CIND 123 - Data Analytics: Basic Methods Fall 2021
Assignment 3

Assignment 3 (10%)

[Eseohe Okafor]

[CIND 123 Fall 2021 Student Number: 501143898]

Instructions

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. Review this website for more details on using R Markdown http://rmarkdown.rstudio.com.

Use RStudio for this assignment. Complete the assignment by inserting your R code wherever you see the string "#INSERT YOUR ANSWER HERE".

When you click the **Knit** button, a document (PDF, Word, or HTML format) will be generated that includes both the assignment content as well as the output of any embedded R code chunks.

Submit **both** the rmd and generated output files. Failing to submit both files will be subject to mark deduction.

Sample Question and Solution

Use `seq()` to create the vector (2,4,6, … ,20).

```
#Insert your code here.
seq(2,20,by = 2)
```

```
##  [1]  2  4  6  8 10 12 14 16 18 20
```

Note:

You will use 'Admission_Predict.csv' for Assignment-3. This dataset includes the data of the applicants of an academic program. Each application has a unique serial number, which represents a particular student. The dataset contains several parameters which are considered important during the application for Masters Programs. The parameters included are :

1) GRE Scores (out of 340)

2) TOEFL Scores (out of 120)

3) University Rating (out of 5)

4) Statement of Purpose (SOP) (out of 5)

5) Letter of Recommendation (LOR) Strength (out of 5)

6) Undergraduate GPA (out of 10)

7) Research Experience (either 0 or 1)

8) Chance of Admit (ranging from 0 to 1)

Download "Admission_Predict.csv" dataset and load it as 'data'.

```
data<-read.csv("Admission_Predict.csv")
```

Question 1 (30 points in total)

a) i- Display the first three rows in this dataset.(1 point)

```
data[1:3,]
```

```
##   Serial.No. GRE.Score TOEFL.Score University.Rating SOP LOR CGPA Research
## 1          1       337         118                 4 4.5 4.5 9.65        1
## 2          2       324         107                 4 4.0 4.5 8.87        1
## 3          3       316         104                 3 3.0 3.5 8.00        1
##   Chance.of.Admit
## 1            0.92
## 2            0.76
## 3            0.72
```

ii - Display the structure of all variables.(1 point)

```
str(data)
```

```
## 'data.frame':    400 obs. of  9 variables:
##  $ Serial.No.       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ GRE.Score        : int  337 324 316 322 314 330 321 308 302 323 ...
##  $ TOEFL.Score      : int  118 107 104 110 103 115 109 101 102 108 ...
##  $ University.Rating: int  4 4 3 3 2 5 3 2 1 3 ...
##  $ SOP              : num  4.5 4 3 3.5 2 4.5 3 3 2 3.5 ...
##  $ LOR              : num  4.5 4.5 3.5 2.5 3 3 4 4 1.5 3 ...
##  $ CGPA             : num  9.65 8.87 8 8.67 8.21 9.34 8.2 7.9 8 8.6 ...
##  $ Research         : int  1 1 1 1 0 1 1 0 0 0 ...
##  $ Chance.of.Admit  : num  0.92 0.76 0.72 0.8 0.65 0.9 0.75 0.68 0.5 0.45
...
```

iii - Print the descriptive statistics of the admission data to understand the data a little better (min, max, mean, median, 1st and 3rd quartiles). (1 point)

