Statistical Inference

Project phase #2

Erfan Mirhaji

Student ID: 810196568

Question 1:

- A) Two variables of University and SOP are selected. With levels “b” and “3.5”, from each one respectively being compared. The confidence interval of the difference of the proportions is calculated as below:

```
CI <- function(mean, SE, alpha) {
  z <- qnorm(alpha/2 + 0.5)
  return(c(mean - z*SE, mean + z*SE))
}
n <- nrow(UniversityAdmissions)
p1 <- nrow(UniversityAdmissions[which(UniversityAdmissions$University == "b"),])/n
p2 <- nrow(UniversityAdmissions[which(UniversityAdmissions$SOP == "3.5"),])/n
mean <- p1 - p2
SE <- sqrt(p1*(1-p1) + p2*(1-p2))/sqrt(n)
CI <- CI(mean, SE, 0.95)
cat("\nConfidence Interval:", CI, "\n")
```

```
Confidence interval: -0.01707137 0.08277979
```

We are 95% sure that the compared proportions have a difference in means between -0.01707137 & 0.08277979

- B) A hypothesis test is performed & based on the chi-square independence test, we have:

```
tbl = table(UniversityAdmissions$University, UniversityAdmissions$SOP)
tbl
chisq.test(tbl)
```

```
> tbl

    1 1.5  2 2.5  3 3.5  4 4.5  5
a   0   0  0   0  2   3 18  23 27
b   0   2  0   3 12  11 31  32 13
c   0   0 12  15 34  61 30   5  2
d   1  11 19  41 27  11 10   3  0
e   5   7  7   3  4   2  0   0  0
> chisq.test(tbl)

        Pearson's Chi-squared test

data:  tbl
X-squared = 462.68, df = 32, p-value < 2.2e-16
```

Regarding the p-value, we fail to reject the null hypothesis, showing that the two variables are independent.

Also, the conditions for inference are mostly satisfied:

- Observations are independent within each group.
- Each set meets the success-failure condition.
- The independence of the two groups is also assumed.

Question 2:

- A) A sample size of 10 is selected from the data and the proportion of Research variable as “1” is calculated, then a simulation is performed 100 times to calculate the p-value for the hypothesis test.

```
sample <- sample_n(na.omit(UniversityAdmissions), 10)$Research
psub <- sum(sample == "1")/10
cat("\n\nSample proportion =", psub, "\n")
sim <- c()
for (i in 1:1000) {
  sim <- append(sim, sum(sample(c(0, 1), 10, replace = TRUE))/10)
}
cat("P-value =", sum(sim >= psub)/1000, "\n")
```

```
Sample proportion = 0.7    P-value = 0.161
```

It can be seen that the proportion of simulations with outcomes at least as extreme as the sample proportion is not big enough to reject null hypothesis with 95% level.

Question 3:

- A) We calculate the probability distribution:

```
data <- UniversityAdmissions$SOP
n <- nrow(data)
x <- summary(data)
x
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.500   3.500   3.412   4.000   5.000
```

The variable “SOP” is selected which contains 9 levels. A random sample of x_1 and a biased sample of $x_2$with sizes of 100 are selected from the data. $x_2$is only selected from admissions with an SOP of “2.5” and lower, therefore is biased.

```
pop <- sample(data , size = 100)
pop
x1 <- sample(data , size = 100)
x1
x2 <- sample(data[which(data < "2.5")] , size =100)
x2
chisq.test(pop , x1)
chisq.test(pop , x2)
```

```
> chisq.test(pop , x1)

        Pearson's Chi-squared test

data:  pop and x1
X-squared = 70.696, df = 56, p-value = 0.08936

Warning message:
In chisq.test(pop, x1) : Chi-squared approximation may be incorrect
> chisq.test(pop , x2)

        Pearson's Chi-squared test

data:  pop and x2
X-squared = 14.792, df = 24, p-value = 0.9267
```

It clearly can be seen that null hypothesis of having the same distributions can't be rejected for either samples, as the p-value is high enough.

- B) We have:

```
data1 <- UniversityAdmissions$SOP
data2 <- UniversityAdmissions$LOR
tbl <- table(data1 , data2)
tbl
chisq.test(tbl)
```

The table is shown below, of which the following chi-squared test was performed

```
> tbl
     data2
data1  1 1.5  2 2.5  3 3.5  4 4.5  5
  1    1   1  2   2  0   0  0   0  0
  1.5  0   1 11   3  3   2  0   0  0
  2    0   3  4  15  7   4  4   1  0
  2.5  0   4 12   7 18   9  9   3  0
  3    0   2  7   9 23  23  9   4  2
  3.5  0   0  3   8 30  16 22   4  5
  4    0   0  1   4 10  20 19  21 14
  4.5  0   0  0   0  3   9 24  18  9
  5    0   0  0   0  1   3  6  12 20
> chisq.test(tbl)

        Pearson's Chi-squared test

data:  tbl
X-squared = 435.06, df = 64, p-value < 2.2e-16
```

Regarding the p-value, it can be said that the two variables are independent as expected, and the null hypothesis is rejected.

Question 4:

- A) We choose CGPA as response variable. We predict the variables TOEFL and SOP are the most significant predictors, as they are less dependent on each other, and also other variables are more dependent on these two variables and can be roughly predicted based on these two explanatories.

