ST 437/537: Applied Multivariate and Longitudinal Data Analysis

Longitudinal Data Analysis: Introduction

Arnab Maity
NCSU Department of Statistics
SAS Hall 5240 919-515-1937 amaity[at]ncsu.edu

References:

- Modeling Longitudinal Data by Robert E. Weiss. New York: Springer.
- Linear Mixed Models for Longitudinal Data by Geert Verbeke and Geert Molenberghs. New York: Springer.
- Applied Longitudinal Analysis by Fitzmaurice by G.M., Laird, N.M., and Ware, J.H. New York: Wiley (on reserve at NCSU library)

Introduction

The simplest design for a longitudinal study consists of a random sample of subjects/item, where multiple observations correspond to a variable observed at multiple follow-up times. Longitudinal data analysis refers to statistical techniques for studying the behavior of the variable over time. The need often arises in agriculture and the life sciences, medical and public health research, and physical science and engineering, among other fields.

Recall that longitudinal data are different from multivariate data, as the **order of the repeated measurements is essential in the analysis of longitudinal data**, whereas permuting the order of the variables in multivariate analysis yields same results.

Consider the following examples. Pay attention to what is the response variable, what is the observational unit, how many measurements are collected per unit.

Example 1: Treatment of Lead Exposed Children (TLC) Trial

The dataset and its description are available at **[https://content.sph.harvard.edu/fitzmaur/ala2e/] (https://content.sph.harvard.edu/fitzmaur/ala2e/)**. The TLC trial was a placebo-controlled, randomized study of **a chelating agent (https://en.wikipedia.org/wiki/Chelation)** (succimer) in children with blood lead levels of 20-44 micrograms/dL. We only consider an subsample of size 50 ($N = 50$) from the children who received succimer. The dataset consist of four repeated measurements of blood lead levels obtained at baseline (week 0), week 1, week 4, and week 6 on each of the 50 children.

```
tab <- read.table("data/lead-data.txt", header = F)
colnames(tab) <- c("ID", "Week 0", "Week 1", "Week 4", "Week 6")
head(tab)
```

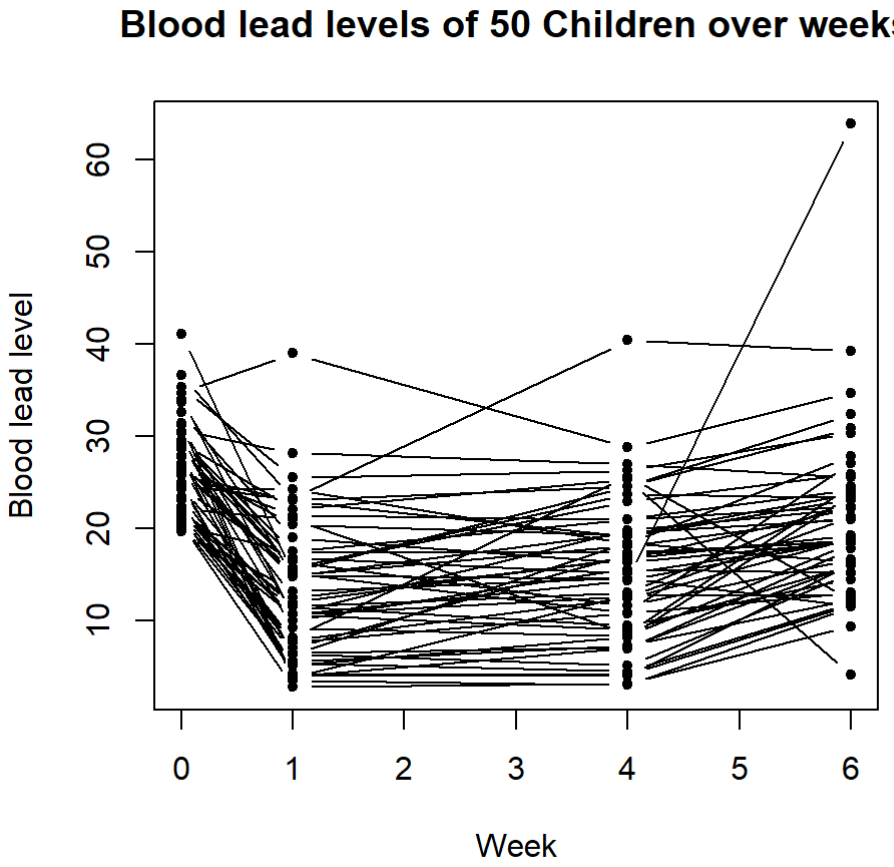
```
##   ID Week 0 Week 1 Week 4 Week 6
## 1  1   26.5   14.8   19.5   21.0
## 2  2   25.8   23.0   19.1   23.2
## 3  3   20.4    2.8    3.2    9.4
## 4  4   20.4    5.4    4.5   11.9
## 5  5   24.8   23.1   24.6   30.9
## 6  6   27.9    6.3   18.5   16.3
```

```
tail(tab)
```

```
##    ID Week 0 Week 1 Week 4 Week 6
## 45 45   31.2   10.8   19.8   22.2
## 46 46   31.4    3.9    7.0   17.8
## 47 47   41.1   15.1   10.9   27.1
## 48 48   29.4   22.1   25.3    4.1
## 49 49   21.9    7.6   10.8   13.0
## 50 50   20.7    8.1   25.7   12.3
```

A plot of the dataset is shown below; each line corresponds to one child.

```
# Time of measurements
time <- c(0, 1, 4, 6)

# Plot
matplot(time, t(tab[, -1]), type = "b",
        lty=1, pch=19, col = "black", cex=0.7,
        xlab = "Week", ylab = "Blood lead level", main = "Blood lead levels of 50 Childr
en over weeks")
```



Notice that we defined a `time` variable to account for the unequal spacing of the weeks (0, 1, 4, 6) to properly display the data.

In this example,

- **response variable:** Blood lead levels,
- **observational unit:** a child,
- **number of measurements collected per unit:** 4.

Example 2: Six Cities Study of Air Pollution and Health

The dataset and its description are available at **[https://content.sph.harvard.edu/fitzmaur/ala2e/] (https://content.sph.harvard.edu/fitzmaur/ala2e/)**. The dataset contains a subset of the pulmonary function data collected in the Six Cities Study. The data consist of all measurements of FEV1, height and age obtained from a randomly selected subset of the female participants living in Topeka, Kansas. The random sample consists of 300 girls, with a minimum of one and a maximum of twelve observations over time.

```
tab <- read.table("data/fev1-data.txt", header = F)
colnames(tab)=c("id", "Height", "age", "height_base", "age_base", "logFEV1")
head(tab)
```

```
##   id Height     age height_base age_base logFEV1
## 1  1   1.20  9.3415         1.2   9.3415 0.21511
## 2  1   1.28 10.3929         1.2   9.3415 0.37156
## 3  1   1.33 11.4524         1.2   9.3415 0.48858
## 4  1   1.42 12.4600         1.2   9.3415 0.75142
## 5  1   1.48 13.4182         1.2   9.3415 0.83291
## 6  1   1.50 15.4743         1.2   9.3415 0.89200
```

```
tail(tab)
```

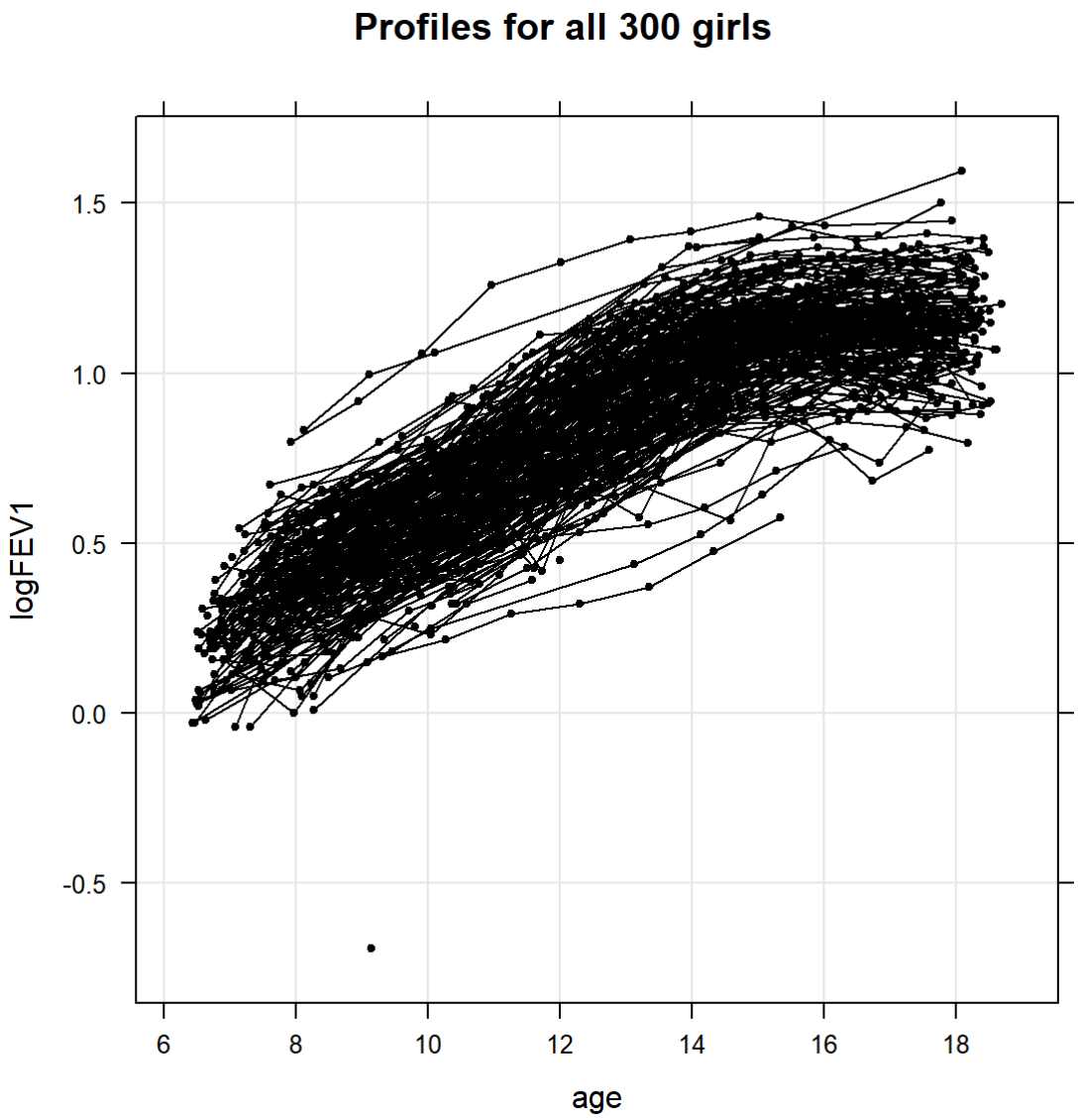
```
##       id Height     age height_base age_base logFEV1
## 1989 300   1.50 12.9993        1.44  11.9617 0.85015
## 1990 300   1.57 13.9055        1.44  11.9617 0.81536
## 1991 300   1.61 14.9596        1.44  11.9617 1.11841
## 1992 300   1.62 15.9398        1.44  11.9617 1.08181
## 1993 300   1.62 17.0075        1.44  11.9617 1.12817
## 1994 300   1.63 17.8645        1.44  11.9617 1.16938
```

