# longitudinal analysis - how to average

## 1 Rational

Consider a dataset with repeated measurement over time for each individual. Suppose we are interested by the effect of a variable that is at the individual level (e.g. male vs. female). For instance we have:
Level 0 : Gender
Level 1 : Individual
Level 2 : Time

Then, under some assumptions (linearity assumptions, balanced design), we can average over level 2 and perform the analysis on the averaged data.

# 2 Simulation

## 2.1 Settings

```r
require(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.2.5
```

```r
require(lme4) ; require(nlme)
```

```
## Warning: package 'lme4' was built under R version 3.2.5
```

```r
require(data.table)

set.seed(9)

n.times <- 20
n.patients <- c(20,30)

diff_group <- 1.5
sdRF <- 2
sdNoise <- 1.5

mean1 <- seq(1,2, length.out = n.times)
```

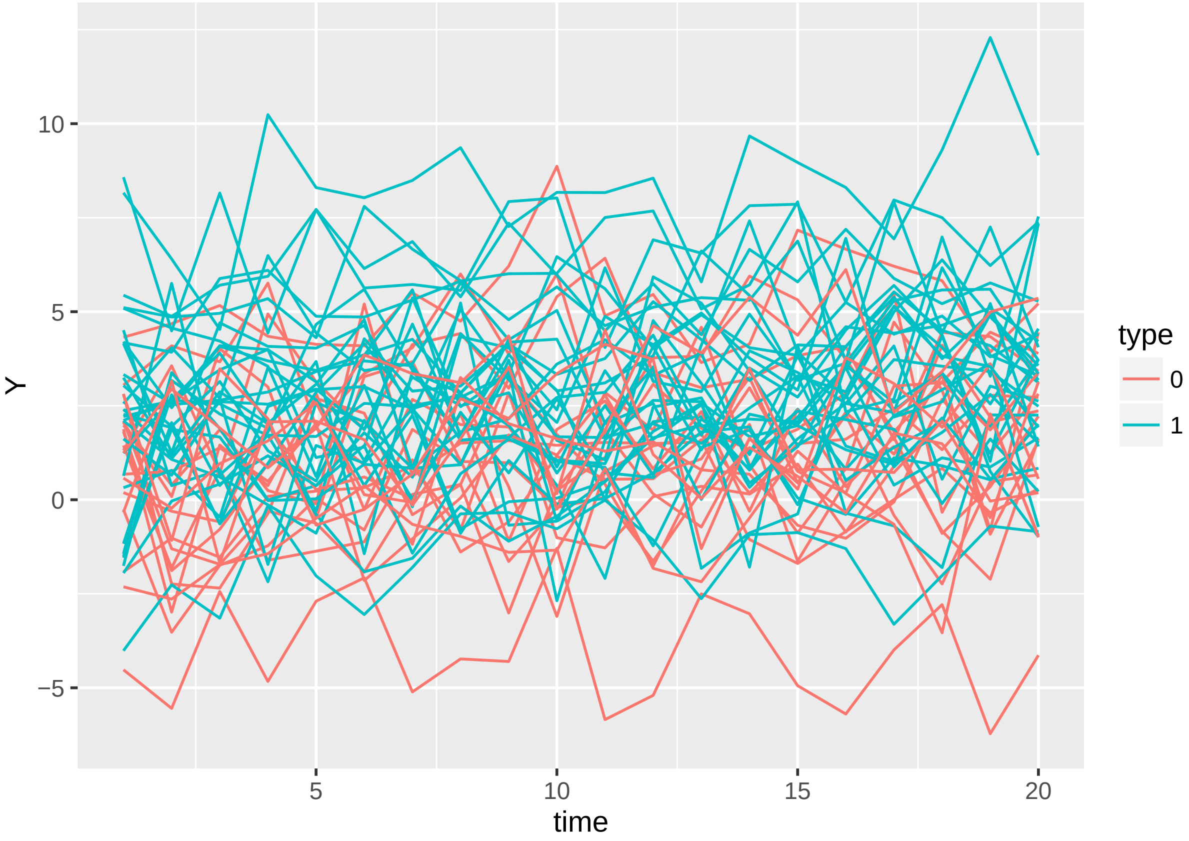
## 2.2 Generation of the dataset

```r
mean2 <- diff_group + mean1
randomEffect <- as.vector(sapply(rnorm(sum(n.patients), sd = sdRF),
                                 rep, n.times))
Id <- as.vector(sapply(1:sum(n.patients),
                rep, n.times))

Y <- c(rnorm(n.patients[1]*n.times, mean = mean1, sd = sdNoise),
       rnorm(n.patients[2]*n.times, mean = mean2, sd = sdNoise))

dt.data <- data.table(Id = as.character(Id), Y = Y + randomEffect,
                      time = 1:n.times,
                      type = c(rep("0", n.times*n.patients[1]),
                                              rep("1", n.times*n.patients[2])) )

ggplot(dt.data, aes(y = Y, x = time, group = Id, col = type)) + geom_line()
```



