STA 135 Project 3

Grant Gambetta

2/22/2022

Introduction

For this project, the goal was to perform statistical analysis on a dataset consisting of GPA and GMAT scores of students from three different schools. First, exploratory analysis was conducted to gain initial insights about the dataset. Exploratory analysis was performed using several graphical and numerical techniques such as sample mean vectors, correlation and covariance matrices, scatterplots, boxplots, and histograms. Next, to support the visual inspection from the exploratory analysis, statistical analysis was performed using a one-way MANOVA and 95% simultaneous confidence intervals to determine if there was a statistically significant difference in mean GPA and GMAT scores across the three schools, as well as to determine whether there were differences in components of the treatment effects.

This report consists of four main sections: introduction, materials and methods, results, and appendix. In the materials and methods section, I provide an overview of the dataset and discuss the statistical, numerical, and visualization methods I used to describe and analyze the data. Additionally in this section, I detail the specific formulas behind the statistical calculations. In the results section, I discuss the main findings of the analysis and interpret some of the R outputs that I generated. Within the results section, there are two subsections: exploratory analysis and statistical analysis, which is where the results of the exploratory analysis and statistical analysis are discussed, respectively. Lastly, the appendix is where the entire R code can be found.

Materials and Methods

To begin with, it is important to understand the dataset behind this project. The dataset consisted of three variables and 85 rows, where two of the variables were numeric/continuous and the third was a categorical variable with three levels. The two continuous variables were `GPA` and `GMAT`, which denoted the GPAs and GMAT scores for students in each school, respectively. The categorical variable represented the three schools, denoted as schools 1, 2, and 3. There were 31 observations for school 1 (n_1), 28 observations for school 2 (n_2), and 26 observations for school 3 (n_3).

For the statistical analysis, a one-way MANOVA was conducted to test if mean GPA and GMAT scores were the same across the three schools. The model equation for one-way MANOVA is

$$\underline{X}_{\ell j} = \underline{\mu} + \underline{\tau}_{\ell} + \underline{e}_{\ell j},$$

where $\sum_{\ell=1}^{g} n_{\ell} \underline{\tau}_{\ell} = 0$ for identifiable τ_{ℓ} and $\underline{e}_{\ell j} \overset{iid}{\sim} N_p(0, \Sigma)$.

For hypothesis testing, Wilk's Lambda test statistic was used with Bartlett's large sample result. The formulas for Wilk's Lambda test statistic and Bartlett's large sample result are below.

$$\Lambda^* = \frac{|SSR \text{ (or W)}|}{|SSR \text{ (or W)} + SST \text{ (or B)}|}$$

$$-(n-1-\frac{p+g}{2})\log_e(\Lambda^*) \approx \chi^2_{p(g-1)}$$

Also, 95% simultaneous confidence intervals were constructed for differences in components of the treatment effects $(\tau_{ki} - \tau_{\ell i})$. The formula is

$$\overline{X}_{ki} - \overline{X}_{\ell i} \pm t_{df}\left(\frac{\alpha}{pg(g-1)}\right)\sqrt{\left(\frac{1}{n_k}+\frac{1}{n_\ell}\right)\frac{SSR_{ii}}{df}}$$

where $df = \sum_{\ell=1}^{g} n_\ell - p$.

Results

Exploratory Analysis

To begin with, exploratory analysis was performed to gain an initial understanding of the dataset and preliminary insights about GPA and GMAT scores for each population (schools 1, 2, and 3). I began by computing summary statistics such as the mean vectors, variance-covariance matrices, and correlation matrices for each population, which can be found below.

