Lab 7

Enter the names of the group members here: Austine Do, Graceanne Becker, Catherine Zhong

This assignment is due by the end of the lab. Only one student in the group submits a pdf file on Gradescope.

For all questions, include the R commands/functions that you used to find your answer (show R chunk). Answers without supporting code will not receive credit. Write full sentences to describe your findings.

In this lab, you will explore a dataset that contains information about the chances of admissions into graduate school for international students. Let's first load the `tidyverse` package:

```
library(tidyverse)
```

Let's upload the data from Github, do a little bit of cleaning, and take a quick look:

```
# Upload data from GitHub
admissions <- read_csv("https://raw.githubusercontent.com/laylaguyot/datasets/main//Admission_Predict.cs
  # Rename variables for easier manipulation
  rename(Serial_No = `Serial No.`,
         GRE_Score = `GRE Score`,
         TOEFL_Score = `TOEFL Score`,
         University_Rating = `University Rating`,
         Admission_Chance = `Chance of Admit`)

# Take a quick look
head(admissions)
```

```
## # A tibble: 6 x 9
##   Serial_No GRE_Score TOEFL_Score University_Rating   SOP   LOR  CGPA Research
##       <dbl>     <dbl>       <dbl>             <dbl> <dbl> <dbl> <dbl>    <dbl>
## 1         1       337         118                 4   4.5   4.5  9.65        1
## 2         2       324         107                 4   4     4.5  8.87        1
## 3         3       316         104                 3   3     3.5  8           1
## 4         4       322         110                 3   3.5   2.5  8.67        1
## 5         5       314         103                 2   2     3    8.21        0
## 6         6       330         115                 5   4.5   3    9.34        1
## # i 1 more variable: Admission_Chance <dbl>
```

This dataset contains the following variables: GRE Scores (out of 340), TOEFL Scores (out of 120), University Rating (out of 5), the strength of the Statement of Purpose `SOP` and Letter of Recommendation `LOR` (out of 5), Undergraduate GPA (out of 10), Research Experience (either yes = 1 or no = 0), and the Admission chance (ranging from 0 to 1).

The goal of the lab is to make predictions for graduate school admission based on other features of a student's application.

Question 1: (3 pts)

Which variable in the `admissions` dataset should be considered as the `outcome` variable?

The Admission_Chance variable should be considered the outcome variable.

Which variable in the `admissions` dataset should we NOT use to predict the `Admission_Chance`? Why?

Serial_No should most definitely not be used in the model to predict Admission_Chance since it is just an ID of the row

