

Longitudinal Data Analysis

Case study of Trenal.XLS using Linear Mixed Effect Model

Group2: Wanchang Zhang; Hugo Blain; Oscar Cabanelas

2023-03-21

Contents

0. Introduction	3
1. Task for week one	3
1.1 Import data	3
1.2 Table structure analysis and variable understanding	5
1.3 List of Hypotheses to be tested by the data	10
Correlation analysis for different HC levels along time	35
Third model Random Slope	46
Only fixed effects	48
Take the intercept as random effects	49
Take	50

List of Figures

List of Tables

0. Introduction

The dataset contains information on patients who received renal graft(kidney transplant) The patients have been followed for at most 10 years.

Background: People with end-stage kidney disease who receive a kidney transplant generally live longer than people with ESRD who are on dialysis.

However, kidney transplant recipients must remain on immunosuppressants (medications to suppress the immune system) for the rest of their life to prevent their body from rejecting the new kidney. The long-term immunosuppression puts them at risk for infections and cancer.

The Haematocrit level (HC level) usually differs with gender, Also the health condition of a person.

1. Task for week one

1.1 Import data

```
#install.packages("readxl")
library(readxl)
trenal <- read_excel("Trenal.XLS")
summary(trenal)

##      HC0          HC06         HC1          HC2          HC3
##  Min.   :14.00   Min.   :22.00   Min.   :20.00   Min.   :17.0   Min.   :20.00
##  1st Qu.:28.00  1st Qu.:35.00  1st Qu.:36.00  1st Qu.:36.0  1st Qu.:36.00
##  Median :32.00  Median :38.55  Median :39.00  Median :40.0  Median :39.00
##  Mean   :31.86  Mean   :38.83  Mean   :39.71  Mean   :39.7  Mean   :39.17
##  3rd Qu.:36.00 3rd Qu.:42.00 3rd Qu.:43.00 3rd Qu.:43.0 3rd Qu.:43.00
##  Max.   :60.00  Max.   :61.70  Max.   :63.00  Max.   :65.0  Max.   :60.00
##  NA's    :12      NA's    :12      NA's    :12      NA's   :1044  NA's   :2460
##      HC4          HC5          HC6          HC7
##  Min.   :23.00   Min.   :17.00   Min.   :20.00   Min.   :17.00
##  1st Qu.:35.00  1st Qu.:35.00  1st Qu.:36.00  1st Qu.:35.00
##  Median :39.00  Median :39.00  Median :39.00  Median :39.00
##  Mean   :39.16  Mean   :39.02  Mean   :39.11  Mean   :38.85
##  3rd Qu.:43.00 3rd Qu.:43.00 3rd Qu.:43.00 3rd Qu.:42.00
##  Max.   :55.00  Max.   :56.00  Max.   :55.00  Max.   :60.00
##  NA's    :3768  NA's    :5016  NA's    :6096  NA's   :7140
##      HC8          HC9          HC10         id
##  Min.   :23.00   Min.   :17.00   Min.   :24.10   Min.   :  1.0
##  1st Qu.:35.00  1st Qu.:35.00  1st Qu.:35.00  1st Qu.:290.8
##  Median :38.05  Median :38.50  Median :38.00  Median : 580.5
##  Mean   :38.35  Mean   :38.57  Mean   :38.49  Mean   : 580.5
##  3rd Qu.:42.00 3rd Qu.:42.00 3rd Qu.:42.00 3rd Qu.: 870.2
##  Max.   :55.00  Max.   :55.00  Max.   :54.00  Max.   :1160.0
##  NA's    :8064  NA's    :8988  NA's    :9744
##      age          male         cardio        reject       const
##  Min.   :15.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :1
##  1st Qu.:36.00  1st Qu.:0.0000  1st Qu.:0.0000  1st Qu.:0.0000  1st Qu.:1
##  Median :48.00  Median :1.0000  Median :0.0000  Median :0.0000  Median :1
##  Mean   :46.43  Mean   :0.5741  Mean   :0.1784  Mean   :0.3164  Mean   :1
##  3rd Qu.:57.00 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:1
##  Max.   :76.00  Max.   :1.0000  Max.   :1.0000  Max.   :1.0000  Max.   :1
##  NA's    :12
```

```

##      j      respons      time
##  Min.   : 1.00   Min.   :14.00   Min.   : 0.000
##  1st Qu.: 3.75   1st Qu.:34.00   1st Qu.: 1.750
##  Median : 6.50   Median :38.00   Median : 4.500
##  Mean   : 6.50   Mean   :38.24   Mean   : 4.625
##  3rd Qu.: 9.25   3rd Qu.:42.00   3rd Qu.: 7.250
##  Max.   :12.00   Max.   :65.00   Max.   :10.000
##           NA's   :4362

remove a noninformative column const
trenal=trenal[,-18]
summary(trenal)

##      HC0      HC06      HC1      HC2      HC3
##  Min.   :14.00   Min.   :22.00   Min.   :20.00   Min.   :17.0   Min.   :20.00
##  1st Qu.:28.00   1st Qu.:35.00   1st Qu.:36.00   1st Qu.:36.0   1st Qu.:36.00
##  Median :32.00   Median :38.55   Median :39.00   Median :40.0   Median :39.00
##  Mean   :31.86   Mean   :38.83   Mean   :39.71   Mean   :39.7   Mean   :39.17
##  3rd Qu.:36.00   3rd Qu.:42.00   3rd Qu.:43.00   3rd Qu.:43.0   3rd Qu.:43.00
##  Max.   :60.00   Max.   :61.70   Max.   :63.00   Max.   :65.0   Max.   :60.00
##  NA's   :12          NA's   :12          NA's   :12        NA's   :1044  NA's   :2460
##      HC4      HC5      HC6      HC7
##  Min.   :23.00   Min.   :17.00   Min.   :20.00   Min.   :17.00
##  1st Qu.:35.00   1st Qu.:35.00   1st Qu.:36.00   1st Qu.:35.00
##  Median :39.00   Median :39.00   Median :39.00   Median :39.00
##  Mean   :39.16   Mean   :39.02   Mean   :39.11   Mean   :38.85
##  3rd Qu.:43.00   3rd Qu.:43.00   3rd Qu.:43.00   3rd Qu.:42.00
##  Max.   :55.00   Max.   :56.00   Max.   :55.00   Max.   :60.00
##  NA's   :3768   NA's   :5016   NA's   :6096   NA's   :7140
##      HC8      HC9      HC10      id
##  Min.   :23.00   Min.   :17.00   Min.   :24.10   Min.   :  1.0
##  1st Qu.:35.00   1st Qu.:35.00   1st Qu.:35.00   1st Qu.: 290.8
##  Median :38.05   Median :38.50   Median :38.00   Median : 580.5
##  Mean   :38.35   Mean   :38.57   Mean   :38.49   Mean   : 580.5
##  3rd Qu.:42.00   3rd Qu.:42.00   3rd Qu.:42.00   3rd Qu.: 870.2
##  Max.   :55.00   Max.   :55.00   Max.   :54.00   Max.   :1160.0
##  NA's   :8064   NA's   :8988   NA's   :9744
##      age      male      cardio      reject
##  Min.   :15.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:36.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :48.00   Median :1.0000   Median :0.0000   Median :0.0000
##  Mean   :46.43   Mean   :0.5741   Mean   :0.1784   Mean   :0.3164
##  3rd Qu.:57.00   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :76.00   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  NA's   :12

##      j      respons      time
##  Min.   : 1.00   Min.   :14.00   Min.   : 0.000
##  1st Qu.: 3.75   1st Qu.:34.00   1st Qu.: 1.750
##  Median : 6.50   Median :38.00   Median : 4.500
##  Mean   : 6.50   Mean   :38.24   Mean   : 4.625
##  3rd Qu.: 9.25   3rd Qu.:42.00   3rd Qu.: 7.250
##  Max.   :12.00   Max.   :65.00   Max.   :10.000
##           NA's   :4362

```

```
dim(trenal)
```

1.2 Table structure analysis and variable understanding

The table contains observation of HC level on 1160 patients who have gone through kidney transplant. Each patient will have maximum 12 measurements in the 12 time point (0, 0.5, 1, 2, ..., 10) years.

if we just look at the first 12 columns, they are all Haematocrit level at the corresponding time. Thus our response variable is Haematocrit level. If we just look at first 17 columns from HC0 to reject, then the subtable looks like a wide table; If we start from column id to column time, the part of table is a long table. From now on we focus on the long table:

```
trenal.long = trenal[,13:20]  
summary(trenal.long)
```

```

##          id           age         male       cardio
##  Min.   : 1.0   Min.   :15.00   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:290.8 1st Qu.:36.00   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :580.5  Median :48.00   Median :1.0000   Median :0.0000
##  Mean    :580.5  Mean    :46.43   Mean    :0.5741   Mean    :0.1784
##  3rd Qu.:870.2  3rd Qu.:57.00   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1160.0  Max.   :76.00   Max.   :1.0000   Max.   :1.0000
##               NA's   :12
##      reject            j      respons        time
##  Min.   :0.0000   Min.   : 1.00   Min.   :14.00   Min.   : 0.000
##  1st Qu.:0.0000   1st Qu.: 3.75   1st Qu.:34.00   1st Qu.: 1.750
##  Median :0.0000   Median : 6.50   Median :38.00   Median : 4.500
##  Mean    :0.3164   Mean    : 6.50   Mean    :38.24   Mean    : 4.625
##  3rd Qu.:1.0000   3rd Qu.: 9.25   3rd Qu.:42.00   3rd Qu.: 7.250
##  Max.   :1.0000   Max.   :12.00   Max.   :65.00   Max.   :10.000
##               NA's   :4362

```

```
dim(trenal.long)
```

```
## [1] 13920 8
```

Besides the time $0, 0.5, 1, 2, 3, 4, 5, \dots, 10$ is one-to-one correspondent to $j = 1, 2, 3, \dots, 12$. But we can still leave it in the data frame. Our response variable is the HC level (The percentage of red cells in the blood, normal levels of hematocrit for men range from 41% to 50%, normal level for women is 36% to 48%) the explanatory variables are age, we can change the structure of the table as we are used to: Identity, time, response, explanatory variables (time dependent), explanatory variables (time independent). The response variables are some continuous integer values? The explanatory variables have binary type: male, cardio, reject, and integer type: age

```
#install.packages("magrittr") # package installations are only needed the first time you use it
#install.packages("dplyr")    # alternative installation of the %>%
library(magrittr) # needs to be run every time you start R and want to use %>%
library(dplyr)

##  
## Attaching package: 'dplyr'  
## The following objects are masked from 'package:stats':  
##  
##     filter, lag
```

```

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

trenal.long$id = as.factor(trenal.long$id)
trenal.long$j = as.factor(trenal.long$j)
trenal.long$male = as.factor(trenal.long$male)
trenal.long$cardio = as.factor(trenal.long$cardio)
trenal.long$reject = as.factor(trenal.long$reject)

data <- trenal.long %>%
  relocate(id) %>%
  relocate(j, .after=id)%>%
  relocate(time, .after = j)%>%
  relocate(respons, .after=time)

summary(data)

##          id           j        time       respons
## 1      : 12    1 :1160   Min.   : 0.000   Min.   :14.00
## 2      : 12    2 :1160   1st Qu.: 1.750   1st Qu.:34.00
## 3      : 12    3 :1160   Median : 4.500   Median :38.00
## 4      : 12    4 :1160   Mean   : 4.625   Mean   :38.24
## 5      : 12    5 :1160   3rd Qu.: 7.250   3rd Qu.:42.00
## 6      : 12    6 :1160   Max.   :10.000   Max.   :65.00
## (Other):13848 (Other):6960             NA's   :4362
##      age       male     cardio     reject
##  Min.   :15.00  0:5928  0:11436  0:9516
##  1st Qu.:36.00  1:7992  1: 2484  1:4404
##  Median :48.00
##  Mean   :46.43
##  3rd Qu.:57.00
##  Max.   :76.00
##  NA's   :12

sum(!is.na(data$respons))

## [1] 9558
length(unique(data$id))

## [1] 1160
# Plot the raw data
#install.packages("tigerstats")
require(tigerstats)

## Loading required package: tigerstats
## Loading required package: abd
## Loading required package: nlme
##
## Attaching package: 'nlme'
## The following object is masked from 'package:dplyr':
## 
##     collapse

```

```

## Loading required package: lattice
## Loading required package: grid
## Loading required package: mosaic

## Registered S3 method overwritten by 'mosaic':
##   method           from
##   fortify.SpatialPolygonsDataFrame ggplot2

##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features. The original behavior of these functions should not be affected by this.

##
## Attaching package: 'mosaic'

## The following object is masked from 'package:Matrix':
##
##   mean

## The following object is masked from 'package:ggplot2':
##
##   stat

## The following objects are masked from 'package:dplyr':
##
##   count, do, tally

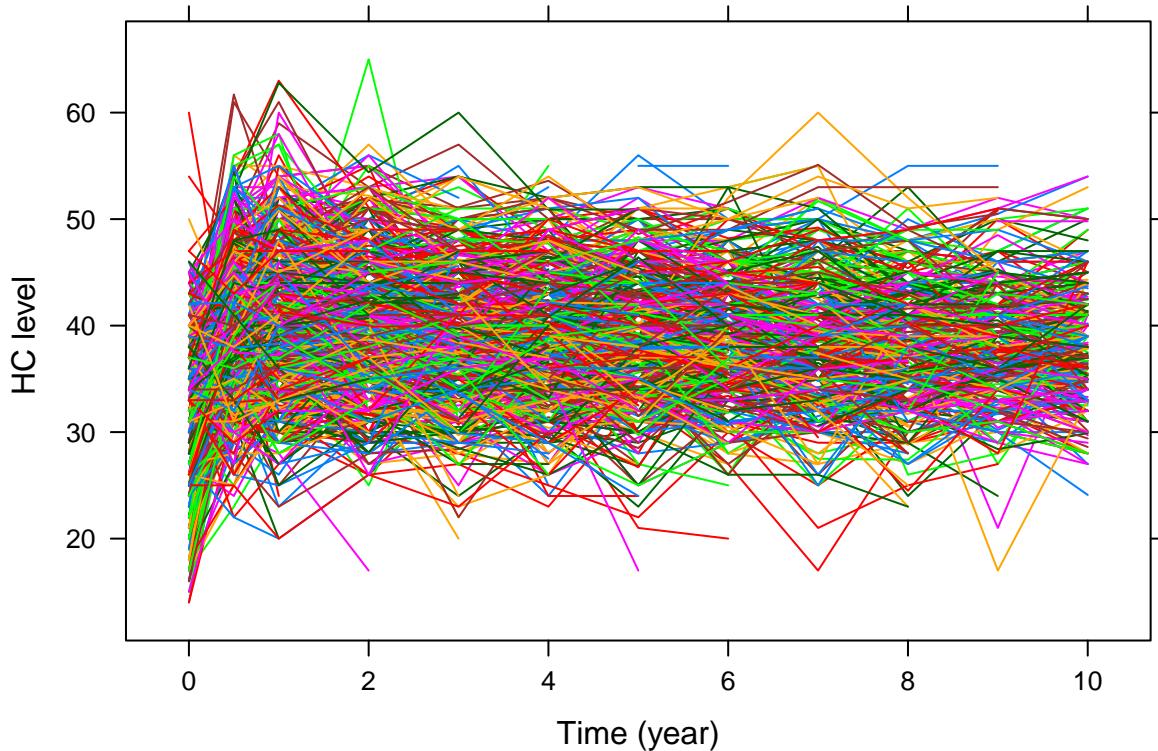
## The following objects are masked from 'package:stats':
##
##   binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##   quantile, sd, t.test, var

## The following objects are masked from 'package:base':
##
##   max, mean, min, prod, range, sample, sum

## Welcome to tigerstats!
## To learn more about this package, consult its website:
##   http://homerhanumat.github.io/tigerstats

xyplot(respons ~ time, groups = id, data=data, type="l",xlab="Time (year)",ylab="HC level " )

