Linear Mixed Models (LMMs) for Correlated Data

LMM Characteristics:

- A mixed model is one that contains both fixed and random effects.
- Mixed models for longitudinal data explicitly identify individual (random effects) and population characteristics (fixed effects).
- Mixed models are very flexible since they can accommodate any degree of imbalance in the data. That is, we do not necessarily require the same number of observations on each subject or that the measurements be taken at the same times.
- Also, the use of random effects allows us to model the covariance structure as a continuous function of time.
- Main Idea: To explicitly model sources of correlation using random effects at the modeling stage.

Example 1: Longitudinal Data (Clustered Data):

subject	time	random effects
1	$Y_{11}, Y_{12}, \cdots, Y_{1n_1}$	$\mathbf{b}_1$
2	$Y_{21}, Y_{22}, \cdots, Y_{2n_2}$	$\mathbf{b}_2$
$\vdots$	$\vdots$	$\vdots$
m	$Y_{m1}, Y_{m2}, \cdots, Y_{mn_m}$	$\mathbf{b}_m$

- Assume m independent subjects (clusters)
- For the ith of m subjects ($i = 1, \cdots, m$), there are n_i observations over time.

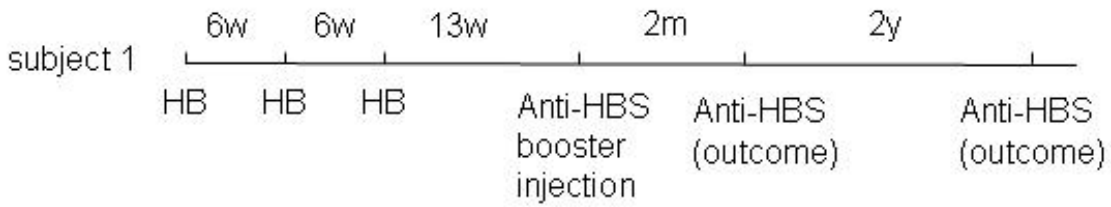
$$\mathbf{Y}_i = (Y_{i1}, \cdots, Y_{in_i})^T \quad - \quad \text{outcome } n_i \times 1$$
$$\mathbf{X}_i = (\mathbf{X}_{i1}, \cdots, \mathbf{X}_{in_i})^T \quad - \quad \text{covariate matrix } n_i \times p$$

where Y_{ij} is the normally distributed outcome and $\mathbf{X}_{ij}$ is a $p \times 1$covariate vector at the jth time point of the ith subject.

- $\mathbf{b}_i$: random effects for the *ith* subject

Example: Hepatitis Data

- 118 infants in Senegal
- Anti-HBs titer measures the degree of immunity to HB
- total # of observation =259



Scientific questions:

1. How does the Anti-HBs titer change after booster injection?
2. How does the pre-immunization Anti-HBs titer affect the Anti-HBs levels after booster injection?

The Linear Mixed Model

$$Y_{ij} = \mathbf{X}_{ij}^T \beta + \mathbf{Z}_{ij}^T \mathbf{b}_i + \epsilon_{ij}$$

- Y_{ij} : the *jth* outcome of the *ith* subject.
- β : regression coefficient vector $(p \times 1)$.
- $\mathbf{b}_i$: random effects for the *ith* subject, $\mathbf{b}_i \sim N\{0, \mathbf{D}(\theta)\}$, and θ is a $q \times 1$ vector of variance components.
- ϵ_{ij} : residual, and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \cdots, \epsilon_{in_i})^T \sim N\{0, \mathbf{R}(\theta)\}$
- $(\mathbf{X}_{ij}, \boldsymbol{Z}_{ij})$: covariate design matrices.

LMM Features

1. The observations of the same subject share the same random effects, which are used to model the correlation of repeated measures.

2. The random effects $\mathbf{b}_i$ vary from one subject to another.

3. Assume $\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_m$ $i.i.d \sim N\{0, \mathbf{D}(\boldsymbol{\theta})\}$. The observations from different subjects are independent.

