Exercise 1: Consider the following data,

```
group1 <- structure(c(16, 11, 14, 20, 13, 12, 13, 21, 11, 15, 16, 11,
14, 20, 13, 12, 13, 21, 11, 15, 28, 27, 27, 23, 26, 25, 26, 35, 21,
26, 26, 29, 27, 24, 29, 23, 22, 22, 16, 23, 54, 50, 36, 41, 51,
49, 39, 48, 44, 44, 54, 50, 36, 41, 51, 49, 39, 48, 44, 44), .Dim = c(20L,
3L), .Dimnames = list(NULL, c("G1", "H1", "J1")))
#
group2 <- structure(c(21, 11, 16, 19, 15, 22, 16, 19, 19, 18, 21,
11, 16, 19, 15, 22, 16, 19, 19, 18, 30, 33, 35, 23, 32, 37, 31, 24, 21,
28, 24, 20, 31, 38, 24, 28, 19, 30, 25, 33, 44, 45, 30, 29, 38,
47, 35, 44, 50, 35, 44, 45, 30, 29, 38, 47, 35, 44, 50, 35),
.Dim = c(20L, 3L), .Dimnames = list(NULL, c("G2", "H2", "J2")))
#
group1; group2
```

a What are the null and alternative hypotheses? What are the assumptions of Hotelling's T2 test?

Solution:
The null hypothesis of Hotelling's T2 test assumes that the means of the two groups are equal, the alternative hypothesis says that they are different,

$$H_0 : \mu_1 = \mu_2,$$

$$H_a : \mu_1 \neq \mu_2.$$

Hotelling's T2 test assumes that the groups are sampled from a multivariate normal distribution, the samples in each group are independent, and the two groups have the same covariance matrix.

b Use any functions you wish to perform a Hotelling's T2 test (0.05 significance level) on the data in group1 and group2.

Solution:
Using the HotellingsT2() function from the ICSNP package we get the following,
Code:

```
> library(ICSNP)
Loading required package: mvtnorm
Loading required package: ICS

> HotellingsT2(group1, group2)

        Hotelling's two sample T2-test

data:  group1 and group2
T.2 = 6.3084, df1 = 3, df2 = 36, p-value = 0.001498
alternative hypothesis: true location difference is not equal to c(0,0,0)
```

With a p-value of 0.001498 we would reject the null hypothesis and say that at a $\alpha = .05$ significance level there is a difference between the means of the two groups.

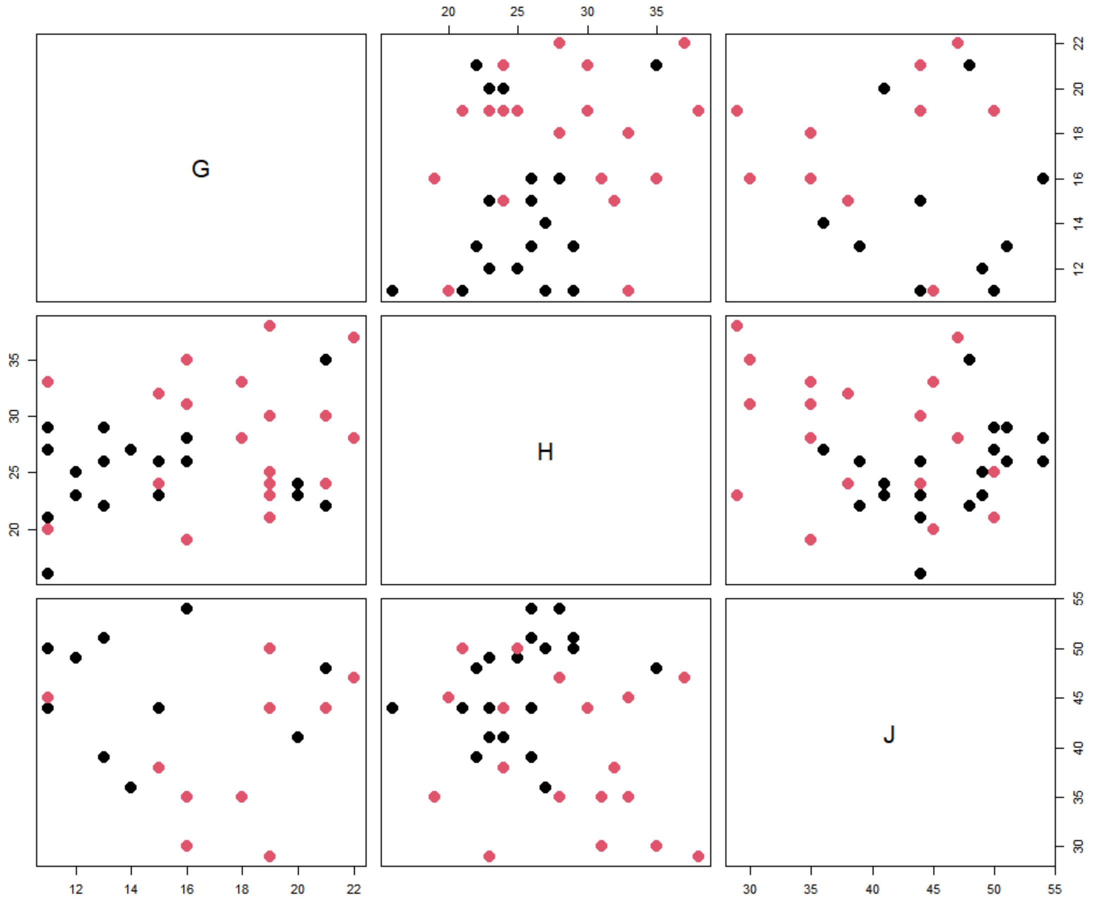
c Perform a t-test on each of the three variables and a graphic of your choice. What are your conclusions? Assume the 0.05 significance level.