Let us consider the variable `logFEV1` only for this discussion. We plot `logFEV1` vs `age` of each individual (top panel), and the same plot for a subset of the data for better visualization (bottom panel).

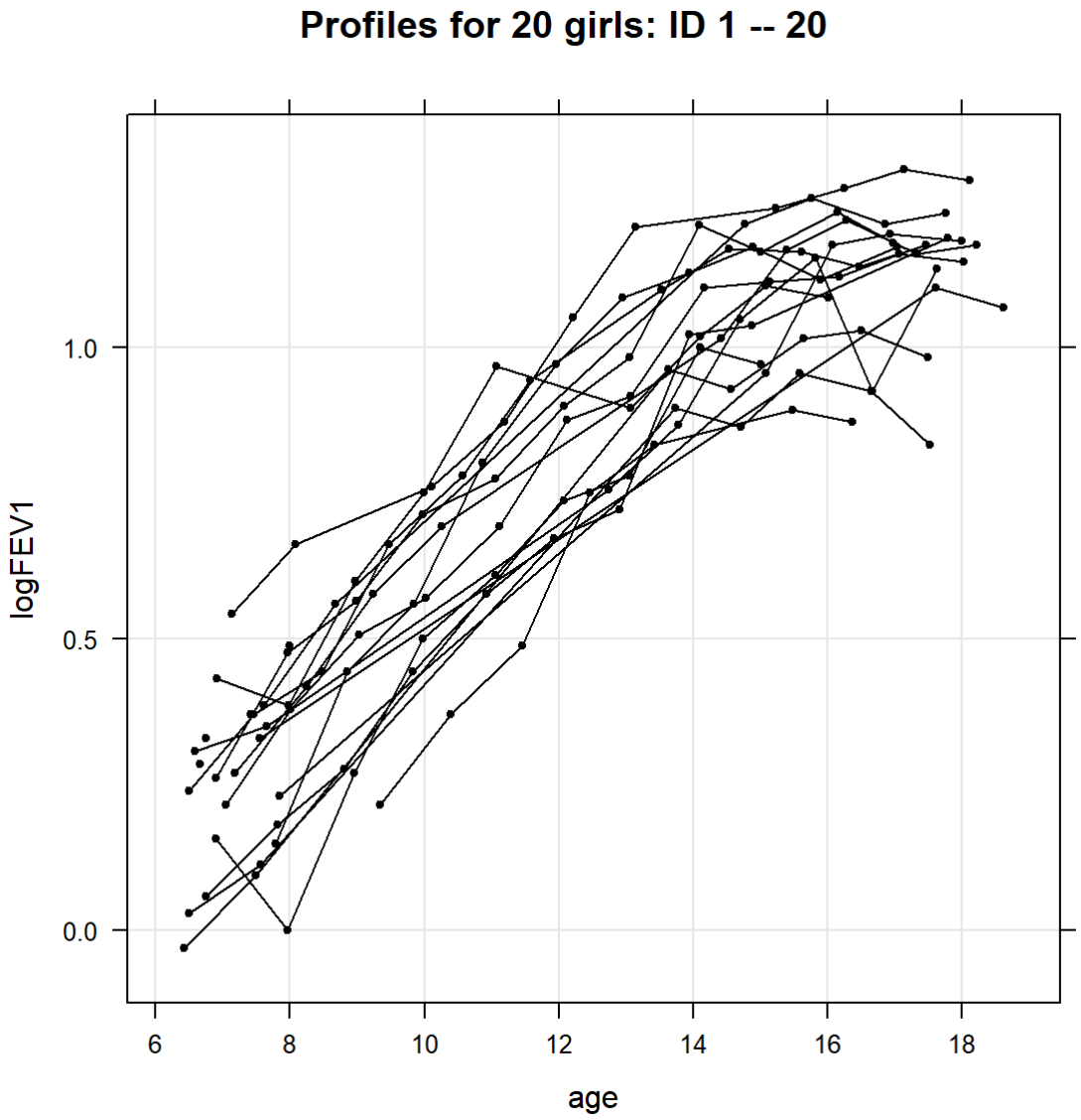
```
library(lattice)
library(latticeExtra)
```

```
## Loading required package: RColorBrewer
```

```
xyplot(logFEV1 ~ age, data=tab, groups = id, type = c("b", "g"), col="black", pch=19, ce
x=0.5, main = "Profiles for all 300 girls")
```




```
tabsub <- subset(tab, id <=20)
xyplot(logFEV1 ~ age, data=tabsub, groups = id, type = c("b", "g"), col="black", pch=19,
cex=0.5, main = "Profiles for 20 girls: ID 1 -- 20")
```



Notice that the `age` variable takes dufferent values for each girl.

In this example,

- **response variable:** log(FEV1),
- **observational unit:** a girl,
- **number of measurements collected per unit:** varies between 1 to 12.

```
# Number of observations per unit
numobs <- tabulate(tab$id)
summary(numobs)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   7.000   6.647  10.000  12.000
```

In the examples above we see that the multiple observations within each subject can be ordered across time; this is the main feature of longitudinal data.

Basic concepts and Notations

Let us first introduce some notation that will be used throughout the course.

- **Response** is the outcome of interest (denoted typically by Y).
- **Unit** (object or subject) is the object on which repeated measurements are taken; typically they are individuals (i indexes units and j indexes the repeated measurement).
- Y_{ij} - denotes the jth repeated measurement taken on the ith subject or unit; n denotes the total number of units and m_i denotes the number of repeated measurements for unit i.
- The **response vector** of measurements for unit i is

$$\boldsymbol{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{im_i} \end{bmatrix}$$

- The responses are typically assumed independent across units (e.g. $\boldsymbol{Y}_i, \boldsymbol{Y}_{i'}$ are independent for $i \neq i'$). However within the unit the responses are correlated (e.g. Y_{ij} and $Y_{ij'}$ are typically correlated). Many statistical models consider modeling the response vector $\boldsymbol{Y}_i$ and not the Y_{ij}'s separately; nevertheless it is not uncommon to model Y_{ij}'s separately. We'll discuss modes that exploit both representations.
- **Time** is the generic term for the condition of measurement (t is used to denote time). Time is considered an important covariate in longitudinal data. Both the mean of the response vector $\boldsymbol{Y}_i$ and the covariance matrix of $\boldsymbol{Y}_i$ may be depend on the time variable. We use t_{ij} to denote the time corresponding to the Y_{ij}.
- Although not specified explicitly, it is assumed that times occur in an increasing order $t_{i1} < t_{i2} < \ldots < t_{im_i}$.

Depending on the observation times, we might have a balanced/regular design.

- We say the design is **balanced** when $m_i = m$ (same number of repeated measurements across units). Otherwise we say the design is *unbalanced*. In our two examples, the TLC trial data is balanced, but the six cities air pollution data is unbalanced.
- We say the design is **regular** if $t_{ij} = t_j$ (the times of measurements are the same for all the units). Otherwise we say the design is *irregular*. In our two examples, the TLC trial data is regular, but the six cities air pollution data is irregular.

General data structure for a balanced, regular design ($m_i = m$ and $t_{ij} = t_i$) is:

	t_1	t_2	t_3	$\dots$	t_m
Units					
1	Y_{11}	Y_{12}	Y_{13}	$\dots$	Y_{1m}
2	Y_{21}	Y_{22}	Y_{23}	$\dots$	Y_{2m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\dots$	$\vdots$
n	Y_{n1}	Y_{n2}	Y_{n3}	$\dots$	Y_{nm}

For an irregular and unbalanced design, we may use the “long format”:

Unit	Time	Response
1	t_{11}	Y_{11}
	$\vdots$	
1	t_{1m_1}	Y_{1m_1}
	$\vdots$	
n	t_{n1}	Y_{n1}
	$\vdots$	
n	t_{nm_n}	Y_{nm_n}

Notice that the long format is also applicable for the balanced and regular case as well. Also, is a covariate is present, we simply add extra columns; see for example the six cities data, where baseline height, baseline weight can be considered as covariates (do not change over time). If we consider height as a covariate as well, this is an example of a covariate that changes over time.

```
##   id Height     age height_base age_base logFEV1
## 1  1   1.20  9.3415         1.2   9.3415 0.21511
## 2  1   1.28 10.3929         1.2   9.3415 0.37156
## 3  1   1.33 11.4524         1.2   9.3415 0.48858
## 4  1   1.42 12.4600         1.2   9.3415 0.75142
## 5  1   1.48 13.4182         1.2   9.3415 0.83291
## 6  1   1.50 15.4743         1.2   9.3415 0.89200
```

Inferences about longitudinal data

In the following, consider the observed data: $\{(Y_{ij}, t_{ij}) : j = 1, \ldots, m_i\}_i$, where Y_{ij} **is assumed to be continuous**. For simplicity we assume $t_{ij} = t_j$ and $m_i = m$ (**balanced** and **regular** design). We are interested in studying the typical behavior of the outcome over time, and furthermore in studying the way the outcome vary over time.

