Longitudinal Data Analysis: Models for mean and covariance

Arnab Maity

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

Contents

Introduction

Modeling longitudinal data is more complex than modeling independent data since we need to model the correlation among the repeated measurements within each subject, the mean trend across time requires attention, and typically the effect of the various covariates is modeled in the mean.[1]

[1] This is called the systematic part

Conceptual model

For continuous data we can write the *conceptual model* as

$$Y_{ij} = \mu(t_j) + e_{ij},$$

where $\mu(t_j)$ is the popolation mean response at time t_j and e_{ij} describes the deviation of the data Y_{ij} from the mean $\mu(t_j)$.

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.

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.

Main steps in modeling longitudinal data

The three main steps in modeling longitudinal data are:

- modeling the mean $\mu(t)$,
- modeling the variance-covariance, $\sigma^2(t)$ and $\sigma(s,t)$, and
- selecting the distribution of the data $\boldsymbol{Y}$.

In each of the steps above, it is imperative that we *look at the data*.[2] 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 for distribution is *multivariate normality*). We will discuss binary/count responses in later chapters.

[2] using techineues described in the previous chapter

Modeling the mean

Let us start with considering a balanced design, that is, $m_i = m$ and $t_{ij} = t_j$. Assume that the time points are in increasing order $t_1 < t_2 < \ldots < t_m$. We often represent the mean function $\mu(t)$ using a finite set of parameters, that is, we put a *parametric model*.

Example A: The orthodontic study data of Potthoff and Roy (1964)

