KU LEUVEN
ANALYSIS OF VARIANCE

Seminar 1

Group 8

Students:

Hankun Ding (r0772442)

Kendall Brown (r0773111)

Shoera Sels (q0473426)

Xinyi Xu (r0766854)

Yang Hu (r0773663)

Yue Huang (r0768557)

Professor:

Ariel Alonso Abad

13 November, 2019

R Seminar 1: Discussion of the analysis

Task:

A model is to be built by examining the relationship between social anxiety and a person's state of mental health. Social anxiety is measured on a continuous spectrum with higher values relating directly to higher levels of social anxiety. The test sample is divided into four groups: the healthy, the depressed, the social anxious but not depressed, and those that suffer from both social anxiety and depression.

To begin the analysis, the data shall be explored graphically with boxplots. Doing so will provide a clear representation of the between group relationships and provide insight into what to expect from future results.
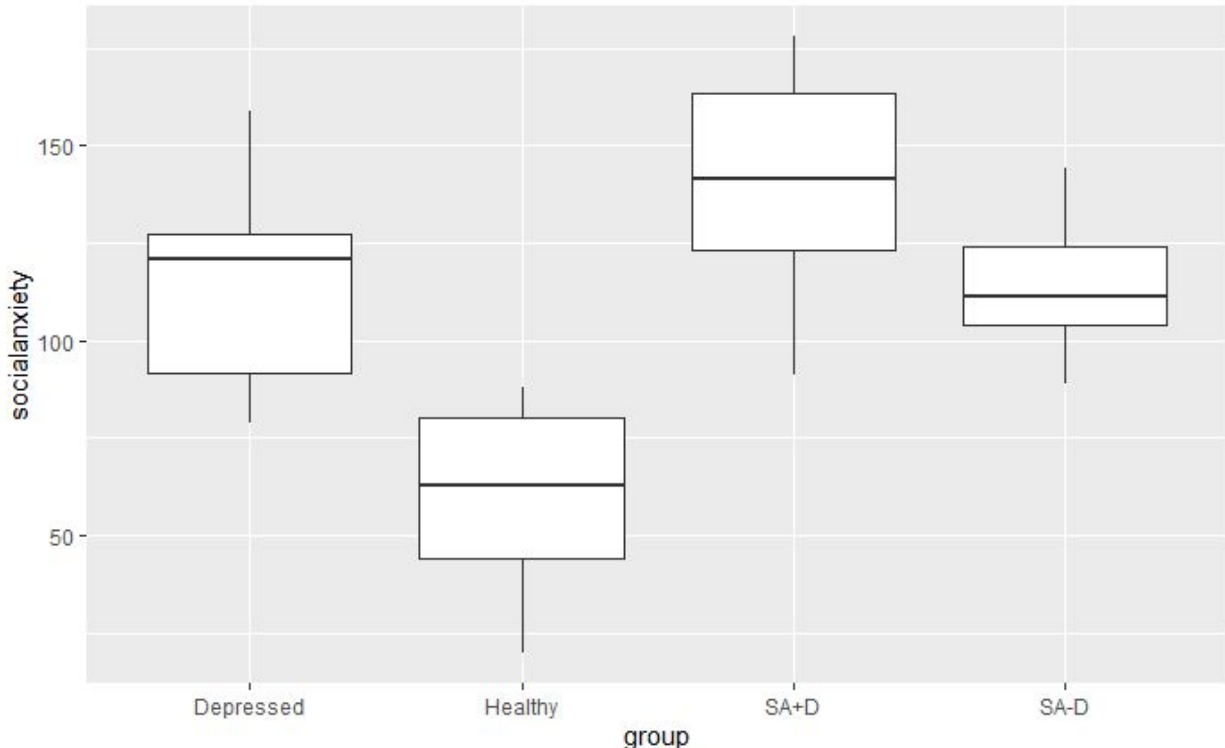
We can derive that healthy people have the lowest level of social anxiety, the levels for depressed group and anxiety without depression group are basically the same and the anxiety with depression group's level is slightly higher than them. The variability of social anxiety in the groups of depressed group, healthy group and anxiety with depression group has no obvious difference, yet compared to them, the variability of anxiety with depression group is much smaller. Besides, there is one outlier being detected in the anxiety with depression group, which is the 69th observation. The outlier will be removed for the following analyses.

