Multivariate Statistics: Group Assignment 1

Max Hicks,Sepehr Anzabipour

April 2020

Contents

1 Q1

The dataset being analysed in our case ILPD "Indian Liver Patient Dataset" measures various metrics between males and females belonging to 6 different cohorts. These metrics include total and direct Bilirubin levels within the liver. Total protein level present in the liver. Albumin and Albumin Globulin levels (A/G ratio). Along with a selector field with labels 1 for liver patient and 2 for non-liver patient.
This data set contains 416 liver patient records and 167 non liver patient records and also contains 441 male patient records and 142 female patient records.

Now let's explain what these metrics actually mean:

TB stands for total bilirubin. Bilirubin is the chemical in our bodies responsible for breaking down haem which is found in red blood cells. If you have high total bilirubin levels this could indicate a problem with your liver, such as cirrhosis or possibly even cancer. However it may not be the case that something is wrong with the liver, it could also be a symptom of gallstones or Gilbert's syndrome.

DB stands for direct bilirubin. Sometimes called conjugated bilirubin as this is the measurement of bilirubin after it has conjugated in the liver. This conjugation converts bilirubin into a water soluble form that can then be excreted by the body. High levels can cause jaundice, resulting in pigmentation of the skin meaning that there is impaired liver function.

TP stands for total protein levels. Albumin and globulin are two types of protein in your body. The total protein test measures the total amount albumin and globulin in your body. This test is useful in detecting things like liver disease. High levels could mean further tests need to be done for hepatitis b and c and also HIV. However a low total protein level can mean there may be liver dysfunction. However there are many other causes for low protein levels such as malnutrition,bleeding and celiac disease.

ALB or albumin is a protein and one half of the total protein test. Low levels of albumin can be a sign of either malnutrition or liver dysfunction. Which when used in combination with our total protein test could further bolster the claim of liver disease.

A/G ratio or the albumin globulin ratio is just the ratio of albumin to globulin. A low A/G ratio may indicate an under production of Albumin, where albumin is produced in the liver, thus could be an indicator of cirrhosis or liver cancer. A high A/G level could mean that the patient could have a genetic deficiency where they produce lower amounts of globulin, or it can possibly mean that leukemia tests should be performed.

2 Q2

Simply running the command in the r script ”ILPD¡- read.csv(”ILPD.csv”) will be used to do any and all transformations to our data.

2.1 Q2i

In the R script there is a simple for loop which prints the total number of patients in each cohort. Where ” print(paste(”Number of patients in cohort”,i,”is:”,length(which(ILPD[9]==i))))” is looped over 6 times and produces the output.

Cohorts 1,2,3,4,5 all have 100 patients and cohort 6 has 83.

2.2 Q2ii

Running the command ”print(length(which(ILPD[2]==”Female”)))” will produce the number of Females in the study which is 142. We can then write $\frac{142}{583}$ for the proportion which is 0.244 or 24.4%.

2.3 Q2iii

Running the command ”which(ILPD$TB==max(ILPD$TB))” produces the row number of the patient with the highest TB count. The patient is located on row 167 a male belonging to cohort 2 who is a liver patient.

2.4 Q2iv

Run the commands which strip the factor variables out of our data set. then run ”colMeans(ILPD_pre_mean)” which will produce our sample mean vector:

$$\vec{\bar{y}} = \begin{bmatrix} 3.2987993 \\ 1.4861063 \\ 6.4831904 \\ 3.1418525 \\ 0.9489708 \end{bmatrix}$$

2.5 Q2v

Basically the same as the last question except this time variance.

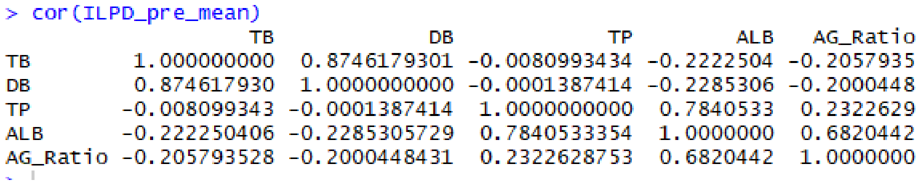
Run the command ”variance_vector¡- apply(ILPD_pre_mean,2,var)” and then ”variance_vector” to output:

$$\vec{\sigma}^2 = \begin{bmatrix} 38.5581601 \\ 7.8876589 \\ 1.1782049 \\ 0.6328502 \\ 0.1041079 \end{bmatrix}$$

We can see here the Aratio has the lowest variation of only 0.104 which makes sense given that A/G ratios are usually less than 1 and have small variations anyway.

2.6 Q2vi

Run the command ”cor(ILPD_pre_mean)” to produce the correlation matrix for our quantitative data:



```
> cor(ILPD_pre_mean)
                   TB            DB            TP        ALB   AG_Ratio
TB        1.000000000  0.8746179301 -0.0080993434 -0.2222504 -0.2057935
DB        0.874617930  1.0000000000 -0.0001387414 -0.2285306 -0.2000448
TP       -0.008099343 -0.0001387414  1.0000000000  0.7840533  0.2322629
ALB      -0.222250406 -0.2285305729  0.7840533354  1.0000000  0.6820442
AG_Ratio -0.205793528 -0.2000448431  0.2322628753  0.6820442  1.0000000
```

Figure 1: Correlation plot matrix

We can see here that TB and DB are most highly positively correlated together and is the strongest correlation in the matrix.

2.7 Q2vii

The pairs.panels function from the psych package is a more powerful scatter plot matrix function, similar to cor.plot.

It returns a scatter plot matrix consisting of the following: The diagonals are the density histograms, how each variable is distributed. The lower squares, below the diagonal are scatter plots consisting of a line of best fit by default however i have changed this by adding a confidence interval.

Also using the smoother argument set to TRUE we get this very nice soft density plot instead of lots of dots everywhere. It is recommended in the documentation that if the number of observations is higher than 100-200 then using pch=”.”

may be a good idea, this does mean we can no longer use the smoother command however. Also the stars command which can show the significance of each correlation in the top right quadrants is a nice addition, but not for everyone. There is also a rug argument which puts a rug under the histograms, but leaving it to false is a little cleaner as you cant really see it anyway.

