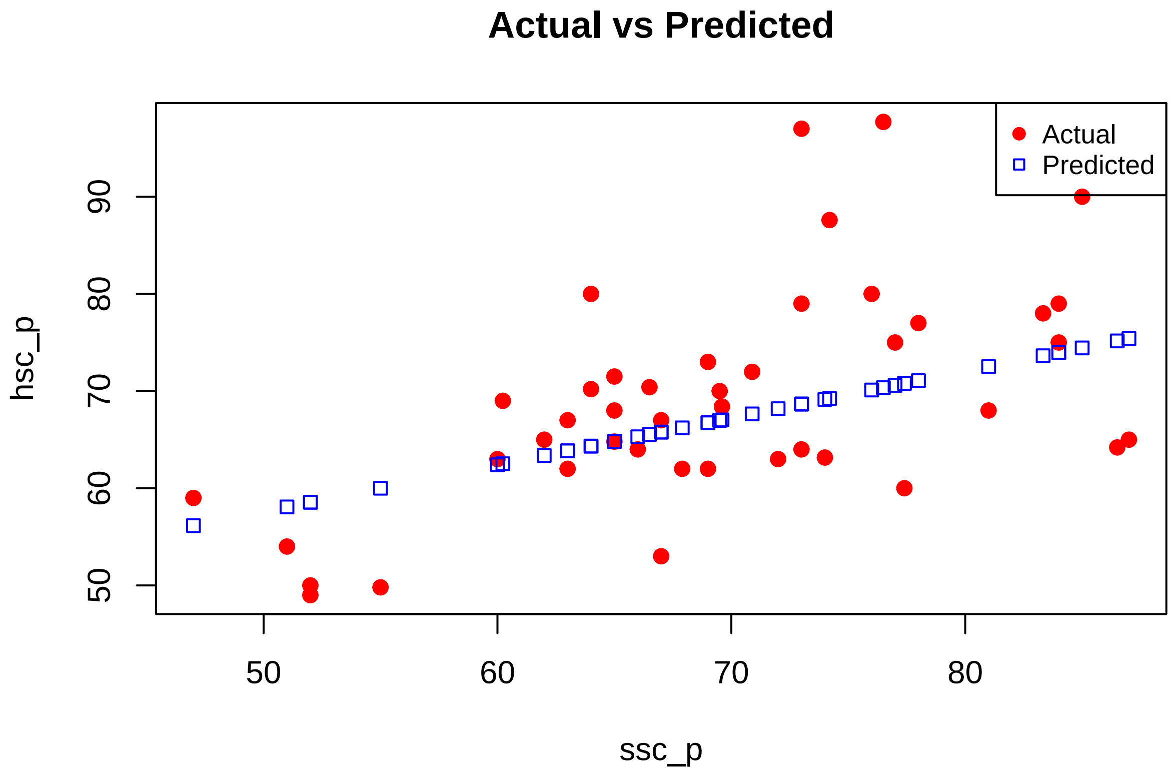
```
##
## Call:
## lm(formula = hsc_p ~ ssc_p, data = train80)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.4462  -6.1637  -0.0186   5.6479  25.2220
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.53042    4.40688   7.609 1.81e-12 ***
## ssc_p        0.48131    0.06512   7.391 6.32e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.381 on 170 degrees of freedom
## Multiple R-squared:  0.2432, Adjusted R-squared:  0.2387
## F-statistic: 54.62 on 1 and 170 DF,  p-value: 6.317e-12
```

Model diagnostic plots for train80_fit:

```
par(mfrow = c(2, 2))
plot(train80_fit)
```



```
pred20 = predict.lm(train80_fit,newdata = test20)
plot(hsc_p~ssc_p,data = test20,col='red',pch=19, main= "Actual vs
↪  Predicted")
points(test20$ssc_p,pred20,col='blue',pch=22)
legend("topright",legend=c("Actual","Predicted"),col=c("red","blue"),
       pch=c(19,22),cex = 0.8)
```