- B)

    a) The variable CGPA is selected for response and SOP and TOEFL are selected as explanatories. A linear regression model is fitted on data as shown below:

```
dataSOP <- UniversityAdmissions$SOP
dataTOEFL <-UniversityAdmissions$TOEFL
dataCGPA <-UniversityAdmissions$CGPA
model1 <- lm(dataCGPA ~ dataSOP   , UniversityAdmissions )
model2 <- lm(dataCGPA ~ dataTOEFL , UniversityAdmissions )
summary(model1)
summary(model2)
```

```
Call:
lm(formula = dataCGPA ~ dataSOP, data = UniversityAdmissions)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6693 -0.0900  0.0458  0.1849  0.5658

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.09830    0.06848   45.24   <2e-16 ***
dataSOP      0.26164    0.01930   13.55   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4143 on 485 degrees of freedom
Multiple R-squared:  0.2747,    Adjusted R-squared:  0.2732
F-statistic: 183.7 on 1 and 485 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = dataCGPA ~ dataTOEFL, data = UniversityAdmissions)

Residuals:
     Min       1Q   Median       3Q      Max
-3.3626 -0.0725   0.0374   0.1424   0.6500

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.300371   0.317051  -4.101 4.81e-05 ***
dataTOEFL    0.049261   0.002947  16.715  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3875 on 485 degrees of freedom
Multiple R-squared:  0.3655,    Adjusted R-squared:  0.3642
F-statistic: 279.4 on 1 and 485 DF,  p-value: < 2.2e-16
```

b)

Equation of the models can be written as:

$$CGPA = 3.09830 + 0.26164(SOP)$$
$$CGPA = -1.300371 + 0.049261(TOEFL)$$

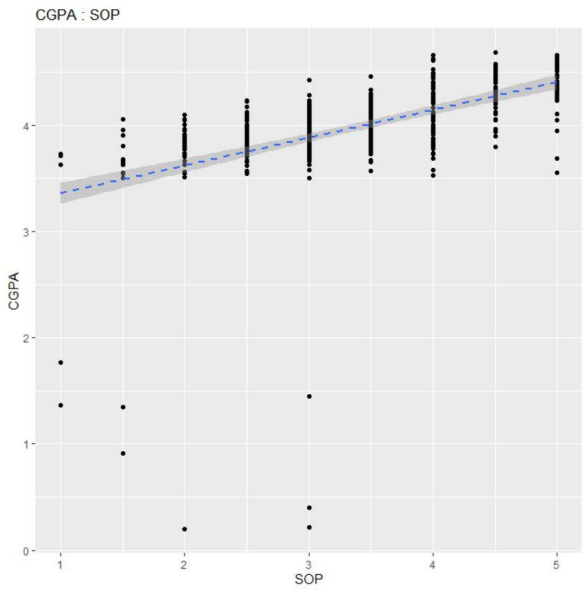
One unit increase in SOP will increase the CGPA by 0.26164units. Also, an interpretation for the intercept parameter is that an admission with SOP as 0 would have 3.09830 units on its CGPA. Also, the adjusted r-squared is 0.2732 meaning 27.3% of the variability of the CGPA is explained by the model.

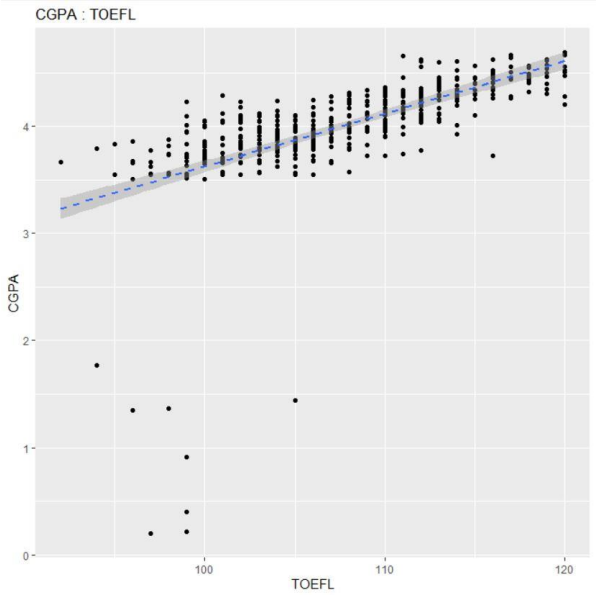
One unit increase in TOEFL will increase the CGPA by 0.049261. Also, an interpretation for the intercept parameter is that an admission with TOEFL as 0 would have -1.300371 units on its CGPA. Also, the adjusted r-squared is 0.3642 meaning 36.4% of the variability of the CGPA is explained by the model.

c) The following code is executed for plotting the scatter plot and the fitted line:

```
ggplot(data, aes(SOP, CGPA)) +
 geom_point() +
 geom_smooth(method = "lm", linetype = "dashed") +
 ggtitle("CGPA : SOP")
```

```
ggplot(data, aes(TOEFL, CGPA)) +
  geom_point() +
  geom_smooth(method = "lm", linetype = "dashed") +
  ggtitle("CGPA : TOEFL")
```





- C) Using the adjusted r-squared and also by taking a look at the plots above , we conclude that TOEFL is a better explanatory variable for CGPA as a response variable, and TOEFL is more significant. TOEFL has higher adjusted r-squared value, and also has a steeper slope in the linear regression plot.

