STAT 6348
Applied Multivariate Analysis
Fall 2022
Project 2

This project is individual work. So do not consult with anybody in or out of class. You can ask me questions.

Sign on this page below and attach with your project. You project will not be graded without it.

This project is entirely my work. I have not discussed about this project with anybody in or out of class. I understand and have complied with the academic integrity policies written in the *Handbook of Operating Procedures* **of UT Dallas** https://policy.utdallas.edu/utdsp5003.

YOUR NAME ____________________________

DATE ____________________________

YOUR SIGNATURE ____________________________

Ashish Mani Acharya 6348 project 2 axa190076

Q1)To calculate the difference between different kinds of bones I have created a separate variable which calculates the difference between dominant and non-dominant part of radius , humerus and ulna bones. To see if these are different I have created vector mu0=[0 0 0]'. Then I have compared the confidence region of the means contain the vector mu0 = [0 0 0]'.

In other words,

H_0: $\delta' = [\delta_1 \ \ \delta_2]$ = [0 0 0]

H_A: $\delta' = [\delta_1 \ \ \delta_2] \neq$ [0 0 0]

$\alpha = 0.05$

As per my calculations 95 % confidence region which is calculated by Hotelling's T^2 . Hotelling's T^2 is calculated with the formula $T^2 = n((\bar{d}-\delta)^T)(Sd-1)(\bar{d}-\delta)$.

Whereas critical value is calculated by $((n-1)p/(n-p))(Fp,n-p\ (\alpha))$

And we reject the null hypothesis (H_0) if $T^2 \geq$ critical value

As per the R output Hotelling's T^2 is 5.945972 for mu0 = [0 0 0]' whereas the c^2 is 9.978955 for the confidence ellipse. Here sample size is 25 while no of variables of p = 3.

```
> T2 <- 25*t(as.matrix(mean) - mu0) %*% var_cov_inverse %*% (as.matrix(mean) - mu0)
> T2
         [,1]
[1,] 5.945972
>
> c2 = (25-1)*3/(25-3)*qf(0.05, 3, 22, lower.tail = FALSE) # alpha = 0.05
> c2
[1] 9.978955
```

Thus, we conclude, the means of the types of bones (dominant vs non-dominant) do not differ. 95% Bonferroni simultaneous confidence interval is given by, and each one contains zero(0) in it, leading to the same conclusion, the means of the types of bones (dominant vs non-dominant) do not differ.

	radius difference	humerus difference	ulna difference
Upper Limit	0.056652566	0.123579969	0..05054213
Lower Limit	-0.005692566	0.007899969	-0.02942213

Exploratory Data Analysis was performed on these variables namely radius difference , humerus difference and ulna difference using histogram and bivariate scatterplots. We do not see strong normality, but this could be the result of small sample, n = 25 .So, we went ahead with constructing Hotelling's T^2 confidence interval and Bonferroni simultaneous CI as these tests are relatively robust to normality assumption.

QUESTION 2 Begins

Q2.(a). The number of data points in each category is just 20, a small sample. I have created dotplot of each the variables, not much normality. Next, I have created pairwise scatterplots in each category, elliptical shape of points in the graph that shows normality, is not there. Following table shows the optimal lambdas for Box-Cox transformation, and we see there is no transformations that work across all categories, so no transformation is proposed. In the QQ plots, we see the deviations are not extreme, except for RUN in Category2, hence we go along with this data only, no transformation is proposed. And as the question itself states As T^2 tests for means are relatively robust to normality assumption, transformation is not needed unless the violation is extreme.

Q2.(b). H_0: $\mu_1 - \mu_2 = 0$ vs

H_A: $\mu_1 - \mu_2 \neq 0$

where μ_i refer to the population mean for category i .

We know that $T^2 = (\bar{X}_1 - \bar{X}_2 - \mu_{D0})' [((1/n_1) + (1/n_2) S_{pooled}]^{-1} (\bar{X}_1 - \bar{X}_2 - \mu_{D0})$,

where $S_{pooled} = (((n_1-1)/(n_1+n_2-2)) S_1) + (((n_2-1)/(n_1+n_2-2)) S_2)$

similarly critical value $c^2 = (((n_1+n_2-2)p)/(n_1+n_2-p-1)) (F_{p,n1+n2-p-1}(\alpha))$

we reject the H_0 if $T^2 \geq c^2$.

We see from the pairwise category plots that the two categories do not overlap much.

```
> # Hotelling's T square value #
> T2 <- (n1*n2)/(n1+n2)*t(means1-means2) %*% s_pooled_inv%*% (means1-means2)
> T2
         [,1]
[1,] 56.3985
>
> #  test statistic #
> c2<-qf(0.95,p,n1+n2-p-1)*((n1+n2-2))*p/(n1+n2-p-1)
> c2
[1] 9.076508
>
> F_stat<-(n1+n2-p-1)*T2/((n1+n2-2)*p)
> F_stat
          [,1]
[1,] 17.81005
> # p value #
> p_value<-1-pf(F_stat_1,p,n1+n2-p-1)
> p_value
              [,1]
[1,] 2.960536e-07
> 
```

We performed Hotelling's T^2 test, which resulted in the test statistic being 56.3985 against the critical value being 9.076508. also p-value = $2.960536e^{-7}$ which is less than significance level of .05 . Hence we reject null hypothesis H_0:. We have assumed equal covariance matrices, and the assumption seems reasonable.

Q2 C) Pairwise graphs across all categories has been created. We can see there is an overlap of few values of CAT2 and CAT3.MANOVA to test whether the 3 continuous variables are similar or different across categories has been conducted. H_0: mean difference across the categories equals zero

H_A: at least one of mean differences o is not equal to zero

In other words;

$H0$: $\tau_1 = \tau_2 = \tau_3 = 0$

$H1$: At least one $\tau_1 \neq 0$

Here mean differences is represented by Greek letter tau (τ)

We have significance level or α at .05 or 5 %

Source of variation	SSCP Matrix	Degrees of freedom
Treatments B	[4709.2 15561.70 977.2 15561.7 51696.63 3864.0 977.2 3864.00 1681.6]	g-1 3-1=2
Residuals W	[1159.8 5798.80 490.30 5798.8 86884.35 8648.15 490.3 8648.15 43970.05]	$\sum^{g}_{\ell=1} n_\ell - g$ 20(3)-3=57
Total W+B	[5869.0 21360.50 1467.50 21360.5 138580.98 12512.15 1467.5 12512.15 45651.65]	$\sum^{g}_{\ell=1} n_\ell - 1$ 20(3)-1=59

Test statistic $= -(n - 1 - ((p+g)/2))^*$ ($\ln\Lambda^*$) where Lambda star or ($\ln\Lambda^*$) =(W)/(B+W)

Ashish Mani Acharya 6348 project 2 axa190076

We have lambda star as .1823719

```
> Wilkslambda_or_lambdastar<-det(W)/det(T)
> Wilkslambda_or_lambdastar
[1] 0.1823719
>
> test_statistic<--(n-1-(p+g)/2)*(log(Wilkslambda_or_lambdastar))
> test_statistic
[1] 95.2956
>
> critical_value<-qchisq(.95,p*(g-1))
> critical_value
[1] 12.59159
>
> p_value<-pchisq(qchisq(.95,p*(g-1)),4,lower.tail=FALSE)
> p_value
[1] 0.01345377
> |
```

we have the value of the test statistic is approximately 95.2956, and that the p-value is 0.01345377.

Hence the conclusion is we reject H_0 at .05 or 5 % significance level as test statistic> p-value . Hence, we can conclude that at least one $\tau_1 \neq 0$ is available, meaning that the treatment differences do exist .

For 95% simultaneous confidence intervals for differences in mean components ,the formula we have is;

$$\tau_{ki} - \tau_{\ell i} \quad \text{belongs to} \quad \bar{x}_{ki} - \bar{x}_{\ell i} \pm t_{n-g}\left(\frac{\alpha}{pg(g-1)}\right)\sqrt{\frac{w_{ii}}{n-g}\left(\frac{1}{n_k}+\frac{1}{n_\ell}\right)}$$

By using that formula in R calculations and as per the resulting R output attached in the output section, we have 95% simultaneous confidence intervals for differences in mean components as follows:-

$\tau_{11} - \tau_{21}$ belongs to (6.888353,15.11165) (95% simultaneous CI tau1-tau2 for swim variable)

$\tau_{12} - \tau_{22}$ belongs to (−3.737301,67.4373) (95% simultaneous CI tau1-tau2 for bike variable)

$\tau_{13} - \tau_{23}$ belongs to (−33.51645,17.11645) (95% simultaneous CI tau1-tau2 for run variable)

$\tau_{11} - \tau_{31}$ belongs to (17.58835,25.81165) (95% simultaneous CI tau1-tau3 for swim variable)

$\tau_{12} - \tau_{32}$ belongs to (36.1627,107.3373) (95% simultaneous CI tau1-tau3 for bike variable)

$\tau_{13} - \tau_{33}$ belongs to (−20.71645,29.91645) (95% simultaneous CI tau1-tau3 for run variable)

$\tau_{21} - \tau_{31}$ belongs to (6.588353,14.81165) (95% simultaneous CI tau2-tau3 for swim variable)

$\tau_{22} - \tau_{32}$ belongs to (4.312699,75.4873) (95% simultaneous CI tau2-tau3 for bike variable)

$\tau_{23} - \tau_{33}$ belongs to (−12.51645,38.11645) (95% simultaneous CI tau2-tau3 for run variable)

Interpretation of scatterplot of all 3 categories:-

Age seems to affect swim and bike but not so much run. For both swim and run age group 1 trends to scatter on top right of the graph indicating that age group 1 has largest value. Similarly, Age group 3 is scattering on the bottom left of the graph indicating that it has smallest value while age group seems to be right in the middle.

