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")



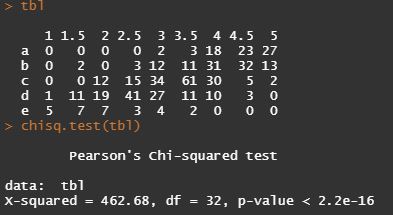
We are 95% sure that the compared proportions have a difference in means between -0.01707137 & 0.08277979

* 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)



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")

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



The variable “SOP” is selected which contains 9 levels. A random sample of and a biased sample of with sizes of 100 are selected from the data. 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)

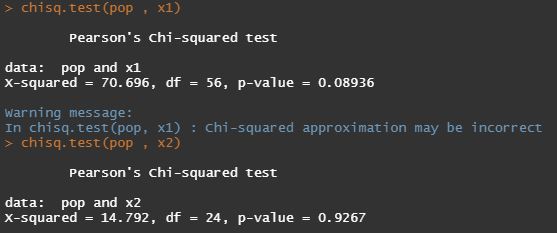
x1

x2 <- sample(data[which(data < "2.5")] , size =100)

x2

chisq.test(pop , x1)

chisq.test(pop , x2)



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

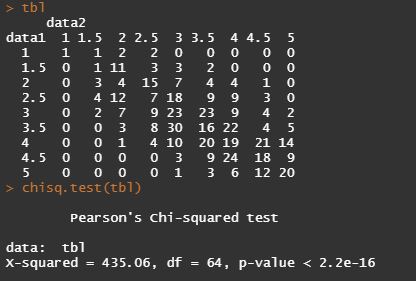
data2 <- UniversityAdmissions$LOR

tbl <- table(data1 , data2)

tbl

chisq.test(tbl)

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



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)

1. 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

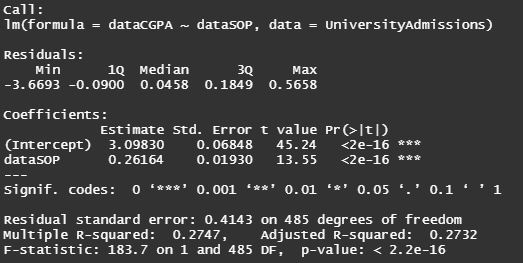
dataCGPA <-UniversityAdmissions$CGPA

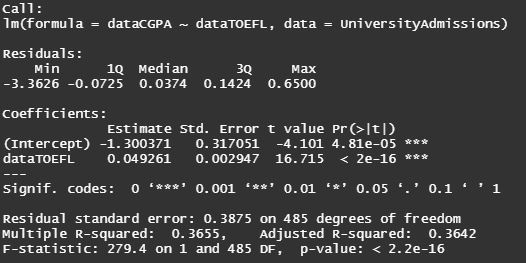
model1 <- lm(dataCGPA ~ dataSOP , UniversityAdmissions )

model2 <- lm(dataCGPA ~ dataTOEFL , UniversityAdmissions )

summary(model1)

summary(model2)





Equation of the models can be written as:

One unit increase in SOP will increase the CGPA by units. Also, an interpretation for the intercept parameter is that an admission with SOP as 0 would have 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 . Also, an interpretation for the intercept parameter is that an admission with TOEFL as 0 would have 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.

