LongitudinalDataAnalysis

Group2: Wanchang Zhang; Hugo Blain; Oscar Cabanelas

2023-03-05

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   :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   :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] 13920    20
```

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 meassurements in the 12 time point $(0, 0.5, 1, 2, \cdots, 10)$ 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 $0, 0.5, 1, 2, 3, 4, 5, \cdots, 10$ is one-to-one correspondent to j $1, 2, 3, \cdots, 12$. But we can still leave it in the dataframe. Our response variable is the HC level(The percentage of red cells in the blood, normal levels of hermatocrit 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, respons, 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 %>%
  relocate(id) %>%
  relocate(j,.after=id)%>%
  relocate(time,.after = j)%>%
  relocate(respons,.after=time)
```

```
## # A tibble: 13,920 x 8
##       id     j  time respons   age  male cardio reject
##    <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl>  <dbl>  <dbl>
##  1     1     1   0        26    25     1      0      1
##  2     1     2   0.5      41    25     1      0      1
##  3     1     3   1        42    25     1      0      1
##  4     1     4   2        44    25     1      0      1
##  5     1     5   3        44    25     1      0      1
##  6     1     6   4        45    25     1      0      1
##  7     1     7   5        43    25     1      0      1
##  8     1     8   6        42    25     1      0      1
##  9     1     9   7        39    25     1      0      1
## 10     1    10   8        NA    25     1      0      1
## # ... with 13,910 more rows
```

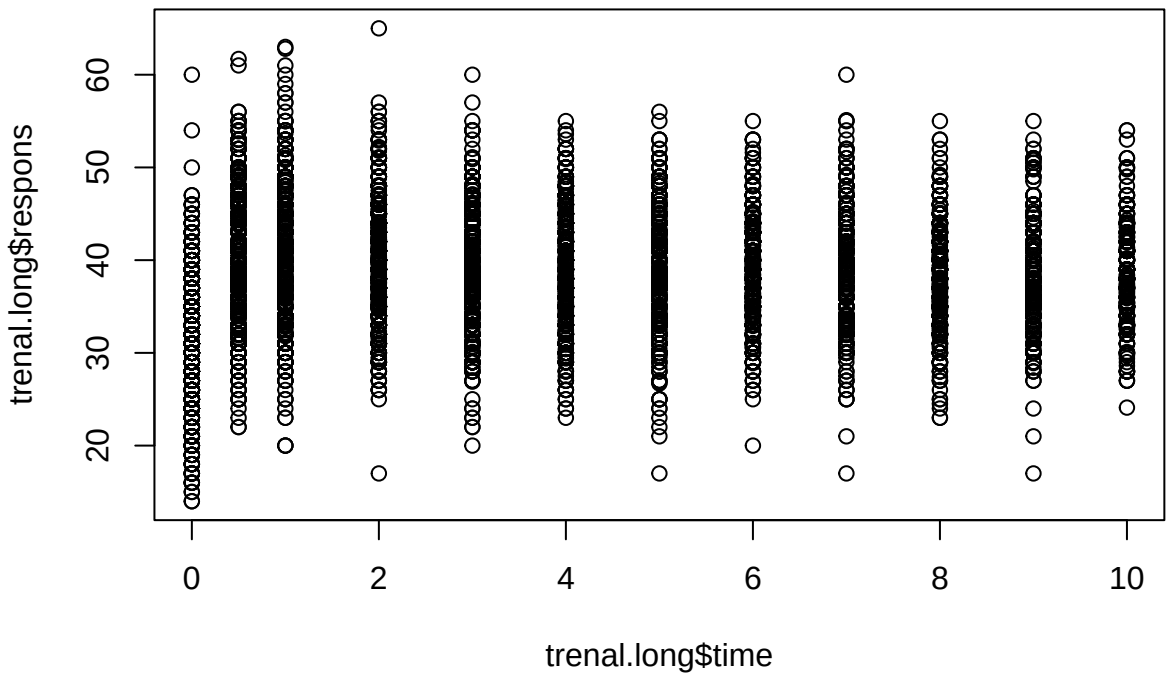
```
trenal.long$id = as.factor(trenal.long$id)
trenal.long$j = as.factor(trenal.long$j)
trenal.long$male = as.factor(trenal.long$male)
trenal.long$cardio = as.factor(trenal.long$cardio)
trenal.long$reject = as.factor(trenal.long$reject)
summary(trenal.long)
```

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

```
length(unique(trenal.long$id))
```

```
## [1] 1160
```

```
# Plot the raw data
plot(trenal.long$time,trenal.long$respons)
```




```
library(ggplot2)
library(nlme)
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
library(lme4)
```