Researchers were interested in the development of children over time.[3] They collected 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. 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
```

The plot in Figure 1 shows the profiles for girls and boys, respectively, and their sample mean trajectory (solid lines). Figure 2 only plots the sample mean trajetories (solid lines) and their best linear approximations (dashed lines), that is, $\mu(age) = a + b(age)$.

The plots strongly suggest that the sample mean trajectories for both the groups are very well approximated by straight lines. Overall, we observe the following:

- Each groups has a *linear trend* in the their mean trajectory,

[3] 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.

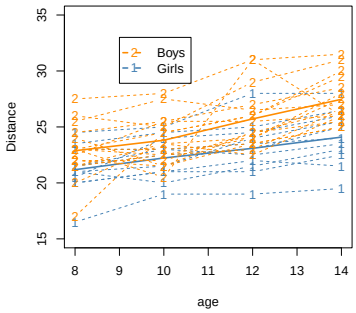


Figure 1: Subject and mean profiles for girls (blue 1) and boys (orange 2) for orthodontic data.

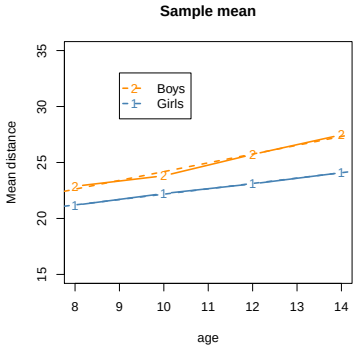


Figure 2: Mean profiles for girls (blue 1) and boys (orange 2) for orthodontic data.

- 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 $\beta_0 + \beta_1 t_j$. Thus, *if unit i is a girl*, we can write

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

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

$$\text{Boys group: } Y_{ij} = \beta_2 + \beta_3 t_j + e_{ij},$$

where β_2 and β_3 are possibly different from β_0 and β_1, respectively.

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 "indicator variable":[4] for $i = 1, \ldots, n$, we define a binary variable

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

[4] Also known as "dummy variable" or "one hot encoding".

Formulation 1

Notice that G_i is the indicator variable for the girls group. Since $G_i = 0$ if and only if $1 - G_i = 1$, we can say $1 - G_i$ is the indicator variable for the boys group. Thus we can combine the two models

$$\begin{aligned} \text{Girls group: } Y_{ij} &= \beta_0 + \beta_1 t_j + e_{ij} \\ \text{Boys group: } Y_{ij} &= \beta_2 + \beta_3 t_j + e_{ij}, \end{aligned}$$

by the following single model

$$Y_{ij} = G_i(\beta_0 + \beta_1 t_j) + (1 - G_i)(\beta_2 + \beta_3 t_j) + e_{ij}.$$

Here the index i refers to the i-th *individual*, G_i indicates whether the i-th individual is a girl or not. Check that the group means are indeed preserved. We interpret the mean parameters as follows.

	Girls group	Boys group
Intercept	β_0	β_2
Slope	β_1	β_3

We can simplify the the model as[5]

$$Y_{ij} = G_i\beta_0 + (1 - G_i)\beta_2 + G_i t_j \beta_1 + (1 - G_i)t_j\beta_3 + e_{ij}.$$

[5] Thus we have *main effects* of G_i and $1 - G_i$ (corresponding to intercepts of the two groups), and *interaction effects* $G_i t_j$ and $(1 - G_i)t_j$ (corresponding to slopes of the two groups).

This model is essentially a linear regression with Y_{ij} as response, and $[G_i, (1 - G_i), G_i t_j, (1 - G_i)t_j]$ as the set of covariates.

We can write subject level model as well. Suppose that we observe data on time points $t_1, \ldots, t_m$. We can collect the responses of the i-th subject as follows.

$$\underbrace{\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{im} \end{pmatrix}}_{\boldsymbol{Y}_i} = \underbrace{\begin{pmatrix} G_i & (1-G_i) & G_i t_1 & (1-G_i)t_1 \\ G_i & (1-G_i) & G_i t_2 & (1-G_i)t_2 \\ & \vdots & & \\ G_i & (1-G_i) & G_i t_m & (1-G_i)t_m \end{pmatrix}}_{\boldsymbol{X}_i} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_2 \\ \beta_1 \\ \beta_3 \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{im} \end{pmatrix}}_{\boldsymbol{e}_i}.$$

Recall that $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im})^T$ is the response vector for the i-th unit. The matrix $\boldsymbol{X}_i$ is called the *model matrix* for subject i, and the vector $\boldsymbol{\beta} = (\beta_0, \beta_2, \beta_1, \beta_3)^T$ is the corresponding *parameter vector*.[6] Thus we can then write the model for the i-th unit as

[6] Notice that we defined the parameter vector according to the columns of the model matrix.

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

where $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{im})^T$ is the vector of errors.

Formulation 2 (modeling the differences)

We are typically interested interested in comparing the slopes β_1 and β_3 of the two groups. We can directly model the difference of these slopes. Let us take the "boys" group as *reference*. Then we write the following model:

$$Y_{ij} = \eta_0 + t_j\eta_1 + G_i\eta_2 + G_i t_j \eta_3 + e_{ij}.$$

Let us plug-in the value of G_i for the groups and check the corresponding intercepts and slopes:[7]

[7] Notice the group means: for boys ($G_i = 0$)

$$E(Y_{ij}) = \eta_0 + t_j\eta_1$$

and for girls ($G_i = 1$)

$$E(Y_{ij}) = (\eta_0 + \eta_2) + t_j(\eta_1 + \eta_3).$$

	Girls group ($G_i = 1$)	Boys group ($G_i = 0$)
Intercept	$\eta_0 + \eta_2$	η_0
Slope	$\eta_1 + \eta_3$	η_1

Here the parameters η_0 and η_1 are the intercept and slope of the "boys" group (these have the same interpretation as β_2 and β_3 in Formulation 1) – this is the *reference group*. However, η_2 is the *difference between the intercepts* of girls and boys groups (that is $\beta_0 - \beta_2$ in Formulation 1). Similarly, η_3 is the *difference between the slopes* of girls and boys groups (that is $\beta_1 - \beta_3$ in Formulation 1). *This formulation allows us to estimate the differences of slope and intercept directly.*[8]

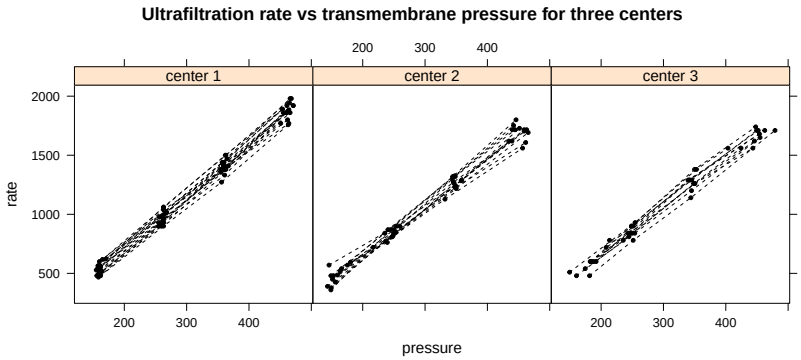
[8] Practice: Write the parameter vector and model matrix $\boldsymbol{X}_i$ in formulation 2.

Both the formulations presented above 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 unbalanced. Consider the following example.

Example B: Ultrafiltration Data For Low Flux Dialyzers in

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.[9] 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.

[9] Source: Vonesh and Carter (1987). Efficient inference for random coefficient growth curve model with unbalanced data. Biometrics, 43, 617—628.



Concept ckeck

Is this a longitudinal data? Justify your answer by clearly specifying the observational unit, and the 'time' variable.

Clearly, the 4 pressure levels (='time') at which each dialyzer was observed are not necessarily the same. However, observing the profile plots, we see that each profile (and thus their average) shows 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. Denote the j-th pressure level of the i-th dialyzer by t_{ij}. We use the following model specification for different centers:

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

Again, we would like to combine the three separate mean functions into one. Since we have three centers, we need two dummay variables.[10] For the i-th dialyzer (of the entire sample), we define:

[10] In general, if we have I groups, we need $I - 1$ dummy variables.

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

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

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 reference group. We can write the combined model: for the i dialyzer,[11]

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

[11] 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 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

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

Thus η_1 and η_4 are the intercept and slope for center 3, respectively. Thus η_1 and η_4 are essentially β_5 and β_6, respectively.

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

$$E(Y_{ij}) = (\eta_1 + \eta_3) + (\eta_2 + \eta_5)t_{ij}.$$

Here $(\eta_1 + \eta_3)$ and $(\eta_2 + \eta_5)$ are the intercept and slope for center 1, respectively. Thus η_3 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

$$E(Y_{ij}) = (\eta_1 + \eta_4) + (\eta_2 + \eta_6)t_{ij}.$$

Here $(\eta_1 + \eta_4)$ and $(\eta_2 + \eta_6)$ are the intercept and slope for center 2, respectively. Thus η_4 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 reference).

	Center 1 ($C_{1i} = 1$ and $C_{2i} = 0$)	Center 2 ($C_{1i} = 0$ and $C_{2i} = 1$)	Center 3 ($C_{1i} = 0$ and $C_{2i} = 0$)
Intercept	$\eta_1 + \eta_3$	$\eta_1 + \eta_4$	η_1
Slope	$\eta_2 + \eta_5$	$\eta_2 + \eta_6$	η_2

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 the `lm()` function to do so. *We want to only get a point estimate of the mean trajectories – we do want to NOT make any inference since we have not modeled the covariance properly yet.*