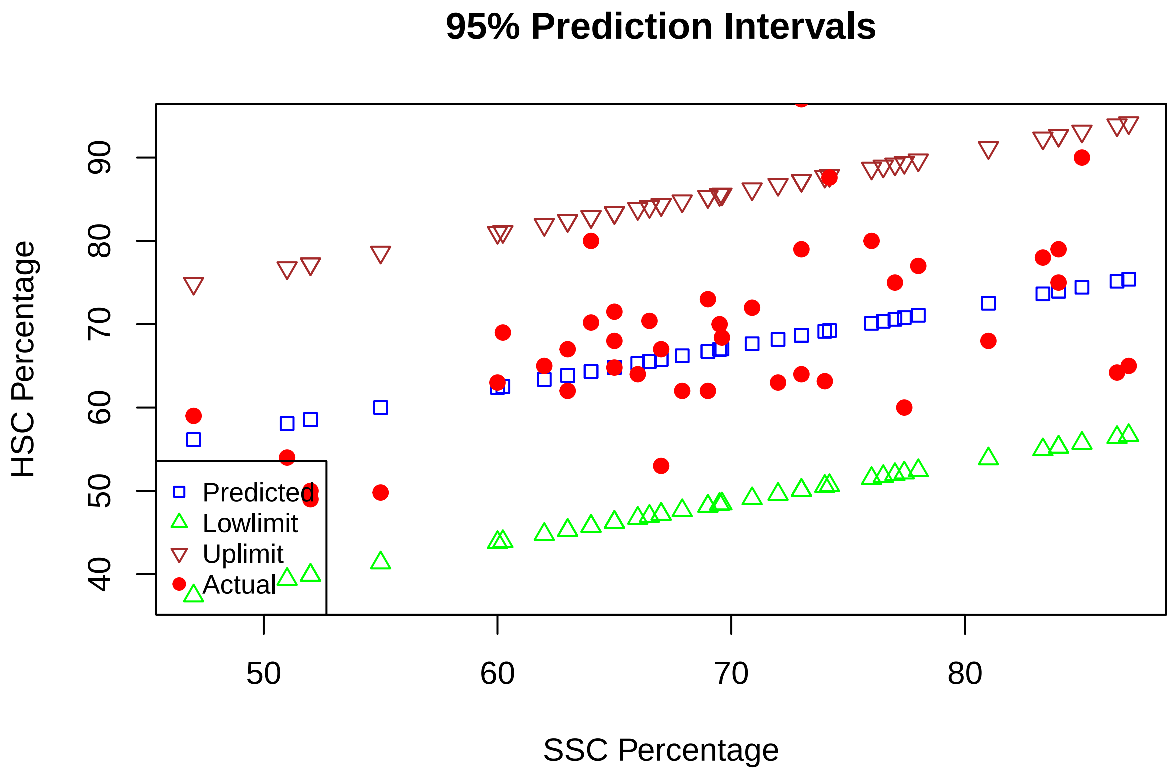
Now we compute MAE,MAPE,MSE and 95% confidence interval.

```
mae = mean(abs(test20$hsc_p - pred20))
mape = mean(abs((test20$hsc_p-pred20)/test20$hsc_p))
mse =mean((test20$hsc_p- pred20)^2)

cat("MAE=",mae,",","MAPE=",mape,",","MSE=",mse)
```

```
## MAE= 7.152492 , MAPE= 0.1006365 , MSE= 89.99581
```

```
pred20pred =predict.lm(train80_fit,newdata =
 ↪  test20,level=0.95,interval = "prediction")

matplot(test20$ssc_p,pred20pred,pch=c(22,2,6),col =
 ↪  c("blue","green","brown"), main="95% Prediction
 ↪  Intervals",xlab = "SSC Percentage",ylab = "HSC Percentage")
points(test20$ssc_p,test20$hsc_p,col="red",pch=19)
```

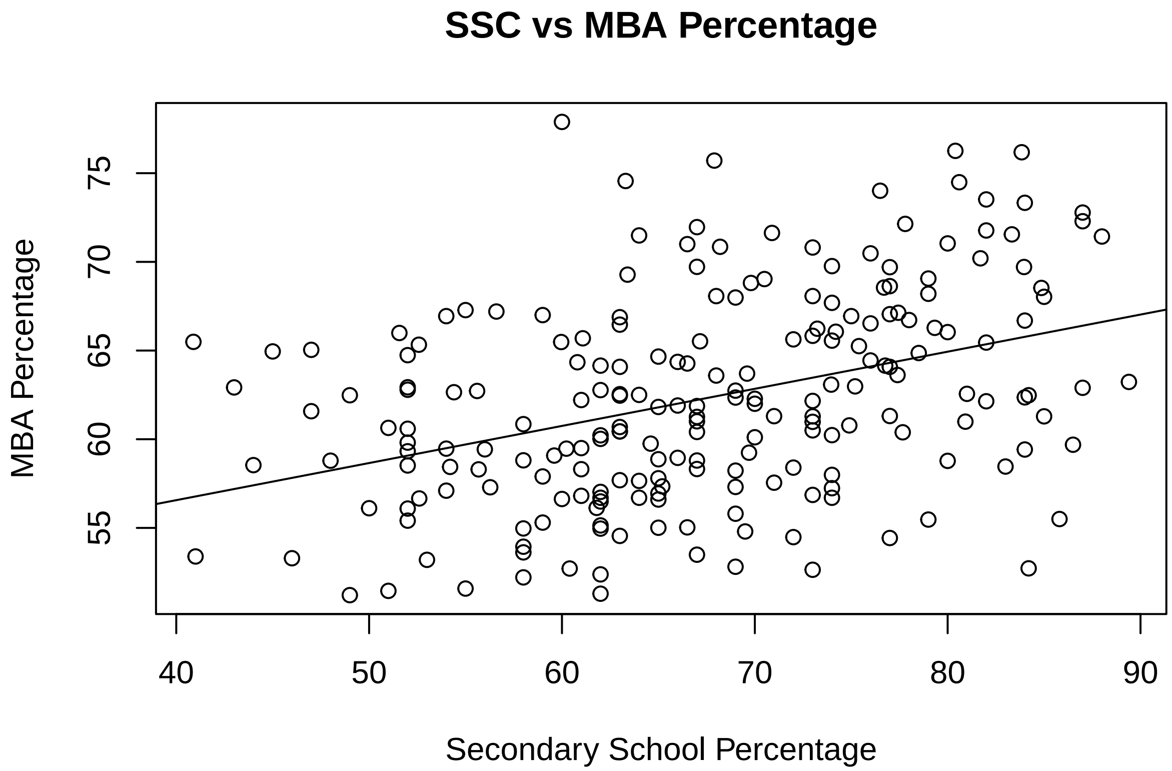
```
legend("bottomleft",legend=c("Predicted","Lowlimit","Uplimit","Actual"),
       ↪  col=c("blue","green","brown","red"),pch=c(22,2,6,19),cex=0.8)
```