1. 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") +

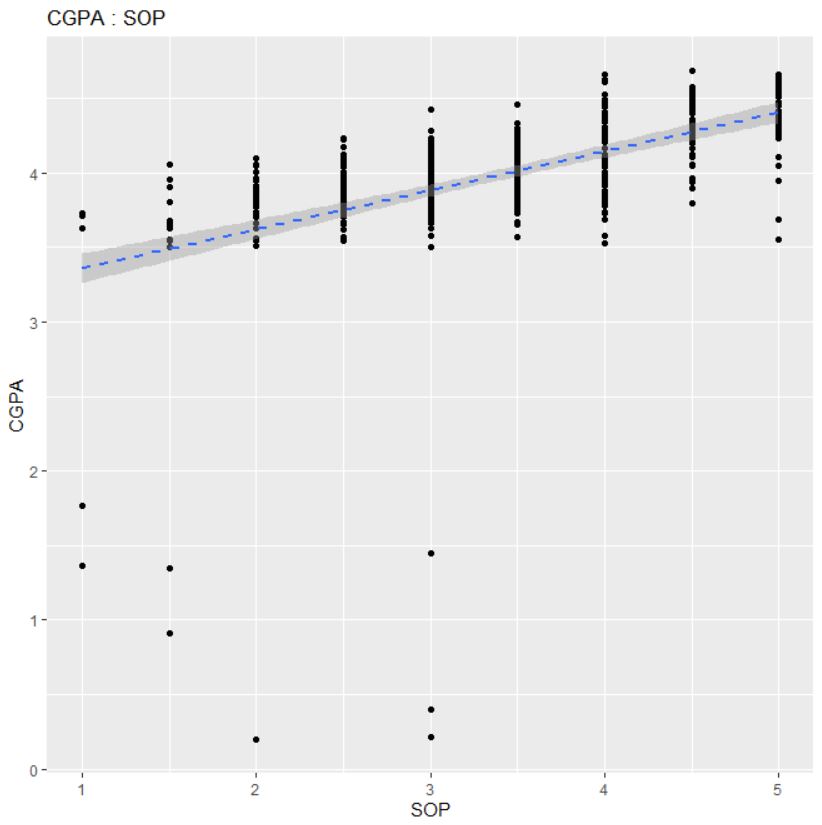
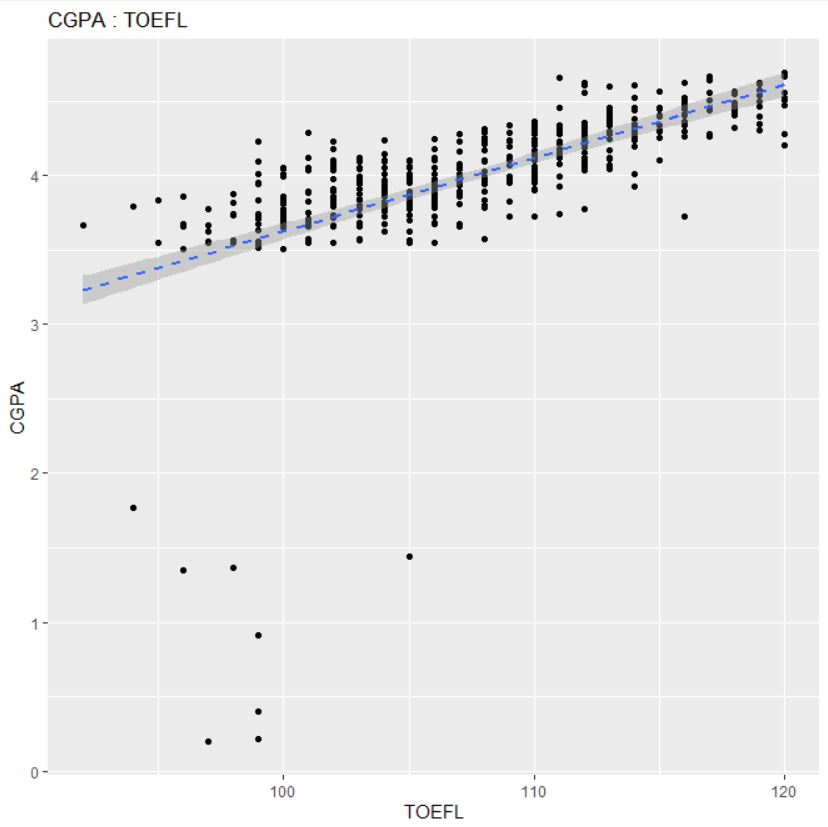
ggtitle("CGPA : SOP")

ggplot(data, aes(TOEFL, CGPA)) +

geom\_point() +

geom\_smooth(method = "lm", linetype = "dashed") +

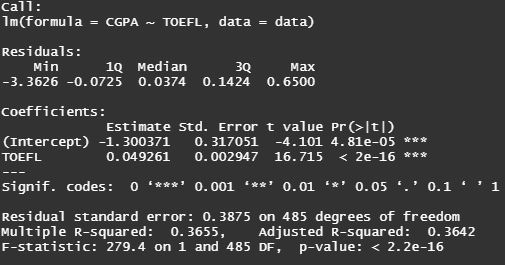
ggtitle("CGPA : TOEFL")

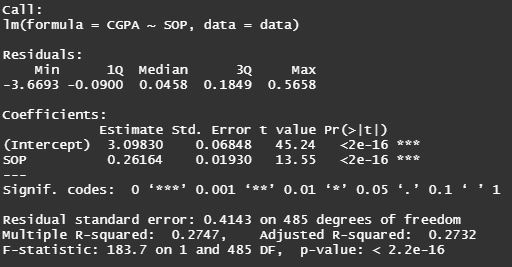
 

* C) Using the adjusted r-squared and also by taking a look at the plots above , we conclude that TOEFL is a better explanatory variable for CGPA as a response variable, and TOEFL is more significant. TOEFL has higher adjusted r-squared value, and also has a steeper slope in the linear regression plot.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

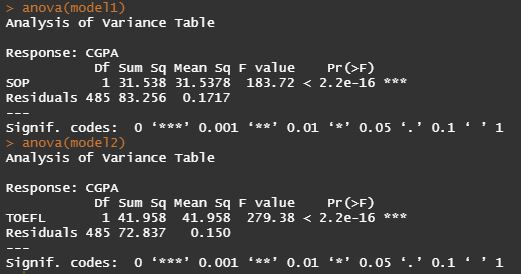
anova(model1)

anova(model2)