Shown below is a table containing the calculated means and standard deviations of each group.

Group	Mean	Standard Deviation
Healthy Control	60.73	21.00
Depressed	112.36	23.30
Socially Anxious and not Depressed	113.59	14.34
Socially Anxious and Depressed	141.92	24.37

With the values shown in the table, the conclusions drawn from the previous graphical exploration are confirmed. For the average social anxiety level, the healthy group is the lowest, the socially anxious and depressed group is the highest and the other two are almost the same. The variabilities of the groups are about the same except for the socially anxious and depressed group, which is much smaller than others.

To advance the analysis a one way ANOVA model will be used to determine if the observed variation between groups is significant. To accomplish this a null and alternative hypothesis are to be established.

Null hypothesis: The mean of each group is the same.
Alternative hypothesis: At least one group has a different mean value.

Provided below is a summary of the one way ANOVA model's results.

```
          Df Sum Sq Mean Sq F value Pr(>F)
group      3 119276   39759   94.59 <2e-16 ***
Residuals 135  56746     420
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results of the ANOVA model prove to be significant. The treatment mean square (with the degree of freedom as r - 1 = 4 - 1 = 3) is about 94.59 times of error mean square (with the degree of freedom as N - r = 139 - 4 = 5), which is almost impossible (the p-value is extremely small) under the null hypothesis. Therefore, at the 5% confidence level, the null hypothesis that there is no difference between group means can be rejected.

To have a further look at the relationship between each pair of the four variables, Tukey, Scheffe and Bonferroni multiple comparisons of means were implemented and the comparison plot was drawn:

```
Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = socialanxiety ~ group, data = SA_binary)

$group
                                                                         diff
healthy controls-depressed                                           -51.63636
social anxiety with depression-depressed                              29.55742
social anxiety without depression-depressed                            1.22332
social anxiety with depression-healthy controls                       81.19378
social anxiety without depression-healthy controls                    52.85968
social anxiety without depression-social anxiety with depression     -28.33410
                                                                          lwr
healthy controls-depressed                                           -66.31568
social anxiety with depression-depressed                              15.26962
social anxiety without depression-depressed                          -12.60142
social anxiety with depression-healthy controls                       68.50344
social anxiety without depression-healthy controls                    40.69305
social anxiety without depression-social anxiety with depression     -40.02537
                                                                          upr
healthy controls-depressed                                           -36.95705
social anxiety with depression-depressed                              43.84521
social anxiety without depression-depressed                           15.04806
social anxiety with depression-healthy controls                       93.88412
social anxiety without depression-healthy controls                    65.02632
social anxiety without depression-social anxiety with depression     -16.64283
                                                                        p adj
healthy controls-depressed                                           0.0000000
social anxiety with depression-depressed                             0.0000019
social anxiety without depression-depressed                          0.9956835
social anxiety with depression-healthy controls                      0.0000000
social anxiety without depression-healthy controls                   0.0000000
social anxiety without depression-social anxiety with depression     0.0000000
```

```
Posthoc multiple comparisons of means : Scheffe Test
    95% family-wise confidence level

$group
                                                                         diff
healthy controls-depressed                                           -51.63636
social anxiety with depression-depressed                              29.55742
social anxiety without depression-depressed                            1.22332
social anxiety with depression-healthy controls                       81.19378
social anxiety without depression-healthy controls                    52.85968
social anxiety without depression-social anxiety with depression     -28.33410
                                                                        lwr.ci
healthy controls-depressed                                           -67.61224
social anxiety with depression-depressed                              14.00764
social anxiety without depression-depressed                          -13.82250
social anxiety with depression-healthy controls                       67.38256
social anxiety without depression-healthy controls                    39.61843
social anxiety without depression-social anxiety with depression     -41.05800
                                                                        upr.ci
healthy controls-depressed                                           -35.66049
social anxiety with depression-depressed                              45.10719
social anxiety without depression-depressed                           16.26914
social anxiety with depression-healthy controls                       95.00500
social anxiety without depression-healthy controls                    66.10094
social anxiety without depression-social anxiety with depression     -15.61019
                                                                         pval
healthy controls-depressed                                           4.2e-14
social anxiety with depression-depressed                             8.1e-06
social anxiety without depression-depressed                          0.9968
social anxiety with depression-healthy controls                      < 2e-16
social anxiety without depression-healthy controls                   < 2e-16
social anxiety without depression-social anxiety with depression     1.2e-07
```

