Longitudinal Data Analysis

Case study of Trenal.XLS using Linear Mixed Effect Model

Group2: Wanchang Zhang; Hugo Blain; Oscar Cabanelas

2023-03-24

Contents

List of Figures

List of Tables

1. Describe the data, and use graphical techniques to explore the mean structure, the variance structure and the correlation structure. Summarize your conclusions. What are the implications with respect to statistical modeling?

Conclusion:

The response variable Hematocrit (the percentage of red cells in your blood level) varies a lot on different subjects, but generally it shows that the time when just going through the kidney transplant, the HC level is quite low around 31.86%, it could goes back to normal half year later and remains at the similar level for the rest of the 10 years observations.

The correlation structure shows that HC level are more correlated to the consecutive HC levels, so we may use a autoregression correlation structure in the statistical modeling. The Correlation structure suitable could be Autoregressive, time dependent or unstructured

- Autoregressiove (ρ)

$$R = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{12} \\ \rho & 1 & \rho^2 & \cdots & \rho^{11} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{12} & \rho^{11} & \rho^{10} & \cdots & 1 \end{bmatrix}$$

- Time dependent (ρ)

$$R = \begin{bmatrix} 1 & \rho^{t_{1,2}} & \rho^{1,3} & \cdots & \rho^{1,12} \\ \rho^{t_{2,1}} & 1 & \rho^{t_{2,3}} & \cdots & \rho^{t_{2,11}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{t_{12,1}} & \rho^{t_{12,2}} & \rho^{t_{12,3}} & \cdots & 1 \end{bmatrix}$$

- Unstructured ($\rho_{1,2}, \rho_{1,3}, \cdots, \rho_{1,12}, \cdots \rho_{11,12}$ totally $\frac{(11+1)\times 11}{2} = 66$ parameters)

$$R = \begin{bmatrix} 1 & \rho_{1,2} & \rho_{1,3} & \cdots & \rho_{1,12} \\ \rho_{2,1} & 1 & \rho_{2,3} & \cdots & \rho_{2,11} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{12,1} & \rho_{12,2} & \rho_{12,3} & \cdots & 1 \end{bmatrix}$$

2. What summary statistics are appropriate for the analysis of these data? Why? Do they yield the same results? Summarize your conclusions. Summary statistics: summarize and provide information about the sample data. It tells you about the values in your data set. This includes where the mean lies and whether your data is skewed. Summary statistics have three main categories from https://www.statisticshowto.com/summary-statistics/ 1). Measures of location(central tendency)

2). Measures of spread(range, interquartile range, quartiles, skewed, Kurtosis)

this two can be answered from the `summary(trenal.wide)`

- The input data:
  - id: totall 1160 persons
  - age to perform the operation: from 15 to 76 years old, average is 46.43 years old
  - male: we observe 494 females and 666 males
  - cardio: 953 persons has experienced a cardio-vascular problem during the years preceding the transplant, 207 did not.
  - reject: 793 patients shown symptoms of graft rejection during the first three months after the transportation, 367 has not.
- The response variable HC level continous from min 14% to max 65%. It is complex to analyse without considering different persons with different characters. The HC level is dependent on the meassured time, individual's age to perform the operation, gender, cardio history and reject history

Thus we plot separatly to see if the age, gender, cardio and reject would influence the HC level change with time. 3). Graphs/charts(Histogram, Frequency Distribution Table, Box plot, Bar chart, Scatter plot, Pie char.t)

3. Fit a multivariate model and find the most parsimonious mean structure which can be used to describe the average evolutions in the data. What covariance structures are applicable in this case? What is the most parsimonious structure you can find?
4. Use an explicit two-stage analysis to get an initial impression about trends and effects of covariates.
5. Formulate a plausible random-effects model. Fit your model and compare the results with those from the multivariate model. Check the appropriateness of your random-effects model. Calculate the subject-specific intercepts/slopes and compare them with the ones you obtained from a two-stages analysis. What do you conclude?
6. Pay attention to the missing, especially the ones presented by the outcome variable. Do your results still hold despite the missingness?

Theory of Linear Mixed Effects Model(LMM)

Index description

Let us assume that a given input data set X has a dimension $N \times p$, with N observations and p predictors.

For each subject indexed with $i, i = 1, \cdots, I$, we can build a linear mixed effect model

$$\mathbf{Y_i} = \mathbf{X_i}\beta + \mathbf{Z_i}\mathbf{b_i} + \epsilon_\mathbf{i}$$

The application of LMM

Linear Mixed Effects Model is used to analyse a data set, where the observations are not fully independent, while the top level clusters are assumed independent. Inside each cluster, the observations are correlated

Data set `Trenal.XLS` pre-analysis

The summary of the data set

Import data

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

Data Preprocessing

```
trenal= trenal[,-18] #remove a noninformative column const

# Continuous or discrete variables
trenal$id = as.factor(trenal$id)
trenal$j = as.factor(trenal$j)
trenal$time = as.factor(trenal$time)
trenal$male = as.factor(trenal$male)
trenal$cardio = as.factor(trenal$cardio)
trenal$reject = as.factor(trenal$reject)
# Change the name of respons
colnames(trenal)[19] <- "HC"
```

```
trenal.long = trenal[,13:20] # long table form
# Remove j
trenal.long = trenal.long[,-6]
trenal.long.noNA <- na.omit(trenal.long)# reordered
trenal.wide = as.data.frame(subset(trenal,trenal$j=="1"))[,1:18]
summary(trenal.wide)
```

```
##       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   :1                       NA's   :1       NA's   :87     NA's   :205
##       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   :314     NA's   :418     NA's   :508     NA's   :595
##       HC8              HC9             HC10              id             age
##  Min.   :23.00   Min.   :17.00   Min.   :24.10   1      :   1   Min.   :15.00
##  1st Qu.:35.00   1st Qu.:35.00   1st Qu.:35.00   2      :   1   1st Qu.:36.00
##  Median :38.05   Median :38.50   Median :38.00   3      :   1   Median :48.00
##  Mean   :38.35   Mean   :38.57   Mean   :38.49   4      :   1   Mean   :46.43
##  3rd Qu.:42.00   3rd Qu.:42.00   3rd Qu.:42.00   5      :   1   3rd Qu.:57.00
##  Max.   :55.00   Max.   :55.00   Max.   :54.00   6      :   1   Max.   :76.00
##  NA's   :672     NA's   :749     NA's   :812     (Other):1154   NA's   :1
##  male    cardio  reject        j
##  0:494   0:953   0:793   1      :1160
##  1:666   1:207   1:367   2      :   0
##                          3      :   0
##                          4      :   0
##                          5      :   0
##                          6      :   0
##                          (Other):   0
```

```
library(magrittr) # needs to be run every time you start R and want to use %>%
library(dplyr)

data.long <- trenal.long %>% # reordered long table
  relocate(id) %>%
  relocate(time,.after = id)%>%
  relocate(HC,.after=time)
#summary(data.long)
sum(!is.na(data.long$HC))
```

```
## [1] 9558
```

```
data.long.noNA <- na.omit(data.long)# reordered long table without NAs
summary(data.long.noNA)
```

```
##        id            time            HC              age        male
##  3      :  12   0.5    :1159   Min.   :14.00   Min.   :15.00   0:4213
##  5      :  12   0      :1158   1st Qu.:34.00   1st Qu.:35.00   1:5338
##  6      :  12   1      :1158   Median :38.00   Median :46.00
##  8      :  12   2      :1072   Mean   :38.24   Mean   :45.27
##  9      :  12   3      : 954   3rd Qu.:42.00   3rd Qu.:56.00
##  10     :  12   4      : 845   Max.   :65.00   Max.   :76.00
##  (Other):9479   (Other):3205
##  cardio   reject
##  0:7927   0:6314
##  1:1624   1:3237
##
##
##
##
##
```

```
data_summary_long = data.frame(unclass(summary(data.long.noNA,maxsum=1160)),check.names=FALSE)
data.long.noNA$id[length(data.long.noNA$id)]
```

```
## [1] 1160
## 1160 Levels: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 ... 1160
```