Q2 D) Here H_0: $\gamma 11 = \gamma 12 = \gamma 13 = 0$

H_A: $\gamma 11 = \gamma 12 = \gamma 13 \neq 0$

$\alpha = 0.05$

Ashish Mani Acharya 6348 project 2 axa190076

```
                                                       Df   Wilks approx F num Df den Df    Pr(>F)    
as.factor(triathlon$CATEGORY)                          2 0.12952  30.8289      6    104 < 2.2e-16 ***
as.factor(triathlon$GENDER)                            1 0.90547   1.8095      3     52 0.1568890    
as.factor(triathlon$CATEGORY):as.factor(triathlon$GENDER)  2 0.62497   4.5923      6    104 0.0003562 ***
Residuals                                             54                                            
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Here we have p-value for age category , p-value for gender and p-value for interaction as 2.2e $^{-16}$,.1569 and .0003562.

Only P value for gender which is .1569 is greater than significance level of .05 .

Conclusion:-

the participant's age category and interaction have significant effects, while the participant's gender does not have any significant effects. we can also notice that there is an interaction between the two factors, age category and gender.

Scatter Plot interpretation:-

We can see that for Swim and Bike relation between age and swim & age and bike seems to be dependent upon gender on both cases. Scatterplot also indicates male participants took longer time than female in age 1 and age 2 categories for swim and bike. Similarly for run , scatterplot shows that male participants at the age group1 run slower than females at same age group .Similarly it also shows that male participants at age group 2 and age group 3 run faster than females at same categories.

Interaction Plots:-

Note:- Here Cat 1 Cat 2 and Cat 3 means first , second and third age category respectively.

Interaction plots for Swim and Bike were similar. In both cases lines are not parallel .We can see that for these two variables relation between age and swim & age and bike depend on gender on both cases. We can also see that the lines meet between Cat 2 and Cat 3. Meaning male participants took longer time than female in age 1 and age 2 categories. While males were faster in third category.

Interaction plot of Run is different though. Here even though similar to previous plots lines are not parallel, lines have an upward slope from Cat 1 to Cat 2 but takes a dive from Cat 2 towards Cat3. Hence even though this also means relation between age and depend on gender(like previous plots) , it also indicates that male participants at the Cat 1 run slower than females at Cat 1 .Similarly it also shows that male participants at Cat 2 and Cat 3 run faster than females at same categories.

For last part of the question :- In triathlon data, we saw the effect of interaction between the two factors, gender and age category, is significant as the p-value (<.001) was very small. This leads us to reject the null hypothesis and conclude that there is interaction between the two factors, gender and age category and hence both have an effect. Univariate ANOVA tests has been conducted on the data and at 5% significance level .

H_0: There is no interaction between factors
H_1: There is a significant interaction between factors
We can see that the effect of interaction between the two factors, gender and age category, is significant for SWIM and BIKE, however the exception was RUN . Meaning effect of interaction between the two factors, gender and age category was not significant for RUN. Following output has been attached to justify my conclusion about the interaction.

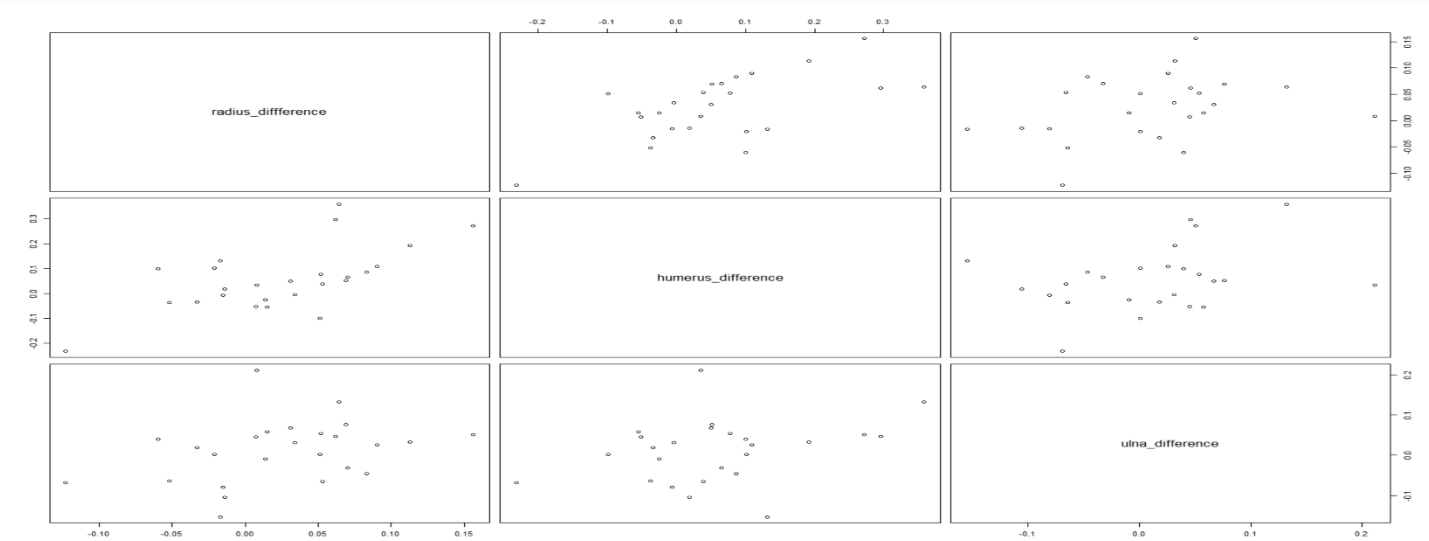
```
> summary(tri3anova)
                             Df Sum Sq Mean Sq F value   Pr(>F)    
as.factor(triathlon$GENDER)  1     24    24.1   0.675    0.415    
as.factor(tri$CATEGORY)      2   3850  1924.9  54.026 8.51e-14 ***
Residuals                   56   1995    35.6                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #for bike#
> summary(tri3anova)
                             Df Sum Sq Mean Sq F value   Pr(>F)    
as.factor(triathlon$GENDER)  1     24    24.1   0.675    0.415    
as.factor(tri$CATEGORY)      2   3850  1924.9  54.026 8.51e-14 ***
Residuals                   56   1995    35.6                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #for run #
> summary(tri3anova)
                             Df Sum Sq Mean Sq F value   Pr(>F)    
as.factor(triathlon$GENDER)  1     24    24.1   0.675    0.415    
as.factor(tri$CATEGORY)      2   3850  1924.9  54.026 8.51e-14 ***
Residuals                   56   1995    35.6                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> 
```

Ashish Mani Acharya 6348 project 2 axa190076

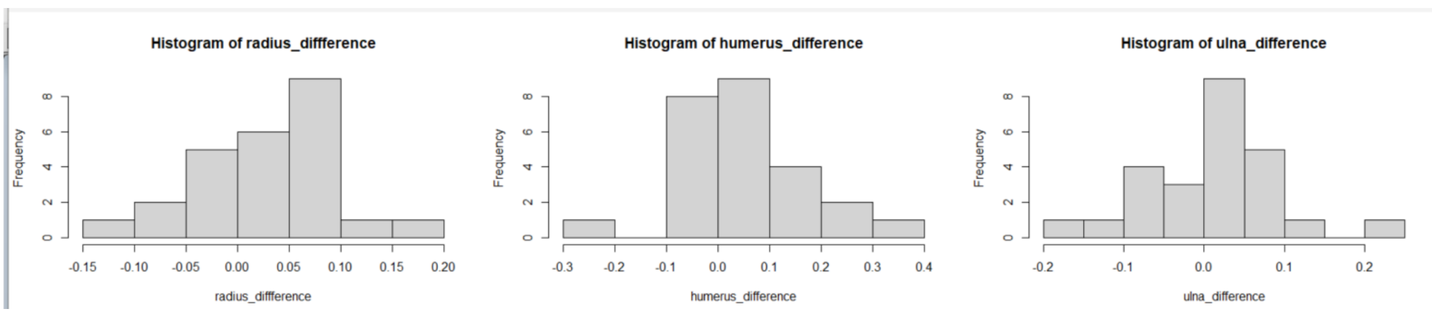
OUTPUTS:-

Q1)

Scatterplot for bone differences:-



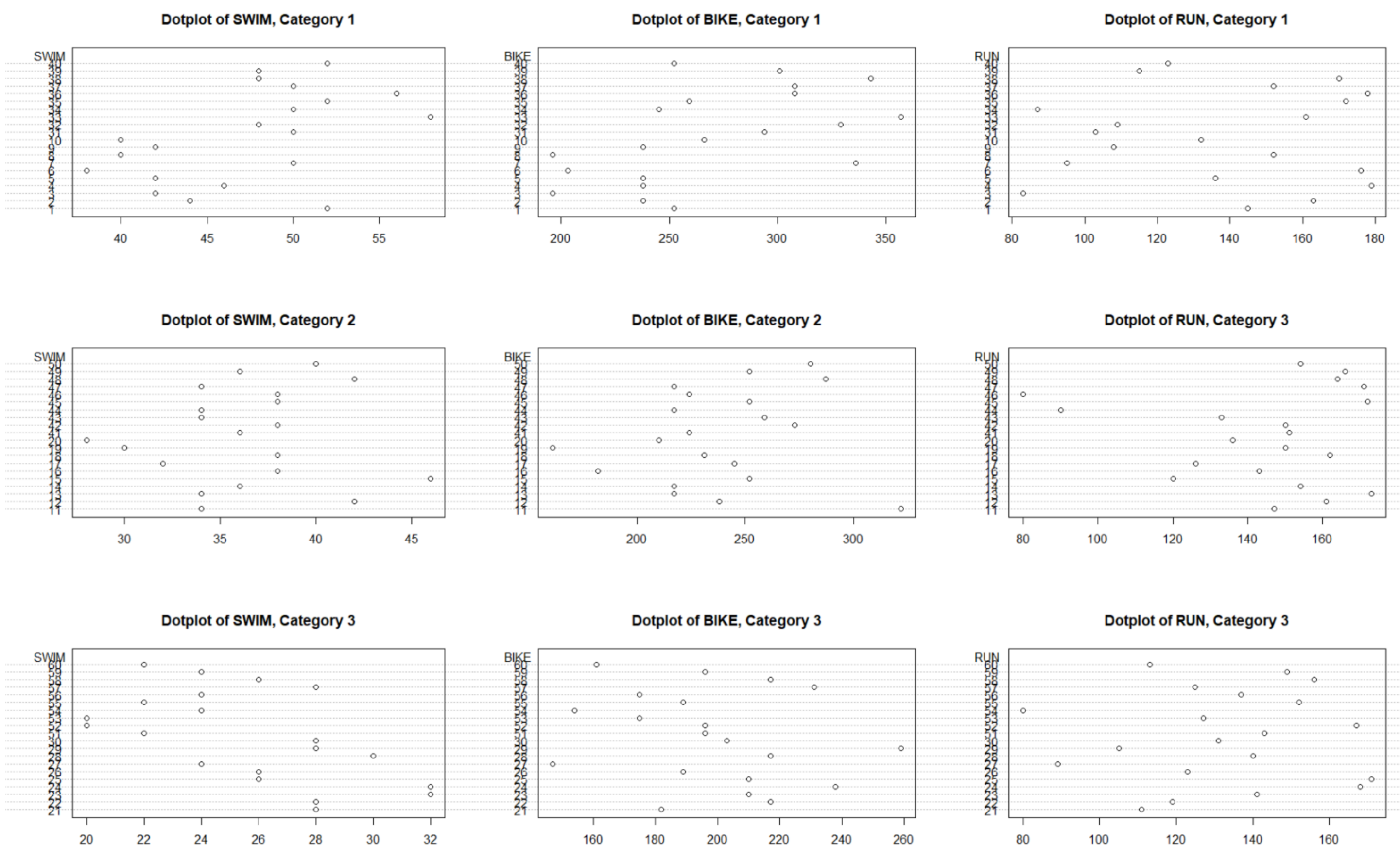
Histogram for differences:-

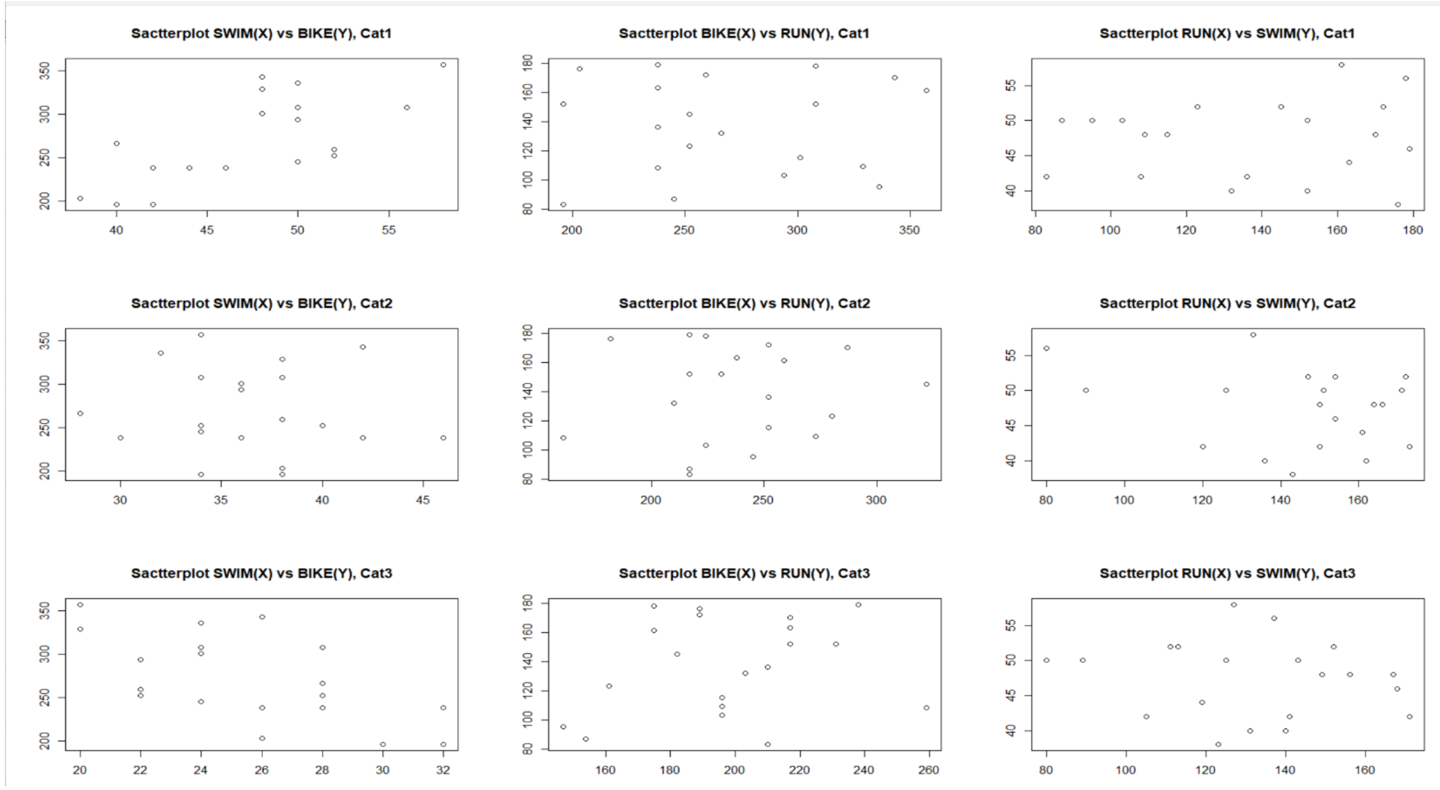


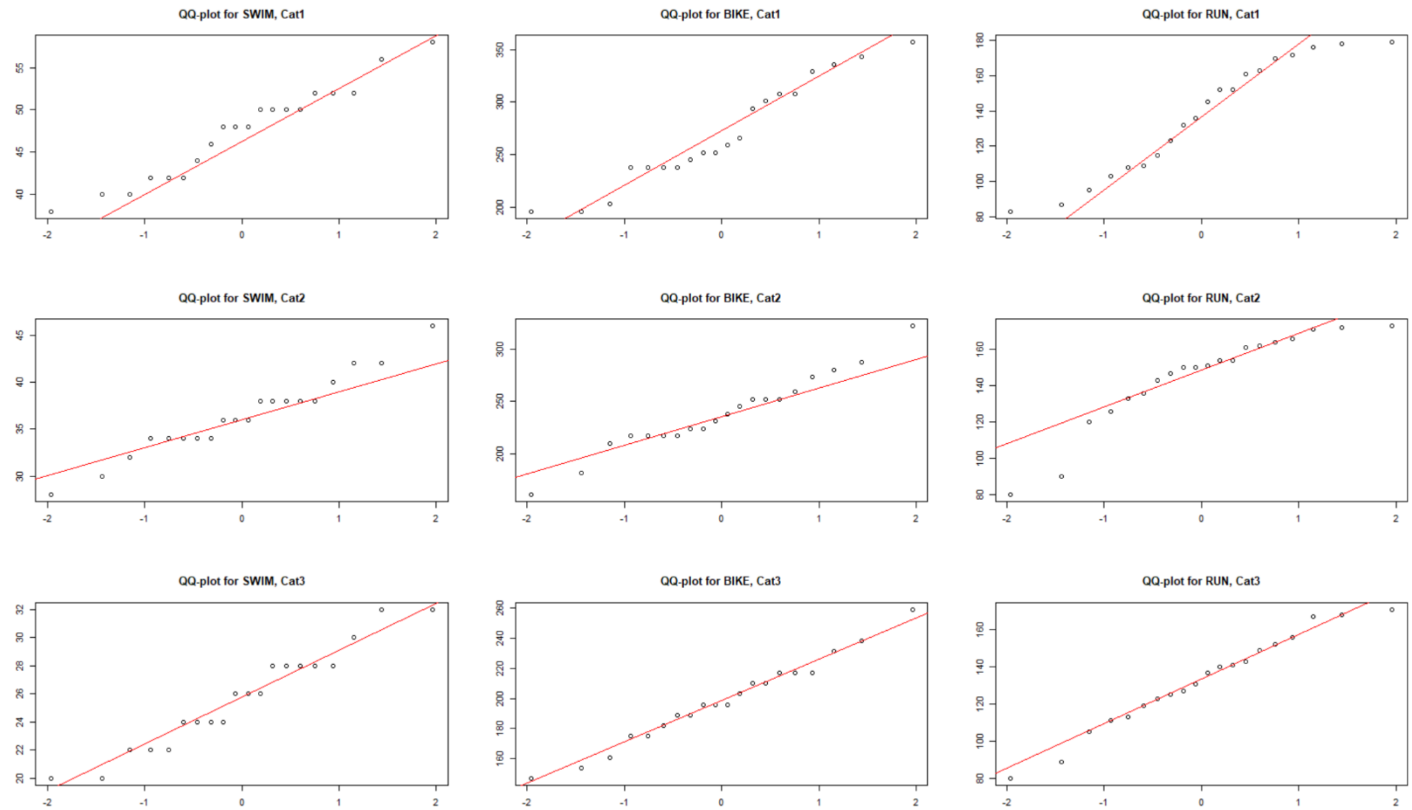