4. The variance components $\boldsymbol{\theta}$ model the between-subject variation. e.g. $\boldsymbol{\theta} = 0 \Rightarrow$ no correlation.

Examples $\mathbf{b}_i$: Random Intercept (1)

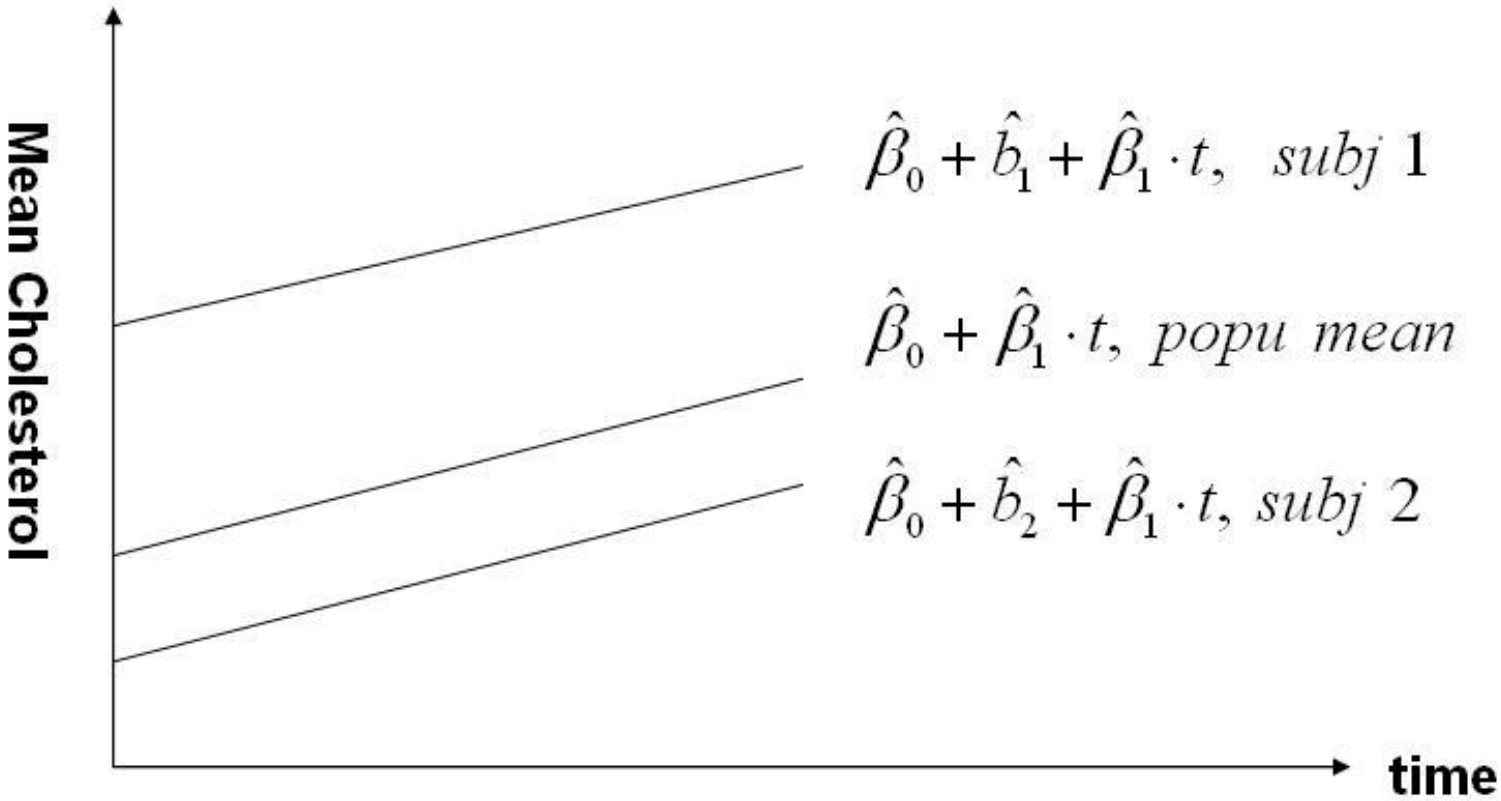
$$Y_{ij} = \mathbf{X}_{ij}^T \beta + b_i + \epsilon_{ij}$$