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



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:

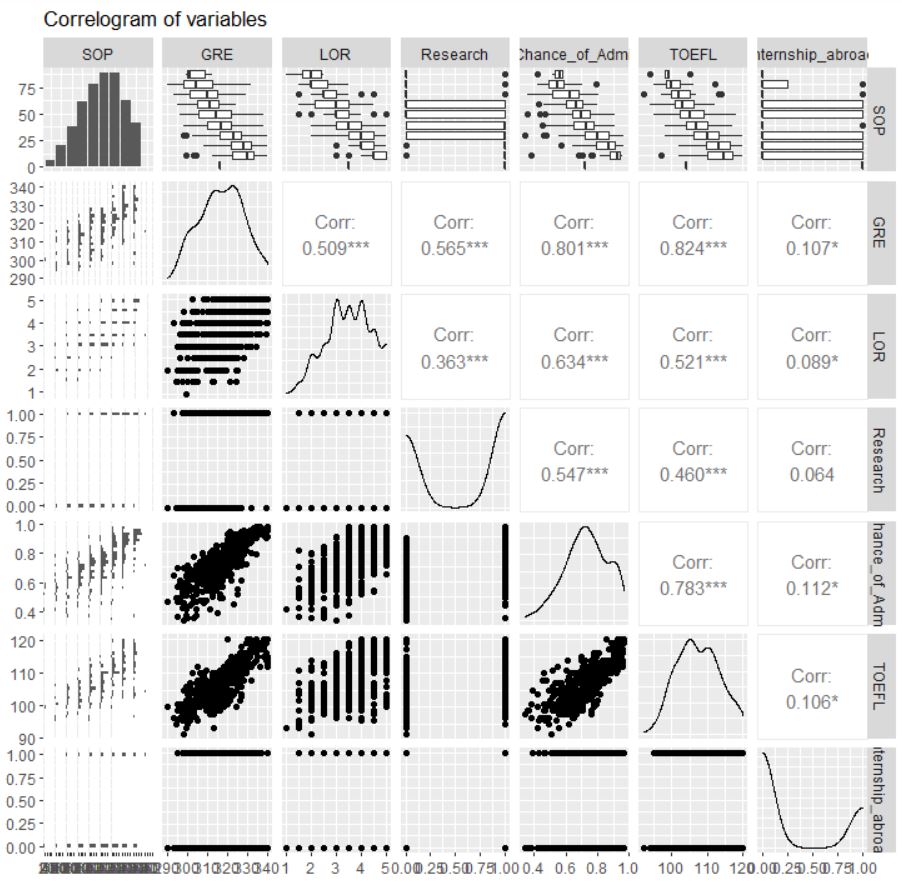
* A) R code to plot correlogram:

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:

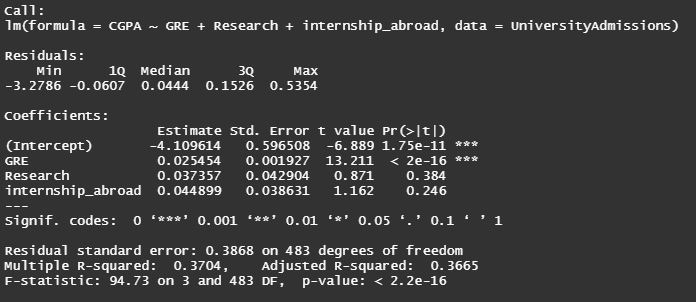
GRE <-UniversityAdmissions$GRE

Research <-UniversityAdmissions$Research

internship\_abroad <-UniversityAdmissions$internship\_abroad

modelg <- lm( CGPA ~ GRE + Research + internship\_abroad , UniversityAdmissions )

summary(modelg)



* 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 +

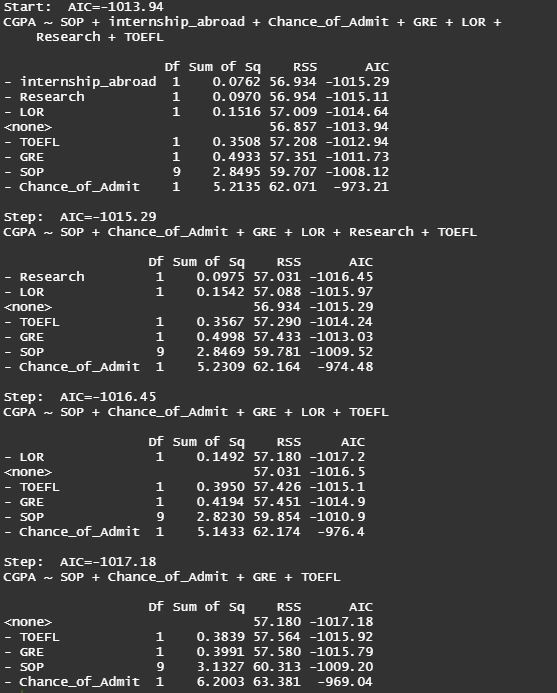
GRE + LOR + Research + TOEFL, data)

modelnull <- lm(CGPA ~ 1, data)