Here are two examples, one with smoother on and one with smoother off and pch=”.” on:

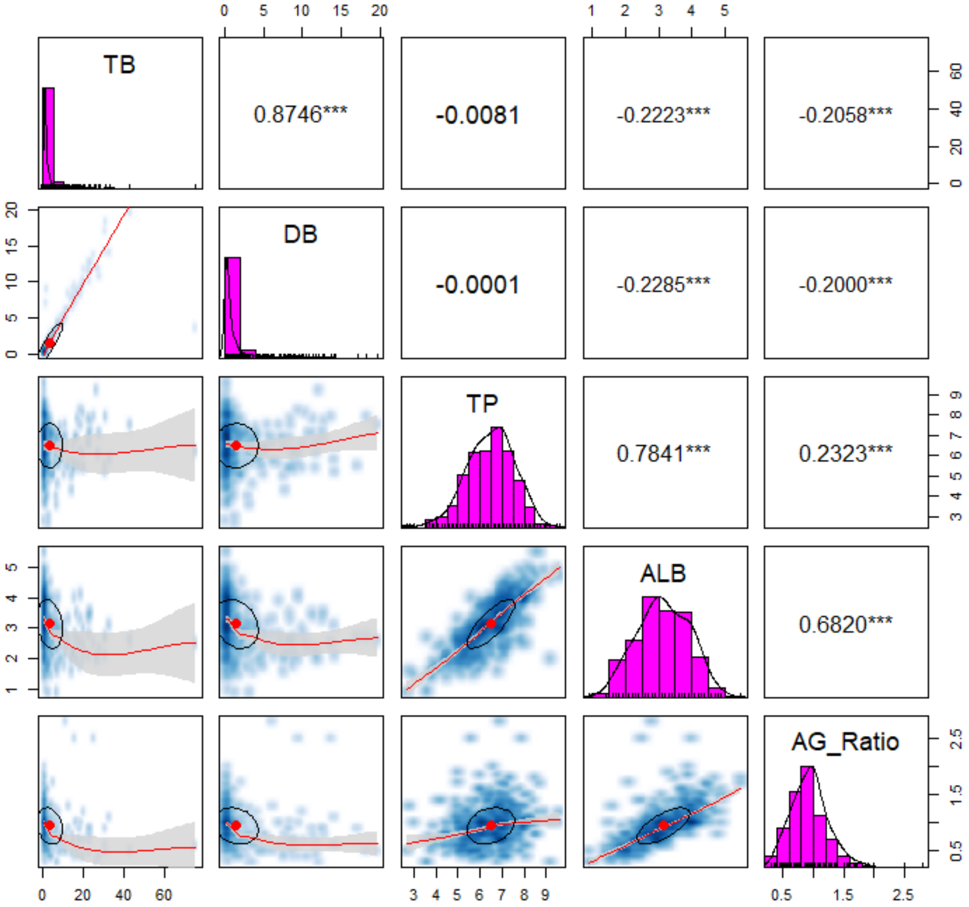


Figure 2: Pairs.panels smoother

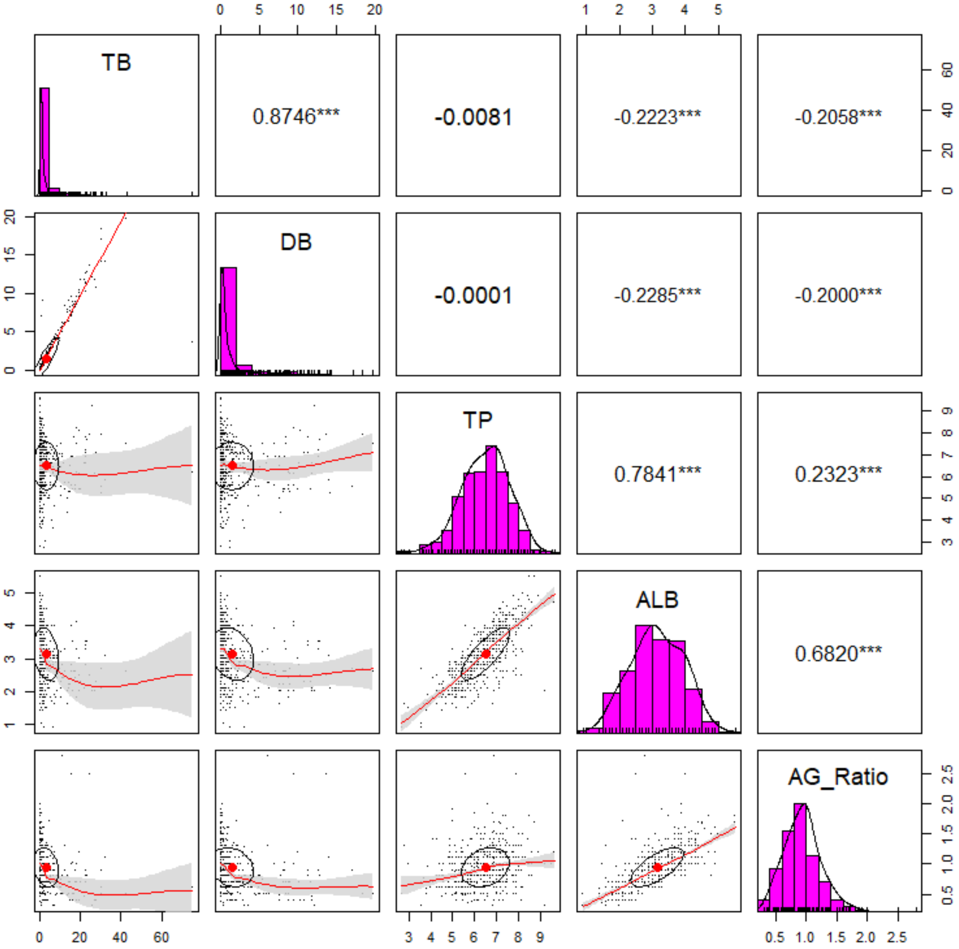
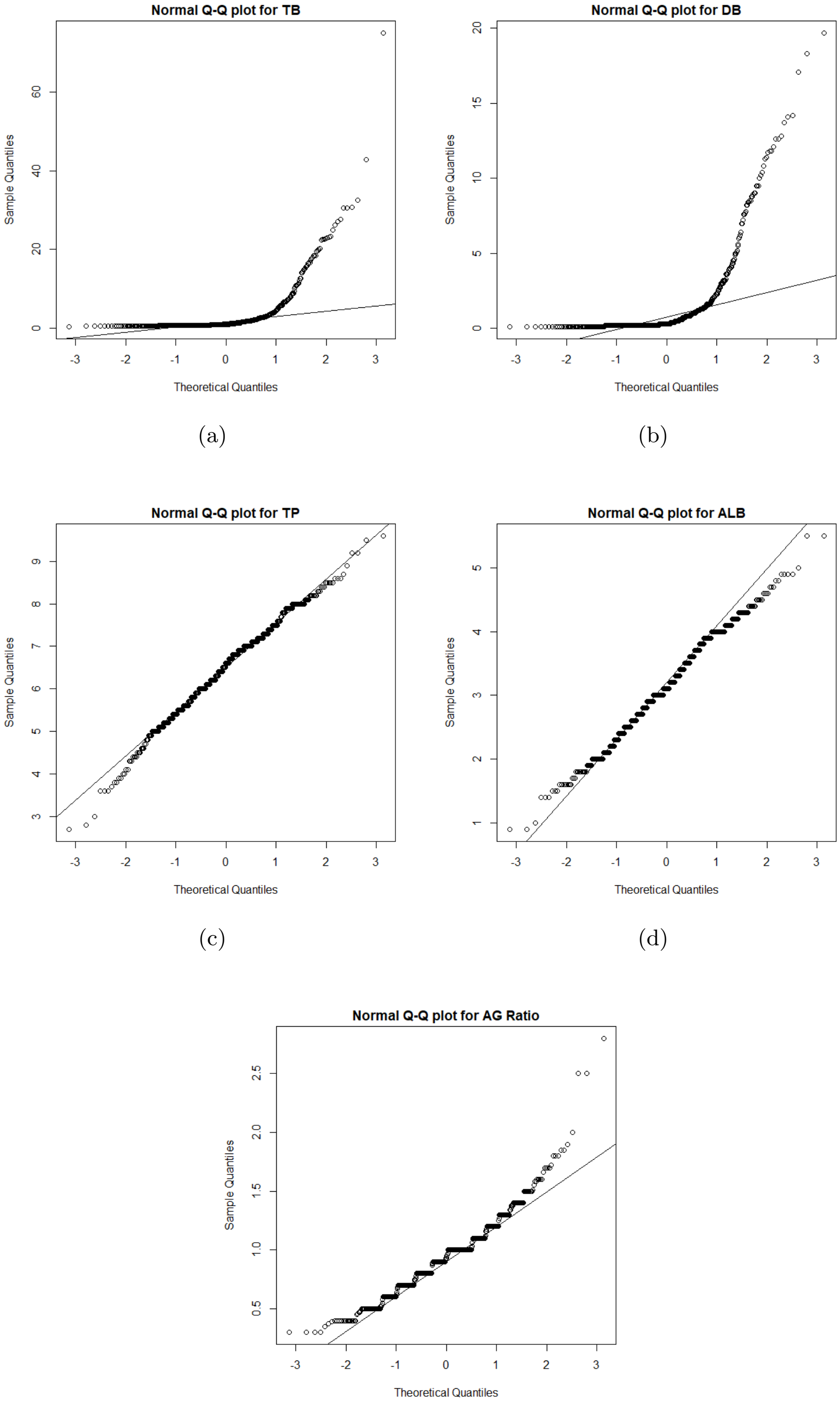