```



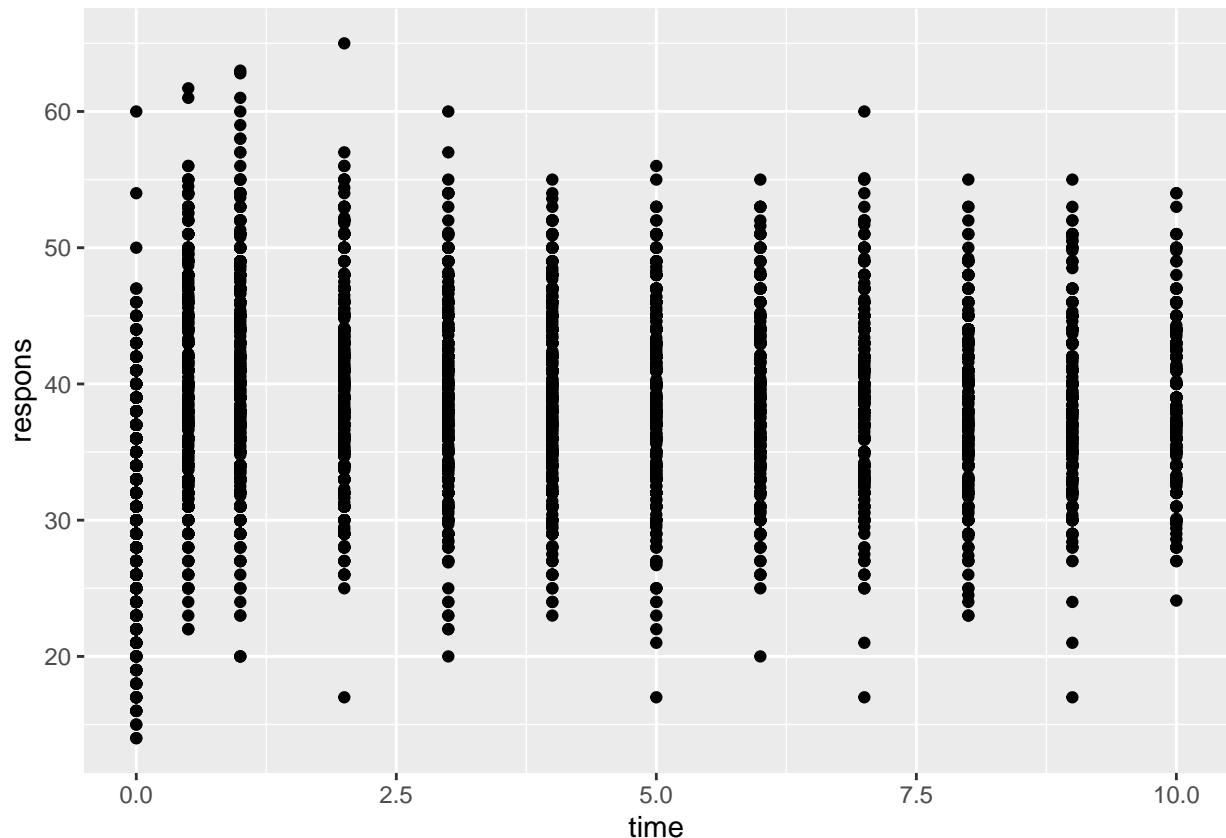
```

library(ggplot2)
library(nlme)
library(lme4)

##
## Attaching package: 'lme4'
## The following object is masked from 'package:mosaic':
##     factorize
## The following object is masked from 'package:nlme':
##     lmList
#Plot data
ggplot(data, aes(x=time, y=respons)) + geom_point()

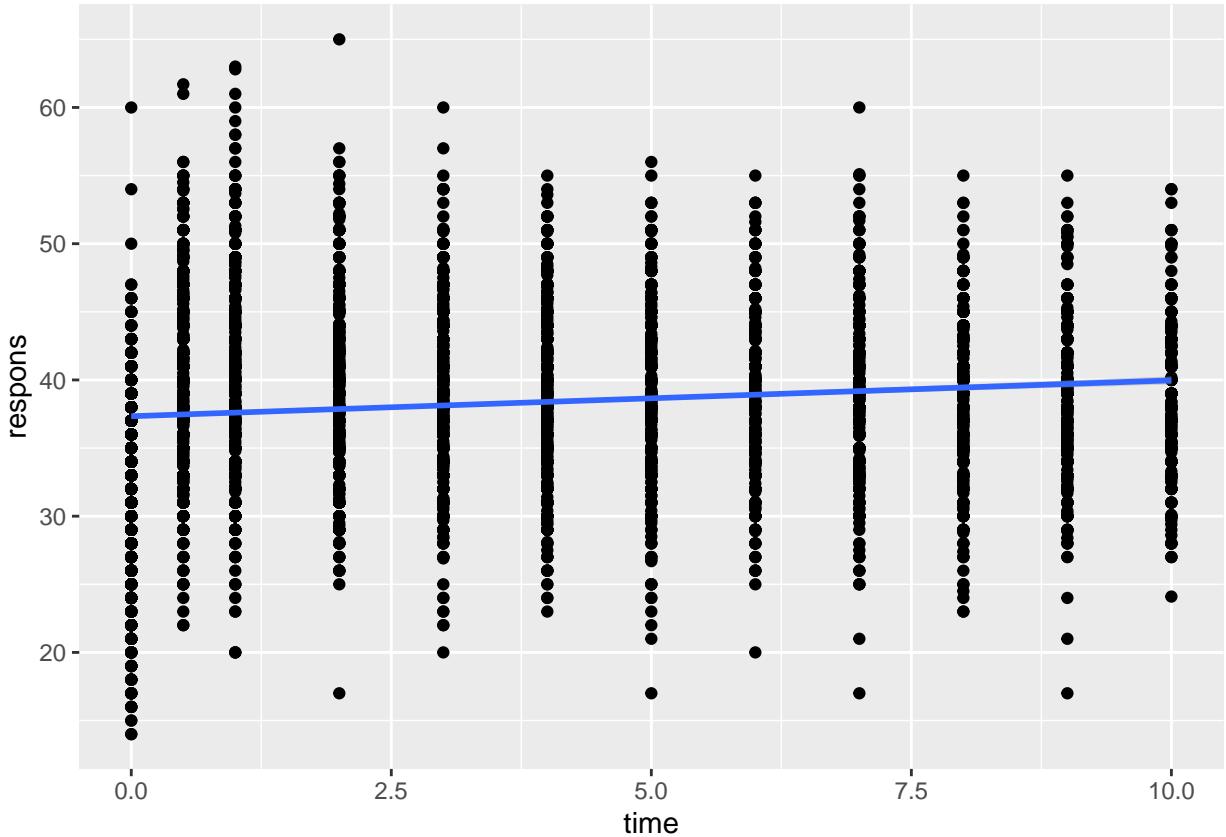
## Warning: Removed 4362 rows containing missing values (`geom_point()`).

```



```
#Plot data with lm line
ggplot(data, aes(x=time, y=responses)) + geom_point() + geom_smooth(method="lm")

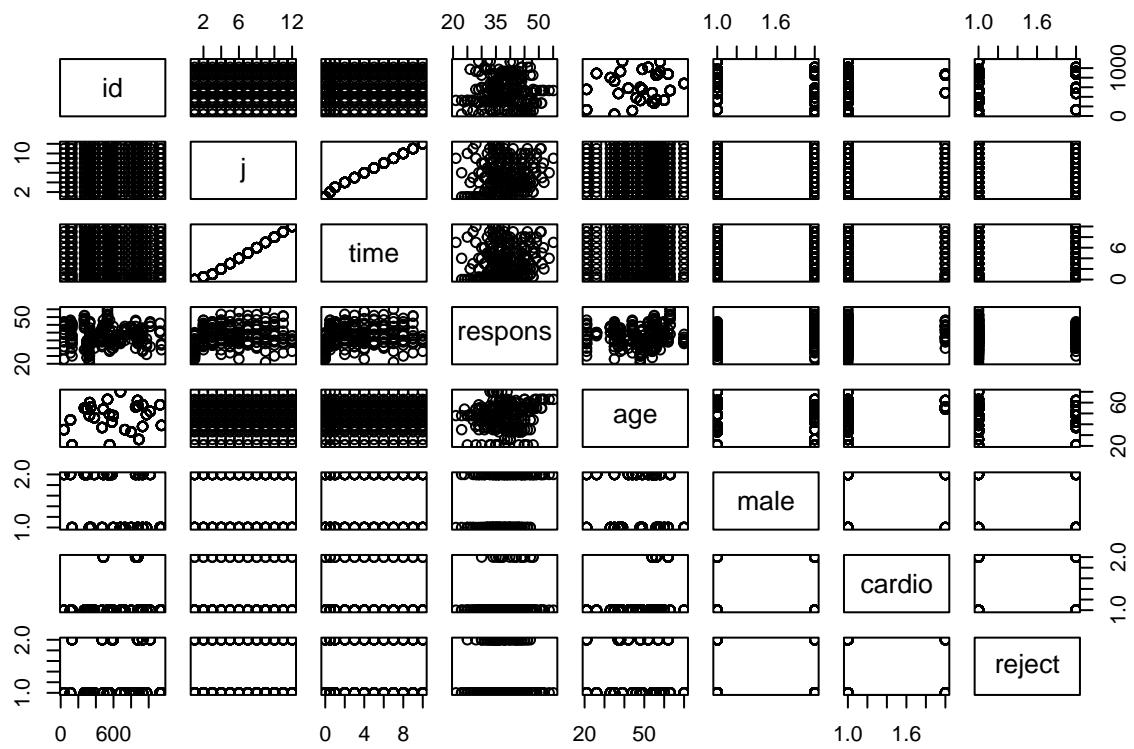
## `geom_smooth()` using formula = 'y ~ x'
## Warning: Removed 4362 rows containing non-finite values (`stat_smooth()`).
## Removed 4362 rows containing missing values (`geom_point()`).
```



1.3 List of Hypotheses to be tested by the data

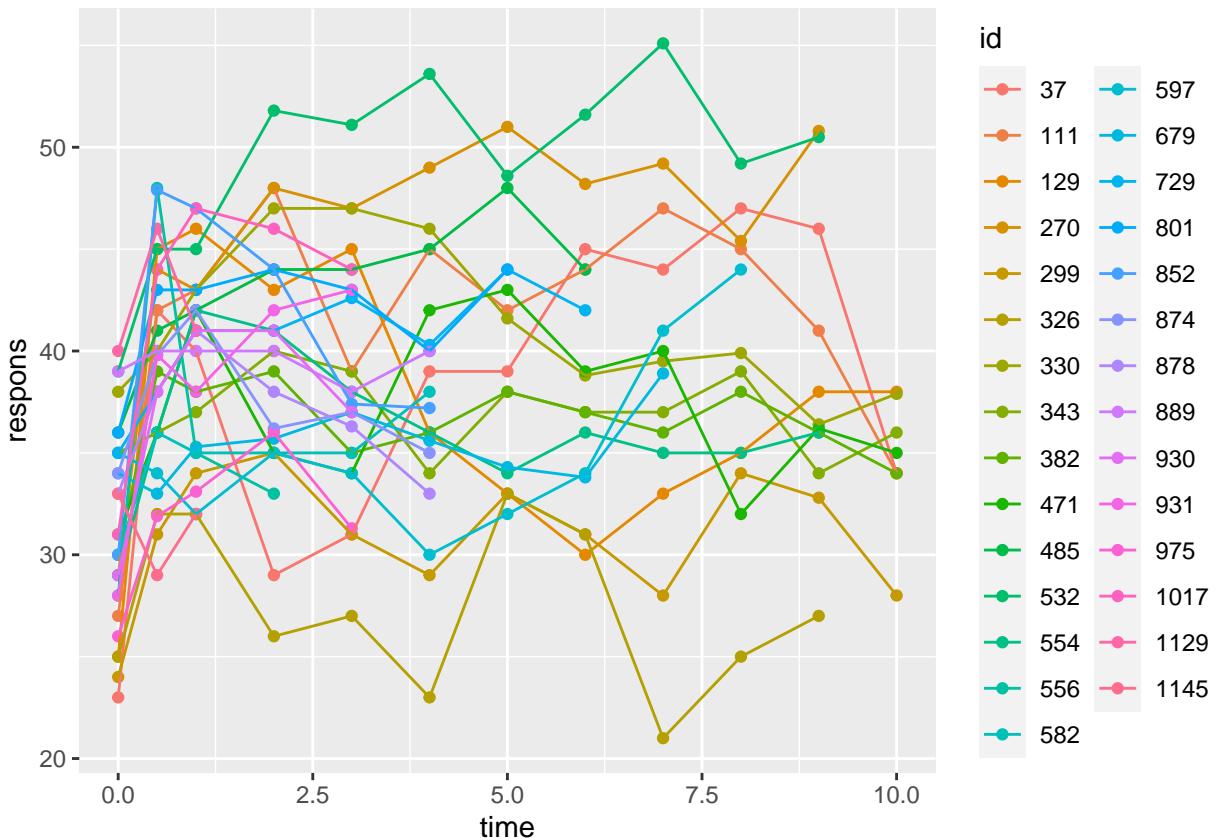
```
#Select a sample of data to plot
set.seed(1)
selected <- sample(1:length(unique(trenal.long$id)), 30, replace=T) # random samples and permutations
#selected.vector = as.vector(selected)
data.selected = data[(data$id %in% c(selected)), ]

# Individual plots
plot(data.selected) # WHAT I WILL GET FROM THE PLOT(DATA), HOW to plot a scatter plot of HC level chang
```



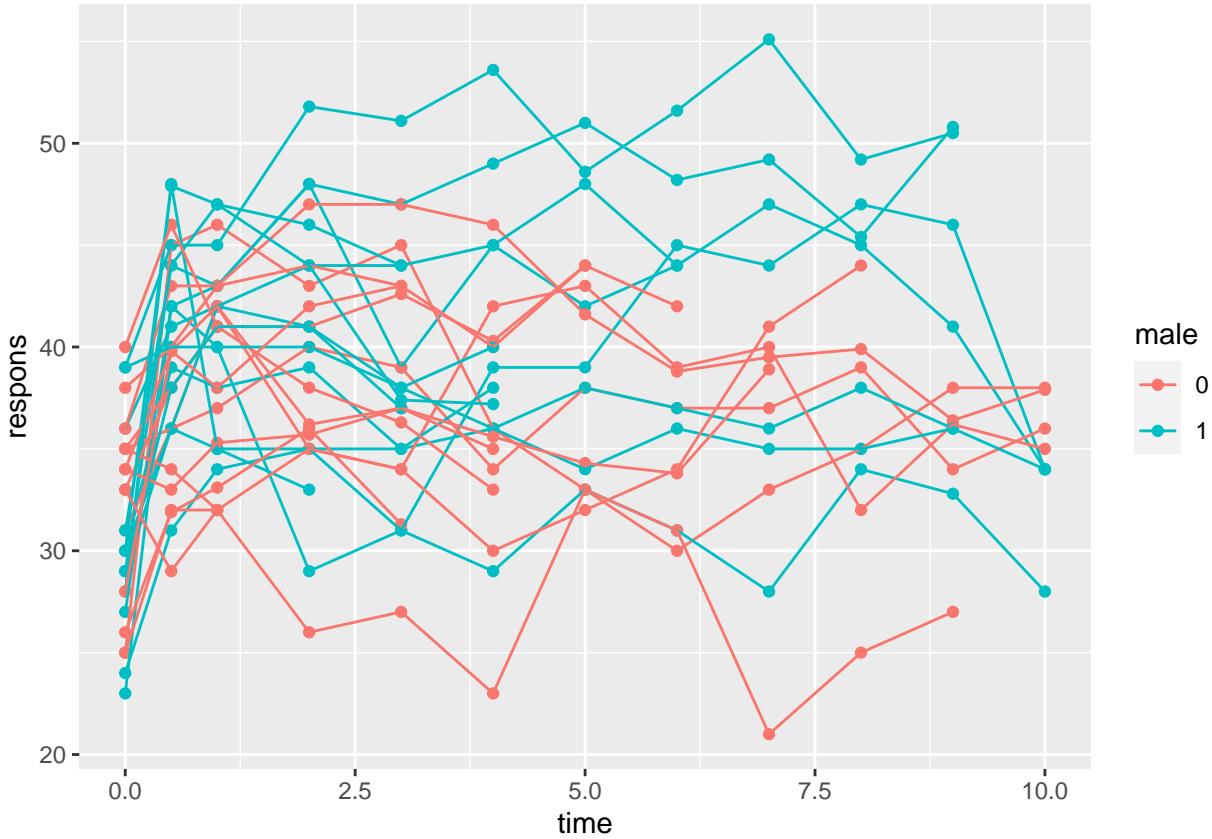
```
# spaghetti plot
ggplot(data.selected,aes(x=time,y=respons,group=id,color=id))+geom_point()+ geom_line()

## Warning: Removed 106 rows containing missing values (`geom_point()`).
## Warning: Removed 106 rows containing missing values (`geom_line()`).
```



```
# Plot individual data by sex
ggplot(data.selected,aes(x=time,y= respons,group=id, color= male))+geom_point() + geom_line()

## Warning: Removed 106 rows containing missing values (`geom_point()`).
## Warning: Removed 106 rows containing missing values (`geom_line()`).
```