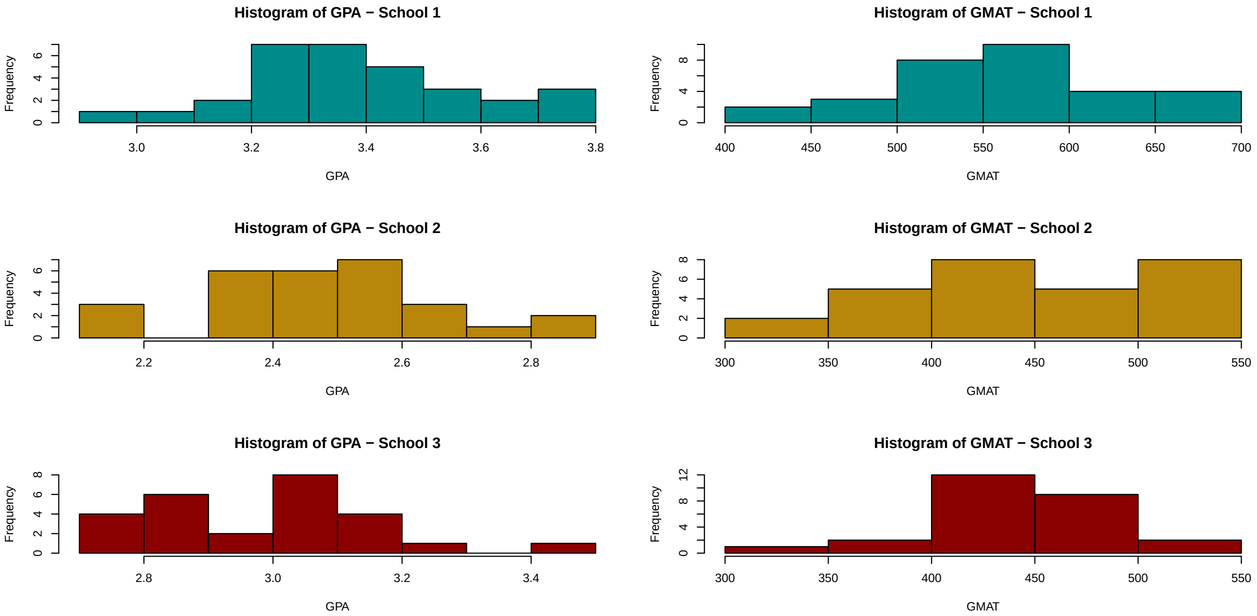
$$\overline{\mathbf{X}}_{\text{school1}} = \begin{bmatrix} 3.4 \\ 561.2 \end{bmatrix} \qquad \overline{\mathbf{X}}_{\text{school2}} = \begin{bmatrix} 2.5 \\ 447.1 \end{bmatrix} \qquad \overline{\mathbf{X}}_{\text{school3}} = \begin{bmatrix} 3 \\ 446 \end{bmatrix}$$

$$\mathbf{S}_{\text{school1}} = \begin{bmatrix} 4.4e-02 & 5.8e-02 \\ 5.8e-02 & 4.6e+03 \end{bmatrix} \qquad \mathbf{S}_{\text{school2}} = \begin{bmatrix} 0.034 & -1.192 \\ -1.192 & 3891.254 \end{bmatrix} \qquad \mathbf{S}_{\text{school3}} = \begin{bmatrix} 0.03 & -5.40 \\ -5.40 & 2246.90 \end{bmatrix}$$