Figure 3: Pairs.panels pch=”.”

3 Q3

For this question we will produce 5 normal Q-Q plots for our 5 quantitative variables along with 5 Shapiro-Wilk tests all to see if our data are normally distributed.



(a)

(b)

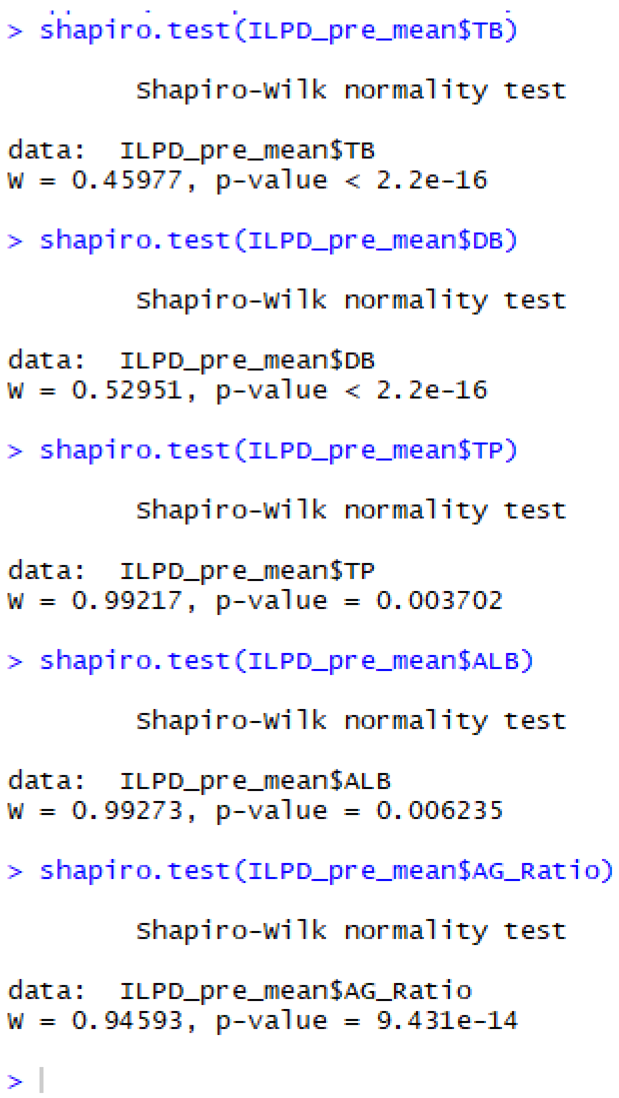
(c)

(d)

(e)

Figure 4: Q-Q plots for quantitative variables

We can see here from inspection of these univariate Q-Q plots that it is certainly the case that TB and DB are not normally distributed due to the points being very far away from our qqline. The other plot that stands out is for A/G ratio looks to be not so normally distributed either. However TP and ALB, from graphical inspection. Do appear on the face to be normally distributed somewhat except for nearer the tails of the quantities where it deviates from the line. Using a Shapiro-Wilk test on both of these further to check there normality, so let's do that!



```
> shapiro.test(ILPD_pre_mean$TB)

        Shapiro-Wilk normality test

data:  ILPD_pre_mean$TB
W = 0.45977, p-value < 2.2e-16

> shapiro.test(ILPD_pre_mean$DB)

        Shapiro-Wilk normality test

data:  ILPD_pre_mean$DB
W = 0.52951, p-value < 2.2e-16

> shapiro.test(ILPD_pre_mean$TP)

        Shapiro-Wilk normality test

data:  ILPD_pre_mean$TP
W = 0.99217, p-value = 0.003702

> shapiro.test(ILPD_pre_mean$ALB)

        Shapiro-Wilk normality test

data:  ILPD_pre_mean$ALB
W = 0.99273, p-value = 0.006235

> shapiro.test(ILPD_pre_mean$AG_Ratio)

        Shapiro-Wilk normality test

data:  ILPD_pre_mean$AG_Ratio
W = 0.94593, p-value = 9.431e-14

> 
```

Figure 5: Caption

Here we can see that all quantities, even those that appeared to be normally distributed on there face are not so. All null hypotheses are rejected at the 5% and 1% significance levels. We can say with confidence our data are not normally distributed.

As an aside it might be worth noting that we can do just a bit more by using the mvn() function from the MVN package to produce a chi-squared Q-Q plot and also an output for Mardia's multivariate statistic to test for multivariate normality.

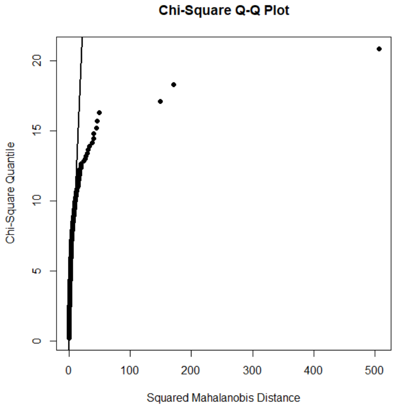
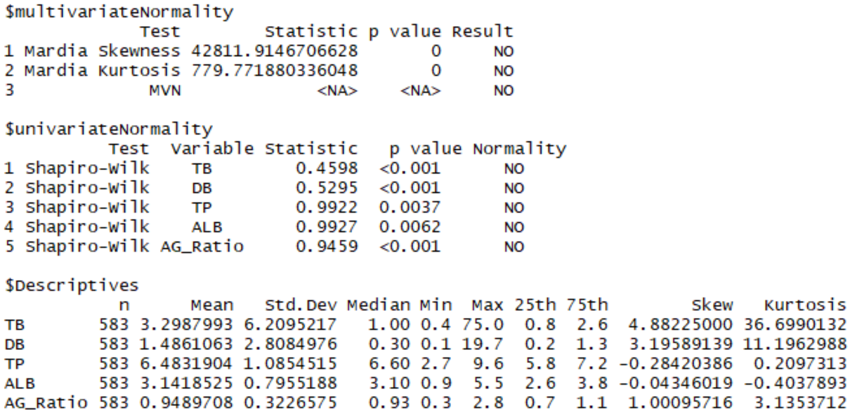