All the actual values lie between the 95% prediction intervals.

(b) Finding out how each percentage, contributes to the MBA percentage.
(i) Secondary Percentage vs MBA Percentage

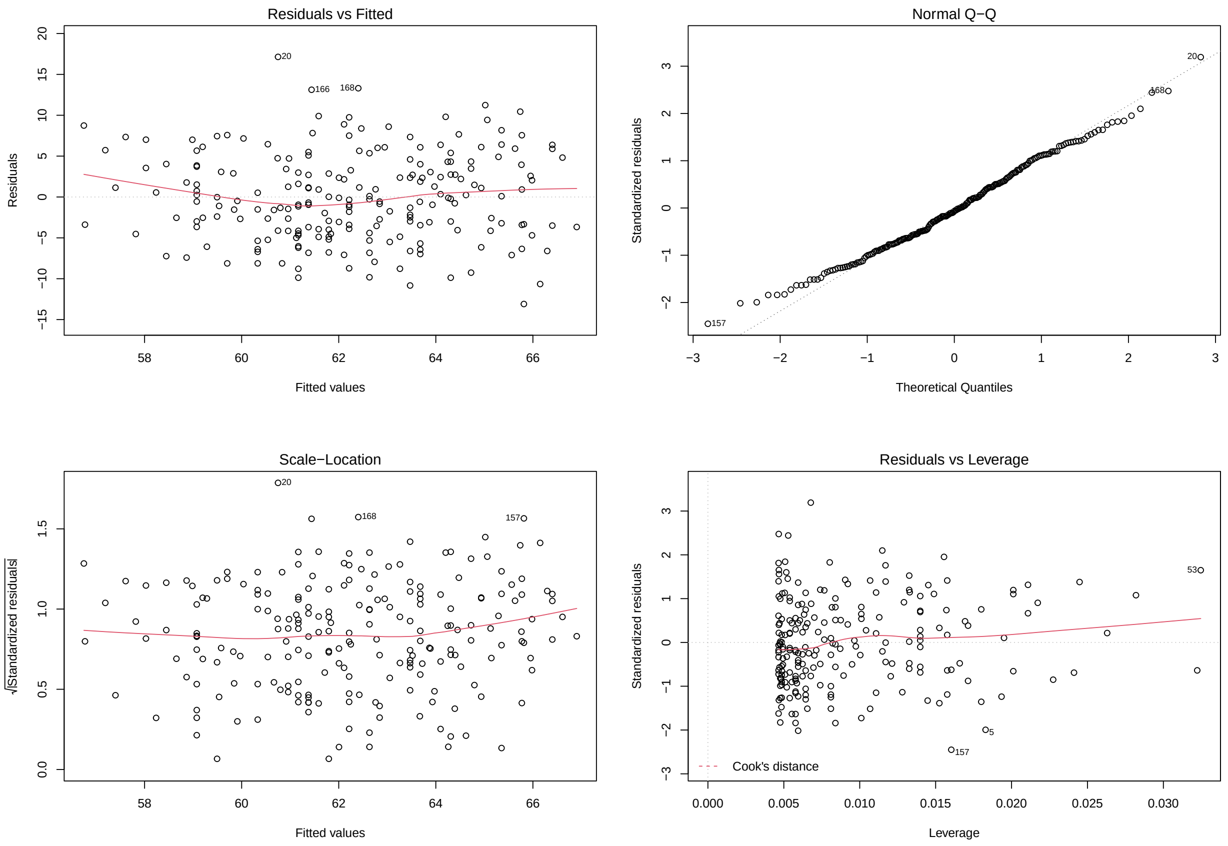
```
plot(data$mba_p~data$ssc_p,xlab="Secondary School
 ↪  Percentage",ylab="MBA Percentage",main="SSC vs MBA
 ↪  Percentage")
reg = lm(data$mba_p~data$ssc_p,data = data)
abline(reg)
```



```
summary(reg)
```

```
##
## Call:
## lm(formula = data$mba_p ~ data$ssc_p, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.0947  -3.9664  -0.3447   3.9002  17.1404
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 48.19156    2.31873  20.784  < 2e-16 ***
## data$ssc_p   0.20930    0.03402   6.153 3.72e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.388 on 213 degrees of freedom
## Multiple R-squared:  0.1509, Adjusted R-squared:  0.1469
## F-statistic: 37.86 on 1 and 213 DF,  p-value: 3.719e-09
```

```r
par(mfrow = c(2, 2))
plot(reg)
```

