ST 437/537: Applied Multivariate and Longitudinal Data Analysis

Longitudinal Data Analysis: Models for mean and covariance

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

Modeling longitudinal data is more complex than modeling independent data;

- need to model the correlation among the repeated measurements within each subject
- modeling the mean trend across time requires attention
- typically the effect of the various predictors is modeled in the mean (systematic part).

Conceptual model: For continuous data we write

$$Y_{ij} = \mu_j + e_{ij},$$

where μ_j is the average response corresponding time t_j and e_{ij} describes the deviation of the data Y_{ij} from the mean μ_j.

- The **mean describes how the response changes on average over time**. If additional factors (or covariate info such as group, additional subject information) are available then the mean may depend on these factors. Overall, the mean can be thought of as a function of time: $\mu_j = \mu(t_j)$.

- The **residual determines how far the data deviate from its mean**. It determines the distribution of the response (commonly assumed to be normal). It also determines how the repeated observations correlate over time.

The three main steps in modeling longitudinal data are:

- modeling the mean,
- modeling the covariance,
- and selecting the distribution of the data Y.

In each of these it is imperative that we *look at the data* (using techineues described in the **previous chapter (Lecture07_LDA_Introduction.html)**). We discuss below how to model the mean and covariance of the data. For now we will assume the response variable is continuous (a common assumption is to assume multivariate normality). We will discuss binary/count responses in later chapters.

Modeling the mean

Let us start with considering a balanced and regular design. Because the elements of the mean vector $\boldsymbol{\mu} = (\mu_1 \ldots, \mu_m)^T$ are arranged in time increasing order $t_1 < t_2 < \ldots < t_n$ (μ_j corresponds to t_j) , we refer to the mean μ as a **mean trajectory** instead and less as a mean vector (as is common in multivariate statistics). We often represent the mean trajectory using a finite set of parameters, that is, we put a parametric structure on the mean function.

Example A: Dental study [The orthodontic study data of Potthoff and Roy (1964).]

Source: Potthoff, R. F. and Roy, S. N. (1964), “A generalized multivariate analysis of variance model useful especially for growth curve problems”, Biometrika, 51, 313–326.

Researchers are interested in the development of children over time. They collect dental growth measurements of the distance (mm) from the center of the pituitary gland to the pterygomaxillary fissure for 27 children (11 girls and 16 boys) at ages 8, 10, 12, and 14. A picture of the pterygomaxillary fissure can be found at **http://www.stat.ncsu.edu/people/davidian/courses/st732/examples/pterygomaxillary-**

fissure.jpg (http://www.stat.ncsu.edu/people/davidian/courses/st732/examples/pterygomaxillary-fissure.jpg). The interest is in

- how the dental growth measurements vary over time,
- if they are different in boys and girls, and
- if the rate of change is different for boys than girls.

The dataset is available as `Orthodont` in the `nlme` package.

