Homework 1

Xinyi Lin

9/22/2019

Part A

In order to use Hotelling T^2 statistic, three assumptions need to be meet.

1. Each population follows the multivariate normal distribution.
2. Samples are independent
3. Sample have equal variance-covariance matrices.

For 13 control and 20 obese patients with 0, 0.5, 1, 1.5, 2 and 3 hours time points, we can set up the model as following:

Let $y_i, i = 1, \ldots, 33$ be independent 6-values vectors. $y_i \sim N_6(\mu_j, \Sigma), j = 1,2$.

$$Y_{33\times6} = \begin{bmatrix} y_{1,1} \cdots y_{1,6} \\ \vdots \quad \ddots \quad \vdots \\ y_{33,1} \cdots y_{33,6} \end{bmatrix} = \begin{bmatrix} y'_1 \\ \vdots \\ y'_{33} \end{bmatrix}$$

$$X_{33\times2} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix}$$

$$B_{2\times6} = \begin{bmatrix} \mu_{11} \cdots \mu_{16} \\ \mu_{21} \cdots \mu_{26} \end{bmatrix}$$

$$E_{33\times6} = \begin{bmatrix} \epsilon_{1,1} \cdots \epsilon_{1,6} \\ \vdots \quad \ddots \quad \vdots \\ \epsilon_{33,1} \cdots \epsilon_{33,6} \end{bmatrix} = \begin{bmatrix} \epsilon'_1 \\ \vdots \\ \epsilon'_{33} \end{bmatrix}$$

$$Y_{33\times6} = X_{33\times2} B_{2\times6} + E_{33\times6}$$

Problem a

$H_0: ABC = D$

$A = [1 \ -1], C = I_6, D = 0_{1\times6}$

Problem b

$H_0: ABC = D$

$$A = [1\ -1], C_{6\times5} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ -1 & 1 & \dots & 0 \\ 0 & -1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -1 \end{bmatrix}, D_{1\times5} = 0_{1\times5}$$

Problem c

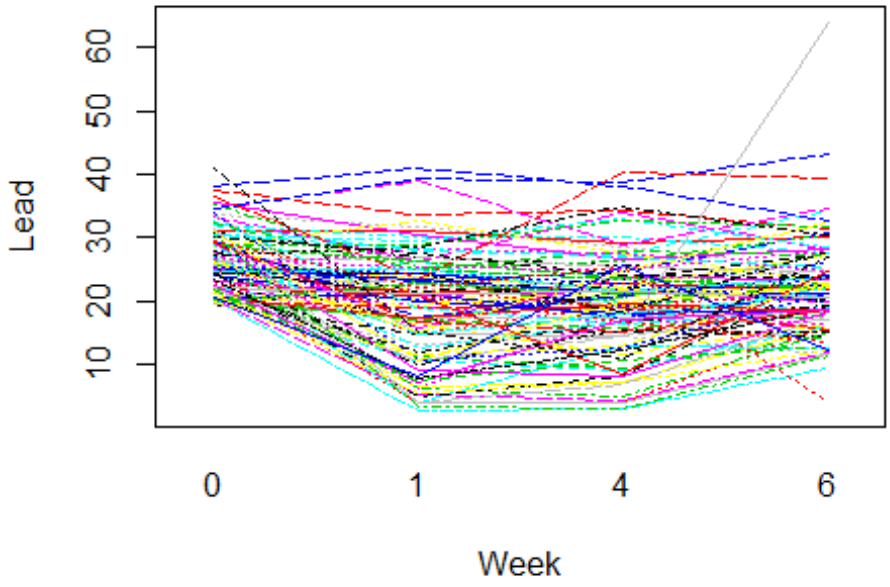
H_0: The differences in means at 2 and 3 hours after an oral glucose challenge are the same between the control and obese patients.