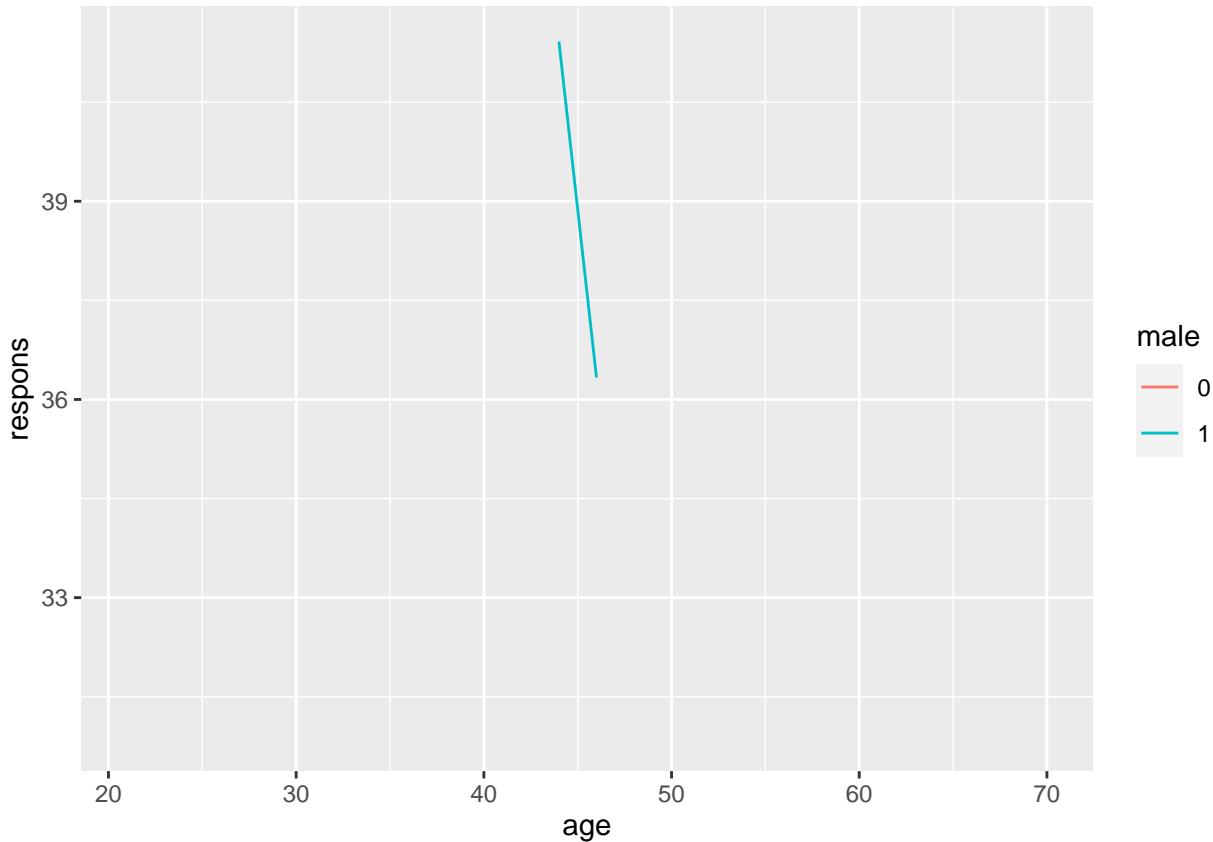
```
# Plot mean of male and mean of female
library(dplyr)
MEAN <- data.selected %>%
  group_by(male, age, cardio, reject) %>%
  summarise(respons = mean(respons))

## `summarise()` has grouped output by 'male', 'age', 'cardio'. You can override
## using the `.groups` argument.

MEAN

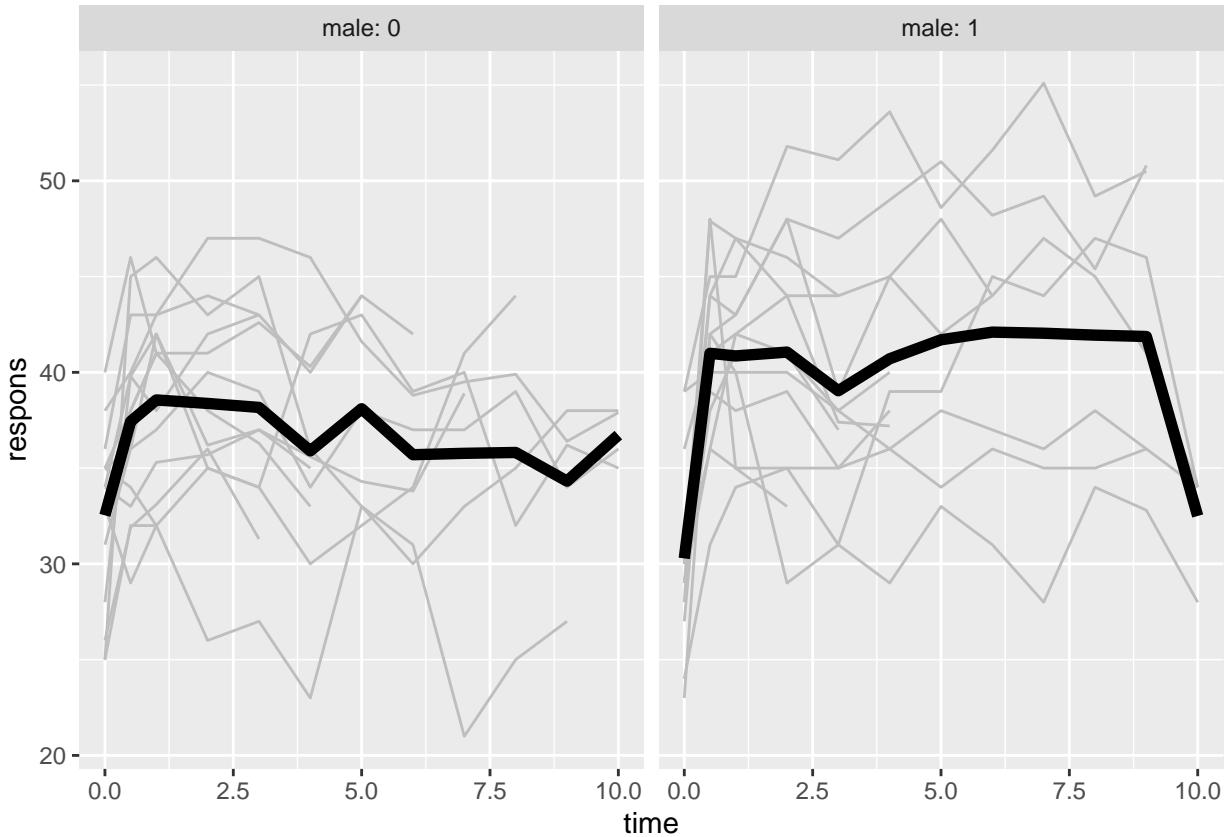
## # A tibble: 29 x 5
## # Groups:   male, age, cardio [28]
##   male    age  cardio reject respons
##   <fct> <dbl> <fct>   <fct>    <dbl>
## 1 0        21  0       1      37.2
## 2 0        33  0       0      NA
## 3 0        35  0       0      NA
## 4 0        37  0       1      37.1
## 5 0        38  0       1      NA
## 6 0        39  0       0      NA
## 7 0        48  0       0      NA
## 8 0        48  0       1      NA
## 9 0        49  0       0      NA
## 10 0       56  0       0      36.8
## # ... with 19 more rows
```

```
ggplot(data.selected,aes(x=age,y=respons,color=maile)) + geom_line(data=MEAN)
## Warning: Removed 9 rows containing missing values (`geom_line()`).
```



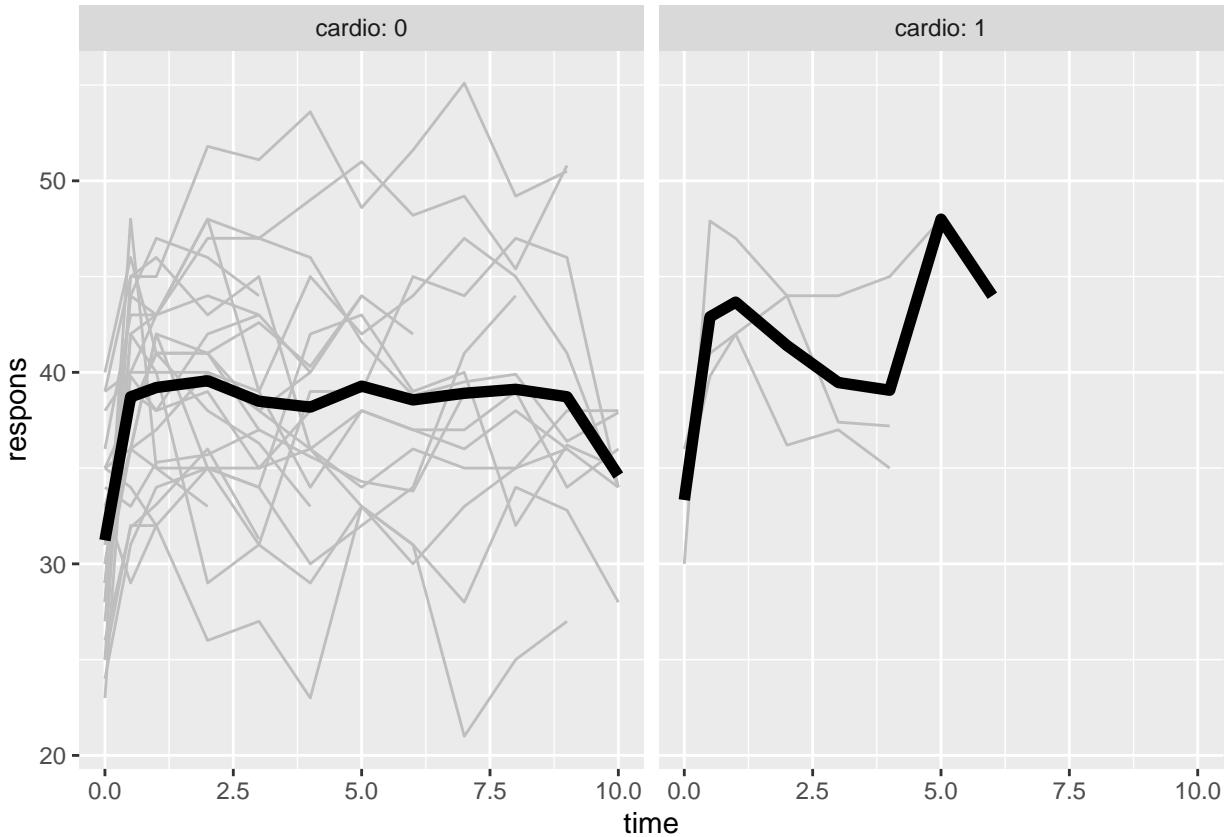
```
# Spaghetti Ggplot separated by male =1
p <- ggplot(data=data.selected,aes(x=time,y=responses,group=id))
p <- p + geom_line(col="grey") + stat_summary(aes(group=1),geom="line",fun=mean,linewidth=2)
p + facet_grid(~male,labeller=label_both)

## Warning: Removed 106 rows containing non-finite values (`stat_summary()`).
## Warning: Removed 106 rows containing missing values (`geom_line()`).
```



```
# Spaghetti Ggplot separated by cardio
p <- ggplot(data=data.selected,aes(x=time,y=respons,group=id))
p <- p + geom_line(col="grey") + stat_summary(aes(group=1),geom="line",fun=mean,linewidth=2)
cardio.labs <- c("Cardio = 0","Cardio = 1")
p + facet_grid(~cardio,labeler = label_both)

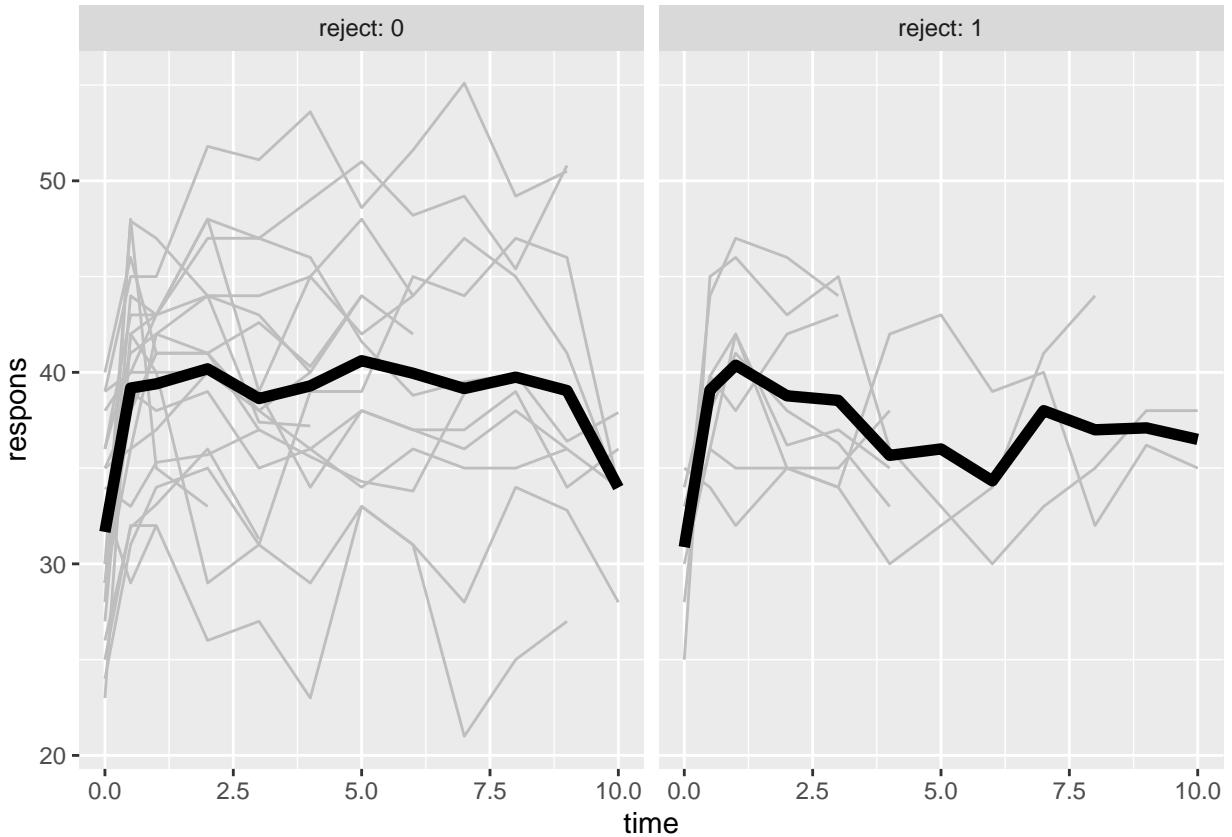
## Warning: Removed 106 rows containing non-finite values (`stat_summary()`).
## Warning: Removed 106 rows containing missing values (`geom_line()`).
```



```
#p + facet_grid(~cardio, labeller = labeller (cardio = cardio.labs))

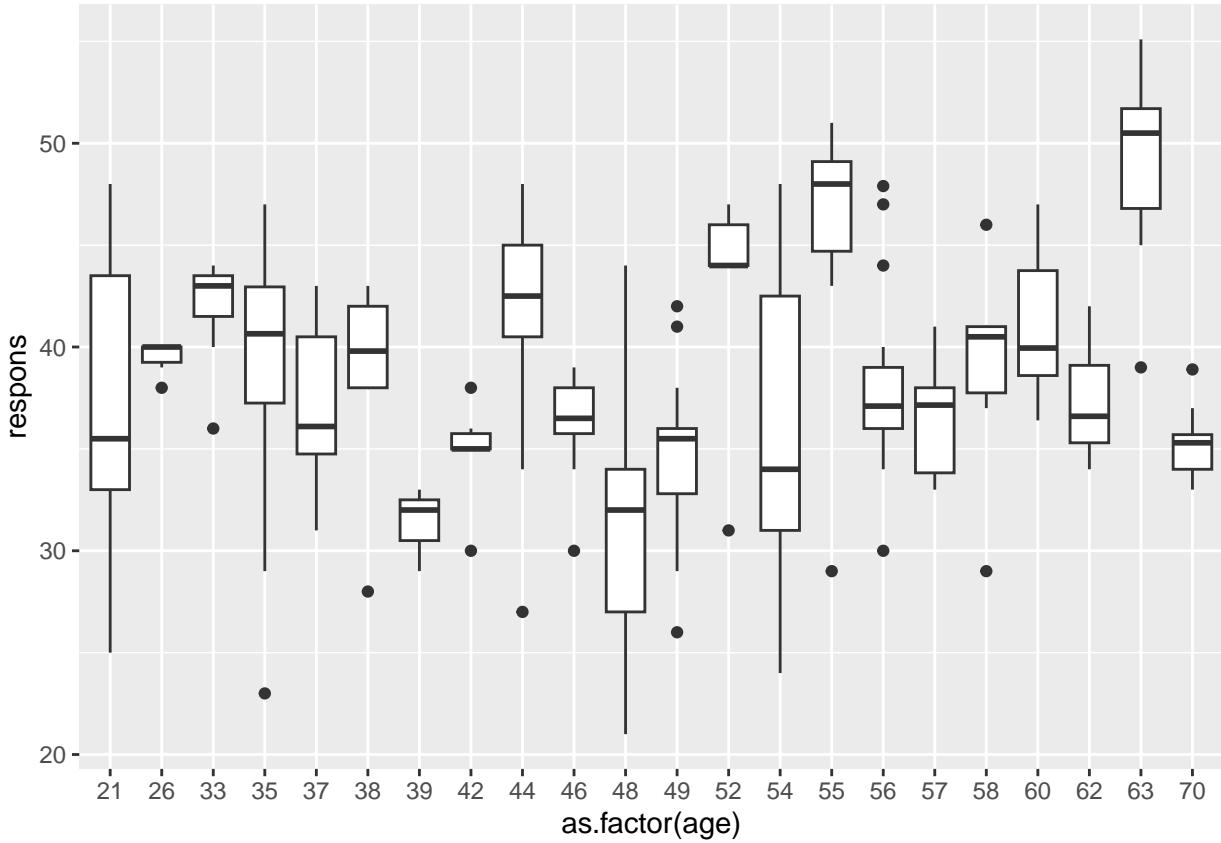
# Spaghetti Ggplot separated by reject =1
p <- ggplot(data=data.selected,aes(x=time,y=respons,group=id))
p <- p + geom_line(col="grey")+stat_summary(aes(group=1),geom="line",fun=mean,linewidth=2)
p + facet_grid(~reject,labeller=label_both)

## Warning: Removed 106 rows containing non-finite values (`stat_summary()`).
## Warning: Removed 106 rows containing missing values (`geom_line()`).
```



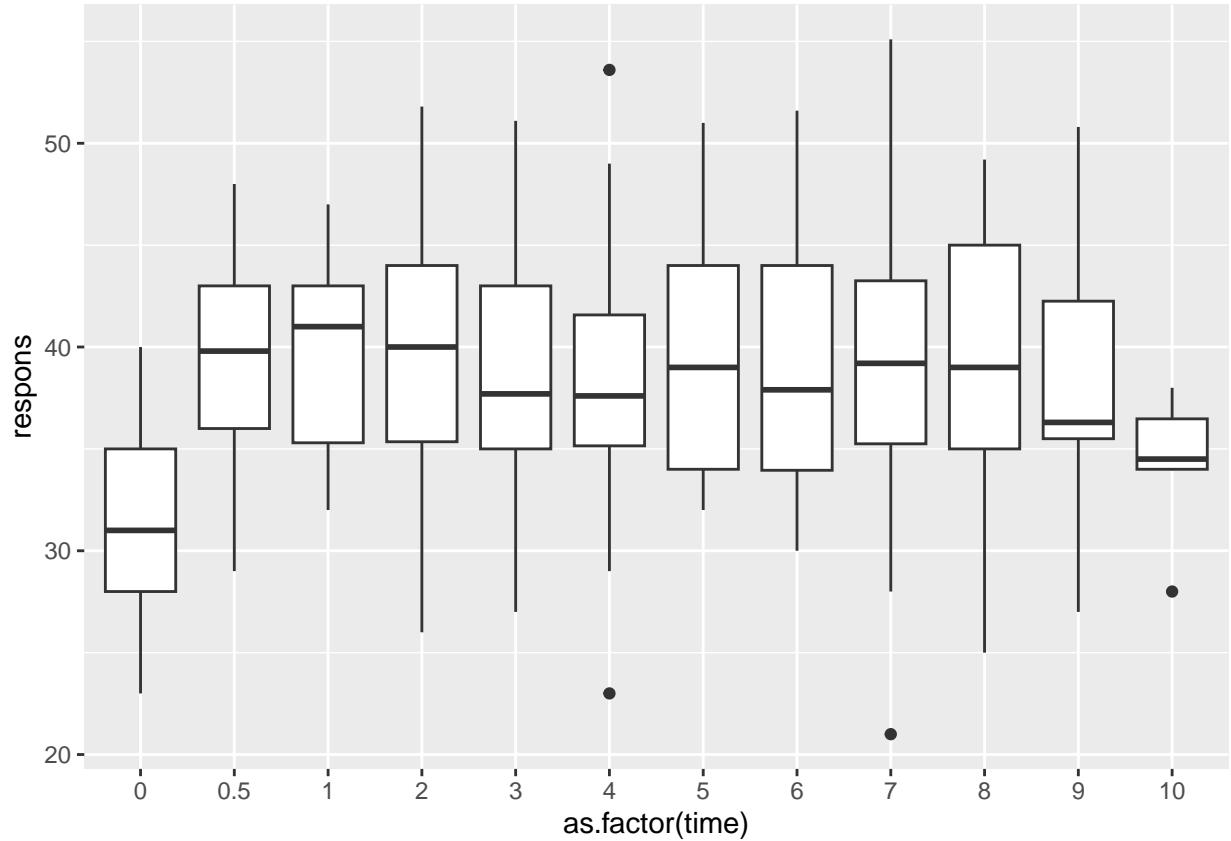
```
# BoxPlot
ggplot(data.selected,aes(x=as.factor(age),y=responses))+ geom_boxplot(position=position_dodge(1))

## Warning: Removed 106 rows containing non-finite values (`stat_boxplot()`).
```