```
>
> #95% Bonferroni simultaneous confidence interval #
> radius_diffference_upper= mean[1] + sqrt(var_cov[1,1]/25)*qt(0.0083, 24, lower.tail = FALSE) #upper limit
> radius_diffference_upper
radius_diffference
        0.05665257
>
> humerus_difference_upper = mean[2] + sqrt(var_cov[2,2]/25)*qt(0.0083, 24, lower.tail = FALSE) #upper limit
> humerus_difference_upper
humerus_difference
           0.12358
>
> ulna_difference_upper = mean[3] + sqrt(var_cov[3,3]/25)*qt(0.0083, 24, lower.tail = FALSE) #upper limit
> ulna_difference_upper
ulna_difference
     0.05054213
>
> radius_diffference_lower = mean[1] - sqrt(var_cov[1,1]/25)*qt(0.0083, 24, lower.tail = FALSE) #lower limit
> radius_diffference_lower
radius_diffference
       -0.005692566
>
> humerus_difference_lower = mean[2] - sqrt(var_cov[2,2]/25)*qt(0.0083, 24, lower.tail = FALSE) #lower limit
> humerus_difference_lower
humerus_difference
       -0.007899969
>
> ulna_difference_lower = mean[3] - sqrt(var_cov[3,3]/25)*qt(0.0083, 24, lower.tail = FALSE) #lower limit
> ulna_difference_lower
ulna_difference
     -0.02942213
>
```

Q 2 Outputs

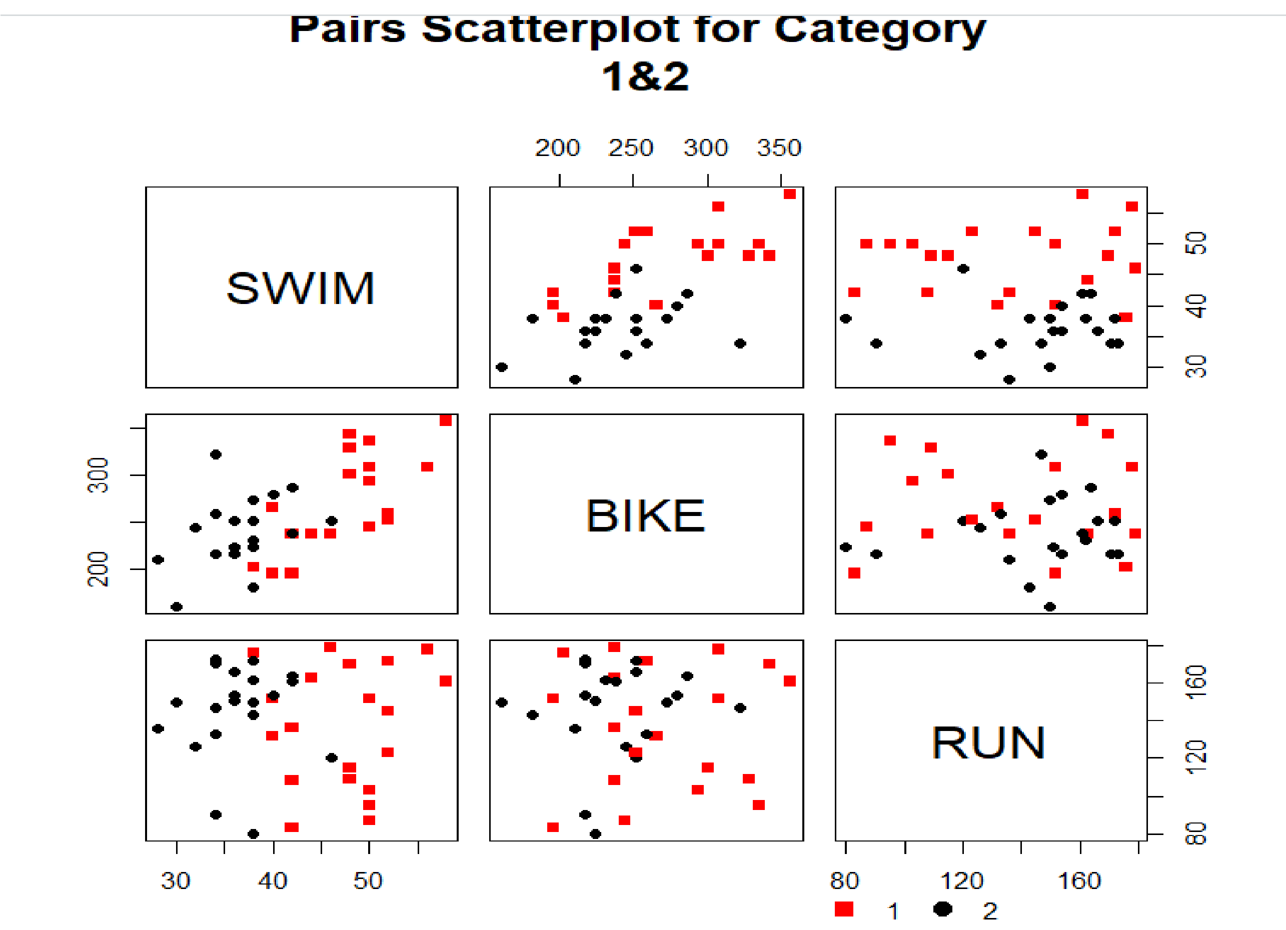
Q2A)






QQ-plot for SWIM, Cat1
QQ-plot for BIKE, Cat1
QQ-plot for RUN, Cat1
QQ-plot for SWIM, Cat2
QQ-plot for BIKE, Cat2
QQ-plot for RUN, Cat2
QQ-plot for SWIM, Cat3
QQ-plot for BIKE, Cat3
QQ-plot for RUN, Cat3

Q2b)

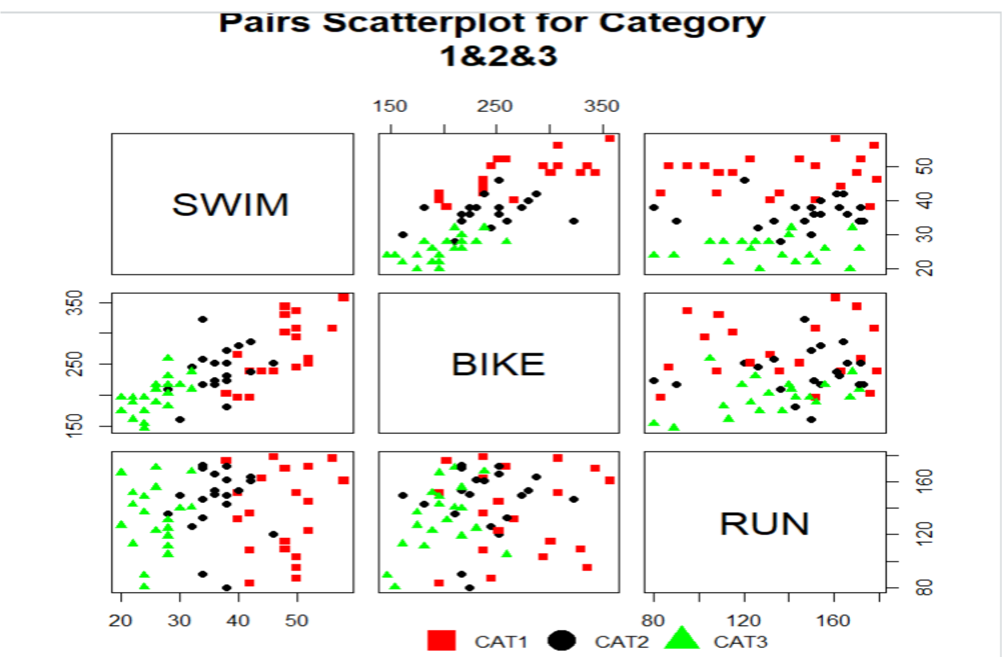

Pairs Scatterplot for Category 1&2
SWIM
BIKE
RUN
200 250 300 350
30 40 50
80 120 160
1
2

Q 2 C

Manova table