Figure 6: Chi-Squared Q-Q Plot



```
$multivariateNormality
             Test        Statistic p value Result
1 Mardia Skewness 42811.9146706628       0     NO
2 Mardia Kurtosis 779.771880336048       0     NO
3             MVN             <NA>    <NA>     NO

$univariateNormality
          Test  Variable Statistic   p value Normality
1 Shapiro-Wilk    TB        0.4598  <0.001      NO
2 Shapiro-Wilk    DB        0.5295  <0.001      NO
3 Shapiro-Wilk    TP        0.9922  0.0037      NO
4 Shapiro-Wilk    ALB       0.9927  0.0062      NO
5 Shapiro-Wilk AG_Ratio     0.9459  <0.001      NO

$Descriptives
           n      Mean   Std.Dev Median Min  Max 25th 75th        Skew   Kurtosis
TB       583 3.2987993 6.2095217   1.00 0.4 75.0  0.8  2.6  4.88225000 36.6990132
DB       583 1.4861063 2.8084976   0.30 0.1 19.7  0.2  1.3  3.19589139 11.1962988
TP       583 6.4831904 1.0854515   6.60 2.7  9.6  5.8  7.2 -0.28420386  0.2097313
ALB      583 3.1418525 0.7955188   3.10 0.9  5.5  2.6  3.8 -0.04346019 -0.4037893
AG_Ratio 583 0.9489708 0.3226575   0.93 0.3  2.8  0.7  1.1  1.00095716  3.1353712
```

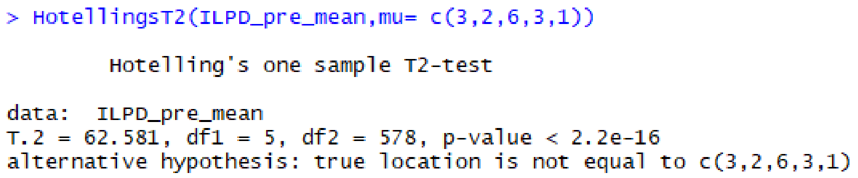
Figure 7: Mardia test

We can see from this plot and further from the statistical output that our data is also not multivariate normal nor univriate normal.

4 Q4

4.1 Q4i

under the assumption that multivriate normality does hold we can test if our means are equal to this $\vec{\mu} = (3, 2, 6, 3, 1)^T$. To do this we can do a one sample Hotellings test to see if our mean is within this vector.



```
> HotellingsT2(ILPD_pre_mean,mu= c(3,2,6,3,1))

        Hotelling's one sample T2-test

data:  ILPD_pre_mean
T.2 = 62.581, df1 = 5, df2 = 578, p-value < 2.2e-16
alternative hypothesis: true location is not equal to c(3,2,6,3,1)
```

Figure 8: Hotellings T2

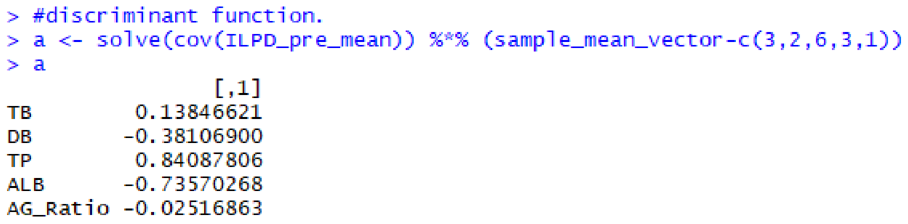
As we can clearly see our mean is not in this vector and we reject the null hypothesis.

4.2 Q4ii

Now that we have reject our hypothesis that the mean does not lie within our vector $\vec{\mu} = (3, 2, 6, 3, 1)^T$ so now we want to see what variable would contribute to this rejection the most.

Previously we used a two sample version with a pooled covariance matrix but since it is one sample now we are effectively using two different means on the same data set. in other words. we have two mean vectors. Our true mean as seen in section 2.4 and our theoretical mean $\vec{\mu} = (3, 2, 6, 3, 1)^T$.

So our discriminant will just be the inverse of our covariance matrix of our quantitative data set and our difference vector will be the true mean - the theoretical mean. and we can see just what variable contributes the most to this rejection.



```
> #discriminant function.
> a <- solve(cov(ILPD_pre_mean)) %*% (sample_mean_vector-c(3,2,6,3,1))
> a
                [,1]
TB        0.13846621
DB       -0.38106900
TP        0.84087806
ALB      -0.73570268
AG_Ratio -0.02516863
```

Figure 9: Caption

We can see from the discriminant function that the TP variable contributed most to the rejection.

4.3 Q4iii

Next we can do a profile analysis of our variables and test them with the paos function to see both numerically and in this case visibly that our profile is not flat.

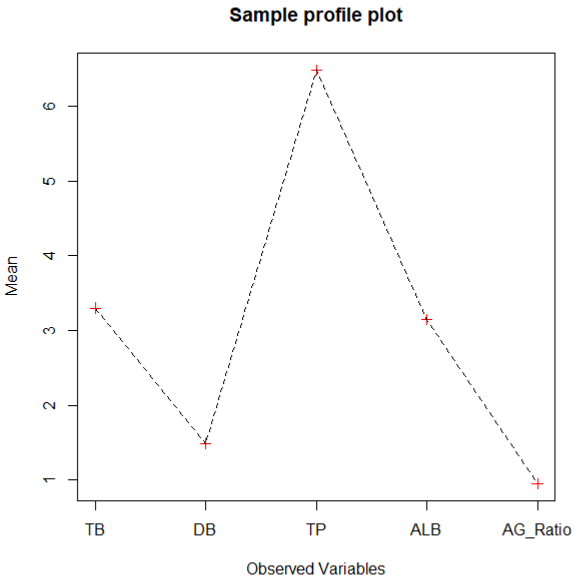
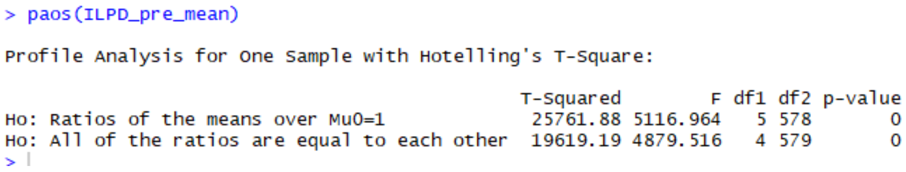


Figure 10: Caption



```
> paos(ILPD_pre_mean)

Profile Analysis for One Sample with Hotelling's T-Square:

                                             T-Squared        F df1 df2 p-value
Ho: Ratios of the means over Mu0=1            25761.88 5116.964   5 578       0
Ho: All of the ratios are equal to each other  19619.19 4879.516   4 579       0
> 
```

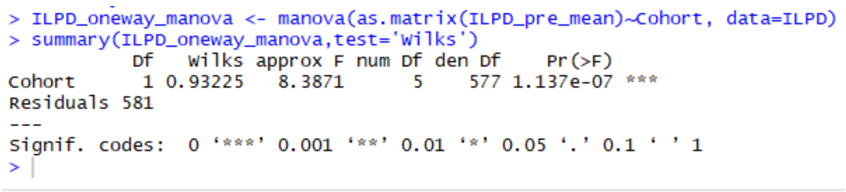
Figure 11: Caption

We can see here both graphically and from our statistical test that our profile is not flat and thus we can reject the null hypothesis as our output has a p-value of 0 and graphically... wel because it isnt flat.