where $b_i \overset{i.i.d}{\sim} N(0, \theta)$, and $\epsilon_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2)$.

- $var(Y_{ij}) = \theta + \sigma^2$, constant.
- $corr(Y_{ij}, Y_{ik}) = \frac{\theta}{\theta+\sigma^2}$, when $j \neq k$.
  - Constant correlation; exchangeable correlation; compound symmetry.

Note: Hard to test $H_0 : \theta = 0$ v.s. $H_1 : \theta > 0$ since H_0 is on the boundary, s.t. LR no longer a usual chi-square distribution.

Examples $\mathbf{b}_i$: Random Intercept (2)



- Subject-specific lines are parallel to the population line.
- Subject-specific intercepts are obtained by estimating the random effects b_i, which are unobserved.

Examples $\mathbf{b}_i$: Random Intercept and Slope (1)

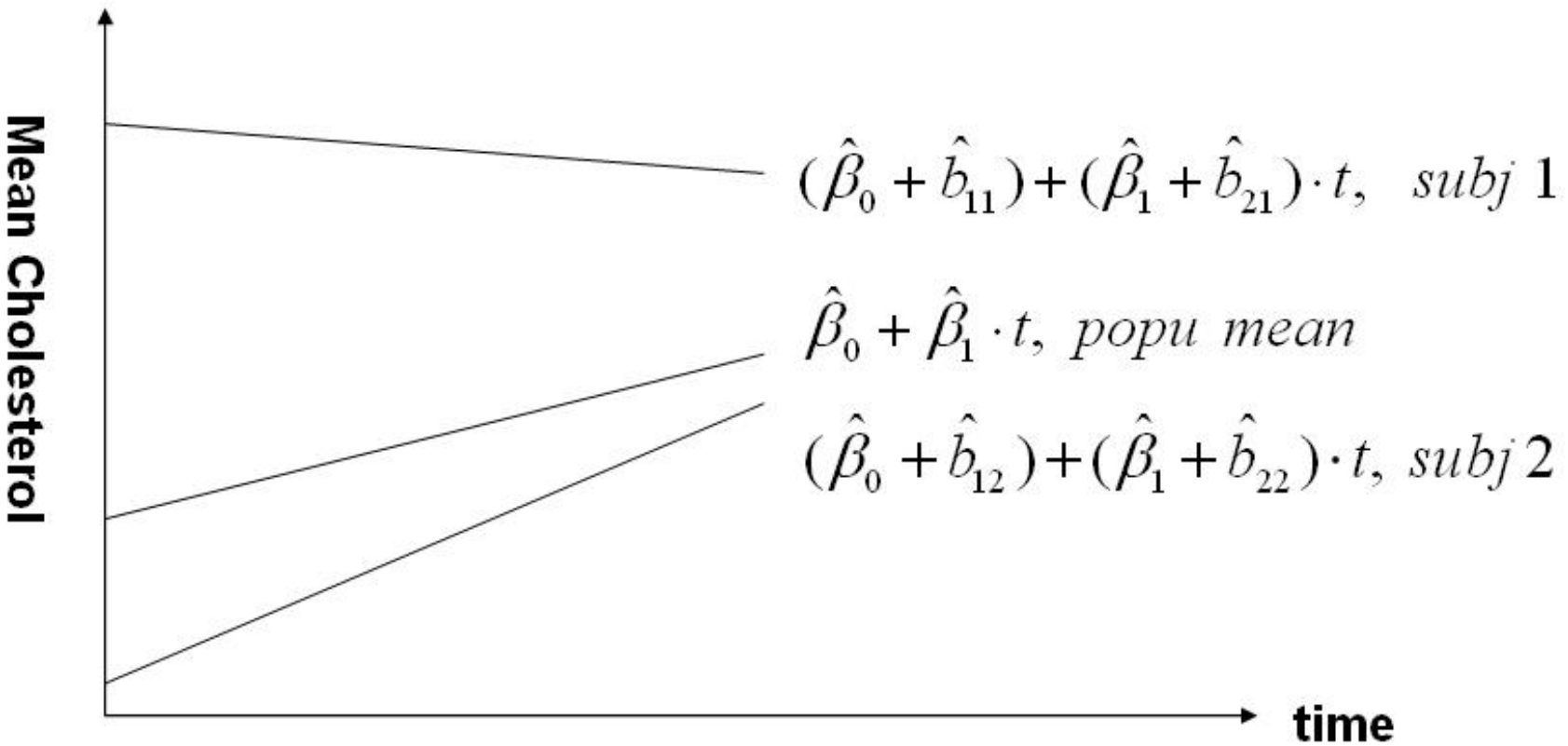
$$Y_{ij} = \boldsymbol{X}_{ij}^T \beta + b_{1i} + t_{ij} b_{2i} + \epsilon_{ij},$$

