MANOVA Lab Practice

Baybayon, Darlyn Antoinette B.

```
suppressPackageStartupMessages({
  library(tidyverse)
  library(readr)
  library(MVN)
  library(mice)
  library(heplots)
})
```

The Data

```
df <- read_csv("manova_mancova_practice.csv", show_col_types = FALSE)
head(df)
```

```
## # A tibble: 6 x 12
##   ID    Treatment School Gender   Age SES_Index Pre_Math Pre_Reading Pre_Science
##   <chr> <chr>     <chr>  <chr>  <dbl>     <dbl>    <dbl>       <dbl>       <dbl>
## 1 C000~ Control   East   F       17.1     0.163     50.3        64.1        63.8
## 2 C001~ Control   North  F       15.9     0.586     41.9        54.7        53.1
## 3 C002~ Control   West   F       17.5     0.711     54.9        60.7        37.3
## 4 C003~ Control   North  M       17.6     0.793     65.8        55.8        58.9
## 5 C004~ Control   North  F       15.0    -0.349     63.2        60.3        62
## 6 C005~ Control   West   M       15.6    -0.462     53.9        58.1        56.2
## # i 3 more variables: Post_Math <dbl>, Post_Reading <dbl>, Post_Science <dbl>
```

- Design: 3 instructional methods (Treatment = Control, MethodA, MethodB), 4 Schools (North/South/East/West), Gender, Age, SES_Index (z-score).
- Outcomes: Post_Math, Post_Reading, Post_Science (0–100).
- Covariates: Pre_Math, Pre_Reading, Pre_Science, SES_Index, Age.

Data Preprocessing

Convert data types

```
df <- df %>%
  select(-"ID") %>%
  mutate(across(c("Treatment", "Gender", "School"), as.factor))
```

Check nulls and handle them

```
colSums(is.na(df))
```

```
##    Treatment       School       Gender          Age    SES_Index     Pre_Math
##            0            0            0            0            9            5
##  Pre_Reading  Pre_Science    Post_Math Post_Reading Post_Science
##            6            8            8           10            4
```

There are some missing data in numeric columns. According to the design of the data, the missingness is MCAR, so we can impute the missing values with a simple mean imputation.

```
df_imputed <- df %>%
  mutate(
    SES_Index = ifelse(is.na(SES_Index), mean(SES_Index, na.rm=TRUE), SES_Index),
    Pre_Math = ifelse(is.na(Pre_Math), mean(Pre_Math, na.rm=TRUE), Pre_Math),
    Pre_Reading = ifelse(is.na(Pre_Reading), mean(Pre_Reading, na.rm=TRUE), Pre_Reading),
    Pre_Science = ifelse(is.na(Pre_Science), mean(Pre_Science, na.rm=TRUE), Pre_Science),
    Post_Math = ifelse(is.na(Post_Math), mean(Post_Math, na.rm=TRUE), Post_Math),
    Post_Reading = ifelse(is.na(Post_Reading), mean(Post_Reading, na.rm=TRUE), Post_Reading),
    Post_Science = ifelse(is.na(Post_Science), mean(Post_Science, na.rm=TRUE), Post_Science),
  )
colSums(is.na(df_imputed))
```

```
##    Treatment       School       Gender          Age    SES_Index     Pre_Math
##            0            0            0            0            0            0
##  Pre_Reading  Pre_Science    Post_Math Post_Reading Post_Science
##            0            0            0            0            0
```