- D) We write the R code:

```
anova(model1)
anova(model2)
```

```
Call:
lm(formula = CGPA ~ TOEFL, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3626 -0.0725  0.0374  0.1424  0.6500

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.300371   0.317051  -4.101 4.81e-05 ***
TOEFL        0.049261   0.002947  16.715  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3875 on 485 degrees of freedom
Multiple R-squared:  0.3655,    Adjusted R-squared:  0.3642
F-statistic: 279.4 on 1 and 485 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = CGPA ~ SOP, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6693 -0.0900  0.0458  0.1849  0.5658

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.09830    0.06848   45.24   <2e-16 ***
SOP          0.26164    0.01930   13.55   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4143 on 485 degrees of freedom
Multiple R-squared:  0.2747,    Adjusted R-squared:  0.2732
F-statistic: 183.7 on 1 and 485 DF,  p-value: < 2.2e-16
```

TOEFL has a higher adjusted r-squared value, so is a better predictor.

```
> anova(model1)
Analysis of Variance Table

Response: CGPA
           Df Sum Sq Mean Sq F value    Pr(>F)
SOP         1 31.538 31.5378  183.72 < 2.2e-16 ***
Residuals 485 83.256  0.1717
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(model2)
Analysis of Variance Table

Response: CGPA
           Df Sum Sq Mean Sq F value    Pr(>F)
TOEFL       1 41.958  41.958  279.38 < 2.2e-16 ***
Residuals 485 72.837   0.150
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

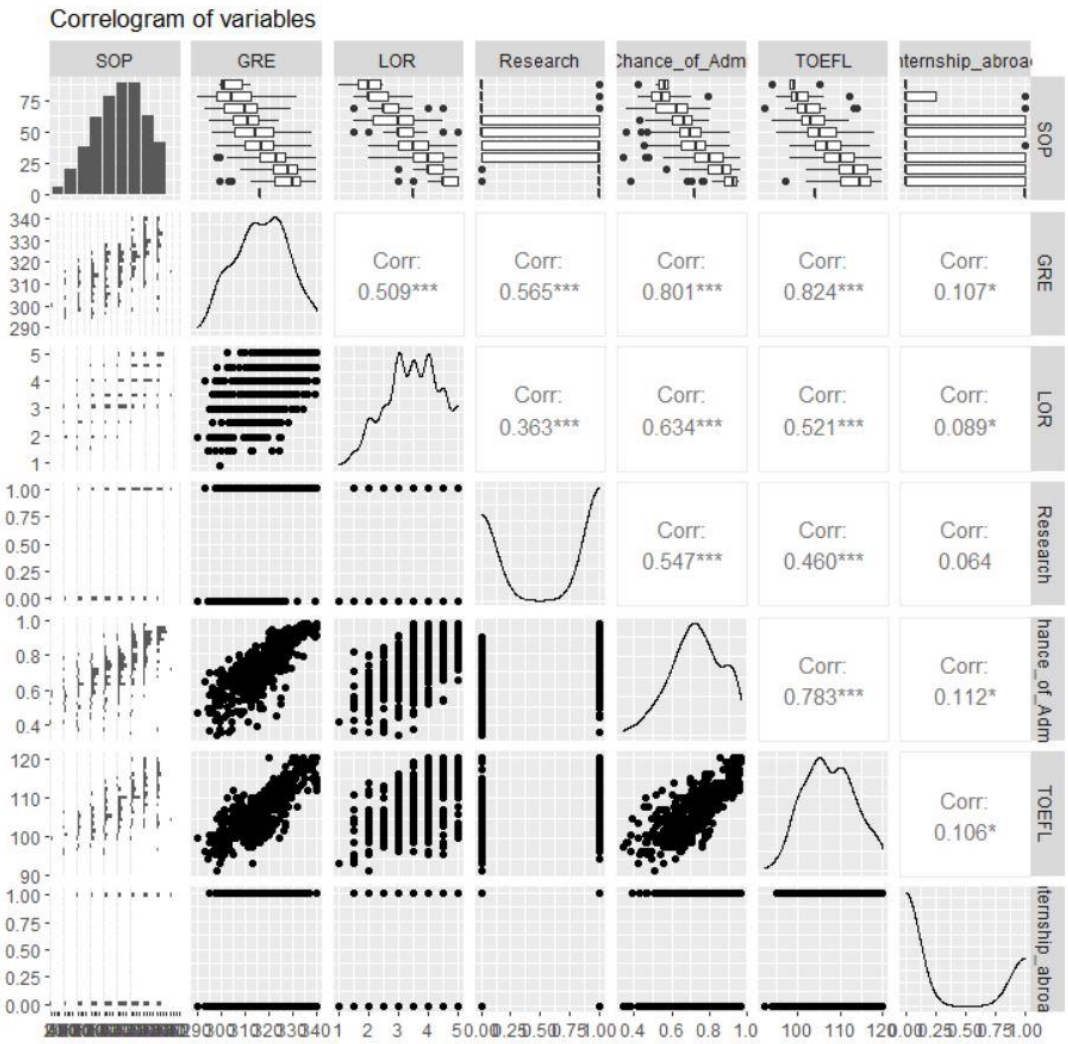
TOEFL has a higher F-value, so is a better predictor.

- E) A good predictor has high F-value, High p-value, also has a steeper slope on the regression line.

Question 5:

- A) R code to plot correlogram:

```
data <- UniversityAdmissions
expvars <- data[c("SOP", "GRE", "LOR", "Research",
                  "Chance_of_Admit", "TOEFL", "internship_abroad")]
ggpairs(expvars, title="Correlogram of variables")
```



Because internship abroad has low correlation values with other variables, it seems like it has the most significance in the prediction. After that, GRE and research also have high significances.

- B) We use the code:

```
GRE <-UniversityAdmissions$GRE
Research <-UniversityAdmissions$Research
internship_abroad <-UniversityAdmissions$internship_abroad
modelg <- lm( CGPA ~ GRE + Research + internship_abroad ,
UniversityAdmissions )
summary(modelg)
```

```
Call:
lm(formula = CGPA ~ GRE + Research + internship_abroad, data = UniversityAdmissions)

Residuals:
     Min       1Q   Median       3Q      Max
-3.2786 -0.0607  0.0444  0.1526  0.5354

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -4.109614   0.596508  -6.889 1.75e-11 ***
GRE               0.025454   0.001927  13.211  < 2e-16 ***
Research          0.037357   0.042904   0.871    0.384
internship_abroad 0.044899   0.038631   1.162    0.246
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3868 on 483 degrees of freedom
Multiple R-squared:  0.3704,    Adjusted R-squared:  0.3665
F-statistic: 94.73 on 3 and 483 DF,  p-value: < 2.2e-16
```

$$CGPA = -4.109614 + 0.025454(GRE) + 0.037357(Research) + 0.044899(Internship\ abroad)$$

- C) Based on the adjusted r-squared value, which is 0.3665, around 36.65% of the variation in the response variable is explained by the model.

- D) Because the adjusted r-squared value isn't big, the response is moderately explained well by the model, but the model isn't the best.

- E) Two models are created for the model selection process, an empty model and one with all the explanatory variables. We use forward and backward selection methods for the best model.