Response variable and predictors

Response variable From the `summary(data.long.noNA)`, we can read that the response variable is a continuous variable `HC`level from $(15, 76)$ with Mean 38.24. (Hematocrit is the percentage of red cells in your blood %)

We have totally $I = 1160$ ids for subjects. Ideally each id would have 12 (start from $HC_0, HC_{0.5}, HC_1, HC_2, \cdots, HC_{10}$) HC level measurement, but in really not all subjects have all of the 12 measurements (Actually only 348 patients all 12 measurements). We have totally $N = 9558 = \sum_{i=1}^{I} n_i$ missing values.

Predictors The explaining variables are

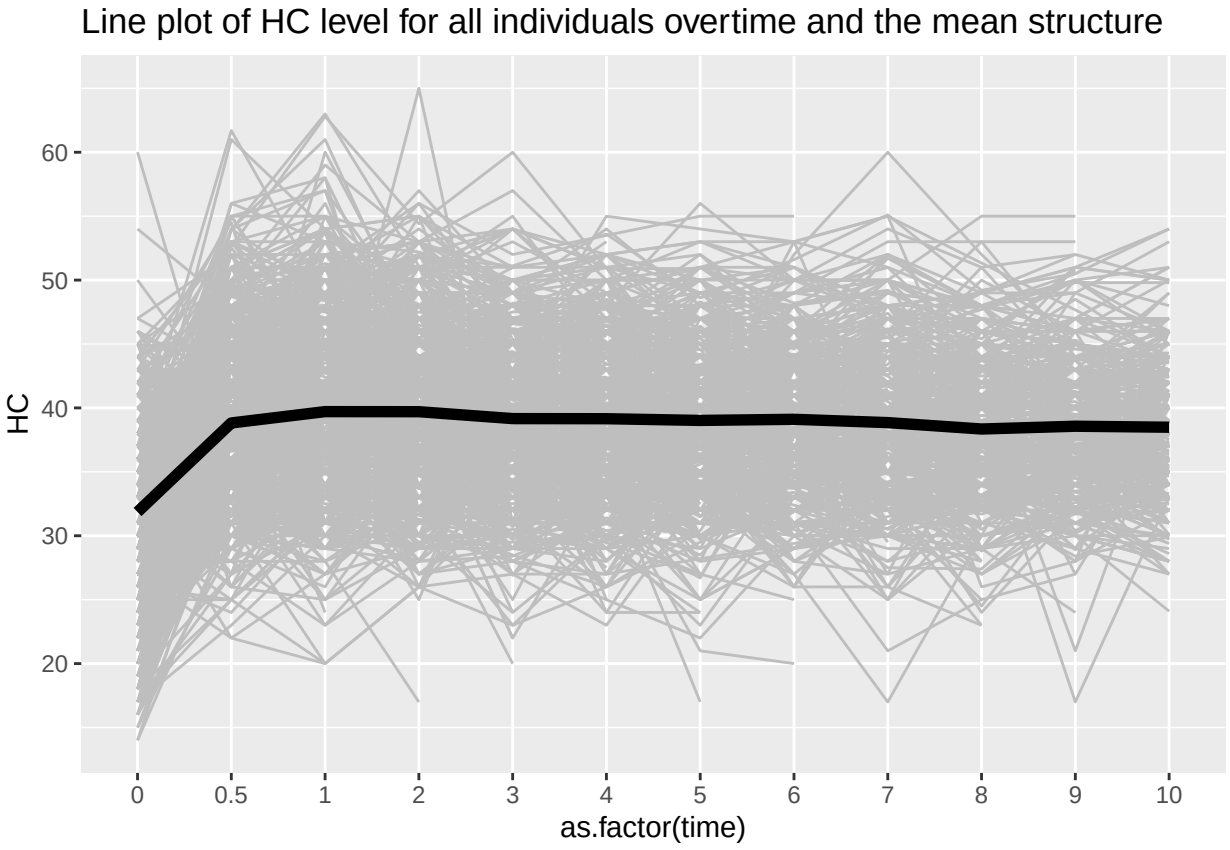
1. $X_1 = time$ in year as discrete values, only changes with j,$j = 1, \cdots, n_i$
2. $X_2 = age$ in year with 12 NAs; will only change with subject id i
3. $X_3 = male$ 0 = female, 1= male;will only change with subject id i
4. $X_4 = cardio$ 0 = no, 1= yes;will only change with subject id i
5. $X_5 = reject$ 0 = accept, 1= reject;will only change with subject id i
6. fixed intercept, continuous

In the data analysis part, we need to try out different variables accounting for fixed effects and random effects.

Data visualization and the information from the data

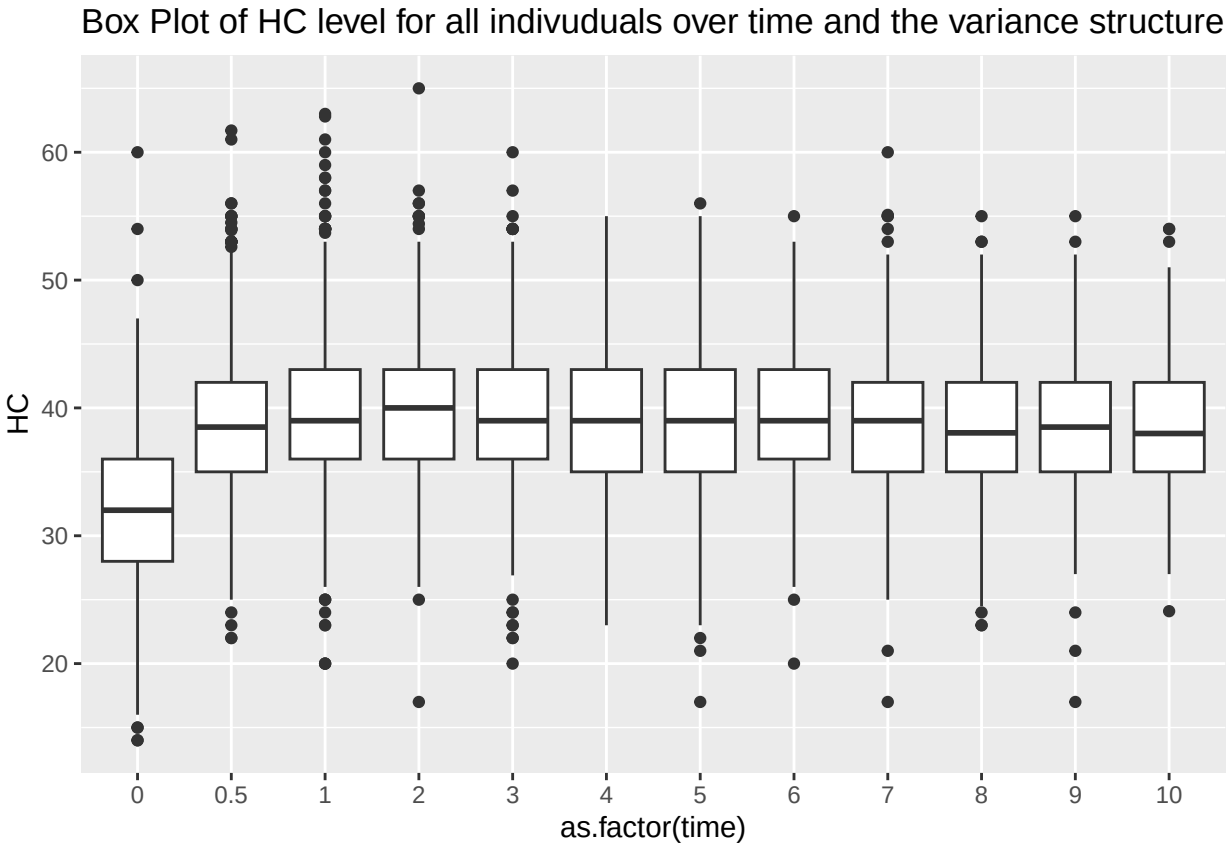
Mean Structure

```
library(ggplot2)
# To view the mean structure of the HC for all individuals
ggplot(data.long.noNA,aes(x=as.factor(time),y=HC,group=id))  + geom_line(col="grey")+stat_summary(aes(g
  labs(title="Line plot of HC level for all individuals overtime and the mean structure")
```