First, backward elimination:

cat("\n\n\*\*\* Backward elimination \*\*\*\n\n")

bestbw <- step(modelfull, direction = "backward")

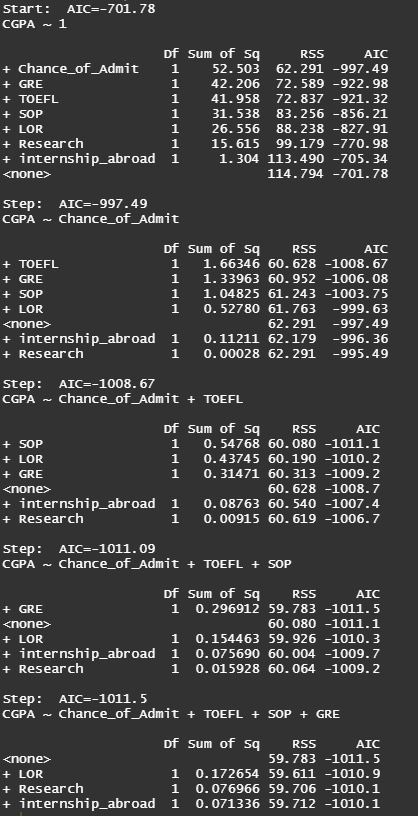


And the code for forward selection:

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) )



Which resulted the same final model with backward elimination.

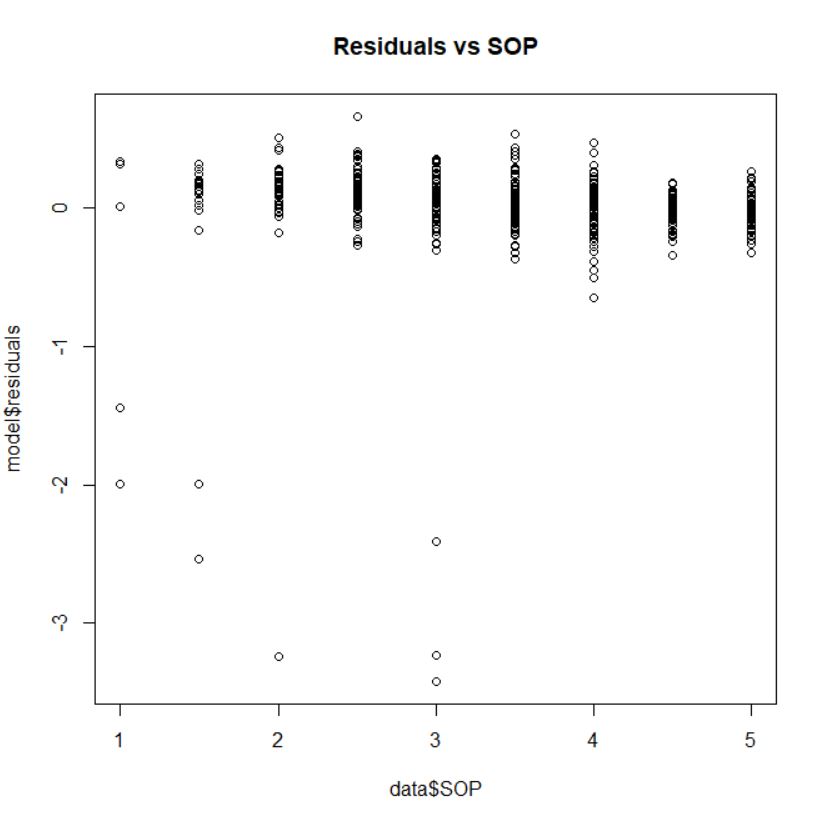
* F) The three conditions are checked with a number of residual plots.