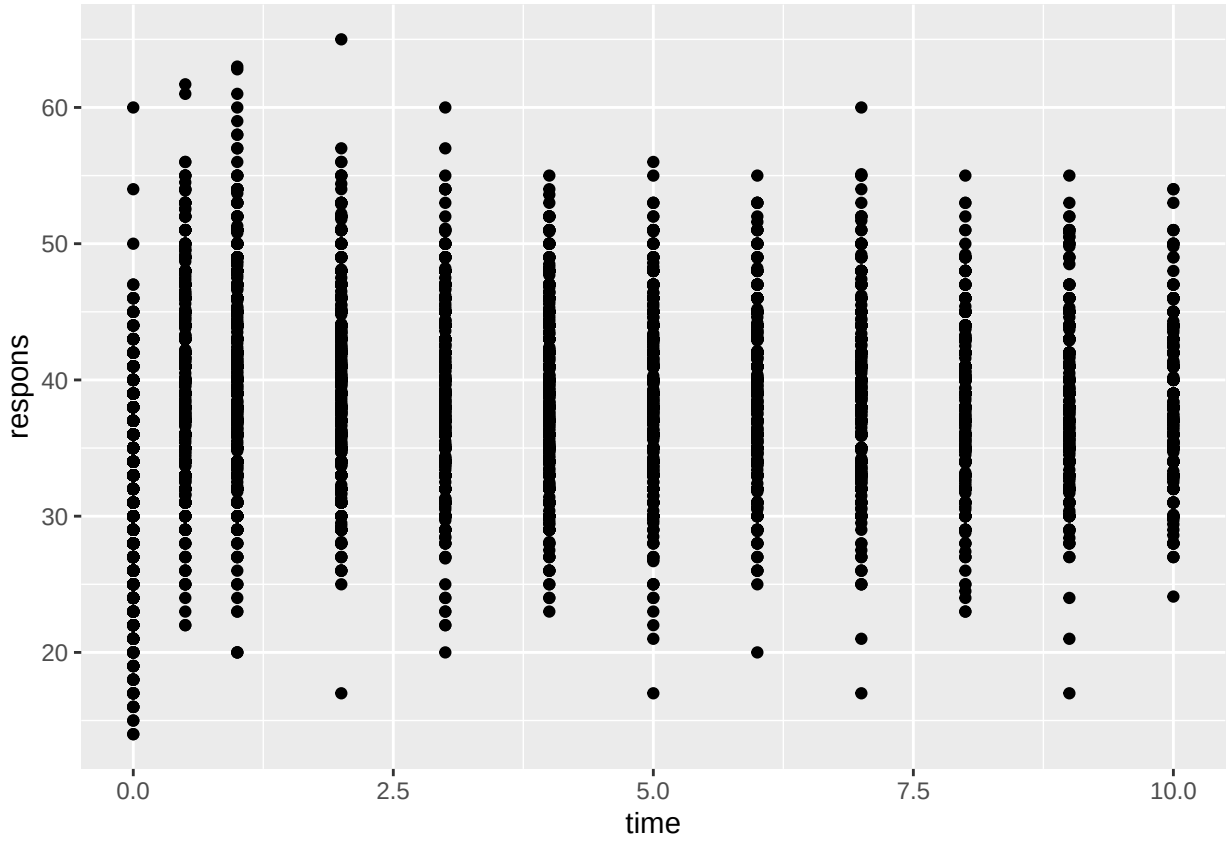
```
## Loading required package: Matrix
```

```
##
## Attaching package: 'lme4'
```

```
## The following object is masked from 'package:nlme':
##
##     lmList
```

```
data = trenal.long
#Plot data
ggplot(data, aes(x=time, y=respons)) + geom_point()
```

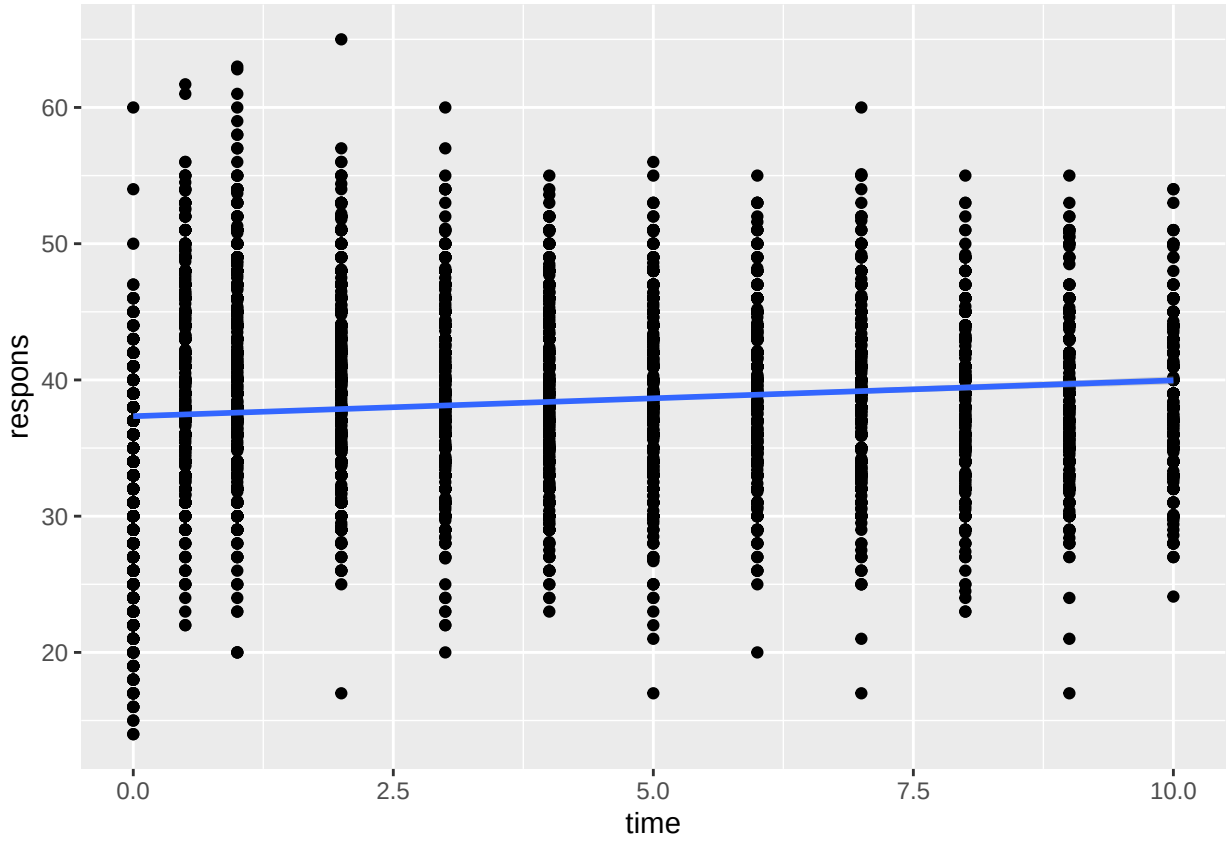
```
## Warning: Removed 4362 rows containing missing values (`geom_point()`).
```




```
#Plot data with lm line
ggplot(data, aes(x=time, y=respons)) + 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()`).
```