```
>
> B<-manova$`as.factor(triathlon_qc$CATEGORY)`
> B
        [,1]     [,2]   [,3]
[1,]  4709.2 15561.70  977.2
[2,] 15561.7 51696.63 3864.0
[3,]   977.2  3864.00 1681.6
>
> # residuals #
> W<-manova$Residuals
> W
       [,1]     [,2]     [,3]
[1,] 1159.8  5798.80   490.30
[2,] 5798.8 86884.35  8648.15
[3,]  490.3  8648.15 43970.05
>
> # total #
> T<-B+W
> T
        [,1]      [,2]     [,3]
[1,]  5869.0  21360.50  1467.50
[2,] 21360.5 138580.98 12512.15
[3,]  1467.5  12512.15 45651.65
>
```

Scatterplot with all three categories.



Ashish Mani Acharya 6348 project 2 axa190076

95 % simultaneous CI

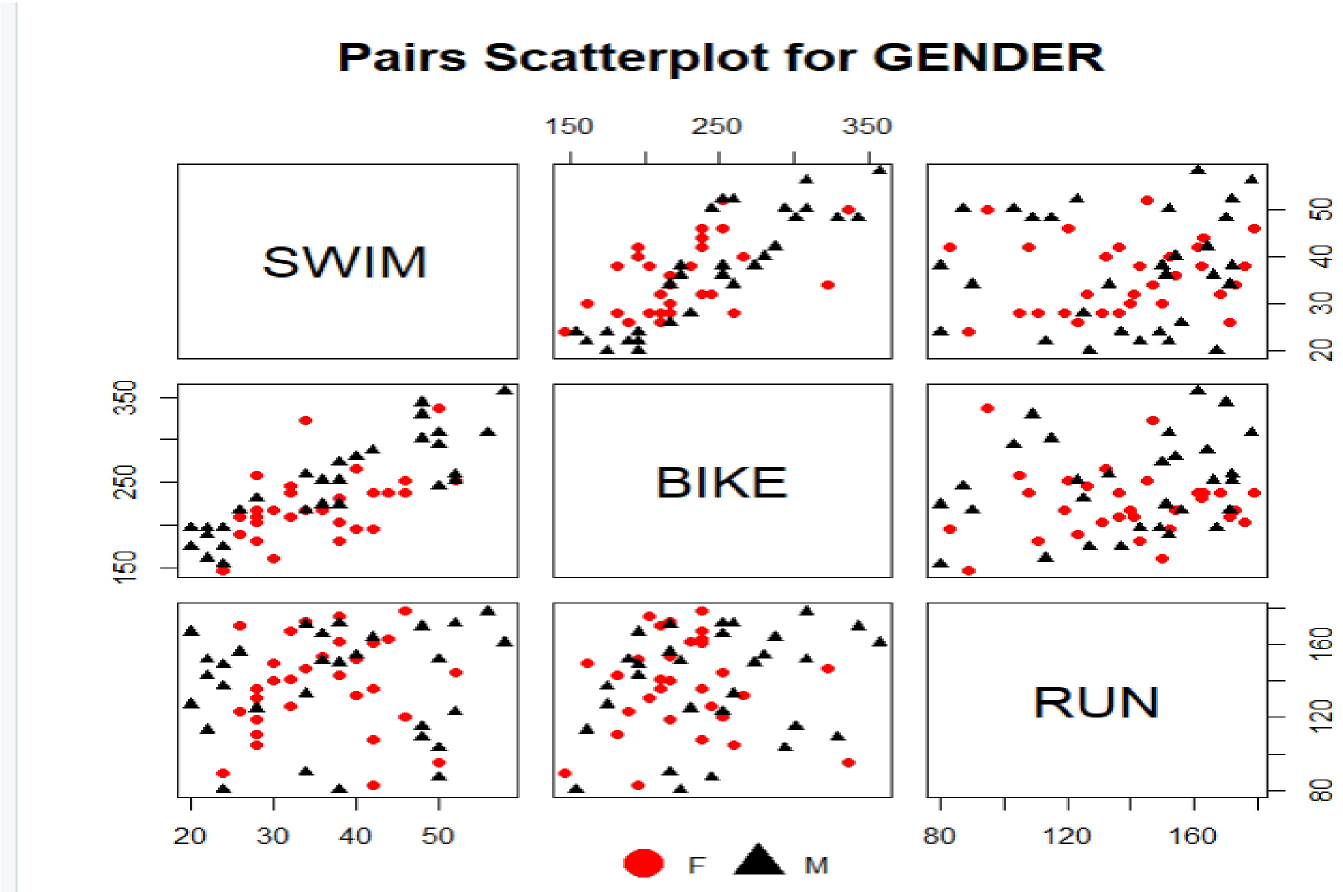
```
> #for swim variable taul-tau2 is #
> lowerlimit_swim_variable_taul_tau2
    SWIM
6.888353
> upperlimit_swim_variable_taul_tau2
    SWIM
15.11165
>
> #for bike  variable taul-tau2 is #
> lowerlimit_bike_variable_taul_tau2
     BIKE
-3.737301
> upperlimit_bike_variable_taul_tau2
   BIKE
67.4373
>
> #for run  variable taul-tau2 is #
> lowerlimit_run_variable_taul_tau2
      RUN
-33.51645
> upperlimit_run_variable_taul_tau2
     RUN
17.11645
```

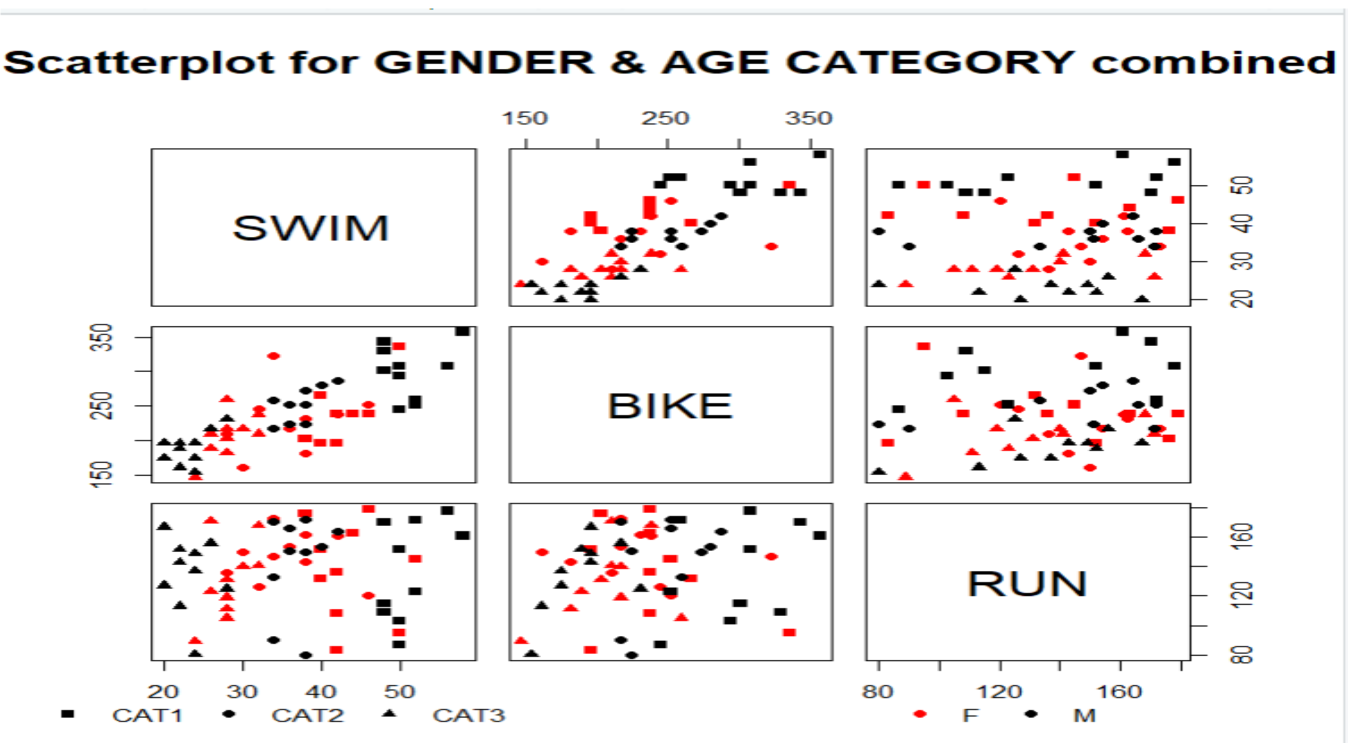
```
> #for swim variable taul-tau3 is #
> lowerlimit_swim_variable_taul_tau3
    SWIM
17.58835
> upperlimit_swim_variable_taul_tau3
    SWIM
25.81165
>
> #for bike  variable taul-tau3 is #
> lowerlimit_bike_variable_taul_tau3
   BIKE
36.1627
> upperlimit_bike_variable_taul_tau3
    BIKE
107.3373
>
> #for run  variable taul-tau3 is #
> lowerlimit_run_variable_taul_tau3
      RUN
-20.71645
> upperlimit_run_variable_taul_tau3
     RUN
29.91645
```

