MANOVA TEST ANALYSIS

Data Cleaning

```
# loading student performance data
performance <- read.csv(file = 'Assignment_1.csv',
                        col.names = c('Id', 'ses', 'engach', 'mathsach'))
```

```
head(performance)
```

```
##   Id ses engach mathsach
## 1  1   2  16.67    16.42
## 2  2   1  14.67    18.52
## 3  3   1  19.67    17.05
## 4  4   1  19.33    17.05
## 5  5   3   9.34    22.86
## 6  6   2  13.33    13.45
```

```
str(performance)
```

```
## 'data.frame':    240 obs. of  4 variables:
##  $ Id      : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ ses     : int  2 1 1 1 3 2 1 2 1 1 ...
##  $ engach  : num  16.67 14.67 19.67 19.33 9.34 ...
##  $ mathsach: num  16.4 18.5 17.1 17.1 22.9 ...
```

Step 1: Check for MANOVA Assumptions

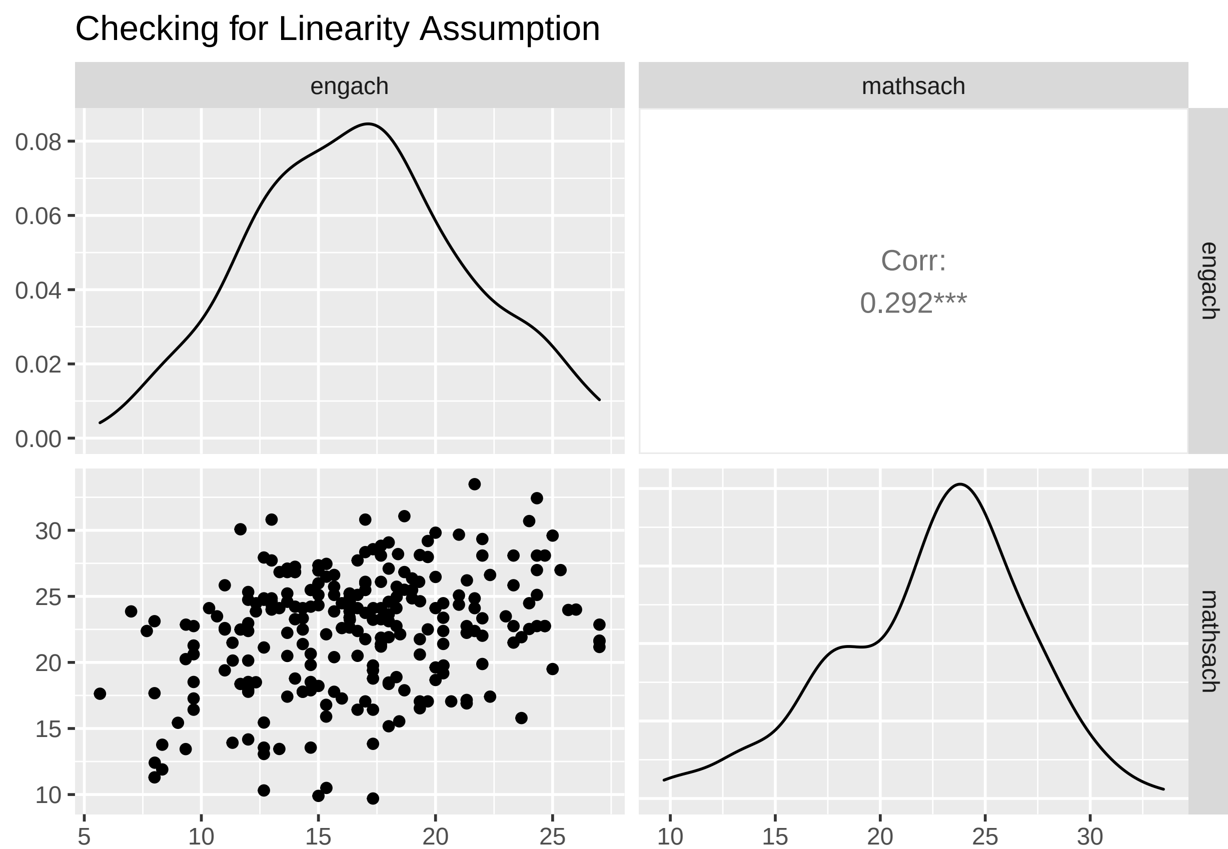
Assumptions of MANOVA

- There must be linearity between pairs of dependent or response variables

```
require(GGally)

# checking for Linearity

# pairs plot
ggpairs(performance[,c("engach", "mathsach")],) +
  ggtitle('Checking for Linearity Assumption')
```




```r
# correlation coefficients
corr <- cor.test(performance$engach, performance$mathsach)
corr$p.value # pvalue
```

```
## [1] 4.282818e-06
```

```r
corr$estimate # correlation coefficient
```

```
##       cor
## 0.2917861
```

- Normality of multivariate variables and groups

```r
# checking for normality
library(MVN)
```

```
## Warning: package 'MVN' was built under R version 4.2.1
```

```r
library(mvnormtest)
library(nortest)


# Checking for multivariate normality
mvnormtest::mshapiro.test(t(performance[,3:4]))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.98202, p-value = 0.003892
```

```r
res <- mvn(performance[,3:4], mvnTest = 'mardia')
res
```

```
## $multivariateNormality
##              Test         Statistic            p value Result
## 1 Mardia Skewness  11.3402079285109 0.0229952290402109     NO
## 2 Mardia Kurtosis -1.61080952342355  0.107221244752767    YES
## 3             MVN              <NA>               <NA>     NO
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling  engach      0.4272    0.3106    YES