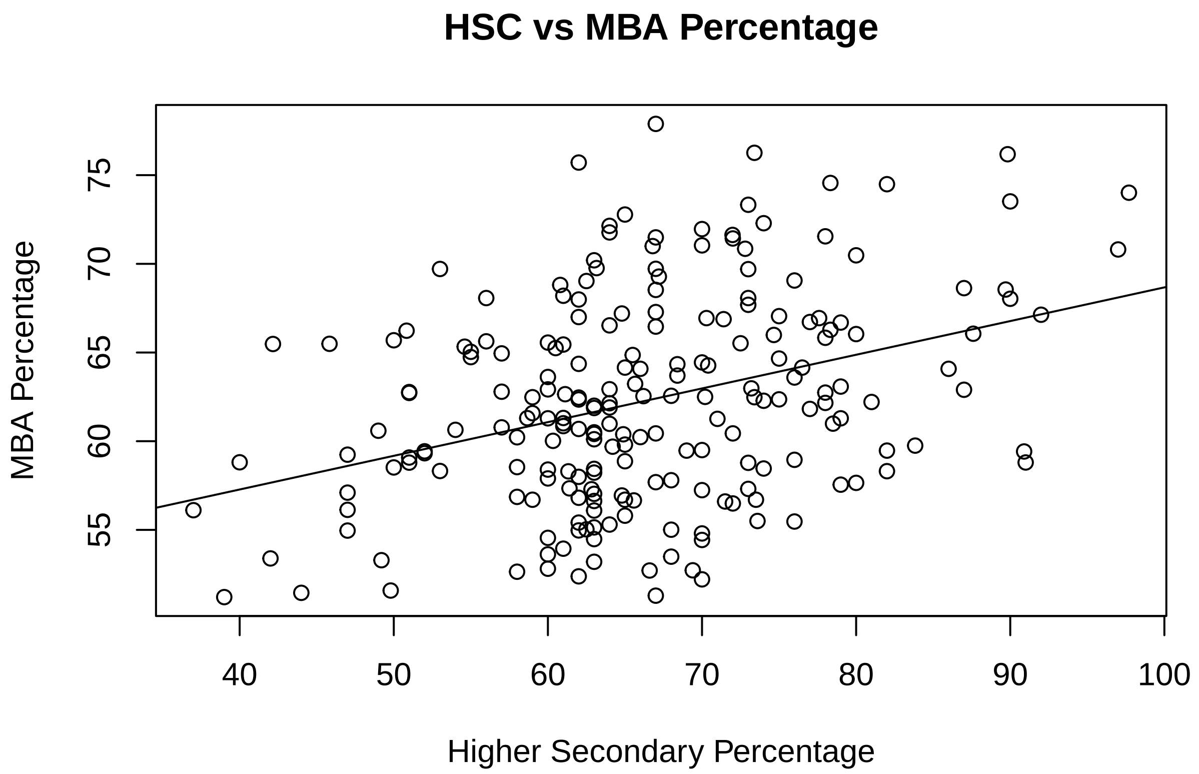


The plot of the residuals vs fitted, shows that the relationship between secondary school percentage and MBA percentage is linear since the residuals “bounce randomly” around the 0 line. Secondly, although the red line in the scale-location plot is not perfectly horizontal, the points are randomly spread out. This shows heteroscedasticity, which is a good sign for the model. In addition, the straight line in the normal Q-Q plot shows that the distribution is approximately normal, although with a few departures at the upper and lower tail. Furthermore, since no point falls outside of the Cook’s distance in the residuals vs leverage plot, we can assume that there are no influential points in our model.

Finally, based on these analysis, the fitted model is statistically significant and diagnostic plots reveal no major concerns. Hence we recommend using the model for predicting students MBA performance, based on their secondary school performance amount removed.