```
Multiple Comparisons of Means: User-defined Contrasts


Fit: aov(formula = socialanxiety ~ group, data = SA_binary)

Linear Hypotheses:
                          Estimate Std. Error t value Pr(>|t|)
healthy - depressed == 0    -51.636      5.643  -9.150 5.33e-15 ***
SA with De.-depressed == 0   29.557      5.493   5.381 1.90e-06 ***
SA without De.-depressed == 0 1.223      5.315   0.230        1
SA with De.-healthy == 0     81.194      4.878  16.643  < 2e-16 ***
SA without De.-healthy == 0  52.860      4.677  11.302  < 2e-16 ***
SA without De.-SA with De. == 0 -28.334  4.494  -6.304 2.29e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- bonferroni method)
```
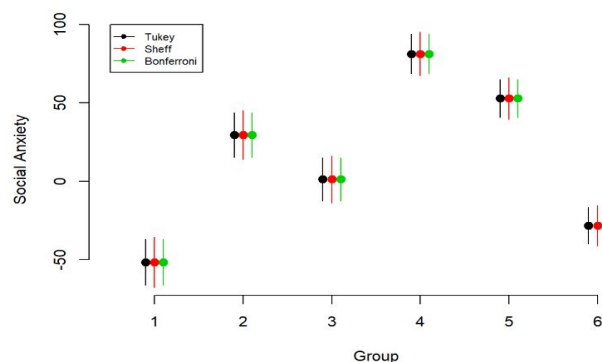
The three tests give similar results. Only the difference in social anxiety between socially anxious and not depressed group and depressed group is not significant. Differences between all other pairs are highly significant.

The left plot compares the confidence intervals given by the three methods. There are no obvious differences between them.

In addition, the plot shows that the Bonferroni method gives the narrowest interval in the case, while Scheffe method gives the widest interval because it is for all hypotheses.

On the analysis above, it can be concluded that the state of a person's mental health has a relation with their level of social anxiety. To determine the validity of the model, an analysis of the residuals plots and several formal tests shall be provided below.
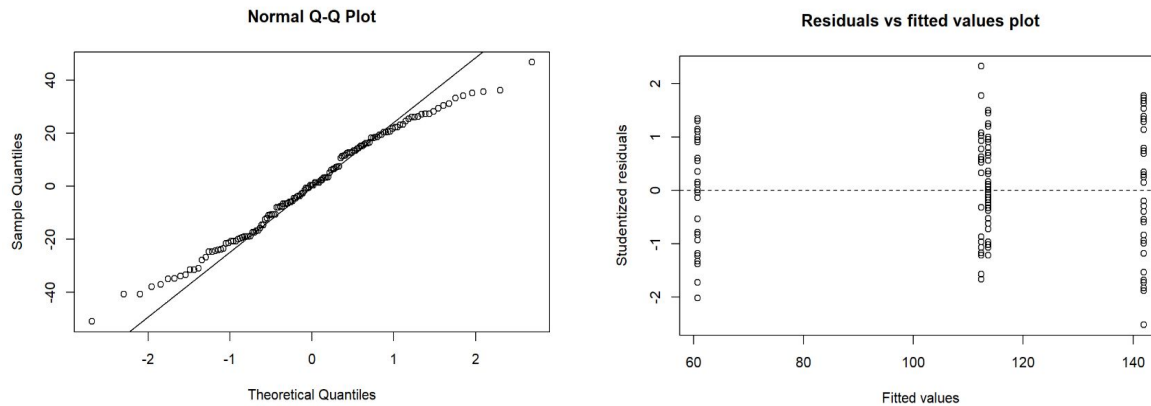
From the normal Q-Q plot, we can see that the first quantiles occurring at larger than expected values and the last quantiles occuring at less than expected values. That is, the sample distribution has thinner tails than a normal distribution. So the assumption of normality may be violated.

In the residual plot, for the given fitted values, the standardized residuals seem not to be the same. Hence, there is also the need to test the homogeneity assumption. To test the assumptions of homogeneity, non-normality and independence of errors formally, Levene's test, Shapiro-Wilk test and Durbin Watson test were applied.

