

**KU LEUVEN**  
**ANALYSIS OF VARIANCE**

**Seminar 2**

***Group 8***

***Students:***

Hankun Ding (r0772442)

Kendall Brown (r0773111)

Xinyi Xu (r0766854)

Yang Hu (r0773663)

Yue Huang (r0768557)

***Professor:***

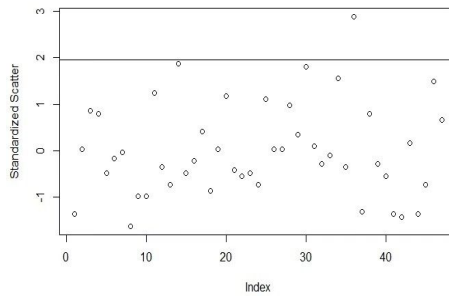
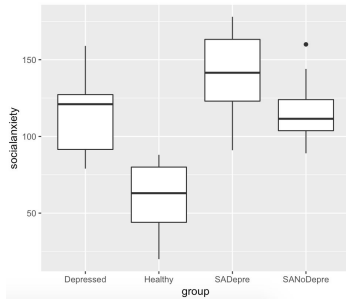
Ariel Alonso Abad

**05 December 2019**



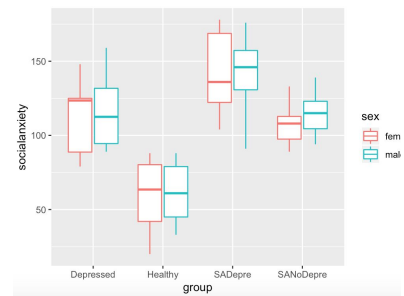
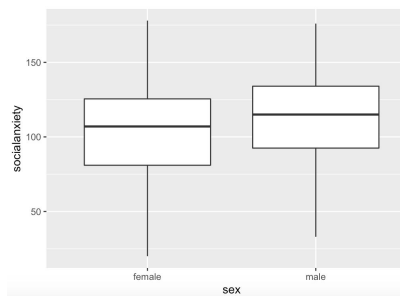
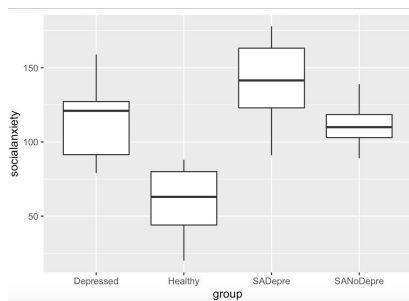
## R Seminar 2: Discussion of the analysis

In seminar 2, the same dataset as seminar 1 has been provided. Before we start, we should notice that this was an unbalanced study, so that the samples in different groups were not of the same number.



In the pre-processing, an observation that has a null value in the “sex” variable was removed (observation 58). In addition, when we first plotted the boxplot of social anxiety levels by

different groups, an outlier in the group “social anxiety without depression” was detected, which is the observation 69. After such data cleaning, boxplots of the levels of social anxiety by groups, gender and interaction of them were plotted separately as follows.



From the left figure, 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 the 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. Regarding the middle one, there is no obvious difference in anxiety levels across gender. When we consider the two factors together, as the right figure shows, for each group the mean and variation of anxiety level are similar for both males and females.

Mean and Standard Deviation Within Groups				
	Healthy	Depressed	Social Anxiety with no depressed	Social Anxiety with depressed
Mean	60.727	112.364	111.767	141.921
Standard Deviation	21.000	23.296	12.893	24.375

Mean and Standard Deviation Within Gender		
	Male	Female
Mean	110.9	105.539
Standard Deviation	33.385	37.639

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

Mean and Standard Deviation of Interaction between Groups and Gender								
	Healthy		Depressed		Social Anxiety with no depressed		Social Anxiety with depressed	
	Male	Female	Male	Female	Male	Female	Male	Female
Mean	62	59.9	115.75	110.429	115.261	107.75	141.938	141.909
Standard Deviation	19.088	22.602	24.511	23.283	12.585	12.341	20.930	27.091

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. Besides, the average social anxiety levels of males and females are basically the same. There is also no obvious interaction effect being found.



From the interaction plot, the curves of social anxiety levels for males and females are almost parallel, so it can be inferred that there is no significant interaction effect between gender and anxiety groups.

	female	male
depressed	14	8
healthy controls	20	13
social anxiety with depression	22	16
social anxiety without depression	22	23

The left frequency table shows that this study design is unbalanced, therefore, the variance of observations can not be partitioned by main effects and interaction effect directly:

$$SSTR \neq SSA + SSB + SSAB$$

$$SSTO \neq SSA + SSB + SSAB + SSE$$

In this case, the cell mean model, factor effect model, Type I Sum of Squares and Type III Sum of Squares Analyses will be discussed in the following to test the main effect[ of social anxiety group, sex and their interaction effect.