(ii) Higher Secondary Percentage vs MBA Percentage

```
plot(data$mba_p~data$hsc_p, xlab="Higher Secondary
 ↪  Percentage",ylab="MBA Percentage",main="HSC vs MBA
 ↪  Percentage")
reg = lm(data$mba_p~data$hsc_p,data = data)
abline(reg)
```

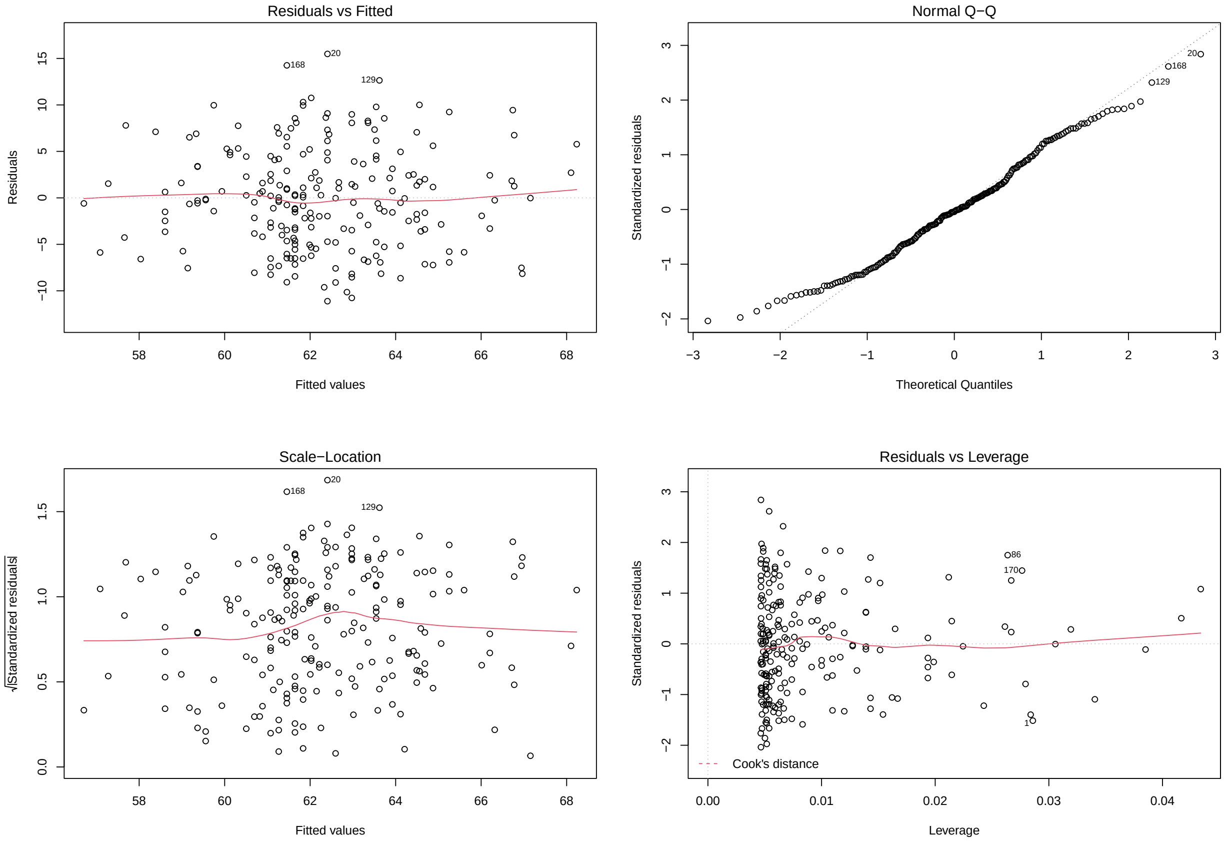


```
summary(reg)
```

```
##
## Call:
```

```
## lm(formula = data$mba_p ~ data$hsc_p, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1148  -4.2259  -0.2358   3.9827  15.4852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 49.67921    2.30500  21.553  < 2e-16 ***
## data$hsc_p   0.18993    0.03429   5.539 8.92e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.467 on 213 degrees of freedom
## Multiple R-squared:  0.1259, Adjusted R-squared:  0.1218
## F-statistic: 30.68 on 1 and 213 DF,  p-value: 8.923e-08
```