```
data <- UniversityAdmissions
modelfull <- lm(CGPA ~ SOP + internship_abroad+ Chance_of_Admit +
                GRE + LOR + Research + TOEFL, data)
modelnull <- lm(CGPA ~ 1, data)
```

First, backward elimination:

```
cat("\n\n*** Backward elimination ***\n\n")
bestbw <- step(modelfull, direction = "backward")
```

```
Start:  AIC=-1013.94
CGPA ~ SOP + internship_abroad + Chance_of_Admit + GRE + LOR +
    Research + TOEFL

                   Df Sum of Sq    RSS      AIC
- internship_abroad  1    0.0762 56.934 -1015.29
- Research           1    0.0970 56.954 -1015.11
- LOR                1    0.1516 57.009 -1014.64
<none>                          56.857 -1013.94
- TOEFL              1    0.3508 57.208 -1012.94
- GRE                1    0.4933 57.351 -1011.73
- SOP                9    2.8495 59.707 -1008.12
- Chance_of_Admit    1    5.2135 62.071  -973.21

Step:  AIC=-1015.29
CGPA ~ SOP + Chance_of_Admit + GRE + LOR + Research + TOEFL

                  Df Sum of Sq    RSS      AIC
- Research         1    0.0975 57.031 -1016.45
- LOR              1    0.1542 57.088 -1015.97
<none>                        56.934 -1015.29
- TOEFL            1    0.3567 57.290 -1014.24
- GRE              1    0.4998 57.433 -1013.03
- SOP              9    2.8469 59.781 -1009.52
- Chance_of_Admit  1    5.2309 62.164  -974.48

Step:  AIC=-1016.45
CGPA ~ SOP + Chance_of_Admit + GRE + LOR + TOEFL

                  Df Sum of Sq    RSS     AIC
- LOR              1    0.1492 57.180 -1017.2
<none>                        57.031 -1016.5
- TOEFL            1    0.3950 57.426 -1015.1
- GRE              1    0.4194 57.451 -1014.9
- SOP              9    2.8230 59.854 -1010.9
- Chance_of_Admit  1    5.1433 62.174  -976.4

Step:  AIC=-1017.18
CGPA ~ SOP + Chance_of_Admit + GRE + TOEFL

                  Df Sum of Sq    RSS      AIC
<none>                        57.180 -1017.18
- TOEFL            1    0.3839 57.564 -1015.92
- GRE              1    0.3991 57.580 -1015.79
- SOP              9    3.1327 60.313 -1009.20
- Chance_of_Admit  1    6.2003 63.381  -969.04
```

And the code for forward selection:

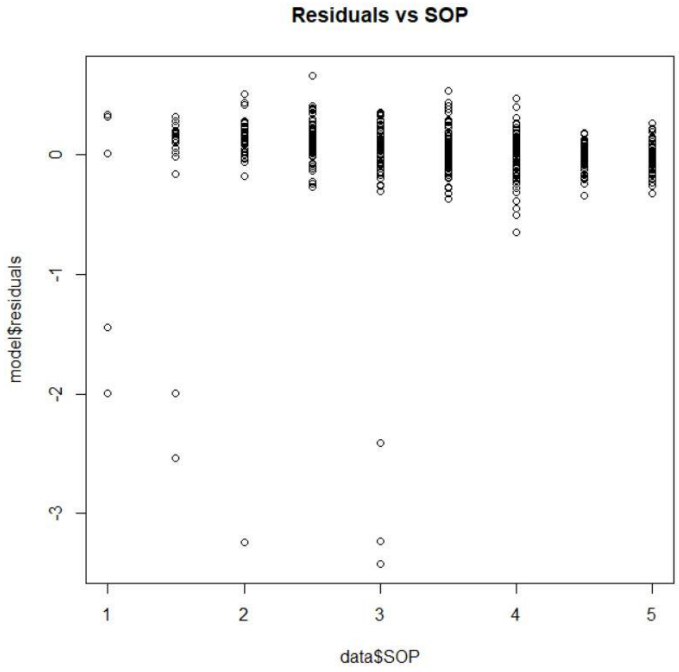
```
data <- UniversityAdmissions

cat("\n\n*** Forward selection ***\n\n")