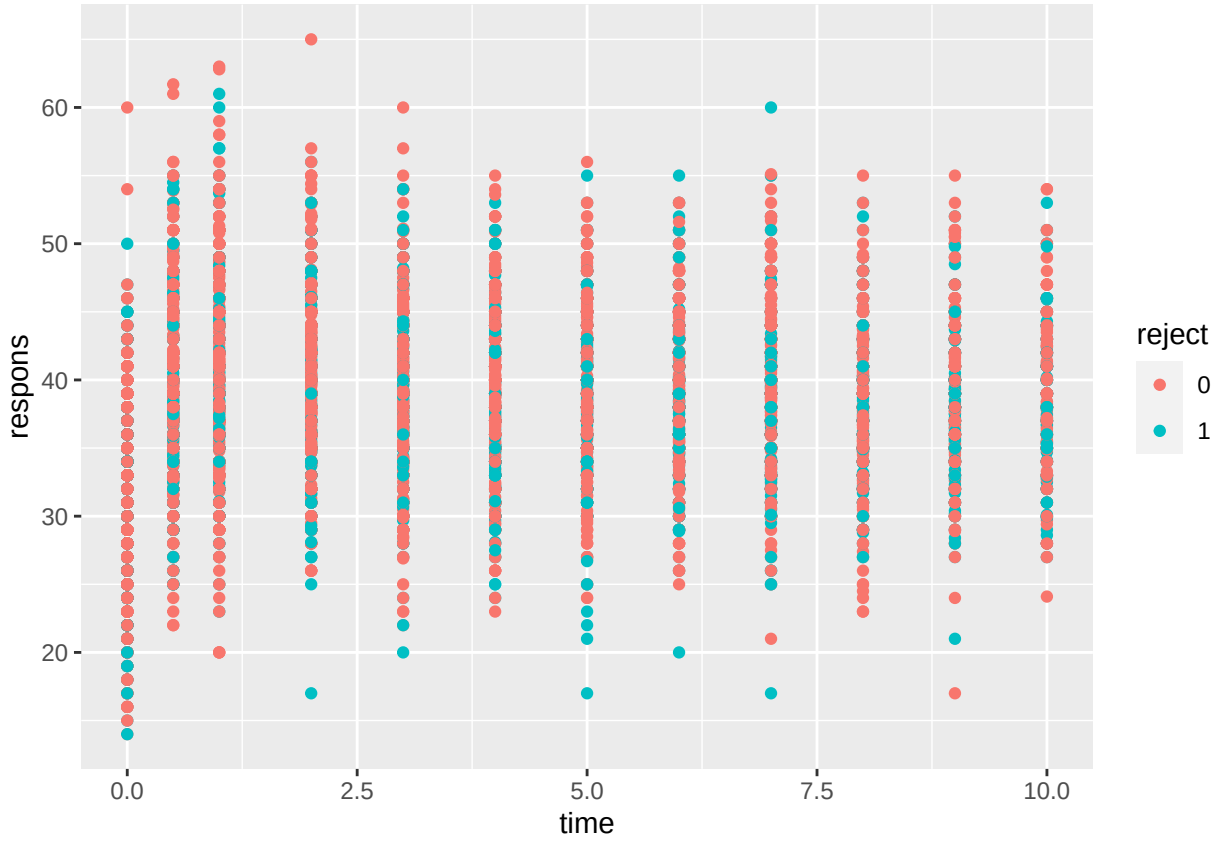


1.3 List of Hypotheses to be tested by the data

Hypothese one HC level will change with time differently if the REJECT is different