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

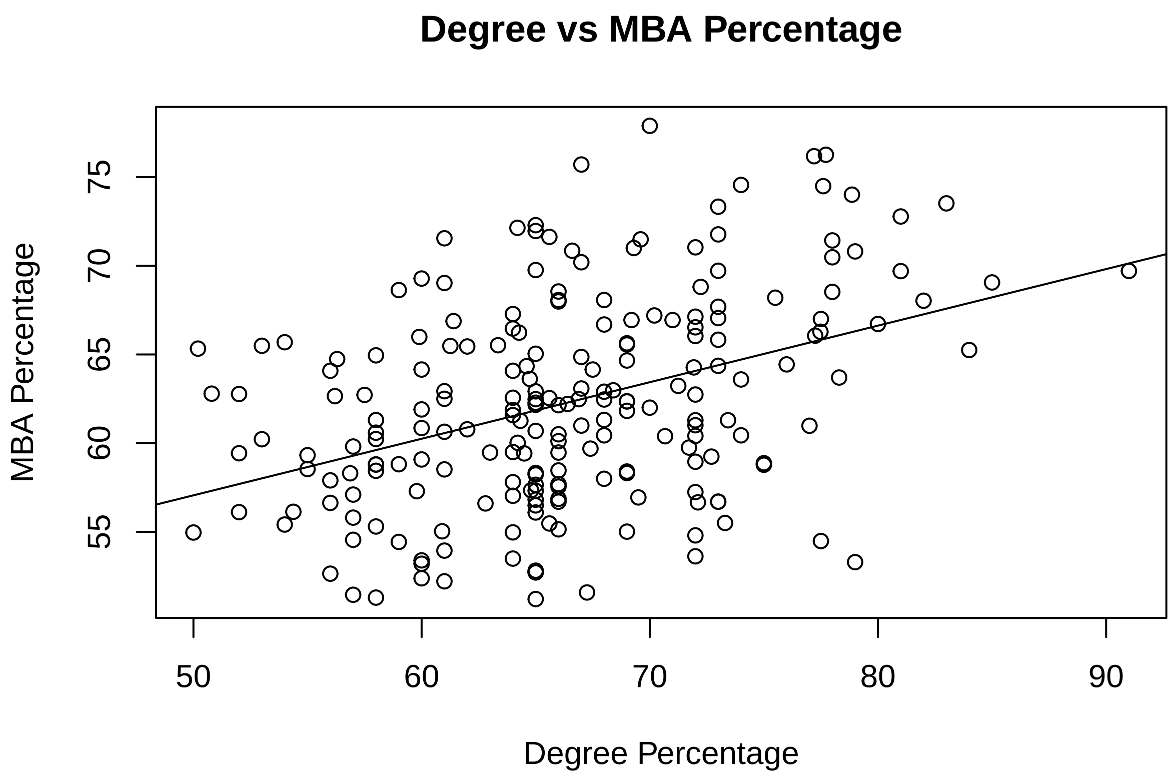


The plot of the residuals vs fitted, does not exactly show linearity, since the bounce around the 0 line is not really random. Secondly, the red line in the scale-location plot is almost horizontal, the points seem to spread out. This is a sign of Heteroscedasticity which is a good sign for the model. In addition, the line of the normal Q-Q plot is almost straight, with a few departures at the lower and upper tail. This indicates the distribution of the model variables, is approximately normal. Conversely, the residuals vs leverage plot, seem to have a couple of influential points in our model, which is not a good sign.

Finally, based on these analysis, we recommend further analysis on the student high school performance vs MBA performance model, as the statistical significance of the fitted model based on the diagnostic plots is not satisfactory. Hence I don't recommend using the model for predicting MBA performance.