---

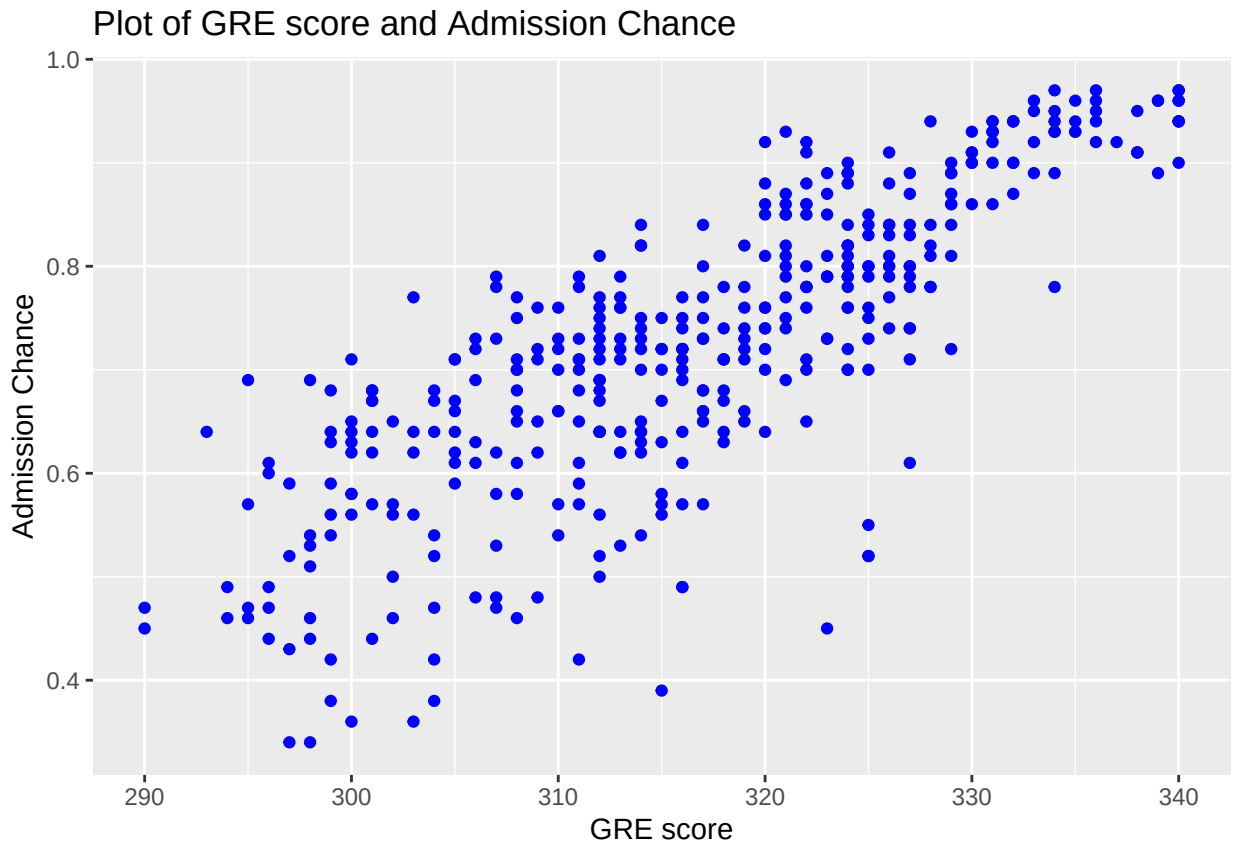
Question 2: (8 pts)

Pick one variable to predict a student's admission chance into graduate school. What potential relationship between this variable and the outcome variable do you anticipate? Answer that question before exploring the data!

The variable we pick to predict admission chance into graduate school is GRE_Score. We expect a higher GRE score to predict a higher chance of admission and vice versa.

Visualize the relationship between the predictor you chose and the outcome variable. Does your visualization match the relationship that you had anticipated?

```
# Roughly visualizing the plot of GRE_Score vs Admission_Chance
admissions |>
    ggplot(aes(x = GRE_Score, y = Admission_Chance)) +
        geom_point(color = 'blue') +
        labs(title = 'Plot of GRE score and Admission Chance',
             x = 'GRE score',
             y = 'Admission Chance')
```