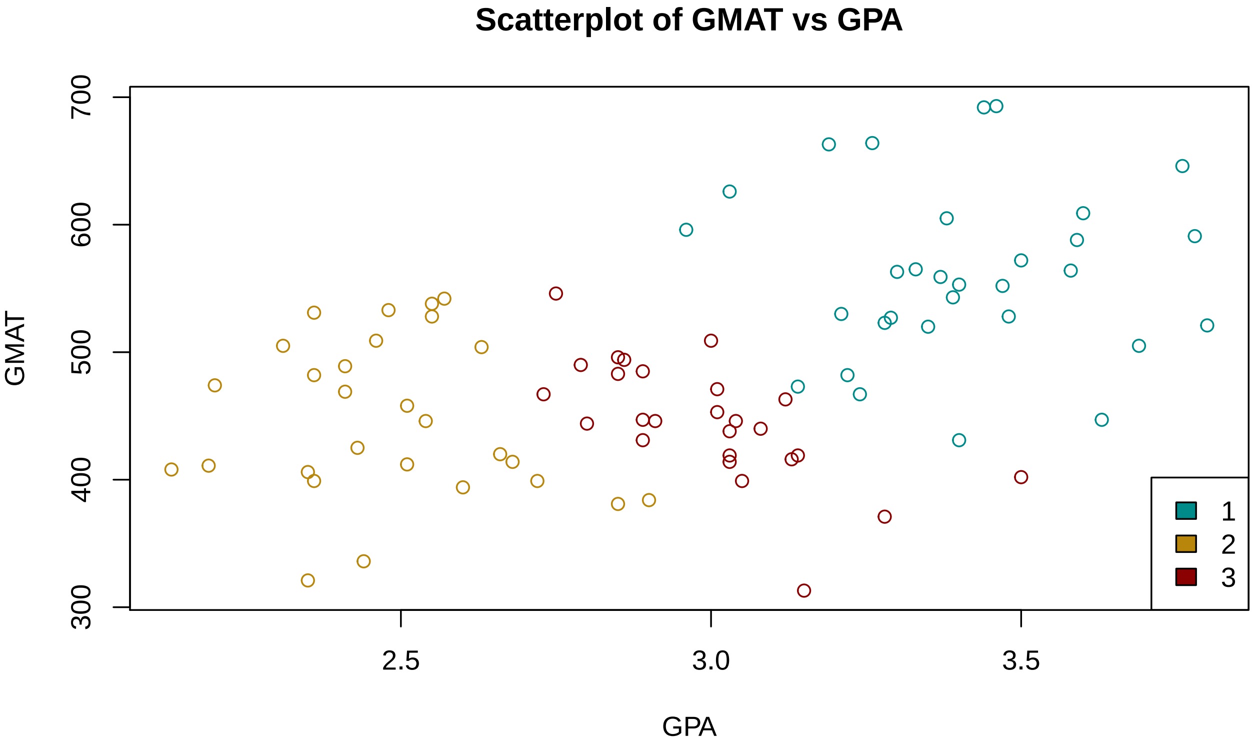
$$\mathbf{r}_{\text{school1}} = \begin{bmatrix} 1.0000 & 0.0041 \\ 0.0041 & 1.0000 \end{bmatrix} \qquad \mathbf{r}_{\text{school2}} = \begin{bmatrix} 1.0 & -0.1 \\ -0.1 & 1.0 \end{bmatrix} \qquad \mathbf{r}_{\text{school3}} = \begin{bmatrix} 1.00 & -0.66 \\ -0.66 & 1.00 \end{bmatrix}$$

Looking at the sample mean vectors, we notice that school 1 has the highest average GPA and the highest average GMAT score out of the three schools. On the other hand, school 2 has the lowest average GPA among the three schools and schools 2 and 3 have similar average GMAT scores (447.1 and 446, respectively). Looking at the variance-covariance matrices, we see that the covariance between GPA and GMAT for school 1 is positive, however for schools 2 and 3, the covariance is negative. To investigate the strength of relationship, correlation matrices are observed. It appears that GPA and GMAT have a correlation of -0.66 for school 3, which means that in school 3, GPA and GMAT have a negative linear relationship (as GPA increases, GMAT scores decrease). For schools 1 and 2, there is little to no correlation between GPA and GMAT scores since the correlation coefficients are small.

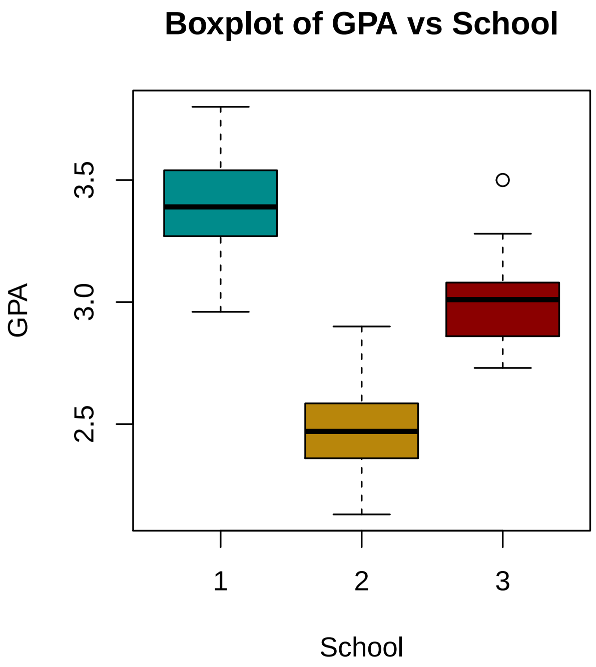
Next, a series of visualizations were created to analyze the data through visual inspection. First, side-by-side histograms were created to observe the distribution of GPA and GMAT for each school.