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

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

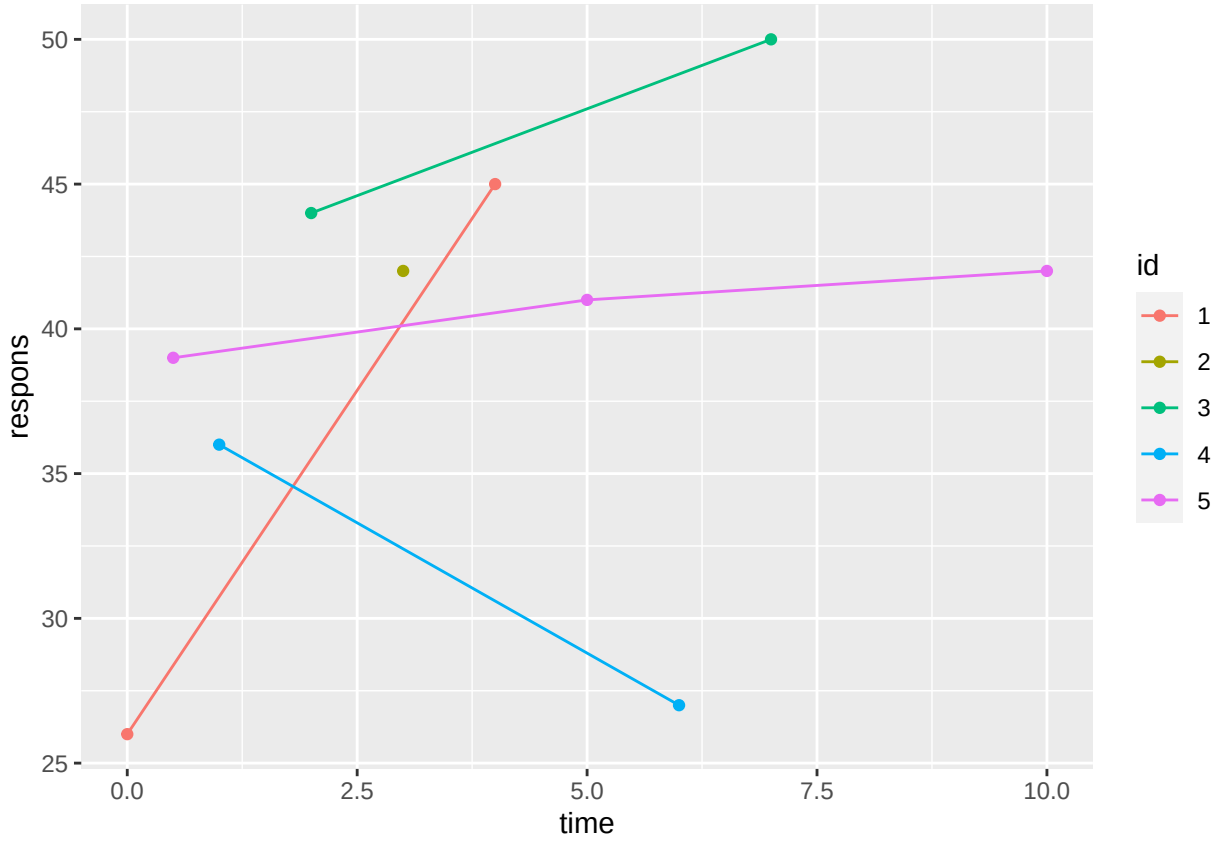



```
#Plot individual data

ggplot(data[data$id == c('1','2','3','4','5'),], aes(x=time, y=respons, group=id,color=id)) + geom_poin
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```