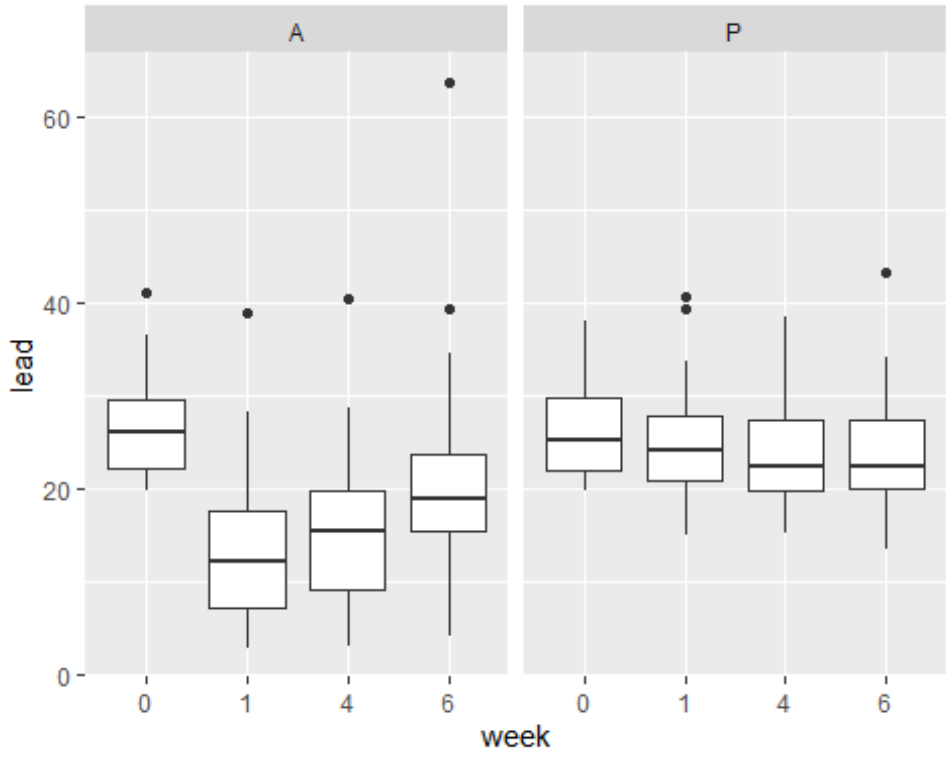
$$A = [1, -1], C_{6\times1} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ -1 \end{bmatrix}, D = 0$$

Part B

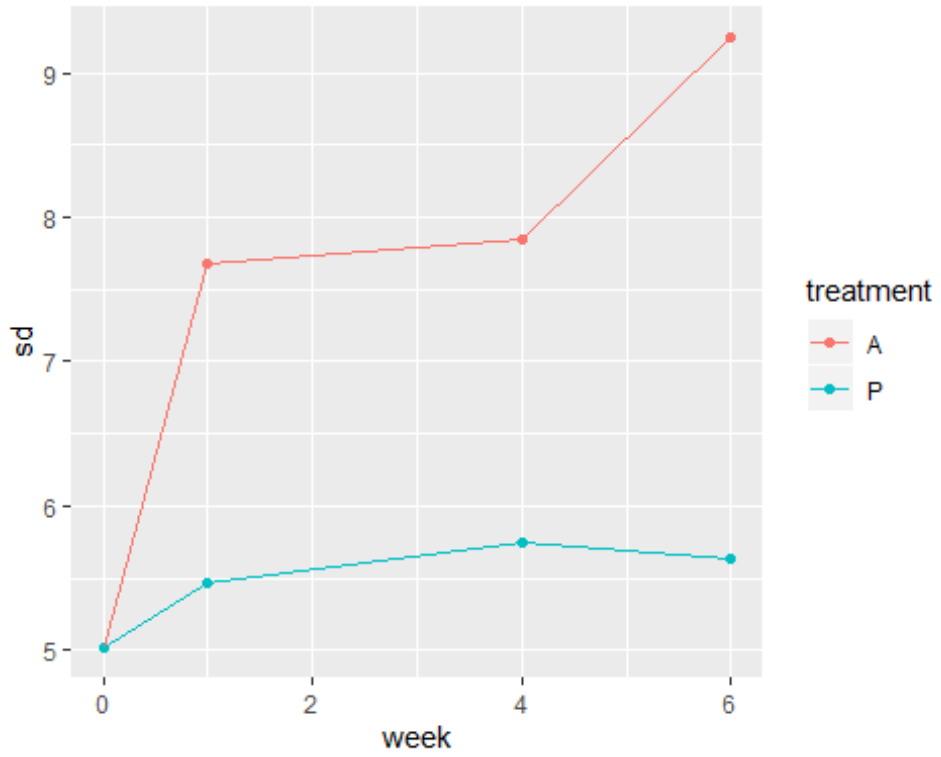
Question 1

The spaghetti plot of TLC data.