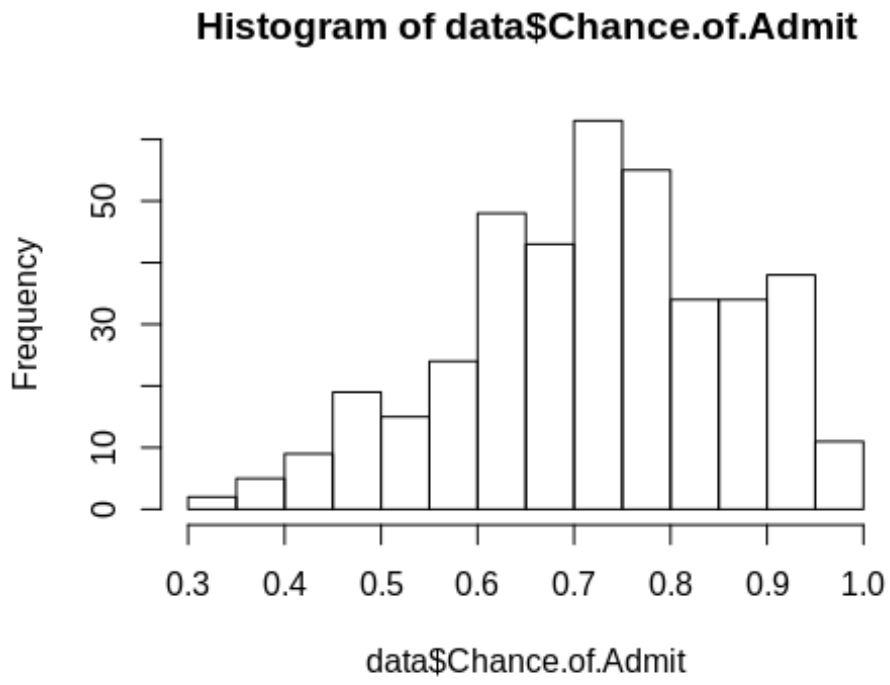
```
summary(data)
```

```
##    Serial.No.      GRE.Score      TOEFL.Score    University.Rating
##  Min.   :  1.0   Min.   :290.0   Min.   : 92.0   Min.   :1.000
```

```
##  1st Qu.:100.8   1st Qu.:308.0   1st Qu.:103.0   1st Qu.:2.000
##  Median :200.5   Median :317.0   Median :107.0   Median :3.000
##  Mean   :200.5   Mean   :316.8   Mean   :107.4   Mean   :3.087
##  3rd Qu.:300.2   3rd Qu.:325.0   3rd Qu.:112.0   3rd Qu.:4.000
##  Max.   :400.0   Max.   :340.0   Max.   :120.0   Max.   :5.000
##       SOP             LOR             CGPA           Research
##  Min.   :1.0   Min.   :1.000   Min.   :6.800   Min.   :0.0000
##  1st Qu.:2.5   1st Qu.:3.000   1st Qu.:8.170   1st Qu.:0.0000
##  Median :3.5   Median :3.500   Median :8.610   Median :1.0000
##  Mean   :3.4   Mean   :3.453   Mean   :8.599   Mean   :0.5475
##  3rd Qu.:4.0   3rd Qu.:4.000   3rd Qu.:9.062   3rd Qu.:1.0000
##  Max.   :5.0   Max.   :5.000   Max.   :9.920   Max.   :1.0000
##  Chance.of.Admit
##  Min.   :0.3400
##  1st Qu.:0.6400
##  Median :0.7300
##  Mean   :0.7244
##  3rd Qu.:0.8300
##  Max.   :0.9700
```

iv - Use a histogram to assess the normality of the 'Chance.of.Admit' variable and explain whether it appears normally distributed or not and why? (1 point)