### Cell mean model

Firstly, we consider each combination of gender and social anxiety group as a treatment, the model conducted as:

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \quad i = 1, 2, 3, 4, \quad j = 1, 2, \quad k = 1, \dots, n_{ij}$$

Here,  $i$  denotes gender,  $j$  denotes social anxiety group and  $k$  denotes the  $k$ th observation in the  $ij$  group. To test the main effect of group, the null hypothesis and the alternative hypothesis are made as:

$$H_0 : \mu_{11} + \mu_{12} = \mu_{21} + \mu_{22}, \mu_{21} + \mu_{22} = \mu_{31} + \mu_{32}, \mu_{31} + \mu_{32} = \mu_{41} + \mu_{42}$$

vs.  $H_1$  : Not all of them are equal.

To test the main effect of sex, the null hypothesis and the alternative hypothesis are made as:

$$H_0 : \mu_{11} + \mu_{21} + \mu_{31} + \mu_{41} = \mu_{12} + \mu_{22} + \mu_{32} + \mu_{42} \text{ vs. } H_1 : \text{They are not equal.}$$

To test the interaction effect of sex and social anxiety group, the null hypothesis and the alternative hypothesis are made as:

$$H_0 : \mu_{11} - \mu_{12} = \mu_{21} - \mu_{22}, \mu_{21} - \mu_{22} = \mu_{31} - \mu_{32}, \mu_{31} - \mu_{32} = \mu_{41} - \mu_{42}$$

vs.  $H_1$  : Not all of them are equal.

General Linear Hypotheses	General Linear Hypotheses	General Linear Hypotheses
Linear Hypotheses: Estimate 1 == 0   -11.71	Linear Hypotheses: Estimate 1 == 0   104.28 2 == 0   -161.95 3 == 0   57.59	Linear Hypotheses: Estimate 1 == 0   -3.221 2 == 0   -2.072 3 == 0   4.232
Global Test: F DF1 DF2 Pr(>F) 1 0.6118   1 130 0.4355	Global Test: F DF1 DF2 Pr(>F) 1 88.21   3 130 3.388e-31	Global Test: F DF1 DF2 Pr(>F) 1 0.1015   3 130 0.959

The above tables show the significant results corresponding to the hypotheses stated before. The left table tests the main effect of the sex factor in the case, suggesting that there is no significant difference in social anxiety levels for males and females. That is, no significant main effect of sex was shown. Similarly, from the right table, we can also derive there is no significant interaction effect of sex and anxiety group here, as we inferred from the interaction plot. However, highly significant differences among social anxiety groups are detected, indicating the strong main effect of social anxiety group.

## Factor effect model

Now, let us consider the factor effect model, whose form is as follows:

$$Y_{ijk} = \bar{\mu}_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \quad i = 1, 2, 3, 4, \quad j = 1, 2, \quad k = 1, \dots, n_{ij}$$

Here,  $\alpha_i$  denotes the main effect of group,  $\beta_j$  denotes the main effect of sex and  $(\alpha\beta)_{ij}$  denotes their interaction effect. Besides,  $\bar{\mu}_{..}$  is a reference parameter here, as the total mean of all observations. In this model, when testing effect we compare the full model with the reduced model in which the tested effect is deleted.

#### Analysis of Variance Table

```
Model 1: anxiety ~ (group1 + group2 + group3 + sex1 + 'group1:sex1' +
'group2:sex1' + 'group3:sex1') - group1 - group2 - group3
Model 2: anxiety ~ group1 + group2 + group3 + sex1 + 'group1:sex1' + 'group2:sex1' +
'group3:sex1'
Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1      133 170044
2      130  56016  3    114028 88.211 < 2.2e-16 ***
```

#### Analysis of Variance Table

```
Model 1: anxiety ~ group1 + group2 + group3 + sex1
Model 2: anxiety ~ group1 + group2 + group3 + sex1 + 'group1:sex1' + 'group2:sex1' +
'group3:sex1'
Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1      133  56147
2      130  56016  3    131.27 0.1015  0.959
```

#### Analysis of Variance Table

```
Model 1: anxiety ~ (group1 + group2 + group3 + sex1 + 'group1:sex1' +
'group2:sex1' + 'group3:sex1') - sex1
Model 2: anxiety ~ group1 + group2 + group3 + sex1 + 'group1:sex1' + 'group2:sex1' +
'group3:sex1'
Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1      131  56280
2      130  56016  1    263.61 0.6118  0.4355
```

These three analyses of variance tables show the results of significance testing based on the factor effect model. They are corresponding to the main effect of anxiety group, the main effect of sex and

their interaction effect, respectively. They give the same result as the tests based on the cell mean model, that is, not significant sex main effect, interaction effect and highly significant group main effect.