```
## Warning: Removed 2 rows containing missing values (`geom_line()`).
```

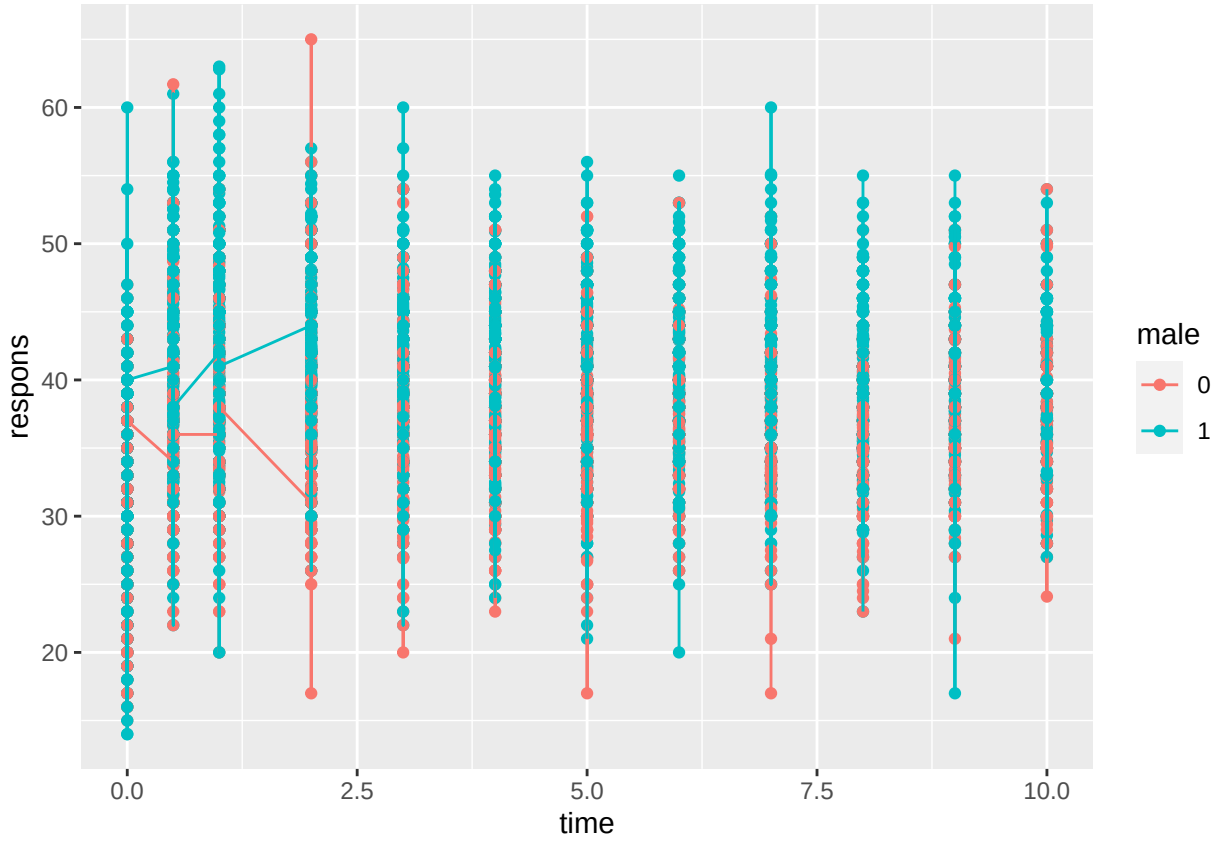


Hypothese two HC level will change with time differently if the sex is different, male has generally higher HC level than female

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

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

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

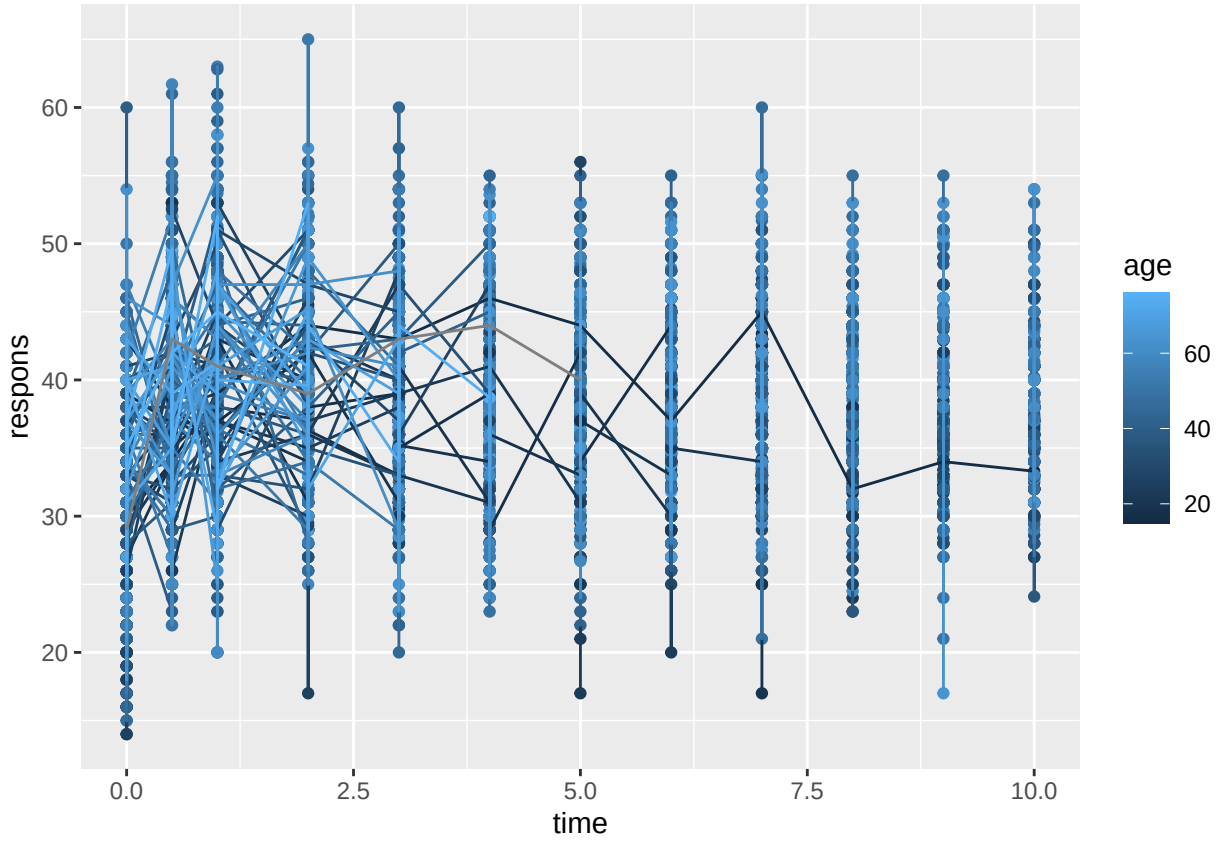


Hypothese three HC level will change with time differently if the age when performing the kidney transplant is younger

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

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