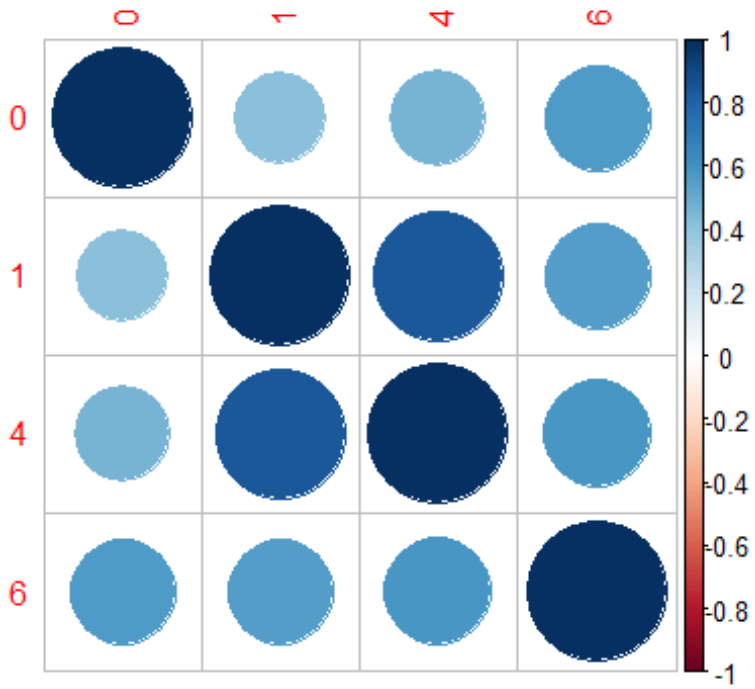




The boxplot of lead levels in different weeks among different groups are shown above. We can find that for the placebo group, lead levels are relatively stable while for control group, lead levels decrease first than increase.



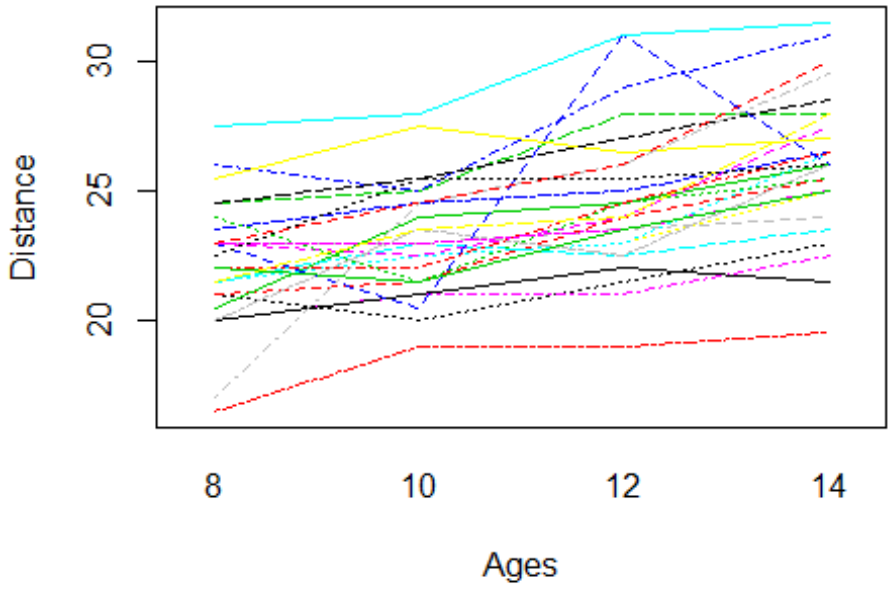
The sd of control group increase while the sd of placebo group is relatively stable.