Variance Structure

```
# To view to variance structure
ggplot(data.long.noNA,aes(x=as.factor(time),y=HC))+
  geom_boxplot(position=position_dodge(1))+
  labs(title="Box Plot of HC level for all indivuduals over time and the variance structure")
```



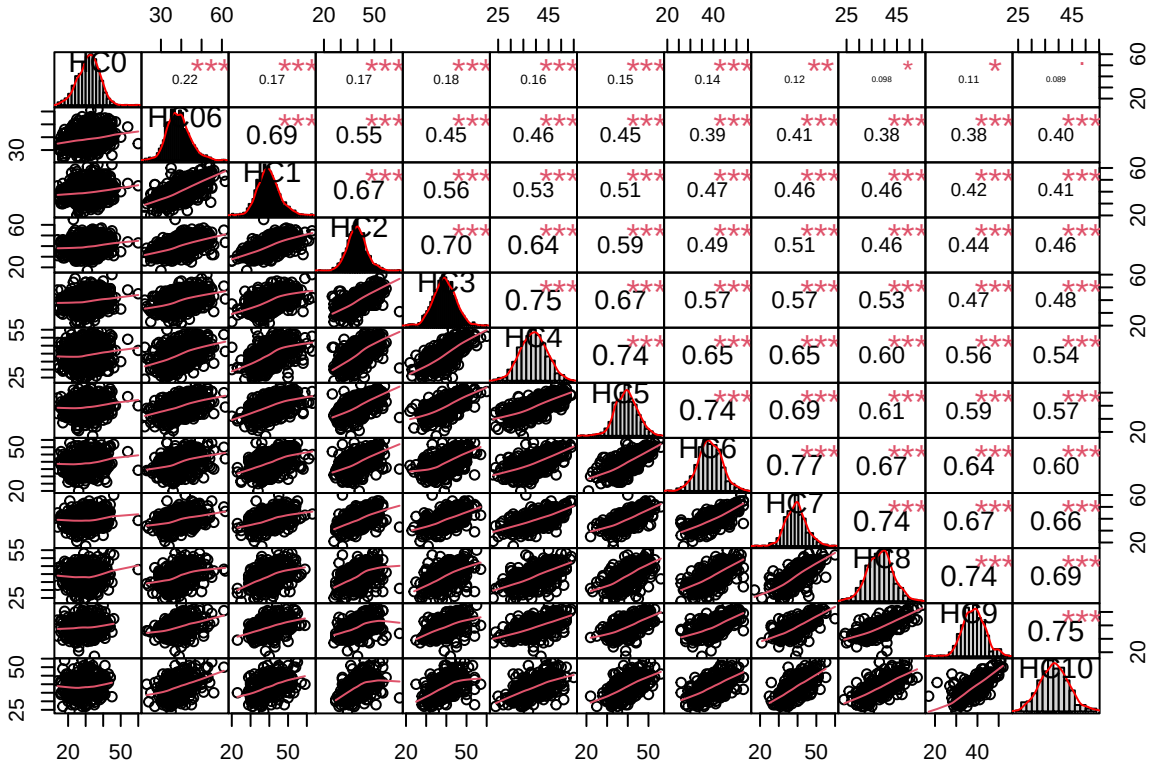
Covariance Structure

```
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")
chart.Correlation(HcCorr,historgram=TRUE)
```

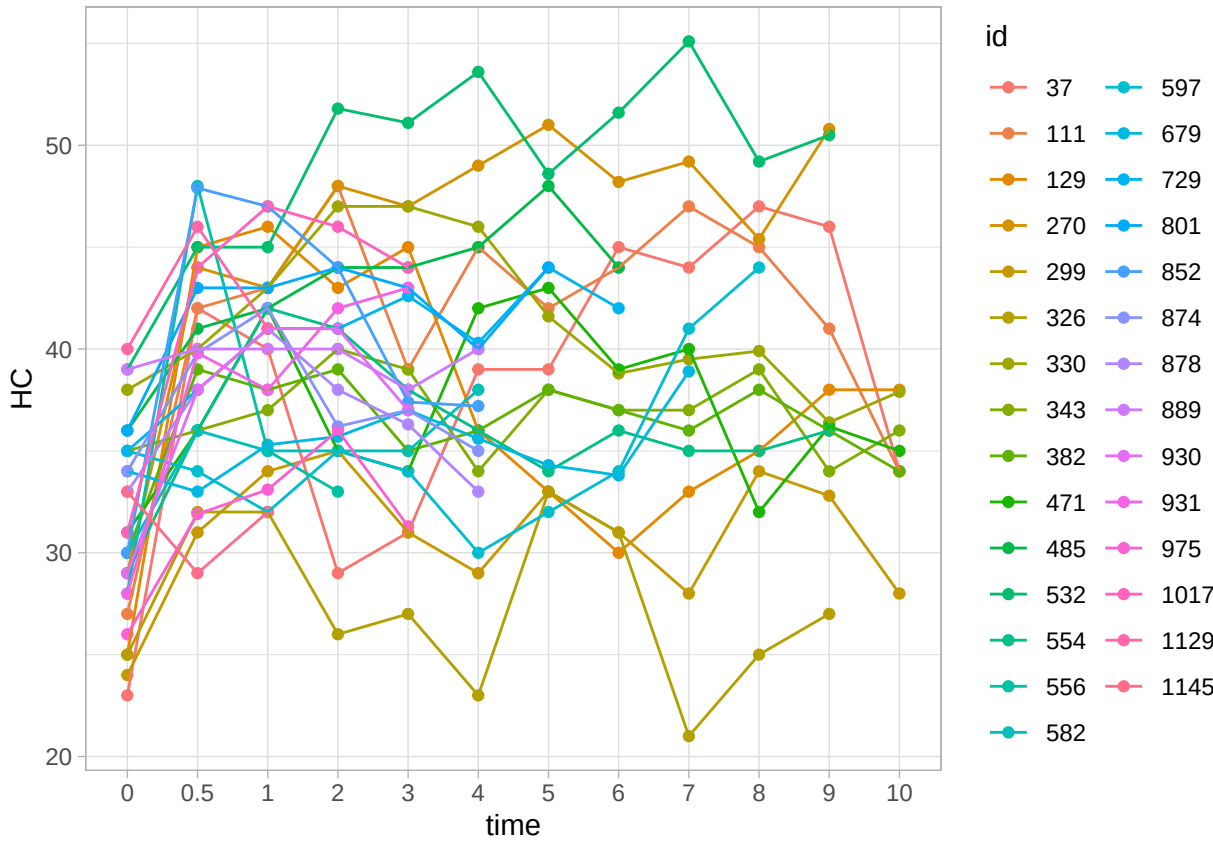



```
dim(data.long.noNA)
```

```
## [1] 9551    7
```

```
# since the data dimension is large 9551 x 8, we can select random 30 data to have a look
set.seed(1)
selected <- sample(1:length(unique(data.long.noNA$id)),30,replace=T) # random samples and permutations
#selected.vector = as.vector(selected)
data.selected = data.long.noNA[(data.long.noNA$id %in% c(selected)), ]
```