## Type I and Type III sum of squares analysis

```

      Df Sum Sq Mean Sq F value Pr(>F)
sex      1      653      653   1.515  0.221
group    3 118670  39557  91.802 <2e-16 ***
sex:group 3      131       44   0.102  0.959
Residuals 130  56016      431
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```

      Df Sum Sq Mean Sq F value Pr(>F)
group    3 119071  39690  92.112 <2e-16 ***
sex       1      252      252   0.584  0.446
group:sex 3      131       44   0.102  0.959
Residuals 130  56016      431
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```

      Df Sum Sq Mean Sq F value Pr(>F)
group    3 119071  39690  92.112 <2e-16 ***
sex       1      252      252   0.584  0.446
group:sex 3      131       44   0.102  0.959
Residuals 130  56016      431
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The table on the upright is the Type I analysis. Because Type I test is order dependent. Thus different cases of group\*sex and sex\*group were shown as above. In both cases, the main effect of

group is significant and both the main effect of sex and the interaction effect of sex and group are insignificant.

The left table shows the result of Type III ANOVA sum of square test. The results of the significance tests are consistent with the previous tests. Only factor group has a significant main effect.

## Tukey Procedure

1. Comparison of simple effect: Group

<b>Difference among groups for Female</b>	Difference	Lower	Upper	P adj
healthy controls: female-depressed: female	-50.529	-72.814	-28.243	0.000
social anxiety with depression: female-depressed: female	31.481	9.616	53.345	0.001
social anxiety without depression: female-depressed: female	0.571	-21.293	22.436	1.000
social anxiety with depression: female-healthy controls: female	82.009	62.250	101.768	0.000
social anxiety without depression: female-healthy controls: female	51.100	31.341	70.859	0.000
social anxiety without depression: female-social anxiety with depression: female	-30.909	-50.192	-11.626	0.000

<b>Difference among groups for Male</b>	Difference	Lower	Upper	P adj
healthy controls: male-depressed: male	-53.750	-82.488	-25.012	0.000
social anxiety with depression: male-depressed: male	26.188	-1.505	53.880	0.078
social anxiety without depression: male-depressed: male	-0.489	-26.739	25.761	1.000
social anxiety with depression: male-healthy controls: male	79.938	56.058	103.817	0.000
social anxiety without depression: male-healthy controls: male	53.261	31.070	75.452	0.000
social anxiety without depression: male-social anxiety with depression: male	-26.677	-47.496	-5.857	0.003

For female's social anxiety levels, five conclusions can be made: Firstly, the social anxiety levels of healthy-controls are significantly lower than the depressed group, social anxiety with depression group and social anxiety without depression group. Besides, the social anxiety level

of anxiety with depression group is significantly higher than the depressed group and the social anxiety without depression group.

For male, except that the difference in anxiety level between depressed group and social anxiety with depression group is not significant anymore, the relationships between any other pairs of groups keep the same with the female's.

## 2. Comparison of simple effect: Sex

	Difference	Lower	Upper	P adj
Depressed:male-depressed:female	5.321	-23.023	33.666	0.999
healthy controls:male-healthy controls:female	2.100	-20.684	24.884	1.000
social anxiety with depression:male-social anxiety with depression:female	0.028	-20.984	21.041	1.000
social anxiety without depression:male-social anxiety without depression:female	4.261	-14.811	23.333	0.997

From the comparison table, we can confirm that men and women have no difference in social anxiety levels across groups.

## Assumption Testing

To test whether the essential assumptions are met, homogeneity of variance, normality and independence of errors will be checked by Levene's, Durbin-Waston's, and Shapiro-Wilk's test respectively.

```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group   7  2.7754 0.0101 *
      130
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