Estimating mean and variance

Since we data are collected over multiple time points, the population mean is not a single number. Instead, the population mean can be thought of as a function of time: $\mu_j = \mu(t_j)$. Thus, at the time point t_j, we can estimate mu_j by the sample mean of the data observed at t_j:

$$\widehat{\mu}_j = \bar{Y}_j = \frac{1}{n}\sum_{i=1}^{n} Y_{ij}.$$

In general, $\mu(t)$ can be constant over time, or have a trend over time. It can also depend on other covariates. We will discuss such scenarios later.

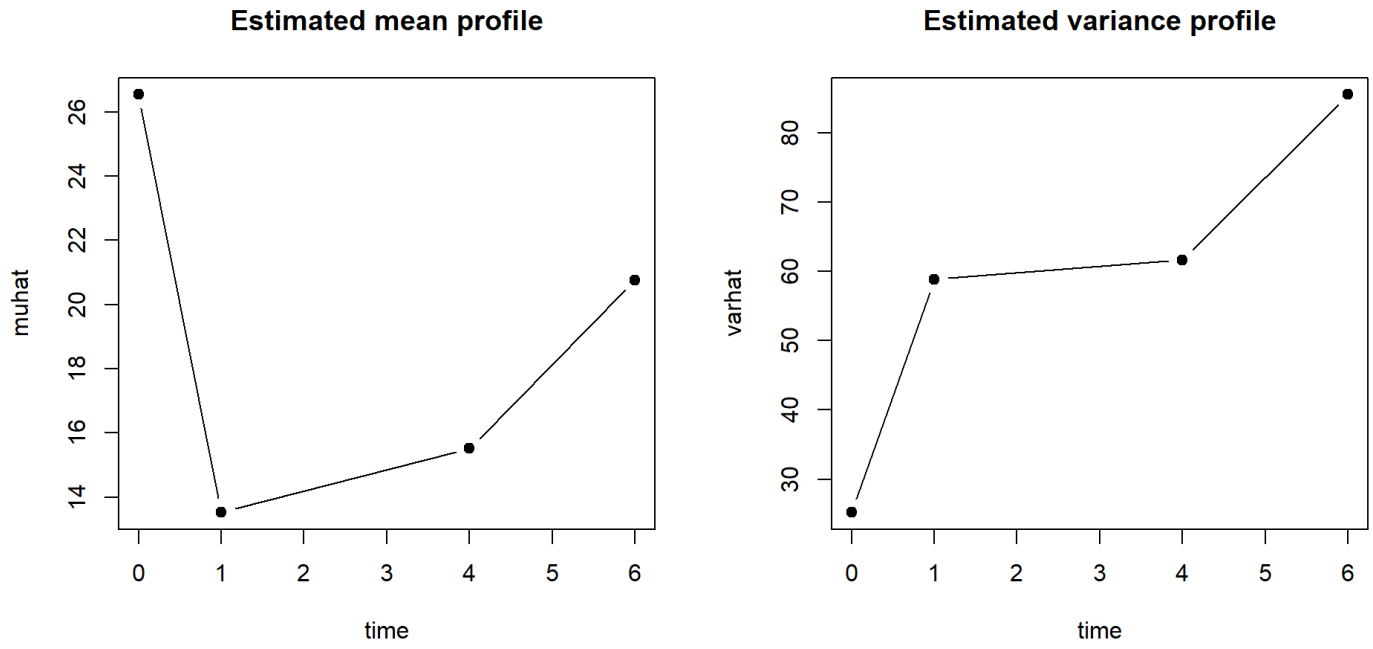
Simularly, the population variance is also not a single number but a function of time. At time t_j, the population variance is σ_j^2, and can be estimated by the sample variance of the data observed at t_j:

$$\widehat{\sigma}_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_{ij} - \bar{Y}_j)^2.$$

The population variance σ_j^2 describes how the response varies at a particular time.

In our blood lead level example (recall, the design is balanced and regular), the estimated mean and variance profiles are shown below.

```
# Read the TLC data
tab <- read.table("data/lead-data.txt", header = F)
colnames(tab) <- c("ID", "Week 0", "Week 1", "Week 4", "Week 6")

# Time of measurements
time <- c(0, 1, 4, 6)

# Estimated mean
muhat <- colMeans(tab[, -1])

# Estimated variance
varhat <- apply(tab[, -1], 2, var)

# Plot
par(mfrow = c(1,2))
plot(time, muhat, type = "b", pch=19, main = "Estimated mean profile")
plot(time, varhat, type = "b", pch=19, main = "Estimated variance profile")
```



Covariance and correlation

Since there are multiple measurements for each subject, there might be correlation between these measurements. Specifically, for any subject i, the measurements Y_{ij} and $Y_{i\ell}$, taken at t_j and t_ℓ, can be correlated; the covariance is denoted as $\sigma_{j\ell} = cov(Y_{ij}, Y_{i\ell})$, and thus the correlation $\rho_{j\ell} = \sigma_{j\ell}/(\sigma_j \sigma_\ell)$.

These quantities can be estimated by the sample covariance and correlation, respectively:

$$\widehat{\sigma}_{j\ell} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_{ij} - \bar{Y}_j)(Y_{i\ell} - \bar{Y}_\ell),$$

$$\widehat{\rho}_{j\ell} = \frac{\widehat{\sigma}_{j\ell}}{\widehat{\sigma}_j\widehat{\sigma}_\ell}.$$

In general, an unbiased estimator for the population covariance is the sample covariance

$$\widehat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}})^T;$$

the numerator $\sum_{i=1}^{n}(\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{Y}_i - \widehat{\boldsymbol{\mu}})^T$ is also known as the sums of square and cross-product matrix (SS & CP).

For the TLC data, the estimated covariance and correlation coefficients are plotted below.

```
# Estimated covariance
covhat <- cov(tab[, -1])

# Estimated correlation
corrhat <- cor(tab[, -1])
corrhat
```

```
##           Week 0    Week 1    Week 4    Week 6
## Week 0 1.0000000 0.4014589 0.3839654 0.4951063
## Week 1 0.4014589 1.0000000 0.7308221 0.5069743
## Week 4 0.3839654 0.7308221 1.0000000 0.4548224
## Week 6 0.4951063 0.5069743 0.4548224 1.0000000
```

Autocorrelation

Another measure that describes the association is the **autocorrelation**: the correlation between the repeated measurements when the `lag', or distance between the time, is constant. Stationarity is a property of a stochastic processes that is related to the first/second/etc. moments being constant over time. Examining the autocorrelation is done with the purpose of checking for stationarity assumption (whether the covariance varies with the lag between the observations $|t_j - t_\ell|$ instead of the actual times, t_j, t_ℓ).

Autocorrelation is formally defined as:

$$\rho(u) = corr\{Y_{ij}, Y_{i\ell}\}, \quad \text{, where} \quad |t_j - t_\ell| = u;$$

this measure describes the stationarity nature of the dependence. Here u is commonly referred as the **lag**.

To study this behavior, we plot for each lag u, the following standardized residuals

$$\frac{Y_{ij} - \widehat{\mu}_j}{\widehat{\sigma}_j} \quad \frac{Y_{i\ell} - \widehat{\mu}_\ell}{\widehat{\sigma}_\ell}, \quad |t_j - t_\ell| = u;$$

Equivalently one can calculate a sample autocorrelation estimator $\widehat{\rho}(u)$, based on these standardized residuals. Notice however that the estimator is based on different number of pairs, hence is characterized by different theoretical properties at various lags, and thus caution should be used in interpreting it.

One alternative of using autocorrelatio is to use a **variogram**. The variogram is defined as

$$V(u) = \frac{1}{2}E\{(Y_{ij} - Y_{i\ell})\}^2, \quad \text{, where} \quad |t_j - t_\ell| = u.$$

For stationary processes (mean and variance constant over time) we have $V(u) = \tau^2 + \sigma^2\{1 - \rho(u)\}$, where τ^2 is the noise variance (known from spatial statistics as `nugget' effect). When data are unbalanced it is easier to estimate $V(u)$ than $\rho(u)$.

To estimate the variogram, we need

$$v_{ij\ell} = \frac{1}{2}(Y_{ij} - Y_{i\ell})^2,$$

and estimate $V(u)$ by $\widehat{V}(u) = Ave_{|t_{ij} - t_{i\ell}| \approx u}(v_{ij\ell})$.

We will discuss more about correlogram and variogram later in the course.

What kind of questions are we looking to answer?