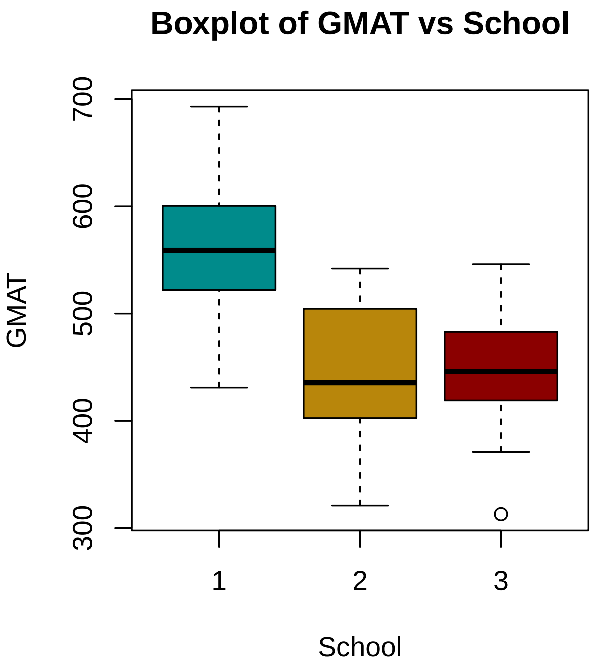


Looking at the histograms, it appears that GPA and GMAT for school 1 and GMAT for school 3 follow a normal distribution. The rest of the histograms do not appear to follow a normal distribution. Next, I created a scatterplot of GMAT vs GPA, color coded by school which is below.



Looking at the scatterplot, it is obvious that school 1 has the largest values for both GPA and GMAT, while school 2 has the overall smallest values for GPA and school 3 has the overall smallest values for GMAT. Also, from visual inspection, the data for school 3 has a negative linear relationship, which corresponds with the correlation coefficient of -0.66 from earlier. Lastly, side-by-side boxplots were created to analyze the data for outliers and discriminating effects among the three schools.





Looking at the boxplot of GPA vs school, it appears there is a discriminating effect because the values for each school are not the same. Specifically, school 1 has the highest GPA and school 2 has the lowest GPA. Looking at the boxplot of GMAT vs school, there appears to be a discriminating effect because school 1 clearly has the highest GMAT scores. Additionally, there appears to be one outlier in school 3 for GPA and one outlier in school 3 for GMAT.

Statistical Analysis

To verify and support the preliminary results from the exploratory analysis, statistical analysis was performed. First, a one-way MANOVA and the Wilk's Lambda test statistic were used to determine if the mean GPA and GMAT scores were the same across the three schools. The null and alternative hypotheses for the test are as follows

$$H_0 : \tau_1 = \cdots = \tau_g = 0 \text{ vs } H_a : \tau_j \neq 0 \text{ for some j.}$$

The MANOVA table is below.

```
##  Response GPA :
##            Df  Sum Sq Mean Sq F value    Pr(>F)    
## SCHOOL      2 12.5015  6.2508  173.31 < 2.2e-16 ***
## Residuals  82  2.9576  0.0361                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
##  Response GMAT :
##            Df Sum Sq Mean Sq F value    Pr(>F)    
## SCHOOL      2 258471  129236   35.35 8.492e-12 ***
## Residuals  82 299784    3656                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wilk's Lambda test statistic can be found in the table below.