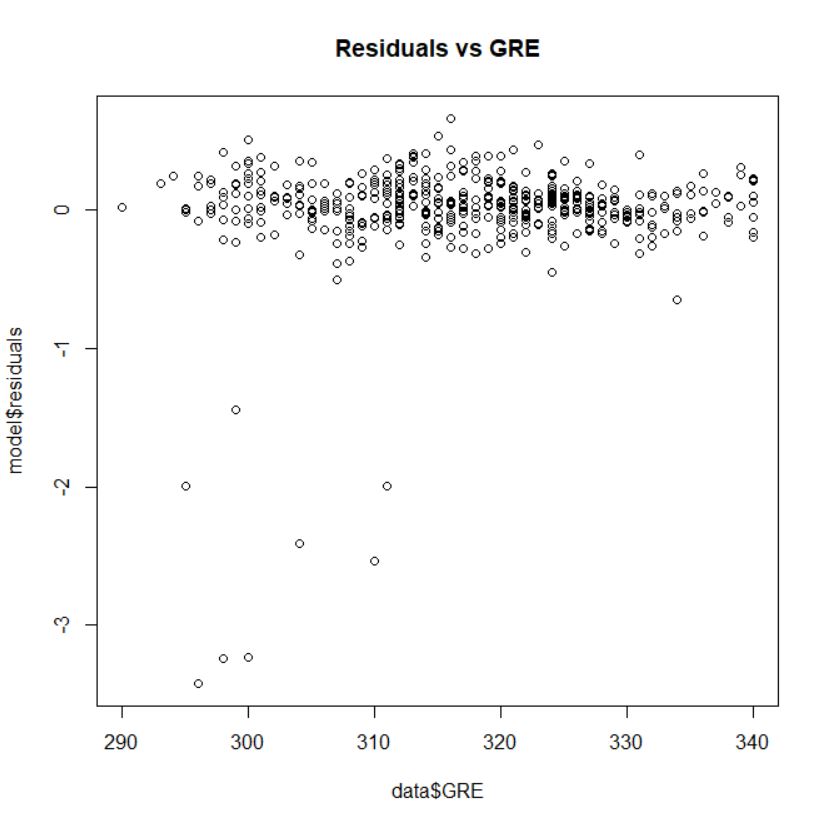
plot(model$residuals ~ data$SOP, main = "Residuals vs SOP")

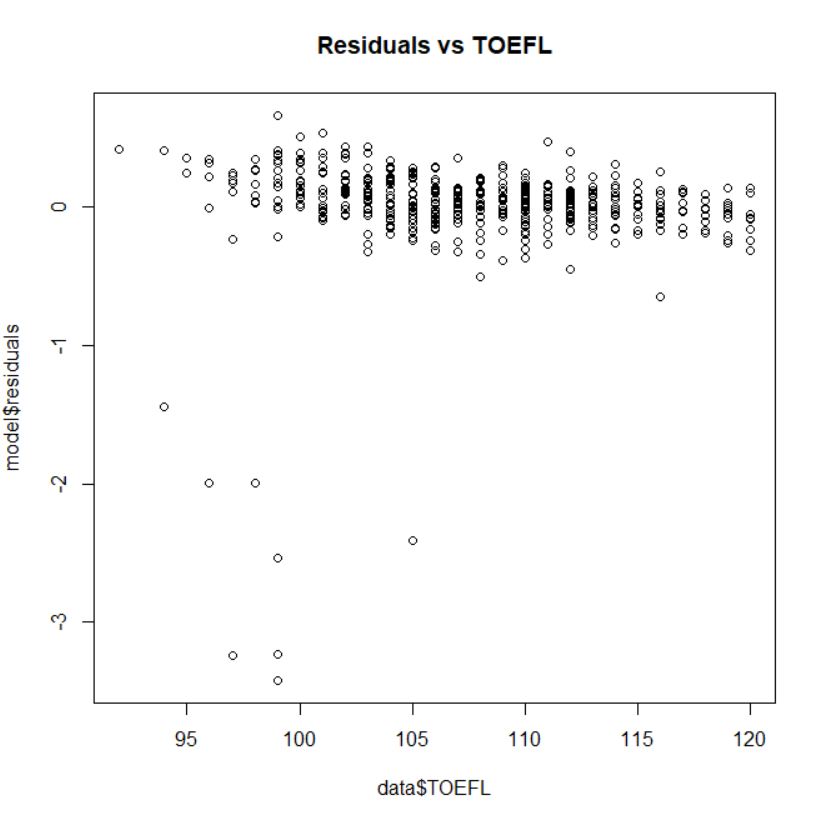
plot(model$residuals ~ data$GRE, main = "Residuals vs GRE")