where b_{1i} is the random intercept, b_{2i} is the random slope,

$$\begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{D}(\boldsymbol{\theta}) \right\} \text{ and } \mathbf{D}(\boldsymbol{\theta}) = \begin{pmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{pmatrix}.$$

- $Var(Y_{ij}) = \begin{pmatrix} 1, & t_{ij} \end{pmatrix} \cdot \mathbf{D}(\boldsymbol{\theta}) \cdot \begin{pmatrix} 1 \\ t_{ij} \end{pmatrix} + \sigma^2$, changes over time.
- $Cov(Y_{ij}, Y_{ik}) = \begin{pmatrix} 1, & t_{ij} \end{pmatrix} \cdot \mathbf{D}(\boldsymbol{\theta}) \cdot \begin{pmatrix} 1 \\ t_{ik} \end{pmatrix}$

Examples $\mathbf{b}_i$: Random Intercept and Slope (2)



Examples $\mathbf{b}_i$: AR(1)/Ornstein-Uhlenbeck (OU) Process

$$Y_{ij} = \mathbf{X}_{ij}^T \boldsymbol{\beta} + b_{ij} + \epsilon_{ij},$$

where $Cov(b_{ij}, b_{ik}) = \theta \cdot \rho^{|t_{ij} - t_{ik}|}$, and $\epsilon_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2)$.

- $Var(Y_{ij}) = \theta + \sigma^2$.
- $Corr(Y_{ij}, Y_{ik}) = \frac{\theta}{\theta + \sigma^2} \rho^{|t_{ij} - t_{ik}|}$, exponential decay.

Note:

1. Here we always assume a stationary process, because σ^2 is a constant.
2. We can assume models for $\sigma^2(t_{ij})$, e.g. assume $ln\{\sigma^2(t_{ij})\} = \alpha_0 + \alpha_1 t_{ij}$.

Example 2: Hierarchical Data

Suppose we collect CD4 counts (y_{ijk}) at time k, from patient j, in center i:

$$Y_{ijk} = \mathbf{X}_{ijk}^T \beta + b_{1i} + b_{2j} + \epsilon_{ijk}$$

- $b_{1i} \sim N(0, \theta_1)$: center effect, θ_1 is the between-center variation.
- $b_{2j} \sim N(0, \theta_2)$: patient effect, θ_2 is the between-subject variation.
- $\epsilon_{ijk} \sim N(0, \sigma^2)$: measurement error, σ^2 is the within-subject variation.

note: Here we model the sources of variation explicitly