```
# Read data
ultra <- read.table("data/ultra.dat", header = F)
colnames(ultra) <- c("Id", "pressure", "rate", "center")
head(ultra)
```

```
##   Id pressure rate center
## 1  1    160.0  600      1
## 2  1    265.0 1026      1
## 3  1    365.0 1470      1
## 4  1    454.0 1890      1
## 5  2    164.0  516      1
## 6  2    260.5  930      1
```

```
# Response, time and dummy variables
Y <- ultra$rate
pressure <- ultra$pressure
C1 <- as.numeric(ultra$center == 1)
C2 <- as.numeric(ultra$center == 2)
# Fit the Least Squares model
out <- lm(Y ~ pressure + C1 + C2 + C1:pressure + C2:pressure)

out
```

```
##
## Call:
## lm(formula = Y ~ pressure + C1 + C2 + C1:pressure + C2:pressure)
##
## Coefficients:
## (Intercept)     pressure           C1           C2  pressure:C1  pressure:C2
##  -148.03509      4.05534    -27.09087    -20.73469      0.35648      0.05975
```

The six numbers in the output table "Coefficients" (Intercept, pressure, – pressure:C2) correspond to *estimated values* of $\eta_1, \eta_2, \ldots, \eta_6$, respectively.

From the output, the mean trajectory for center 3 (reference) is

$$-148.05 + 4.05t.$$

The coefficients corresponding C1 (-27.09) and pressure:C1 (0.36) denote the change in intercept and slope between center 1 and center

3. Thus the mean trajectory for center 1 is

$$(-148.05 - 27.09) + (4.05 + 0.36)t = -175.14 + 4.41t.$$

Similarly, the mean trajectory for center 2 is

$$(-148.05 - 20.73) + (4.05 + 0.06)t = -168.78 + 4.11t.$$

A plot of the three estimated mean trajectories is shown in Figure 3. Since we do not want to extrapolate, we will take the minimum and maximum pressure for each center for plotting the mean trends.

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 $\hat{\eta}_3 + 425\hat{\eta}_5 = 124.41$ units **higher**.

In contrast, compared to center 3, the mean ultrafiltration rate in center 2 for a pressure level of $t = 425$ is only $\hat{\eta}_4 + 425\hat{\eta}_6 = 4.66$ units higher.[12]

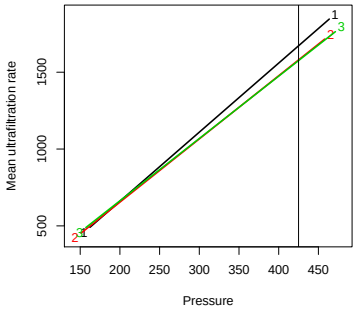


Figure 3: Estimated mean trajectories for the three centers.

[12] Note that these are only point estimates. We will need to discuss modeling covariance to get a proper estimate of standard errors.

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.

Thirty patients (13 males and 17 females) underwent hip-replacement surgery.[13] 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. The age of each participant is also recorded. The primary interest was to determine whether there are possible differences in mean response following replacement for men and women.

[13] These data are adapted from Crowder and Hand (1990, section 5.2).