(iii) Degree Percentage vs MBA Percentage

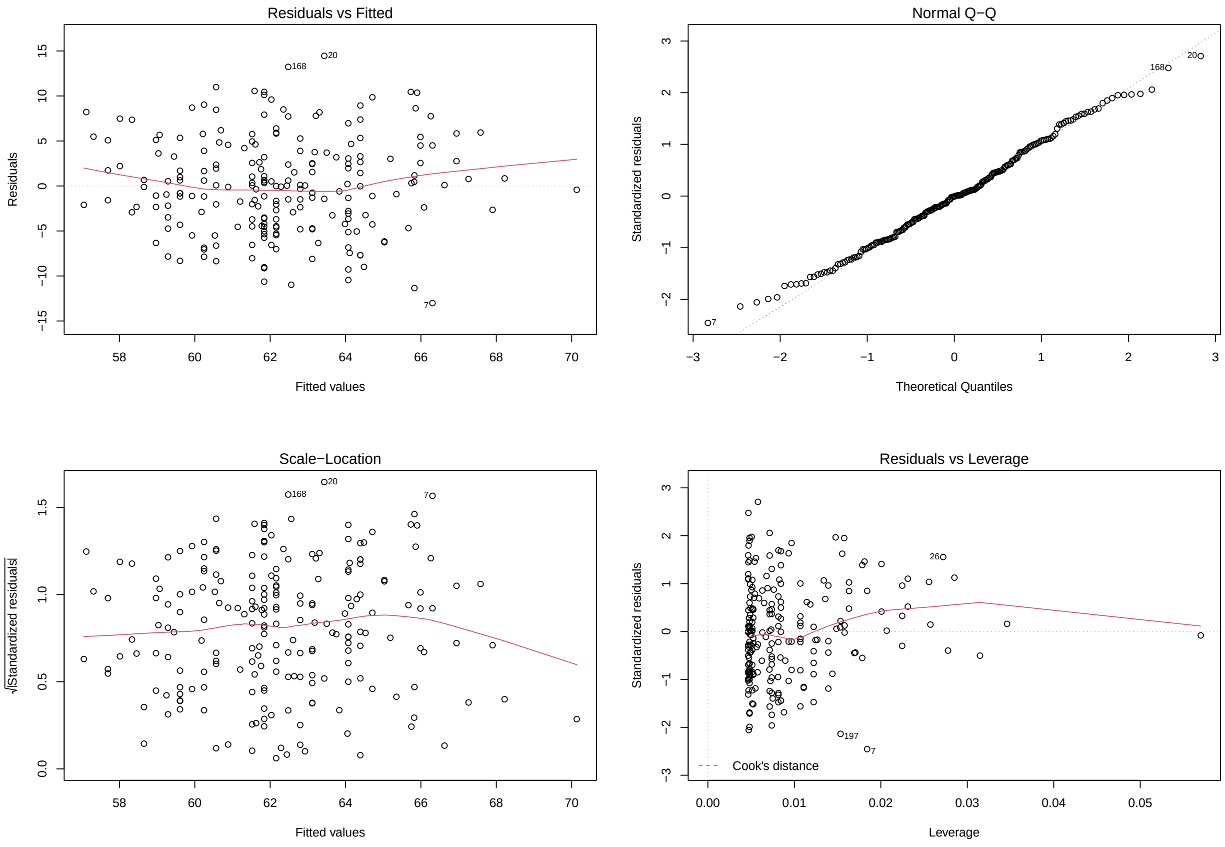
```
plot(data$mba_p~data$degree_p,xlab="Degree Percentage",ylab="MBA
 ↪  Percentage",main="Degree vs MBA Percentage")
reg = lm(data$mba_p~data$degree_p,data = data)
abline(reg)
```



```
summary(reg)
```

```
##
## Call:
## lm(formula = data$mba_p ~ data$degree_p, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.0166  -3.9567  -0.0328   3.6580  14.4540
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    41.10877    3.32040  12.381  < 2e-16 ***
## data$degree_p  0.31896    0.04973   6.414 8.99e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.353 on 213 degrees of freedom
## Multiple R-squared:  0.1619, Adjusted R-squared:  0.158
## F-statistic: 41.15 on 1 and 213 DF,  p-value: 8.993e-10
```

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



The residuals vs fitted plot, does not exactly show linearity, since the bounce around the 0 line is not really random. Secondly, the red line in the scale-location plot is not horizontal, the points seem to converge to the left of the plot This is a sign of Homoscedasticity which is not a good sign for the model. In addition, the line of the normal Q-Q plot is almost straight, with a few departures at the lower and upper tail. This indicates the distribution of the model variables, is approximately normal. Conversely, the residuals vs leverage plot, seem to have a couple of influential points in our model, which is not a good sign.

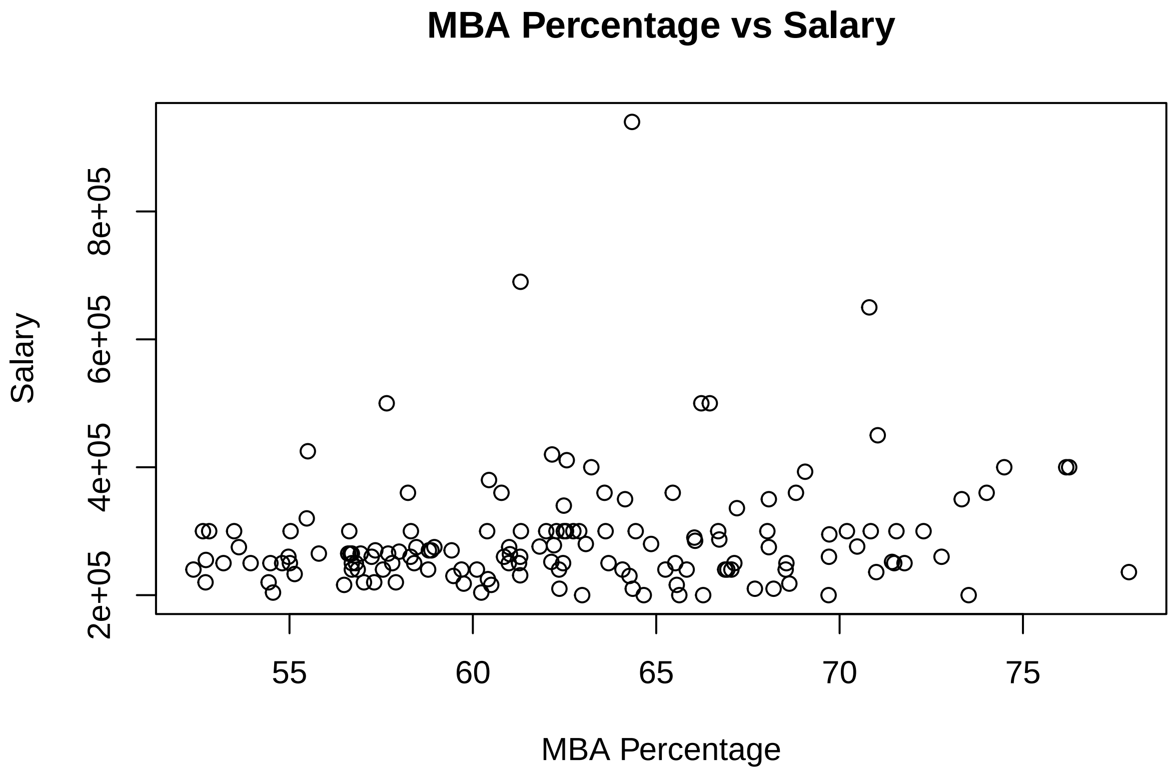
Finally, based on these analysis, we find the model lacking statistical significance of the fitted model based on the diagnostic plots is not satisfactory. Hence we don't recommend using the model for predicting MBA performance.