5 Q5

5.1 Q5i

In this question we are being asked to do a one way MANOVA tset on our data set to see if there are differences between multivariate means of all cohorts are equal, such that $H_0\colon \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_g.$ and $H_a\colon \mu_{ik} \neq \mu_{jk}$ This says that the null hypothesis is false if at least one pair of treatments is different on at least one variable.



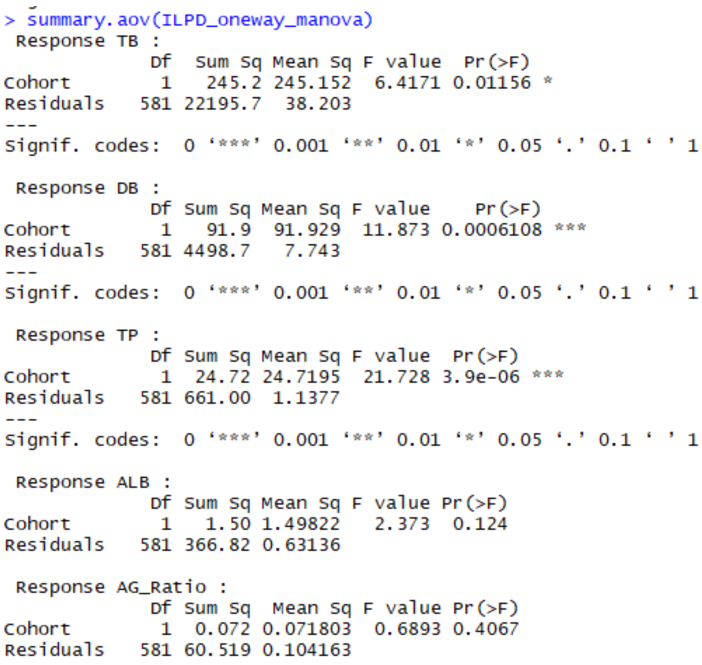
```
> ILPD_oneway_manova <- manova(as.matrix(ILPD_pre_mean)~Cohort, data=ILPD)
> summary(ILPD_oneway_manova,test='Wilks')
           Df   Wilks approx F num Df den Df    Pr(>F)    
Cohort      1 0.93225   8.3871      5    577 1.137e-07 ***
Residuals 581
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

Figure 12: One-Way MANOVA model

We can see here our null hypothesis has been rejected. Therefore we can say at least one of our multivariate mean vectors is not the same as the other mean vectors.

5.2 Q5ii

Seeing as we have rejected H_0 we can begin our individual ANOVA tests on these variables to see what what variables were significantly different as seen here in this figure:



```
> summary.aov(ILPD_oneway_manova)
 Response TB :
             Df  Sum Sq Mean Sq F value  Pr(>F)
Cohort        1   245.2 245.152  6.4171 0.01156 *
Residuals   581 22195.7  38.203
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response DB :
             Df Sum Sq Mean Sq F value    Pr(>F)
Cohort        1   91.9  91.929  11.873 0.0006108 ***
Residuals   581 4498.7   7.743
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response TP :
             Df Sum Sq Mean Sq F value  Pr(>F)
Cohort        1  24.72 24.7195  21.728 3.9e-06 ***
Residuals   581 661.00  1.1377
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response ALB :
             Df Sum Sq Mean Sq F value Pr(>F)
Cohort        1   1.50 1.49822   2.373  0.124
Residuals   581 366.82 0.63136

 Response AG_Ratio :
             Df Sum Sq  Mean Sq F value Pr(>F)
Cohort        1  0.072 0.071803  0.6893 0.4067
Residuals   581 60.519 0.104163
```

Figure 13: Individual ANOVA tests

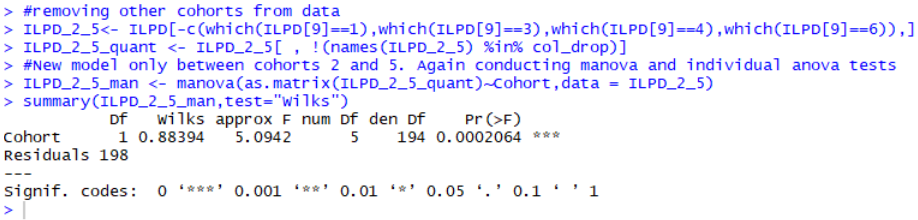
So what these tests are saying to us is, is there a difference between means of this variable between cohorts. So for example in the first test we are doing an ANOVA test on the TB variable between cohorts 1 to 6. We have a p-value of 0.01156, therefore we reject the null hypothesis that the means are same between cohorts and therefore accept that there is a difference between the means of TB levels between cohorts. We repeat this process and we fin that we accept the alternate hypothesis for TB,DB and TP levels between cohorts. Meaning that there is a difference in means between cohorts with regard to these variables. However for ALB and A/G ratio we cannot reject the null hypothesis as the p-values are greater than 0.05 in both cases.

The downside to doing individual ANOVA tests like this is that MANOVA is more powerful in the way that it can find the truly important factors in our data as these variables can express multi co-linearity and thus be linked to each other, ANOVA cannot capture this. There will also be a Type 1 error inflation when using multiple independent ANOVA tests. Therefore more true null hypothesis will be rejected even though we should accept them. So in summary ANOVA cannot capture the predictors interacting in the presence of other predictor variables and thus loses more of the picture whilst also increasing the chance that you can potentially reject true null hypotheses .

5.3 Q5iii

In this question all we have to now is remove all cohorts except for 2 and 5. So we can write some code to remove these rows which contain a cohort values of of not 2 and not 5. We now can start analysing the differences between cohorts 2 and 5, to assess the claims made: That the averages (means) are the same between cohorts 2 and 5 and also test what variables differ significantly between said cohorts.

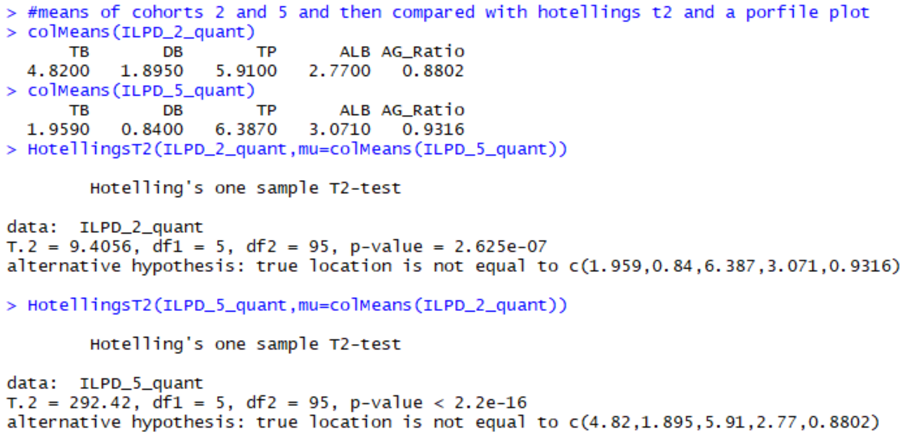
This means we will perform a MANOVA test to asses whether or not our mean vectors are different:



```
> #removing other cohorts from data
> ILPD_2_5<- ILPD[-c(which(ILPD[9]==1),which(ILPD[9]==3),which(ILPD[9]==4),which(ILPD[9]==6)),]
> ILPD_2_5_quant <- ILPD_2_5[ , !(names(ILPD_2_5) %in% col_drop)]
> #New model only between cohorts 2 and 5. Again conducting manova and individual anova tests
> ILPD_2_5_man <- manova(as.matrix(ILPD_2_5_quant)~Cohort,data = ILPD_2_5)
> summary(ILPD_2_5_man,test="Wilks")
           Df   Wilks approx F num Df den Df    Pr(>F)    
Cohort      1 0.88394   5.0942      5    194 0.0002064 ***
Residuals 198                                              
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

Figure 14: MANOVA test between cohorts 2 and 5

We can say with confidence that the means between cohorts are in fact different. We can then reject the claim made that the average levels are the same. We can further prove this by separating this data set of cohorts 2 and 5 to 2 new data sets with one counting cohort 2 and the other cohort 5 perform a Hotelling's test on 1 or both of them to see if the means are equal and also plot a sample profile using the pbg function.



```
> #means of cohorts 2 and 5 and then compared with hotellings t2 and a porfile plot
> colMeans(ILPD_2_quant)
      TB       DB       TP      ALB  AG_Ratio
  4.8200   1.8950   5.9100   2.7700   0.8802
> colMeans(ILPD_5_quant)
      TB       DB       TP      ALB  AG_Ratio
  1.9590   0.8400   6.3870   3.0710   0.9316
> HotellingsT2(ILPD_2_quant,mu=colMeans(ILPD_5_quant))

        Hotelling's one sample T2-test

data:  ILPD_2_quant
T.2 = 9.4056, df1 = 5, df2 = 95, p-value = 2.625e-07
alternative hypothesis: true location is not equal to c(1.959,0.84,6.387,3.071,0.9316)

> HotellingsT2(ILPD_5_quant,mu=colMeans(ILPD_2_quant))

        Hotelling's one sample T2-test

data:  ILPD_5_quant
T.2 = 292.42, df1 = 5, df2 = 95, p-value < 2.2e-16
alternative hypothesis: true location is not equal to c(4.82,1.895,5.91,2.77,0.8802)
```

Figure 15: Cohort 2 vs Cohort 5 Hotelling's tests

We can see here that the Hotelling's function is telling us the means are not the same.

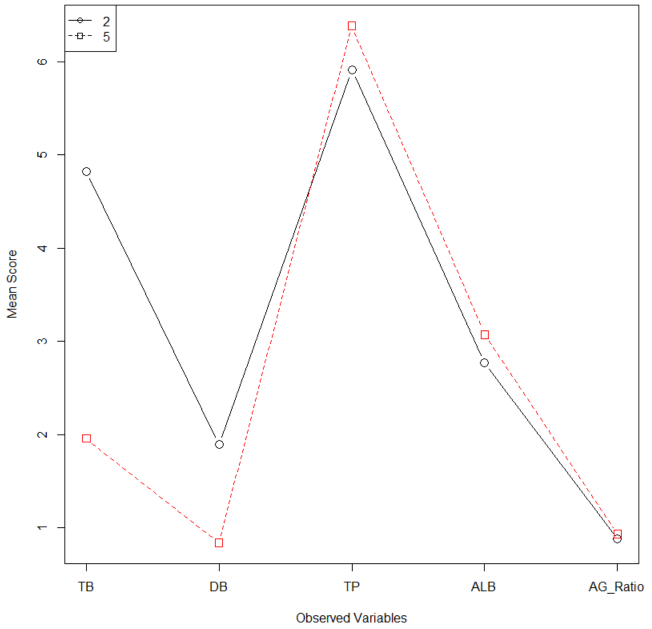
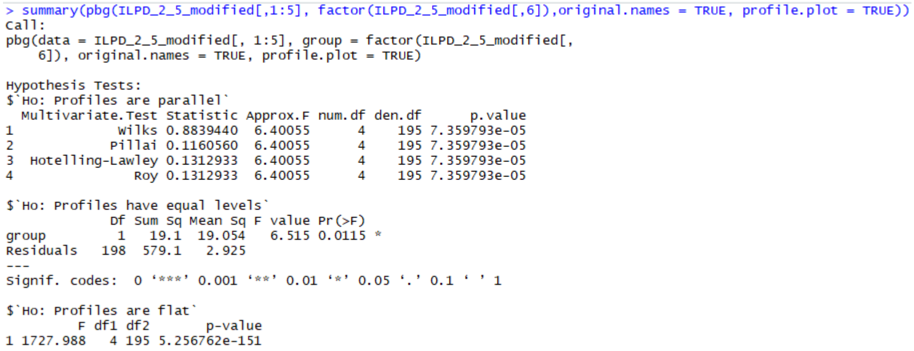


Figure 16: Profiles of Cohorts 2 and 5

We can also see here that are profile plot show that we do not have a flat mean, bu we may suspect that they have equal levels.

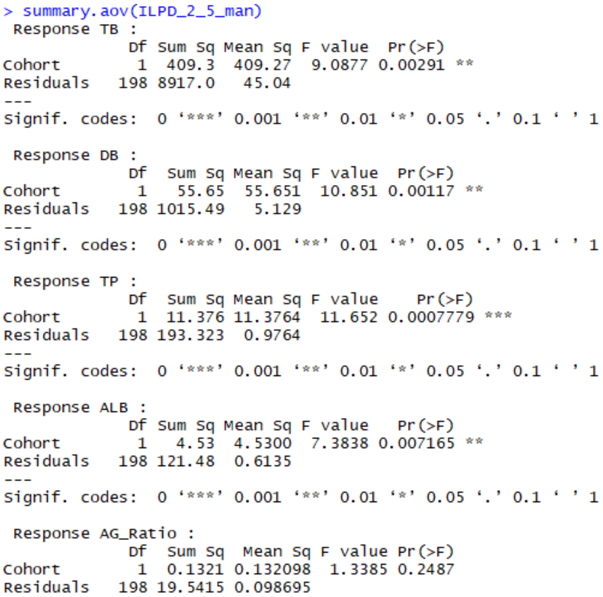


```
> summary(pbg(ILPD_2_5_modified[,1:5], factor(ILPD_2_5_modified[,6]),original.names = TRUE, profile.plot = TRUE))
Call:
pbg(data = ILPD_2_5_modified[, 1:5], group = factor(ILPD_2_5_modified[,
    6]), original.names = TRUE, profile.plot = TRUE)

Hypothesis Tests:
$`Ho: Profiles are parallel`
  Multivariate.Test Statistic Approx.F num.df den.df      p.value
1             Wilks 0.8839440  6.40055      4    195 7.359793e-05
2            Pillai 0.1160560  6.40055      4    195 7.359793e-05
3  Hotelling-Lawley 0.1312933  6.40055      4    195 7.359793e-05
4               Roy 0.1312933  6.40055      4    195 7.359793e-05

$`Ho: Profiles have equal levels`
             Df Sum Sq Mean Sq F value Pr(>F)
group         1   19.1  19.054   6.515 0.0115 *
Residuals   198  579.1   2.925
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

$`Ho: Profiles are flat`
         F df1 df2       p-value
1 1727.988   4 195 5.256762e-151
```