```
## Warning: Removed 808 rows containing missing values (`geom_line()`).
```

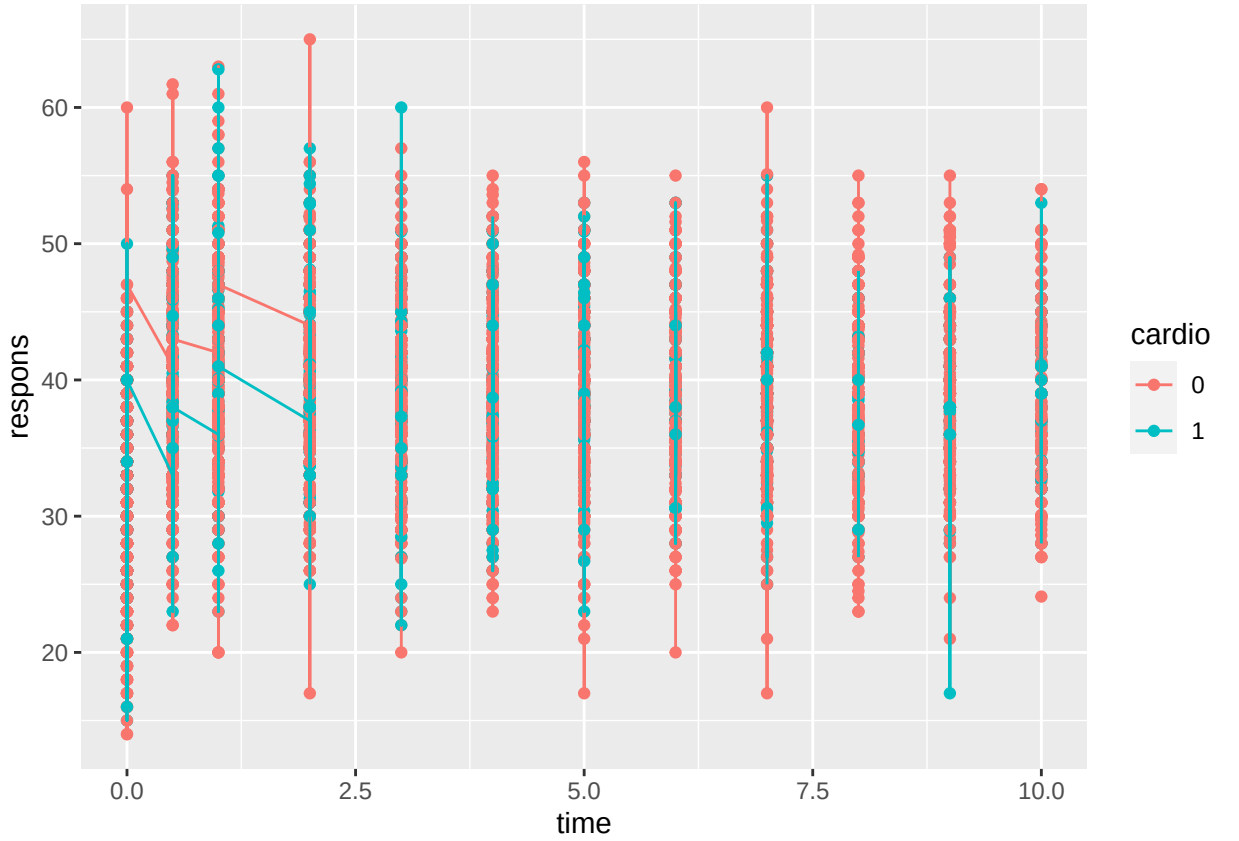


Hypothese four HC level will change with time differently if the patient has experienced cardio-vascular problem during the years preceding the tranplantation

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

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