Consider the TLC trial data presented before. We can ask:

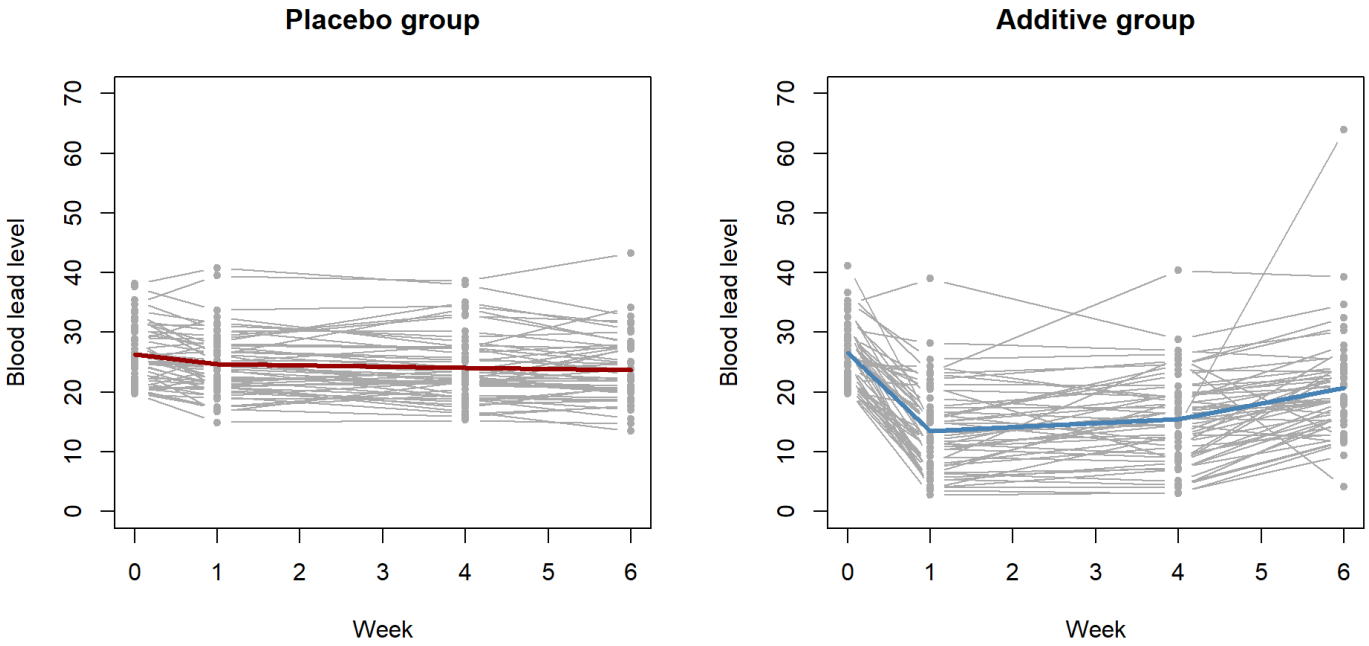
- How does the average response (blood lead level) change over time?
- How can we predict the future of a new subject given previous measurements?

Suppose we have samples of two groups: Subsample of size $N = 100$ of data on Blood Lead Levels from the Treatment of Lead Exposed Children (TLC) Trial. Specifically 50 children from each arm: placebo(P) and succimer (A).

```
tab <- read.table("data/tlc-data.txt", header = F)
colnames(tab) <- c("ID", "Treatment", "Week 0", "Week 1", "Week 4", "Week 6")
head(tab)
```

```
##   ID Treatment Week 0 Week 1 Week 4 Week 6
## 1  1         P   30.8   26.9   25.8   23.8
## 2  2         A   26.5   14.8   19.5   21.0
## 3  3         A   25.8   23.0   19.1   23.2
## 4  4         P   24.7   24.5   22.0   22.5
## 5  5         A   20.4    2.8    3.2    9.4
## 6  6         A   20.4    5.4    4.5   11.9
```

```
# Time of measurements
time <- c(0, 1, 4, 6)

# Placebo group, only the lead level
placebo <- subset(tab, Treatment == "P")[, -c(1,2)]
mean.plc <- colMeans(placebo)

# Succimer group, only the lead level
additive <- subset(tab, Treatment == "A")[, -c(1,2)]
mean.add <- colMeans(additive)

# Plot
par(mfrow = c(1,2))
matplot(time, t(placebo), type = "b", ylim = c(0, 70),
        lty=1, pch=19, col = "darkgrey", cex=0.7,
        xlab = "Week", ylab = "Blood lead level",
        main = "Placebo group")
lines(time, mean.plc, col = "#990000", lwd=3)

matplot(time, t(additive), type = "b", ylim = c(0, 70),
        lty=1, pch=19, col = "darkgrey", cex=0.7,
        xlab = "Week", ylab = "Blood lead level",
        main = "Additive group")
lines(time, mean.add, col = "steelblue", lwd=3)
```



The red and blue lines are the mean profiles of the placebo and additive groups, respectively. In this scenario, we can ask:

- Are the mean profiles of the two groups same at all the time points?
- If there is a difference, is it increasing or decreasing over time?
- Is the difference is not monotone, how does the difference change over time?

In general, the groups in the above data is an example of a covariate measured in the longitudinal study. As a general question, we may ask:

- How does the covariate impact the mean profile? For example, in the six cities data example described above, how does the covariate `Hight` impact the mean `logFEV1` level?

Exploring Longitudinal Data: Plots

We want to create plots that reveal different features of the longitudinal data that relate to modeling choices. Modeling longitudinal data is more complex than modeling independent data:

- need to model the correlation among the repeated measurements
- modeling the mean trend across time requires attention
- variance of the response may differ across time as well.

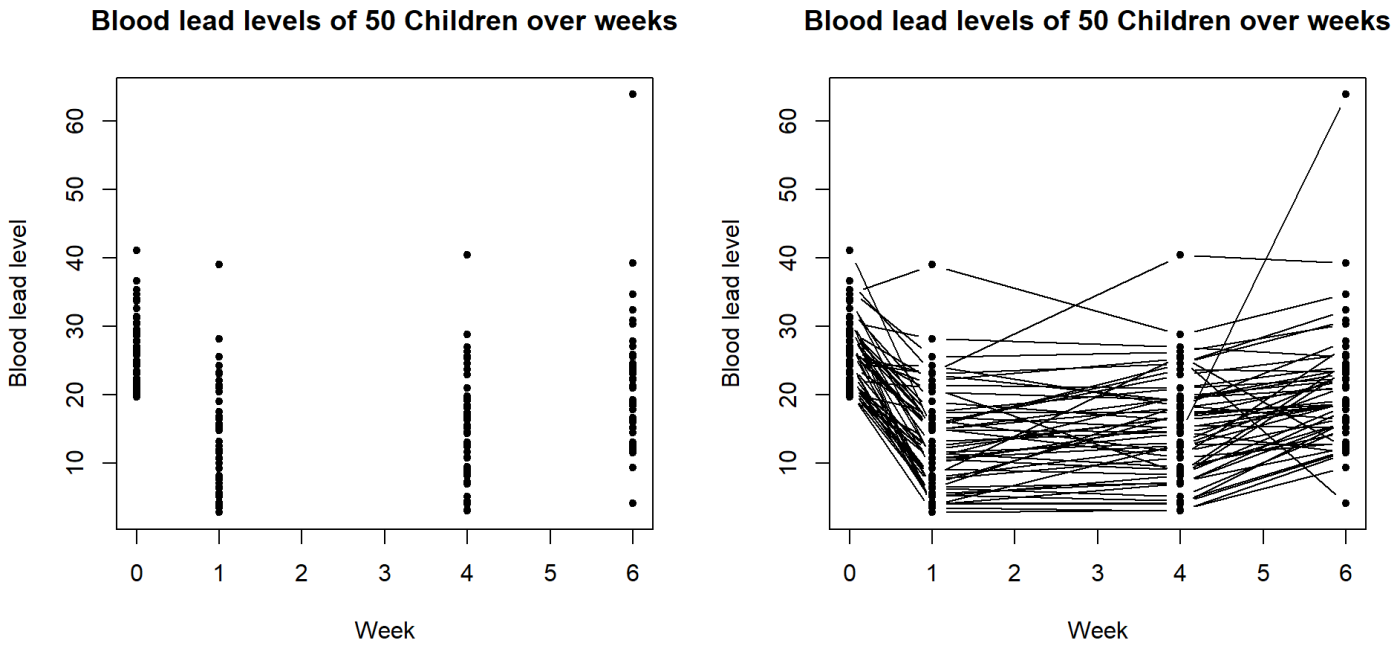
- typically the effect of the various predictors is modeled in the mean (systematic part).

We primarily want to make qualitative judgement from the plots. Once we select and fit a model, we can then make quantitative judgements from the output of the model fit.

Profile plot / Spaghetti plots

For longitudinal data, the usual scatterplot is of little use. Consider the TLC data desctibed in the introduction section. A simple scatter plot (left panel below) does not show individual trajectories at all.

```
tab <- read.table("data/lead-data.txt", header = F)
colnames(tab) <- c("ID", "Week 0", "Week 1", "Week 4", "Week 6")
time <- c(0, 1, 4, 6)


par(mfrow = c(1,2))
#  scatter Plot
matplot(time, t(tab[, -1]), type = "p",
        lty=1, pch=19, col = "black", cex=0.7,
        xlab = "Week", ylab = "Blood lead level", main = "Blood lead levels of 50 Childr
en over weeks")

#  Profile Plot
matplot(time, t(tab[, -1]), type = "b",
        lty=1, pch=19, col = "black", cex=0.7,
        xlab = "Week", ylab = "Blood lead level", main = "Blood lead levels of 50 Childr
en over weeks")
```



Profile/Spaghetti plots (right panel above) are a method of viewing data to visualize the dynamic behavior over time, corresponding to each unit/subject. Notice the measurements for each subject are connected with line; but measurements from diffeerent subjects are not connected.