```
##           Df   Wilks approx F num Df den Df    Pr(>F)
## SCHOOL     2 0.12638   73.426      4    162 < 2.2e-16 ***
## Residuals 82
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] 0.1263766
```

It is worth noting that through manual calculation, I obtained the same value for Wilk's Lambda test statistic as the MANOVA output did. The Wilk's Lambda test statistic was then used to calculate Bartlett's large sample result using the formula stated in the materials and methods section. I obtained a value of 168.58 for Bartlett's large sample result and a critical value of $\chi^2_{p(g-1)}(0.05) = 9.49$.

Bartlett's large sample result is greater than $\chi^2_{p(g-1)}(0.05)$, which allows us to reject H_0 and conclude that there is a population effect $\tau_j \neq 0$ at $\alpha = 0.05$. In other words, there is enough evidence to statistically conclude that the mean GPA and GMAT scores across the three schools are not equal.

Next, 95% simultaneous confidence intervals were constructed for differences in components of the treatment effects, $\tau_{ki} - \tau_{\ell i}$. The confidence intervals allow us to determine the specific differences in mean GPA and GMAT scores between the three schools. The formula for the general $(1-\alpha)100\%$ simultaneous confidence intervals can be found in the materials and methods section. The 95% simultaneous confidence intervals are below, where "L" denotes the lower bound and "U" denotes the upper bound.

Table 1: 95% confidence interval for School1 - School2.

	GPA	GMAT
L12	0.787	71.532
U12	1.055	156.776

Table 2: 95% confidence interval for School1 - School3.

	GPA	GMAT
L13	0.275	71.520
U13	0.548	158.470

Table 3: 95% confidence interval for School2 - School3.

	GPA	GMAT
L23	−0.650	−43.684
U23	−0.370	45.365

For the difference in mean GPA and GMAT scores between school 1 and school 2, we can be 95% confident that the true difference in GPA lies between 0.79 and 1.06, and the true difference in GMAT scores lies between 71.53 and 156.78.

Next, for the difference in mean GPA and GMAT scores between school 1 and school 3, we can be 95% confident that the true difference in GPA lies between 0.27 and 0.55, and the true difference in GMAT scores lies between 71.52 and 158.47.

Lastly, for the difference in mean GPA and GMAT scores between school 2 and school 3, we can be 95% confident that the true difference in GPA lies between -0.65 and -0.37, and the true difference in GMAT scores lies between -43.68 and 45.37. Since the confidence interval for difference in GMAT scores contains 0, we are unable to conclude that there is a difference in GMAT scores between schools 2 and 3.

Summarizing the results from the confidence intervals, it is obvious that school 1 has higher mean GPA and GMAT scores than schools 2 and 3. Additionally, school 3 has a higher mean GPA than school 2, however, we cannot conclude that there is a significant difference in GMAT scores between schools 2 and 3.

Appendix

```
library(dplyr)
library(ggplot2)
library(ggpubr)
library(ggforce)

# read in the data and convert school to factor variable
data <- read.table('Project_3_Data.txt', header = T)
data$SCHOOL <- as.factor(data$SCHOOL)
head(data)

dim(data)
summary(data$SCHOOL)

# split dataset based on population
school1 <- data %>% filter(SCHOOL == 1) %>% select(GPA, GMAT)
school2 <- data %>% filter(SCHOOL == 2) %>% select(GPA, GMAT)
school3 <- data %>% filter(SCHOOL == 3) %>% select(GPA, GMAT)

write_matex2 <- function(x) {
  begin <- "\\begin{bmatrix}"
  end <- "\\end{bmatrix}"
  X <-
    apply(x, 1, function(x) {
      paste(
        paste(x, collapse = "&"),
        "\\\\"
      )
    })
  paste(c(begin, X, end), collapse = "")
}
# mean vectors
mean_vecs <- lapply(split(data[, c('GPA', 'GMAT')], data$SCHOOL), function(x) colMeans(x))
# var-cov matrices
var_cov_mat <- lapply(split(data[, c('GPA', 'GMAT')], data$SCHOOL), cov)
# correlation matrices
cor_mat <- lapply(split(data[, c('GPA', 'GMAT')], data$SCHOOL), cor)