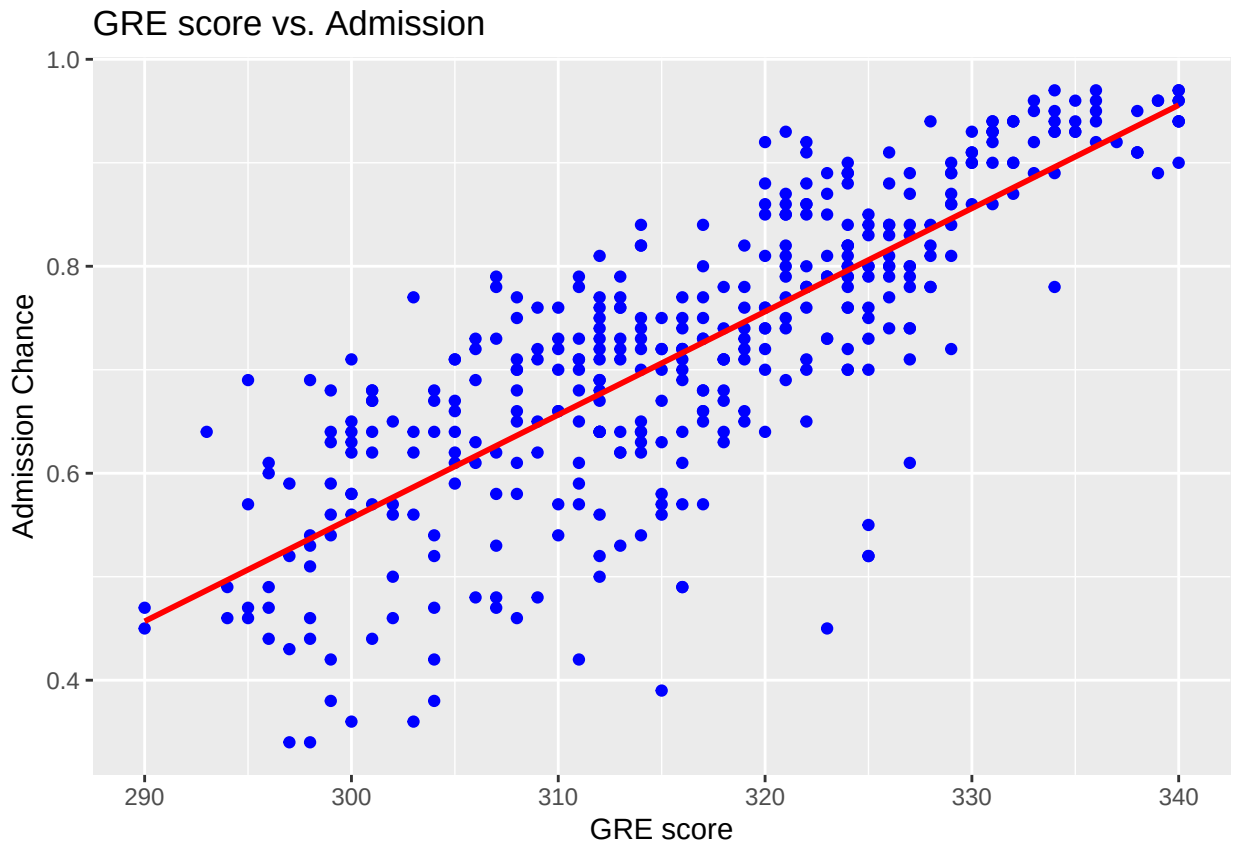


The visualization does roughly match our expectation.

Fit a linear regression model to predict the outcome based on the predictor you chose. Write the expression of the linear model.

```
# Plot of GRE_Score and Admission_Chance with a linear regression model plotted over the data
admissions |>
    ggplot(aes(x = GRE_Score, y = Admission_Chance)) +
        geom_point(color = 'blue') +
        geom_smooth(method = "lm", se = FALSE, color = "red", size = 1) +
        labs(title = 'GRE score vs. Admission',
             x = 'GRE score',
             y = 'Admission Chance')
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```




```
lin_model <- lm(Admission_Chance ~ GRE_Score, data = admissions)
summary(lin_model)
```

```
##
## Call:
## lm(formula = Admission_Chance ~ GRE_Score, data = admissions)
##
## 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
```

Expression of our linear model: Admission_Chance = 0.0099759 * GRE_Score - 2.4360842

Find predicted values for the lowest and the highest possible values of your predictor (for example, CGPA varies from 0 to 10 in theory). Do the predicted values make sense in context? Why/Why not?

```r
# Lowest and highest possible GRE score model prediction
lowest_value_GRE <- data.frame(GRE_Score = 260)
predict(lin_model, newdata = lowest_value_GRE)
```

```
##         1
## 0.1576451
```

```r
highest_value_GRE <- data.frame(GRE_Score = 340)
predict(lin_model, newdata = highest_value_GRE)
```

```
##         1
## 0.9557156
```

Yes these predicted values do make sense because lower GRE score means you might have a tougher chance of getting into grad school while a higher GRE increase your chances of getting into grad school.

Evaluate the performance of the model with two appropriate measures. *Note: no need to comment on the values for now.*

```r
# RSME of linear model with GRE score predictor
sqrt(mean(resid(lin_model)^2))
```

```
## [1] 0.08496057
```

```r
# Adjusted R-squared of linear model with GRE score predictor
summary(lin_model)$adj.r.squared
```