```
> #for swim variable tau2-tau3 is #
> lowerlimit_swim_variable_tau2_tau3
    SWIM
6.588353
> upperlimit_swim_variable_tau2_tau3
    SWIM
14.81165
>
> #for bike  variable taul-tau3 is #
> lowerlimit_bike_variable_tau2_tau3
    BIKE
4.312699
> upperlimit_bike_variable_tau2_tau3
   BIKE
75.4873
>
> #for run  variable taul-tau3 is #
> lowerlimit_run_variable_tau2_tau3
     RUN
-12.51645
> upperlimit_run_variable_tau2_tau3
    RUN
38.11645
```

Ashish Mani Acharya 6348 project 2 axa190076

Q 2 D outputs





Ashish Mani Acharya 6348 project 2 axa190076

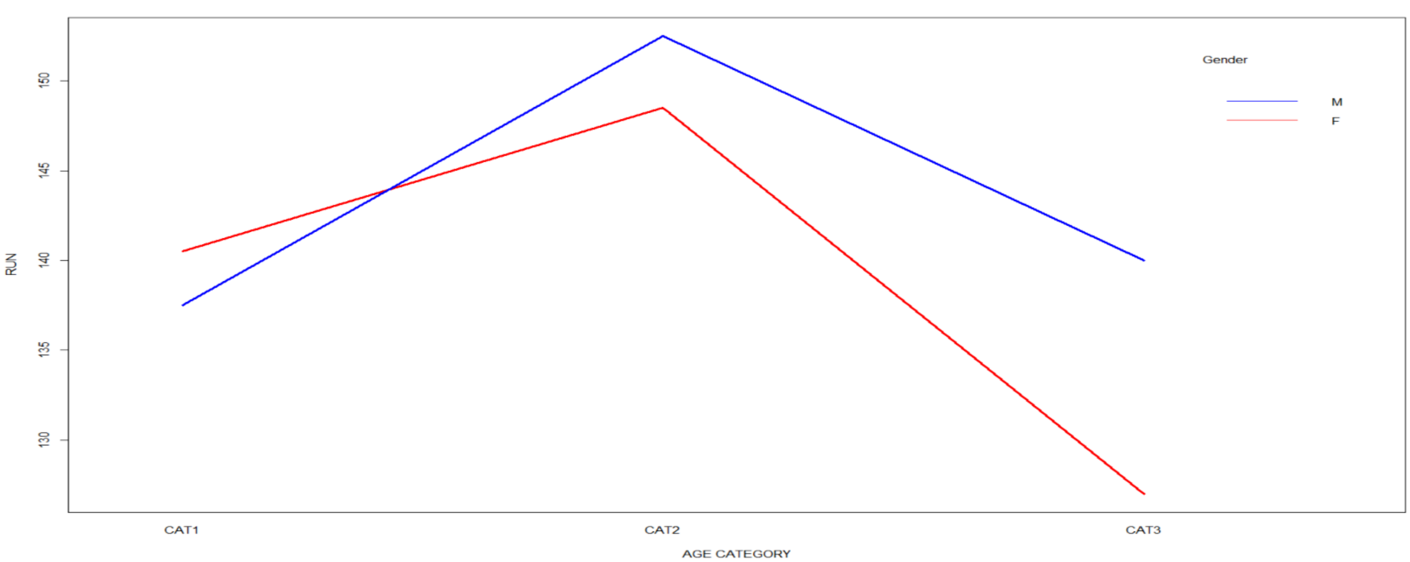
MANOVA to test whether category, gender, and their interaction have significant effects.

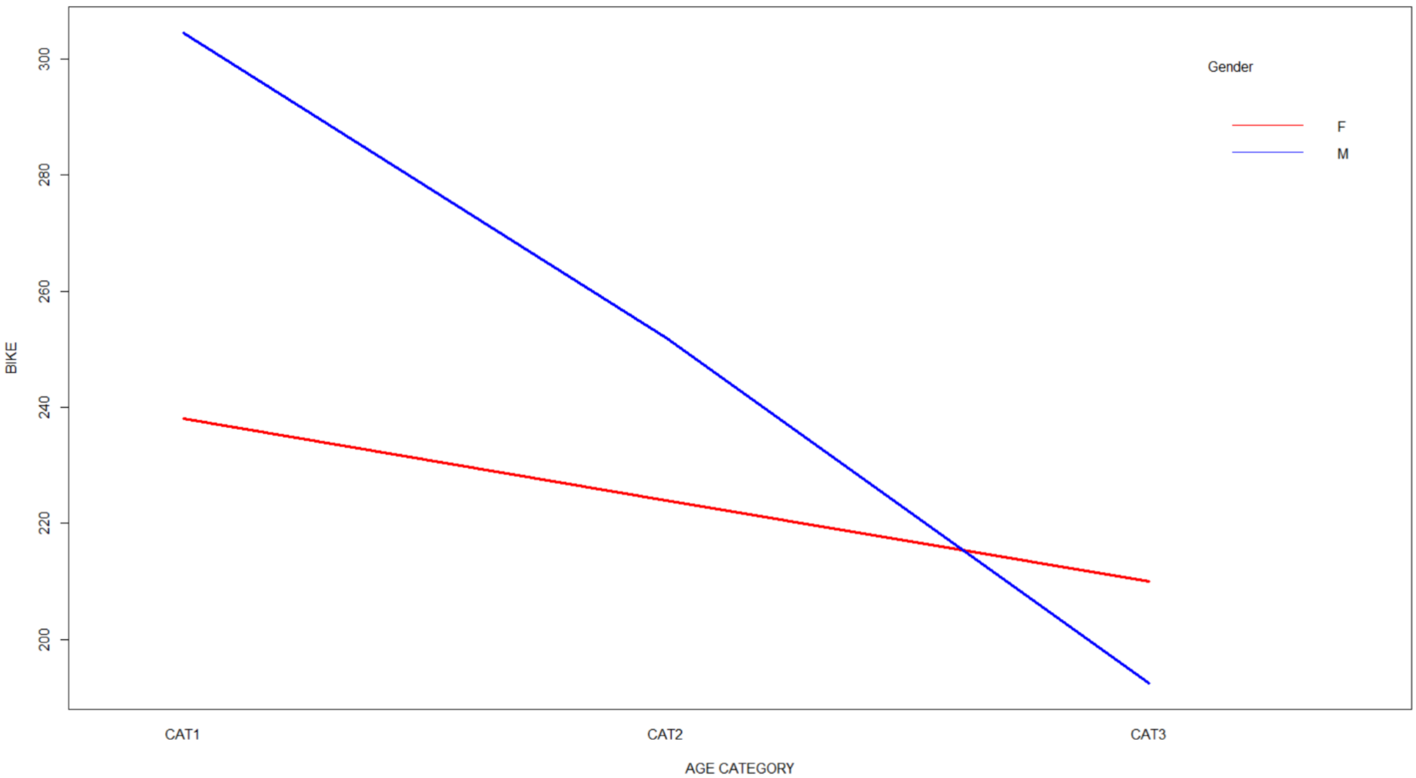
```
                                                        Df   Wilks approx F num Df den Df    Pr(>F)    
as.factor(triathlon$CATEGORY)                            2 0.12952  30.8289      6    104 < 2.2e-16 ***
as.factor(triathlon$GENDER)                              1 0.90547   1.8095      3     52 0.1568890    
as.factor(triathlon$CATEGORY):as.factor(triathlon$GENDER) 2 0.62497   4.5923      6    104 0.0003562 ***
Residuals                                               54                                            
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