# 3 Analysis

## 3.1 random effect model

```
lme.Simul <- lme(Y ~ type + time,
                 data = dt.data,
                 random = ~1 | Id)
summary(lme.Simul)
```

```
## Linear mixed-effects model fit by REML
##  Data: dt.data
##         AIC      BIC    logLik
##   3765.091 3789.615 -1877.545
##
## Random effects:
##  Formula: ~1 | Id
##         (Intercept) Residual
## StdDev:    1.993295 1.439684
##
## Fixed effects: Y ~ type + time
##                  Value Std.Error  DF  t-value p-value
## (Intercept) 0.6128784 0.4590376 949 1.335138  0.1822
## type1       1.7492381 0.5828706  48 3.001074  0.0043
## time        0.0542741 0.0078954 949 6.874188  0.0000
##  Correlation:
##       (Intr) type1
## type1 -0.762
## time  -0.181  0.000
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.11741252 -0.67137820  0.01568415  0.66632928  2.75879287
##
## Number of Observations: 1000
## Number of Groups: 50
```

```
anova(lme.Simul, type = "marginal")
```

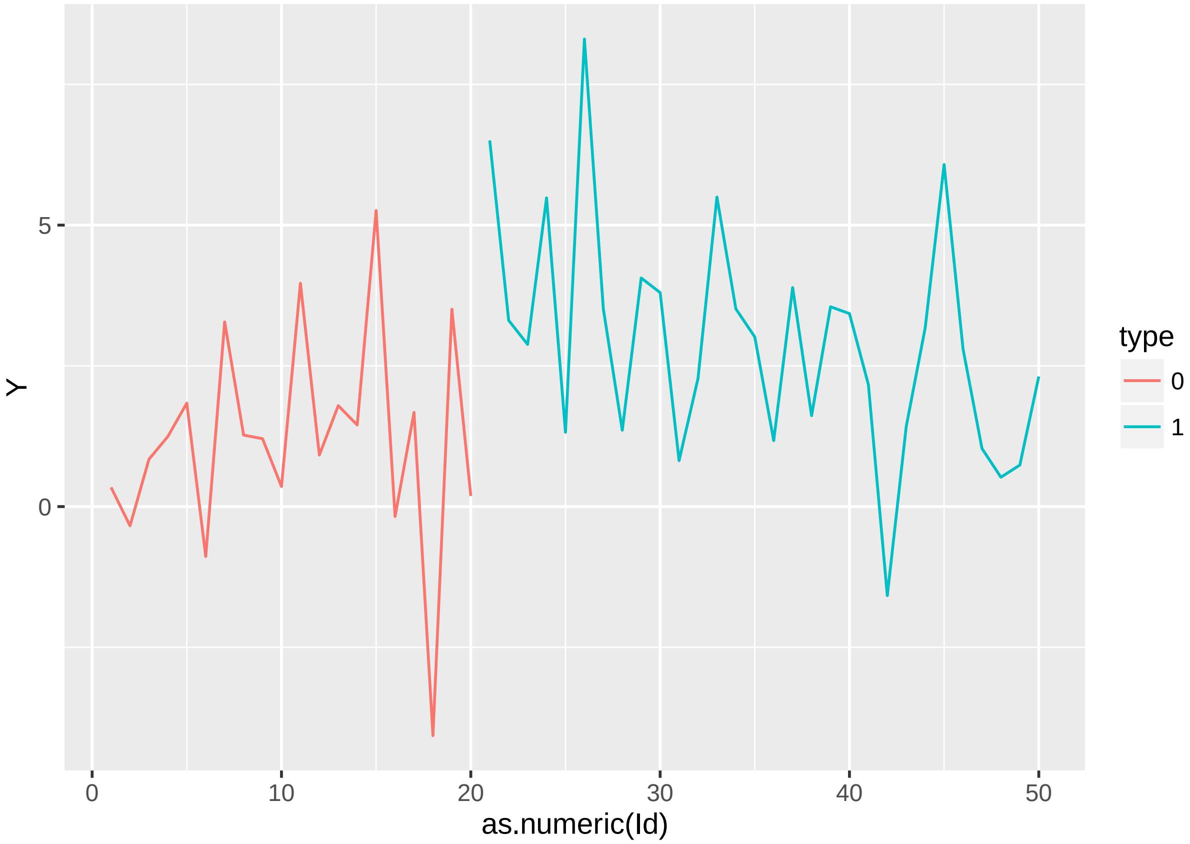
```
##             numDF denDF  F-value p-value
## (Intercept)     1   949  1.78259  0.1822
## type            1    48  9.00645  0.0043
## time            1   949 47.25446  <.0001
```