bestfw <- step(modelnull, direction = "forward", scope = (~ LOR + SOP + GRE + Chance_of_Admit  + Research + internship_abroad  + TOEFL) )
```

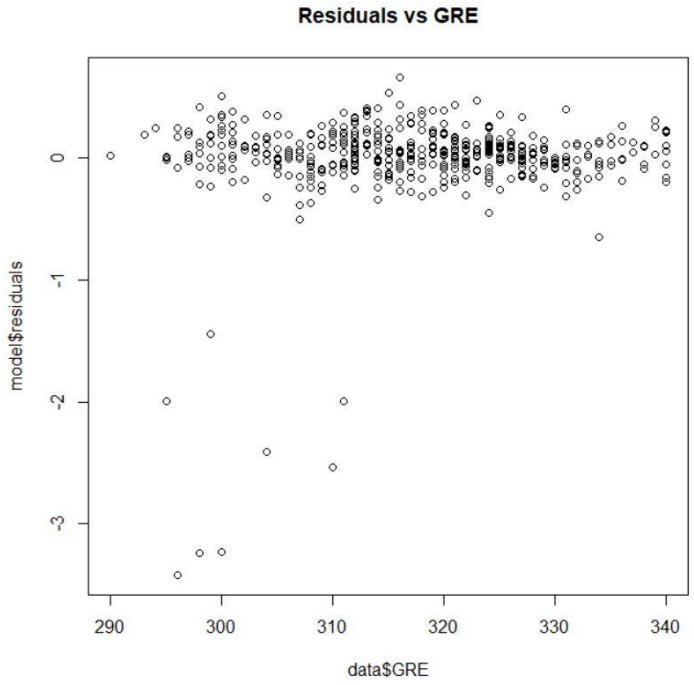
```
Start:  AIC=-701.78
CGPA ~ 1

                    Df Sum of Sq     RSS     AIC
+ Chance_of_Admit    1    52.503  62.291 -997.49
+ GRE                1    42.206  72.589 -922.98
+ TOEFL              1    41.958  72.837 -921.32
+ SOP                1    31.538  83.256 -856.21
+ LOR                1    26.556  88.238 -827.91
+ Research           1    15.615  99.179 -770.98
+ internship_abroad  1     1.304 113.490 -705.34
<none>                           114.794 -701.78

Step:  AIC=-997.49
CGPA ~ Chance_of_Admit

                    Df Sum of Sq    RSS      AIC
+ TOEFL              1   1.66346 60.628 -1008.67
+ GRE                1   1.33963 60.952 -1006.08
+ SOP                1   1.04825 61.243 -1003.75
+ LOR                1   0.52780 61.763  -999.63
<none>                           62.291  -997.49
+ internship_abroad  1   0.11211 62.179  -996.36
+ Research           1   0.00028 62.291  -995.49

Step:  AIC=-1008.67
CGPA ~ Chance_of_Admit + TOEFL

                    Df Sum of Sq    RSS     AIC
+ SOP                1   0.54768 60.080 -1011.1
+ LOR                1   0.43745 60.190 -1010.2
+ GRE                1   0.31471 60.313 -1009.2
<none>                           60.628 -1008.7
+ internship_abroad  1   0.08763 60.540 -1007.4
+ Research           1   0.00915 60.619 -1006.7

Step:  AIC=-1011.09
CGPA ~ Chance_of_Admit + TOEFL + SOP

                    Df Sum of Sq    RSS     AIC
+ GRE                1  0.296912 59.783 -1011.5
<none>                           60.080 -1011.1
+ LOR                1  0.154463 59.926 -1010.3
+ internship_abroad  1  0.075690 60.004 -1009.7
+ Research           1  0.015928 60.064 -1009.2

Step:  AIC=-1011.5
CGPA ~ Chance_of_Admit + TOEFL + SOP + GRE

                    Df Sum of Sq    RSS     AIC
<none>                           59.783 -1011.5
+ LOR                1  0.172654 59.611 -1010.9
+ Research           1  0.076966 59.706 -1010.1
+ internship_abroad  1  0.071336 59.712 -1010.1
```

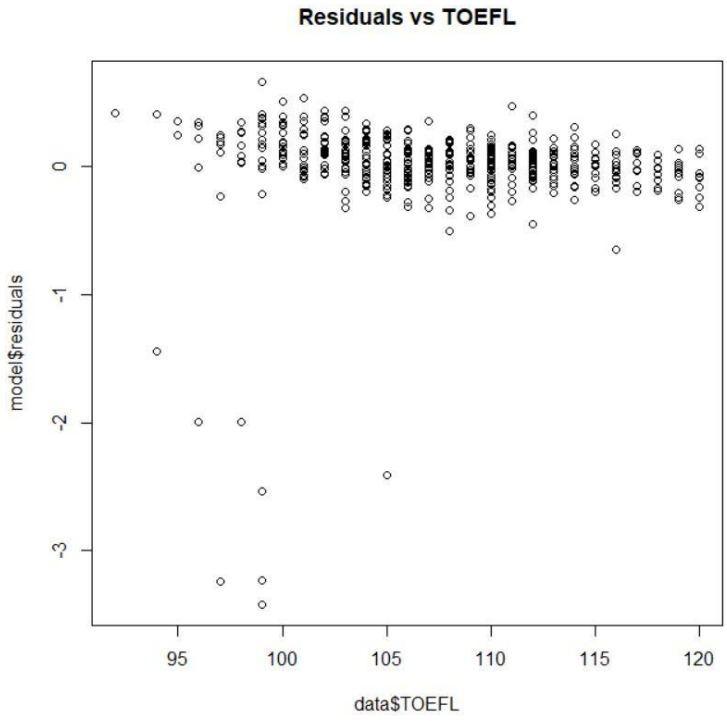
Which resulted the same final model with backward elimination.

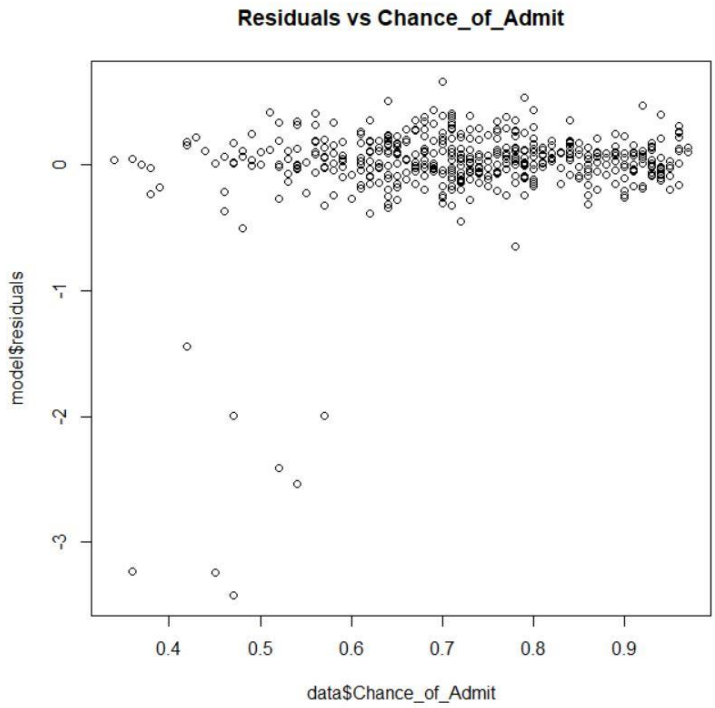
- F) The three conditions are checked with a number of residual plots.