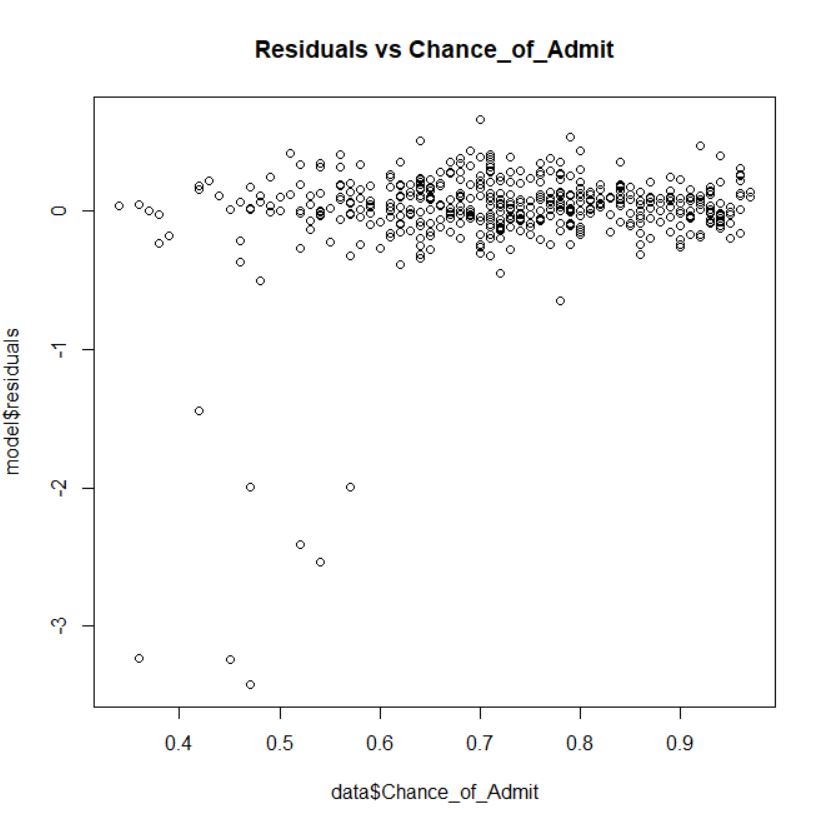
plot(model$residuals ~ data$TOEFL, main = "Residuals vs TOEFL")

plot(model$residuals ~ data$Chance\_of\_Admit, main = "Residuals vs Chance\_of\_Admit")









Question 6:

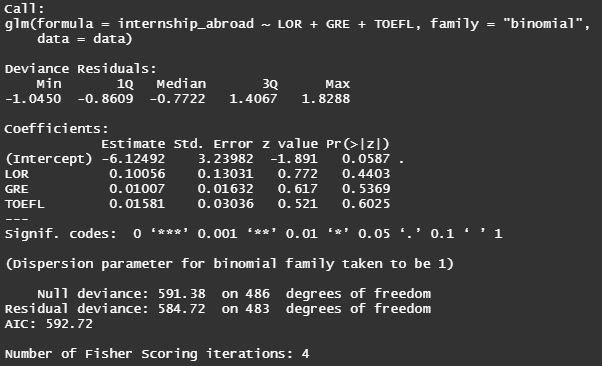
* A) A logistic regression model for the response variable of internship abroad is created using the explanatory variables of LOR, GRE & TOEFL.

data <- UniversityAdmissions

gmodel <- glm(internship\_abroad ~ LOR + GRE + TOEFL ,

data, family = "binomial")

summary(gmodel)



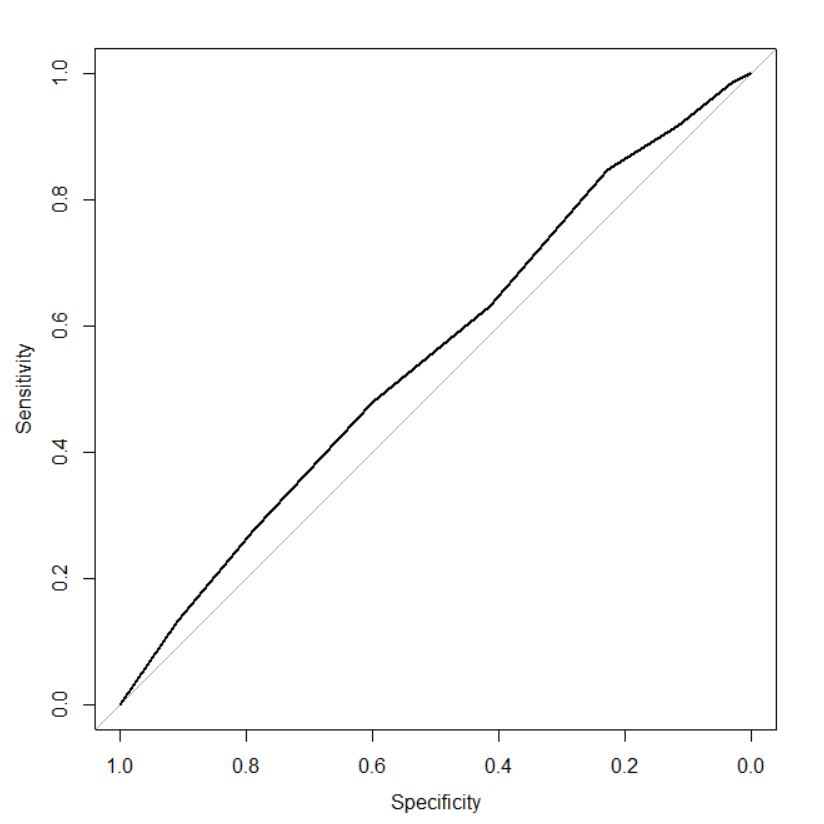
For each 1 unit increase in any of the variables, the odds ratio of internship abroad is equal to , in which the estimate is the estimated coefficient for any of the variables in a logistic regression. Also , we can say the log odds ratio for an additional unit in each numerical variable is its respective 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.