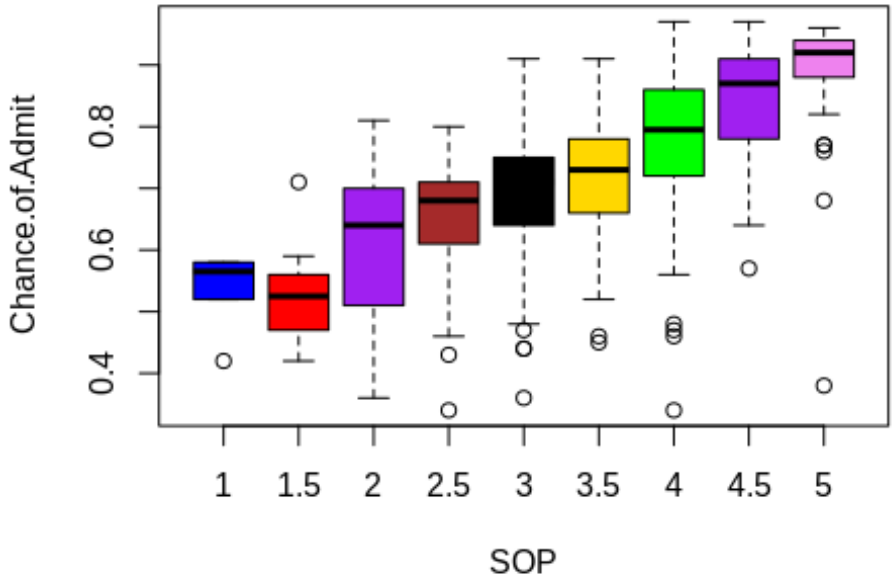
```
hist(data$Chance.of.Admit)
```




```
#This histogram does not appear to be normal but has a left-skewed shape with
the mean of the distribution appearing on the left of the peak.
```

b) Create a set of boxplots that shows the distribution of Chance.of.Admit on SOP variables. Use different colors for different SOP score marks. Hint: SOP scores are changing between 1,1.5, to 5, therefore you can use different box colours for each score likewise; 1 (red), 1.5(green), etc. (8 points)

```
colors <- c("blue", "red", "purple", "brown", "black", "gold", "green", "purple", "violet")
boxplot(Chance.of.Admit ~ SOP, data=data, col=colors)
```



i- Find the covariance between the "GRE.Score" and the "Chance.of.Admit". (3 points)

```
cov(data$GRE.Score, data$Chance.of.Admit, method="pearson")

## [1] 1.313271
```

ii- Print or plot the correlation matrix of the data and write down the correlations between the GRE.Score, TOEFL.Score, CGPA and the Chance.of.Admit. (3 points)

```
cor(data)

##                    Serial.No.   GRE.Score TOEFL.Score University.Rating
## Serial.No.          1.00000000 -0.09752579  -0.1479317        -0.1699479
## GRE.Score          -0.09752579  1.00000000   0.8359768         0.6689759
## TOEFL.Score        -0.14793170  0.83597680   1.0000000         0.6955898
## University.Rating  -0.16994786  0.66897585   0.6955898         1.0000000
## SOP                -0.16693236  0.61283074   0.6579805         0.7345228
## LOR                -0.08822139  0.55755452   0.5677209         0.6601235
## CGPA               -0.04560845  0.83306045   0.8284174         0.7464787
```

```
## Research           -0.06313754  0.58039064   0.4898579         0.4477825
## Chance.of.Admit     0.04233586  0.80261046   0.7915940         0.7112503
##                           SOP         LOR        CGPA    Research
## Serial.No.         -0.1669324 -0.08822139 -0.04560845 -0.06313754
## GRE.Score           0.6128307  0.55755452  0.83306045  0.58039064
## TOEFL.Score         0.6579805  0.56772092  0.82841742  0.48985785
## University.Rating   0.7345228  0.66012345  0.74647869  0.44778251
## SOP                 1.0000000  0.72959254  0.71814396  0.44402881
## LOR                 0.7295925  1.00000000  0.67021130  0.39685926
## CGPA                0.7181440  0.67021130  1.00000000  0.52165423
## Research            0.4440288  0.39685926  0.52165423  1.00000000
## Chance.of.Admit     0.6757319  0.66988879  0.87328910  0.55320214
##                   Chance.of.Admit
## Serial.No.             0.04233586
## GRE.Score              0.80261046
## TOEFL.Score            0.79159399
## University.Rating      0.71125025
## SOP                    0.67573186
## LOR                    0.66988879
## CGPA                   0.87328910
## Research               0.55320214
## Chance.of.Admit        1.00000000
```

```
cor(data$GRE.Score, data$TOEFL.Score)
```

```
## [1] 0.8359768
```

```
cor(data$CGPA,data$Chance.of.Admit)
```