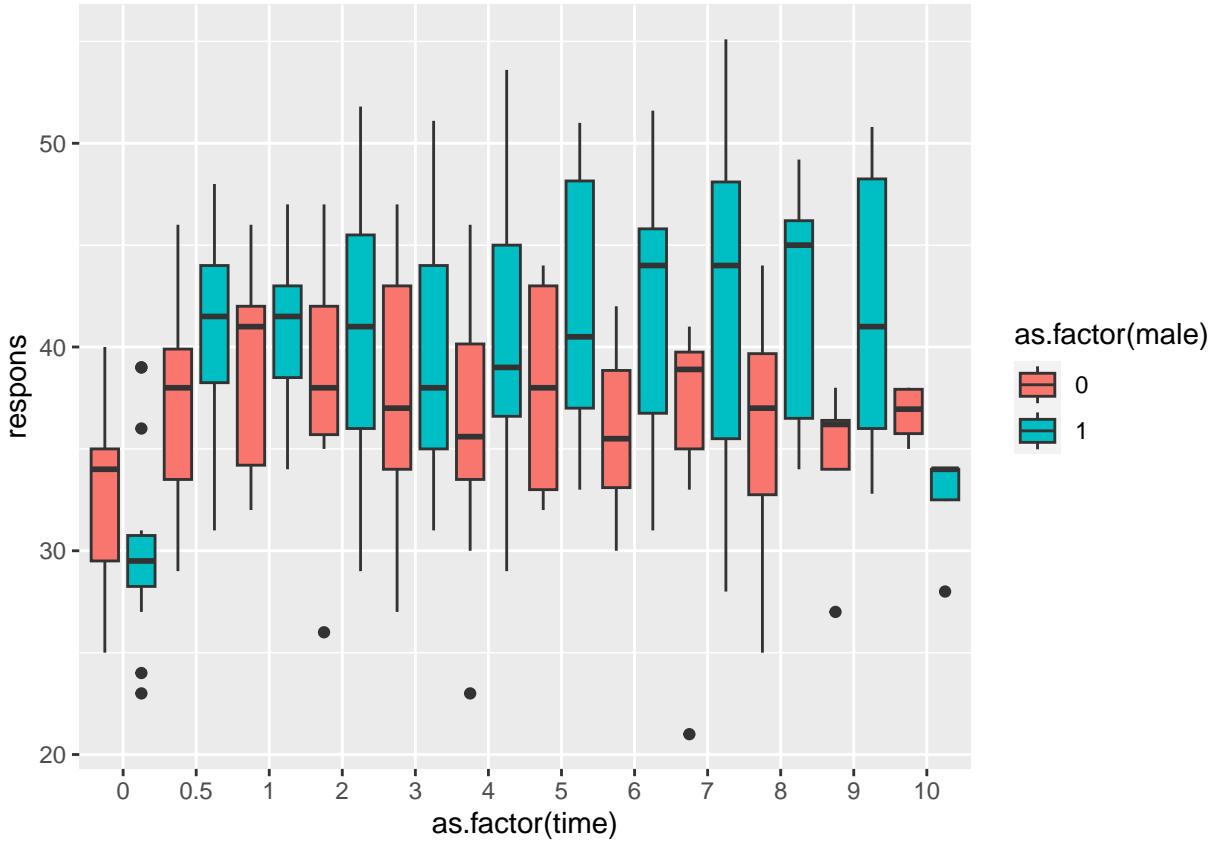
```
# BoxPlot
ggplot(data.selected,aes(x=as.factor(time),y=responses))+ geom_boxplot(position=position_dodge(1))

## Warning: Removed 106 rows containing non-finite values (`stat_boxplot()`).
```



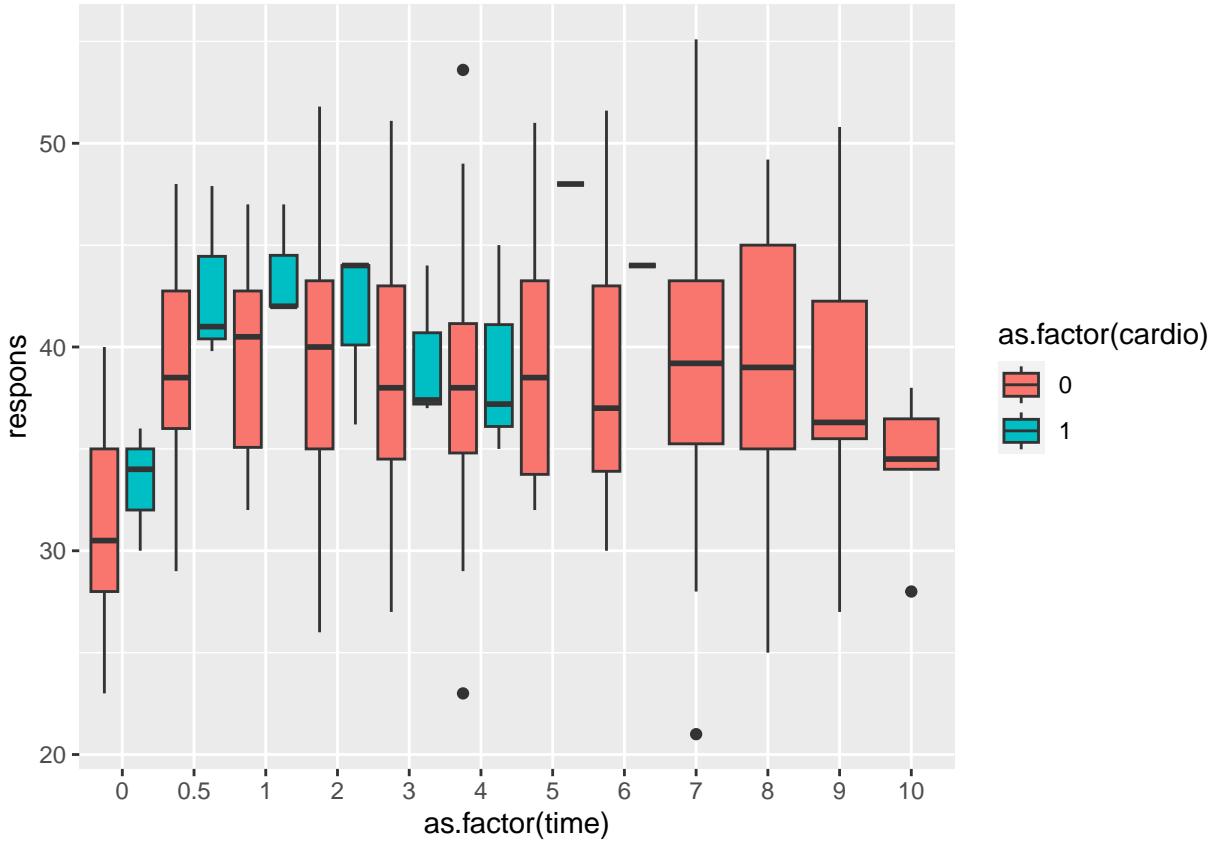
```
# Box plot by sex
ggplot(data.selected,aes(x=as.factor(time),y=responses,fill=as.factor(male)))+
  geom_boxplot(position=position_dodge(1))

## Warning: Removed 106 rows containing non-finite values (`stat_boxplot()`).
```



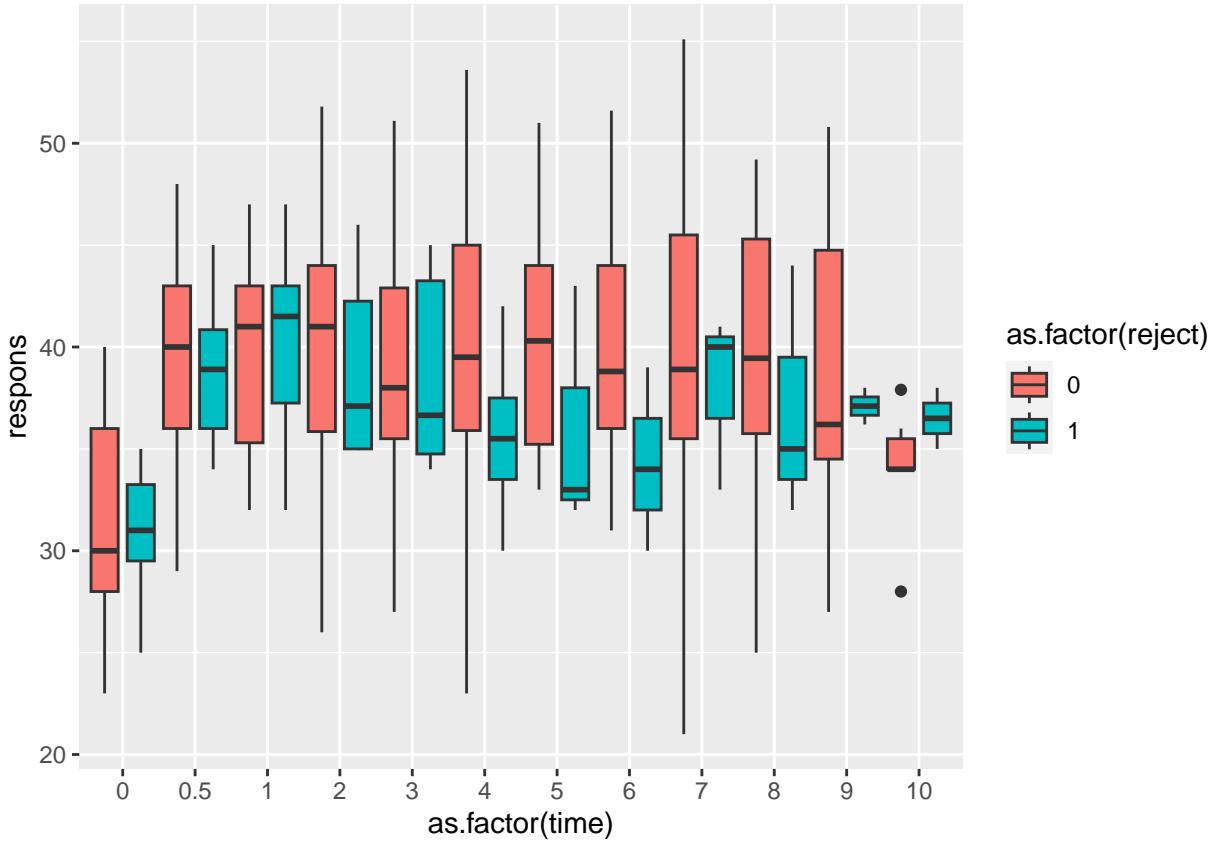
```
# Box plot by cardio
ggplot(data.selected,aes(x=as.factor(time),y=responses,fill=as.factor(cardio)))+
  geom_boxplot(position=position_dodge(1))
```

```
## Warning: Removed 106 rows containing non-finite values (`stat_boxplot()`).
```



```
# Box plot by reject
ggplot(data.selected,aes(x=as.factor(time),y=responses,fill=as.factor(reject)))+
  geom_boxplot(position=position_dodge(1))

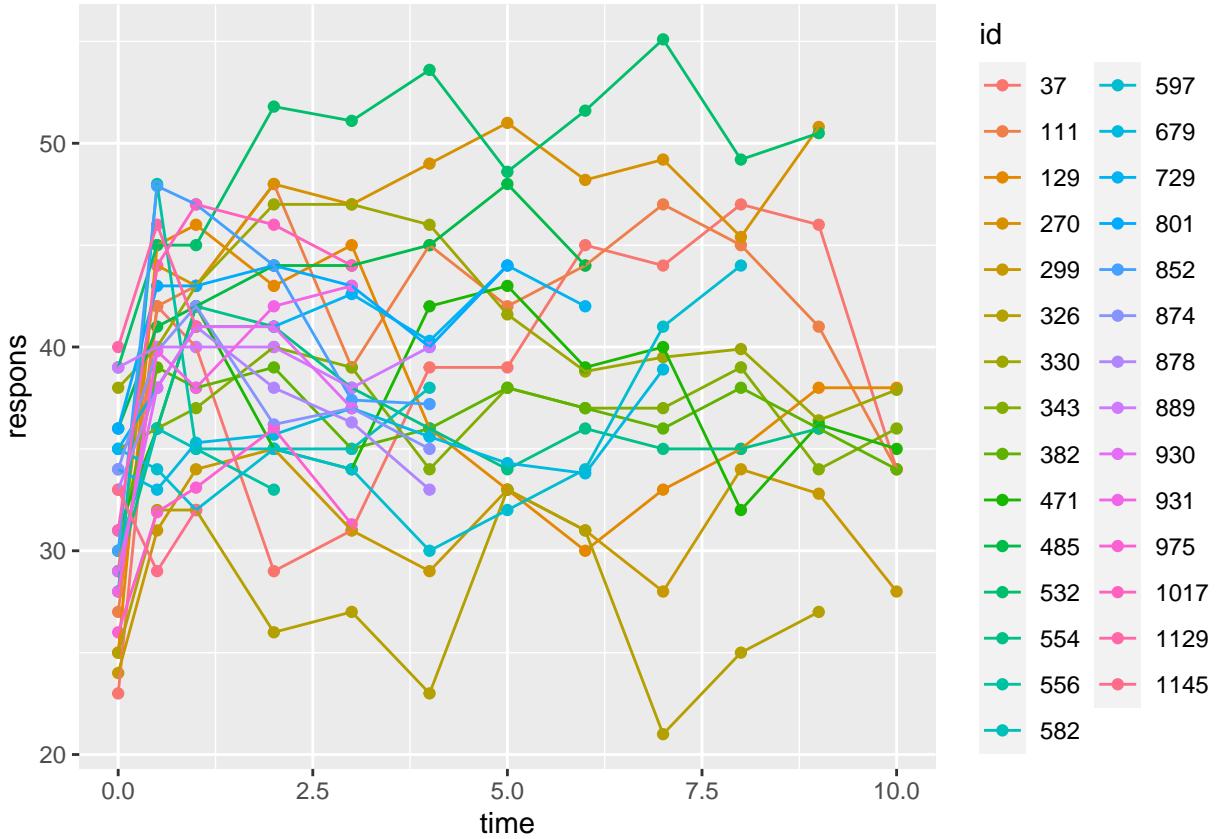
## Warning: Removed 106 rows containing non-finite values (`stat_boxplot()`).
```



```
#### Spaghetti Plot the response over time with the different persons
```

```
#data.selected = data[data$id == selected.vector,]# why the dim(data.selected) = 12 x 8
# Plot the responses over time for different id
# If some responses are not available, NA
ggplot(data.selected, aes(x=time, y=responses, group=id, color=id)) + geom_point() + geom_line()

## Warning: Removed 106 rows containing missing values (`geom_point()`).
## Warning: Removed 106 rows containing missing values (`geom_line()`).
```

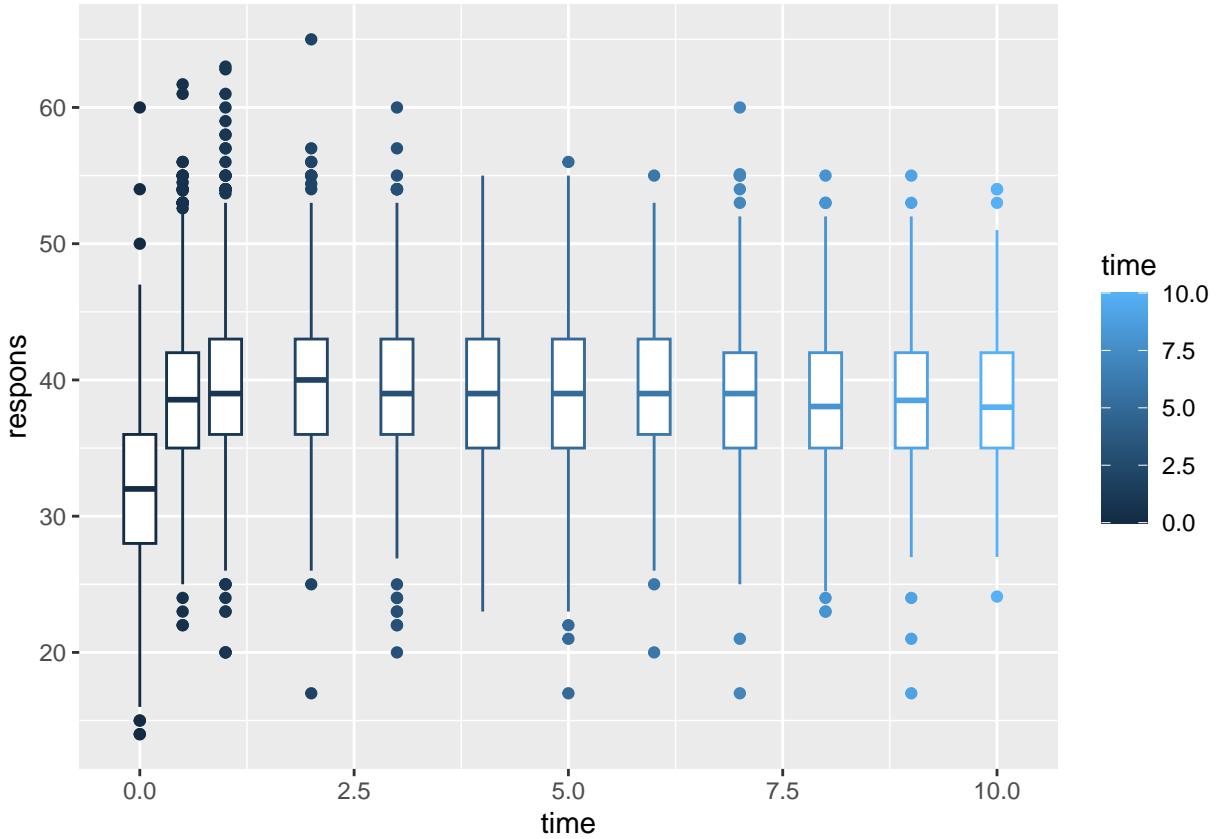


```
#ggplot(data.selected, aes(x=time, y=na.pass(respons), group=id,color=id)) + geom_point() +geom_line()
```

```
# Box plot
p <- ggplot(data, aes(x=time, y=respons,group =time, color = time)) +
  geom_boxplot()
p
```