```
plot(model$residuals ~ data$SOP, main = "Residuals vs SOP")
plot(model$residuals ~ data$GRE, main = "Residuals vs GRE")
plot(model$residuals ~ data$TOEFL, main = "Residuals vs TOEFL")
plot(model$residuals ~ data$Chance_of_Admit, main = "Residuals vs
Chance_of_Admit")
```




Residuals vs GRE
model$residuals
0
-1
-2
-3
290
300
310
320
330
340
data$GRE


Residuals vs TOEFL
model$residuals
0
-1
-2
-3
95
100
105
110
115
120
data$TOEFL



Question 6:

- A) A logistic regression model for the response variable of internship abroad is created using the explanatory variables of LOR, GRE & TOEFL.

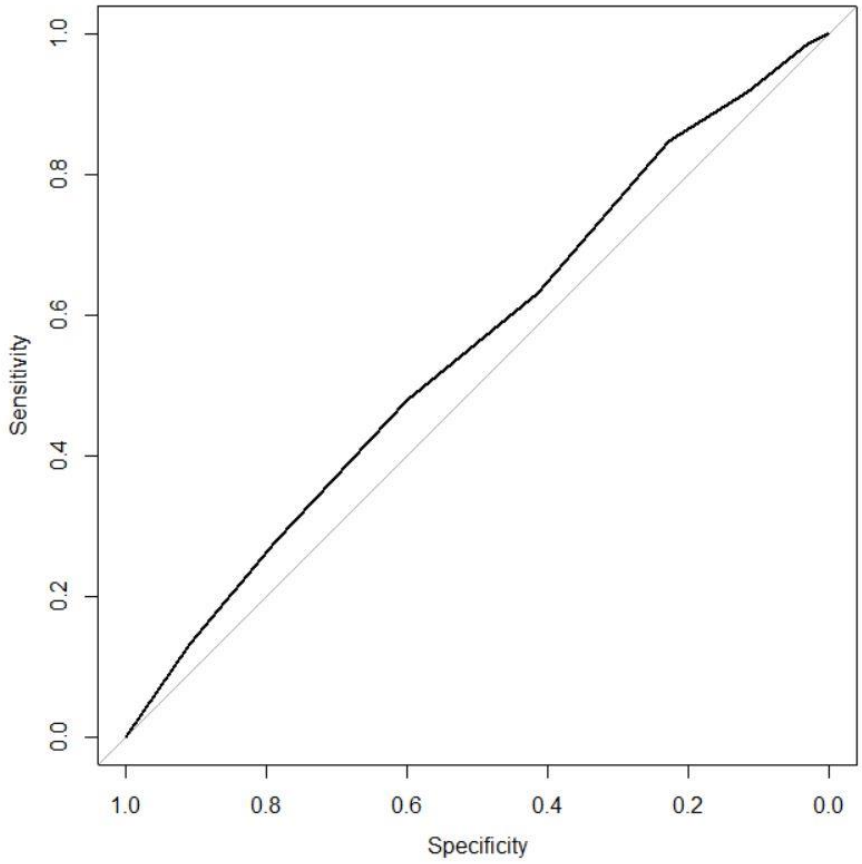
```
data <- UniversityAdmissions
gmodel <- glm(internship_abroad ~ LOR + GRE + TOEFL ,
              data, family = "binomial")
summary(gmodel)
```

```
Call:
glm(formula = internship_abroad ~ LOR + GRE + TOEFL, family = "binomial",
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0450  -0.8609  -0.7722   1.4067   1.8288

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.12492    3.23982  -1.891   0.0587 .
LOR          0.10056    0.13031   0.772   0.4403
GRE          0.01007    0.01632   0.617   0.5369
TOEFL        0.01581    0.03036   0.521   0.6025
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 591.38  on 486  degrees of freedom
Residual deviance: 584.72  on 483  degrees of freedom
AIC: 592.72

Number of Fisher Scoring iterations: 4
```

For each 1 unit increase in any of the variables, the odds ratio of internship abroad is equal to $e^{estimate}$, in which the estimate is the estimated coefficient for any of the variables in a logistic regression. Also , we can say the log odds ratio for an additional unit in each numerical variable is its respective slope, and for the categorical variables the slop can be interpreted as the log odds ratio for being in the category versus not being in the category. Also, the intercept shoes the log odds of response variable is -6.12492, when all explanatory variables are set to 0.

- B) We choose LOR as the categorical variable. The plot shows how good a model is in capturing the response variable.