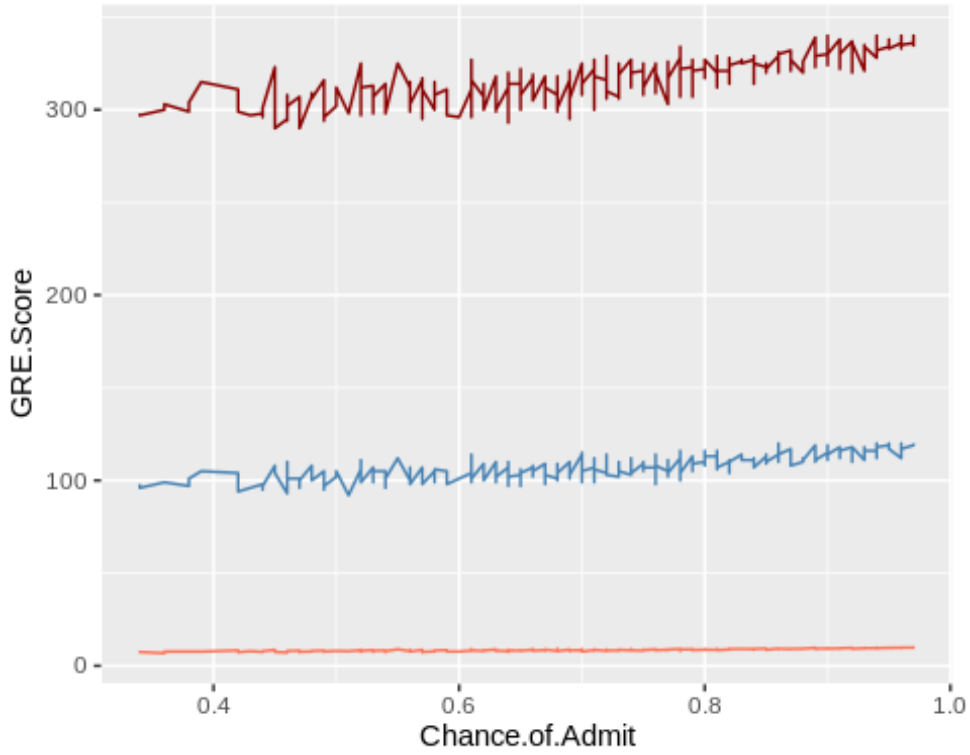
```
## [1] 0.8732891
```

iii - Interpret the covariance and correlation results obtained from c(i) and c(ii) in terms of the strength and direction of the relationship. (4 points)

```
#c(i)The covariance result of 1.313271 indicates a strong positive linear rel
ationship between "GRE.Score" and the "Chance.of.Admit" meaning that the valu
e of "GRE.Score" will increase along with an increase in the value of "Chance
.of.Admit"
#c(ii)The correlation result of 0.8359768 and 0.8732891 indicate also a stron
g positive linear relationship between "GRE.Score" and the "Chance.of.Admit"
,CGPA and "Chance.of.Admit meaning that an increase in the value of one will
result in an increase in the value of the other.
```

d) Use ggplot() to plot the graphs to see the relationship between each of three variables (GRE.Score, TOEFL.Score, CGPA) with Chance.of.Admit. (8 points)

```
library(ggplot2)
ggplot(data, aes(x=Chance.of.Admit)) +
  geom_line(aes(y = GRE.Score), color = "darkred") +
  geom_line(aes(y = TOEFL.Score), color="steelblue")+
  geom_line(aes(y = CGPA), color="coral1")
```



Question 2 (40 points in total)