```
Levene's Test for Homogeneity of Variance (center = median)
       Df F value   Pr(>F)
group   3  4.5241 0.004671 **
       136
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


        Shapiro-Wilk normality test

data:  residuals(anovfit)
W = 0.98766, p-value = 0.2473



 lag Autocorrelation D-W Statistic p-value
   1      -0.0929765       2.170589   0.468
 Alternative hypothesis: rho != 0
```

For Levene's test, the p-value equals 0.004671 which is smaller than 0.05, the value of α we prefer, thus the null hypothesis can be rejected at 5% level of significance. Therefore, homoscedasticity does not exist. Solving approached like taking transformation or using weighted least squares could be considered for further improvement.

For Shapiro-Wilk test, the p-value equals 0.2473 which is greater than 0.05, the value of α we prefer, thus the null hypothesis can not be rejected at 5% level of significance. Therefore, the assumption of the error's normality is valid.

For Durbin Watson test, the p-value equals 0.468 which is greater than 0.05, the value of α we prefer, thus the null hypothesis can not be rejected at 5% level of significance. Therefore, the assumption of the error's independence is valid.

APPENDIX

R Code
```r
```{r, eval=FALSE, include=FALSE}
rm(list=ls())
library(dplyr)
library(MASS)
library(tidyverse)
library(ggplot2)
library(multcomp)
library(car)
```

```{r}
#SA=read.csv(choose.files(),header = T)
summary(SA)
#SA_binary=SA[SA$sex=="male" | SA$sex=="female",]
#single unidentified sex observation removed.
SA_binary=SA[-69,]
summary(SA_binary)
ggplot(data=SA_binary)+
  geom_point(mapping=aes(x=group,y=socialanxiety))
ggplot(data=SA_binary)+
  geom_boxplot(mapping=aes(x=sex,y=socialanxiety,color=group))
ggplot(data=SA_binary)+
  geom_boxplot(mapping=aes(x=group,y=socialanxiety))+
  scale_x_discrete(labels = c("Depressed","Healthy","SA+D","SA-D"))
#plot(SA_binary$socialanxiety[SA_binary$group=="social anxiety without depression"])
```

```{r}
print("Healthy Group")
print(c("Mean",mean(SA_binary[SA_binary$group=='healthy controls',]$socialanxiety)))
print(c("Standarad Deviaiton",sd(SA_binary[SA_binary$group=='healthy
controls',]$socialanxiety)))
cat("\n")

print("Depressed Group")
print(c('Mean',mean(SA_binary[SA_binary$group=='depressed',]$socialanxiety)))
```

```
print(c('Standarad Deviation',sd(SA_binary[SA_binary$group=='depressed',]$socialanxiety)))
cat("\n")