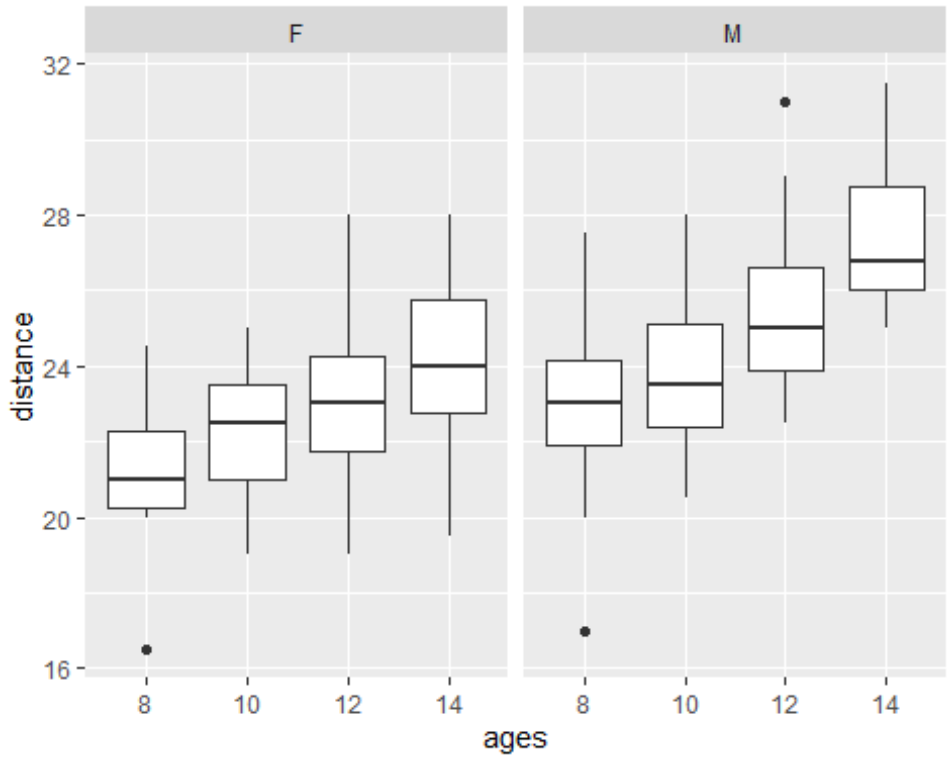


According to this data, we can find that correlations of lead levels between different week are all significant and correlation among lead levels in week 1 and week 4 is higher.