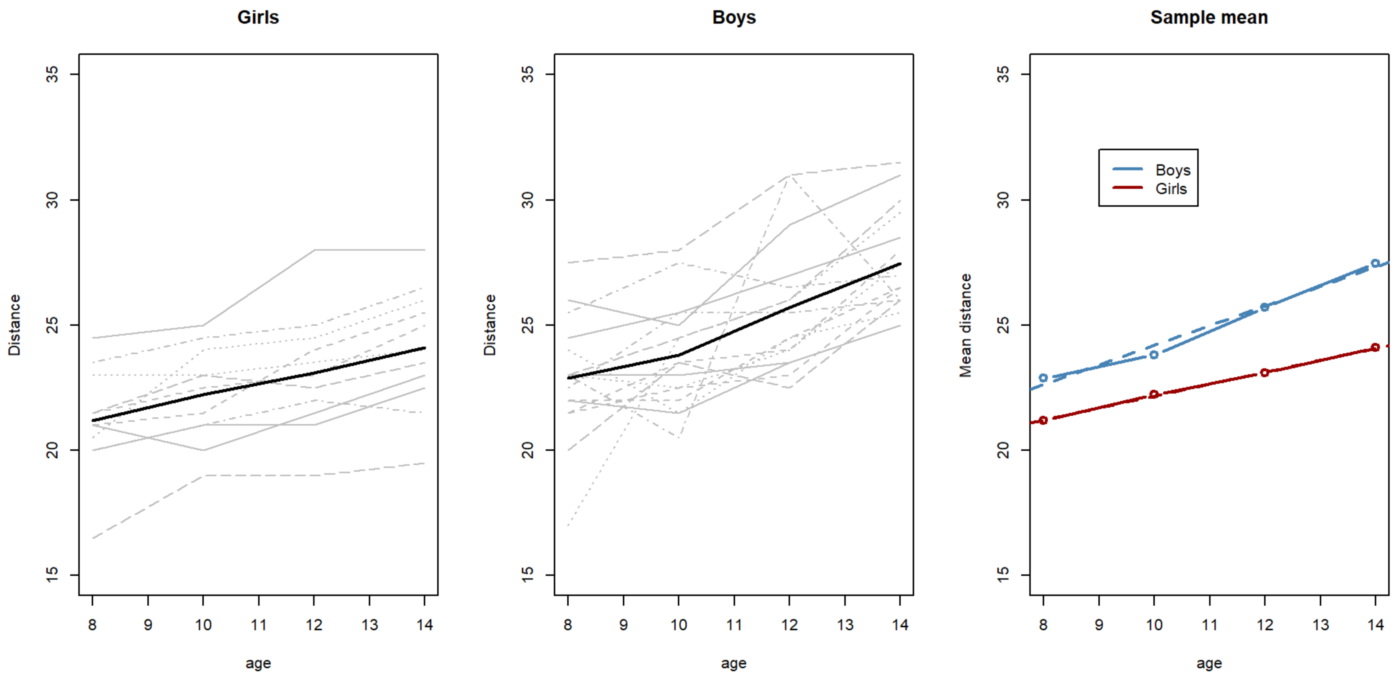
```
library(nlme)
head(Orthodont)
```

```
## Grouped Data: distance ~ age | Subject
##   distance age Subject  Sex
## 1     26.0   8     M01 Male
## 2     25.0  10     M01 Male
## 3     29.0  12     M01 Male
## 4     31.0  14     M01 Male
## 5     21.5   8     M02 Male
## 6     22.5  10     M02 Male
```

```
tail(Orthodont)
```

```
## Grouped Data: distance ~ age | Subject
##     distance age Subject    Sex
## 103     19.0  12     F10 Female
## 104     19.5  14     F10 Female
## 105     24.5   8     F11 Female
## 106     25.0  10     F11 Female
## 107     28.0  12     F11 Female
## 108     28.0  14     F11 Female
```

```
# Subset girls/boys (only take data and ID)
girls <- subset(Orthodont, Sex == "Female")[, 1:3]
boys <- subset(Orthodont, Sex == "Male")[, 1:3]

# Balance and regular data, so put in a matrix form for easy access
mat.Gr <- matrix(girls$distance, ncol=4, byrow = T)
mat.By <- matrix(boys$distance, ncol=4, byrow = T)
age = c(8, 10, 12, 14)

# Sample mean trajectories
muhat.girls <- colMeans(mat.Gr)
muhat.boys <- colMeans(mat.By)

# plot
par(mfrow = c(1,3))
matplot(age, t(mat.Gr), type="l", ylim = c(15, 35),
        col="grey", ylab = "Distance", main = "Girls")
lines(age, muhat.girls, lwd=2)

matplot(age, t(mat.By), type="l", ylim = c(15, 35),
        col="grey", ylab = "Distance", main = "Boys")
lines(age, muhat.boys, lwd=2)

# only means and linear approximation
plot(age, muhat.girls, type = "b", lwd=2, ylim = c(15, 35), col="#990000", ylab = "Mean distance", main = "Sample mean")
abline( lm(muhat.girls~age), col="#990000", lwd=2, lty=2 )

lines(age, muhat.boys, type = "b", lwd=2, col="steelblue")
abline( lm(muhat.boys~age), col="steelblue", lwd=2, lty=2 )

legend(9, 32, legend = c("Boys", "Girls"), lwd=2, col = c("steelblue", "#990000"))
```



The plots in the left and middle panel plots the profiles for girls and boys, respectively and their sample mean trajectory (black solid lines). The right panel only plots the sample mean trajetories with the best linear approximation, that is, $\mu(age) = a + b(age)$.

The plots strongly suggests that the sample mean trajectories for both the groups are verywell approximated by a straight line. Overall, we observe the following:

- Each groups has a linear trend in the their mean trajectory,
- The lines have different intercepts for each group, and
- The lines have different slopes for each group.

Thus for a single group (say girls), it is reasonable to model the *mean function* as

$$\mu_G(t_j) = \beta_0 + \beta_1 t_j,$$

where the subject G indicates that we are modeling the mean trajectory of the girls; t_j now denotes the age corresponding to the j-th measurement for the i-th girl. Thus we have

$$\text{Girls group: } Y_{ij} = \beta_0 + \beta_1 t_j + e_{ij},$$

where e_{ij} the random deviations. Similarly, write a mean model for the boys group:

$$\text{Boys group: } Y_{ij} =$$

keeping in mind the observations we made from the plots.

Ideally, we do not want to fit models to each group one-at-a-time. Thus we would like to combine both mean models into one function. There are various ways of doing so; we present two such formulations below. Both these approaches involve defing a “dummy variable”: for $i = 1, \dots, n$, we define

$$G_i = \begin{cases} 1 \text{ if the } i\text{-th child is a girl} \\ 0 \text{ if the } i\text{-th child is a boy} \end{cases}.$$

Formulation 1

Formulation 2 (modeling the differences)

Both these formulations are equivalent; they just lead to different interpretation of the parameters involved in the model.

We can easily extend this idea to the case where the design is irregular and/or unbalanced. Consider the following example.