print("Socially Anxious and not Depressed Group")
print(c("Mean",mean(SA_binary[SA_binary$group=="social anxiety without
depression",]$socialanxiety)))
print(c('Standard Deviation',sd(SA_binary[SA_binary$group=="social anxiety without
depression",]$socialanxiety)))
cat("\n")

print("Socially Anxious and Depressed Group")
print(c("Mean",mean(SA_binary[SA_binary$group=="social anxiety with
depression",]$socialanxiety)))
print(c("Standard Deviation",sd(SA_binary[SA_binary$group=="social anxiety with
depression",]$socialanxiety)))
```

```{r}
SA_aov=aov(socialanxiety~group,data=SA_binary)
summary(SA_aov)
model.tables(SA_aov,"means")

##Tukey
TukeyHSD(SA_aov)
library(multcomp)
tukey = glht(SA_aov, linfct = mcp(group= "Tukey"))
summary(tukey)
tukey_ci = confint(tukey)

## ScheffeTest
library(DescTools)
all_pairwise = rbind("healthy - depressed" = c(-1, 1, 0, 0),"SA with De.-depressed" = c(-1, 0,
1,0),"SA without De.-depressed" = c(-1, 0, 0, 1),"SA with De.-healthy"=c(0,-1,1,0),"SA without
De.-healthy"=c(0,-1,0,1),"SA without De.-SA with De."=c(0,0,-1,1))

(sheffe=ScheffeTest(SA_aov, contrasts=t(all_pairwise)))


##Bonferroni
```

```
bonferroni = glht(SA_aov, linfct = mcp(group = all_pairwise),test = adjusted("bonferroni"))
summary(bonferroni, test = adjusted("bonferroni"))
bonferroni_ci = confint(bonferroni)

##Plot
plot(tukey_ci$confint[,1]~c(0.9,1.9,2.9,3.9,4.9,5.9),xlab="Group",ylab="Social
Anxiety",axes=F,ylim=c(min(tukey_ci$confint,sheffe$ci_sheffe,bonferroni_ci$confint),max(tuk
ey_ci$confint,sheffe$ci_sheffe,bonferroni_ci$confint)), xlim=c(0.5,6.5))
group_means = c(aggregate(SA_binary$socialanxiety,list(SA_binary$group),mean))
points(tukey_ci$confint[,1]~c(0.9,1.9,2.9,3.9,4.9,5.9),pch=19,cex=1.2)
points(sheffe$group[,1]~c(1:6),pch=19,cex=1.2,col=2)
points(bonferroni_ci$confint[,1]~c(1:6+0.1),pch=19,cex=1.2,col=3)
axis(side=1)
axis(side=2)
for (i in 1:6) {
  segments(i-0.1,tukey_ci$confint[i,2],i-0.1,tukey_ci$confint[i,3])
segments(i,sheffe$group[i,2],i,sheffe$group[i,3],col=2)
segments(i+0.1,bonferroni_ci$confint[i,2],i+0.1,bonferroni_ci$confint[i,3],col=3)
}

legend(0.5,100,pch=c(19,19,19),lty=c(1,1,1),col=c(1,2,3),legend=c("Tukey","Sheff","Bonferroni
"),cex = 0.7)
```

```{r}
plot(fitted.values(SA_aov),rstandard(SA_aov),
xlab="Fitted values",
ylab="Studentized residuals",
main="Residuals vs fitted values plot")
abline(h=0,lty="dashed")
qqnorm(residuals(SA_aov))
qqline(residuals(SA_aov))
```