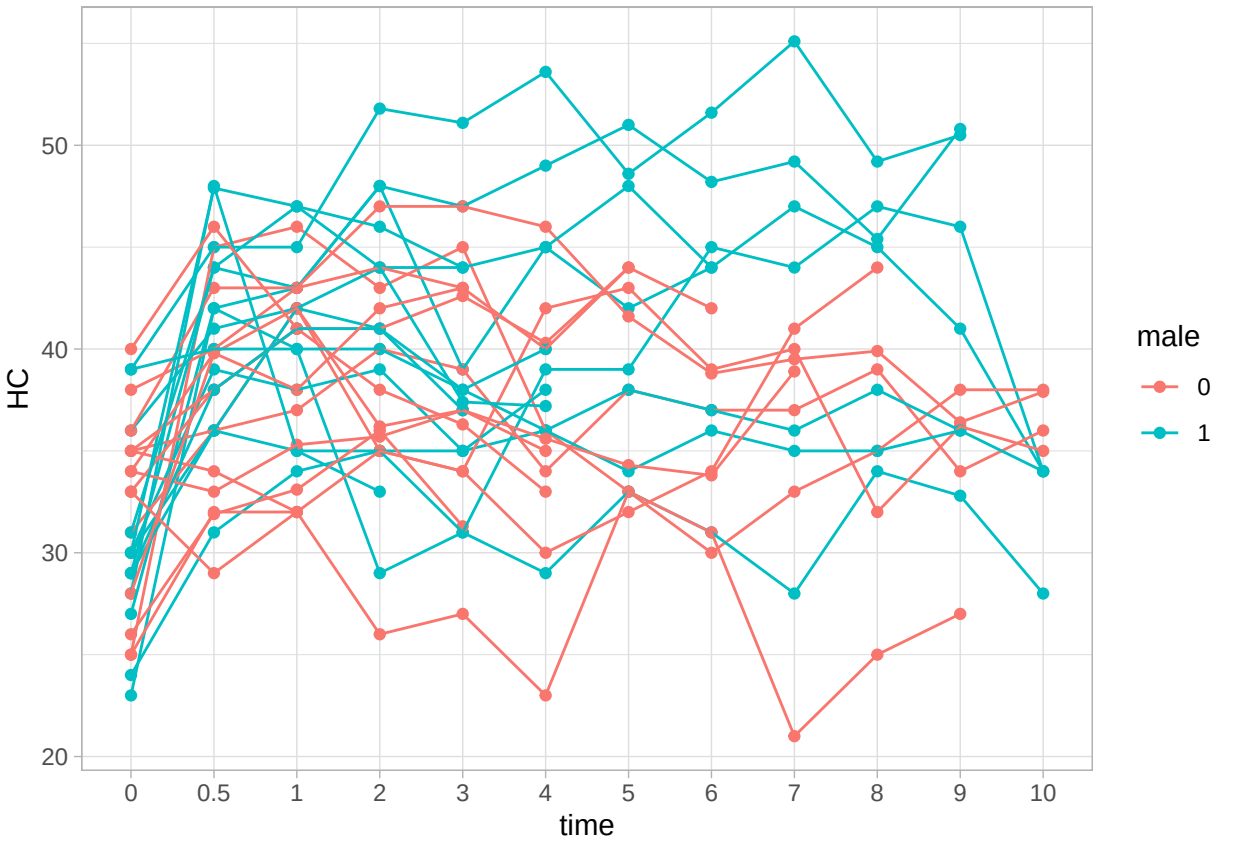
```
ggplot(data.selected,aes(x=time,y=HC,group=id,color=id))+geom_point()+ geom_line()+theme_light()
```