Bax plot response over time

```
## Warning: Removed 4362 rows containing non-finite values (`stat_boxplot()`).
```

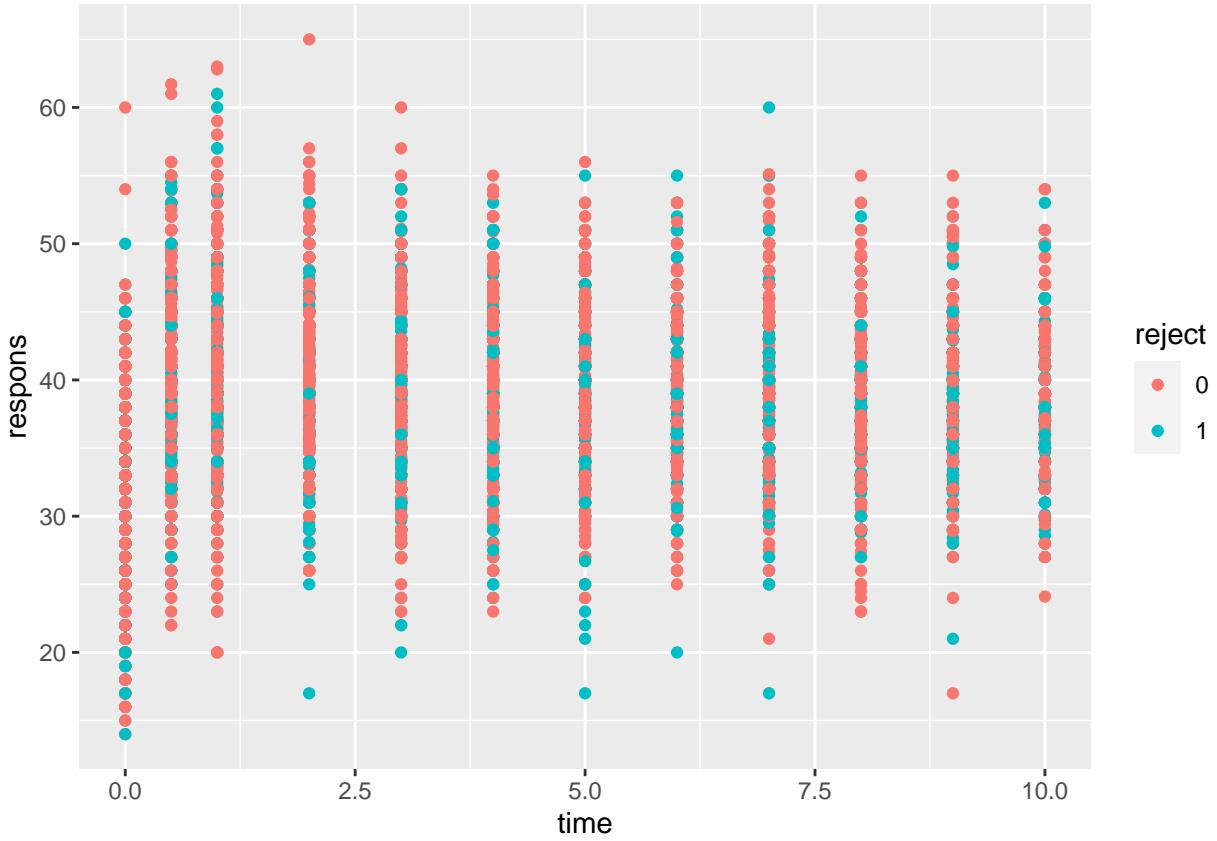


```
#### Hypothesis one HC level will change with time differently if the REJECT is different
```

```
#Plot individual data
```

```
ggplot(data, aes(x=time, y=responses, group=reject, color=reject)) + geom_point()
```

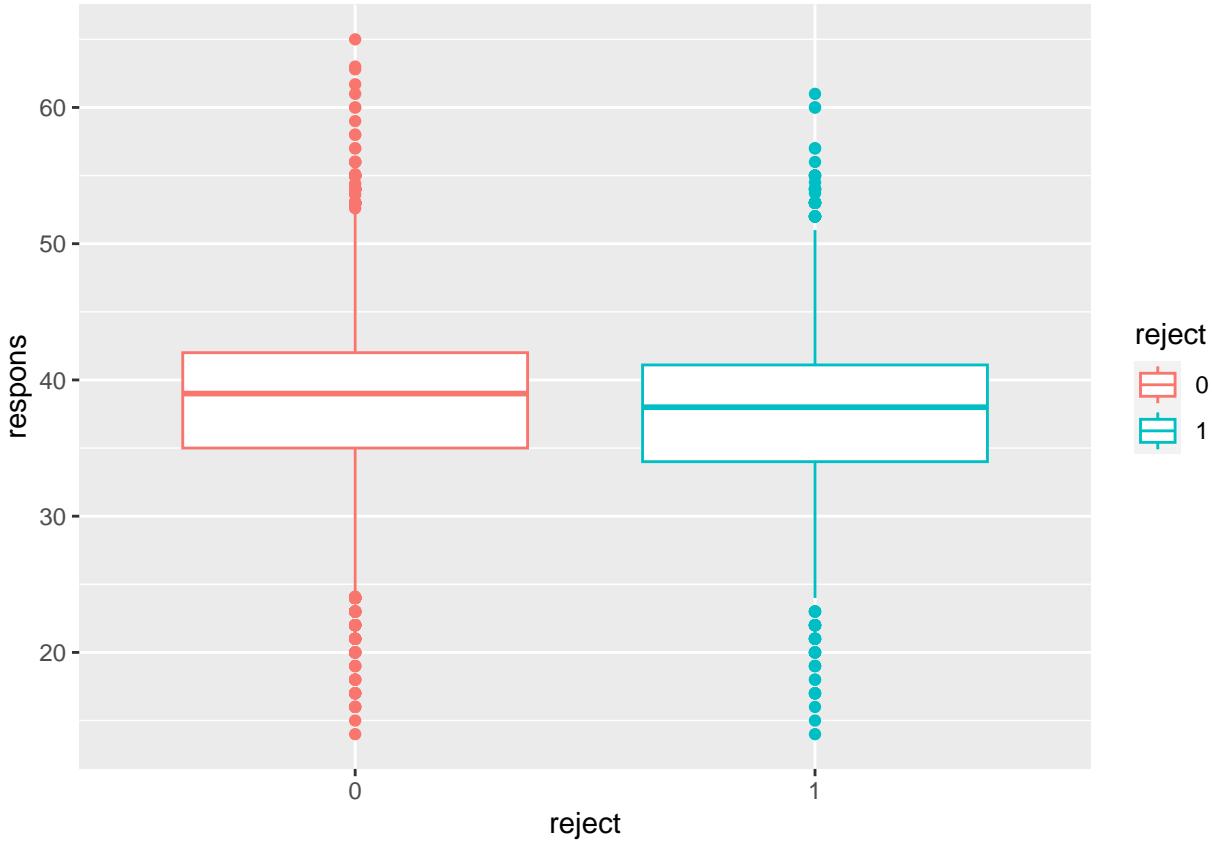
```
## Warning: Removed 4362 rows containing missing values (`geom_point()`).
```



```
##### Box plot
```

```
# Box plot
p <- ggplot(data, aes(x=reject, y=responses, group = reject, color = reject)) +
  geom_boxplot()
p
```

```
## Warning: Removed 4362 rows containing non-finite values (`stat_boxplot()`).
```



Hypothese two HC level will change with time differently if the gender is different, male (1) has generally higher HC level than female (0)

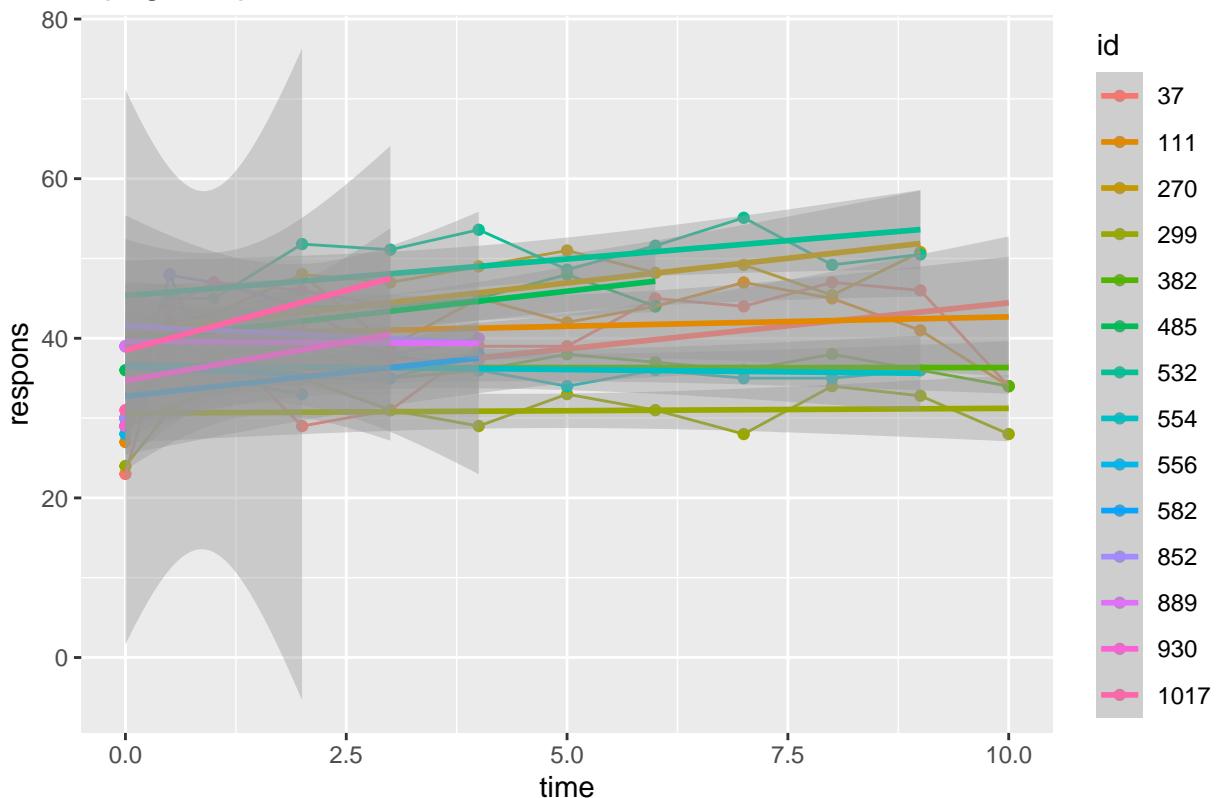
```
data.male = data[(data$male == "1"), ]
data.female = data[(data$male == "0"), ]
data.male.selected = data[(data$male == "1" & data$id %in% c(selected)), ]
data.female.selected = data[(data$male == "0" & data$id %in% c(selected)), ]
```

```
ggplot(data.male.selected, aes(x=time, y =respons, group=id,color=id)) + geom_point() +geom_line() +ggt
```

Spaghetti plots stratified by variables gender male

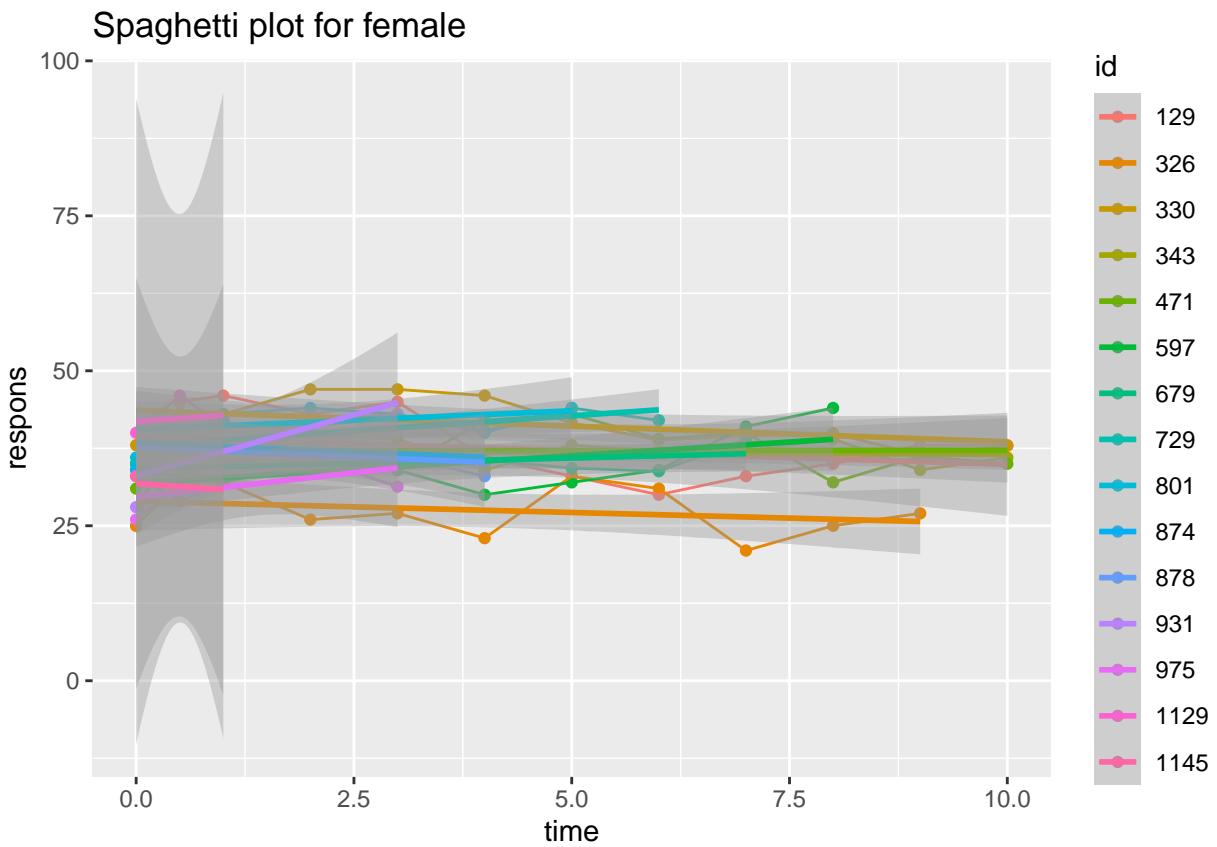
```
## `geom_smooth()` using formula = 'y ~ x'
## Warning: Removed 47 rows containing non-finite values (`stat_smooth()`).
## Warning: Removed 47 rows containing missing values (`geom_point()`).
## Warning: Removed 47 rows containing missing values (`geom_line()`).
```

Spaghetti plot for male



```
ggplot(data.female.selected, aes(x=time, y= respons, group=id,color=id)) + geom_point() +geom_line() +geom_smooth() +stat_smooth()

## `geom_smooth()` using formula = 'y ~ x'
## Warning: Removed 59 rows containing non-finite values (`stat_smooth()`).
## Warning: Removed 59 rows containing missing values (`geom_point()`).
## Warning: Removed 59 rows containing missing values (`geom_line()`).
```

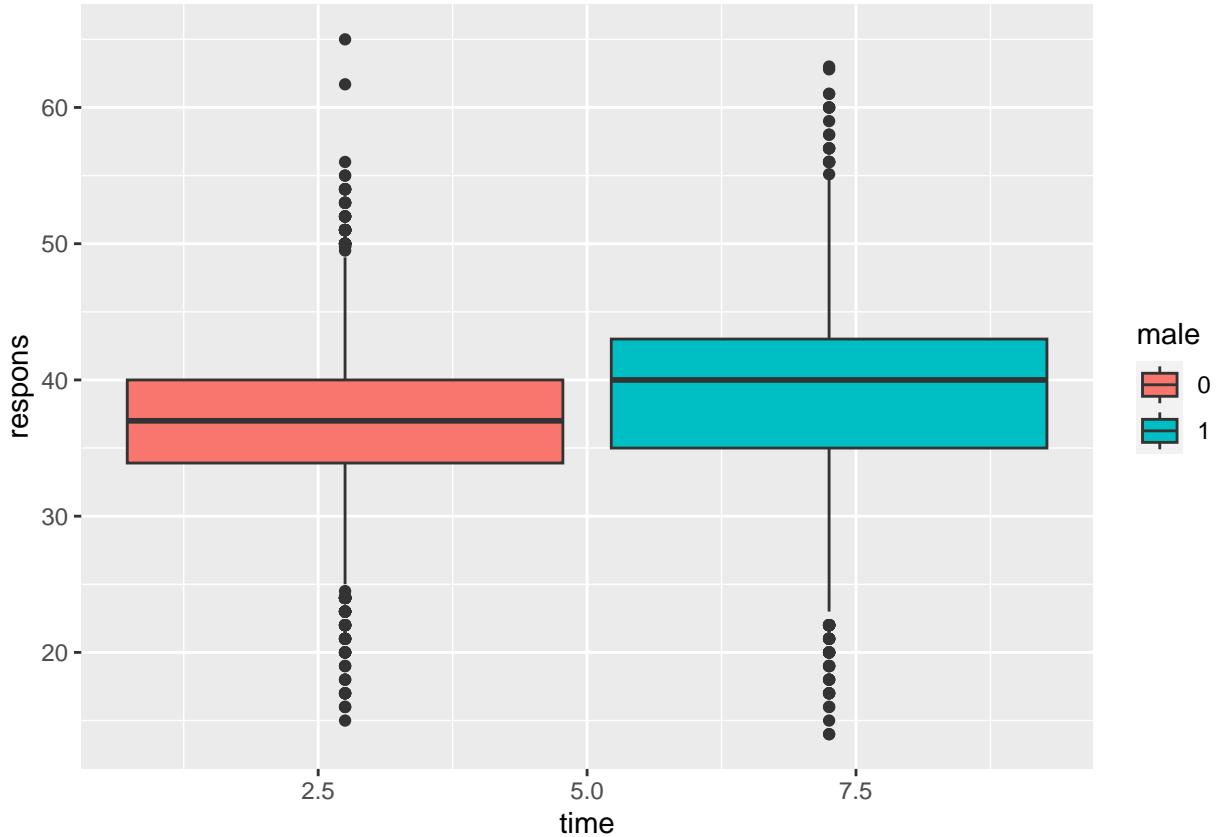


```
##### Box plot
```

```
p <- ggplot(data, aes(x=time, y=responses, fill= male)) +
  geom_boxplot()
```

```
p
```