Testing Assumptions
```{r}
leveneTest(socialanxiety~group,data=SA_binary)
#reject the null, reject homoscedasicity
shapiro.test(residuals(SA_aov))
```

```
#fail to reject null, fail to reject non-normality
durbinWatsonTest(SA_aov, alternative="two.sided", data=SA_binary)
#fail to reject null, fail to reject independence
```
```

Additional Exercises

1. Answer: d

Justification:

The tested hypotheses under F-test for one-way ANOVA are:

H_0: $\mu_1 = \mu_2 = \ldots = \mu_r$

H_1: at least one μ_i is different, for i =1, 2,......r

Since F-statistic is significant, which means the effect being tested is significant and the null hypothesis can be rejected so that the means being tested are not equal. The hypotheses can be rejected either the test statistic is greater than the critical statistic, or the associated p-value is smaller than α. Consequently, as the alternative to test the hypotheses, the latter is expected to be the default once the former has been true. Therefore, all the statements have been proved.

2. Answer: c

Justification:

Here, the Levene's test statistic (F=0.0069) should be compared with the corresponding critical statistics rather than α. While, p-value here is 0.9656 which is greater than α, normally be 0.05, indicating that there is no evidence to reject the null hypothesis that the variances are the same. Thus, (b) (d) (e) are false. Moreover, test statistic and p-value have nothing to so with validity of test, so (a) is false.

From the ANOVA table, the test statistic equals 0 and the p-value is 0.9578, which is much greater than 0.05. Hence, we can not reject the null hypothesis that the group mean is the same. (c) is the right choice.

3. Answer: b

Justification:

The number of treatments or groups K is equal to the degree of freedom of model plus 1, which formulated as

$K = r = 3+1 = 4$

4. Answer: d

Justification:

N indicates the total number of observations which equal to the degree of freedom of total plus 1, which formulated as

$N = n_T = 23+1 = 24$

5. Answer: a

Justification:

According to the scenario, we could know that

SSTR = SSTO - SSE = 51.380 - 9.205 = 42.175

DF of ERROR = n_T - r = 24 - 4 = 20

$$F = \frac{MSTR}{MSE} = \frac{SSTR/r-1}{SSE/nT-r} = \frac{42.175/3}{9.205/20} = \frac{14.0583}{0.4603} = 30.545$$

6. <u>Answer:</u> b

<u>Justification:</u>

The critical statistic of F-test is F(1 - α, r - 1, n_T - r). Thus,

df_1 = r - 1 = 3, df_2 = n_T - r = 20

7. <u>Answer:</u> b

<u>Justification:</u>

The plot of residuals versus fitted values is used to check the homogeneity assumption. The assumption requires that each error terms should have the same variance, as the estimate of error terms, residuals should also have the same variability. From the plot, given fitted values, variances of residuals differ significantly from each other. Therefore, the homogeneity assumption is violated.

8. <u>Answer:</u> d

<u>Justification:</u>

One-way ANOVA F-test is used to test whether the means are equal or not. The F statistic is conducted as F = $\frac{MSTR}{MSE}$, the fraction of two mean sum of squares so it should never be negative. In summary, all of the above are true, so (d) is the right choice.

9. <u>Answer:</u> c

<u>Justification:</u>

The normality assumption and homoscedasticity assumption are two independent assumptions, so one assumption being violated has no effect on the other. Hence, (a) is false. Departures from the homoscedasticity assumption can not be offset by sample size, in contrast, in the unbalanced sample where only one of the groups has a large sample size may lead to the easy violation of homoscedasticity assumption, (b) is false. Levene test is used to check homoscedasticity, (c) is right.

10. <u>Answer:</u> NO

<u>Justification:</u>

As the scenario mentioned, all the research participants are in the same room and it is possible that some people overhead the answer of nearby respondents. Consequently, the

answers from these three groups are highly lack of independence. Therefore, the assumption of independence is violated in the present situation.

11.
a. Preferred Model: one-way ANOVA
Reasons: Basically, the basic idea here is that the comparison is between the mean of different groups in the same factor. Thus, one-way ANOVA will be used for the analysis.

b. Hypothesis:
H0: $\mu_1 = \mu_2 = \ldots = \mu_r$
H1: at least one μ_i is different, for i =1, 2,......r

c. Assumptions:
1) Independent;
2) Normality;
3) Homoscedasticity.
Plausibility:
i. All the persons have been randomly assigned to four groups of treatment and each group has the same number of people, so the assumptions of independence and homoscedasticity are satisfied;
ii. The normality assumption is also plausible here since the Central Limit Theory might be used to ensure normality when the size of the sample (n = 400) is large enough.
Additionally, there might be a need to discuss the outlier problem as well because the size of the given sample is large (n = 400) so that it is highly possible exists some individual cases which are not in agreement with the vast majority of data.

12. Answer: d
Justification:
When t-test is applied for each group at the 5% significance level, the type I error that at least one test gives false result will be greater than 5%, then (a) is false. To avoid data snooping, we should test all possibilities rather than just what has been observed. Therefore, (b) and (c) are false. Compared to Tukey procedure, Bonfferoni gives narrower confidence interval of group mean when comparing all possible pairwise of groups, so Bonfferoni procedure has more power in the case. Hence, (d) was chosen.