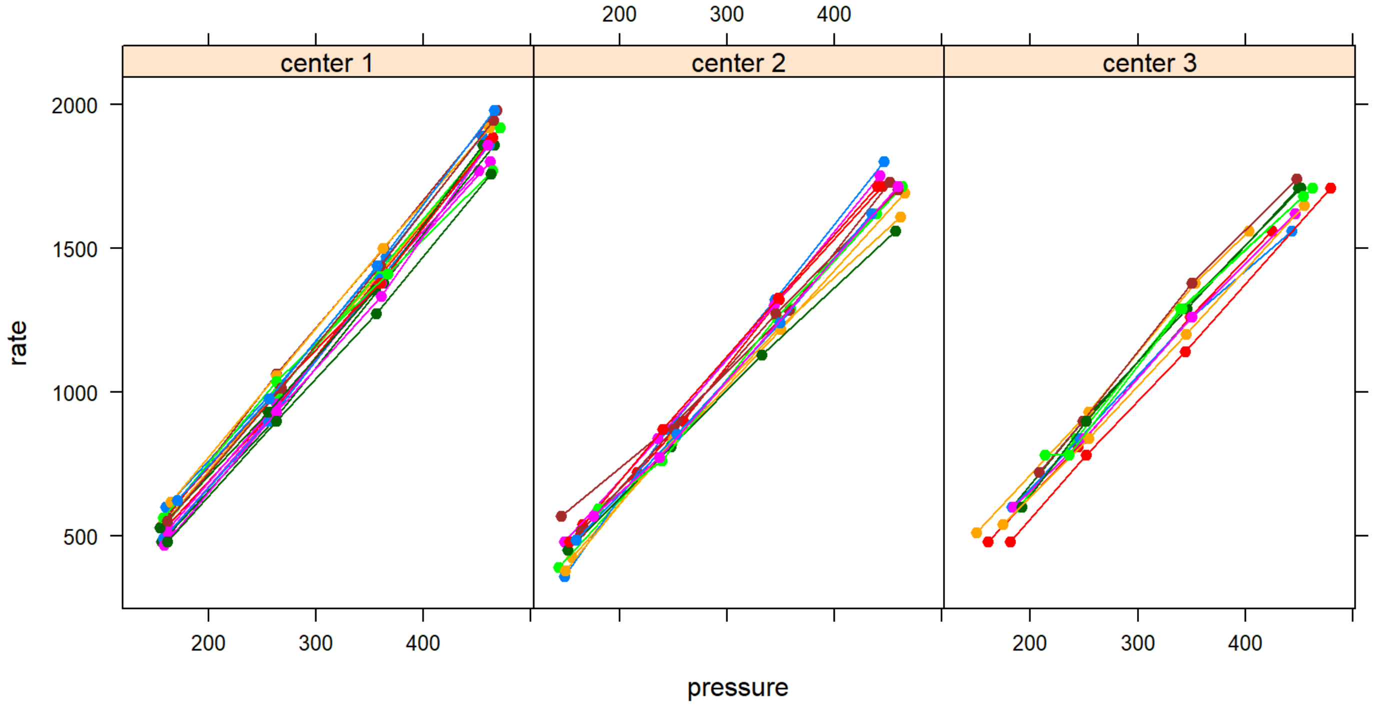
Example B: Dialysis study (Ultrafiltration Data For Low Flux Dialyzers presented in Vonesh and Chinchilli, 1997)

Source: Vonesh and Carter (1987). Efficient inference for random coefficient growth curve model with unbalanced data. Biometrics, 43, 617—628. The datafile is available at **https://www.stat.ncsu.edu/people/davidian/courses/st732/#examples (https://www.stat.ncsu.edu/people/davidian/courses/st732/#examples)** under the name “ultrafiltration data”.

Low flux dialyzers are used to treat patients with end stage renal disease to remove excess fluid and waste from their blood. In low flux hemodialysis, the **ultrafiltration rate (ml/hr)** at which fluid is removed is thought to follow a straight line relationship with the **transmembrane pressure (mmHg)** applied across the dialyzer membrane. A study was conducted to compare the average ultrafiltration rate (the response) of such dialyzers across **three dialysis centers** where they are used on patients. A total of 41 dialyzers (units) were involved. The experiment involved recording the ultrafiltration rate at 4 transmembrane pressures (depicted by dots in the Figure below) for each dialyzer.

```
library(latticeExtra)
ultra <- read.table("data/ultra.dat", header = F)
colnames(ultra) <- c("Id", "pressure", "rate", "center")
ultra$center.factor <- factor(ultra$center, levels = 1:3, labels = paste("center", 1:3))
xyplot(rate ~ pressure | center.factor, data = ultra, groups = Id, type="b", pch=19, main
 = "Ultrafiltration rate vs transmembrane pressure for three centers")
```

Ultrafiltration rate vs transmembrane pressure for three centers



Is this a longitudinal data? Justify your answer:

- what is the observational unit?

- what is the `time' variable?

Clearly, the 4 pressure levels (=`time') at which each dialyzer was observed are not necessarily the same, that is, the design is irregular. However, just observing the profile plots, we see that each profile (and thus their average) shows a almost linear trend. In this case too, we can model the mean trajectory as a linear function of pressure level. However, we need to account for the fact that each dialyzer has their own pressure levels. We use the following model specification:

$$\begin{aligned}\text{center 1: } Y_{ij} &= \beta_1 + \beta_2 t_{ij} + e_{ij}\\ \text{center 2: } Y_{ij} &= \beta_3 + \beta_4 t_{ij} + e_{ij}\\ \text{center 3: } Y_{ij} &= \beta_5 + \beta_6 t_{ij} + e_{ij}\end{aligned}$$

Again, we would like to combine the three separate mean functions into one. Since we have three centers, we need two dummay variables. For the i-th dialyzer (of the entire sample)

$$C_{1i} = \begin{cases} 1 & \text{if the } i\text{-th dialyzer is in center 1}\\ 0 & \text{otherwise}\end{cases}$$

$$C_{2i} = \begin{cases} 1 & \text{if the } i\text{-th dialyzer is in center 2}\\ 0 & \text{otherwise}\end{cases}$$