Note that $n.times * coef(lme.Simul)[1, "time"] = 1.0854824$ is close to $diff(range(mean1)) = 1$, as fixed by the simulation.

## 3.2 Average over time (level 2)

```
dt.data.Id <- dt.data[,.(Y = mean(Y)), by = c("type","Id")]
setkeyv(dt.data.Id, c("type","Id"))

ggplot(dt.data.Id,
       aes(y = Y, x = as.numeric(Id), group = type, col = type)) + geom_line()
```



```
aov.Id <- lm(Y ~ type, data = dt.data.Id)
summary(aov.Id)
```

```
##
## Call:
## lm(formula = Y ~ type, data = dt.data.Id)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2526 -1.3488  0.0428  0.6455  5.3725
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.1828     0.4515   2.620  0.01175 *
## type1         1.7492     0.5829   3.001  0.00426 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.019 on 48 degrees of freedom
## Multiple R-squared:  0.158,  Adjusted R-squared:  0.1404
## F-statistic: 9.006 on 1 and 48 DF,  p-value: 0.004259
```

```
anova(aov.Id)
```

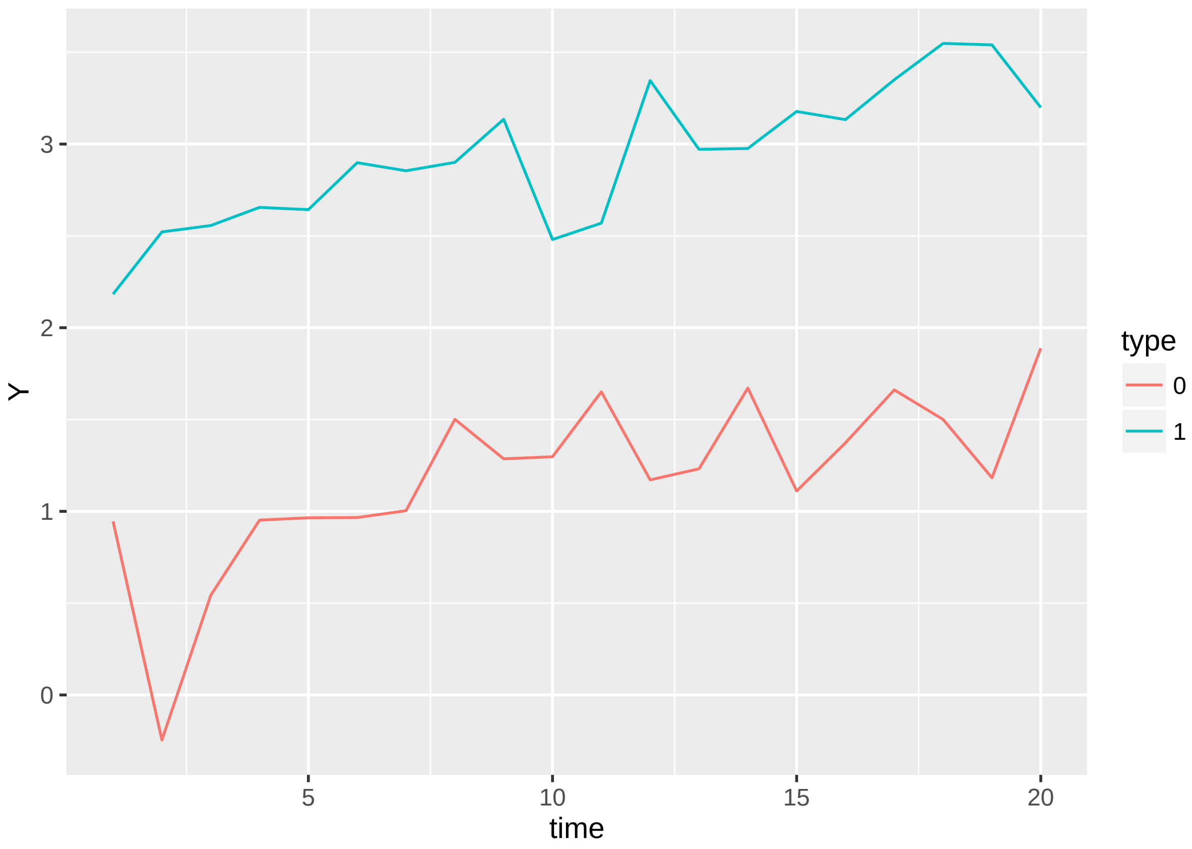
```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## type       1  36.718  36.718  9.0064 0.004259 **
## Residuals 48 195.689   4.077
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

CORRECT: same results as lme

## 3.3 Average over individual

```
dt.data.time <- dt.data[,.(Y = mean(Y)), by = c("type","time")]
setkeyv(dt.data.time, c("type","time"))

ggplot(dt.data.time,
       aes(y = Y, x = time, group = type, col = type)) + geom_line()
```



```r
aov.time <- lm(Y ~ type, data = dt.data.time)
summary(aov.time)
```

```
##
## Call:
## lm(formula = Y ~ type, data = dt.data.time)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.42814 -0.23228  0.01962  0.27978  0.70567
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.18276    0.09478   12.48 5.17e-15 ***