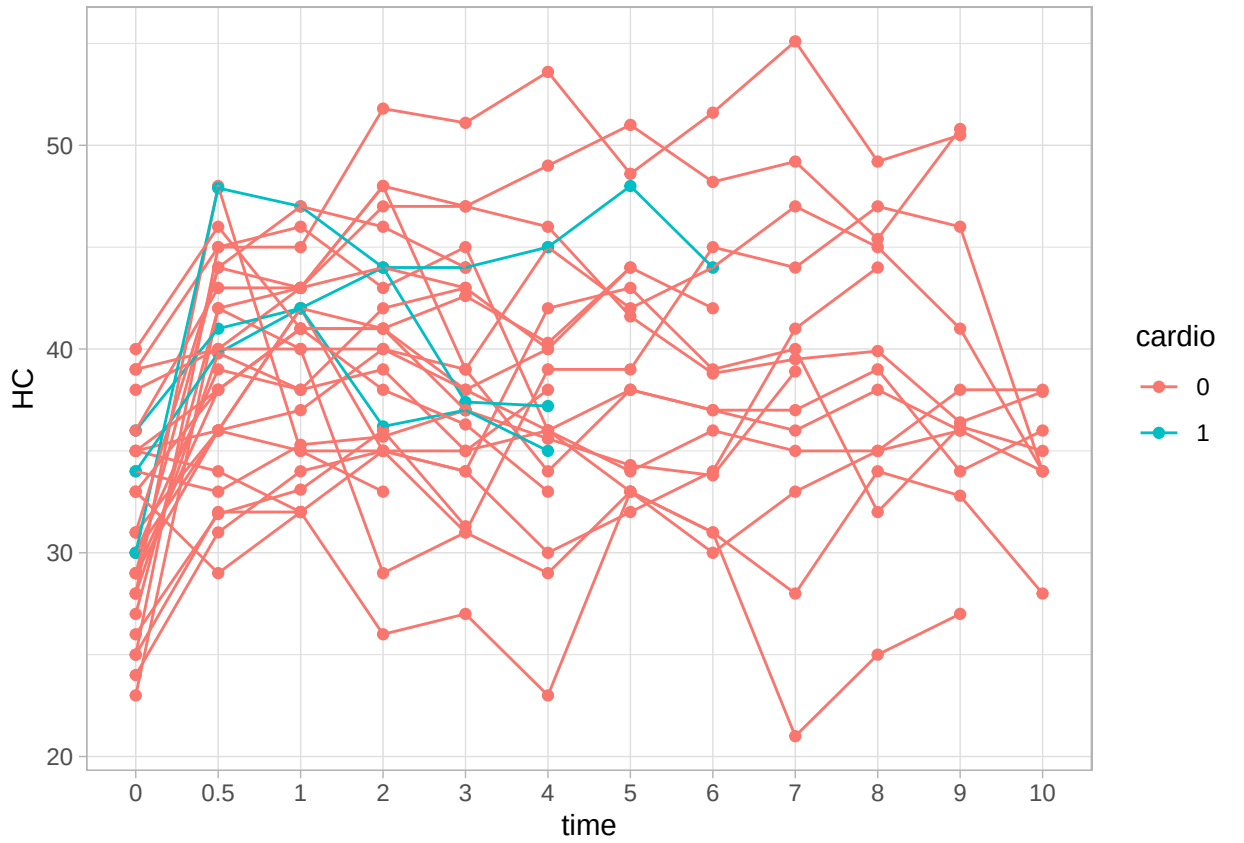
Spaghettic plot

Plot Info: The intercept may vary according to each individual The slope is not very easy to see

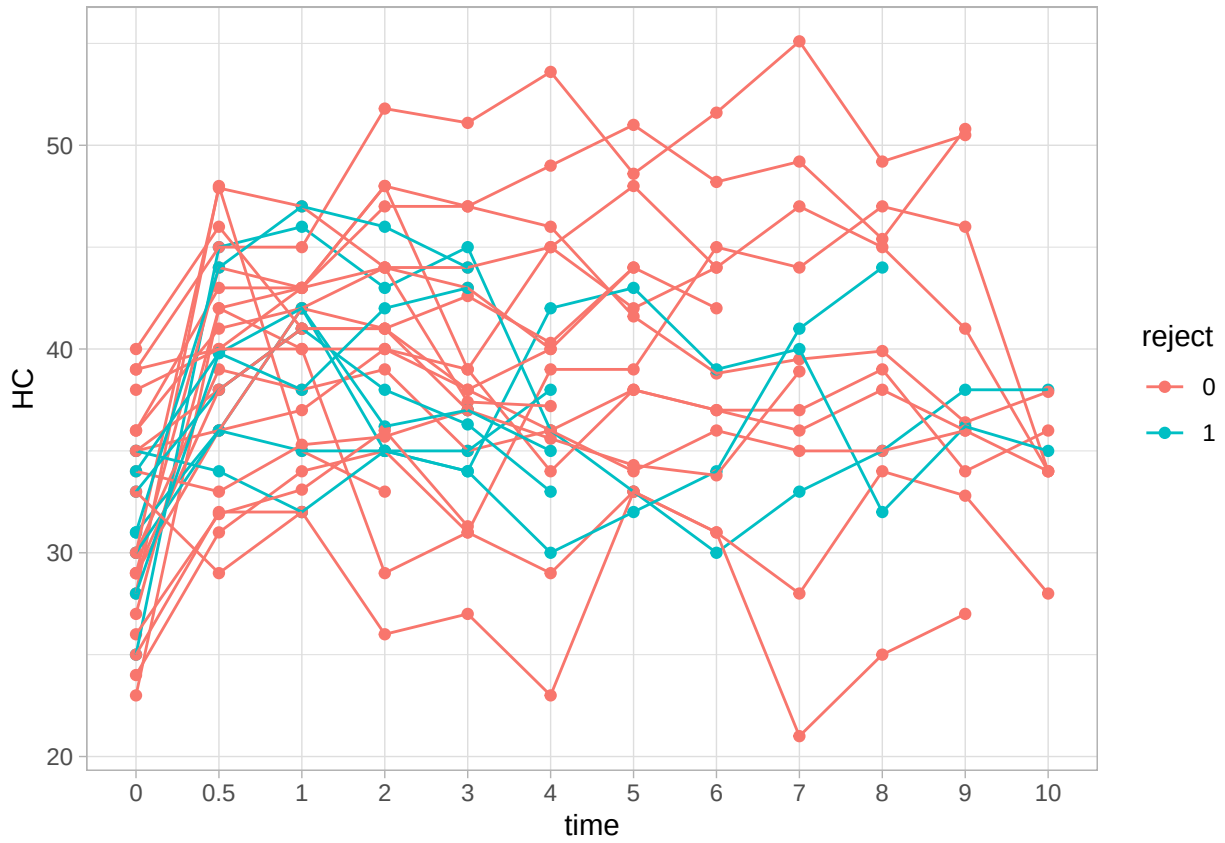
```
ggplot(data.selected,aes(x=time,y=HC,group=id,color=male)) +geom_point()+ geom_line()+theme_light()
```



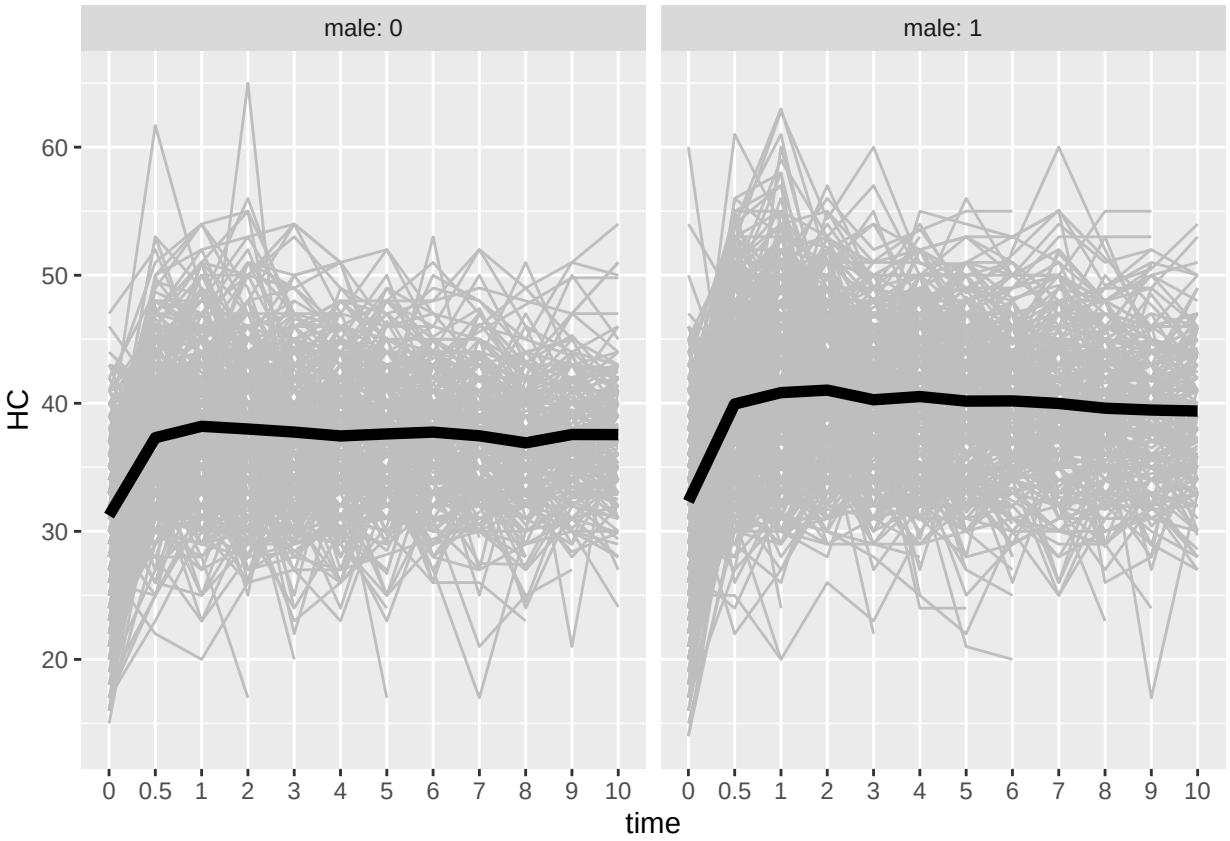

```
ggplot(data.selected,aes(x=time,y=HC,group=id,color=cardio)) +geom_point()+ geom_line()+theme_light()
```