We do not need the third dummy variable since a dialyzer would be in center 3 if $C_{1i} = 0$ and $C_{2i} = 0$. With these two dummy variabled defined, we are essentially using the third center (center 3) as our baseline. We can write the combined model: for the i dialyzer,

$$Y_{ij} = \eta_1 + C_{1i}\eta_2 + C_{2i}\eta_3 + t_{ij}\eta_4 + C_{1i}t_{ij}\eta_5 + C_{2i}t_{ij}\eta_6 + e_{ij}.$$

Let us now interpret the parameters $\eta_1, \ldots, \eta_6$ and compare them to the original parameters, $\beta_1, \ldots, \beta_6$. Recall that each center has its own intercept and slope.

For dialyzers in **center 3** (that is, those values of i where $C_{1i} = 0$ and $C_{2i} = 0$), the mean trajectory is

$$E(Y_{ij}) = \eta_1 + \eta_4 t_{ij}.$$

Thus η_1 and η_4 are the intercept and slope for center 3, respectively.

For dialyzers in **center 1** (that is, those values of i where $C_{1i} = 1$ and $C_{2i} = 0$), the mean trajectory is

$$E(Y_{ij}) = (\eta_1 + \eta_2) + (\eta_4 + \eta_5)t_{ij}.$$

Thus $(\eta_1 + \eta_2)$ and $(\eta_4 + \eta_5)$ are the intercept and slope for center 1, respectively. Thus η_2 denotes the **change in intercept** between center 1 and center 3, and η_5 denotes the **change in slope** between center 1 and center 3.

Similarly for **center 2** (that is, those values of i where $C_{1i} = 0$ and $C_{2i} = 1$), the mean trajectory is

$$E(Y_{ij}) = (\eta_1 + \eta_3) + (\eta_4 + \eta_6)t_{ij}.$$

Thus $(\eta_1 + \eta_3)$ and $(\eta_4 + \eta_6)$ are the intercept and slope for center 2, respectively. Thus η_3 denotes the **change in intercept** between center 2 and center 3, and η_6 denotes the **change in slope** between center 2 and center 3.

In summary, the formulation above, **directly models the change** in intercept and slope paraneter between centers 1 and 2 from those of center 3 (which is used as the baseline).

This is simply a linear regression with six covariates: intercept, main effects of C_{1i}, C_{2i} and t_{ij}, and interaction terms between C_{1i}, C_{2i} and t_{ij}.

Let us fit a linear regression model (altough we have not discussed about possible covariance models) to the data just to visualize the ideas we discussed so far. We can simply use `lm` function to do so. **We want to only get a point estimate of the mean trajectories, NOT make any inference since we have not modeled the covariance properly yet.**

```
Y <- ultra$rate
time <- ultra$pressure
C1 <- (ultra$center == 1)
C2 <- (ultra$center == 2)

# Fit the LS model
out <- lm(Y ~ C1 + C2 + time + C1:time + C2:time)
summary(out)
```

```
## 
## Call:
## lm(formula = Y ~ C1 + C2 + time + C1:time + C2:time)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -149.77  -33.22    1.09   35.95  137.97 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -148.03509   25.65224  -5.771 4.05e-08 ***
## C1TRUE       -27.09087   31.91040  -0.849 0.397184    
## C2TRUE       -20.73469   33.27797  -0.623 0.534133    
## time           4.05534    0.07958  50.958  < 2e-16 ***
## C1TRUE:time    0.35648    0.09810   3.634 0.000377 ***
## C2TRUE:time    0.05975    0.10382   0.576 0.565766    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 53.19 on 158 degrees of freedom
## Multiple R-squared:  0.9876, Adjusted R-squared:  0.9872 
## F-statistic:  2510 on 5 and 158 DF,  p-value: < 2.2e-16
```

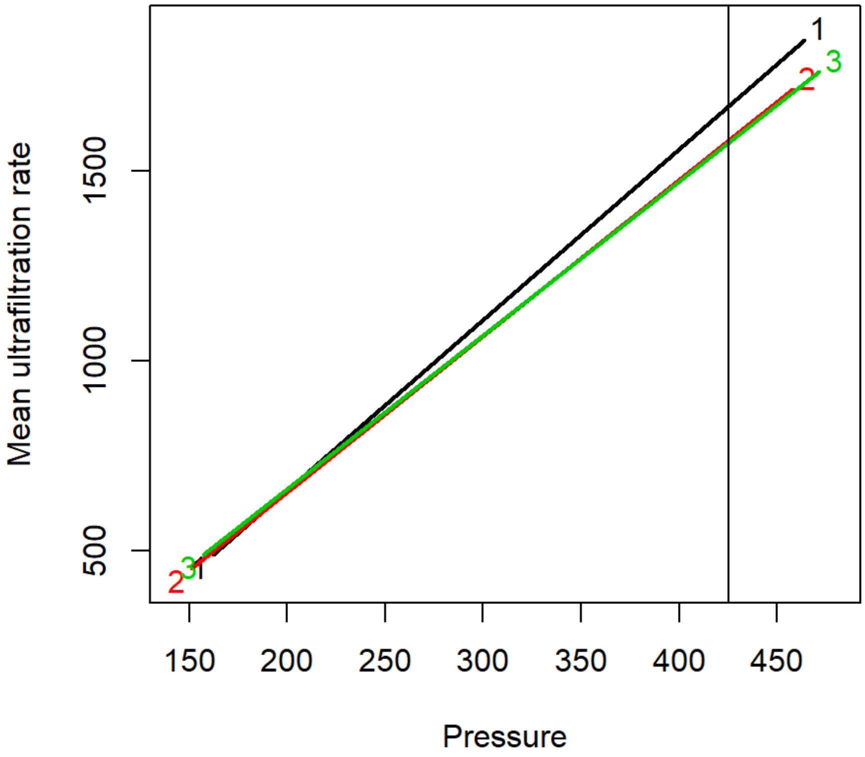
The six rows in the output table `Coefficients` (Intercept – C2:time) correspond to η_1 – η_6, respectively. **Again, do NOT pay attention to p-values as we have not properly modeled the covariance yet; just examine the point estimates.**