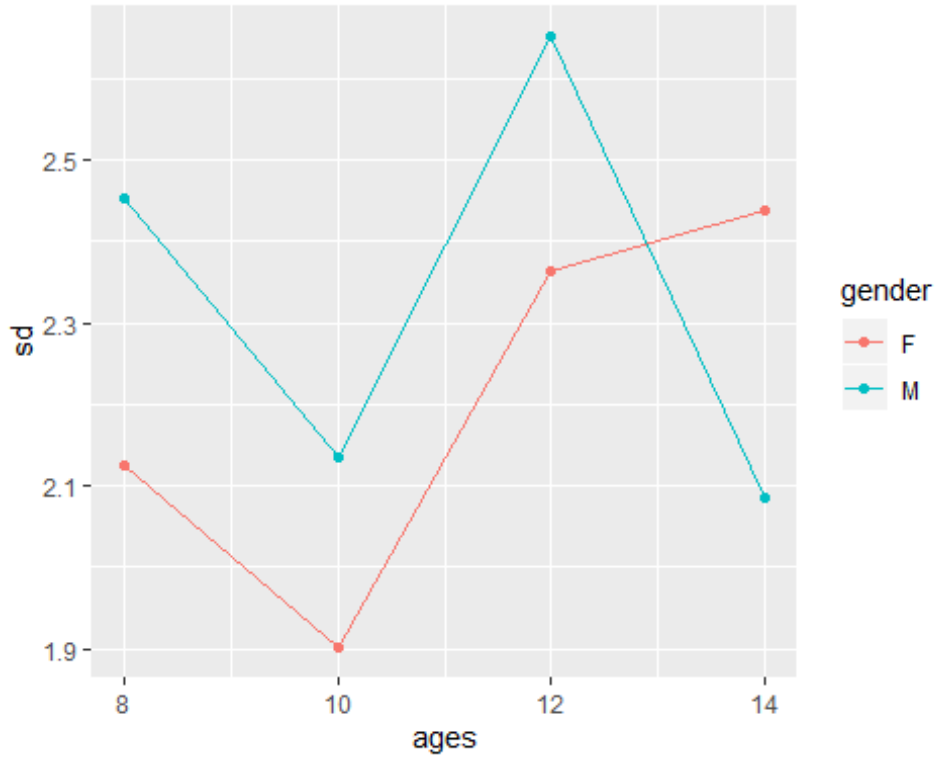
Question 2

The spaghetti plot of dental data.

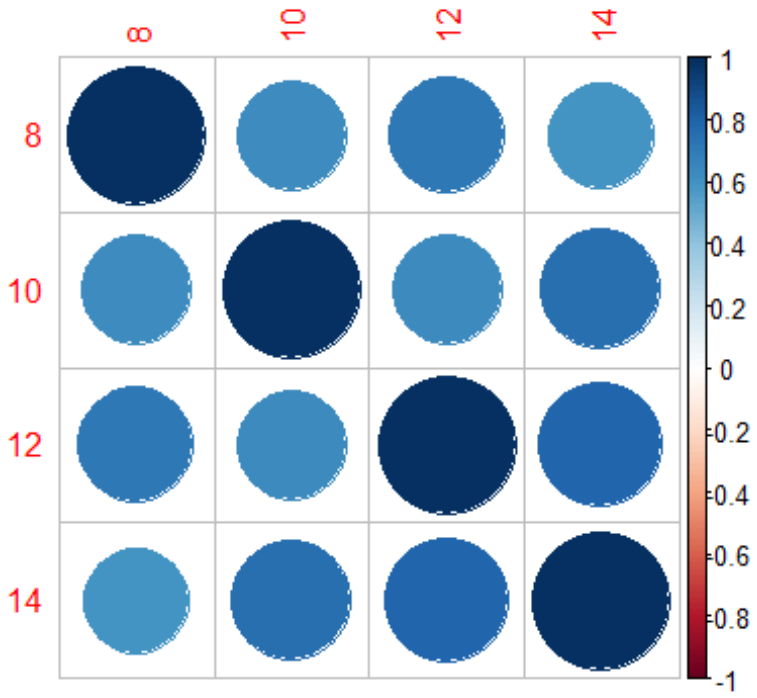




The boxplot of distance in different ages among different groups are shown above. We can find that with ages increase, distance tend to increase. Patterns among genders are similar.



The sd of female group decrease first then increase while the sd of male group is relatively stable.



According to this data, we can find that correlations of distance between different ages are all significant and correlation between age 10,14 and age 10,12 are higher.

Question 3

For a)

```
## Loading required package: mvtnorm

## Loading required package: ICS

##
##  Hotelling's two sample T2-test
##
## data:  x1 by y
## T.2 = 8.5531, df1 = 6, df2 = 26, p-value = 3.495e-05
## alternative hypothesis: true location difference is not equal to
c(0,0,0,0,0,0)
```

As p-value is smaller than 0.05, we can reject the null hypothesis and conclude that the means at all 6 time points between 2 groups are not all the same.

For b)