Solution:
We can perform a t-test on each of the variables in the groups using t.test(). Doing so we find that both the J and G variable are significantly different with p-values around .006. The H variable had a p-value of 0.05312 which would fail to reject the null at the $\alpha = .05$ level, but it is close enough that accross the whole group we could likely say that our findings with the individual t-tests match our Hotelling's T2.
Code:

```
group1 <- data.frame(group1)
group2 <- data.frame(group2)
------------------------------------------------------------------
> t.test(group1$G1, group2$G2)

        Welch Two Sample t-test

data:  group1$G1 and group2$G2
t = -2.9038, df = 37.703, p-value = 0.006132
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.0920142 -0.9079858
sample estimates:
mean of x mean of y
     14.6      17.6



------------------------------------------------------------------
> t.test(group1$H1, group2$H2)

        Welch Two Sample t-test

data:  group1$H1 and group2$H2
t = -2.0043, df = 33.745, p-value = 0.05312
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.14338425  0.04338425
sample estimates:
mean of x mean of y
    25.25     28.30


------------------------------------------------------------------
> t.test(group1$J1, group2$J2)

        Welch Two Sample t-test
```

```
data:  group1$J1 and group2$J2
t = 2.9147, df = 36.037, p-value = 0.006085
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  1.794887 10.005113
sample estimates:
mean of x mean of y
     45.6      39.7
--------------------------------------------------------------------
```

Figure 1: Scatterplot Matrix of the 2 groups. black is group1, red is group2.



d Use a Box's M test (you can use the function in biotools, or any other one in R you wish to use) to test... something. What IS it's null and alternative hypothesis? What is your conclusion? Discuss how this influences your analysis. Assume the 0.05 significance level.

Solution:
Box's M test is a diagnostic test to check the assumption of equal covariance matrices among the groups. The null hypothesis assumes that the covariance matrices are equal and the alternative hypothsis assumes that they are not. Wrangling the data to use the boxM() function from biotools we get a p-value of .3079, therefore at the $\alpha = 0.05$ significance level we would fail to reject the null hypothesis and conclude that the covariance matrices for these groups are the same.
Code:

```
%% Data Wrangle
> group1 <- data.frame(group1)
> group2 <- data.frame(group2)
> group1$class <- 1
> group2$class <- 2
%% Input data needs to be a single dataframe
> Fulldata <- rbind(group1, group2)
%% Class label must be a factor variable
> Fulldata$class <- as.factor(Fulldata$class)


%% Running boxM() test.
> boxM(Fulldata[, 1:3], Fulldata$class)

        Box's M-test for Homogeneity of Covariance Matrices

data:  Fulldata[, 1:3]
Chi-Sq (approx.) = 7.1427, df = 6, p-value = 0.3079
```

e What IS a significance level?