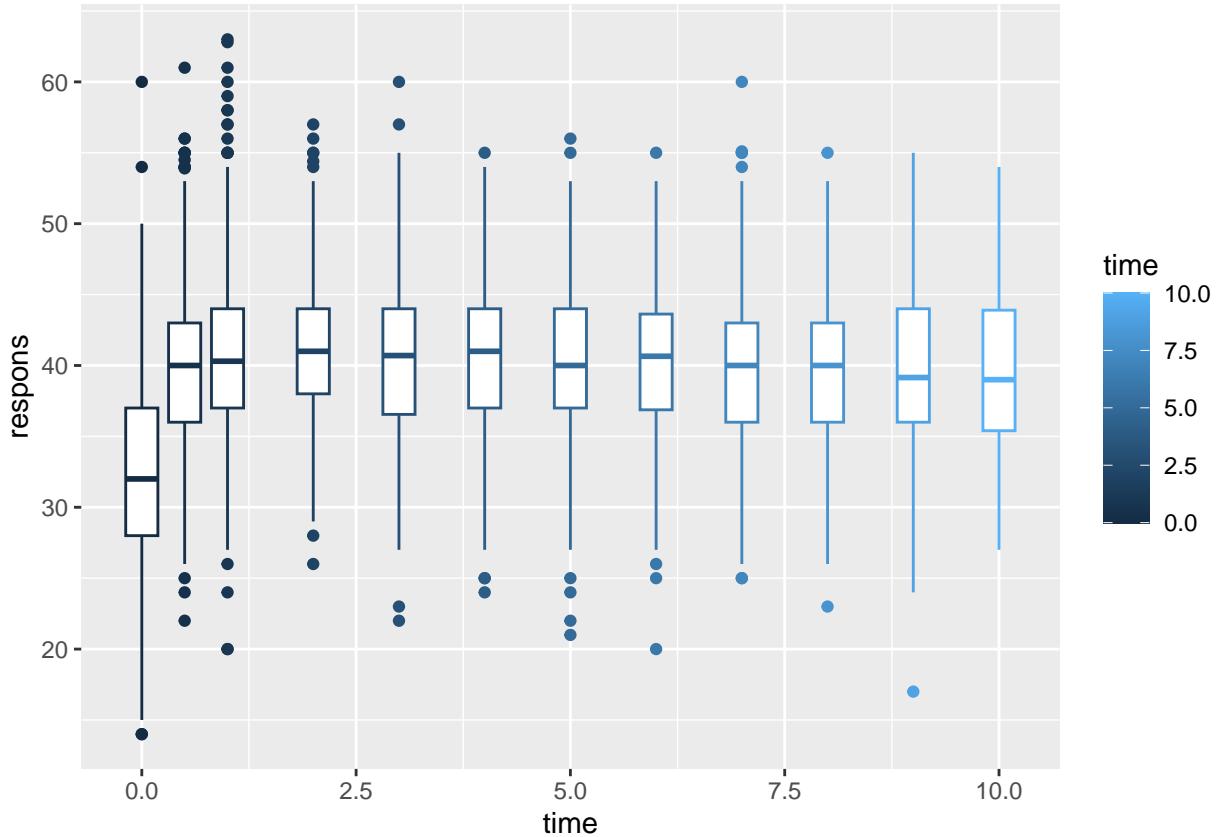
```
## Warning: Removed 4362 rows containing non-finite values (`stat_boxplot()`).
```



Box plot for male and female COULD I PLOT THEM IN THE SAME FIGURE?

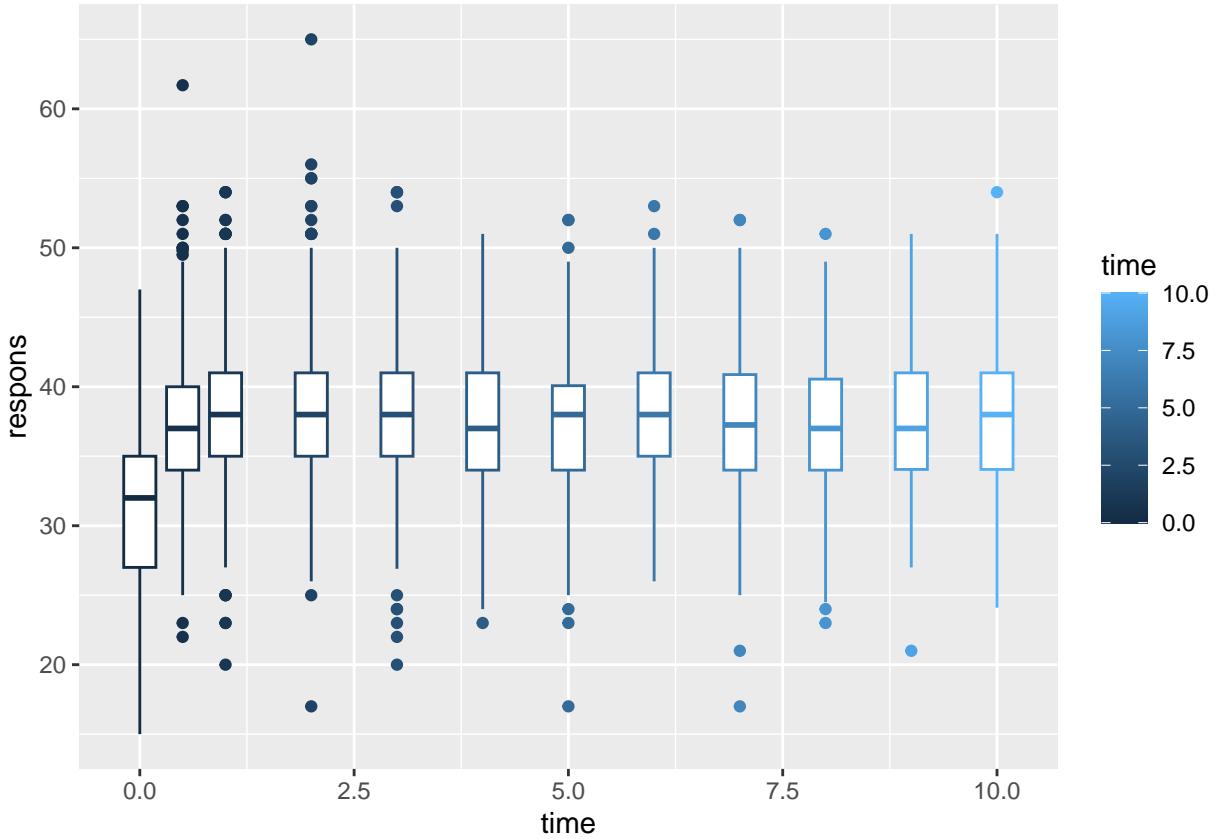
```
p.box.male <- ggplot(data.male, aes(x=time, y=responses, group = time, color = time)) +
  geom_boxplot()
p.box.male
```

```
## Warning: Removed 2647 rows containing non-finite values (`stat_boxplot()`).
```



```
p.box.female <- ggplot(data.female, aes(x=time, y=responses, group = time, color = time)) +
  geom_boxplot()
p.box.female
```

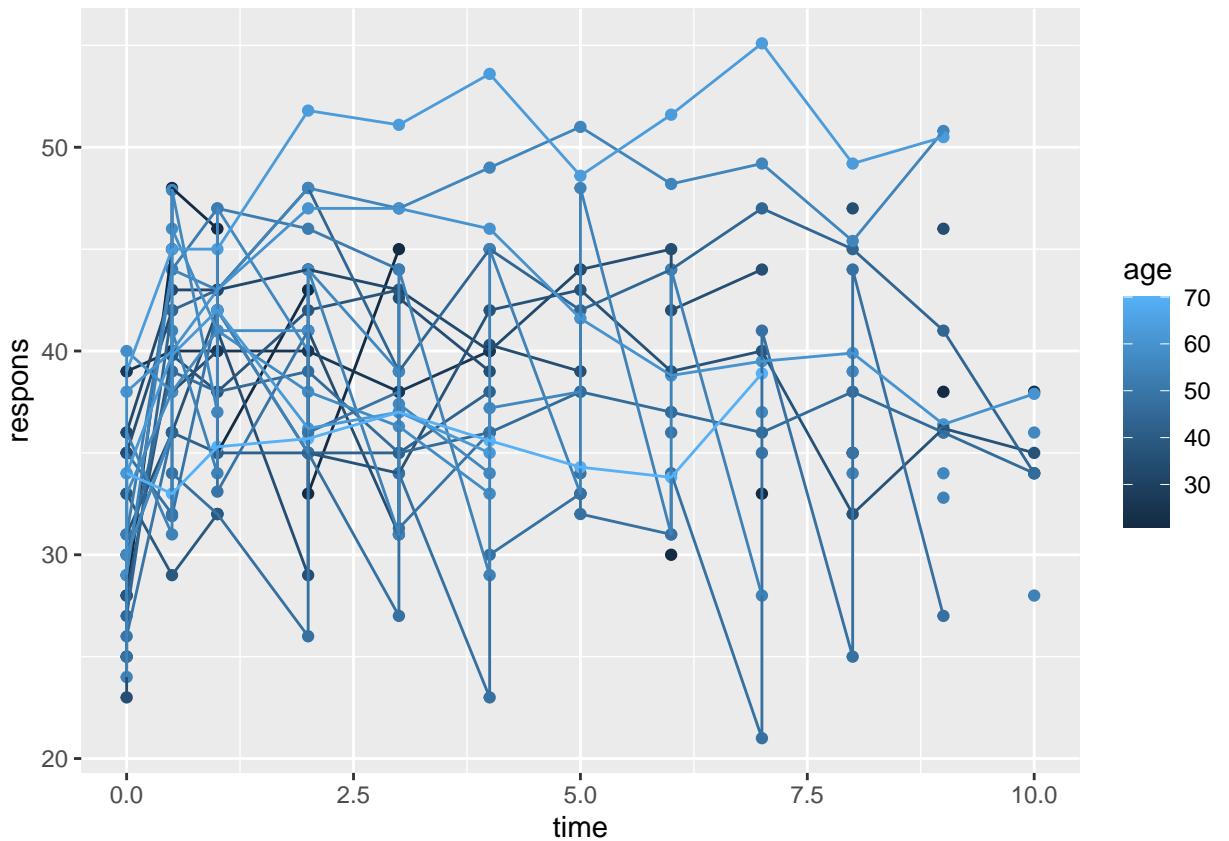
```
## Warning: Removed 1715 rows containing non-finite values (`stat_boxplot()`).
```



Hypothese three HC level will change with time differently if the age when performing the kidney transplant is younger

```
#Plot individual data
ggplot(data.selected, aes(x=time, y=responses, group=age, color=age)) + geom_point() + geom_line()

## Warning: Removed 106 rows containing missing values (`geom_point()`).
## Warning: Removed 82 rows containing missing values (`geom_line()`).
```



```

p <- ggplot(data, aes(x=age, y=responses, group = age, color = age)) +
  geom_boxplot()
p

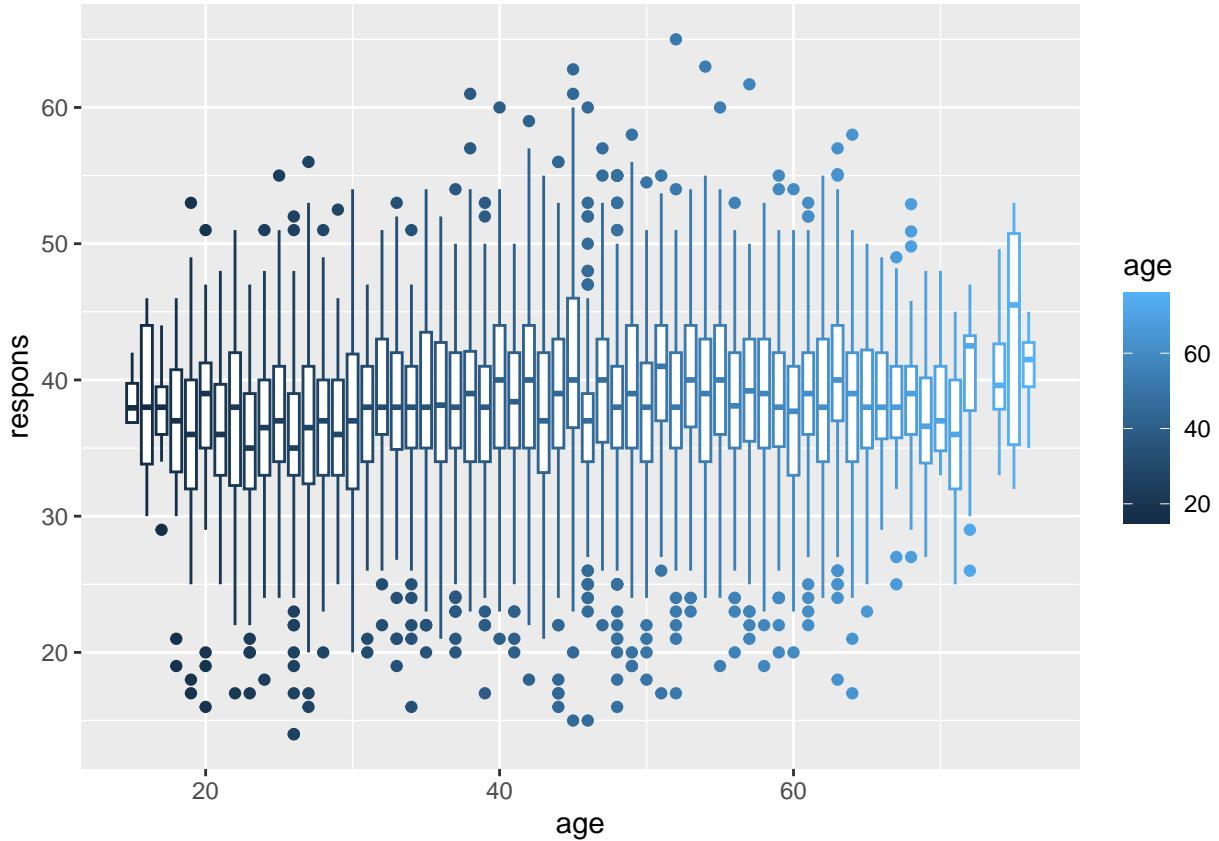
```

Box plot

```

## Warning: Removed 12 rows containing missing values (`stat_boxplot()`).
## Warning: Removed 4357 rows containing non-finite values (`stat_boxplot()`).

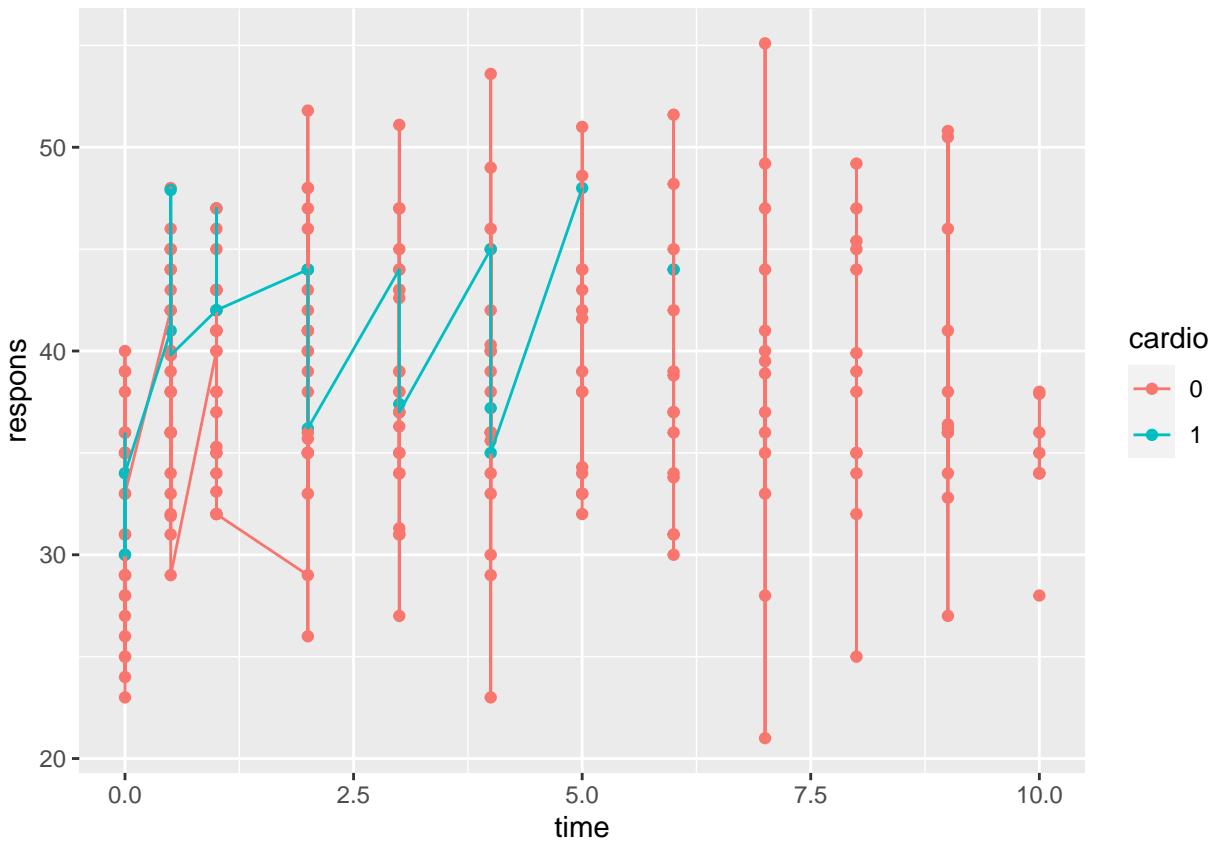
```