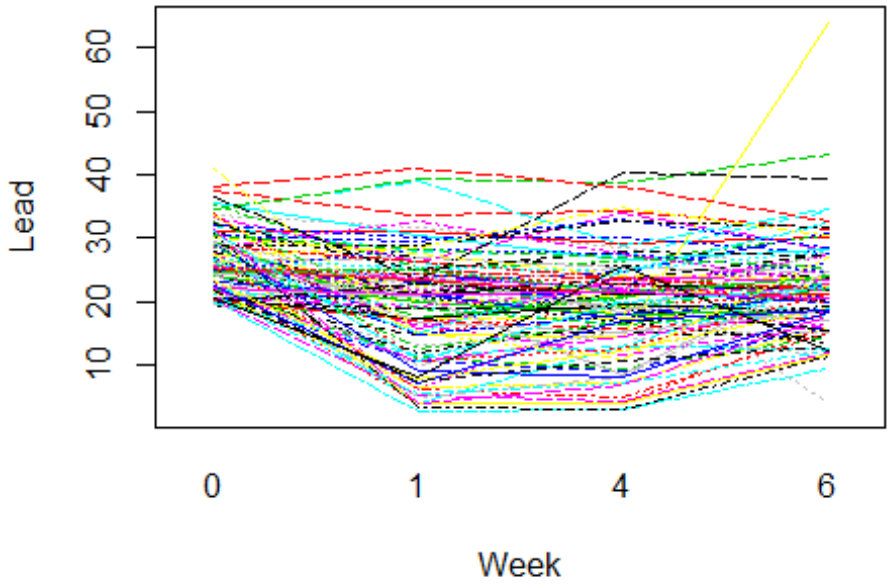
```
##
##  Hotelling's two sample T2-test
##
## data:  x2 by y
## T.2 = 8.1805, df1 = 5, df2 = 27, p-value = 8.344e-05
## alternative hypothesis: true location difference is not equal to
c(0,0,0,0,0)
```

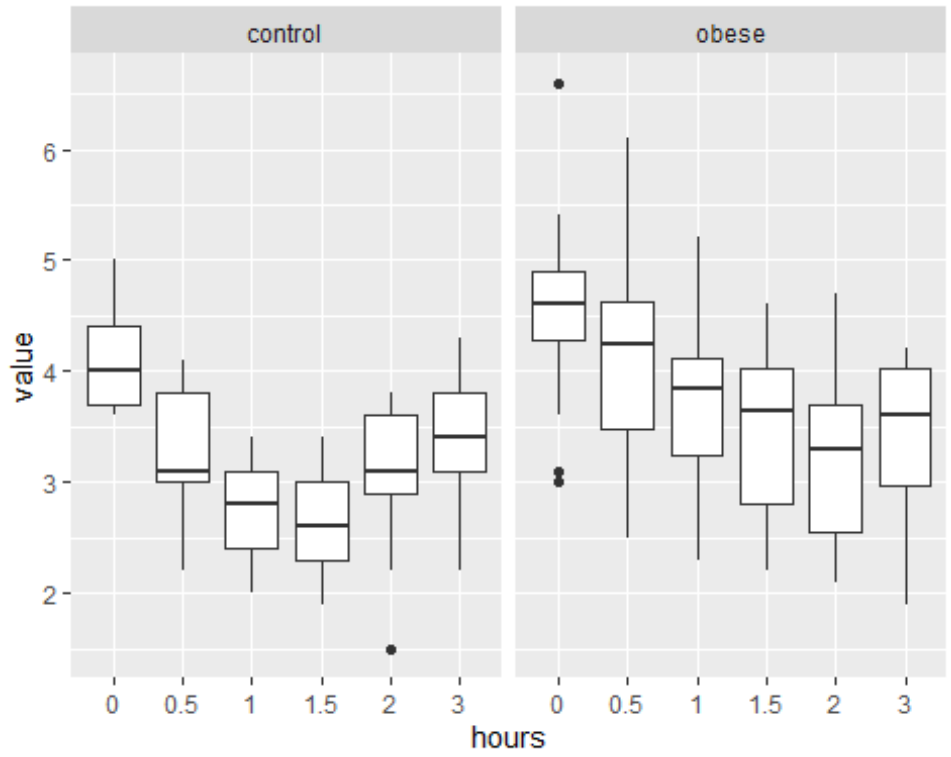
As p-value is smaller than 0.05, we can reject the null hypothesis and conclude that the profiles in the two groups are parallel.