From the output, the mean trajectory for center 3 (baseline) is $-148.05 + 4.05t.$.

The coefficients corresponding C1TRUE (-27.09) and C1TRUE:time (0.36) denote the change in intercept and slope between center 1 and center 3. Thus the mean trajectory for center 1 (baseline) is $(-148.05 - 27.09) + (4.05 + 0.36)t = -175.14 + 4.41t.$

A plot of the three estimated mean trajectories is shown below.

```
# Coefficients
eta <- out$coefficients

# time grid to plot (only need two points since
# the mean is a straight line)
# Since we do not want to extrapolate, we will take
# the minimum and maximum times (pressure) for each center
tgrid <- matrix(NA, 2, 3)
for(ii in 1:3){
  tgrid[,ii] <- range(ultra$pressure[ultra$center==ii])
}

line <- matrix(NA, 2, 3)
# Center 3
cf <- c(eta[1], eta[4])
line[, 3] <- cf[1] + tgrid[, 3]*cf[2]

# Center 1
cf <- c(eta[1], eta[4]) + c(eta[2], eta[5])
line[, 1] <- cf[1] + tgrid[, 2]*cf[2]

# Center 2
cf <- c(eta[1], eta[4]) + c(eta[3], eta[6])
line[, 2] <- cf[1] + tgrid[, 2]*cf[2]

# Plot
matplot(tgrid, line, lwd=2, lty=1, pch = as.character(1:3), type="b",
        ylab = "Mean ultrafiltration rate", xlab = "Pressure")
abline(v = 425)
```



It seems that at higher pressure levels center 1 dialyzers have somewhat larger mean ultrafiltration rate compared to centers 2 and 3. For example, compared to center 3, the mean ultrafiltration rate in center 1 for a pressure level of $t = 425$ is $\hat{\eta}_2 + 425\hat{\eta}_5 = 124.413868$ units **higher**.

In contrast, compared to center 3, the mean ultrafiltration rate in center 2 for a pressure level of $t = 425$ is only $\hat{\eta}_3 + 425\hat{\eta}_6 = 4.6578787$ units higher.

Polynomial mean models

In general, depending on the data at hand, we can posit other polynomial models for mean as well. For example, a **quadratic trend over time** can be represented as

$$E[Y_{ij}] = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2.$$

In presence of groups (e.g., centers) or other continuous covariates (e.g., age), one can easily include more terms to this model including interaction terms (and higher order terms) as well.

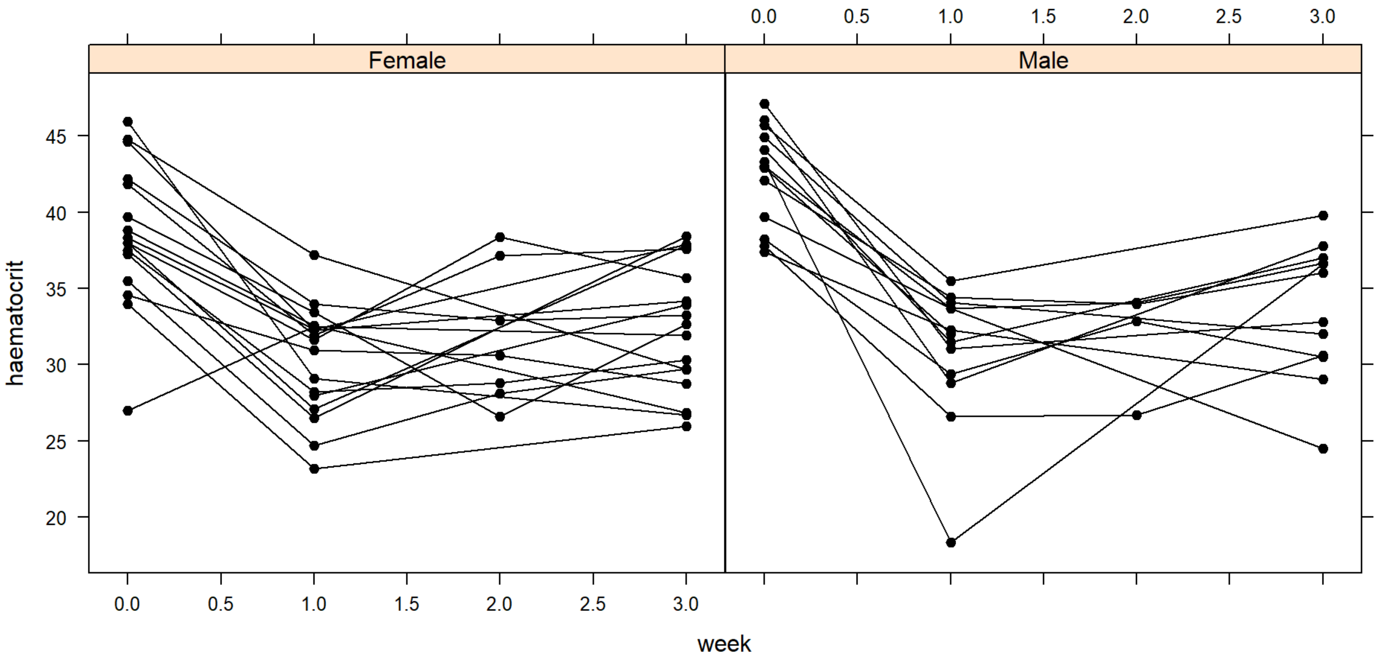
Example C: Hip replacement study.