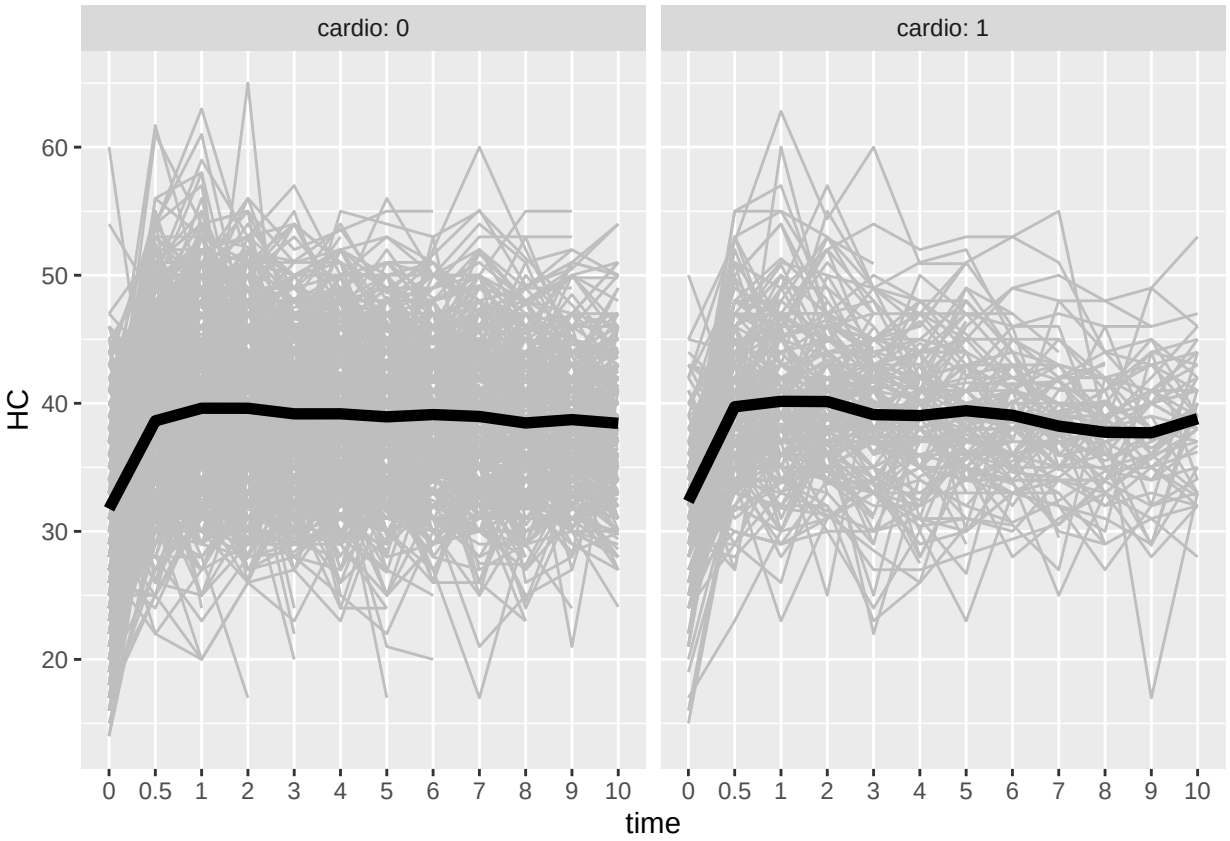

```
ggplot(data.selected,aes(x=time,y=HC,group=id,color=reject)) +geom_point()+ geom_line()+theme_light()
```



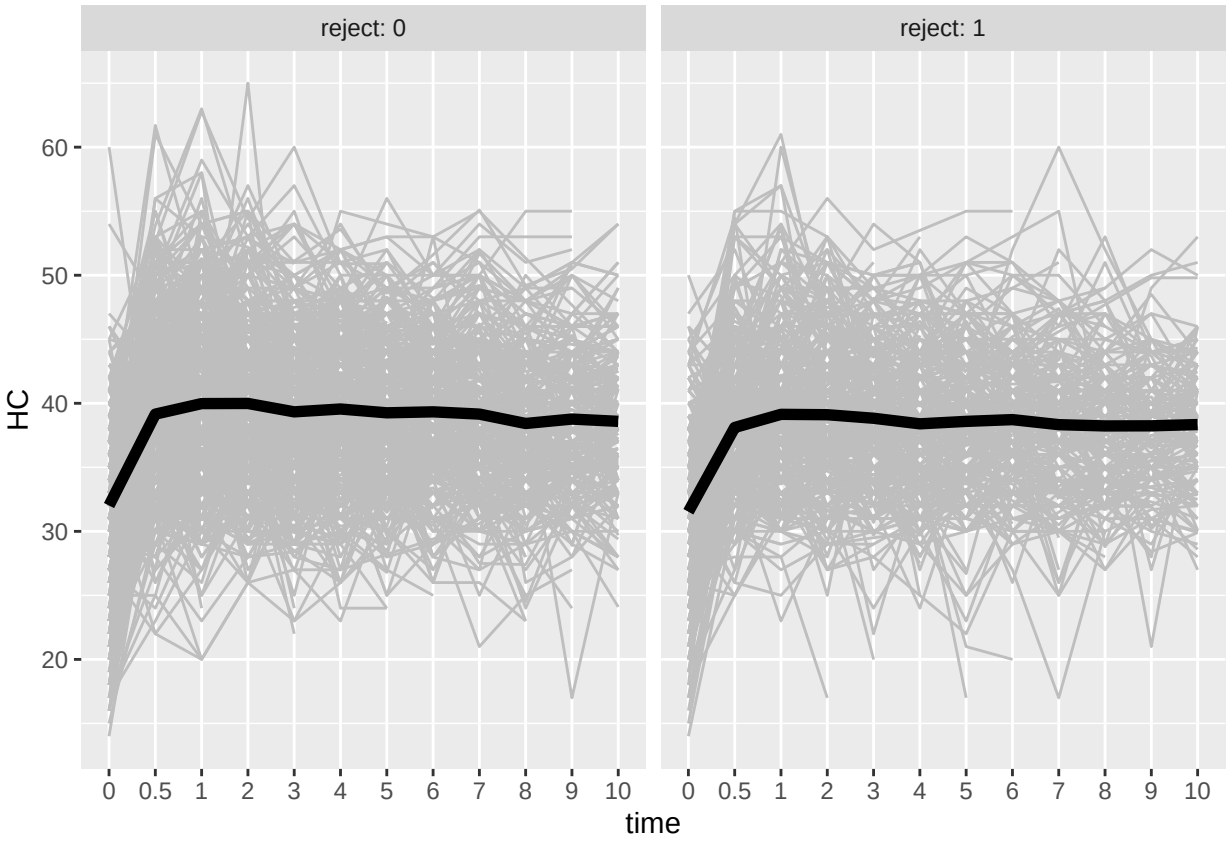

```
# Spaghetti Ggplot separated by male =1
p <- ggplot(data=data.long.noNA,aes(x=time,y=HC,group=id))
p <- p + geom_line(col="grey")+stat_summary(aes(group=1),geom="line",fun=mean,linewidth=2)
p + facet_grid(~male,labeller=label_both)
```