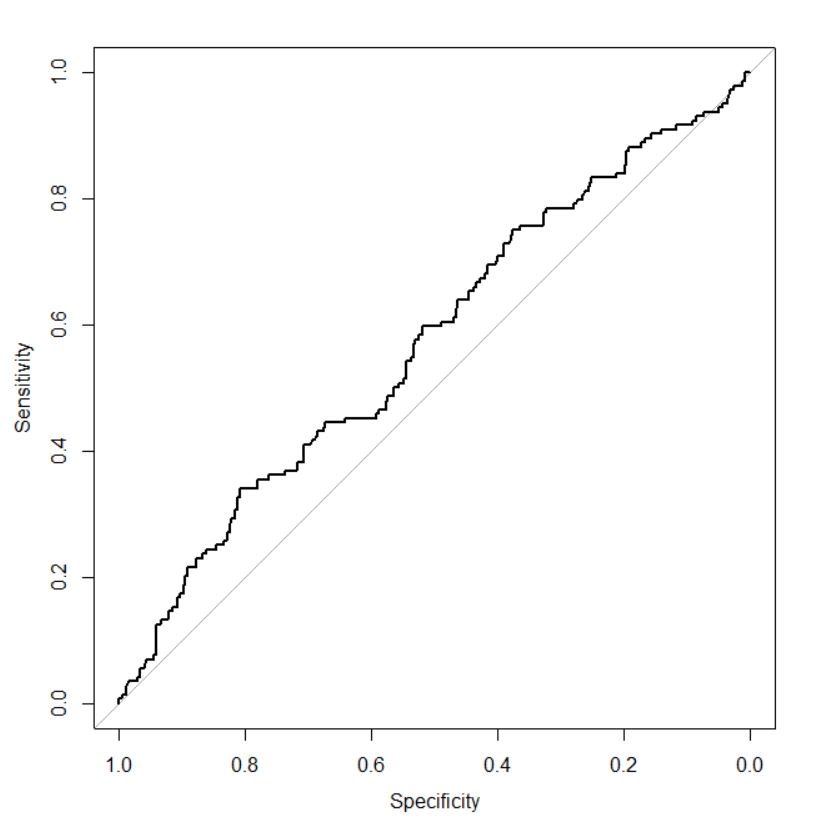
C) We write the R code as follows:

roc <- roc(data$internship\_abroad, gmodel$fitted)

plot.roc(roc)

auc <- roc$auc

auc





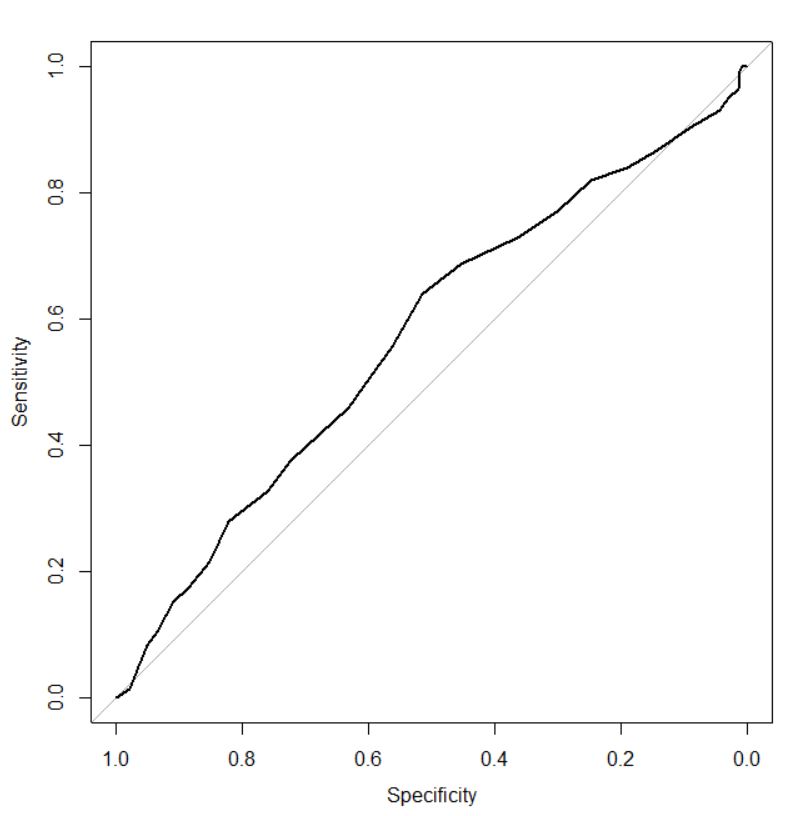
Since AUC shows the predictive ability of the model, and is around 0.58, it can be said the model doesn’t have a good performance for predicting the response variable (internship abroad).