shapiro-wilk normality test

data:  residuals(SA_aov_mvf)
W = 0.98502, p-value = 0.1373

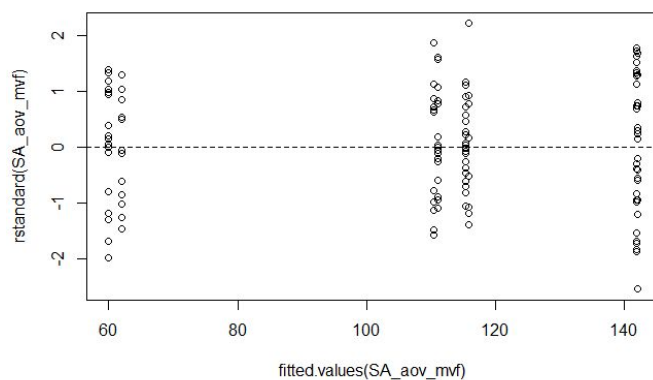
lag Autocorrelation D-W Statistic p-value
1 -0.07942622 2.140551 0.608
Alternative hypothesis: rho != 0
```

For Levene's test, the p-value equals 0.0101 which is smaller than 0.05 so that the null hypothesis can be rejected at 5% level of significance. Therefore, homoscedasticity does not

exist. Comparing with Seminar 1, the variances do not show an effective change which becomes homogeneity.

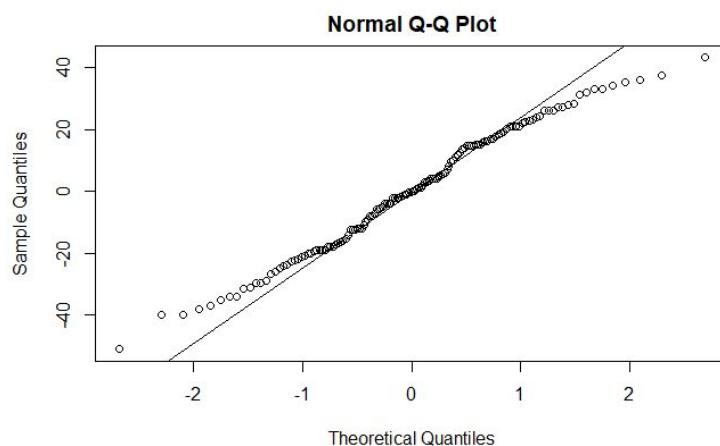
For Shapiro-Wilk's test, the p-value is 0.1373 which is greater than 0.05. Thus, not enough evidence has been shown to reject the null hypothesis, at 5% level of significance, indicating that the assumption of the error's normality is valid which consistent with Seminar 1.

The p-value has been got from the Durbin-Waston's test is 0.608 which is larger than 0.05 which let the null hypothesis could not be rejected at 5% level of significance. As a consequence, similar to Seminar 1, the assumption of the error's independence is valid.



Meanwhile, visualizing the distribution and situations of the relevant tests is also helpful. Firstly, checking the assumption of homoscedasticity is being considered by showing the plot of residuals versus fitted values. The variations are expected to, more or less, be the same with each other. However, moving back to this plot, the variation differences between four groups are not hard to be distinguished.

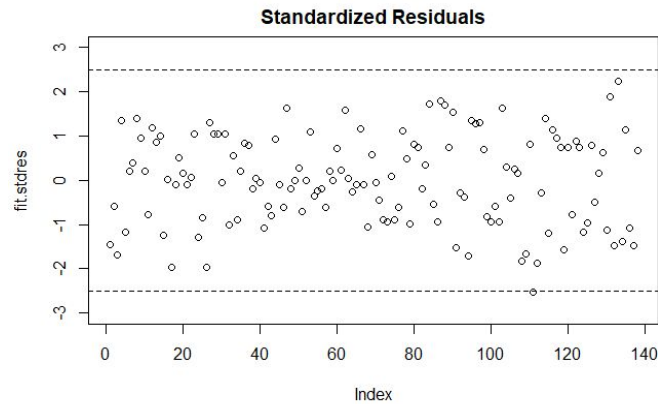
Comparing with Seminar, the distributed situation might have improved but not enough which leads to the homogeneity assumption still have difficulties to meet.



In order to provide more evidence of the validation of the normality assumption, the normal Q-Q plot will be used. The first quantiles occurring at larger than the expected values while the last quantiles present as the opposite which lower than the expected values. Thus, even the other points line up as a line, this plot might still

have a little bit of lack of showing strong evidence to support the assumption of normality.





Lastly, it is also necessary to check whether existing more outliers need to be further considered. With adding horizontal cut-off values to the plot of standardized residuals, the observations whose absolute standardized residuals are larger than 2.5 could be pinpointed as outliers. Under these considerations, shown as the plot, there is no more point out of the boundaries so that no more

outliers need to be removed after fitting with this defined model.

## Conclusion

During Seminar 2, we first deal with outliers by plotting boxplot and standardized deviation scores of all the social anxiety values to find an outlier in the group of social anxiety without depressed. We also find a missing value in the healthy group which is also omitted. From our first step of plotting boxplots with respect to groups and gender. Our first conclusion is that there are differences between groups. The healthy group is the lowest, the socially anxious and depressed group is the highest and the other two are almost the same and no such a big difference to be detected between gender. When we consider the interaction of the two factors, no interaction is found. The variabilities of the groups are about the same except for the socially anxious and depressed group. Then, we show that this is an unbalanced design.

In the two-way ANOVA analysis, the cell mean model and the factor effect model both confirm the conclusions we have at the beginning. Based on Type I and Type III tests, there is no significant result to support the interaction. In Tukey procedure, there is one more result to confirm that gender has no difference in all groups but female has stronger social anxiety with depression compared with depressed than male.

In the end, some assumption tests such as homogeneity, normality and outlier tests are carried out again.

## APPENDIX

### **R Code**