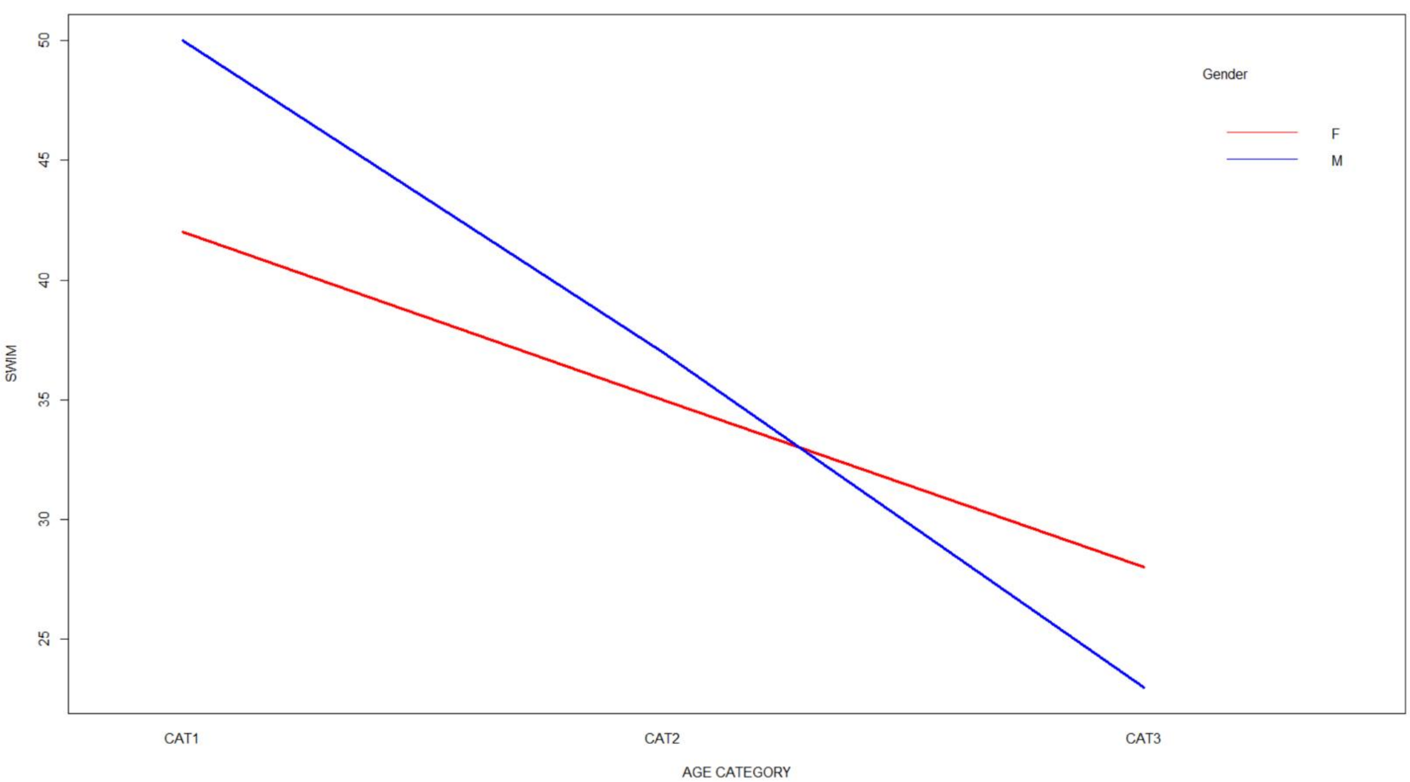
Univariate ANOVA models to find out if interaction is significant for some variables.

```
>
> tri3anova = aov(triathlon[,3]~as.factor(triathlon$GENDER)*as.factor(tri$CATEGORY))
> summary(tri3anova)
                            Df Sum Sq Mean Sq F value   Pr(>F)    
as.factor(triathlon$GENDER)  1     24    24.1   0.675    0.415    
as.factor(tri$CATEGORY)      2   3850  1924.9  54.026 8.51e-14 ***
Residuals                   56   1995    35.6                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> tri4anova = aov(triathlon[,4]~as.factor(triathlon$GENDER)*as.factor(triathlon$CATEGORY))
> summary(tri4anova)
                                                          Df Sum Sq Mean Sq F value   Pr(>F)    
as.factor(triathlon$GENDER)                                1   6469    6469   5.348  0.02459 *  
as.factor(triathlon$CATEGORY)                              2  51697   25848  21.368 1.46e-07 ***
as.factor(triathlon$GENDER):as.factor(triathlon$CATEGORY)  2  15094    7547   6.239  0.00365 ** 
Residuals                                                 54  65322    1210                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> tri5anova = aov(triathlon[,5]~as.factor(triathlon$GENDER)*as.factor(triathlon$CATEGORY))
> summary(tri5anova)
                                                          Df Sum Sq Mean Sq F value Pr(>F)
as.factor(triathlon$GENDER)                                1      2     2.0   0.002  0.960
as.factor(triathlon$CATEGORY)                              2   1682   840.8   1.038  0.361
as.factor(triathlon$GENDER):as.factor(triathlon$CATEGORY)  2    212   106.1   0.131  0.878
Residuals                                                 54  43756   810.3               
```

Interaction plots




BIKE
300
280
260
240
220
200
CAT1
CAT2
CAT3
AGE CATEGORY
Gender
F
M


SWIM
50
45
40
35
30
25
CAT1
CAT2
CAT3
AGE CATEGORY
Gender
F
M

Ashish Mani Acharya 6348 project 2 axa190076

R Codes :-

```
#####################Q 1 ##########################################

bones <- read.table("C:/Users/alexk/Downloads/bones.dat", header = FALSE, col.names =
                    c("radius_dominant", "radius_non_dominant", "humerus_dominant", "humerus_non_dominant",
"ulna_dominant", "ulna"))

View(bones)


# getting differences #

radius_diffference= bones$radius_dominant - bones$radius_non_dominant

humerus_difference = bones$humerus_dominant - bones$humerus_non_dominant

ulna_difference = bones$ulna_dominant - bones$ulna

bones2 <- cbind(radius_diffference, humerus_difference, ulna_difference)

View(bones2)


#multivariate data (p=3) for paired samples created#


par(mfrow=c(3,3))

hist(radius_diffference)

hist(humerus_difference)

hist(ulna_difference)


pairs(bones2) #Exploratory data analysis


mean = colMeans(bones2) #sample mean

var_cov = cov(bones2)#sample variance-covariance matrix S#

var_cov_inverse = solve(var_cov)

var_cov_inverse #inverse(S)#

mu0 <- matrix(c(0, 0, 0), nrow = 3)


#Hotelling's T square #

T2 <- 25*t(as.matrix(mean) - mu0) %*% var_cov_inverse %*% (as.matrix(mean) - mu0)

T2
```

Ashish Mani Acharya 6348 project 2 axa190076

```
#critical value #

c2 = (25-1)*3/(25-3)*qf(0.05, 3, 22, lower.tail = FALSE) # alpha = 0.05

c2


#95% Bonferroni simultaneous confidence interval #

radius_diffference_upper= mean[1] + sqrt(var_cov[1,1]/25)*qt(0.0083, 24, lower.tail = FALSE) #upper limit

radius_diffference_upper


humerus_difference_upper = mean[2] + sqrt(var_cov[2,2]/25)*qt(0.0083, 24, lower.tail = FALSE) #upper limit

humerus_difference_upper


ulna_difference_upper = mean[3] + sqrt(var_cov[3,3]/25)*qt(0.0083, 24, lower.tail = FALSE) #upper limit

ulna_difference_upper


radius_diffference_lower = mean[1] - sqrt(var_cov[1,1]/25)*qt(0.0083, 24, lower.tail = FALSE) #lower limit

radius_diffference_lower


humerus_difference_lower = mean[2] - sqrt(var_cov[2,2]/25)*qt(0.0083, 24, lower.tail = FALSE) #lower limit

humerus_difference_lower


ulna_difference_lower = mean[3] - sqrt(var_cov[3,3]/25)*qt(0.0083, 24, lower.tail = FALSE) #lower limit

ulna_difference_lower


triathlon <- read.csv("C:/Users/alexk/Downloads/triathlon.csv", header = TRUE)

View(triathlon)

##########################Q 2 A #####################################

# subsetting data as per age categories #

triathlon_cat1 <- subset(triathlon, CATEGORY == "CAT1")

triathlon_cat2 <- subset(triathlon, CATEGORY == "CAT2")

triathlon_cat3 <- subset(triathlon, CATEGORY == "CAT3")
```