Mean and variance after imputation

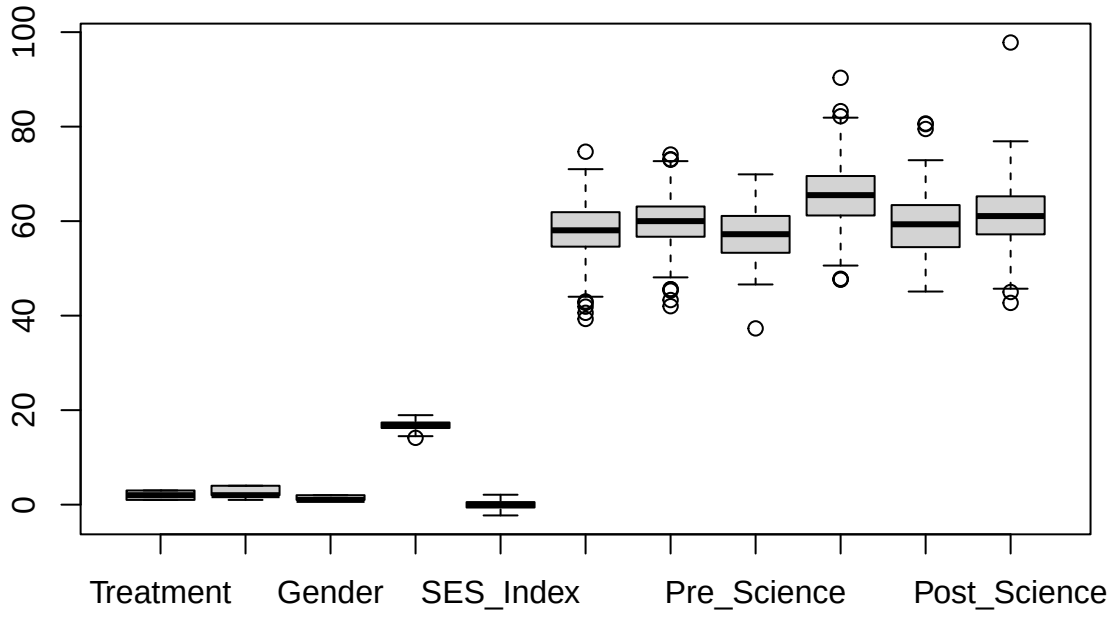
```
round(data.frame(mean = sapply(select(df,where(is.numeric)), mean, na.rm=TRUE),
           mean_imp = sapply(select(df_imputed,where(is.numeric)), mean),
           var = sapply(select(df,where(is.numeric)), var, , na.rm=TRUE),
           var_imp = sapply(select(df_imputed,where(is.numeric)), var)),4)
```

```
##                  mean mean_imp     var var_imp
## Age           16.7966  16.7966  0.7701  0.7701
## SES_Index     -0.0331  -0.0331  0.8640  0.8215
## Pre_Math      58.0575  58.0575 38.5963 37.5417
## Pre_Reading   59.6517  59.6517 32.7753 31.7007
## Pre_Science   57.2375  57.2375 27.7856 26.5709
## Post_Math     65.5115  65.5115 53.5236 51.1838
## Post_Reading  59.3368  59.3368 43.7254 41.3361
## Post_Science  61.0757  61.0757 54.9417 53.7408
```

Check and handle outliers

Univariate outliers

```
boxplot(df_imputed)
```




```
lapply(select(df_imputed,where(is.numeric)), function(x) boxplot.stats(x)$out)
```

```
## $Age
## [1] 14.14
##
## $SES_Index
## numeric(0)
##
## $Pre_Math
## [1] 41.9 43.0 40.6 39.3 42.7 74.7
##
## $Pre_Reading
## [1] 73.0 45.3 43.3 42.0 45.6 74.1 73.1 45.6
##
## $Pre_Science
## [1] 37.3
##
## $Post_Math
## [1] 47.80000 47.70000 47.60000 82.20000 90.34177 83.28444
##
## $Post_Reading
## [1] 80.67652 79.50072 80.52707
##
## $Post_Science
## [1] 45.00000 42.70000 97.80108
```

Multivariate outliers

```r
dvs <-df_imputed[,9:11]
dvs$dist <- mahalanobis(dvs, colMeans(dvs), cov(dvs))
threshold <- qchisq(1 - 0.001, df = 3)
outliers <- dvs[dvs$dist > threshold,]
outliers
```

```
## # A tibble: 1 x 4
##   Post_Math Post_Reading Post_Science  dist
##       <dbl>        <dbl>        <dbl> <dbl>
## 1      90.3         79.5         97.8  28.1
```

```r
df_no_outliers <- df_imputed[-114,]
```