```
```{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)
#remove single unidentified sex observation.
SA_binary=SA[SA$sex=="male" | SA$sex=="female",]
#remove outlier
SA_binary=SA_binary[SA_binary$socialanxiety!=160,]
SA_binary=SA_binary[!(SA_binary$group=="social anxiety without depression" &
SA_binary$socialanxiety>150),]
SA_male=SA_binary[SA_binary$sex=="male",]
SA_female=SA_binary[SA_binary$sex=="female",]

summary(SA_binary)

#Boxplot of SA vs Group
ggplot(data=SA_binary)+
  geom_boxplot(mapping=aes(x=group,y=socialanxiety))+
  scale_x_discrete(labels = c("Depressed","Healthy","SA+D","SA-D"))
#Boxplot of SA vs Sex
ggplot(data=SA_binary)+
  geom_boxplot(mapping=aes(x=sex,y=socialanxiety))

#Boxplot of SA vs Group:Sex
ggplot(data=SA_binary)+
  geom_boxplot(mapping=aes(x=group,y=socialanxiety,color=sex))+
  scale_x_discrete(labels = c("Depressed","Healthy","SA+D","SA-D"))
...

```

#No Immediately obvious relation between sex and social anxiety or between sex and group.

#Social anxiety vs grouping

```
``{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)))
``
```

#Social anxiety vs sex

```
``{r}
print("Male")
print(c("Mean",mean(SA_binary[SA_binary$sex=='male'],$socialanxiety)))
print(c("Standarad Deviaiton",sd(SA_binary[SA_binary$sex=='male'],$socialanxiety)))
cat("\n")
```

```

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

print("Depressed Group")
print(c("Male Mean",mean(SA_male[SA_male$group=='depressed'],$socialanxiety)))
print(c("Male Standarad
Deviaiton",sd(SA_male[SA_male$group=='depressed'],$socialanxiety)))
print(c("Female
Mean",mean(SA_female[SA_female$group=='depressed'],$socialanxiety)))
print(c("Female Standarad
Deviaiton",sd(SA_female[SA_female$group=='depressed'],$socialanxiety)))
cat("\n")

print("Socially Anxious and not Depressed Group")
print(c("Male Mean",mean(SA_male[SA_male$group=='social anxiety without
depression'],$socialanxiety)))
print(c("Male Standarad Deviaiton",sd(SA_male[SA_male$group=='social anxiety
without depression'],$socialanxiety)))
print(c("Female Mean",mean(SA_female[SA_female$group=='social anxiety without
depression'],$socialanxiety)))
print(c("Female Standarad Deviaiton",sd(SA_female[SA_female$group=='social anxiety
without depression'],$socialanxiety)))
cat("\n")

print("Socially Anxious and Depressed Group")

```

```

print(c("Male Mean",mean(SA_male[SA_male$group=='social anxiety with
depression'],$socialanxiety)))
print(c("Male Standarad Deviaiton",sd(SA_male[SA_male$group=='social anxiety with
depression'],$socialanxiety)))
print(c("Female Mean",mean(SA_female[SA_female$group=='social anxiety with
depression'],$socialanxiety)))
print(c("Female Standarad Deviaiton",sd(SA_female[SA_female$group=='social anxiety
with depression'],$socialanxiety)))
...

#One Way anovas
#SA vs group
```{r}
SA_aov=aov(socialanxiety~group,data=SA_binary)
summary(SA_aov)
model.tables(SA_aov,"means")
TukeyHSD(SA_aov)
...

#SA vs sex
```{r}
SA_aov=aov(socialanxiety~sex,data=SA_binary)
summary(SA_aov)
model.tables(SA_aov,"means")
TukeyHSD(SA_aov)
...

#SA vs Sex nd Group
```{r}
SA_aov=aov(socialanxiety~sex*group,data=SA_binary)
summary(SA_aov)
model.tables(SA_aov,"means")
TukeyHSD(SA_aov)
...