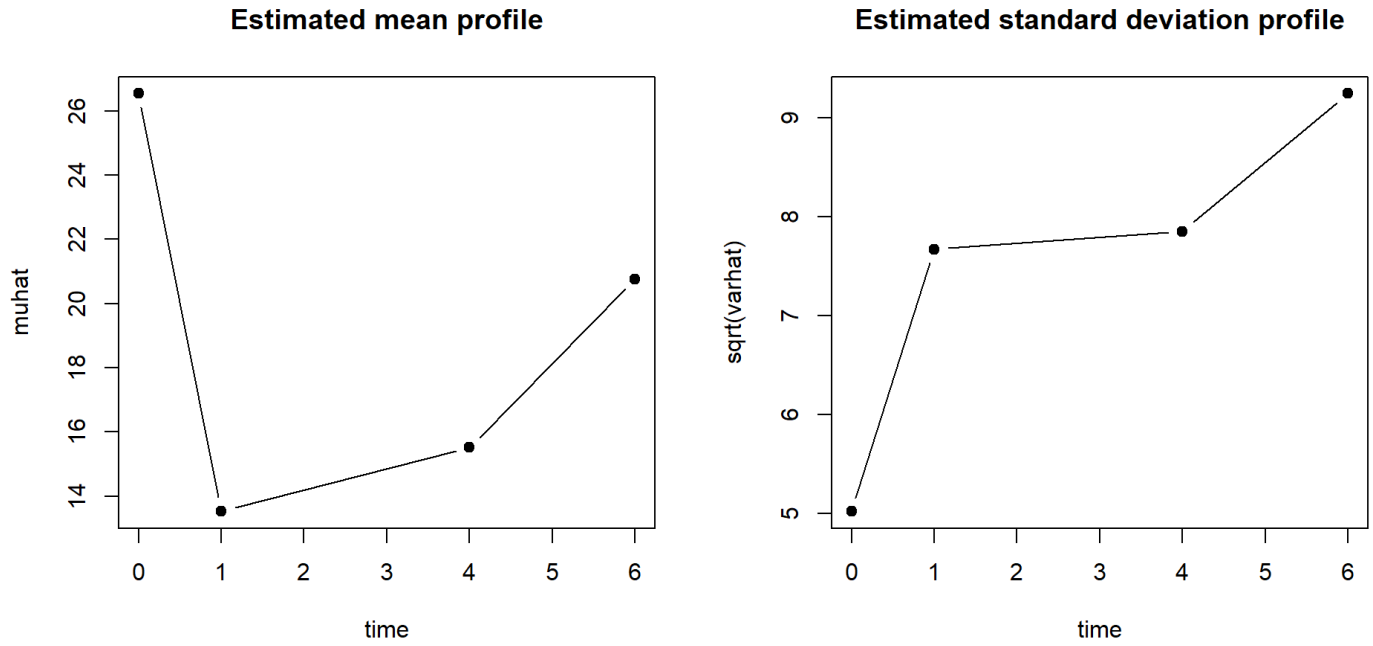
> *In a profile plot, the basic plotting unit is not the observation* (Y_{ij}, t_{ij})*; it is the* ***whole profile*** $\boldsymbol{Y}_i$*.*

Interpreting profile plot may give us valuable insight on how to model the observed data. For instance, the profile plot above indicates that the population mean is not a linear function of time; in fact, it seems to be a *piecewise* linear function with different slope and intercept for each piece. Perhaps we could also try to fit a quadratic function of time as well.

Sample Means and Standard Deviations

As we discussed earlier, the evolution of mean and standard deviation (or variance) is usually of interest. For a dataset with regular design, they are easy to plot. Take the TLC data (only for succimer arm) for instance (we have seen this plot above):

```
# Read the TLC data
tab <- read.table("data/lead-data.txt", header = F)
colnames(tab) <- c("ID", "Week 0", "Week 1", "Week 4", "Week 6")

# Time of measurements
time <- c(0, 1, 4, 6)

# Estimated mean
muhat <- colMeans(tab[, -1])

# Estimated variance
varhat <- apply(tab[, -1], 2, var)

# Plot
par(mfrow = c(1,2))
plot(time, muhat, type = "b", pch=19, main = "Estimated mean profile")
plot(time, sqrt(varhat), type = "b", pch=19, main = "Estimated standard deviation profil
e")
```



The population standard deviation measures the across-subject variability within each time point. It is clear that our modeling strategy for this data should account for increasing variablity of the response over time to ensure correct inference.

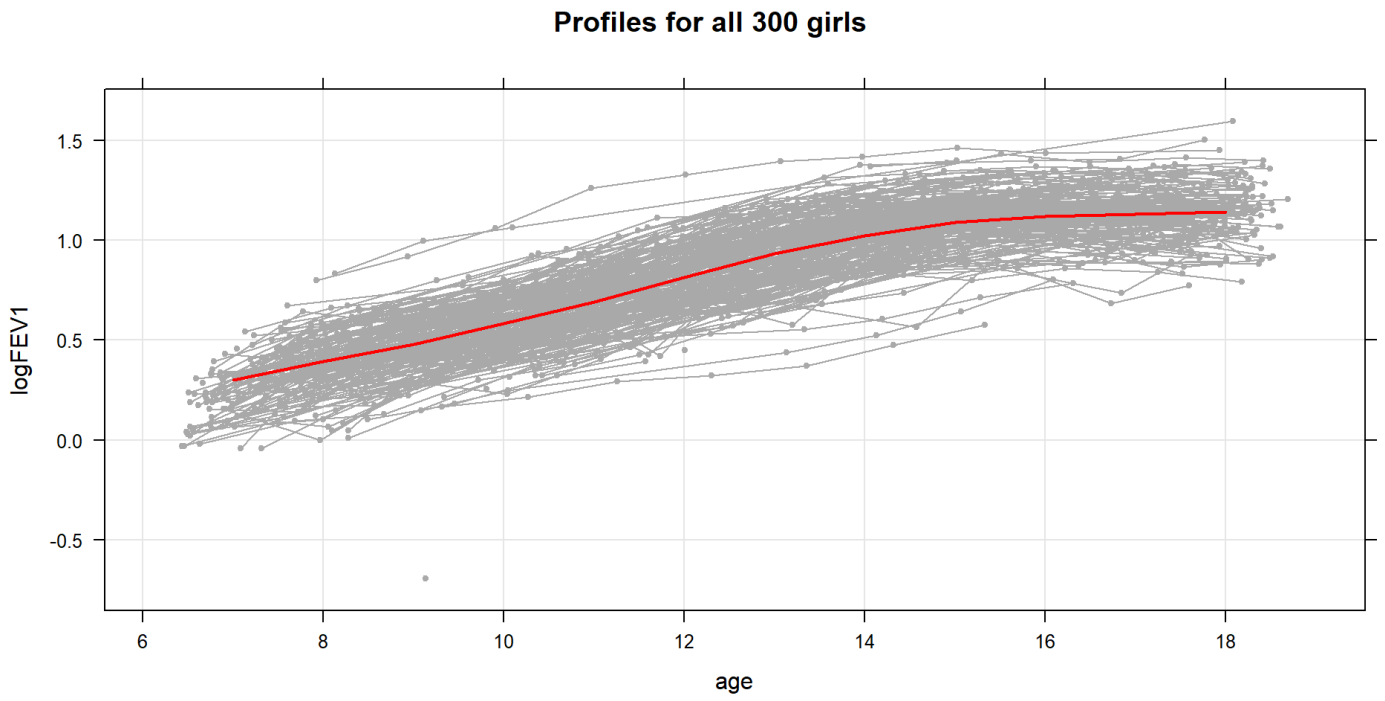
When we have unbalanced and/or irregular design, calculation of sample mean profile or standard deviation profile directly may not be possible. In this situation, one can adopt a **moving window** based approach. Specifically, we may combine responses taken at similar times to calculate our mean or standard deviation. For a given time point t, we then consider the time interval $[t - w/2, t + w/2]$ and calculate mean and standard deviation based on the responses whose observation times are within this interval. The *window width* is subjectively; usually, taken to be just large enough to give us enough data to obtain a reasonable estimate. We then move the time point t along the time domain.

To observe this method, consider the six cities air pollution data:

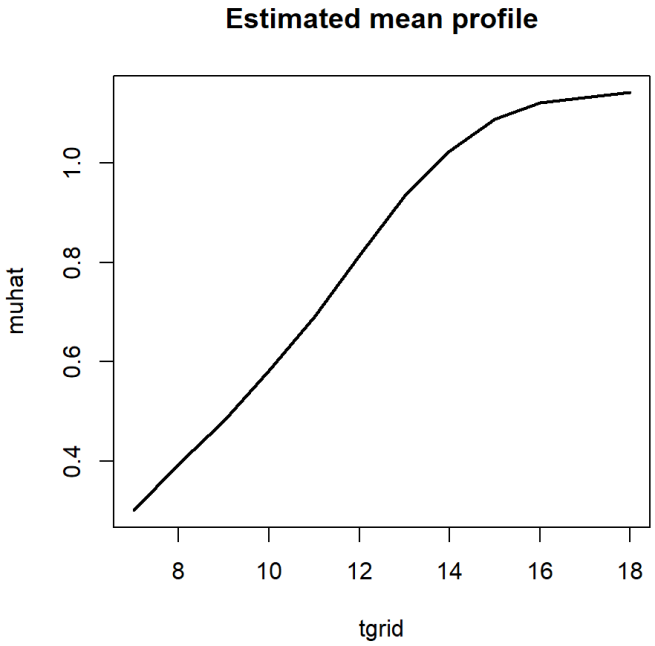
```
tab <- read.table("data/fev1-data.txt", header = F)
colnames(tab)=c("id", "Height", "age", "height_base", "age_base", "logFEV1")
head(tab)
```