Ashish Mani Acharya 6348 project 2 axa190076

```
#creatinf dot charts #
par(mfrow=c(3,3))
dotchart(as.matrix(triathlon_cat1[3]), main="Dotplot of SWIM, Category 1")
dotchart(as.matrix(triathlon_cat1[4]), main="Dotplot of BIKE, Category 1")
dotchart(as.matrix(triathlon_cat1[5]), main="Dotplot of RUN, Category 1")
dotchart(as.matrix(triathlon_cat2[3]), main="Dotplot of SWIM, Category 2")
dotchart(as.matrix(triathlon_cat2[4]), main="Dotplot of BIKE, Category 2")
dotchart(as.matrix(triathlon_cat2[5]), main="Dotplot of RUN, Category 3")
dotchart(as.matrix(triathlon_cat3[3]), main="Dotplot of SWIM, Category 3")
dotchart(as.matrix(triathlon_cat3[4]), main="Dotplot of BIKE, Category 3")
dotchart(as.matrix(triathlon_cat3[5]), main="Dotplot of RUN, Category 3")


#creating scatterplots #
par(mfrow=c(3,3))
plot(as.matrix(triathlon_cat1[3]), as.matrix(triathlon_cat1[4]), main="Sactterplot SWIM(X) vs BIKE(Y), Cat1",
   xlab="", ylab="")
plot(as.matrix(triathlon_cat1[4]), as.matrix(triathlon_cat1[5]), main="Sactterplot BIKE(X) vs RUN(Y), Cat1",
   xlab="", ylab="")
plot(as.matrix(triathlon_cat1[5]), as.matrix(triathlon_cat1[3]), main="Sactterplot RUN(X) vs SWIM(Y), Cat1",
   xlab="", ylab="")
plot(as.matrix(triathlon_cat2[3]), as.matrix(triathlon_cat1[4]), main="Sactterplot SWIM(X) vs BIKE(Y), Cat2",
   xlab="", ylab="")
plot(as.matrix(triathlon_cat2[4]), as.matrix(triathlon_cat1[5]), main="Sactterplot BIKE(X) vs RUN(Y), Cat2",
   xlab="", ylab="")
plot(as.matrix(triathlon_cat2[5]), as.matrix(triathlon_cat1[3]), main="Sactterplot RUN(X) vs SWIM(Y), Cat2",
   xlab="", ylab="")
plot(as.matrix(triathlon_cat3[3]), as.matrix(triathlon_cat1[4]), main="Sactterplot SWIM(X) vs BIKE(Y), Cat3",
   xlab="", ylab="")
plot(as.matrix(triathlon_cat3[4]), as.matrix(triathlon_cat1[5]), main="Sactterplot BIKE(X) vs RUN(Y), Cat3",
   xlab="", ylab="")
```

Ashish Mani Acharya 6348 project 2 axa190076

```
plot(as.matrix(triathlon_cat3[5]), as.matrix(triathlon_cat1[3]), main="Sactterplot RUN(X) vs SWIM(Y), Cat3",
    xlab="", ylab="")


# testing normality #
par(mfrow=c(3,3))
qqnorm(as.matrix(triathlon_cat1[3]), main="QQ-plot for SWIM, Cat1", xlab="", ylab="")
qqline(as.matrix(triathlon_cat1[3]), col = "red")
qqnorm(as.matrix(triathlon_cat1[4]), main="QQ-plot for BIKE, Cat1", xlab="", ylab="")
qqline(as.matrix(triathlon_cat1[4]), col = "red")
qqnorm(as.matrix(triathlon_cat1[5]), main="QQ-plot for RUN, Cat1", xlab="", ylab="")
qqline(as.matrix(triathlon_cat1[5]), col = "red")

qqnorm(as.matrix(triathlon_cat2[3]), main="QQ-plot for SWIM, Cat2", xlab="", ylab="")
qqline(as.matrix(triathlon_cat2[3]), col = "red")
qqnorm(as.matrix(triathlon_cat2[4]), main="QQ-plot for BIKE, Cat2", xlab="", ylab="")
qqline(as.matrix(triathlon_cat2[4]), col = "red")
qqnorm(as.matrix(triathlon_cat2[5]), main="QQ-plot for RUN, Cat2", xlab="", ylab="")
qqline(as.matrix(triathlon_cat2[5]), col = "red")

qqnorm(as.matrix(triathlon_cat3[3]), main="QQ-plot for SWIM, Cat3", xlab="", ylab="")
qqline(as.matrix(triathlon_cat3[3]), col = "red")
qqnorm(as.matrix(triathlon_cat3[4]), main="QQ-plot for BIKE, Cat3", xlab="", ylab="")
qqline(as.matrix(triathlon_cat3[4]), col = "red")
qqnorm(as.matrix(triathlon_cat3[5]), main="QQ-plot for RUN, Cat3", xlab="", ylab="")
qqline(as.matrix(triathlon_cat3[5]), col = "red")

#power transformation #
library(car)
#cat 1 #
powerTransform(as.matrix(triathlon_cat1[3]))
powerTransform(as.matrix(triathlon_cat1[4]))
```

```
powerTransform(as.matrix(triathlon_cat1[5]))


#cat 2 #

powerTransform(as.matrix(triathlon_cat2[3]))

powerTransform(as.matrix(triathlon_cat2[4]))

powerTransform(as.matrix(triathlon_cat2[5]))


#cat 3 #

powerTransform(as.matrix(triathlon_cat3[3]))

powerTransform(as.matrix(triathlon_cat3[4]))

powerTransform(as.matrix(triathlon_cat3[5]))


########################Q 2 B ####################################################

# subsetting triathlon data by removing category 3 from the data #

triathlon2 <- triathlon

triathlon2$CATEGORY = replace(triathlon2$CATEGORY, CATEGORY == "CAT1", 1)

triathlon2$CATEGORY = replace(triathlon2$CATEGORY, CATEGORY == "CAT2", 2)

triathlon2$CATEGORY = replace(triathlon2$CATEGORY, CATEGORY == "CAT3", 3)

triathlon_b <- subset(triathlon2, CATEGORY != 3)

triathlon_c<-subset(triathlon2, CATEGORY ==3)

attach(triathlon_b)

View(triathlon_b)


# plotting new scatterplot with category  1 and 2 only #


mycols1<-c("red","black")


dcol1<-factor(triathlon_b$CATEGORY)


plot.new()

pairs(triathlon_b[,3:5], pch = c(15:16)[as.numeric(dcol1)], cex = 1, col = mycols1[as.numeric(dcol1)] , main = "Pairs
Scatterplot for Category
```

```
1&2")
legend("bottomleft", col = mycols1, legend = levels(dcol1), pch = c(15:16),
    xpd = NA, ncol = 3, bty = "n",inset = c(-0.2,-.25), pt.cex = 1.5)

par(xpd = TRUE)

# determing p variables for future calculations #
p<-dim(triathlon[,3:5])[2]
p

# determining n number of subjects for future calculations #
n1<-dim(triathlon_cat1)[1]
n1
n2<-dim(triathlon_cat2)[1]
n2
n3<-dim(triathlon_cat2)[1]
n3

# getting col means for respective age categories #
means1<-colMeans(triathlon_cat1[,3:5])
means1

means2<-colMeans(triathlon_cat2[,3:5])
means2

means3<-colMeans(triathlon_cat3[,3:5])
means3

# getting S matrix for respective age categories #
var1<-cov(triathlon_cat1[,3:5])
var1
```

```
var2<-cov(triathlon_cat2[,3:5])

var2


var3<-cov(triathlon_cat3[,3:5])

var3


# getting S inverse  matrix for respective age categories #

sinv1<-solve(var1)

sinv1


sinv2<-solve(var2)

sinv2


sinv3<-solve(var3)

sinv3


# getting s pooled matrix #

s_pooled<-((n1-1)*var1+(n2-1)*var2)/(n1+n2-2)

s_pooled


s_pooled_inv <- solve(s_pooled)

s_pooled_inv


# Hotelling’s T square value #

T2 <- (n1*n2)/(n1+n2)*t(means1-means2) %*% s_pooled_inv%*% (means1-means2)

T2


#  test statistic #

c2<-qf(0.95,p,n1+n2-p-1)*((n1+n2-2))*p/(n1+n2-p-1)

c2
```

```
F_stat<-(n1+n2-p-1)*T2/((n1+n2-2)*p)

F_stat

# p value #

p_value<-1-pf(F_stat_1,p,n1+n2-p-1)

p_value


####################Q2C######################################

triathlon_qc<-subset(triathlon,select=-c(GENDER))


manova1<-manova(cbind(triathlon_qc$SWIM,triathlon_qc$BIKE,triathlon_qc$RUN)~as.factor(triathlon_qc$CATEGORY))

summary(manova1,test="Wilks")


manova<-summary(manova1,test="Wilks")$SS

manova



g<-3

g

# treatment matrix #

B<-manova$`as.factor(triathlon_qc$CATEGORY)`

B


# residuals #

W<-manova$Residuals

W


# total #

T<-B+W

T
```

```
# total subjects #

n_total=n1+n2+n3

n_total


#lambda star #

Wilkslambda_or_lambdastar<-det(W)/det(T)

Wilkslambda_or_lambdastar


n<-dim(triathlon_qc)[1]

n

p<-dim(triathlon_qc[,2:4])[2]

p

g<-3

g


# test statistic #

test_statistic<--(n-1-(p+g)/2)*(log(Wilkslambda_or_lambdastar))

test_statistic


# critical value #

critical_value<-qchisq(.95,p*(g-1))

critical_value


# p value #

p_value<-pchisq(qchisq(.95,p*(g-1)),4,lower.tail=FALSE)

p_value


# wvalue #

w_value<-(n1-1)*var1+(n2-1)*var2+(n3-1)*var3

w_value
```