For c)

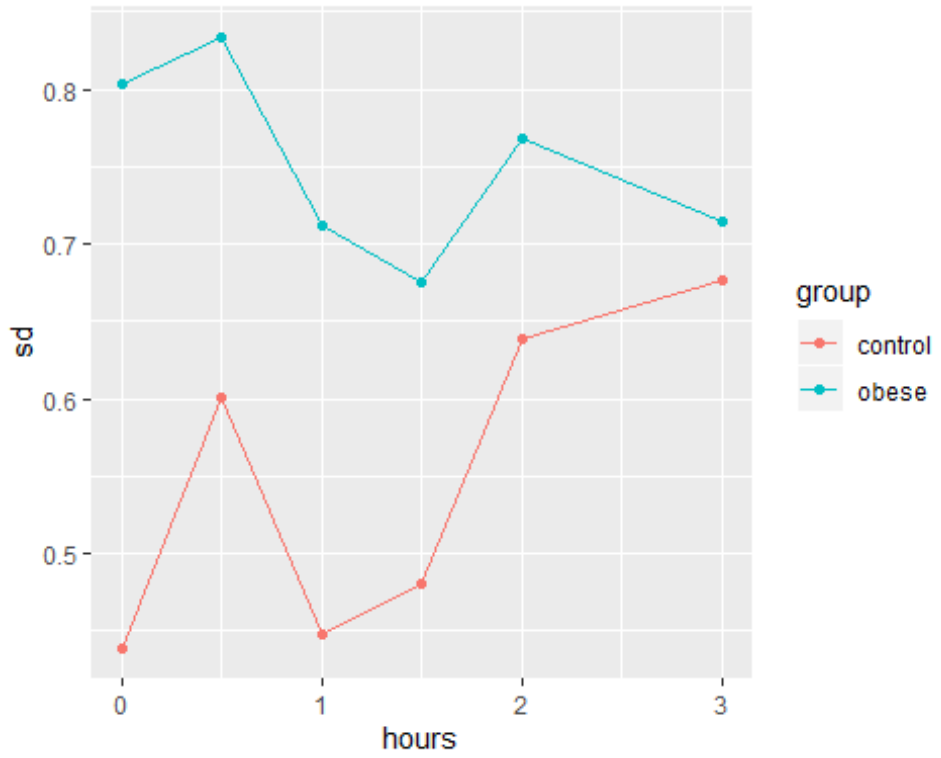
```
##
##  Hotelling's two sample T2-test
##
## data:  x3 by y
## T.2 = 0.41711, df1 = 1, df2 = 31, p-value = 0.5231
## alternative hypothesis: true location difference is not equal to c(0)
```

As p-value is bigger than 0.05, we can accept the null hypothesis and conclude that the differences in means at 2 and 3 hours after an oral glucose challenge are the same between the control and obese patients.

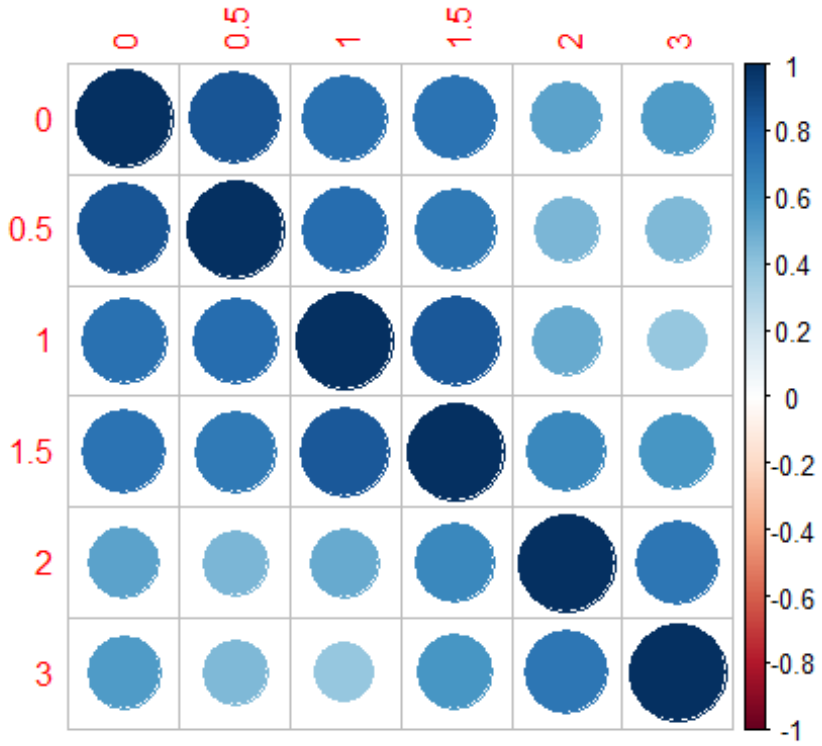




The boxplot of measurements in different hours among different groups are shown above. We can find that for the control group, measurements first decrease than increase, while for obese group, measurements decrease.