## 2 Anderson-Darling mathsach     1.8226    0.0001    NO
##
## $Descriptives
##            n     Mean  Std.Dev Median  Min   Max    25th    75th       Skew
## engach   240 16.70192 4.517747 16.670 5.67 27.00 13.5875 19.6700  0.1195675
## mathsach 240 22.40825 4.544545 23.045 9.70 33.49 19.4000 25.2475 -0.4751019
##             Kurtosis
## engach   -0.478135999
## mathsach  0.002242405
```

Both dependent variables are not normally distributed, with pvalue (0.003892) < 0.05. Therefore, we reject the null that the multivariate samples are normally distributed.

Next we will check the univariate normality of the two dependent variables. We will use the shapiro test, Anderson-Darling, and Lillie tests to do that.

```r
# checking for univariate normality

# Shapiro Test
shapiro.test(performance$engach)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  performance$engach
## W = 0.99088, p-value = 0.1391
```

```r
shapiro.test(performance$mathsach)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  performance$mathsach
## W = 0.97878, p-value = 0.001159
```

```r
# Anderson-Darling Test
ad.test(performance$engach)
```

```
##
##  Anderson-Darling normality test
##
## data:  performance$engach
## A = 0.42718, p-value = 0.3106
```

```r
ad.test(performance$mathsach)
```

```
##
##  Anderson-Darling normality test
##
## data:  performance$mathsach
## A = 1.8226, p-value = 0.0001137
```

```r
# Kolmogorov-Smirnov test
lillie.test(performance$engach)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  performance$engach
## D = 0.037162, p-value = 0.5796
```

```r
lillie.test(performance$mathsach)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  performance$mathsach
## D = 0.093353, p-value = 2.846e-05
```

From the three normality tests applied to test the normality of engach and mathsach, only engach is normally distributed while mathsach is not. From all the tests, the pvalues for engach was greater than the 0.05, hence we will fail to reject the null hypothesis and conclude that the distribution is normal. On the other hand, mathsach has pvalues less than 0.05, as a result, we reject the null hypothesis and conclude that the mathsach distribution is not normal in nature.