Ashish Mani Acharya 6348 project 2 axa190076

```
# qt level #

qtlevel<-qt(1-.05/(p*g*(g-1)),df=n_total-g)

qtlevel


# 95 % simultaneous CI #



#for swim variable tau1-tau2 is #

lowerlimit_swim_variable_tau1_tau2<-(means1[1]-means2[1])-qtlevel*sqrt(w_value[1,1]/(n_total-g)*(1/n1+1/n2))

upperlimit_swim_variable_tau1_tau2<-(means1[1]-means2[1])+qtlevel*sqrt(w_value[1,1]/(n_total-g)*(1/n1+1/n2))



#for bike  variable tau1-tau2 is #

lowerlimit_bike_variable_tau1_tau2<-(means1[2]-means2[2])-qtlevel*sqrt(w_value[2,2]/(n_total-g)*(1/n1+1/n2))

upperlimit_bike_variable_tau1_tau2<-(means1[2]-means2[2])+qtlevel*sqrt(w_value[2,2]/(n_total-g)*(1/n1+1/n2))



#for run  variable tau1-tau2 is #

lowerlimit_run_variable_tau1_tau2<-(means1[3]-means2[3])-qtlevel*sqrt(w_value[3,3]/(n_total-g)*(1/n1+1/n2))

upperlimit_run_variable_tau1_tau2<-(means1[3]-means2[3])+qtlevel*sqrt(w_value[3,3]/(n_total-g)*(1/n1+1/n2))


#for swim variable tau1-tau3 is #

lowerlimit_swim_variable_tau1_tau3<-(means1[1]-means3[1])-qtlevel*sqrt(w_value[1,1]/(n_total-g)*(1/n1+1/n3))

upperlimit_swim_variable_tau1_tau3<-(means1[1]-means3[1])+qtlevel*sqrt(w_value[1,1]/(n_total-g)*(1/n1+1/n3))



#for bike  variable tau1-tau3 is #

lowerlimit_bike_variable_tau1_tau3<-(means1[2]-means3[2])-qtlevel*sqrt(w_value[2,2]/(n_total-g)*(1/n1+1/n3))

upperlimit_bike_variable_tau1_tau3<-(means1[2]-means3[2])+qtlevel*sqrt(w_value[2,2]/(n_total-g)*(1/n1+1/n3))
```

Ashish Mani Acharya 6348 project 2 axa190076

```
#for run  variable tau1-tau3 is #
lowerlimit_run_variable_tau1_tau3<-(means1[3]-means3[3])-qtlevel*sqrt(w_value[3,3]/(n_total-g)*(1/n1+1/n3))
upperlimit_run_variable_tau1_tau3<-(means1[3]-means3[3])+qtlevel*sqrt(w_value[3,3]/(n_total-g)*(1/n1+1/n3))

#for swim variable tau2-tau3 is #
lowerlimit_swim_variable_tau2_tau3<-(means2[1]-means3[1])-qtlevel*sqrt(w_value[1,1]/(n_total-g)*(1/n2+1/n3))
upperlimit_swim_variable_tau2_tau3<-(means2[1]-means3[1])+qtlevel*sqrt(w_value[1,1]/(n_total-g)*(1/n2+1/n3))




#for bike  variable tau1-tau3 is #
lowerlimit_bike_variable_tau2_tau3<-(means2[2]-means3[2])-qtlevel*sqrt(w_value[2,2]/(n_total-g)*(1/n2+1/n3))
upperlimit_bike_variable_tau2_tau3<-(means2[2]-means3[2])+qtlevel*sqrt(w_value[2,2]/(n_total-g)*(1/n2+1/n3))

#for run  variable tau1-tau3 is #
lowerlimit_run_variable_tau2_tau3<-(means2[3]-means3[3])-qtlevel*sqrt(w_value[3,3]/(n_total-g)*(1/n2+1/n3))
upperlimit_run_variable_tau2_tau3<-(means2[3]-means3[3])+qtlevel*sqrt(w_value[3,3]/(n_total-g)*(1/n2+1/n3))

# summary of CI #
#for swim variable tau1-tau2 is #
lowerlimit_swim_variable_tau1_tau2
upperlimit_swim_variable_tau1_tau2

#for bike  variable tau1-tau2 is #
lowerlimit_bike_variable_tau1_tau2
upperlimit_bike_variable_tau1_tau2

#for run  variable tau1-tau2 is #
lowerlimit_run_variable_tau1_tau2
upperlimit_run_variable_tau1_tau2
```

Ashish Mani Acharya 6348 project 2 axa190076

```
#for swim variable tau1-tau3 is #

lowerlimit_swim_variable_tau1_tau3

upperlimit_swim_variable_tau1_tau3


#for bike  variable tau1-tau3 is #

lowerlimit_bike_variable_tau1_tau3

upperlimit_bike_variable_tau1_tau3


#for run  variable tau1-tau3 is #

lowerlimit_run_variable_tau1_tau3

upperlimit_run_variable_tau1_tau3


#for swim variable tau2-tau3 is #

lowerlimit_swim_variable_tau2_tau3

upperlimit_swim_variable_tau2_tau3


#for bike  variable tau1-tau3 is #

lowerlimit_bike_variable_tau2_tau3

upperlimit_bike_variable_tau2_tau3


#for run  variable tau1-tau3 is #

lowerlimit_run_variable_tau2_tau3

upperlimit_run_variable_tau2_tau3


##scatterplot for all age categories ###


mycols2<-c("red","black","green")

dcol2<-factor(triathlon$CATEGORY)#col as per age category #

plot.new()

pairs(triathlon[,3:5], pch =c(15:17)[as.numeric(dcol2)], cex = 1, col = mycols2[as.numeric(dcol2)] , main = "Pairs
Scatterplot for Category
```

```
1&2&3")

legend("bottomleft", col = mycols2, legend = levels(dcol2), pch =c(15:17) ,xpd = NA, ncol = 3, bty = "n", inset = c(-0.9,-.25), pt.cex = 3)

par(xpd = TRUE)


############################Q 2 D###########################################


#a MANOVA to test whether category, gender, and their interaction have significant effects#

manova2<-manova(cbind(triathlon$SWIM,triathlon$BIKE,triathlon$RUN)~as.factor(triathlon$CATEGORY)*as.factor(triathlon$GENDER))

summary(manova2,test="Wilks")


#univariate ANOVA models to find out if interaction is significant for some variables#

# for swim #

tri3anova = aov(triathlon[,3]~as.factor(triathlon$GENDER)*as.factor(tri$CATEGORY))

summary(tri3anova)

#for bike #

tri4anova = aov(triathlon[,4]~as.factor(triathlon$GENDER)*as.factor(triathlon$CATEGORY))

summary(tri4anova)

#for run #

tri5anova = aov(triathlon[,5]~as.factor(triathlon$GENDER)*as.factor(triathlon$CATEGORY))

summary(tri5anova)



##scatterplot for all age categories ###

mycols2<-c("red","black","green")


dcol2<-factor(triathlon$CATEGORY)#col as per age category #

plot.new()

pairs(triathlon[,3:5], pch =c(15:17)[as.numeric(dcol2)], cex = 1, col = mycols2[as.numeric(dcol2)] , main = "Pairs Scatterplot for Category

1&2&3")
```

```
legend("bottomleft", col = mycols2, legend = levels(dcol2), pch =c(15:17) ,xpd = NA, ncol = 3, bty = "n", inset = c(-0.9,-.25), pt.cex = 3)

par(xpd = TRUE)


# scatter plot for all gender  #


mycols3<-c("red","black")

dcol3<-factor(triathlon$GENDER)#col as per age category #

plot.new()

pairs(triathlon[,3:5], pch = c(16:17)[as.numeric(dcol3)], cex = 1, col = mycols3[as.numeric(dcol3)] , main = "Pairs Scatterplot for GENDER")

legend("bottomleft", col = mycols2, legend = levels(dcol3), pch = c(16:17),xpd = NA, ncol = 3, bty = "n", inset = c(-0.9,-.25), pt.cex = 3)

par(xpd = TRUE)


# scatterplot for all categoreis and gender combined #

mycols6<-c("red","black")

pairs(triathlon[,3:5], col = mycols6[as.numeric(dcol3)], pch = c(15:17)[as.numeric(dcol2)],main = "Scatterplot for GENDER & AGE CATEGORY combined")

legend("bottomleft", col = mycols6, legend = levels(dcol3), pch = 20,

   xpd = NA, ncol = 3, bty = "n",inset = c(-0.1,-.25), pt.cex = 1.5)

legend("bottomleft", pch = c(15:17), legend = levels(dcol2), col = "black",

   xpd = NA, ncol = 3, bty = "n", inset = c(-2.2,-.25))

par(xpd=TRUE)


# interaction plot for swim #


interaction.plot(x.factor = triathlon$CATEGORY, #x-axis variable

         trace.factor =triathlon$GENDER, #variable for lines

         response = triathlon$SWIM, #y-axis variable

         fun = median, #metric to plot

         ylab = "SWIM",

         xlab = "AGE CATEGORY",
```

```
        col = c("RED", "BLUE"),
        lty = 1, #line type
        lwd = 3, #line width
        trace.label = "Gender")

# interaction plot for bike #
interaction.plot(x.factor = triathlon$CATEGORY, #x-axis variable
        trace.factor =triathlon$GENDER, #variable for lines
        response = triathlon$BIKE, #y-axis variable
        fun = median, #metric to plot
        ylab = "BIKE",
        xlab = "AGE CATEGORY",
        col = c("RED", "BLUE"),
        lty = 1, #line type
        lwd = 3, #line width
        trace.label = "Gender")

# interaction plot for run #
interaction.plot(x.factor = triathlon$CATEGORY, #x-axis variable
        trace.factor =triathlon$GENDER, #variable for lines
        response = triathlon$RUN, #y-axis variable
        fun = median, #metric to plot
        ylab = "RUN",
        xlab = "AGE CATEGORY",
        col = c("RED", "BLUE"),
        lty = 1, #line type
        lwd = 3, #line width
        trace.label = "Gender")
```

Ashish Mani Acharya	6348 project 2	axa190076