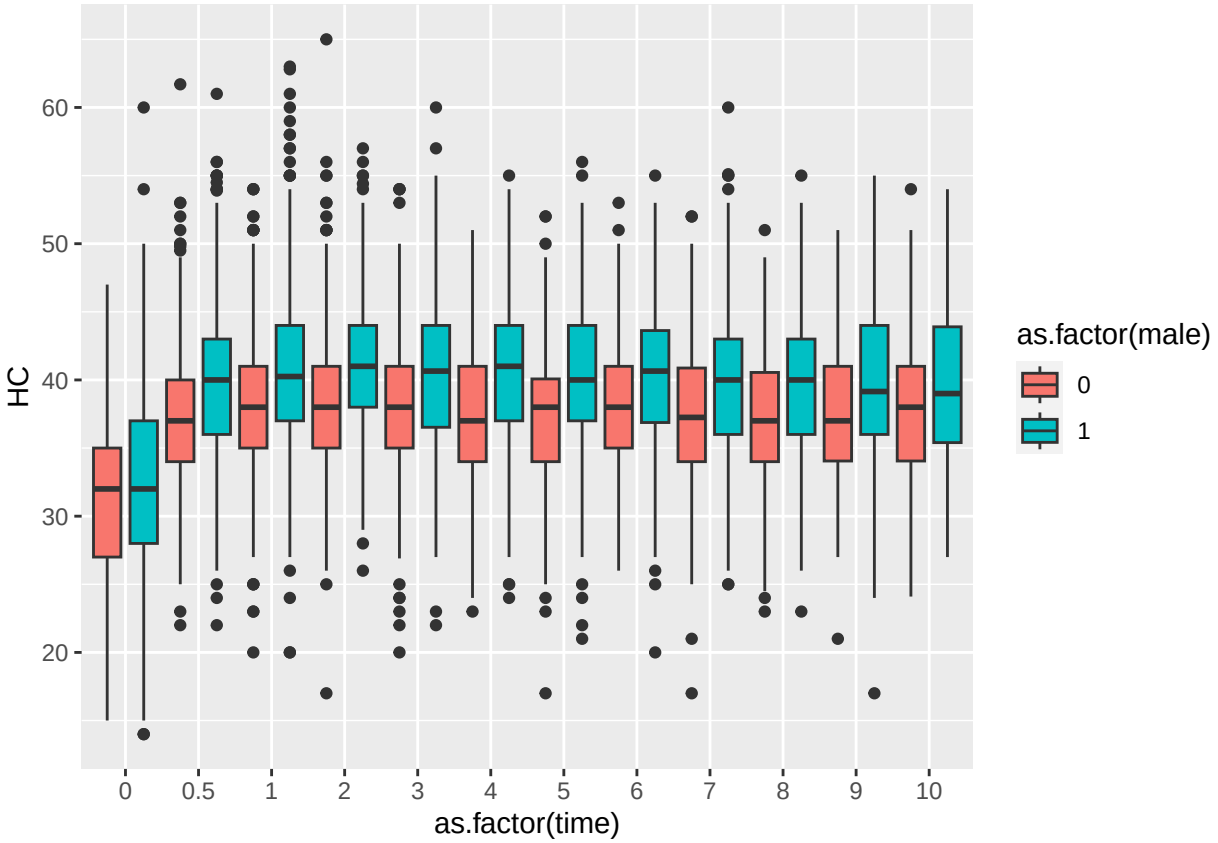

```
# Spaghetti Ggplot separated by cardio
p <- ggplot(data=data.long.noNA,aes(x=time,y=HC,group=id))
p <- p + geom_line(col="grey")+stat_summary(aes(group=1),geom="line",fun=mean,linewidth=2)
cardio.labs <- c("Cardio = 0","Cardio = 1")
p + facet_grid(~cardio,labeller = label_both)
```




```
# Spaghetti Ggplot separated by reject =1
p <- ggplot(data=data.long.noNA,aes(x=time,y=HC,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)
```

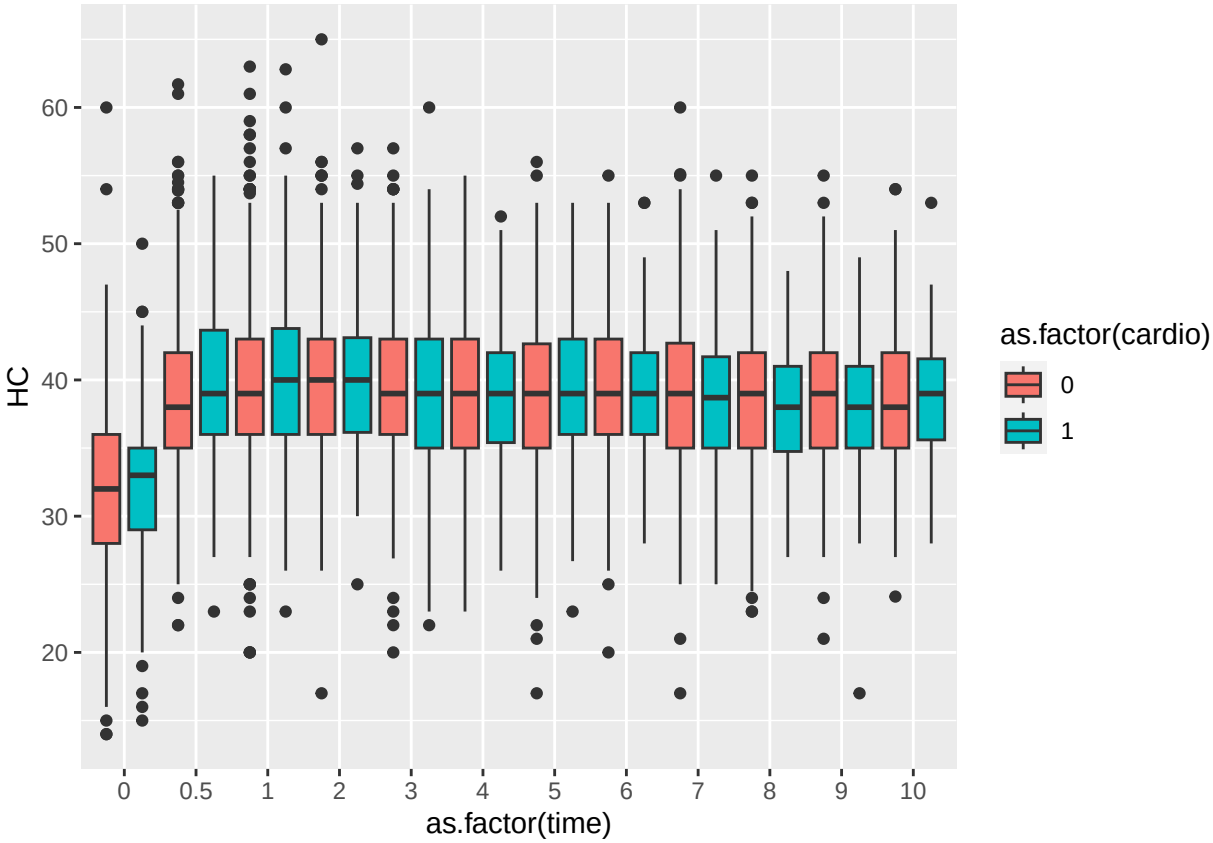



```
# Box plot by sex
ggplot(data.long.noNA,aes(x=as.factor(time),y=HC,fill=as.factor(male)))+
  geom_boxplot(position=position_dodge(1))
```