The sd of control group increase while the sd of obese group is relatively stable.



According to this data, we can find that correlations of measurements between different time points are all significant and correlation among 0.5 hour, 1 hour and 1.5 hour are higher.

Appendix

```
knitr::opts_chunk$set(echo = FALSE)
library(tidyverse)
library(ggplot2)
TLC = read.table("./TLC.dat")
colnames(TLC) = c("subject_id", "treatment","0", "1", "4", "6")
TLC_data = TLC %>%
  gather(key = "week", value = "lead", "0":"6")

interaction.plot(TLC_data$week, TLC_data$subject_id, TLC_data$lead,
xlab="Week", ylab="Lead", col=c(1:100), legend=F)
TLC_data %>%
  ggplot(aes(x=week, y=lead)) +
  geom_boxplot() +
  facet_grid(.~treatment)
TLC_data %>%
  group_by(week, treatment) %>%
  summarize(sd = sd(lead)) %>%
  ungroup() %>%
  mutate(week = as.numeric(week)) %>%
```

```r
  ggplot(aes(x=week, y=sd, color = treatment)) +
  geom_point() +
  geom_line()
x = cor(TLC[,3:6])
corrplot::corrplot(x)
dental = read.table("./dental.dat")
colnames(dental) = c("subject_id", "gender","8", "10", "12", "14")
dental_data = dental %>%
  gather(key = "ages", value = "distance", "8":"14") %>%
  mutate(ages = as.numeric(ages))

interaction.plot(dental_data$ages, dental_data$subject_id,
dental_data$distance, xlab="Ages", ylab="Distance", col=c(1:27), legend=F)
dental_data %>%
  mutate(ages = as.factor(ages)) %>%
  ggplot(aes(x=ages, y=distance)) +
  geom_boxplot() +
  facet_grid(.~gender)
dental_data %>%
  group_by(ages, gender) %>%
  summarize(sd = sd(distance)) %>%
  ungroup() %>%
  ggplot(aes(x=ages, y=sd, color = gender)) +
  geom_point() +
  geom_line()
x = cor(dental[,3:6])
corrplot::corrplot(x)
library(ICSNP)
zerbe2 = read.table("./ZERBE2.DAT")
colnames(zerbe2) = c("group", "subject_id", "hour0", "hour0.5", "hour1",
"hour1.5", "hour2", "hour3")
x1 = as.matrix(zerbe2[,3:8])
y = factor(rep(c(1,2),c(13,20)))
HotellingsT2(x1~y)
zerbe2 = zerbe2 %>%
  mutate(diff1 = hour0.5-hour0,
         diff2 = hour1-hour0.5,
         diff3 = hour1.5-hour1,
         diff4 = hour2-hour1.5,
         diff5 = hour3-hour2)
x2 = as.matrix(zerbe2[9:13])
HotellingsT2(x2~y)
x3 = as.matrix(zerbe2[13])
HotellingsT2(x3~y)
zerbe2 = read.table("./ZERBE2.DAT")
colnames(zerbe2) = c("group", "subject_id", "0", "0.5", "1", "1.5", "2", "3")
zerbe2_data = zerbe2 %>%
  gather(key = "hours", value = "value", "0":"3") %>%
  mutate(group = ifelse(group==1, "control", "obese"))
```

```r
interaction.plot(TLC_data$week, TLC_data$subject_id, TLC_data$lead,
xlab="Week", ylab="Lead", col=c(1:33), legend=F)
zerbe2_data %>%
  ggplot(aes(x=hours, y=value)) +
  geom_boxplot() +
  facet_grid(.~group)
zerbe2_data %>%
  group_by(hours, group) %>%
  summarize(sd = sd(value)) %>%
  ungroup() %>%
  mutate(hours = as.numeric(hours)) %>%
  ggplot(aes(x=hours, y=sd, color = group)) +
  geom_point() +
  geom_line()
x = cor(zerbe2[,3:8])
corrplot::corrplot(x)
```