One-way MANOVA

Treatment → DVs = Post_Math, Post_Reading, Post_Science.

```r
manova_result <- manova(
  cbind(Post_Math, Post_Reading, Post_Science) ~ Treatment,
  data = df_no_outliers
)
```

```r
summary(manova_result)
```

```
##            Df   Pillai approx F num Df den Df  Pr(>F)
## Treatment   2 0.091201   2.8508      6    358 0.01004 *
## Residuals 180
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A one-way multivariate analysis of variance (MANOVA) was conducted to examine the effect of Treatment on scores in Math, Reading, and Science. The results indicate a statistically significant difference in the mean test scores across the different treatment groups (Pillai = 0.091, F(6, 358) = 2.85, p = 0.010). Thus, we reject the null hypothesis, and conclude that the instructional method has a significant effect on the combined outcomes of the scores.

Test Statistics

```r
wilks <- summary(manova_result, test ="Wilks")
pillai <- summary(manova_result, test ="Pillai")
roy <- summary(manova_result, test ="Roy")
lh<- summary(manova_result, test ="Hotelling-Lawley")

data.frame( Wilks = wilks$stats[1,2],
            Pillai = pillai$stats[1,2],
            Lawley_Hotell = lh$stats[1,2],
            Roy = roy$stats[1,2])
```

```
##       Wilks    Pillai Lawley_Hotell        Roy
## 1 0.9097418 0.0912012    0.09817654 0.08614432
```

The Wilks' Lambda quantifies how much of the variance in the combination of the dependent variables is not explained by the independent variable. The Wilks' Lambda of 0.9097 indicates that 90.97% of the variance in test scores is not explained by the treatment group, suggesting a weak effect.

The Pillai's Trace measures how much of the multivariate variance is explained by the independent variable. The value of 0.0912 indicates that about 9.12% of the variance in test scores is explained by the treatment group. Similar to Wilks' Lambda, this suggests that the treatment has a weak effect on the scores.

The Lawley–Hotelling Trace value of 0.0982 similarly indicates that the treatment explains a low percentage of the combined variance across the test measures, again pointing to a small effect.

The Roy's Largest Root value of 0.0861 reflects the maximum variance explained in a single dimension where group differences are strongest. This also suggests a weak effect of the treatment on test scores.

```
summary.aov(manova_result)
```

```
##  Response Post_Math :
##              Df Sum Sq Mean Sq F value Pr(>F)
## Treatment     2   87.1  43.525  0.9047 0.4065
## Residuals   180 8659.7  48.109
##
##  Response Post_Reading :
##              Df Sum Sq Mean Sq F value  Pr(>F)
## Treatment     2  231.5 115.729  3.0085 0.05186 .
## Residuals   180 6924.2  38.468
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Post_Science :
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Treatment     2  573.6 286.820  6.5312 0.001827 **
## Residuals   180 7904.8  43.916
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Follow-up univariate ANOVAs revealed that Post_Science scores differed significantly between treatment groups (p=0.002) while Post_Reading scores differed marginally (p =0.0519).

Check assumptions

Independence of Observations

Based on the design of the data, the observations within and between groups are independent. Each student is in a single treatment group (Control, Method A, or Method B). Each student also contributes only one set of Endline test scores. As such, the observations within and between groups are considered independent, satisfying a key assumption for multivariate analysis.