These data are adapted from Crowder and Hand (1990, section 5.2). 30 patients (13 males and 17 females) underwent hip-replacement surgery. Haematocrit, the ratio of volume packed red blood cells relative to volume of whole blood recorded on a percentage basis, was supposed to be measured for each patient at week 0, before the replacement, and then at weeks 1, 2, and 3, after the replacement. In addition the age of each participant is recorded. The primary interest was to determine whether there are possible differences in mean response following replacement for men and women. Spaghetti plots of the profiles for each patient are shown in the left hand panels of Figure 3. (We will discuss the right-hand panels later.). It may be seen from the figure that a number of both male and female patients are missing the measurement at week 2; in fact, there is one female missing the pre-replacement measurement and week 2. Here, we have a situation where the data vectors Y_i are of possibly different lengths for different units.

```
hip <- read.table("data/hips.dat", header = F)
colnames(hip) <- c("id", "sex", "age", "week", "haematocrit")
head(hip)
```

```
##   id sex age week haematocrit
## 1  1   1  66    0       47.10
## 2  1   1  66    1       31.05
## 3  1   1  66    3       32.80
## 4  2   1  70    0       44.10
## 5  2   1  70    1       31.50
## 6  2   1  70    3       37.00
```

```
hip$sex.factor <- factor(hip$sex, levels = c(0,1), labels = c("Female", "Male"))

xyplot(haematocrit ~ week | sex.factor, data = hip, groups = id,
       pch=19, type="b", col="black")
```



In this example, fitting a straight line for mean trajectory would be inappropriate. Also, the the covariate `age` might impact the mean haematocrit as well. A possible model for this model is given below:

$$\begin{aligned}\text{Male: } & E(Y_{ij}) = \beta_1 + \beta_2 t_{ij} + \beta_3 t_{ij}^2 + \beta_4 a_i\\ \text{Female: } & E(Y_{ij}) = \beta_5 + \beta_6 t_{ij} + \beta_7 t_{ij}^2 + \beta_8 a_i,\end{aligned}$$

where a_i denotes the `age` of the individual. Here the covariate `age` has a linear effect on the mean trajectory.

> *Exc: How do we write one combined model in this example?*

Linear splines

In some applications the longitudinal trends in the mean response cannot be characterized by a simple polynomial (first or second) in time. In some application the trend cannot be well represented by polynomials in time of any order. This will most occur when the mean response increases (or decreases) rapidly for some duration, and then more slowly thereafter (or vice versa). When this type of change pattern occurs the mean trend can be modeled by spline models.