Solution:
A significance can be described as a measure of type I error for hypotheses test. To recall type I error is when we reject the null when it is the truth. A $\alpha = .05$ significance level means that we are accepting a five percent chance of committing a type I error. We compare this α value to the p-value of a test statistic because the p-value is the probability of obtaining out test-statistic under the null hypothesis. Clearly when the p-value is zero we must reject the null, this doesn't always happen so we accept a certain level of uncertainty so when p-values come out lower than our significance level we also reject the null.

Exercise 2: Skim the article (on Blackboard):
WJ Cooper, MW Westneat (2009) Form and function of damselfish skulls: rapid and repeated evolution into a limited number of trophic niches. BMC Evolutionary Biology 9 (24). Among a variety of univariate and multivariate tests, they perform a MANOVA. What are they testing? What are the conclusions? Do they follow this with simpler tests? If so, which ones?

Solution:
The goal of the paper was to use morphological and biomechanical analysis to see if there was a significant difference in the anatomical structure among different species of damselfish mainly grouped together as omnivores, herbivores, and planktivores. The tests described in the results section are used to see if there is a significant difference in the anatomical data among these groups of damselfish. Using the Wilk's Lambda and Goodall's F test they found that there was indeed a significant difference in the anatomy of these different groups of damselfish. A pairwise ANOVA found that the majority of the difference came from the planktivore group, as the pairwise test between omnivores and herbivores failed to reject the null with a p-value of .1977. The paper also mentions further pairwise analysis comparing the individual predictors in each group, stating that of the seven that where found to be significant, the planktivore group was considered in six of them. This further analysis was documented in additional file 4, which was not accessible through the link in the document.

Exercise 3: a In Appendix One I run a MANOVA followed by two ANOVAs, one for the first variable (over all three groups) and one for a second variable. Interpret the output. Is the result obvious from the plotted data?

Solution:
The MANOVA using the Wilk's Lambda test statistic finds that on the $\alpha = .05$ significance level there is a significant difference in the means of each group. The further ANOVAs test each individaul variable for each group. We find that means of the second variable were the most distinguishable, with a p-value of .05473 as apposed to the first variable's .16. Looking at the plot I would say that this result is not obvious, maybe there is a slight clustering which is pushing the means of each group apart, but at a first glance it is not clear to me that the means of each group are significantly different. Even more so the results of the individual ANOVA are not obvious from looking at the plot. Maybe after knowing the results of the tests one could distinguish the difference between the means along the y-axis but without prior knowledge I wouldn't make that conclusion.

b I could also conduct three Hotelling's T2 test, one for each pair of groups. What would this tell me? The output is in Appendix Two.

Solution:
Running pairwise Hotelling's T2 tests tells you which means among your groups are significantly different. The pairwise analysis found that group2 and group3 showed the most significant difference, with a p-value of .02993, while all other test failed to reject the null at a $\alpha = .05$ significance level.

Exercise 4: We want to model growth curves for two responses, y1 and y2. We'll do that by regressing on $x1$ (1,2,3,...,20), $x2 = x1^2$, and $x3 = x1^3$.

```
Y <- structure(c(1958, 1922, 2004, 1924, 1903, 1950, 1864, 1952,
1979, 1954, 1977, 1976, 1938, 2005, 2027, 1932, 2025, 2084, 2082,
2065, 1035, 987, 971, 966, 908, 881, 928, 999, 941, 944, 931, 898,
918, 971, 966, 998, 989, 993, 1060, 1030), .Dim = c(20L, 2L), .Dimnames = list(
    NULL, c("y1", "y2")))
#
X <- structure(c(-9, -8, -7, -6, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4,
5, 6, 7, 8, 9, 10, 81, 64, 49, 36, 25, 16, 9, 4, 1, 0, 1, 4,
9, 16, 25, 36, 49, 64, 81, 100, -729, -512, -343, -216, -125,
-64, -27, -8, -1, 0, 1, 8, 27, 64, 125, 216, 343, 512, 729, 1000
), .Dim = c(20L, 3L), .Dimnames = list(NULL, c("x1", "x2", "x3"
)))
#
cbind(X,Y)
```

a Perform (using any function you wish in R) a multiple linear regression to predict $y1$ from $x1, x2, x3$, then another one predicting $y2$ from $x1, x2, x3$. Write down the slopes and intercepts.