# side by side histograms
par(mfrow=c(3,2))
hist(school1$GPA, xlab = 'GPA', main = 'Histogram of GPA - School 1', col = 'darkcyan')
hist(school1$GMAT, xlab = 'GMAT', main = 'Histogram of GMAT - School 1', col = 'darkcyan')
hist(school2$GPA, xlab = 'GPA', main = 'Histogram of GPA - School 2', col = 'darkgoldenrod')
hist(school2$GMAT, xlab = 'GMAT', main = 'Histogram of GMAT - School 2', col = 'darkgoldenrod')
hist(school3$GPA, xlab = 'GPA', main = 'Histogram of GPA - School 3', col = 'darkred')
hist(school3$GMAT, xlab = 'GMAT', main = 'Histogram of GMAT - School 3', col = 'darkred')
```

```
cols <- c('darkcyan', 'darkgoldenrod', 'darkred')
plot(data$GPA, data$GMAT, col = cols[data$SCHOOL], main = 'Scatterplot of GMAT vs GPA', xlab = 'GPA', yl
legend('bottomright', fill = cols, legend = c(levels(data$SCHOOL)))

# boxplots
par(mfrow=c(1,2))
```

```
boxplot(GPA ~ SCHOOL, data = data, col = cols[unique(data$SCHOOL)], xlab = 'School', ylab = 'GPA', main
boxplot(GMAT ~ SCHOOL, data = data, col = cols[unique(data$SCHOOL)], xlab = 'School', ylab = 'GMAT', mai
```

```
# one way MANOVA
manova <- manova(as.matrix(data[, 1:2]) ~ SCHOOL, data = data)
summary.aov(manova)

# verify the Wilk's Lambda test statistics
p <- ncol(data) - 1
g <- length(unique(data[, p+1]))
school1 <- data[which(data[, 3] == 1), 1:p]
school2 <- data[which(data[, 3] == 2), 1:p]
school3 <- data[which(data[, 3] == 3), 1:p]
n1 <- sum(data[, 3] == 1)
n2 <- sum(data[, 3] == 2)
n3 <- sum(data[, 3] == 3)
SSR <- (n1 - 1) * var(school1) + (n2 - 1) * var(school2) + (n3 - 1) * var(school3)
SST <- n1 * tcrossprod(colMeans(school1) - colMeans(data[, 1:p])) + n2 * tcrossprod(colMeans(school2) -
Wilks_stats <- det(SSR) / det(SSR + SST)
Wilks_stats

# test statistic for large sample
test_stat <- -(sum(n1, n2, n3)-1-(3+2)/2)*log(Wilks_stats)
critical_val <- qchisq(0.05, p*(g-1), lower.tail = FALSE)

N <- nrow(data)
alpha <- .05
SSR <- (n1 - 1) * var(school1) + (n2 - 1) * var(school2) + (n3 - 1) * var(school3)
W_diag <- diag(SSR)

# construct simultaneous CIs
L12 <- colMeans(school1) - colMeans(school2) - qt(alpha/(p*g*(g-1)), N-g, lower.tail = FALSE)*sqrt(c(1/
U12 <- colMeans(school1) - colMeans(school2) + qt(alpha/(p*g*(g-1)), N-g, lower.tail = FALSE)*sqrt(c(1/
L13 <- colMeans(school1) - colMeans(school3) - qt(alpha/(p*g*(g-1)), N-g, lower.tail = FALSE)*sqrt(c(1/
U13 <- colMeans(school1) - colMeans(school3) + qt(alpha/(p*g*(g-1)), N-g, lower.tail = FALSE)*sqrt(c(1/
L23 <- colMeans(school2) - colMeans(school3) - qt(alpha/(p*g*(g-1)), N-g, lower.tail = FALSE)*sqrt(c(1/
U23 <- colMeans(school2) - colMeans(school3) + qt(alpha/(p*g*(g-1)), N-g, lower.tail = FALSE)*sqrt(c(1/
SCI <- list("school1.school2" = rbind(L12, U12), "school1.school3" = rbind(L13, U13), "school2.school3"
stargazer(SCI)
```