```
## Warning: Removed 641 rows containing missing values (`geom_line()`).
```

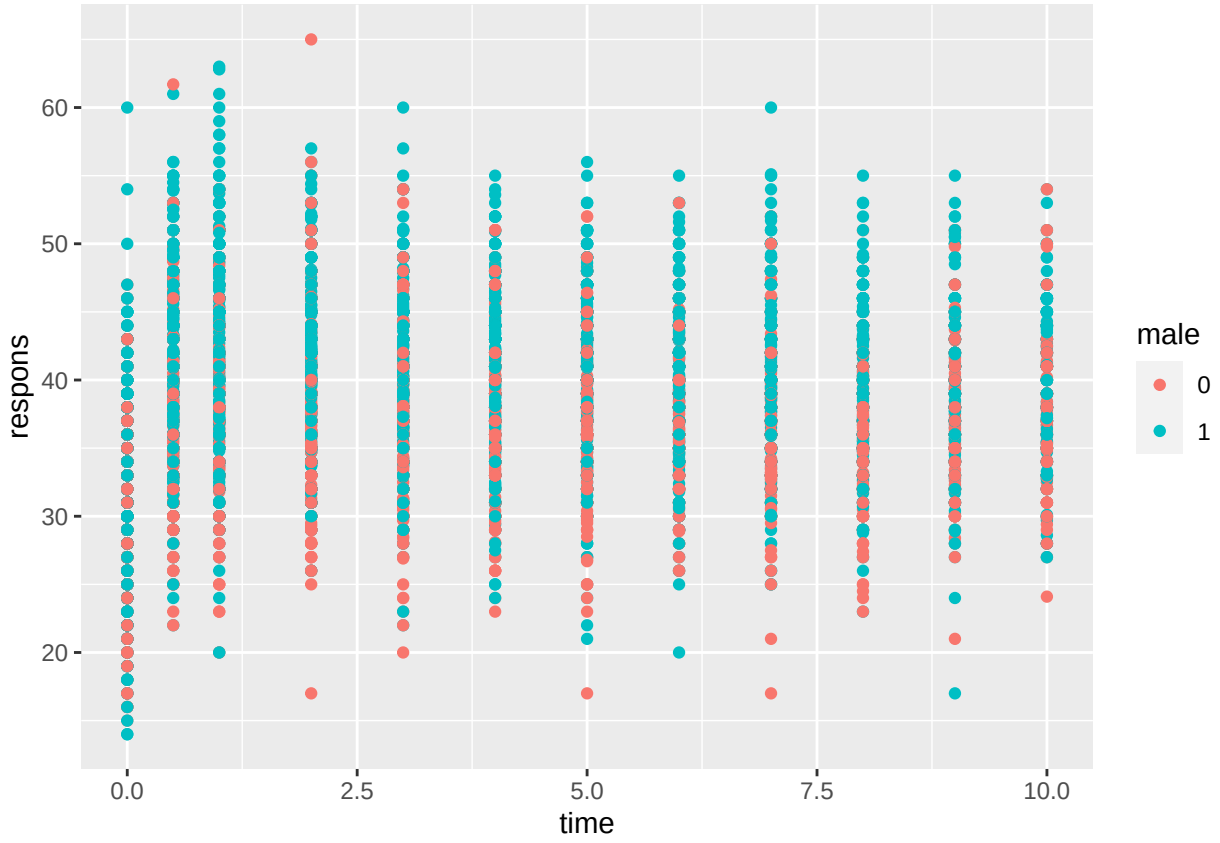