(iv) Checking the relationship between MBA Percentage of placed students and their salaries.

```
dat1 = data[data$status == "Placed", ]
head(dat1)
```

```
##   sl_no gender ssc_p   ssc_b hsc_p   hsc_b    hsc_s degree_p  degree_t w
## 1     1      M 67.00  Others 91.00  Others Commerce    58.00  Sci&Tech
## 2     2      M 79.33 Central 78.33  Others  Science    77.48  Sci&Tech
## 3     3      M 65.00 Central 68.00 Central     Arts    64.00 Comm&Mgmt
## 5     5      M 85.80 Central 73.60 Central Commerce    73.30 Comm&Mgmt
## 8     8      M 82.00 Central 64.00 Central  Science    66.00  Sci&Tech
## 9     9      M 73.00 Central 79.00 Central Commerce    72.00 Comm&Mgmt
##   etest_p specialisation mba_p status salary
## 1   55.00         Mkt&HR 58.80 Placed 270000
## 2   86.50        Mkt&Fin 66.28 Placed 200000
## 3   75.00        Mkt&Fin 57.80 Placed 250000
## 5   96.80        Mkt&Fin 55.50 Placed 425000
## 8   67.00        Mkt&Fin 62.14 Placed 252000
## 9   91.34        Mkt&Fin 61.29 Placed 231000
```

```
plot(dat1$salary~dat1$mba_p,data=dat1,xlab="MBA
 ↪  Percentage",ylab="Salary",main="MBA Percentage vs Salary")
```



From the plot we can see that there is no correlation between MBA percentage of the students and their salaries.
This could be explained from the fact that companies usually have a fixed grade based payment structure, which may include degree possessed but not necessarily the score obtained in such degree.

Checking the correlation value for the same

```
cor(dat1$salary,dat1$mba_p)
```

```
## [1] 0.1750129
```

Question 3 H_0: There is no association between gender and type of specialization in higher secondary education
H_a: There is association between gender and type of specialization in higher secondary education

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
new1 = aggregate(data1$a,
 ↪  by=list(Gender=data1$gender,Specialization=data1$hsc_s),FUN=sum)
new1
```

```
##   Gender Specialization  x
## 1      F           Arts  6
## 2      M           Arts  5
## 3      F       Commerce 40
## 4      M       Commerce 73
## 5      F        Science 30
## 6      M        Science 61
```

```
observed_table = matrix(c(6,40,30,5,73,61),nrow = 2,ncol =
 ↪  3,byrow = T)
rownames(observed_table) =c('Female','Male')
colnames(observed_table) =c('Arts','Commerce','Science')
observed_table
```

```
##        Arts Commerce Science
## Female    6       40      30
## Male      5       73      61
```

```r
res = chisq.test(observed_table)
res
```

```
##
##  Pearson's Chi-squared test
##
## data:  observed_table
## X-squared = 1.9998, df = 2, p-value = 0.3679
```

Since the P-value is greater than alpha, we fail to reject the null hypothesis and conclude that there is no signinficant relationship between gender and specializtion in higher secondary education.

Conclusion

1. After the statistical analysis of our dataset (regression analysis) we conclude that the marks secured in MBA does not affect the chances of getting placed. Moreover, from the analysis we cannot draw relationship between Employability test percentage and the placement status.
2. Also, we observed that students who did well during the secondary school are more likely to perform better in higher secondary school.
3. Although there is no direct relationship between the past performances and the MBA score, each Percentage contributes to certain extent for the variation of the MBA percentage of student. And we were not able to see any correlation between the salaries of the students and their MBA percentage.
4. Again from the test of independence we can see that there is no association which specialization a student goes for based on their gender.

Reference

1. DG[2020]Campus Recruitment Available at: https://www.kaggle.com/benroshan/factors-affecting-campus-placement/metadata (Accessed: 15th Oct 2021)