```
## [1] 0.6432895
```

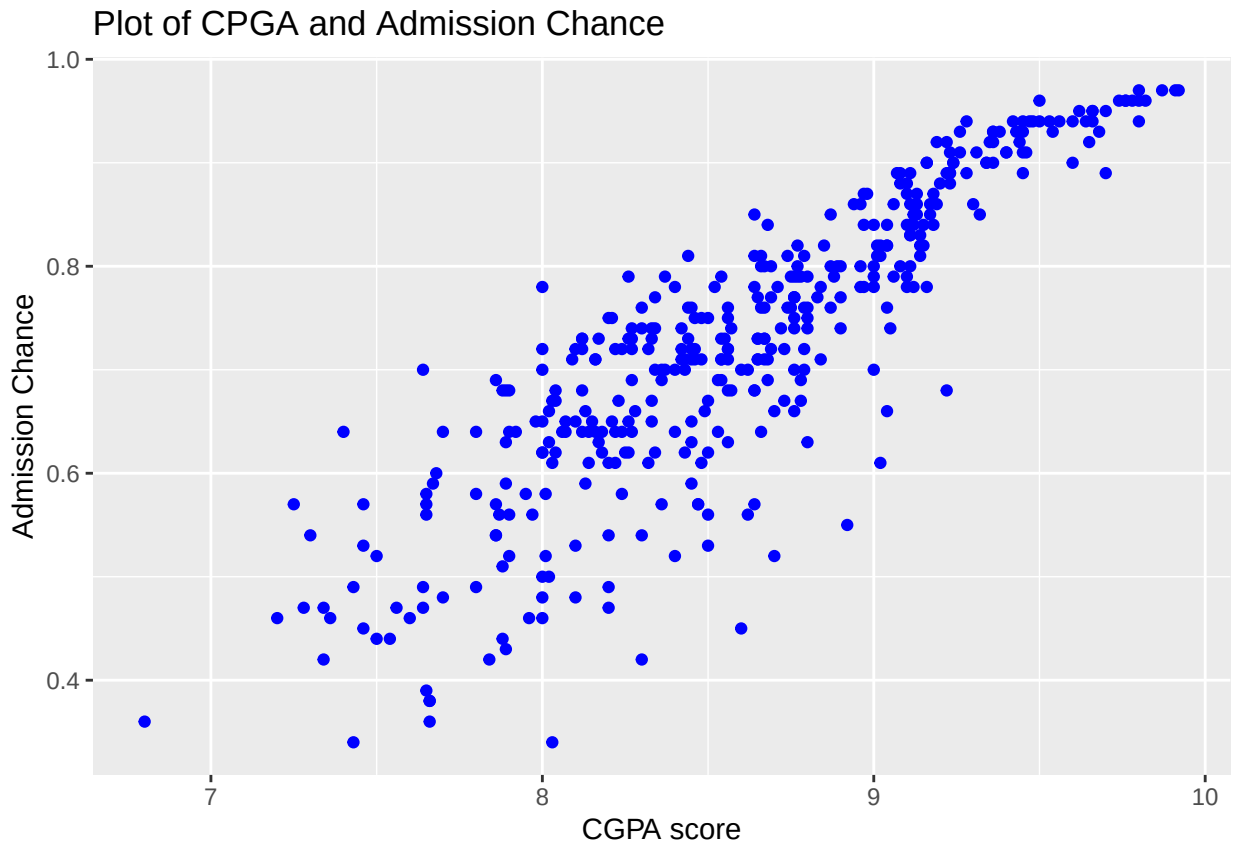
Question 3: (8 pts)

Pick another variable to predict a student's admission chance into graduate school. What potential relationship between this variable and the outcome variable do you anticipate? Answer that question before exploring the data!

The other variable we picked to predict the admission chance is CGPA. We believe that a higher CGPA will predict a higher chance of admission.

Visualize the relationship between the predictor you chose and the outcome variable. Does your visualization match the relationship that you had anticipated?