* 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")

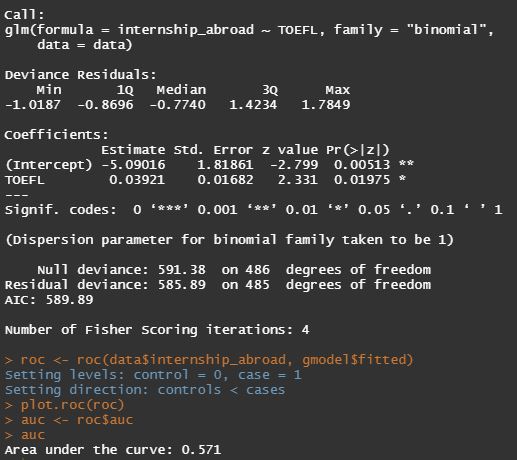
summary(gmodel)

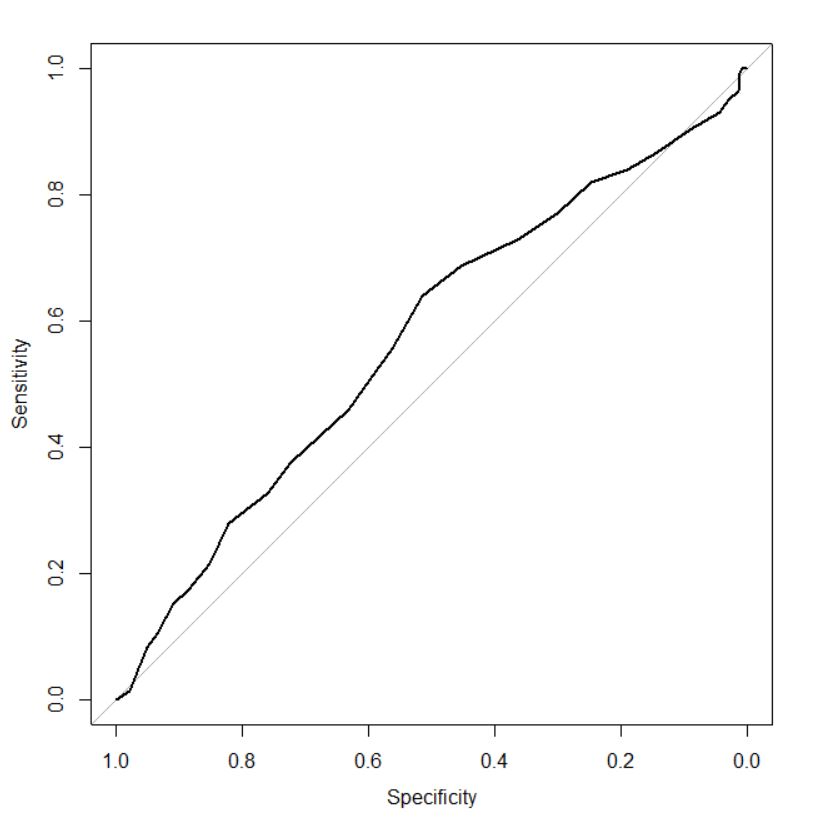
roc <- roc(data$internship\_abroad, gmodel$fitted)

plot.roc(roc)

auc <- roc$auc

auc





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

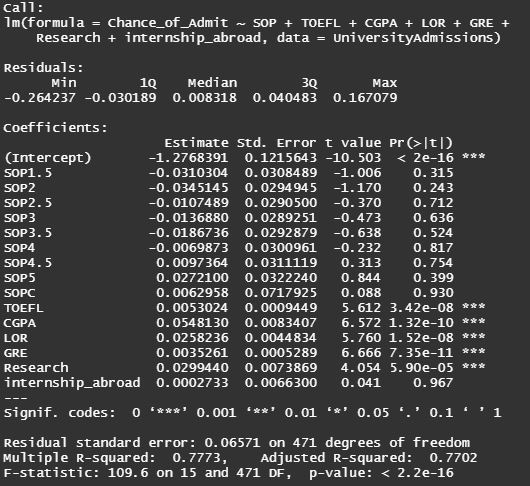
Research <-UniversityAdmissions$Research

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)



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.