# BIOS 755: Linear Mixed Models I

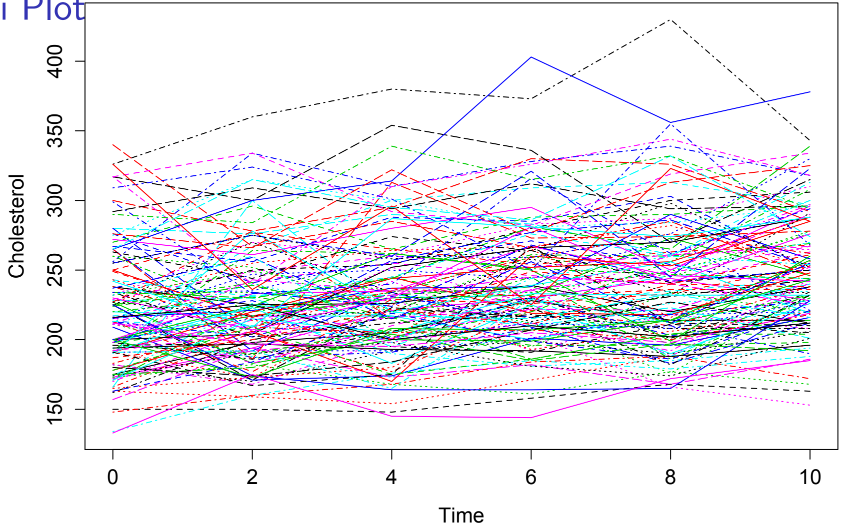
Alexander McLain

January 30, 2024

# Framingham study Cholesterol Data

- In the Framingham study, each of 2634 participants was examined every 2 years for a 10 year period for his/her cholesterol level.
- Study objectives:
  - How does cholesterol level change over time on average as people get older?
  - How is the change of cholesterol level associated with sex and baseline age?
- A subset of 200 subjects' data is used for illustrative purpose.

# Spaghetti Plot



# Introduction to Linear Mixed Models

- In the General Linear Model, we focused our conceptual model on the covariance and correlation of the error terms.
- In linear mixed models, the conceptual model is based on thinking about individual behavior first.
- The possibilities for how this is represented and how the variation in the population is represented are very flexible.
- As we'll see, linear mixed models can incorporate heterogeneity and different correlation structures (even though we don't think about them that way).

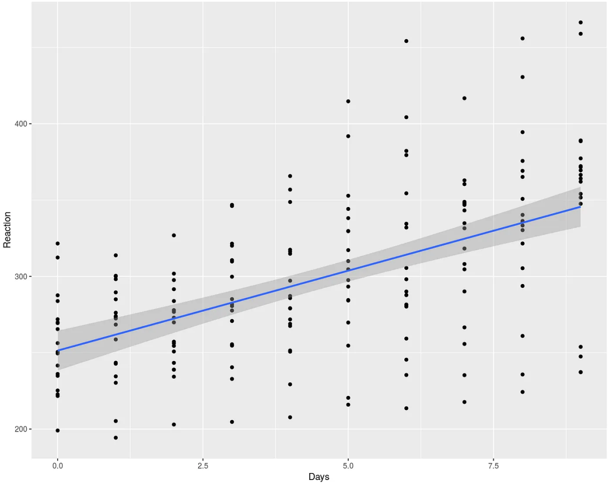
# Linear Mixed (Effects) Models

- The linear mixed model can be expressed as

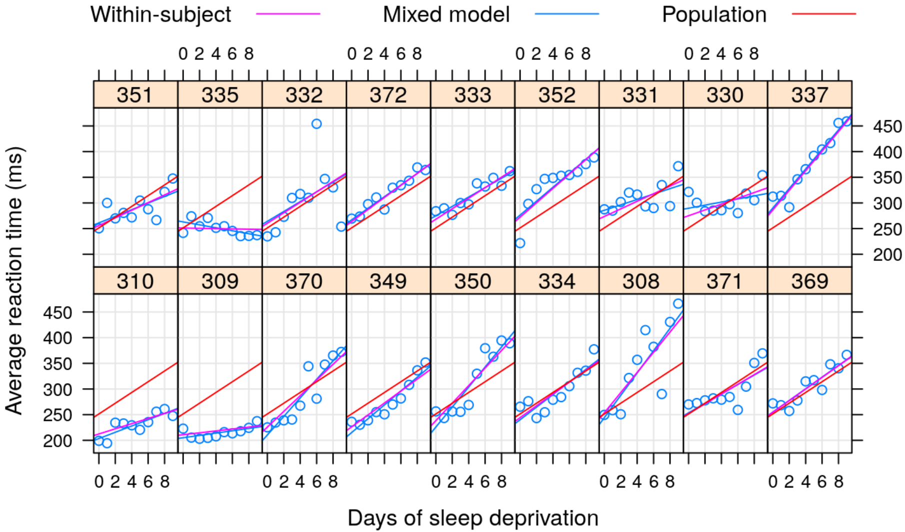
$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i$$

where

- $\boldsymbol{X}_i$ – $n_i \times p$ matrix of fixed effect covariates
- $\boldsymbol{\beta}$ – $k \times 1$ vector of regression coefficients (fixed effects).
- $\boldsymbol{Z}_i$ – $n_i \times q$ matrix of random effect covariates.
- $\boldsymbol{b}_i$ – $q \times 1$ vector of random effects, $\boldsymbol{b}_i \sim N(0, \boldsymbol{G})$,
- $\boldsymbol{e}_i$ – $n_i \times 1$ vector of errors and $\boldsymbol{e}_i \sim N(0, \boldsymbol{R}_i)$.