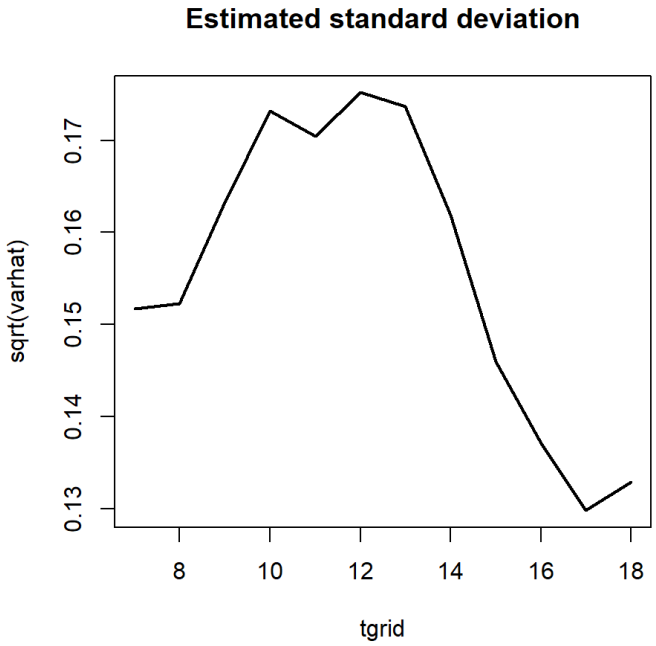
```
##   id Height     age height_base age_base logFEV1
## 1  1   1.20  9.3415         1.2   9.3415 0.21511
## 2  1   1.28 10.3929         1.2   9.3415 0.37156
## 3  1   1.33 11.4524         1.2   9.3415 0.48858
## 4  1   1.42 12.4600         1.2   9.3415 0.75142
## 5  1   1.48 13.4182         1.2   9.3415 0.83291
## 6  1   1.50 15.4743         1.2   9.3415 0.89200
```

```
# create a grid of age (time) to plot mean and sd
tgrid <- 7:18

# windo length
w <- 2

# mean and variance
muhat <- rep(NA, length(tgrid))
varhat <- rep(NA, length(tgrid))

for(ii in 1:length(tgrid)){
  tmp <- subset(tab, abs(age - tgrid[ii]) <= w/2)
  muhat[ii] <- mean(tmp$logFEV1)
  varhat[ii] <- var(tmp$logFEV1)
}

par(mfrow = c(1,2))
# plot original data and overlay mean
xyplot(logFEV1 ~ age, data=tab, groups = id, type = c("b", "g"), col="darkgrey", pch=19,
cex=0.5, main = "Profiles for all 300 girls") + layer(llines(tgrid, muhat, col="red", lw
d=2))
```



```
plot(tgrid, muhat, lwd=2, type="l", main = "Estimated mean profile")
plot(tgrid, sqrt(varhat), lwd=2, type="l", main = "Estimated standard deviation")
```





We can compute other quantities such as sample quantiles using similar technique as well.

Empirical summary and prediction plots

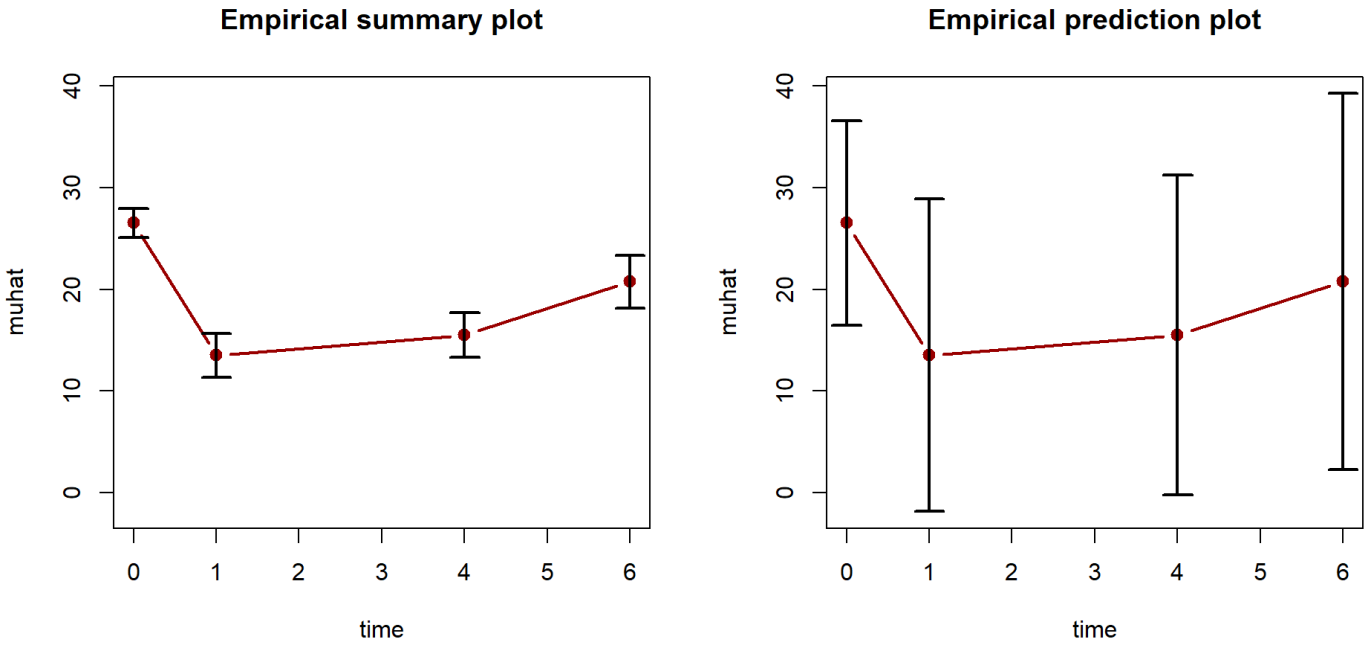
The empirical summary plot presents information (typically point-wise confidence intervals) about the mean profile.

For regular data, recall that the population mean of the j-th time point t_j can be estimated as $\widehat{\mu}_j = \bar{Y}_j$. Thus the standard error os this estimator is $SE(\bar{Y}_j) = \widehat{\sigma}_j/\sqrt{n}$. An **empirical summary plot** can be used to plot the estimated means along with lines depicting $\pm 2\sigma_j/\sqrt{n}$ to show approximate 95% confidence intervals.

Another useful plot is the **empirical prediction plot** where we plot the mean profile along with lines depicting $\pm 2\sigma_j$ to show approximate 95% prediction intervals. These intervals cover majority of the observed data.

For the TLC data, we show the plots below.

```
# Read the TLC data
tab <- read.table("data/lead-data.txt", header = F)
colnames(tab) <- c("ID", "Week 0", "Week 1", "Week 4", "Week 6")

# number of individuals
n <- nrow(tab)

# Time of measurements
time <- c(0, 1, 4, 6)

# Estimated mean
muhat <- colMeans(tab[, -1])

# Estimated variance
varhat <- apply(tab[, -1], 2, var)

# Lower/Upper bounds of the CI
lower <- muhat - 2*sqrt(varhat/n)
upper <- muhat + 2*sqrt(varhat/n)

# Lower/Upper bounds of the prediction
pred.lower <- muhat - 2*sqrt(varhat)
pred.upper <- muhat + 2*sqrt(varhat)

# Summary Plot
par(mfrow = c(1,2))
plot(time, muhat, type = "b", pch=19, main = "Empirical summary plot", ylim = c(min(pre
d.lower), max(pred.upper)), lwd=2, col = "#990000")
for(ii in 1:length(muhat)){
  arrows(time[ii], lower[ii], time[ii], upper[ii], angle=90, code = 3, length = 0.1, lwd
=2)
}

# Prediction plot
plot(time, muhat, type = "b", pch=19, main = "Empirical prediction plot", ylim = c(min(p
red.lower), max(pred.upper)), lwd=2, col = "#990000")
for(ii in 1:length(muhat)){
  arrows(time[ii], pred.lower[ii], time[ii], pred.upper[ii], angle=90, code = 3, length
 = 0.1, lwd=2)
}
```



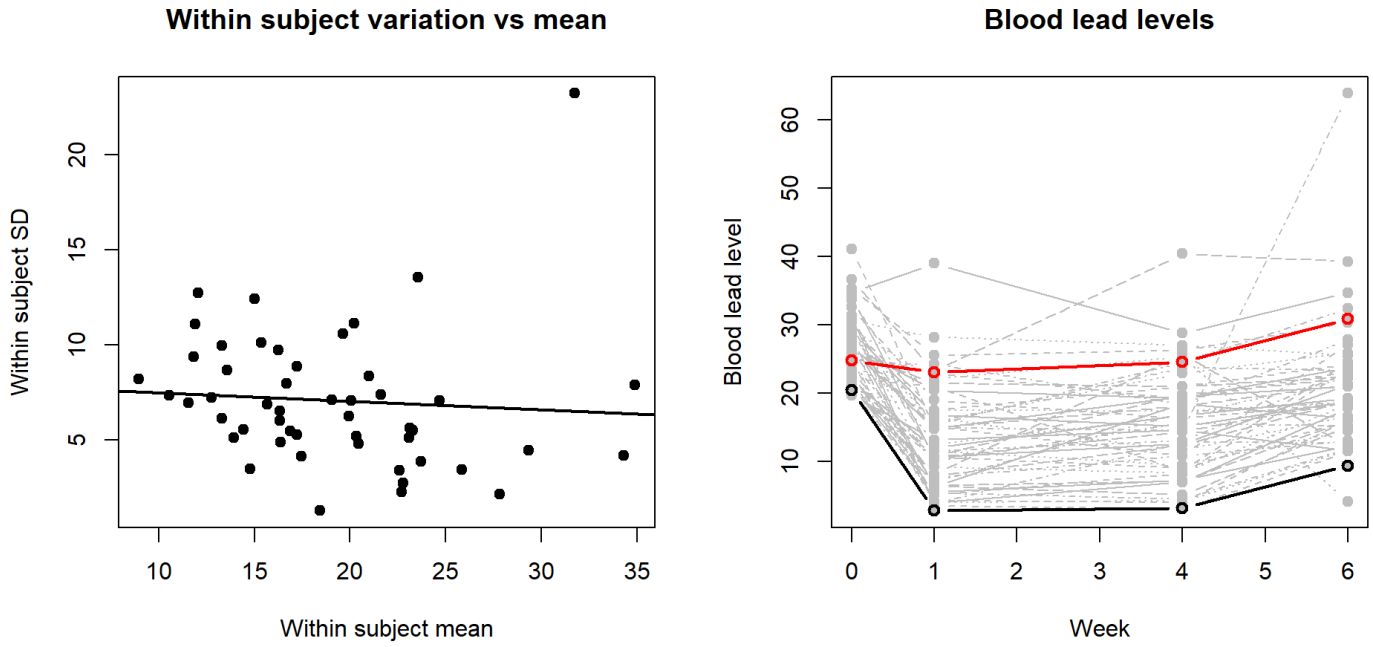
We can see that the prediction intervals are much wider than the condifence intervals. This is because the empirical summary plot is making inference about the *mean profile*, while the prediction intervals are trying to predict a *new observation* (not just the mean profile).