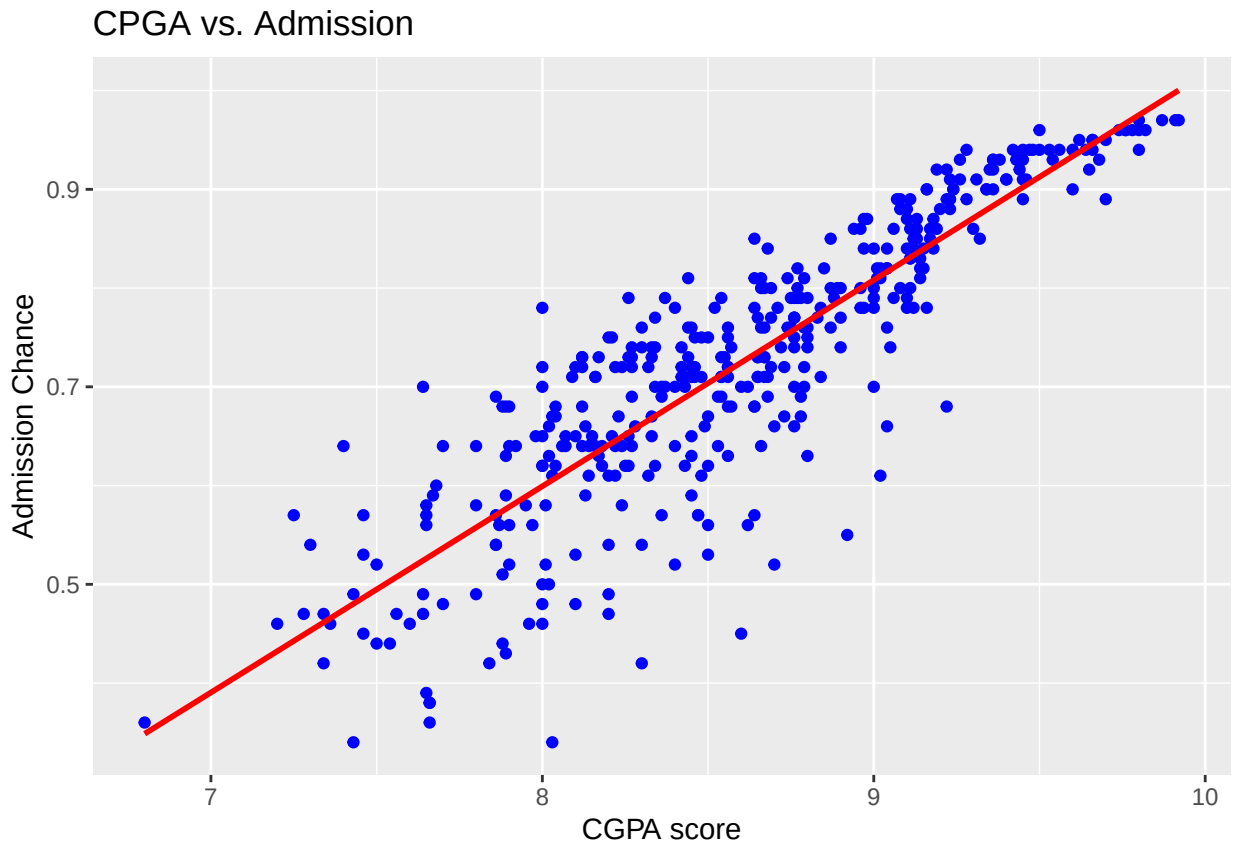
```r
# Roughly visualizing the plot of GRE_Score vs Admission_Chance
admissions |>
    ggplot(aes(x = CGPA, y = Admission_Chance)) +
        geom_point(color = 'blue') +
        labs(title = 'Plot of CPGA and Admission Chance',
             x = 'CGPA score',
             y = 'Admission Chance')
```



The visualization of CGPA score and admission chance did roughly match our expectations.

Fit a linear regression model to predict the outcome based on the predictor you chose. Write the expression of the linear model.