```
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
```

Profile plots for each patient are shown in Figure 4. 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 $\boldsymbol{Y}_i$ are of possibly different lengths for different units.

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:

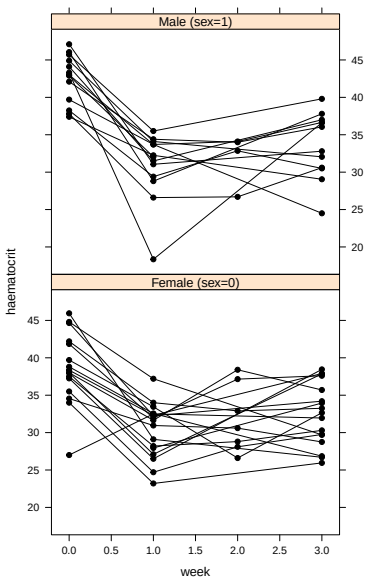


Figure 4: Subject level profile plots for the hip replacement study.

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

where a_i denotes the `age` of the individual. Here the covariate `age` is assumed to have a linear effect on the mean trajectory.

We define the indicator variable for the "male" group as before – this variable has been already stored in the "sex" variable in the data set:

$$M_i = \begin{cases} 1 \text{ if the } i\text{-th unit is a male} \\ 0 \text{ if the } i\text{-th unit is a female} \end{cases}.$$

In this case, we write one combined model (using Formulation 2) as follows, using the "female" group as the reference.

$$\begin{aligned} Y_{ij} \quad = \quad & \eta_1 + t_{ij}\eta_2 + t_{ij}^2\eta_3 + a_i\eta_4 + \\ & M_i\eta_5 + M_i t_{ij}\eta_6 + M_i t_{ij}^2\eta_7 + M_i a_i\eta_8 + e_{ij}. \end{aligned}$$

Thus the mean trends for the two groups become as follows.

$$\begin{aligned} \text{Females:} \quad & E(Y_{ij}) = \eta_1 + t_{ij}\eta_2 + t_{ij}^2\eta_3 + a_i\eta_4 \\ \text{Males:} \quad & E(Y_{ij}) = (\eta_1 + \eta_5) + t_{ij}(\eta_2 + \eta_6) + t_{ij}^2(\eta_3 + \eta_7) + a_i(\eta_4 + \eta_8). \end{aligned}$$

We can interpret the parameters as before. For example, η_4 is the effect of "age" for females, and η_8 is the *change in effect of age* between female and male groups.

Note that we can still write the model as $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{e}_i$, where the parameter vector is $\boldsymbol{\beta} = (\eta_1, \eta_2, \ldots, \eta_8)^T$, and the model matrix for the i-th unit is

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & t_{i1} & t_{i1}^2 & a_i & M_i & M_i t_{i1} & M_i t_{i1}^2 & M_i a_i \\ 1 & t_{i2} & t_{i2}^2 & a_i & M_i & M_i t_{i2} & M_i t_{i2}^2 & M_i a_i \\ & \vdots & & & & & \vdots & \\ 1 & t_{im_i} & t_{im_i}^2 & a_i & M_i & M_i t_{im_i} & M_i t_{im_i}^2 & M_i a_i \end{bmatrix}$$

In our example, take the first unit $i = 1$. The data for the individual is below.

```
##   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
```

Here we have $m_1 = 3$, the response vector is

$$\boldsymbol{Y}_1 = (47.1, 31.05, 32.8)^T.$$

Also, the first individual is a "male", that is, $M_1 = 1$, and "age" is $a_1 = 66$. Thus the model matrix $\boldsymbol{X}_1$ is as follows.