Figure 17: Summary of profile tests for cohorts 2 and 5

And here we can see various null hypotheses tested against our profile and we find that we do not have a flat profile (obviously) a parallel profile or a profile with equal levels. This further back the claim that the means are different between cohorts 2 and 5.

Next we must look if there are significant differences between individual variables between cohorts 2 and 5. We perform multiple ANOVA tests to achieve this, also noting that this has the same drawbacks as mentioned previously.



```
> summary.aov(ILPD_2_5_man)
 Response TB :
             Df Sum Sq Mean Sq F value   Pr(>F)
Cohort        1  409.3  409.27  9.0877 0.00291 **
Residuals   198 8917.0   45.04
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response DB :
             Df  Sum Sq Mean Sq F value   Pr(>F)
Cohort        1   55.65  55.651  10.851 0.00117 **
Residuals   198 1015.49   5.129
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response TP :
             Df  Sum Sq Mean Sq F value    Pr(>F)
Cohort        1  11.376 11.3764  11.652 0.0007779 ***
Residuals   198 193.323  0.9764
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response ALB :
             Df Sum Sq Mean Sq F value   Pr(>F)
Cohort        1   4.53  4.5300  7.3838 0.007165 **
Residuals   198 121.48  0.6135
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response AG_Ratio :
             Df  Sum Sq  Mean Sq F value Pr(>F)
Cohort        1  0.1321 0.132098  1.3385 0.2487
Residuals   198 19.5415 0.098695
```

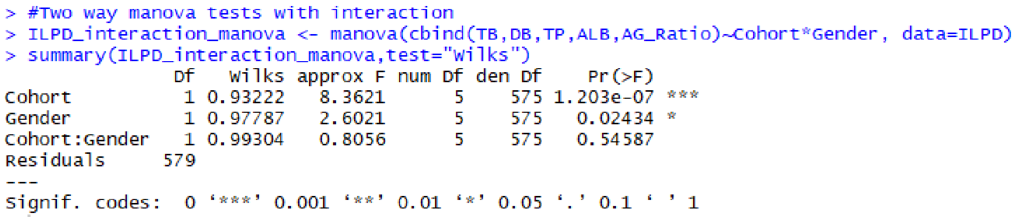
Figure 18: Caption

We can see here that we do indeed see significant differences in almost all variables between cohorts 2 and 5. The only variable which does not appear to be to different is the A/G ratio, however this is somewhat expected as the variation in A/G ratio is so low it is unlikely that it would vary that miuch despite every other variable varying quite a bit between both cohorts. We can then say with confidence that we can accept the claim that there are significant differences between cohorts 2 and 5 with regard to individual variables, however there means are not the same as seen in the MANOVA and profile analysis.

6 Q6

6.1 Q6i

For this question we are asked to do a two way MANOVA test concerning both cohort and gender. There are two ways we will do this. The first way is the interactive model whereby we consider that gender and cohort interact with each other, and secondly a simpler model known as an additive model where we only consider the main effects more independently of one another.
Let's start our model with interactions:

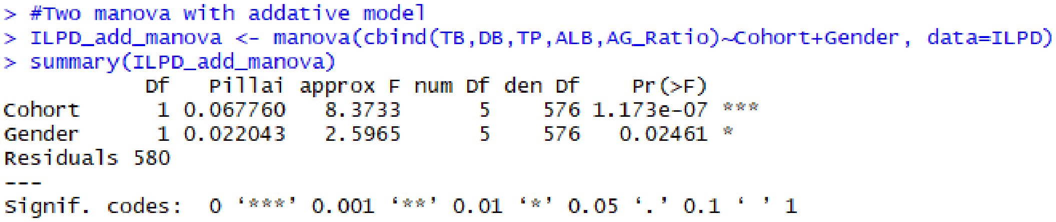


```
> #Two way manova tests with interaction
> ILPD_interaction_manova <- manova(cbind(TB,DB,TP,ALB,AG_Ratio)~Cohort*Gender, data=ILPD)
> summary(ILPD_interaction_manova,test="Wilks")
               Df   Wilks approx F num Df den Df    Pr(>F)    
Cohort          1 0.93222   8.3621      5    575 1.203e-07 ***
Gender          1 0.97787   2.6021      5    575   0.02434 *  
Cohort:Gender   1 0.99304   0.8056      5    575   0.54587    
Residuals     579                                             
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 19: Two way MANOVA interaction model

As we can see here we have told R to include interactions by using the asterisk operator. And as we can see here we can reject that there is any interaction between cohort and gender as the p-value is greater than 0.05. Therefore we cannot reject the null hypothesis for interactions between gender and cohort.

Next we can use our additive model:



```
> #Two manova with addative model
> ILPD_add_manova <- manova(cbind(TB,DB,TP,ALB,AG_Ratio)~Cohort+Gender, data=ILPD)
> summary(ILPD_add_manova)
           Df   Pillai approx F num Df den Df    Pr(>F)    
Cohort      1 0.067760   8.3733      5    576 1.173e-07 ***
Gender      1 0.022043   2.5965      5    576   0.02461 *  
Residuals 580                                              
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

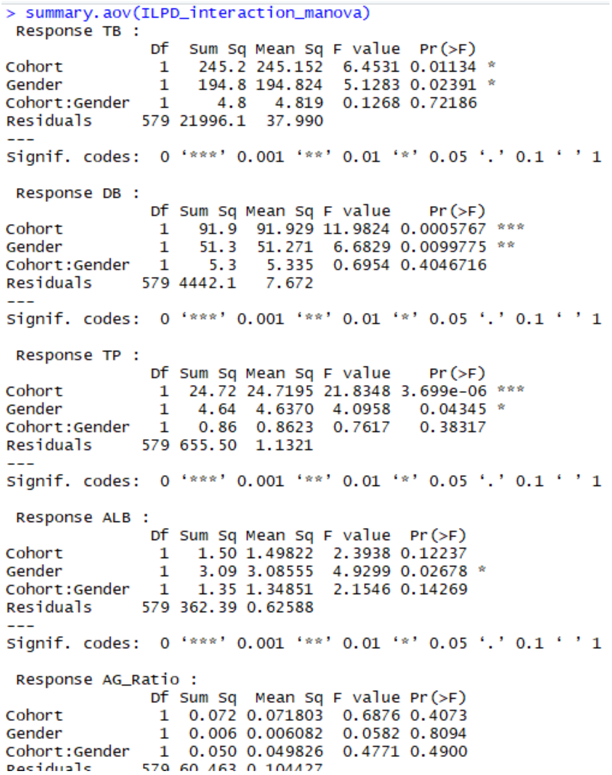
Figure 20: Two way MANOVA additive model

We can see here that there is a difference between cohort and also a difference between gender when it comes to our mean vectors as both p-values are below 0.05 so in this case we can reject the null hypothesis that there is no difference between groups in either gender or cohort.