- Testing normality of univariates across outcomes using the Shapiro-Wilk test

```r
# univariate normality across outcomes

# engach
tapply(performance$engach, performance$ses, shapiro.test)
```

```
## $`1`
##
##  Shapiro-Wilk normality test
```

```
##
## data:  X[[i]]
## W = 0.99029, p-value = 0.8132
##
##
## $`2`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.97767, p-value = 0.1748
##
##
## $`3`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.97846, p-value = 0.1958
```

From the shapiro-Wilk test, engach variable is normally distributed across outcomes, with pvalues > 0.05.

```
# mathsach
tapply(performance$mathsach, performance$ses, shapiro.test)
```

```
## $`1`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.97357, p-value = 0.09589
##
##
## $`2`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.95668, p-value = 0.008478
##
##
## $`3`
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.98631, p-value = 0.5542
```

For mathsach, only the second class (2) is not normally distributed, with the mathsach distribution across other classes normally distributed.

- Checking for adequate sample size

```r
# checking for adequate sample size
table(performance$ses)
```

```
##
##  1  2  3
## 80 80 80
```

Both classes have similar number of samples

- Assumption of homogeneity of covariances

```r
library(biotools)
```

```
## Warning: package 'biotools' was built under R version 4.2.1
```

```
## Loading required package: MASS
```

```
## ---
## biotools version 4.2
```

```r
# Checking for Homogeneity of covariance using a boxM test
boxM(data = performance[,3:4], grouping = performance$ses)
```

```
##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  performance[, 3:4]
## Chi-Sq (approx.) = 5.154, df = 6, p-value = 0.5242
```

Step 2: MANOVA test

```r
# MANOVA test
manova_model <- manova(cbind(engach, mathsach) ~ ses, data=performance)
```

- MANOVA model using various tests: Pillai, Wilk, Hotelling-Lawley and Roy variants.

```r
# Pillai Test
summary(manova_model, test = 'Pillai')
```

```
##            Df   Pillai approx F num Df den Df   Pr(>F)
## ses         1 0.051124   6.3846      2    237 0.001992 **
## Residuals 238
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Wilk Test
summary(manova_model, test = 'Wilk')
```

```
##            Df   Wilks approx F num Df den Df   Pr(>F)
## ses         1 0.94888   6.3846      2    237 0.001992 **
## Residuals 238
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Hotelling-Lawley Test
summary(manova_model, test = 'Hotelling-Lawley')
```

```
##            Df Hotelling-Lawley approx F num Df den Df   Pr(>F)
## ses         1         0.053879   6.3846      2    237 0.001992 **
## Residuals 238
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Roy Test
summary(manova_model, test = 'Roy')
```

```
##            Df      Roy approx F num Df den Df   Pr(>F)
## ses         1 0.053879   6.3846      2    237 0.001992 **
## Residuals 238
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step 3: Univariate ANOVA for each dependent variable

```r
# univariate ANOVA of Dependent variables
summary.aov(manova_model)
```

```
##  Response engach :
##              Df Sum Sq Mean Sq F value Pr(>F)
## ses           1   44.8  44.849  2.2085 0.1386
## Residuals   238 4833.2  20.307
##
##  Response mathsach :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## ses           1  247.3 247.307  12.553 0.0004759 ***
## Residuals   238 4688.7  19.701
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There's no significant difference in the mean engach value for all outcomes. The pvalue (0.1386) is greater than the level of significance (5%), therefore, we will fail to reject the null hypothesis.

On the other hand for mathsach, there's a significant difference in the mean mathsach value for all outcome categories. This is significant at 5% significance level. As a result, we will reject the null hypothesis of no difference in means.

Step 4: Pairwise Test: Post Hoc Tests

```r
# pairwise test using the Bonferroni Adjustments

# for engach
pairwise.t.test(performance$engach, performance$ses,
                p.adjust.method = 'bonferroni')
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  performance$engach and performance$ses
##
##   1    2
## 2 1.00 -
## 3 0.42 1.00
##
## P value adjustment method: bonferroni
```

```r
# for mathsach
pairwise.t.test(performance$mathsach, performance$ses,
                p.adjust.method = 'bonferroni')
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  performance$mathsach and performance$ses
##
##   1      2
## 2 0.2927 -
## 3 0.0015 0.1872
##
## P value adjustment method: bonferroni
```

From the post-hoc test result for engfach, there's no significant difference in engach scores by either of the outcome groups. All groups have pvalues greater than 0.05.

For mathsach, there's a statistical difference in the mathsach score gotten by people in group 1 and 3. For the other groups, 2 and 3, there is none.