Solution:
Performing the two MLRs using the lm() command we get the following summary report,
Code:

```
> summary(lm(Y[,1] ~ X))

Call:
lm(formula = Y[, 1] ~ X)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-81.893 -18.562   9.231  19.792  62.881

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 1947.36929   13.23211 147.170   <2e-16 ***
Xx1            6.90672    3.79740   1.819   0.0877 .
Xx2            0.77170    0.30818   2.504   0.0235 *
Xx3           -0.01249    0.05889  -0.212   0.8347
---

Residual standard error: 39.12 on 16 degrees of freedom
Multiple R-squared:  0.6312,    Adjusted R-squared:  0.562
F-statistic: 9.127 on 3 and 16 DF,  p-value: 0.000937



------------------------------------------------------------
> summary(lm(Y[,2] ~ X))

Call:
lm(formula = Y[, 2] ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-51.245 -13.627  -0.435  15.323  75.958

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 925.51452   10.39014  89.076  < 2e-16 ***
Xx1           3.81656    2.98180   1.280 0.218810
Xx2           1.20569    0.24199   4.982 0.000136 ***
Xx3          -0.04227    0.04624  -0.914 0.374242
---

Residual standard error: 30.72 on 16 degrees of freedom
Multiple R-squared:  0.6476,    Adjusted R-squared:  0.5816
F-statistic: 9.803 on 3 and 16 DF,  p-value: 0.0006572
```

b Run the functions lm(Y X), and then Anova() on the output of the lm(Y X) function. What does this tell you?

Solution:

Running the full multivariate model we get a summary report which combines the previous individual MLR summary reports. This is as expected given the analogous nature of the multivariate model. Running the Anova() function(I ran Anova() from the car package instead of anova()) we get the Pillai test statistic, which with a p-value of 0.0003634 we reject the null and conclude that there is a significant difference

among the response means.
Code:

```
------------------------------------------------------------
> summary(lm(Y ~ X))
Response y1 :

Call:
lm(formula = y1 ~ X)

Residuals:
    Min      1Q  Median      3Q     Max 
-81.893 -18.562   9.231  19.792  62.881 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1947.36929   13.23211 147.170   <2e-16 ***
Xx1            6.90672    3.79740   1.819   0.0877 .  
Xx2            0.77170    0.30818   2.504   0.0235 *  
Xx3           -0.01249    0.05889  -0.212   0.8347    
---

Residual standard error: 39.12 on 16 degrees of freedom
Multiple R-squared:  0.6312,	Adjusted R-squared:  0.562 
F-statistic: 9.127 on 3 and 16 DF,  p-value: 0.000937
------------------------------------------------------------
Response y2 :

Call:
lm(formula = y2 ~ X)

Residuals:
    Min      1Q  Median      3Q     Max 
-51.245 -13.627  -0.435  15.323  75.958 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 925.51452   10.39014  89.076  < 2e-16 ***
Xx1           3.81656    2.98180   1.280 0.218810    
Xx2           1.20569    0.24199   4.982 0.000136 ***
Xx3          -0.04227    0.04624  -0.914 0.374242    
---

Residual standard error: 30.72 on 16 degrees of freedom
Multiple R-squared:  0.6476,	Adjusted R-squared:  0.5816 
F-statistic: 9.803 on 3 and 16 DF,  p-value: 0.0006572
------------------------------------------------------------

> Anova(lm(Y ~ X))

Type II MANOVA Tests: Pillai test statistic
  Df test stat approx F num Df den Df    Pr(>F)    
X  3    1.0405   5.7834      6     32 0.0003634 ***
------------------------------------------------------------
```

c Finally, run the following matrix equation, and check to see if you get the coefficients from both regressions (from a) in one run:

```
X <- cbind(rep(1,20),X)  #adding a column of ones, for the intercept
B <- solve(t(X)%*%X)%*%t(X)%*%Y
```

Solution:
Running the given matrix equation we do get the dame values from a, and b.
Code:

```
> X <- cbind(rep(1,20),X)  #adding a column of ones, for the intercept
> B <- solve(t(X)%*%X)%*%t(X)%*%Y
> B
              y1           y2
   1947.36929408 925.51452032
x1    6.90671653   3.81655963
x2    0.77169586   1.20569180
x3   -0.01248928  -0.04226951
```