Within subject variability

The sample standard deviation (that we plotted before) measures the across-subject variability within each time point. Another quantity of interest is the *within subject* variability across time, that is, how much the response within each subject varies. We can examine this by plotting the within subject mean (in the x axis) and the within subject standard deviation (in the y axis).

For the TLC data, the plot is shown below (left panel).

```
within.mean <- rowMeans(tab[, -1])
within.sd <- apply(tab[, -1], 1, sd)

par(mfrow = c(1,2))
plot(within.mean, within.sd, pch=19,
     main  = "Within subject variation vs mean",
     xlab = "Within subject mean",
     ylab = "Within subject SD")

# trend line
abline( lm(within.sd ~ within.mean), lwd=2 )

# A few subjects higlighted
matplot(time, t(tab[, -1]), type="b", col="grey", pch=19,
        main = "Blood lead levels", xlab = "Week", ylab = "Blood lead level")
lines(time, tab[3,-1], col="black", lwd=2, type="b")
lines(time, tab[5,-1], col="red", lwd=2, type = "b")
```



The line represent the overall trend in the within subject SD as a function of within subject mean. It seems the within subject SD remains almost constant; a slight negative slope of the linear regression line indicates that subjects with higher mean tend to have lower SD. The range of the observations within person is low for the subjects with the highest values. Two specific subjects are highlighted in the plot in the right panel to show this phenomenon.

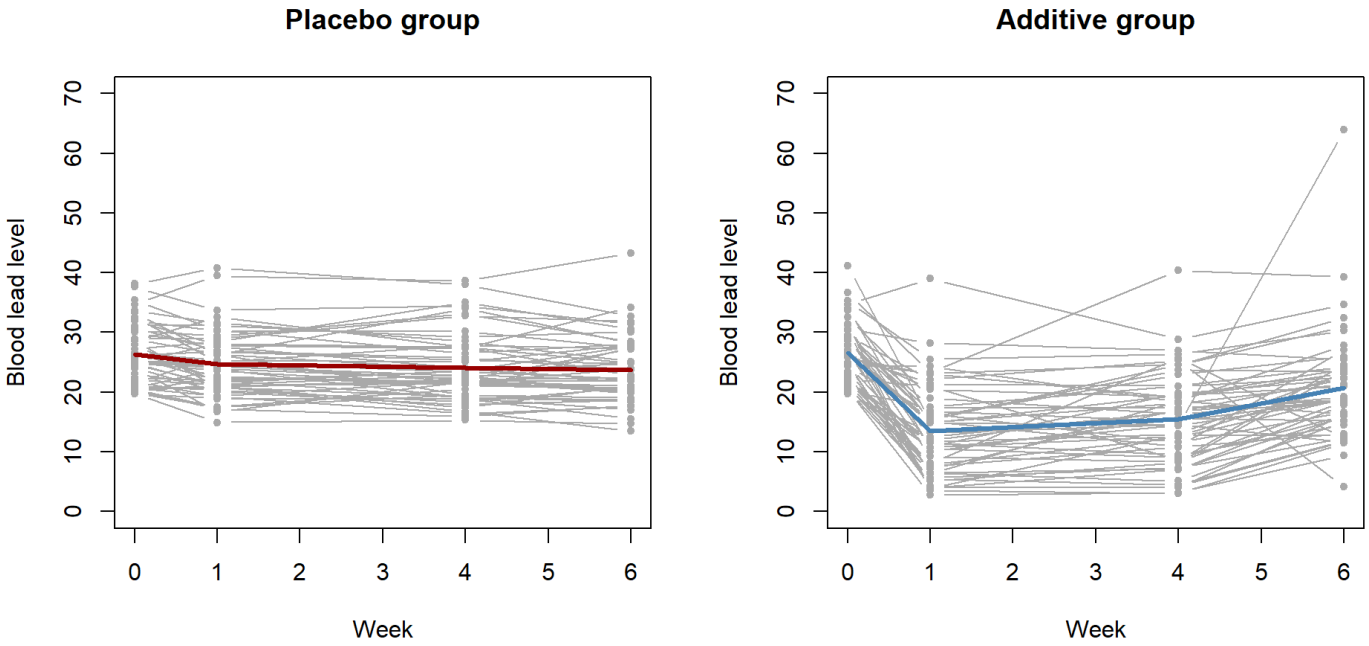
Sometimes such a plot would help us to determine whether to use the raw data for analysis or to use some transformation such as a `log` or `square-root` transformation first. Non-constant variance and skewness of the data are often related. So taking a transformation might help reduce both.

Profile plot with covariates

It is useful to try to incorporate the covariate information to enhance profile plots. When the covariate is discrete (e.g, a grouping factor), it is better to plot the profile for each level/value of the covariate in separate plots/panels.

For the TLC data for two groups (Placebo and additive) described above, we saw such plots already:

```
tab <- read.table("data/tlc-data.txt", header = F)
colnames(tab) <- c("ID", "Treatment", "Week 0", "Week 1", "Week 4", "Week 6")

# Time of measurements
time <- c(0, 1, 4, 6)

# Placebo group, only the lead level
placebo <- subset(tab, Treatment == "P")[, -c(1,2)]
mean.plc <- colMeans(placebo)

# Succimer group, only the lead level
additive <- subset(tab, Treatment == "A")[, -c(1,2)]
mean.add <- colMeans(additive)

# Plot
par(mfrow = c(1,2))
matplot(time, t(placebo), type = "b", ylim = c(0, 70),
        lty=1, pch=19, col = "darkgrey", cex=0.7,
        xlab = "Week", ylab = "Blood lead level",
        main = "Placebo group")
lines(time, mean.plc, col = "#990000", lwd=3)

matplot(time, t(additive), type = "b", ylim = c(0, 70),
        lty=1, pch=19, col = "darkgrey", cex=0.7,
        xlab = "Week", ylab = "Blood lead level",
        main = "Additive group")
lines(time, mean.add, col = "steelblue", lwd=3)
```