Hypothesis four HC level will change with time differently if the patient has experienced cardio-vascular problem during the years preceding the transplantation ##### Spathetti plot ?

```
#Plot individual data
ggplot(data.selected, aes(x=time, y=respons, group=cardio, color=cardio)) + geom_point() +geom_line()

## Warning: Removed 106 rows containing missing values (`geom_point()`).
## Warning: Removed 30 rows containing missing values (`geom_line()`).
```



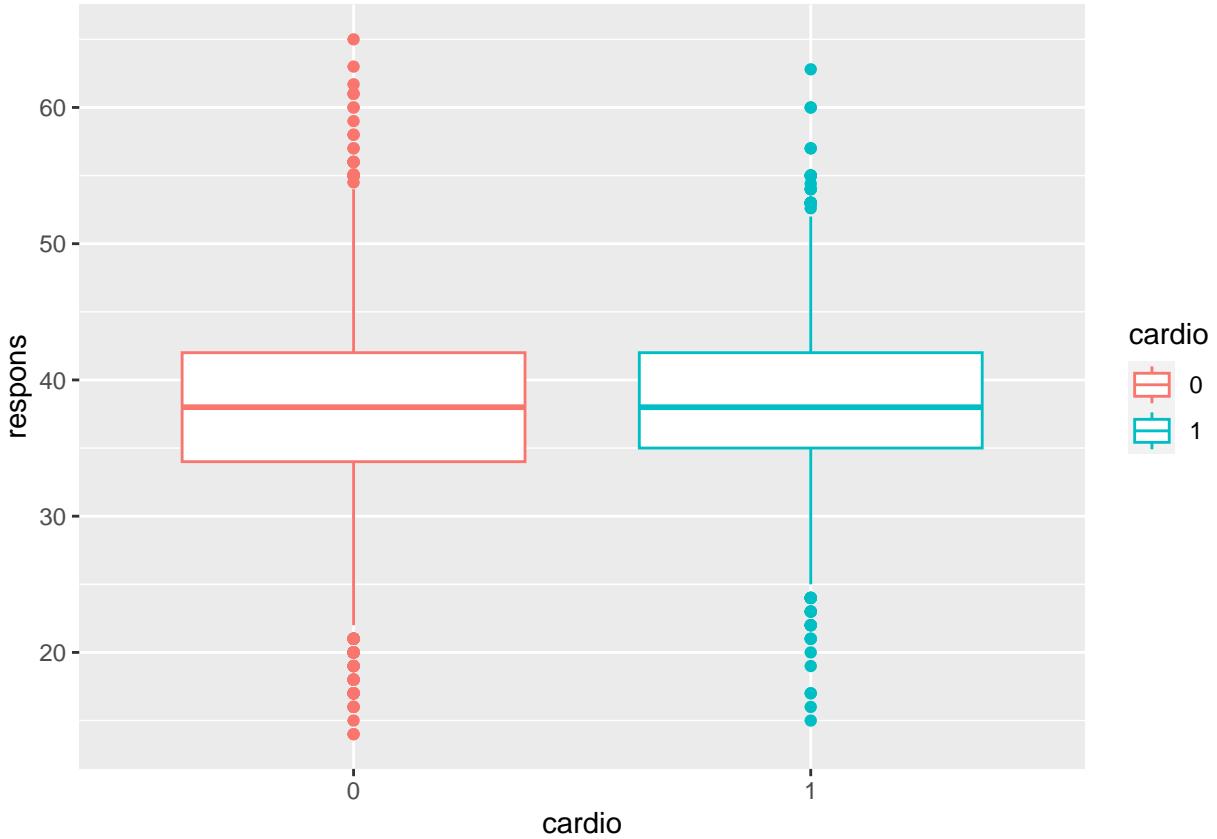
```
##### Bar plot
```

```
p <- ggplot(data, aes(x=cardio, y=responses, group = cardio, color = cardio)) +
```

```
  geom_boxplot()
```

```
p
```

```
## Warning: Removed 4362 rows containing non-finite values (`stat_boxplot()`).
```



Correlation analysis for different HC levels along time

For this purpose we need the wide table

```
trenal.wide = trenal[,1:17]
```

```
summary(trenal.wide)
```

```
##      HCO          HC06         HC1          HC2          HC3
##  Min.   :14.00   Min.   :22.00   Min.   :20.00   Min.   :17.0   Min.   :20.00
##  1st Qu.:28.00   1st Qu.:35.00   1st Qu.:36.00   1st Qu.:36.0   1st Qu.:36.00
##  Median :32.00   Median :38.55   Median :39.00   Median :40.0   Median :39.00
##  Mean   :31.86   Mean   :38.83   Mean   :39.71   Mean   :39.7   Mean   :39.17
##  3rd Qu.:36.00   3rd Qu.:42.00   3rd Qu.:43.00   3rd Qu.:43.0   3rd Qu.:43.00
##  Max.   :60.00   Max.   :61.70   Max.   :63.00   Max.   :65.0   Max.   :60.00
##  NA's    :12       NA's    :12       NA's    :12       NA's   :1044  NA's   :2460
##      HC4          HC5          HC6          HC7
##  Min.   :23.00   Min.   :17.00   Min.   :20.00   Min.   :17.00
##  1st Qu.:35.00   1st Qu.:35.00   1st Qu.:36.00   1st Qu.:35.00
##  Median :39.00   Median :39.00   Median :39.00   Median :39.00
##  Mean   :39.16   Mean   :39.02   Mean   :39.11   Mean   :38.85
##  3rd Qu.:43.00   3rd Qu.:43.00   3rd Qu.:43.00   3rd Qu.:42.00
##  Max.   :55.00   Max.   :56.00   Max.   :55.00   Max.   :60.00
##  NA's    :3768  NA's    :5016  NA's    :6096  NA's   :7140
##      HC8          HC9          HC10         id
##  Min.   :23.00   Min.   :17.00   Min.   :24.10   Min.   : 1.0
##  1st Qu.:35.00   1st Qu.:35.00   1st Qu.:35.00   1st Qu.:290.8
##  Median :38.05   Median :38.50   Median :38.00   Median : 580.5
```

```

##   Mean    :38.35   Mean    :38.57   Mean    :38.49   Mean    : 580.5
## 3rd Qu.:42.00   3rd Qu.:42.00   3rd Qu.:42.00   3rd Qu.: 870.2
## Max.    :55.00   Max.    :55.00   Max.    :54.00   Max.    :1160.0
## NA's    :8064   NA's    :8988   NA's    :9744
##      age          male         cardio        reject
##  Min.   :15.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:36.00  1st Qu.:0.0000  1st Qu.:0.0000  1st Qu.:0.0000
##  Median :48.00  Median :1.0000  Median :0.0000  Median :0.0000
##  Mean   :46.43  Mean   :0.5741  Mean   :0.1784  Mean   :0.3164
##  3rd Qu.:57.00  3rd Qu.:1.0000  3rd Qu.:0.0000  3rd Qu.:1.0000
##  Max.   :76.00  Max.   :1.0000  Max.   :1.0000  Max.   :1.0000
##  NA's   :12

HcCorr = trenal.wide[,c(1:12)]
cor(HcCorr,use="complete.obs" ) # also COV for covariance

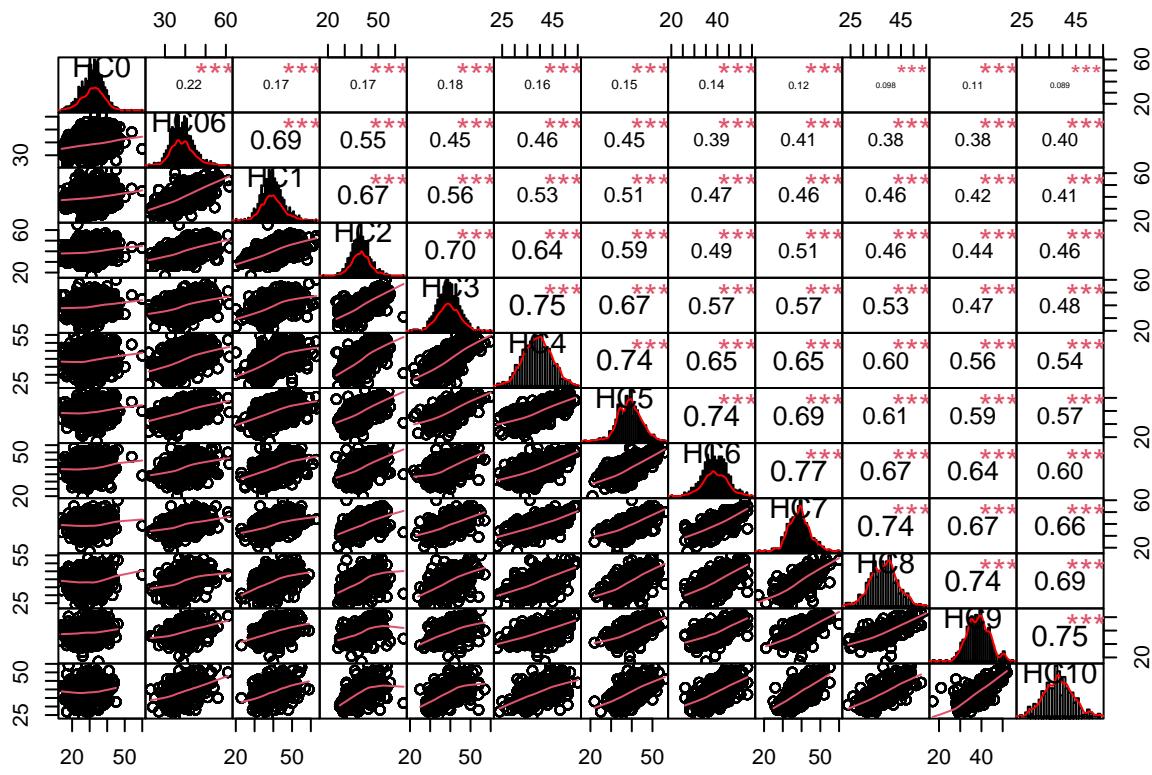
##           HC0       HC06       HC1       HC2       HC3       HC4       HC5
## HC0  1.00000000 0.2264123 0.1587116 0.1724777 0.2139805 0.1732267 0.1557624
## HC06 0.22641235 1.0000000 0.7562367 0.6233688 0.5520591 0.5278499 0.5143061
## HC1  0.15871158 0.7562367 1.0000000 0.7315995 0.6656006 0.6119867 0.5873331
## HC2  0.17247771 0.6233688 0.7315995 1.0000000 0.7284046 0.6382434 0.5996189
## HC3  0.21398049 0.5520591 0.6656006 0.7284046 1.0000000 0.7733522 0.7016965
## HC4  0.17322666 0.5278499 0.6119867 0.6382434 0.7733522 1.0000000 0.7888249
## HC5  0.15576243 0.5143061 0.5873331 0.5996189 0.7016965 0.7888249 1.0000000
## HC6  0.13620085 0.4569881 0.5004036 0.4869519 0.5786122 0.6814132 0.7592203
## HC7  0.10156045 0.3936597 0.4541699 0.4724703 0.5402798 0.6466212 0.7067887
## HC8  0.08419757 0.3687935 0.4454882 0.4244221 0.5030428 0.6040136 0.6080051
## HC9  0.08859254 0.3711560 0.4254622 0.3971477 0.4303661 0.5461579 0.5713338
## HC10 0.09718506 0.4210917 0.4301937 0.4647890 0.4972001 0.5629570 0.5800544
##          HC6       HC7       HC8       HC9       HC10
## HC0  0.1362008 0.1015604 0.08419757 0.08859254 0.09718506
## HC06 0.4569881 0.3936597 0.36879347 0.37115604 0.42109175
## HC1  0.5004036 0.4541699 0.44548815 0.42546216 0.43019368
## HC2  0.4869519 0.4724703 0.42442213 0.39714773 0.46478897
## HC3  0.5786122 0.5402798 0.50304282 0.43036614 0.49720006
## HC4  0.6814132 0.6466212 0.60401365 0.54615793 0.56295695
## HC5  0.7592203 0.7067887 0.60800514 0.57133378 0.58005440
## HC6  1.0000000 0.7414970 0.67347761 0.62938253 0.60329422
## HC7  0.7414970 1.0000000 0.71838142 0.63933448 0.65646214
## HC8  0.6734776 0.7183814 1.00000000 0.70316750 0.68501304
## HC9  0.6293825 0.6393345 0.70316750 1.00000000 0.74259683
## HC10 0.6032942 0.6564621 0.68501304 0.74259683 1.00000000

library("PerformanceAnalytics")

## Warning: package 'PerformanceAnalytics' was built under R version 4.2.3
## Loading required package: xts
## Loading required package: zoo
##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
## 
##      as.Date, as.Date.numeric

```

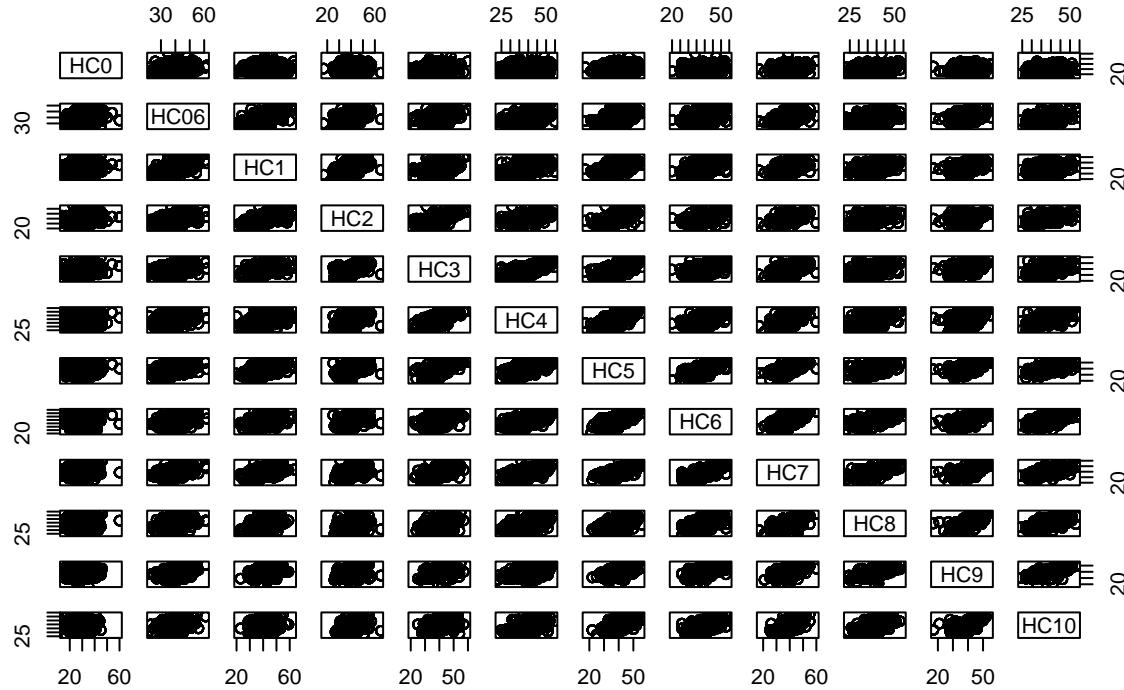



```
cor(trenal.wide$HCO,trenal.wide$HC06)
```

```
## [1] NA
# scatter plot matrix

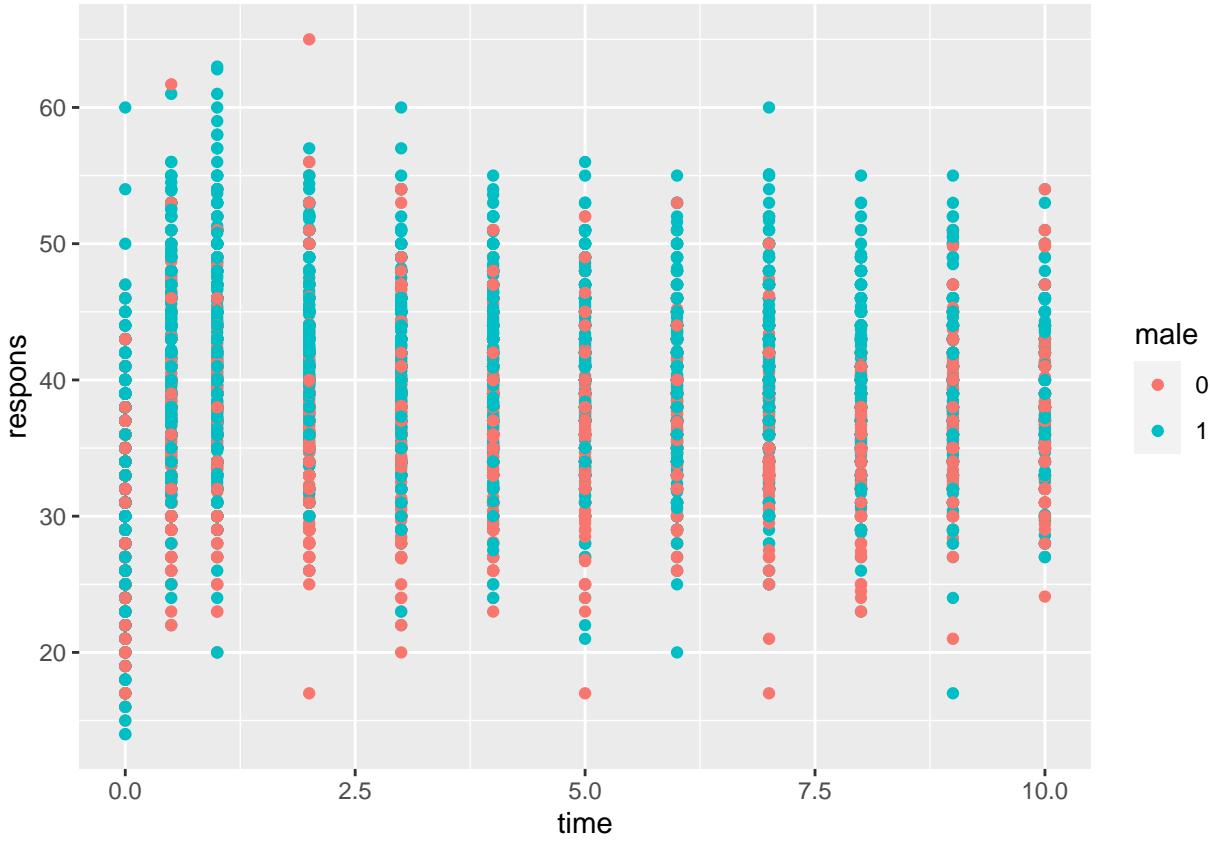
pairs(~HCO+HC06+HC1+HC2+HC3+HC4+HC5+HC6+HC7+HC8+HC9+HC10, data=trenal.wide,
      main="Simple Scatterplot Matrix")
```