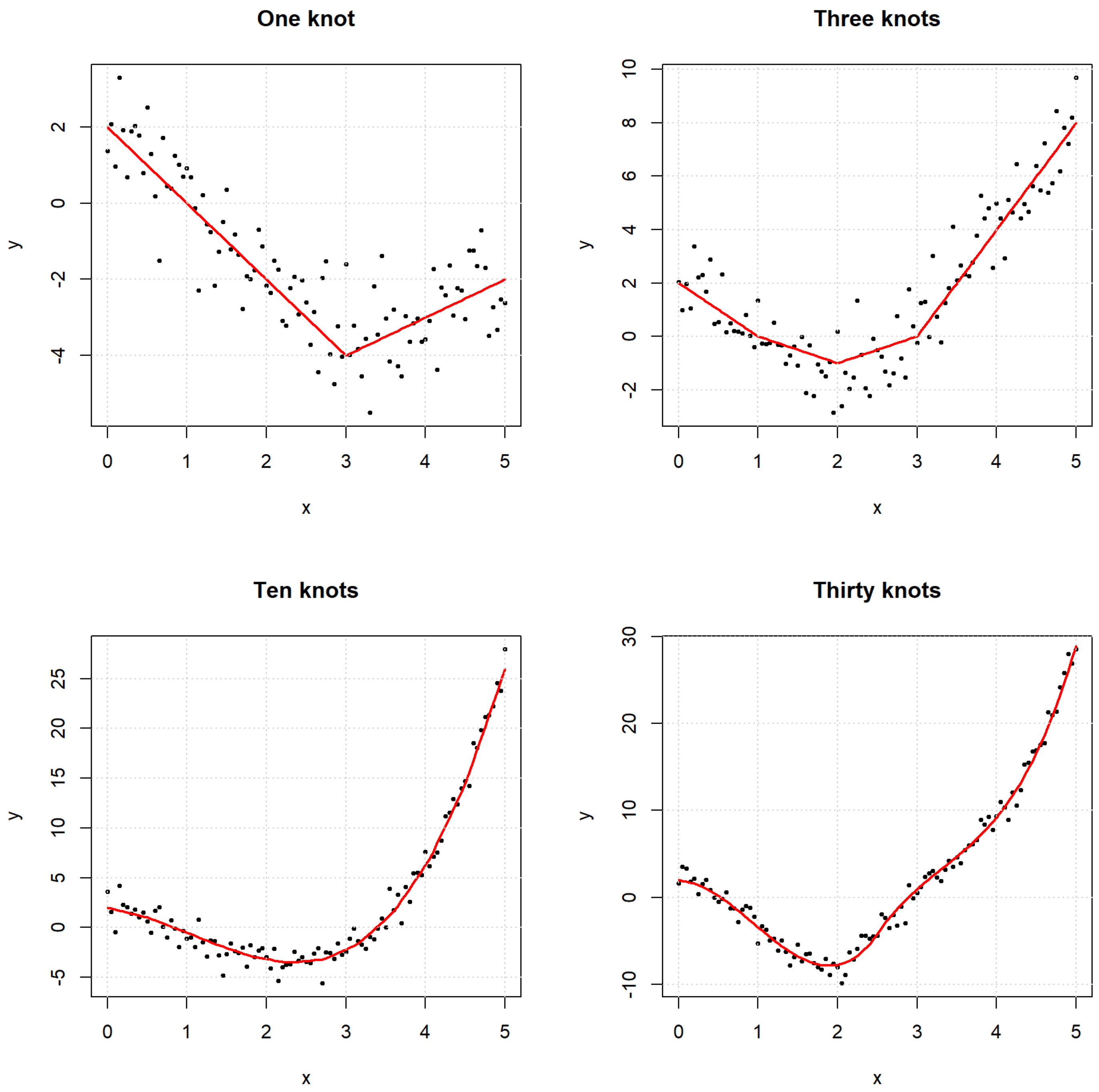
In a nutshell, a spline regression model involves a linear combination of connected or joined piecewise polynomial functions. Splines are defined by *degree* and *knots*. A linear (quadratic, cubic etc.) spline means that the joined polynomials are lines (quadratic functions or cubic functions etc.). Knots are the locations at which the lines meet or are tied together. Linear spline models provide a useful and flexible way to model non-linear trends that cannot be approximated by simple polynomial functions in time.

We defined earlier polynomial models as linear combinations of the power basis functions, $\{1, t, t^2, \dots\}$. Linear spline models rely on the same general idea, except the basis functions are of the form $\{1, t, (t-\kappa_1)_+, \dots, (t-\kappa_k)_+\}$ where $\{\kappa_1, \dots, \kappa_k\}$ are *knots* and k is the number of knots. Here $(x)_+ = x$ if $x > 0$ and 0 if $x \leq 0$.

A linear spline model (using a single knot κ) for the mean trend can be represented as:

$$E[Y_{ij}] = \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij} - \kappa)_+.$$

Depending on the number of knots, linear (and higher order) splines can capture flexible trends in the mean trajectory. Examples of mean trend captured by a linear spline with varying number of knots are shown below. The mean trend is depicted in red solid line while the observed data is shown in black circles.



One knot

Three knots

Ten knots

Thirty knots

The linear spline model is also called a *piece-wise* linear model. Regarding knots, we want to place knots where the function changes most rapidly. As an automatic procedure, we might put equally spaced knots throughout the range of time.

We can also extend this method to incorporate higher order (quadratic or cubic) splines. There are many other choices for basis functions for modeling a mean trend. For example, B-spline functions are one of the most popular (and computationally more efficient) way to model a function.

Mean model in vector form

In general, recall that the response vector for the i-th individual is $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T$. We denote the covariate matrix as $\boldsymbol{X}_i$ (an $m_i \times p$ matrix, where p is the number of covariates). In general, $\boldsymbol{X}_i$ could include intercept (a column of 1's), time effects, and any other covariates like age of dummy variables etc.

In our Hip replacement study, we consider the model for the male group,

$$\text{Male: } E(Y_{ij}) = \beta_1 + \beta_2 t_{ij} + \beta_3 t_{ij}^2 + \beta_4 a_i.$$

This model can be written as

$$E(\boldsymbol{Y}_i) = \boldsymbol{X}_i \beta,$$

where

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & t_{i1} & t_{i1}^2 & a_i \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{im_i} & t_{im_i}^2 & a_i \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}.$$

Thus we can write a linear model for the i-subject:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i \beta + \boldsymbol{e}_i,$$

where $\boldsymbol{e}_i$ are random errors.

> ***We typically assume that*** $\boldsymbol{e}_i \sim N(0, \boldsymbol{\Sigma}_i)$***,*** *where* $\boldsymbol{\Sigma}_i$ *is an unknown* $m_i \times m_i$ *covariance matrix.*

So far we have discussed how to model the mean of $\boldsymbol{Y}_i$, that is to get a proper design matrix $\boldsymbol{X}_i$. Next we will discuss how to model the covariance matrix $\boldsymbol{\Sigma}_i$ properly.