#2-way anova
```{r}
SA_aov_mvf=(aov(socialanxiety~group*sex,data=SA_binary))
summary(SA_aov_mvf)
model.tables(SA_aov_mvf,"means")
...

#Tukey
```{r}

```

```

TukeyHSD(SA_aov_mvf)
...

```{r}
leveneTest(socialanxiety~group*sex,data=SA_binary)
#reject the null, reject homoscedasticity
shapiro.test(residuals(SA_aov_mvf))
#fail to reject null, fail to reject non-normality
durbinWatsonTest(SA_aov_mvf, alternative="two.sided", data=SA_binary)
#fail to reject null, fail to reject independence
...

#For interaction plot
```{r}
interaction.plot(group, sex, socialanxiety, main = "Interaction plot", cex=0.8)
...

table(SA_binary$group,SA_binary$sex)

#Test based on the cell mean model
```{r}
lm.SA <- lm(socialanxiety~-1+group:sex,data=SA_binary)
summary(lm.SA)
group.unbalanced <-
matrix(c(1,0,0,-1,1,0,0,-1,1,0,0,-1,1,0,0,-1,1,0,0,-1),nrow=3)
main.group <- glht(lm.SA, linfct = group.unbalanced)
summary(main.group,test =Ftest())

sex.unbalanced <- matrix(c(1,1,1,1,-1,-1,-1,-1),nrow=1)
main.sex <- glht(lm.SA, linfct = sex.unbalanced)
summary(main.sex,test =Ftest())

int.unbalanced <- matrix(c(1,0,0,-1,1,0,0,-1,1,0,0,-1,-1,0,0,1,-1,0,0,1,-1,0,0,1),nrow=3)
int <- glht(lm.SA, linfct = int.unbalanced)
summary(int,test =Ftest())

#Test based on the factor effect model
SA_binary$sex <- factor(SA_binary$sex,levels = levels(SA_binary$sex)[2:3])
contrasts(SA_binary$group)=contr.sum
contrasts(SA_binary$sex)=contr.sum

```

```

regre.data.unbalanced=as.data.frame(model.matrix(~ group*sex, SA_binary, contrasts =
list(sex="contr.sum",group="contr.sum"))[,2:8])
regre.data.unbalanced$anxiety=SA_binary$socialanxiety
regre.data.unbalanced

```

```

full.unbalanced=lm(anxiety~.,data = regre.data.unbalanced)
summary(full.unbalanced)
anova(full.unbalanced)

```

```

NoInt.unbalanced=lm(anxiety~group1+group2+group3+sex1, data =
regre.data.unbalanced)
summary(NoInt.unbalanced)
anova(NoInt.unbalanced,full.unbalanced)

```

```

NoA.unbalanced=lm(anxiety~.-group1-group2-group3,data = regre.data.unbalanced)
anova(NoA.unbalanced,full.unbalanced)

```

```

NoB.unbalanced=lm(anxiety~.-sex1,data = regre.data.unbalanced)
anova(NoB.unbalanced,full.unbalanced)
...

```

```

#additional plots for assumption checking and possible missed outlier finding

```

```

```{r}
#plot for checking the assumption of homoscedasticity
print(plot(fitted.values(SA_aov_mv),rstandard(SA_aov_mv)))
abline(h=0, lty="dashed")

```

```

#plot for checking the assumption of normality
qqnorm(residuals(SA_aov_mv))
qqline(residuals(SA_aov_mv))

```

```

#plot for checking whether exists any more outliers
fit.stdres <- stdres(SA_aov_mv)
plot(fit.stdres, ylim = c(-3,3), main = "Standardized Residuals")
abline(h = -2.5, lty = 2)
abline(h = 2.5, lty = 2)
fit.stdres
...

```

**#For setting different reference groups**

```

``{r}
##ref 1 = F:D
sex = relevel(sex, ref = "female")
group = relevel(group, ref = "depressed")
contrasts(sex) = contr.treatment
contrasts(group) = contr.treatment
model.matrix(~sex*group, SA_binary)
fit3.type3.ref1 <- aov(socialanxiety~group*sex, data = SA_binary)
drop1(fit3.type3.ref1, ~., test = "F")
##ref 2 = F:H
sex = relevel(sex, ref = "female")
group = relevel(group, ref = "healthy controls")
contrasts(sex) = contr.treatment
contrasts(group) = contr.treatment
model.matrix(~sex*group, SA_binary)
fit3.type3.ref2 <- aov(socialanxiety~group*sex, data = SA_binary)
drop1(fit3.type3.ref2, ~., test = "F")
##ref 3 = F:SA+D
sex = relevel(sex, ref = "female")
group = relevel(group, ref = "social anxiety with depression")
contrasts(sex) = contr.treatment
contrasts(group) = contr.treatment
model.matrix(~sex*group, SA_binary)
fit3.type3.ref3 <- aov(socialanxiety~group*sex, data = SA_binary)
drop1(fit3.type3.ref3, ~., test = "F")
##ref 4 = F:SA-D
sex = relevel(sex, ref = "female")
group = relevel(group, ref = "social anxiety without depression")
contrasts(sex) = contr.treatment
contrasts(group) = contr.treatment
model.matrix(~sex*group, SA_binary)
fit3.type3.ref4 <- aov(socialanxiety~group*sex, data = SA_binary)
drop1(fit3.type3.ref4, ~., test = "F")
##ref 5 = M:D
sex = relevel(sex, ref = "male")
group = relevel(group, ref = "depressed")
contrasts(sex) = contr.treatment
contrasts(group) = contr.treatment
model.matrix(~sex*group, SA_binary)

```