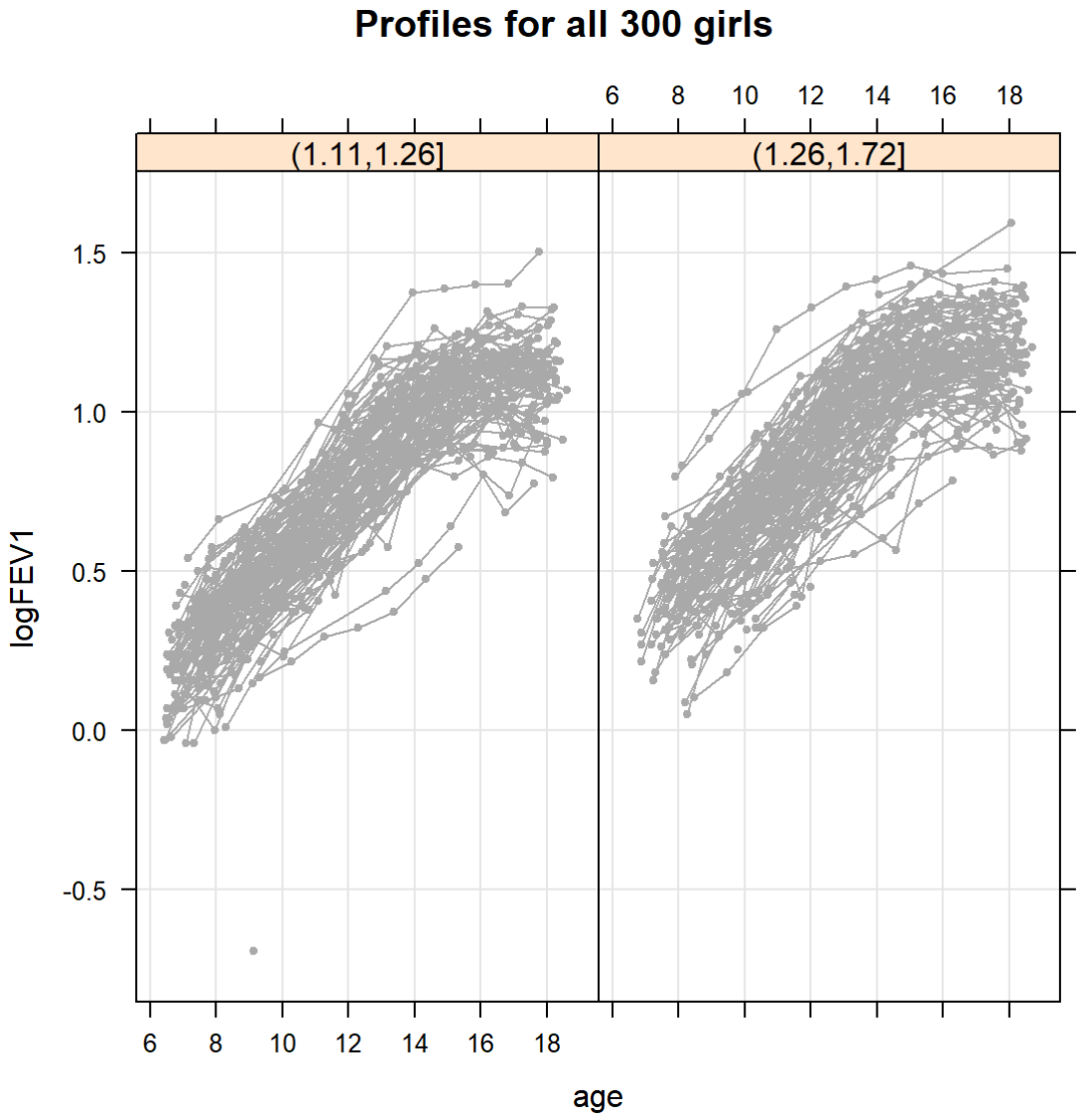


It is evedent from the plots that the placebo group does not show significant (at least visually) change in blood lead levels across time. However, the additive group seems to show reduction in blood lead levels at week 1.

When we have a continuous covariate, we can slice the covariate values into a few intervals and create a separate profile plot for subjects with covariate values that fall into each interval. If no prior knowledge is available as to how one should slice the covariate, we often use just two groups: covariate values below and above the median. This strategy is called the `median split`.

We demonstrate this strategy for the sixcities airpollution data, where we take the baseline height, `height_base`, as the covariate.

```
# Read the data
tab <- read.table("data/fev1-data.txt", header = F)
colnames(tab)=c("id", "Height", "age", "height_base", "age_base", "logFEV1")

# covariate
covar <- tab$height_base

# Median of baseling height
med <- median(covar)

# Creating cutpoints
cutp <- cut(covar,
            c(min(covar), med, max(covar))
            )
# xyplot
xyplot(logFEV1 ~ age | cutp, data=tab, groups = id, type = c("b", "g"), col="darkgrey",
 pch=19, cex=0.5, main = "Profiles for all 300 girls")
```



Not much difference is seen in the plot above between the two groups; perhaps the between subject varibility of the second group is slightly larger than that of the first group.

Correlation structure

As mentioned before, one needs to be careful on modeling the correlation among the measurements coming from a single subject. It is important to exmaine the correlation of the observed data to determine the type of model one would fit. Also, in the presence of a grouping factor, it is important ro see whether the correlation structure is different between the groups.

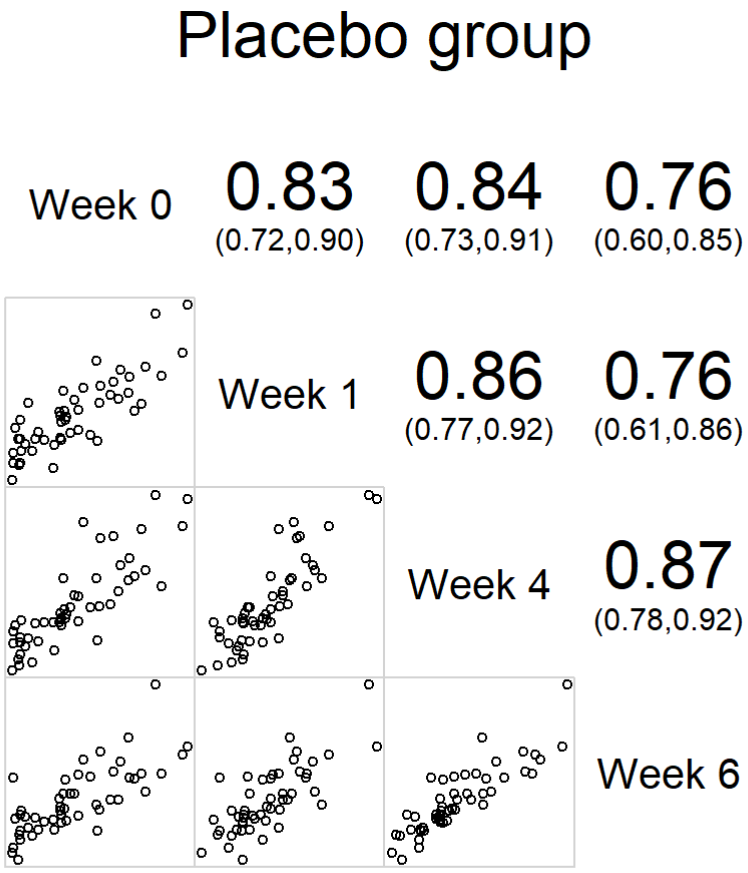
Instead of estimating the correlation ρ_{jk}, it is common to plot the association using a scatterplot matrix (pairs plot) of the standardized data: plot

$$\frac{Y_{ij}-\widehat{\mu}_j}{\widehat{\sigma}_j} \quad \text{vs.} \quad \frac{Y_{ik}-\widehat{\mu}_k}{\widehat{\sigma}_k}.$$

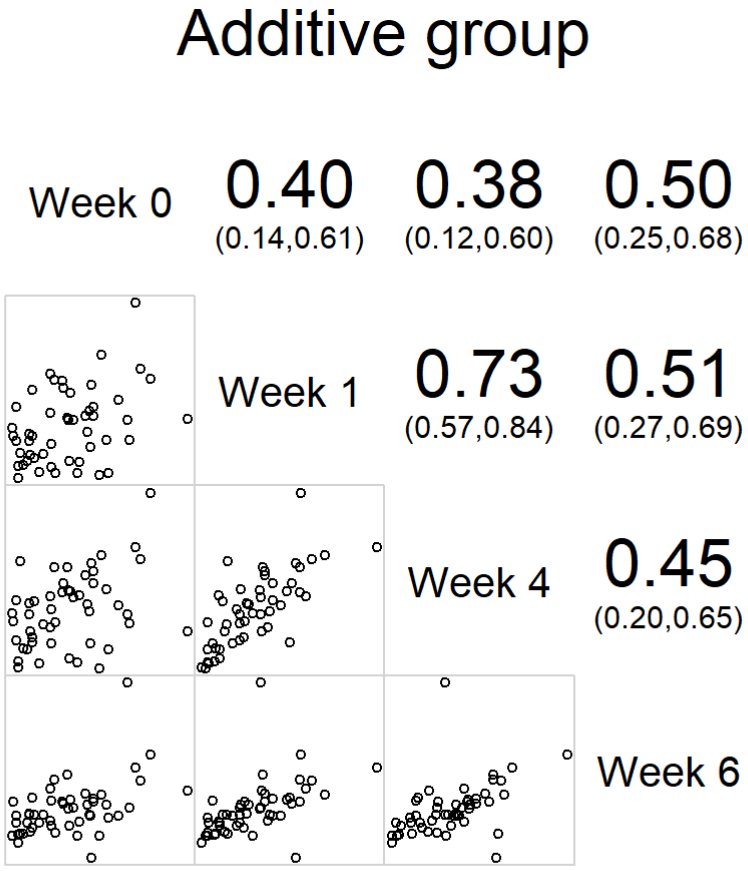
We examine whether the association seem constant across the pairs? the association seem to decay over time? or the association does not vary at all with time? Graphical display of the observations through scatterplot is used for detecting systematic features.

For the TLC data with both placebo and treatment groups, we show these plots and correlation structure below. You can do these using base R (as we have done many times in multivariate data analysis); we demonstrate another option using the `corrgram` package below.

```
library(corrgram)

# scaled responses
scale.placebo = scale(placebo)
scale.additive = scale(additive)

# plots
corrgram(scale.placebo, lower.panel=panel.pts, upper.panel=panel.conf)
mtext("Placebo group", side=3, cex=2, line = -1.5, outer=TRUE, xpd=NA)
```




```
corrgram(scale.additive, lower.panel=panel.pts, upper.panel=panel.conf)
mtext("Additive group", side=3, cex=2, line = -1.5, outer=TRUE, xpd=NA)
```



The option `lower.panel=panel.pts` creates the scatter plots in the lower part of the figures, and `upper.panel=panel.conf` estimates the correlation as well as associated 95% confidence interval.

It seems that the placebo group shows almost equal and high (about 80%) correlation between any two weeks. However, we see a different pattern in the additive group, where the correlations are generally lower except beteen weeks 4 and 6.

Summary

So far we have discussed basic summaries and plots to explore longitudinal data. These are typically used to investigate basic distribution and characteristics of the data.

Our primary goal is to perform inference about the mean response varying over time and to find out how/whether the mean varies as a function of additional covariates.

The exploratory plots help us choose what kind of models we would initially use to analyze the data. Later we will develop formal/quantitative ways to assess goodness-of-fit of various models.

Once we fit a specific model, we can recreate some of these plots (such as the experimental summary plot based on the estimated mean profile from the model fit) to see whether the fitted model is adequate or to check informally whether the model assumptions are reasonable.

Overall, the summary/exploratory plots can be used as an initial visualization tool; they *should not* be used for formal inference.

Main page: **ST 437/537: Applied Multivariate and Longitudinal Data Analysis (https://maityst537.wordpress.ncsu.edu/)**