```
#Lm
lm<-lm(respons~time,data=data)
summary(lm)
```

```
##
## 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
```

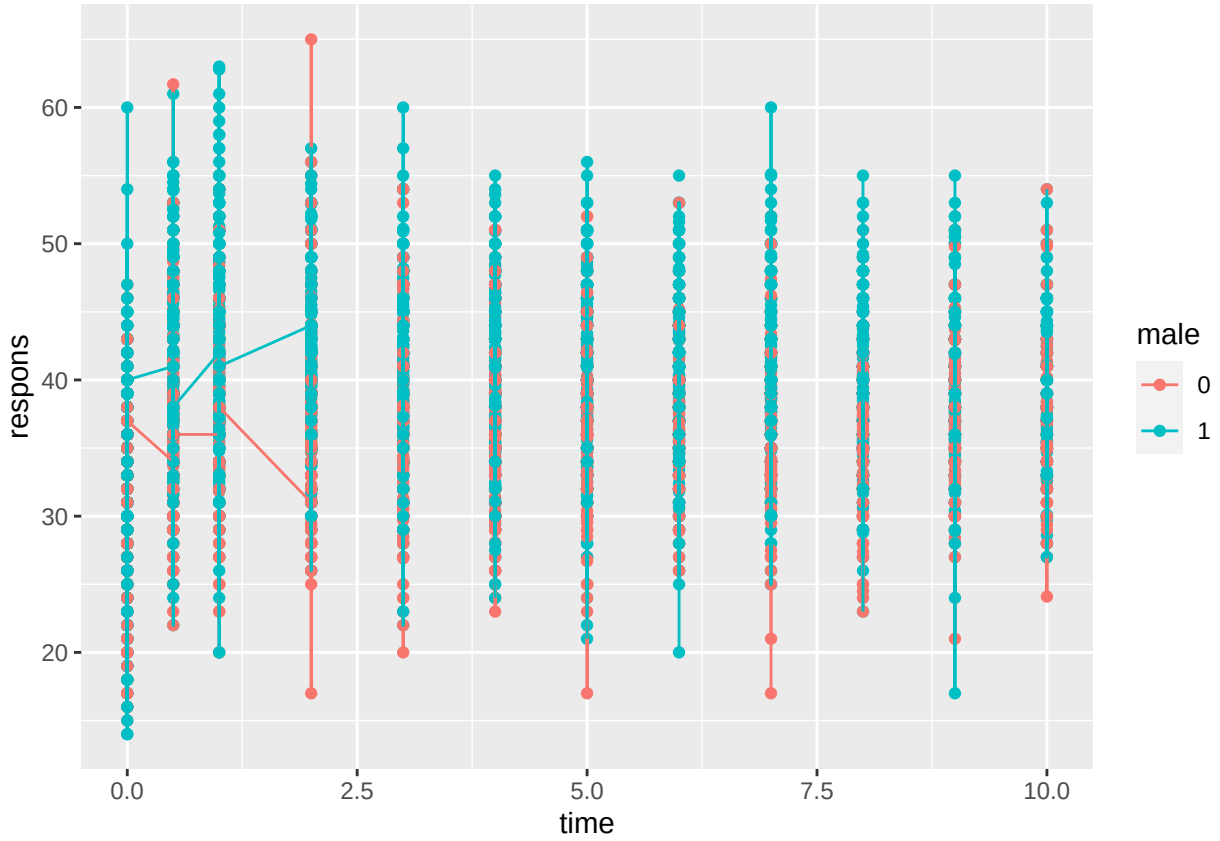
```
#Plot individual data
ggplot(data, aes(x=time, y=respons, group=male,color=male)) + geom_point()
```

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




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

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



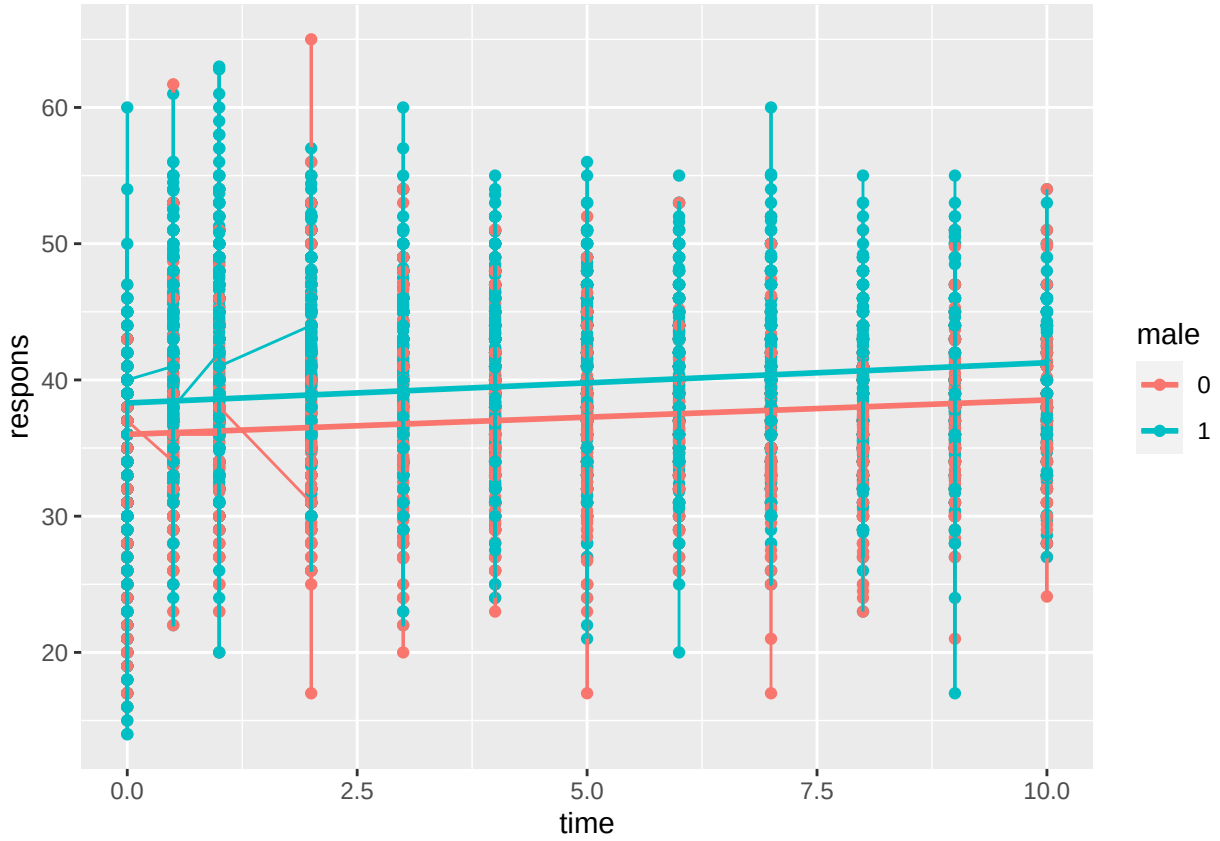

```
#Spaghetti with fitted lines
ggplot(data, aes(x=time, y=respons, group=male,color=male)) + geom_point()+ geom_smooth(method="lm",se=
```

```
## `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 Fixed effect could be time, sex, age, reject, cardio Random effect could be

```
#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)
```