Homogeneity of Variance-Covariance Matrices

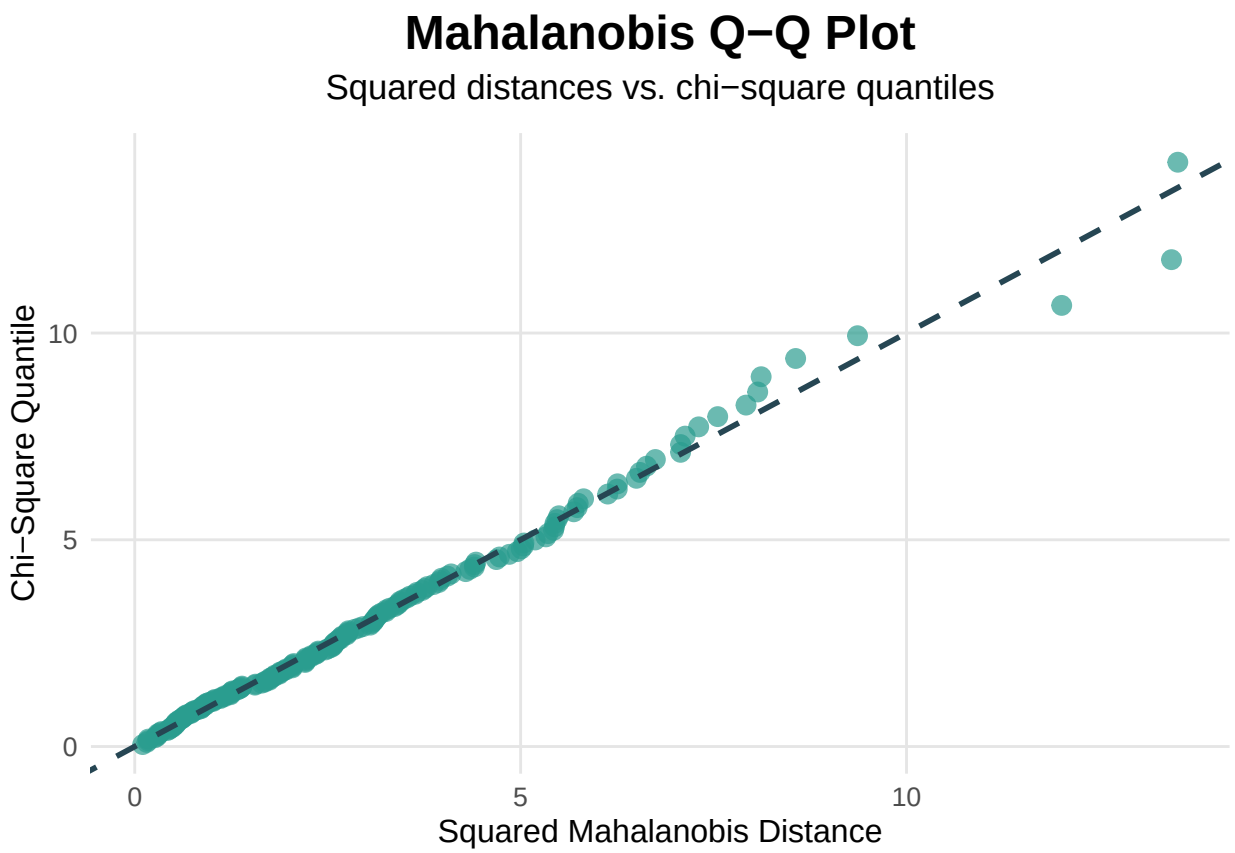
```
boxM(cbind(Post_Math, Post_Reading, Post_Science) ~ Treatment,
  data = df_no_outliers)
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  df_no_outliers
## Chi-Sq (approx.) = 8.1177, df = 12, p-value = 0.7759
```

The Box's M test did not reveal a significant difference ($p > 0.05$) in covariance matrices across the treatment groups. Therefore the assumption of homogeneity of covariance matrices is met and the MANOVA is valid.

Multivariate Normality

```
multivariate_diagnostic_plot(df_no_outliers[,9:11], type="qq")
```



The Q-Q plot shows that most points align with the theoretical quantiles with some observations in the upper tail deviating slightly from the diagonal reference line. This indicates good agreement with the expected normal distribution Hence, based on the Q-Q plot of Mahalanobis distances, the assumption of multivariate normality appears to be reasonably satisfied.

Two-way MANOVA

Treatment, School, and their interaction

```
manova_2w_result <- manova(cbind(Post_Math, Post_Reading, Post_Science) ~
                           Treatment * School, data = df_no_outliers)
summary(manova_2w_result)
```

```
##                   Df   Pillai approx F num Df den Df   Pr(>F)
## Treatment          2 0.097797  2.91339      6    340 0.008763 **
## School             3 0.027128  0.52014      9    513 0.860308
## Treatment:School   6 0.109303  1.07764     18    513 0.371485
## Residuals        171
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A two-way MANOVA was conducted to examine the effect of Treatment, School, and their interaction on combined test scores in Math, Reading, and Science. The results reveal a statistically significant multivariate effect of Treatment (Pillais = 0.098, F(6,340) = 2.91, p = 0.009) on the test scores. Meanwhile School and its interaction with Treatment did not reach statistical significance.