6.2 Q6ii

In this question we just use summary.aov() on both of our models to see whether or not there are any significant differences in terms of the variables being tested in the ANOVA test. This has similar downsides however as Individual ANOVA tests can only take into account 1 variable at a time and cannot account for interactions between these predictor variables. We also may be rejecting true null hypotheses more as type 1 error can tend to inflate with ANOVA tests in this manner.

Here are the ANOVA tests for our interaction model:

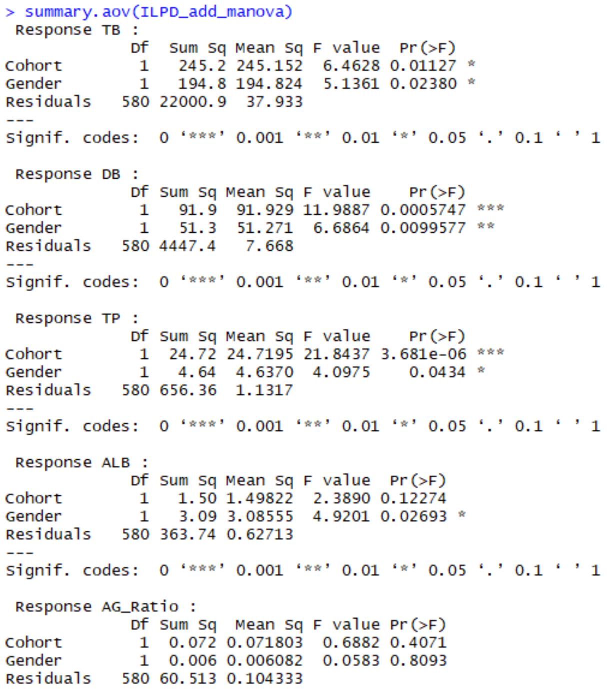


```
> summary.aov(ILPD_interaction_manova)
 Response TB :
               Df  Sum Sq Mean Sq F value  Pr(>F)
Cohort          1   245.2 245.152  6.4531 0.01134 *
Gender          1   194.8 194.824  5.1283 0.02391 *
Cohort:Gender   1     4.8   4.819  0.1268 0.72186
Residuals     579 21996.1  37.990
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response DB :
               Df Sum Sq Mean Sq F value    Pr(>F)
Cohort          1   91.9  91.929 11.9824 0.0005767 ***
Gender          1   51.3  51.271  6.6829 0.0099775 **
Cohort:Gender   1    5.3   5.335  0.6954 0.4046716
Residuals     579 4442.1   7.672
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response TP :
               Df Sum Sq Mean Sq F value    Pr(>F)
Cohort          1  24.72 24.7195 21.8348 3.699e-06 ***
Gender          1   4.64  4.6370  4.0958   0.04345 *
Cohort:Gender   1   0.86  0.8623  0.7617   0.38317
Residuals     579 655.50  1.1321
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response ALB :
               Df Sum Sq Mean Sq F value  Pr(>F)
Cohort          1   1.50 1.49822  2.3938 0.12237
Gender          1   3.09 3.08555  4.9299 0.02678 *
Cohort:Gender   1   1.35 1.34851  2.1546 0.14269
Residuals     579 362.39 0.62588
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response AG_Ratio :
               Df Sum Sq  Mean Sq F value Pr(>F)
Cohort          1  0.072 0.071803  0.6876 0.4073
Gender          1  0.006 0.006082  0.0582 0.8094
Cohort:Gender   1  0.050 0.049826  0.4771 0.4900
Residuals     579 60.463 0.104427
```

Figure 21: ANOVA tests for interaction model

We can see here from our test that for all variables tested we cannot reject the null hypothesis for the interactions term in our model. Meaning that there is no difference between how gender and cohort interact between our variables for these respected groups, IE the interaction between 1 group say gender and cohort 1 is no different than gender and cohort 2 etc etc.

And here are the tests for our additive model:



```
> summary.aov(ILPD_add_manova)
 Response TB :
             Df  Sum Sq Mean Sq F value  Pr(>F)
Cohort        1   245.2 245.152  6.4628 0.01127 *
Gender        1   194.8 194.824  5.1361 0.02380 *
Residuals   580 22000.9  37.933
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response DB :
             Df Sum Sq Mean Sq F value    Pr(>F)
Cohort        1   91.9  91.929 11.9887 0.0005747 ***
Gender        1   51.3  51.271  6.6864 0.0099577 **
Residuals   580 4447.4   7.668
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response TP :
             Df Sum Sq Mean Sq F value    Pr(>F)
Cohort        1  24.72 24.7195 21.8437 3.681e-06 ***
Gender        1   4.64  4.6370  4.0975    0.0434 *
Residuals   580 656.36  1.1317
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response ALB :
             Df Sum Sq Mean Sq F value  Pr(>F)
Cohort        1   1.50 1.49822  2.3890 0.12274
Gender        1   3.09 3.08555  4.9201 0.02693 *
Residuals   580 363.74 0.62713
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response AG_Ratio :
             Df Sum Sq  Mean Sq F value Pr(>F)
Cohort        1  0.072 0.071803  0.6882 0.4071
Gender        1  0.006 0.006082  0.0583 0.8093
Residuals   580 60.513 0.104333
```

Figure 22: ANOVA tests for additive model

As we can see here We cannot reject the null hypothesis for A/G ratio. Again this is most likely due to the low variance in A/G ratio as shown in the variance vector so this shouldn't surprise us. However our first three variables we can reject the null hypothesis and say that there are significant differences between cohort and between gender for TB, DB and TP levels. For Albumin or ALB levels we can only reject the null hypothesis for gender and like we saw before in section 5.2 we cannot reject the null hypothesis for ALB levels being the same between cohorts.

7 Q7

We can conclude in this study the following things.

We have demonstrated the lack of normality in the data moth univeriately and multivaritely.

Under the assumption however that our data could be multivariate normal we have shown there is a significant difference in mean vectors between cohorts and we have also show that three out of 5 quantitative variables differ significantly between cohorts as well as gender using an additive model. We have also demonstrated there is no significant interaction between males and females and the cohorts they are located in.

We have shown consitently that ALB levels and A/G ratios do not differ significantly between cohorts or genders, except in the case of ALB levels in our additive two way MANOVA model where ALB levels did vary significantly between male and female. This is most likely due to the relatively low variances in these variables with 0.633 and 0.104 respectively for ALB and A/G ratios

This testing could further be improved by using the selector variable of liver and non-liver patients to determine just who the most at risk patients are for example the patients with very high TB levels are almost all liver patients. These people will have to seek further testing from medical professionals to asses if they have liver dysfunction as there levels fall way outside normal ranges.

Thank you for reading this document have a good day and stay safe. :)