```

fit3.type3.ref5 <- aov(socialanxiety~group*sex, data = SA_binary)
drop1(fit3.type3.ref5, ~., test = "F")
##ref 6 = M:H
sex = relevel(sex, ref = "male")
group = relevel(group, ref = "healthy controls")
contrasts(sex) = contr.treatment
contrasts(group) = contr.treatment
model.matrix(~sex*group, SA_binary)
fit3.type3.ref6 <- aov(socialanxiety~group*sex, data = SA_binary)
drop1(fit3.type3.ref6, ~., test = "F")
##ref 7 = M:SA+D
sex = relevel(sex, ref = "male")
group = relevel(group, ref = "social anxiety without depression")
contrasts(sex) = contr.treatment
contrasts(group) = contr.treatment
model.matrix(~sex*group, SA_binary)
fit3.type3.ref7 <- aov(socialanxiety~group*sex, data = SA_binary)
drop1(fit3.type3.ref7, ~., test = "F")
##ref 8 = M:SA-D
sex = relevel(sex, ref = "male")
group = relevel(group, ref = "social anxiety without depression")
contrasts(sex) = contr.treatment
contrasts(group) = contr.treatment
model.matrix(~sex*group, SA_binary)
fit3.type3.ref8 <- aov(socialanxiety~group*sex, data = SA_binary)
drop1(fit3.type3.ref8, ~., test = "F")
```

```

### **Type I and Type III analysis :**

```

rm(list=ls())
library(dplyr)
library(MASS)
library(tidyverse)
library(ggplot2)
library(multcomp)
library(car)
library(HH)
library(gplots)
library(plyr)

```

```

SA=read.csv("C:/Users/oscar/Desktop/ANOVA/SocialAnxiety.csv",header = T)
#eliminating the #58 observation with missing gender value
SA<- SA [-58, ]

#rename the level for simpler denotation
levels(group)
levels(group) <- c("Depressed", "Healthy", "SAwithD","SAw/oD")
levels(sex) <- c("F", "M")
sex<- factor(sex)
attach(SA)

#show the number of observations in each cell
table(group,sex)
#showing an unbalanced study

#overall view(mean and standard deviation for each cell)
aggregate(socialanxiety,by = list( group,sex),FUN=mean)
aggregate(socialanxiety,by = list( sex,group),FUN=mean)
aggregate(socialanxiety,by = list( group,sex),FUN=sd)
aggregate(socialanxiety,by = list( sex,group),FUN=sd)

#type I aov:variable in different orders just for comparison with type III below
fit.aov1 <- aov(socialanxiety~group*sex)
summary(fit.aov1)

fit.aov2 <- aov(socialanxiety~sex*group)
summary(fit.aov2)

#type III
fit.type3<-aov(socialanxiety~group*sex,data = SA,
               contrasts=list(sex="contr.sum", group="contr.sum"))
fit.type3
summary(fit.type3)
#only GROUP is Significant

#another method to get type III result in R
fit.type3.2<-Anova(lm(socialanxiety~group*sex,
                     contrasts=list(sex="contr.sum", group="contr.sum"),data = SA),type="III")
fit.type3.2
summary(fit.type3.2)

```

### Additional Exercises

1. Answer: c

Justification: All the p-values in the scenario table, 0.0060, 0.0000 and 0.0100, are highly significant indicating that both factors A and B have the main effects on the built model. Also, the significance of the interaction term means that factors A and B have interacted with each other.

2. Answer: c

Justification: For the main effect terms, only factor A's p-value is significant, which equal to 0.0023, so that the plot is expected to appear changing with the value of A changes while not for B. Also, the interaction term is not significant as well indicating that there is no interaction effect between two factors. Consequently, the two factors do not cross with each other.