Models (marginal models) for the covariance

Recall that the observed data is $\{Y_{ij}, X_{ij} : j = 1, \ldots m_i\}$; let $\boldsymbol{Y}_i$ be the m_i dimensional vector of Y_{ij}'s and X_i be $m_i \times p$ dimensional design matrix (e.g. could include 1's or t_{ij}'s or t_{ij}^2 or other covariates observed for subject i or time-varying covariates etc.). Thus we can write a linear model for the i-subject:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{e}_i,$$

where $\boldsymbol{e}_i$ are random errors.

Assume that

$$\mathrm{cov}(\boldsymbol{e}_i) = \boldsymbol{\Sigma}_i,$$

where $\boldsymbol{\Sigma}_i$ is a covariance matrix. Here the index i is used specifically to allow for different number of repeated measurements per unit n_i. In this part we assume that the covariance model is parametric, that is $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_i(\omega)$ – it is known up to a lower dimensional parameter ω.

Recall the responses measured on the same unit/subject are correlated. Although the correlations, or more generally the covariance, among the repeated responses is not usually of particular interest, we need to account for it in making inferences for the mean parameters. Accounting for the correlations among the repeated measures completes the specification of a (normal) model for the longitudinal data and usually increases precision with which the regression parameters are estimated.

There are three main approaches to describe the covariance among the repeated measures:

1. unstructured;
2. covariance pattern models (to be described below); and
3. random effects covariance models (to be discussed later in the course).

Unstructured

The unstructured covariance is typically used when there is a common sampling design say $\{t_{ij} : j = 1, \ldots, m_i, i = 1, \ldots, n\} = \{t_1, t_2, \ldots, t_r\}$ for not so large r; it involves $r(r-1)/2$ pairwise covariances.

Covariance pattern models

Importantly, *the models considered for the covariance matrix will not explicitly distinguish between the among-units and the between units variation*. Here are few common covariance pattern models that are described by only few parameters.

Compound symmetric: Any two measurements within a specific subject has the same correlation. Here, $\omega = (\sigma^2, \rho)$

$$\Sigma_i(\omega) = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \dots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \dots & \rho\sigma^2 \\ \vdots & \vdots & \vdots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \dots & \sigma^2 \end{bmatrix}$$

One dependent: Covariance between consequtive measurements are nozero, negligible for other measurements. Here $\omega = (\sigma^2, \rho)$

$$\Sigma_i(\omega) = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & 0 & \dots & 0 & 0 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & 0 & \dots & \rho\sigma^2 & \sigma^2 \end{bmatrix}$$

Autoregressive of order 1 (equally-spaced in time): The correlation decreases off as observations get farther apart from each other in time. Here $\omega = (\sigma^2, \rho)$

$$\Sigma_i(\omega) = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 & \dots & \rho^{m-2}\sigma^2 & \rho^{m-1}\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \dots & \rho^{m-3}\sigma^2 & \rho^{m-2}\sigma^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ \rho^{m-2}\sigma^2 & \rho^{m-3}\sigma^2 & \rho^{m-4}\sigma^2 & \dots & \rho\sigma^2 & \sigma^2 \end{bmatrix}$$

This model actually makes sense if the times are equally spaced.

Toeplitz structure: $\omega = (\sigma^2, \rho_1, \dots, \rho_{m-1})$

$$\Sigma(\omega) = \begin{bmatrix} \sigma^2 & \rho_1\sigma^2 & \dots & \rho_{m-1}\sigma^2 \\ \rho_1\sigma^2 & \sigma^2 & \dots & \rho_{m-2}\sigma^2 \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{m-1}\sigma^2 & \rho_{m-2}\sigma^2 & \dots & \sigma^2 \end{bmatrix}$$

Exponential structure: $\omega = (\sigma^2, \rho)$

$$\Sigma_i(\omega) = \begin{bmatrix} \sigma^2 & \rho^{|t_{i1}-t_{i2}|}\sigma^2 & \dots & \rho^{|t_{i1}-t_{im_i}|}\sigma^2 \\ \rho^{|t_{i2}-t_{i1}|}\sigma^2 & \sigma^2 & \dots & \rho^{|t_{i2}-t_{im_i}|}\sigma^2 \\ \vdots & \vdots & \vdots & \vdots \\ \rho^{|t_{im_i}-t_{i1}|}\sigma^2 & \rho^{|t_{im_i}-t_{i2}|}\sigma^2 & \dots & \sigma^2 \end{bmatrix}$$

Note that when the set of time points $\{t_{ij} : i, j\}$ is a set of equispaced time points then the above covariance resembles to AR(1) covariance model corresponding to set of unique points.

The above covariance structure assume the same variance over time. This was used for simplicity, and one can specify covariance structures with different variances over time.

Unbalanced data

While the covariance models present before are well suited for balanced data, thay are largely unsuitable for unbalanced data (e.g., the six cities pollution data or the ultrafiltration data). Only the exponential structure presented above is suitable in such a situaltion.

We will see later in the course that a **random effects** model is more viable in this situation.

We should note that these are just a few examples of the available covariance models. There are many more such models in practice, see the references for more details.

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