```
# Plot of GRE_Score and Admission_Chance with a linear regression model plotted over the data
admissions |>
    ggplot(aes(x = CGPA, y = Admission_Chance)) +
        geom_point(color = 'blue') +
        geom_smooth(method = "lm", se = FALSE, color = "red", size = 1) +
        labs(title = 'CPGA vs. Admission',
             x = 'CGPA score',
             y = 'Admission Chance')
```




```
lin_model_2 <- lm(Admission_Chance ~ CGPA, data = admissions)
summary(lin_model_2)
```

```
##
## Call:
## lm(formula = Admission_Chance ~ CGPA, data = admissions)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.274575 -0.030084  0.009443  0.041954  0.180734
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.07151    0.05034  -21.29   <2e-16 ***
## CGPA         0.20885    0.00584   35.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06957 on 398 degrees of freedom
## Multiple R-squared:  0.7626, Adjusted R-squared:  0.762
## F-statistic:  1279 on 1 and 398 DF,  p-value: < 2.2e-16
```

Expression of our linear model: Admission_Chance = 0.20885 * CGPA - 1.07151

Find predicted values for the lowest and the highest possible values of your predictor (for example, CGPA varies from 0 to 10 in theory). Do the predicted values make sense in context? Why/Why not?

```r
# Lowest and highest possible CGPA model prediction
lowest_value_CGPA <- data.frame(CGPA = 0)
predict(lin_model_2, newdata = lowest_value_CGPA)
```

```
##         1
## -1.071512
```

```r
highest_value_CGPA <- data.frame(CGPA = 10)
predict(lin_model_2, newdata = highest_value_CGPA)
```

```
##        1
## 1.016961
```

No the predicted values do not make sense in the context because admissions chance cannot be a negative value and cannot be greater than 100%.

Evaluate the performance of this second model with two appropriate measures. How does this second model compare to the first model?

```r
# RSME of linear model with CGPA predictor
sqrt(mean(resid(lin_model_2)^2))
```

```
## [1] 0.0693927
```

```r
# Adjusted R-squared of linear model with CGPA predictor
summary(lin_model_2)$adj.r.squared
```

```
## [1] 0.7620375
```

The model with the CGPA predictor variable performed better across most Root Mean Squared Error and Adjusted R-squared metrics of linear model performance compared to the linear model with GRE score as the predictor.

Question 4: (4 pts)

Let's consider a linear regression model with all the potential predictors (using `~ .`). Which predictors do not seem to be so useful to predict the admission to graduate school?

```r
#
lin_model_all_vars <- lm(Admission_Chance ~ ., data = admissions)
summary(lin_model_all_vars)
```

```
##
## Call:
## lm(formula = Admission_Chance ~ ., data = admissions)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
```

```
## -0.233576 -0.026637  0.006226  0.038273  0.140252
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.294e+00  1.201e-01 -10.775  < 2e-16 ***
## Serial_No          1.593e-04  2.769e-05   5.753 1.77e-08 ***
## GRE_Score          1.799e-03  5.749e-04   3.129 0.001885 **
## TOEFL_Score        3.682e-03  1.056e-03   3.487 0.000543 ***
## University_Rating  8.785e-03  4.617e-03   1.903 0.057821 .
## SOP                9.937e-05  5.380e-03   0.018 0.985272
## LOR                2.154e-02  5.330e-03   4.041 6.41e-05 ***
## CGPA               1.053e-01  1.198e-02   8.786  < 2e-16 ***
## Research           2.438e-02  7.653e-03   3.185 0.001561 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06132 on 391 degrees of freedom
## Multiple R-squared:  0.8188, Adjusted R-squared:  0.8151
## F-statistic: 220.9 on 8 and 391 DF,  p-value: < 2.2e-16
```

The variables that do not seem to be significant in predicting admissions chance are University_Rating and SOP (Statement of purpose). It is also worth noting that while it may seem like Serial_No is a significant variable in predicting admission chance it does not make any sense in the context of the data.

Evaluate the performance of this full model (containing all the potential predictors) with two appropriate measures. How does this full model compare to the models with a single predictor?

```
# RSME of linear model of all variables
sqrt(mean(resid(lin_model_all_vars)^2))
```

```
## [1] 0.06062765
```

```
# Adjusted R-squared of linear model of all variables
summary(lin_model_all_vars)$adj.r.squared
```

```
## [1] 0.8151034
```

The linear model containing all potential predictor variables performs better across both Root Square Mean Error and Adjusted R-squared metrics when compared to the linear models containing only a single predictor variable.

Question 5: (1 pt)

After investigating what characteristics of an application seem to affect admission into graduate school for some international students, did the data match your expectations or not? If the data differed from your expectation, provide a possible explanation for why the data differed from what you expected.

After conducting multiple linear model regressions, we found that GRE score, TOEFL score, letter of recommendation, undergraduate GPA, and research experience greatly affected the admissions chance of international students. This matches our expectation since we expected higher admissions chance for applicants with better stats.

Formatting: (1 pt)

Make sure the names of all group members are included at the beginning of the document.

Knit your file! You can knit into pdf directly or into html. Once it knits in html, click on `Open in Browser` at the top left of the window pops out. Print your html file into pdf from your browser.

Any issue? Ask other classmates or TA!

Finally, remember to select pages for each question when submitting your pdf to Gradescope and to identify your group members.