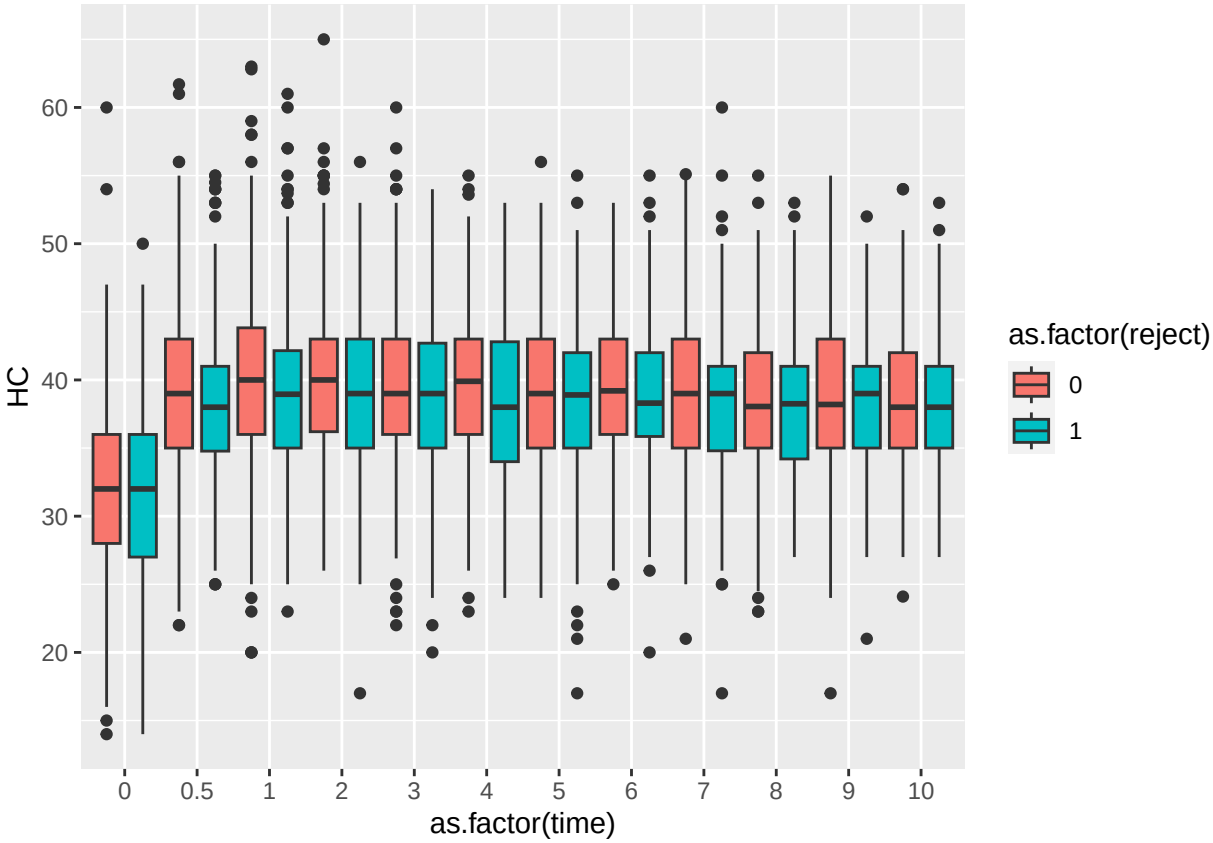


Boxplot

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




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



Hypothesis based on the plot 1.Age 2.Male 3.Cardio 4.Reject

Data set `Trenal.XLS` analysis with the linear mixed effects model

The chosen of fixed effects variable

We can choose all the predictors as the fixed effect variables, plus an intercept ## The chosen of random effects variable

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

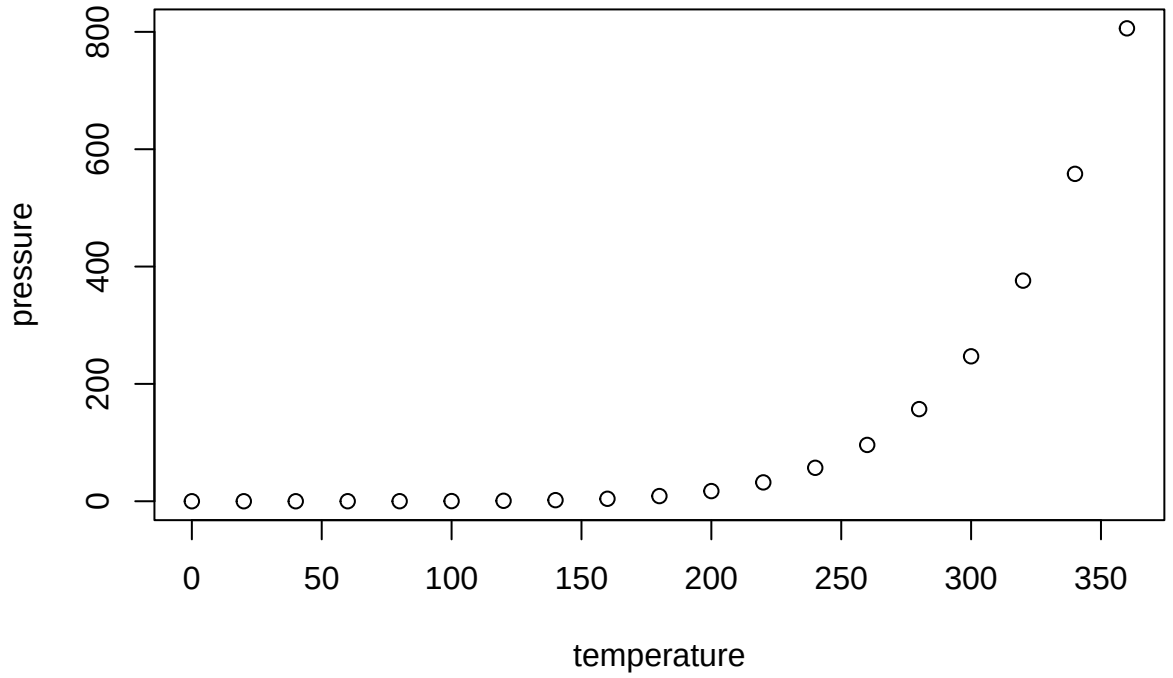
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.