```
##  (Intercept) week I(week^2) age sex week:sex I(week^2):sex age:sex
##            1    0         0  66   1        0             0      66
##            1    1         1  66   1        1             1      66
##            1    3         9  66   1        3             9      66
```

Marginal models for the covariance

Recall that the response vector $\boldsymbol{Y}_i$ is a m_i dimensional vector of Y_{ij}'s and $\boldsymbol{X}_i$ be $m_i \times p$ dimensional model 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.). We write a linear model for the i-subject:

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

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

$$\text{cov}(\boldsymbol{e}_i) = \boldsymbol{\Sigma}_i.$$

Here the index i is used specifically to allow for different number of repeated measurements per unit m_i.[14] 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,

[14] Recall that in an unbalanced design, each subject can have different number of observations. Thus the dimension of the covariance matrix might be different from subject to sibject.

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 covariance

The unstructured covariance is typically used when there is a common sampling design say $\{t_1, t_2, \ldots, t_m\}$ for *not so large* m. Specifically, for an unit with all m measurements, the covariance matrix looks as follows – we only show the upper-triangular part since the covariance matrix is symmetric.

$$\mathbf{\Sigma}(\omega) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1m} \\ & \sigma_2^2 & \ldots & \sigma_{2m} \\ & & \ddots & \vdots \\ & & & \sigma_m^2 \end{bmatrix}.$$

Thus we have m variance parameters $\sigma_1^2, \ldots, \sigma_m^2$, corresponding to m time points.[15] Also, we have $m(m-1)/2$ pairwise covariance parameters, σ_{jk}. In total, we have $m(m-1)/2 + m = m(m+1)/2$ unknown variance-covariance parameters to estimate.

[15] Variance at each time point can be different.

For example, in Example C (the hip replacement study), we have $m = 4$, and thus we need to estimate $4(4+1)/2 = 10$ variance-covariance parameters. The covariance matrix is shown below – we only show the upper-triangular part since the covariance matrix is symmetric.

$$\mathbf{\Sigma}(\omega) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \cdot & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \cdot & \cdot & \sigma_3^2 & \sigma_{34} \\ \cdot & \cdot & \cdot & \sigma_4^2 \end{bmatrix}.$$

Here ω is the set of all the unknown variance components,

$$\omega = (\sigma_1^2, \ldots, \sigma_4^2, \sigma_{12}, \ldots, \sigma_{34}).$$

If any unit has missing observation at a specific time point, the covariance matrix will be the corrresponding submatrix of $\mathbf{\Sigma}(\omega)$. For example, if an individual has missing observation at time point 2, that is, observations are only made at time points 1, 3 and 4, then the corresponding covariance matrix is as follows:

$$\mathbf{\Sigma}_i(\omega) = \begin{bmatrix} \sigma_1^2 & \sigma_{13} & \sigma_{14} \\ \cdot & \sigma_3^2 & \sigma_{34} \\ \cdot & \cdot & \sigma_4^2 \end{bmatrix}.$$

Nevertheless, if we can estimate ω, we can then estimate the covariance matrix of any subject.

Covariance pattern models

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 ρ, and each time point has the same variance σ^2.

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

Here, $\omega = (\sigma^2, \rho)$. The second matrix is the *correlation matrix*.

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

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

This model actually makes sense if the times are equally spaced. For example, correlation beteen observations at t_1 and t_3 is $\rho^{|1-3|} = \rho^2$.

Exponential structure: This covariance structure is applicable to unbalanced (and not equally spaced) data as well.

$$\Sigma_i(\omega) = \sigma^2 \begin{bmatrix} 1 & \rho^{|t_{i1}-t_{i2}|} & \dots & \rho^{|t_{i1}-t_{im_i}|} \\ & 1 & \dots & \rho^{|t_{i2}-t_{im_i}|} \\ & & \ddots & \vdots \\ & & & 1 \end{bmatrix}$$

Here also $\omega = (\sigma^2, \rho)$. This is a generalization of the Autoregressive structure.[16] For example, suppose we observe data at time points $t_{i1} = 1.2$ and $t_{i2} = 3.9$. The correlation between the observed responses at these time points is $\rho^{|t_{i1}-t_{i2}|} = \rho^{|1.2-3.9|} = \rho^{2.7}$.

[16] Note that when the set of time points is a set of equispaced time points then the above covariance resembles to AR(1) covariance model corresponding to set of unique points.

The few covariance structures discussed above assume that the variance is same over time. This was used for simplicity, and one can specify covariance structures with different variances over time as well, as we will see in the next chapter.

Unbalanced data

While the covariance models presented before are well suited for balanced data, many of them 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 situation. 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.