```
data <- UniversityAdmissions
roc <- roc(data$internship_abroad, data$LOR)
plot.roc(roc)
```

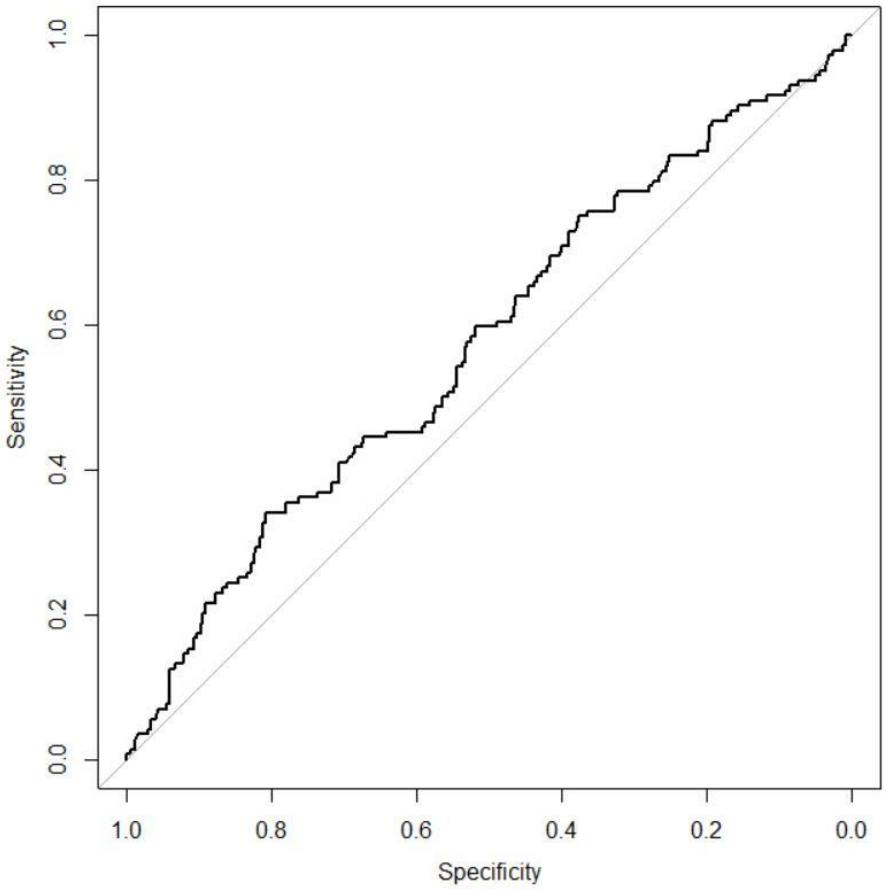


The ROC shows how good a variable is in predicting the response variable, the more curved it is, the more accurately is the variable related to the response variable and is a better predictor. Since the curve is close to a straight line, LOR isn't a good predictor for internship abroad response.

C) We write the R code as follows:

```
roc <- roc(data$internship_abroad, gmodel$fitted)
```

```
plot.roc(roc)

auc <- roc$auc

auc
```

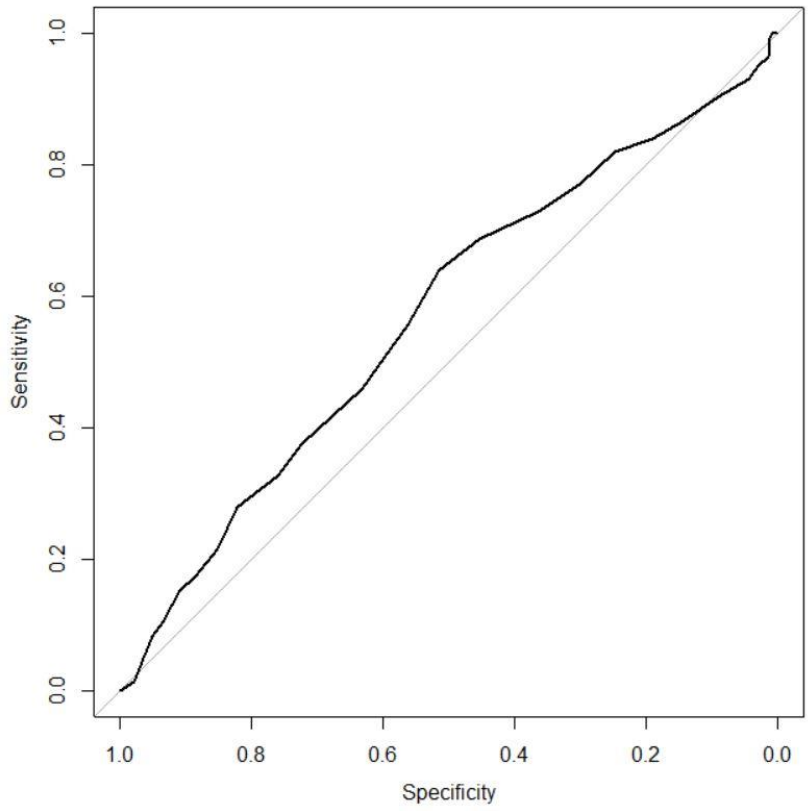



```
Area under the curve: 0.5734
```

Since AUC shows the predictive ability of the model, and is around 0.58, it can be said the model doesn't have a good performance for predicting the response variable (internship abroad).

- D) We plot the ROC for TOEFL and GRE , which TOEFL gives us the most curvaceous plot with highest AUC , so TOEFL is the best predictor, although all variables are bad at predicting the response variable of internship abroad:

```
data <- UniversityAdmissions
roc <- roc(data$internship_abroad, data$TOEFL)
plot.roc(roc)
```

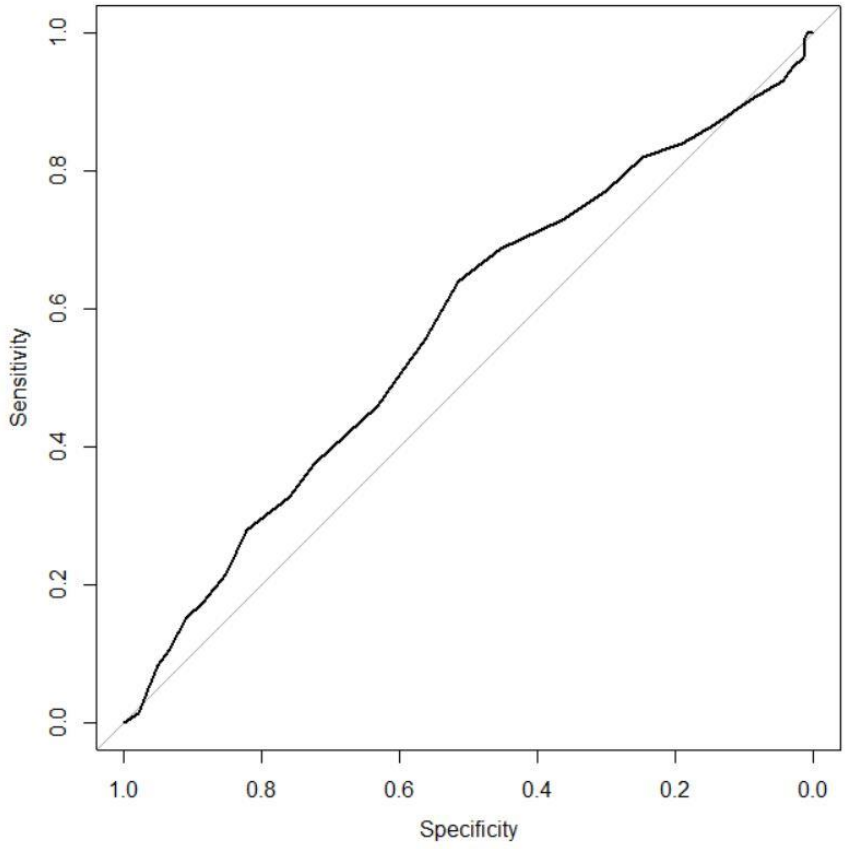


- E) We use TOEFL only for the next model, so we have:

```
data <- UniversityAdmissions
```

```
gmodel <- glm(internship_abroad ~  TOEFL ,
              data, family = "binomial")
summary(gmodel)
roc <- roc(data$internship_abroad, gmodel$fitted)
plot.roc(roc)
auc <- roc$auc
auc
```

```
Call:
glm(formula = internship_abroad ~ TOEFL, family = "binomial",
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0187  -0.8696  -0.7740   1.4234   1.7849

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.09016    1.81861  -2.799  0.00513 **
TOEFL        0.03921    0.01682   2.331  0.01975 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 591.38  on 486  degrees of freedom
Residual deviance: 585.89  on 485  degrees of freedom
AIC: 589.89

Number of Fisher Scoring iterations: 4

> roc <- roc(data$internship_abroad, gmodel$fitted)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
> plot.roc(roc)
> auc <- roc$auc
> auc
Area under the curve: 0.571
```



We used TOEFL this time, but because all the variables were not so significant in in the original model, the result didn't get any better and the ROC is still close to the straight line.

Question 7:

- We create a linear regression model using the R code below:
-

```
SOP <- UniversityAdmissions$SOP
TOEFL <-UniversityAdmissions$TOEFL
CGPA <-UniversityAdmissions$CGPA
LOR <-UniversityAdmissions$LOR
GRE <-UniversityAdmissions$GRE
Research <-UniversityAdmissions$Research
```

```
internship_abroad <-UniversityAdmissions$internship_abroad
Chance_of_Admit <-UniversityAdmissions$Chance_of_Admit
modelx <- lm( Chance_of_Admit ~ SOP + TOEFL +  CGPA + LOR + GRE +
Research + internship_abroad , UniversityAdmissions )
summary(modelx)
```

```
Call:
lm(formula = Chance_of_Admit ~ SOP + TOEFL + CGPA + LOR + GRE +
    Research + internship_abroad, data = UniversityAdmissions)

Residuals:
      Min        1Q    Median        3Q       Max
-0.264237 -0.030189  0.008318  0.040483  0.167079

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.2768391  0.1215643 -10.503  < 2e-16 ***
SOP1.5            -0.0310304  0.0308489  -1.006    0.315
SOP2              -0.0345145  0.0294945  -1.170    0.243
SOP2.5            -0.0107489  0.0290500  -0.370    0.712
SOP3              -0.0136880  0.0289251  -0.473    0.636
SOP3.5            -0.0186736  0.0292879  -0.638    0.524
SOP4              -0.0069873  0.0300961  -0.232    0.817
SOP4.5             0.0097364  0.0311119   0.313    0.754
SOP5               0.0272100  0.0322240   0.844    0.399
SOPC               0.0062958  0.0717925   0.088    0.930
TOEFL              0.0053024  0.0009449   5.612 3.42e-08 ***
CGPA               0.0548130  0.0083407   6.572 1.32e-10 ***
LOR                0.0258236  0.0044834   5.760 1.52e-08 ***
GRE                0.0035261  0.0005289   6.666 7.35e-11 ***
Research           0.0299440  0.0073869   4.054 5.90e-05 ***
internship_abroad  0.0002733  0.0066300   0.041    0.967
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06571 on 471 degrees of freedom
Multiple R-squared:  0.7773,    Adjusted R-squared:  0.7702
F-statistic: 109.6 on 15 and 471 DF,  p-value: < 2.2e-16
```

In our model, the variables with the least p-values have the biggest effect on chance of admission, which is the GRE variable, other significant variables, from most significance to least are shown below:

GRE > CGPA > LOR > TOEFL > Research

Others are not significant and can be omitted from the model.