Simple Scatterplot Matrix



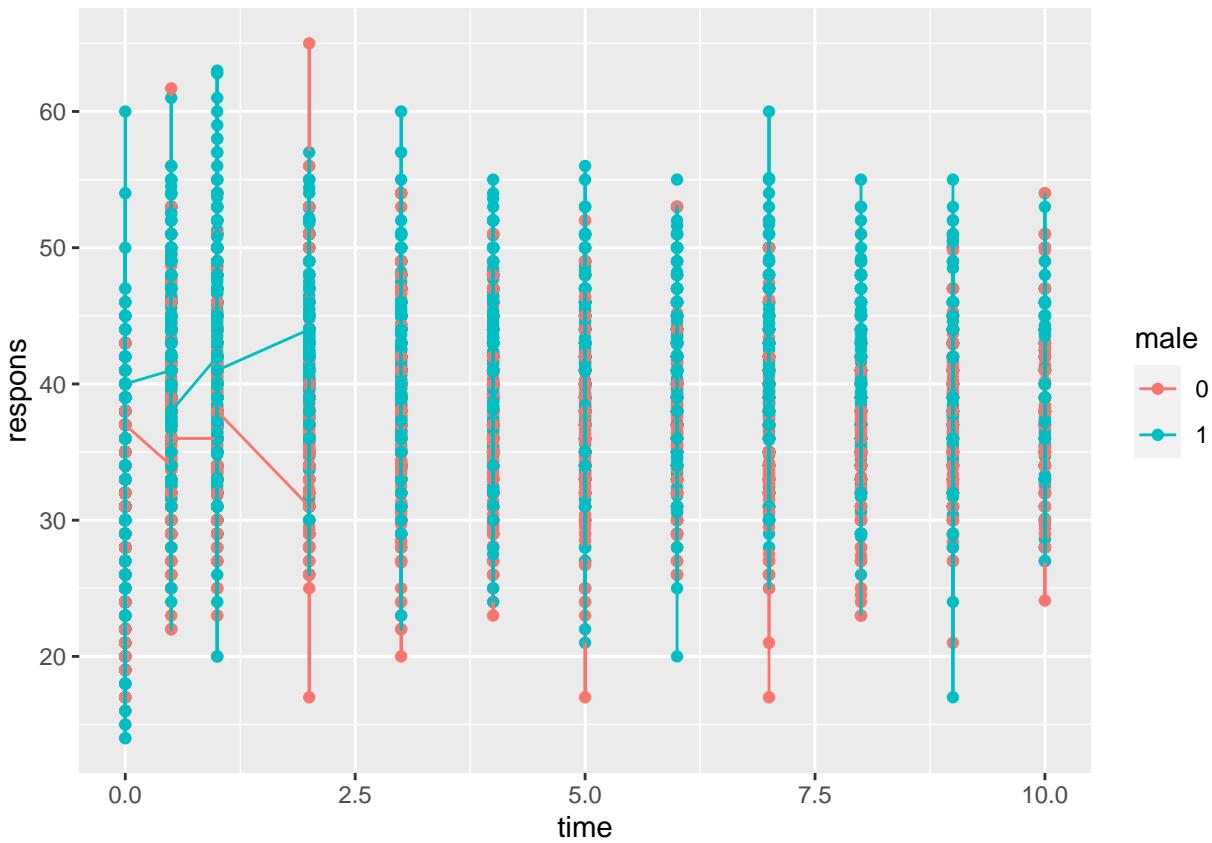
```
#Plot individual data  
ggplot(data, aes(x=time, y=respons, group=male, color=male)) + geom_point()
```

```
## Warning: Removed 4362 rows containing missing values (`geom_point()`).
```



```
#Spaghetti Plot
ggplot(data, aes(x=time, y=responses, group=maale,color=maale)) + geom_point() +geom_line()

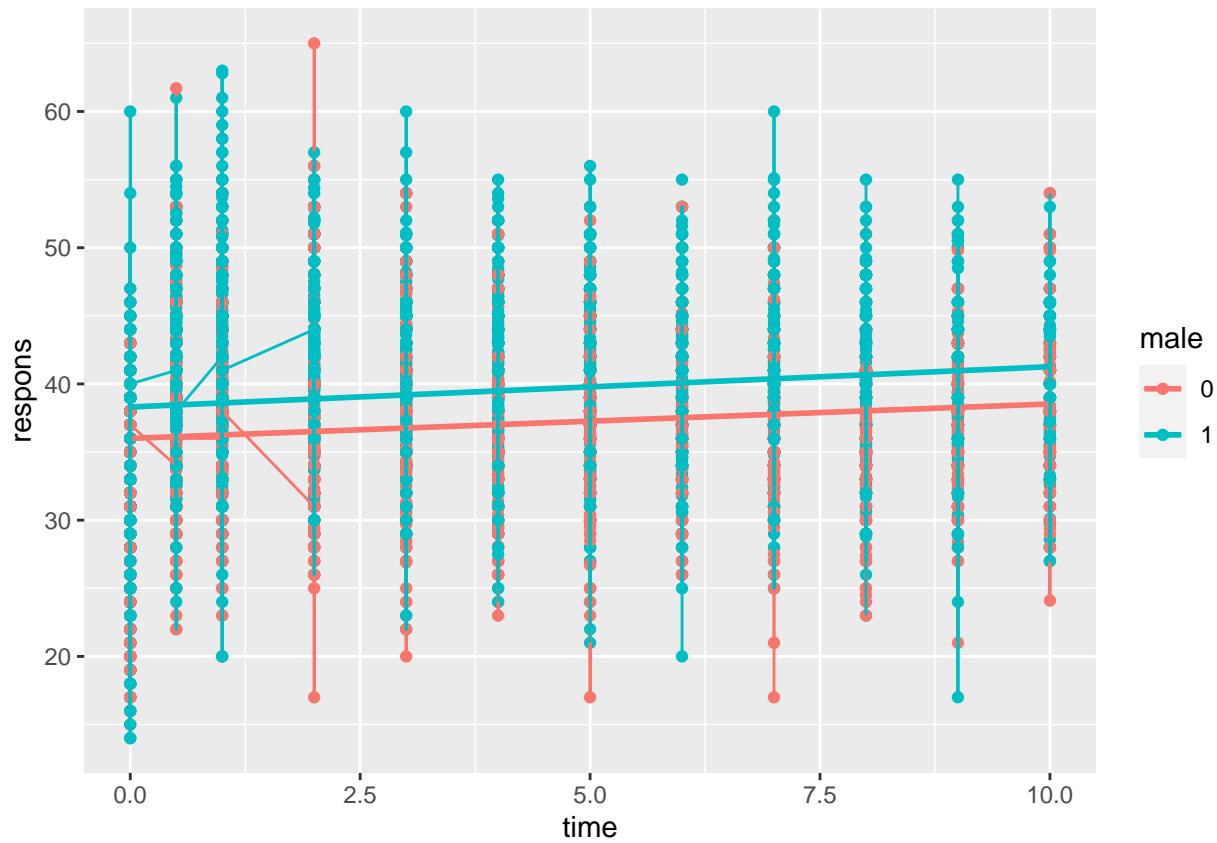
## Warning: Removed 4362 rows containing missing values (`geom_point()`).
## Warning: Removed 638 rows containing missing values (`geom_line()`).
```



#Spaghetti with fitted lines

```
ggplot(data, aes(x=time, y=responses, group=maale, color=maale)) + geom_point() + geom_smooth(method="lm", se=TRUE)

## `geom_smooth()` using formula = 'y ~ x'
## Warning: Removed 4362 rows containing non-finite values (`stat_smooth()`).
## Warning: Removed 4362 rows containing missing values (`geom_point()`).
## Warning: Removed 638 rows containing missing values (`geom_line()`).
```



```
## Linear mixed effect model
library(nlme)
library(lme4)
library(lattice)
library(ggplot2)
```

First model only consider fixed effects, no random effects, use linear model and linear mixed model

```
# First model, no random effects
model.noRandomEffects <- lm(respons ~ time, data=data)
summary(model.noRandomEffects)
```

```
##
## Call:
## lm(formula = respons ~ time, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.3368 -3.8633  0.0393  3.8206 27.1367
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.33685   0.09410 396.8 <2e-16 ***
## time        0.26322   0.02073   12.7 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.023 on 9556 degrees of freedom
## (4362 observations deleted due to missingness)
## Multiple R-squared:  0.01659, Adjusted R-squared:  0.01648
## F-statistic: 161.2 on 1 and 9556 DF, p-value: < 2.2e-16
```

Read from the table, we have a same model for each subject i :

$$\text{Respons}_i = 37.33685 + 0.26322 \times \text{time} + \epsilon_i$$

where ϵ has a large variance Residuals: Min 1Q Median 3Q Max -23.3368 -3.8633 0.0393 3.8206 27.1367

```
# second model random intercept
model.inter <- lmer(respons~time+(1|id), data=data)
summary(model.inter)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: respons ~ time + (1 | id)
## Data: data
##
## REML criterion at convergence: 58849.2
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -5.5197 -0.4741  0.0847  0.5683  6.4131
##
## Random effects:
## Groups   Name        Variance Std.Dev.
## id       (Intercept) 13.69    3.700
## Residual            22.36    4.729
## Number of obs: 9558, groups: id, 1160
```

```

## 
## Fixed effects:
##           Estimate Std. Error t value
## (Intercept) 37.2953     0.1317 283.24
## time        0.2913     0.0178 16.36
## 
## Correlation of Fixed Effects:
##      (Intr) 
## time -0.402

```

Read from the table, we have a different model for each subject i

$$\text{Respons}_i = 37.295 + b_{i0} + (0.2913 + b_{i1}) \times \text{time} + \epsilon_i$$

Now the residual ϵ_i has smaller variance Scaled residuals: Min 1Q Median 3Q Max -5.5197 -0.4741 0.0847 0.5683 6.4131

Is the random intercept giving information?

```

value1 <- as.numeric(2*(logLik(model.inter)-(logLik(model.noRandomEffects))))
value1

```

```

## [1] 2597.099
p1 = 0.5 * (1-pchisq(value1,1))
p1

```

```

## [1] 0
library(nadiv)
#?LRTest
LRTest(logLik(model.inter),logLik(model.noRandomEffects))

```

```

## $lambda
## 'log Lik.' 2597.099 (df=4)
## 
## $Pval
## 'log Lik.' 0 (df=4)
## 
## $corrected.Pval
## [1] FALSE

```

Thus the model.inter outperforms the linear model

Third model Random Slope

```

model.slope = lmer(respons ~ time + (time-1|id), data=data)
summary(model.slope)

```

```

## Linear mixed model fit by REML ['lmerMod']
## Formula: respons ~ time + (time - 1 | id)
##   Data: data
## 
## REML criterion at convergence: 60045.8
## 
## Scaled residuals:
##      Min      1Q  Median      3Q      Max 
## -4.4685 -0.4913  0.0113  0.5322  5.4991

```

```

##
## Random effects:
## Groups      Name Variance Std.Dev.
## id          time  0.584   0.7642
## Residual    27.031   5.1991
## Number of obs: 9558, groups:  id, 1160
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept) 37.23239   0.08413 442.54
## time         0.35231   0.03425 10.29
##
## Correlation of Fixed Effects:
##           (Intr)
## time -0.512

```

Read from the table, we have a different model for each subject i :

$$\text{Respons}_i = 37.23239 + (0.35231 + b_{i1}) \times \text{time} + \epsilon_i$$

Is this random slope giving information?

```
value2 <- as.numeric(2*(logLik(model.slope)-(logLik(model.noRandomEffects))))
value2
```

```

## [1] 1400.468
p2 = 0.5*(1-pchisq(value2,1))
p2

## [1] 0
LRTest(logLik(model.slope),logLik(model.noRandomEffects))
```

```

## $lambda
## 'log Lik.' 1400.468 (df=4)
##
## $Pval
## 'log Lik.' 1.662142e-306 (df=4)
##
## $corrected.Pval
## [1] FALSE
value12 <- as.numeric(2*(logLik(model.slope)-(logLik(model.inter))))
value12
```

```

## [1] -1196.63
p12 = 0.5*(1-pchisq(value12,1))
p12

## [1] 0.5
LRTest(logLik(model.slope),logLik(model.inter))
```

```

## $lambda
## 'log Lik.' -1196.63 (df=4)
##
## $Pval
```

```

## 'log Lik.' 1 (df=4)
##
## $corrected.Pval
## [1] FALSE
value12 <- as.numeric(2*(logLik(model.slope)-(logLik(model.inter))))
value12

## [1] -1196.63
p12 = 0.5*(1-pchisq(value12,1))
p12

## [1] 0.5
LRTest(logLik(model.inter),logLik(model.slope))

## $lambda
## 'log Lik.' 1196.63 (df=4)
##
## $Pval
## 'log Lik.' 3.293312e-262 (df=4)
##
## $corrected.Pval
## [1] FALSE

model.inter outperforms than the model.slope

```

Fourth Model

Fixed effect(changes the same for all subjects) could be time, male, age, reject, cardio Random effect will start from only intercept, to the slope of gender, age, reject, cardio

Only fixed effects

```

model.noRandomEffects.full = lm(respons ~ time+age+male+cardio+reject, data=data)
summary(model.noRandomEffects.full)

##
## Call:
## lm(formula = respons ~ time + age + male + cardio + reject, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.9829  -3.6915   0.1453   3.6927  27.9570
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.103534  0.259284 127.673  <2e-16 ***
## time        0.305654  0.020229  15.109  <2e-16 ***
## age         0.064002  0.004847  13.204  <2e-16 ***
## male1       2.449916  0.120545  20.324  <2e-16 ***
## cardio1    -0.417558  0.165619  -2.521   0.0117 *
## reject1    -0.323401  0.128656  -2.514   0.0120 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.84 on 9545 degrees of freedom

```

```

##  (4369 observations deleted due to missingness)
## Multiple R-squared:  0.07615,   Adjusted R-squared:  0.07566
## F-statistic: 157.3 on 5 and 9545 DF,  p-value: < 2.2e-16

data.noNA <- na.omit(data)
summary(data.noNA)

##      id          j          time        respons       age
## 3 : 12    2 :1159  Min.   : 0.000  Min.   :14.00  Min.   :15.00
## 5 : 12    1 :1158  1st Qu.: 1.000  1st Qu.:34.00  1st Qu.:35.00
## 6 : 12    3 :1158  Median : 3.000  Median :38.00  Median :46.00
## 8 : 12    4 :1072  Mean   : 3.432  Mean   :38.24  Mean   :45.27
## 9 : 12    5 : 954  3rd Qu.: 6.000  3rd Qu.:42.00  3rd Qu.:56.00
## 10: 12   6 : 845  Max.   :10.000  Max.   :65.00  Max.   :76.00
## (Other):9479 (Other):3205
## male  cardio  reject
## 0:4213  0:7927  0:6314
## 1:5338  1:1624  1:3237
##
## 
## 
## 
## 
```

Take the intercept as random effects

```

library(lme4)
model.inter = lmer(respons ~ time+age+male+cardio+reject + (1|id), data=data)
summary(model.inter)

## Linear mixed model fit by REML ['lmerMod']
## Formula: respons ~ time + age + male + cardio + reject + (1 | id)
##   Data: data
##
## REML criterion at convergence: 58668.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.4981 -0.4752  0.0861  0.5753  6.4273
##
## Random effects:
##   Groups   Name        Variance Std.Dev.
##   id       (Intercept) 11.69    3.419
##   Residual           22.37    4.730
##   Number of obs: 9551, groups: id, 1159
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 33.368890  0.469163 71.124
## time        0.300211  0.017783 16.882
## age         0.058963  0.009071  6.500
## male1       2.340836  0.228917 10.226
## cardio1    -0.253096  0.309629 -0.817
## reject1    -0.356231  0.247265 -1.441
## 
```

```

## Correlation of Fixed Effects:
##          (Intr) time   age    male1  cardio
## time      -0.155
## age       -0.898  0.046
## male1     -0.290  0.015  0.005
## cardio1   0.146 -0.002 -0.282 -0.018
## reject1   -0.346 -0.019  0.189  0.048 -0.030

library(nlme)
model.inter.full = lme(respons ~ time+age+male+cardio+reject, random = ~1|id, data=data.noNA)
summary(model.inter.full)

## Linear mixed-effects model fit by REML
## Data: data.noNA
##      AIC      BIC      logLik
## 58684.23 58741.54 -29334.12
##
## Random effects:
## Formula: ~1 | id
##           (Intercept) Residual
## StdDev:    3.418989 4.729746
##
## Fixed effects: respons ~ time + age + male + cardio + reject
##                 Value Std.Error DF t-value p-value
## (Intercept) 33.36889 0.4691635 8391 71.12422 0.0000
## time        0.30021 0.0177826 8391 16.88234 0.0000
## age         0.05896 0.0090710 1154  6.50022 0.0000
## male1       2.34084 0.2289168 1154 10.22571 0.0000
## cardio1    -0.25310 0.3096294 1154 -0.81742 0.4139
## reject1    -0.35623 0.2472655 1154 -1.44068 0.1499
##
## Correlation:
##          (Intr) time   age    male1  cardio
## time      -0.155
## age       -0.898  0.046
## male1     -0.290  0.015  0.005
## cardio1   0.146 -0.002 -0.282 -0.018
## reject1   -0.346 -0.019  0.189  0.048 -0.030
##
## Standardized Within-Group Residuals:
##      Min        Q1        Med        Q3        Max
## -5.49806141 -0.47524100  0.08607945  0.57529958  6.42734540
##
## Number of Observations: 9551
## Number of Groups: 1159

```

Take

Besides we can think which term to consider the fixed effect and random effect Intercept + slope 1st order term + slope 2nd order term

```

#lme
#data = trenal.long
#lme <- lme(repsons ~ time + age ,data=data)
#lme<-lme(respons~time+age+male+reject+cardio,data=data)
#summary(lme)

```

```
#newdata<-data.frame(ID=c(1,2,3,4,5),week=c(3,3,3,3,3))
#newdata$prediction<-predict(lm,newdata=newdata)
#newdata
#predict(lme,newdata=newdata,level=0:1)
```