3. Answer: d

Justification: All the p-values in the scenario table, 0.2356, 0.3568 and 0.8542, are highly nonsignificant indicating that neither factors A nor B does not have the main effects on the built model. Also, the un-significance of the interaction term means that factors A and B do not interact with each other. The plot might be performed as a horizontal line without fluctuating as the values change.

4. Answer: c

Justification: Only the interaction term shown the nonsignificant value, 0.8532, while both of two main effect terms are highly significant, p-values equal to 0.0236 and 0.0268 respectively, which means that there is no interaction exist in the built model. By considering both factors that will impact the plot with the value changes at the same time, the image where two lines are nearly parallel is possible.

5. Answer: e

Justification: The p-values in the scenario table, 0.0060, 0.0000 and 0.0110, are highly significant indicating that both factors A and B have the main effects and there is a significant interaction effect at 10% significance level. At the 1% significance level, the interaction is not significant, but both A and B are.

6. Answer:

Justification: For the two-way ANOVA model, the degree of freedom of the total treatments, followed the fundamental identity as well, which equals to the degree of freedom of total number of observations minus the degree of freedom of errors, while is divided into degrees

of freedom of factor A, B and the interaction term this time. Detailedly, the degree of freedom of factor is themselves level minus one, or for interaction term should the product of two main factors' df-value. So that, notationally,

$$\begin{aligned}
 \text{Level of factor B} &= \text{degree of freedom(B)} + 1 \\
 &= [df(\text{total treatments}) - df(A) - df(AB)] + 1 \\
 &= \{[df(\text{total observations}) - df(\text{errors})] - df(A) - df(AB)\} + 1 \\
 &= [(a-1)(b-1) - (a-1) - (a-1)(b-1)] + 1 \\
 &= [5-2-2] + 1 = 2
 \end{aligned}$$

For two-way ANOVA, the total sum of square consist of treatment sum of square and error sum of square. So that, we could get,

$$SSTR = SSTO - SSE = 1642.0 - 62.0 = 1580.0$$

Meanwhile, the treatment sum of square is the sum of each treatment in the built model. Thus, notationally,

$$SSB = SSTR - SSA - SSAB = 1580.0 - 1544.0 - 24.0 = 12$$

Lastly, the mean squared associated with factor B is the quotient of sum squared of factor B divided by its degree of freedom, indicating

$$MSEB = \frac{SSB}{df(B)} = \frac{12}{1} = 12$$

7. Answer: a

Justification:

Explained in 6;

$$\begin{aligned}
 \text{Level of factor B} &= \text{degree of freedom(B)} + 1 \\
 &= [df(\text{total treatments}) - df(A) - df(AB)] + 1 \\
 &= \{[df(\text{total observations}) - df(\text{errors})] - df(A) - df(AB)\} + 1 \\
 &= [(a-1)(b-1) - (a-1) - (a-1)(b-1)] + 1 \\
 &= [5-2-2] + 1 = 2
 \end{aligned}$$

8. Answer: c

Justification:

Explained in 6;

$$MSEB = \frac{SSB}{df(B)} = \frac{12}{1} = 12$$

9. Answer: c

Justification:

$$F(A) = \text{Mean Square}(A) / \text{Mean Square Error}(A) = \frac{SSA/Df(A)}{MSE_{Error}} = \frac{772}{10.3}$$

$$F(B) = F(AB) = \frac{12}{10.3}$$

Thus, the first and the third p-value is for main effect of factor B and interaction effect of A\*B. They are both insignificant. But the second p-value for main effect of factor A is significant.

10. Answer : b

Justification: The p-values in the scenario table, 0.0017, 0.9220 and 0.9609, indicating that only factors A have the main effect at the 1% significance level. The main effect of B and interaction effect is not significant even at the 10% significance level.

11. Answer:

Justification:

This situation shows a typical case of how disordinal interaction seriously invalidates the interpretation of the main effects, where the differences between levels “intersect” depending how they are combined with levels from the other treatment . Under this context, it means the effects of treatment are positive for some levels (young women) and negative for other levels of the other treatment(old men). It can be illustrated by the line chart below - A1 stands for “young”, A2 for “old”, B1 for “women“ and B2 for “men”, where the lines cross to the degree that the relative ordering of the lines is reversed. In this tricky situation, the main effects are confounded by the disordinal interaction because when looking at the difference of means, either by age or gender, it will show nonsignificance. When disordinal interaction occurs, the tests for difference for the main effects are severely invalidated and should not be trusted. That is also the reason why if there is more than one treatment in the analysis, researchers should always examine the interaction term first to see if they are significant, and if so, do they allow for an interpretation of the main effects or not. Otherwise, it may turn out to be completely meaningless to look at the main effects separately.