## type1        1.74924    0.13404   13.05 1.29e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4239 on 38 degrees of freedom
## Multiple R-squared:  0.8176, Adjusted R-squared:  0.8128
## F-statistic: 170.3 on 1 and 38 DF,  p-value: 1.292e-15
```

```r
anova(aov.time)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## type       1 30.5983 30.5983   170.3 1.292e-15 ***
## Residuals 38  6.8275  0.1797
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
t.test(dt.data.time[type==0,Y],  dt.data.time[type==1,Y])
```

```
##
##  Welch Two Sample t-test
##
## data:  dt.data.time[type == 0, Y] and dt.data.time[type == 1, Y]
## t = -13.05, df = 36.353, p-value = 2.757e-15
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.020995 -1.477481
## sample estimates:
## mean of x mean of y
##  1.182757  2.931995
```

```r
summary(lm(Y ~ type, data = dt.data.time))
```

```
##
## Call:
## lm(formula = Y ~ type, data = dt.data.time)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.42814 -0.23228  0.01962  0.27978  0.70567
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.18276    0.09478   12.48 5.17e-15 ***
## type1        1.74924    0.13404   13.05 1.29e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.4239 on 38 degrees of freedom
## Multiple R-squared:  0.8176, Adjusted R-squared:  0.8128
## F-statistic: 170.3 on 1 and 38 DF,  p-value: 1.292e-15
```

```
t.test(dt.data.time[type==0,Y],  dt.data.time[type==1,Y],
       paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  dt.data.time[type == 0, Y] and dt.data.time[type == 1, Y]
## t = -17.735, df = 19, p-value = 2.799e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.955683 -1.542793
## sample estimates:
## mean of the differences
##               -1.749238
```

INCORRECT: underestimation of the variability