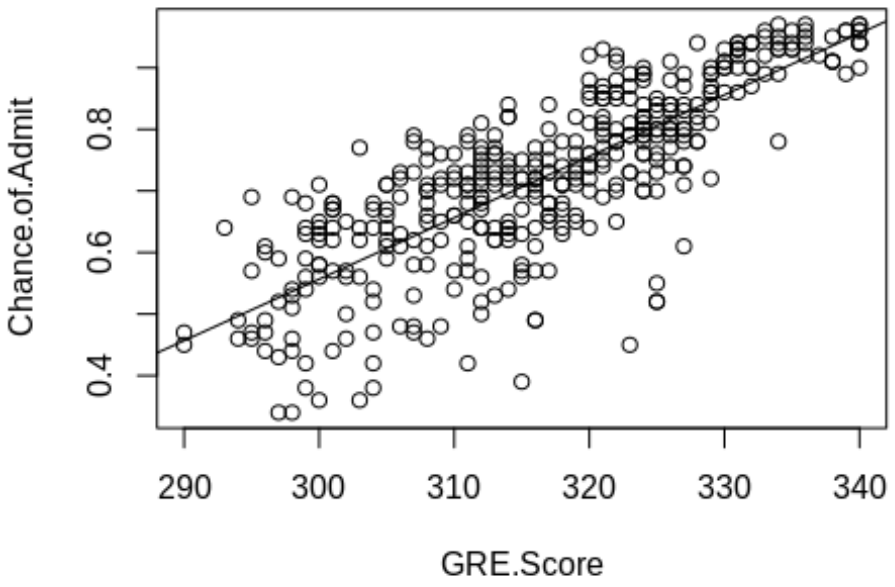
i- Define the linear regression model between GRE.Score and Chance.of.Admit (3 points)

```
lm(Chance.of.Admit ~ GRE.Score, data=data)

##
## Call:
## lm(formula = Chance.of.Admit ~ GRE.Score, data = data)
##
## Coefficients:
## (Intercept)    GRE.Score
##   -2.436084     0.009976
```

ii - Plot the regression (least-square) line on the same plot.(3 points)

```
lmChance.of.Admit=lm(Chance.of.Admit ~ GRE.Score, data=data)
plot(Chance.of.Admit ~ GRE.Score, data=data)
abline(lmChance.of.Admit)
```



ii- - Explain the meaning of the slope and y-intercept for the least-squares regression line in Q2(ii). (3 points)

```
##The slope of 0.009976 indicates that the greater the GRE.score, the higher
the chance of admission, since it is a positive value.


# The y-intercept of -2.436084 shows that the chance of admission for an appl
icant with GRE.Score of 0 is negative.
```

b) Print the results of this model and interpret the results by following questions:

i - What is the number of observations was the regression run on? (3 points)

```
lmChance.of.Admit

##
## Call:
## lm(formula = Chance.of.Admit ~ GRE.Score, data = data)
##
## Coefficients:
## (Intercept)    GRE.Score
##   -2.436084     0.009976

str(lmChance.of.Admit)

## List of 12
##  $ coefficients : Named num [1:2] -2.43608 0.00998
```

```
##   ..- attr(*, "names")= chr [1:2] "(Intercept)" "GRE.Score"
##  $ residuals    : Named num [1:400] -0.00579 -0.0361 0.00371 0.02385 -0.04
634 ...
##   ..- attr(*, "names")= chr [1:400] "1" "2" "3" "4" ...
##  $ effects      : Named num [1:400] -14.487 -2.28633 0.00371 0.02477 -0.04
664 ...
##   ..- attr(*, "names")= chr [1:400] "(Intercept)" "GRE.Score" "" "" ...
##  $ rank         : int 2
##  $ fitted.values: Named num [1:400] 0.926 0.796 0.716 0.776 0.696 ...
##   ..- attr(*, "names")= chr [1:400] "1" "2" "3" "4" ...
##  $ assign       : int [1:2] 0 1
##  $ qr           :List of 5
##   ..$ qr   : num [1:400, 1:2] -20 0.05 0.05 0.05 0.05 0.05 0.05 0.05 0.05
0.05 ...
##   .. ..- attr(*, "dimnames")=List of 2
##   .. .. ..$ : chr [1:400] "1" "2" "3" "4" ...
##   .. .. ..$ : chr [1:2] "(Intercept)" "GRE.Score"
##   .. ..- attr(*, "assign")= int [1:2] 0 1
##   ..$ qraux: num [1:2] 1.05 1.03
##   ..$ pivot: int [1:2] 1 2
##   ..$ tol  : num 1e-07
##   ..$ rank : int 2
##   ..- attr(*, "class")= chr "qr"
##  $ df.residual  : int 398
##  $ xlevels      : Named list()
##  $ call         : language lm(formula = Chance.of.Admit ~ GRE.Score, data
= data)
##  $ terms        :Classes 'terms', 'formula'  language Chance.of.Admit ~ GR
E.Score
##   .. ..- attr(*, "variables")= language list(Chance.of.Admit, GRE.Score)
##   .. ..- attr(*, "factors")= int [1:2, 1] 0 1
##   .. .. ..- attr(*, "dimnames")=List of 2
##   .. .. .. ..$ : chr [1:2] "Chance.of.Admit" "GRE.Score"
##   .. .. .. ..$ : chr "GRE.Score"
##   .. ..- attr(*, "term.labels")= chr "GRE.Score"
##   .. ..- attr(*, "order")= int 1
##   .. ..- attr(*, "intercept")= int 1
##   .. ..- attr(*, "response")= int 1
##   .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##   .. ..- attr(*, "predvars")= language list(Chance.of.Admit, GRE.Score)
##   .. ..- attr(*, "dataClasses")= Named chr [1:2] "numeric" "numeric"
##   .. .. ..- attr(*, "names")= chr [1:2] "Chance.of.Admit" "GRE.Score"
##  $ model        :'data.frame':   400 obs. of  2 variables:
##   ..$ Chance.of.Admit: num [1:400] 0.92 0.76 0.72 0.8 0.65 0.9 0.75 0.68 0
.5 0.45 ...
##   ..$ GRE.Score      : int [1:400] 337 324 316 322 314 330 321 308 302 323
...
##   ..- attr(*, "terms")=Classes 'terms', 'formula'  language Chance.of.Admi
t ~ GRE.Score
##   .. .. ..- attr(*, "variables")= language list(Chance.of.Admit, GRE.Score
```

```
)
##   .. .. ..- attr(*, "factors")= int [1:2, 1] 0 1
##   .. .. .. ..- attr(*, "dimnames")=List of 2
##   .. .. .. .. ..$ : chr [1:2] "Chance.of.Admit" "GRE.Score"
##   .. .. .. .. ..$ : chr "GRE.Score"
##   .. .. ..- attr(*, "term.labels")= chr "GRE.Score"
##   .. .. ..- attr(*, "order")= int 1
##   .. .. ..- attr(*, "intercept")= int 1
##   .. .. ..- attr(*, "response")= int 1
##   .. .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##   .. .. ..- attr(*, "predvars")= language list(Chance.of.Admit, GRE.Score)
##   .. .. ..- attr(*, "dataClasses")= Named chr [1:2] "numeric" "numeric"
##   .. .. .. ..- attr(*, "names")= chr [1:2] "Chance.of.Admit" "GRE.Score"
##  - attr(*, "class")= chr "lm"
```

```
summary(lmChance.of.Admit)
```

```
##
## Call:
## lm(formula = Chance.of.Admit ~ GRE.Score, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33613 -0.04604  0.00408  0.05644  0.18339
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.4360842  0.1178141  -20.68   <2e-16 ***
## GRE.Score    0.0099759  0.0003716   26.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08517 on 398 degrees of freedom
## Multiple R-squared:  0.6442, Adjusted R-squared:  0.6433
## F-statistic: 720.6 on 1 and 398 DF,  p-value: < 2.2e-16
```

```
#The number of observations is 400
```

ii - Interpret the R-squared of this regression? (4 points)

```
#The R-squared value is 0.6442 which means that 64.42% of the observations in
Chance of admit is explained by the GRE score
```

iii - Write the regression equation associated with this regression model? (4 points)

```
# Chance.of.Admit = (-2.4360842) + (0.0099759 * GRE.Score)
```

c) Use the regression line to predict the chance of admit when GRE score 310. (10 points)

```
predict(lmChance.of.Admit, newdata = data.frame(GRE.Score=310))
```

```
##         1
## 0.6564392
```

d) From the given Q2(a) linear model between GRE.Score and Chance.of.Admit, what should be GRE score of a student who has 50% of chance of admission?(10 points)

```
# 0.50 = (-2.4360842) + (0.0099759 * GRE.Score)
# 0.50 + 2.4360842 = 0.0099759 * GRE.Score
# (0.50 + 2.4360842)/0.0099759 = GRE.Score
#                              = 294.3177
```

Question 3 (30 points in total)

a) Use three independent variables ('GRE.Score','TOEFL.Score', 'CGPA') to build a multiple linear regression model to predict dependent variable 'Chance.of.Admit'. Display a summary of your model indicating Residuals, Coefficients, etc. Explain the summary results. (8 points)

```
 lmChance.of.Admit2<- lm(Chance.of.Admit ~ GRE.Score + TOEFL.Score + CGPA, da
ta=data)
summary(lmChance.of.Admit2)

##
## Call:
## lm(formula = Chance.of.Admit ~ GRE.Score + TOEFL.Score + CGPA,
##     data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.290375 -0.023030  0.008255  0.040153  0.143108
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.5856984  0.1058153 -14.986  < 2e-16 ***
## GRE.Score    0.0022660  0.0005929   3.822 0.000154 ***
## TOEFL.Score  0.0031123  0.0011070   2.812 0.005176 **
## CGPA         0.1462844  0.0111770  13.088  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06632 on 396 degrees of freedom
## Multiple R-squared:  0.7854, Adjusted R-squared:  0.7837
## F-statistic:   483 on 3 and 396 DF,  p-value: < 2.2e-16

#The Residual values do not appear to be symmetrical indicating a large varia
tion between observed values and the predicted ones.
#The small P-values indicates that there is a relationship between the chance
of admit and the independent variables.
#The Residual standard error value shows that the actual chance of admit will
deviate from the regression line by 0.06632 points.
#The Multiple R-squared value indicates that 78.5% of the variance in the cha
nce to admit can be explained by the the three independent variables.
#The F statistics of 483 is highly significant.
```

b) Write the regression equation associated with this multiple regression model. (8 points)

```
#Chance.of.Admit = (-1.5856984*0.0022660)+(-1.5856984*0.0031123)+(-1.5856984*
0.1462844)+0
```

c) Using this model:

i- Find the chance of admit for the 3rd student and 23rd students in the dataset. (4 points)

```
predict(lmChance.of.Admit2)[c(3, 23)]

##         3        23
## 0.6242940 0.9082592
```

ii- Identify which student (3rd or 23rd) has higher chance than the other and print the difference between the chance of admit of these two students.(3 points)

```
#The 23rd student has the higher chance of admit and the differnce between th
e two students is 0.2839652
```

d) Explain the difference between the linear regression models in Question 2 and in Question 3. (7 points)

```
#The linear regression model in question 2 has just one independent variable,
GRE.Score as against the model in question 3 which has three independent vari
ables,GRE.Score,TOEFL.Score and CGPA. Having 3 independent variables against
the single independent variable in question 2 will make it a better predictiv
e model.
```