- Consider a sleep deprivation study where the sleeping time of 18 individuals was restricted, and their reaction to a series of tests was measured over 10 days.


Within-subject
Mixed model
Population
351
335
332
372
333
352
331
330
337
310
309
370
349
350
334
308
371
369
Average reaction time (ms)
Days of sleep deprivation
0 2 4 6 8
200
250
300
350
400
450

# Random intercept and slope model

The random intercept and slope model:

$$\boldsymbol{Y}_i = \beta_0 + b_{0i} + (\beta_1 + b_{1i})\boldsymbol{t}_i + \boldsymbol{e}_i$$

where $\boldsymbol{t}_i' = \{t_{i1}, t_{i2}, \ldots, t_{in_i}\}$

- $\beta_0$ is the average intercept and $b_{0i}$ are the deviations from the average intercept.
- $\beta_1$ is the average slope and $b_{1i}$ are the deviations from the average slope.
- We could add other fixed effects to this model (sex, smoking, etc.).

# Random intercept and slope model

The random intercept and slope model:

$$\boldsymbol{Y}_i = \beta_0 + b_{0i} + (\beta_1 + b_{1i})\boldsymbol{t}_i + \boldsymbol{e}_i$$

- $\boldsymbol{R}_i = var(\boldsymbol{e}_i)$ describes the covariance of the residuals
- In the models we've been running in the previous weeks, this is the covariance of the $i$th subject's deviations from $\beta_0 + \beta_1\boldsymbol{t}_i$ (i.e., the overall trend)
- Now it's the covariance of the $i$th subject's deviations from $\beta_0 + b_{0i} + (\beta_1 + b_{1i})\boldsymbol{t}_i$ (i.e., their individual trend)
    - Usually, it is assumed that $\boldsymbol{R}_i = \sigma^2 I$, which is the "conditional independence assumption."

# Linear Mixed (Effects) Models

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

The vector of regression parameters $\boldsymbol{\beta}$ are the fixed effects, which are assumed to be the same for all individuals.

- Fixed effects are constant across individuals, and random effects vary.
- For example, in a growth study, a model with random intercepts $\beta_0 + b_{0i}$ and fixed slope $\beta_1$ corresponds to parallel lines for different individuals $i$, or the model $Y_{ij} = \beta_0 + b_{0i} + \beta_1 t_{ij} + e_{ij}$

# Decomposing the Variation

- In the linear mixed-effects model

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

the error term $e_{ij}$ is decomposed as

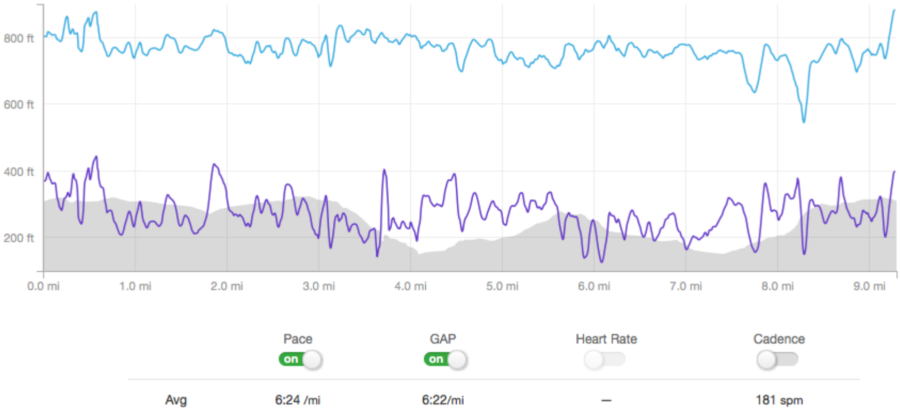
$$e_{ij} = e_{ij1} + e_{ij2}$$

where $e_{ij1}$ represents the deviations due to within-subject fluctuations and $e_{ij2}$ those due to measurement error, where

$$\boldsymbol{e}_i = \boldsymbol{e}_{i1} + \boldsymbol{e}_{i2}.$$

# Decomposing Variation

# Decomposing Variation



# Within-unit Variation

- Some typical scenarios; considerations involved in identifying an appropriate $\boldsymbol{R}_i$.
- There may be biological fluctuations over time, but commonly, the observation times are not close enough to catch these variations.
- Then correlation due to within-subject sources among the $Y_{ij}$ may be considered negligible.
- If we furthermore believe that the magnitude of fluctuations is similar across time and units, we may represent this variance as

$$var(\boldsymbol{e}_{i1}) = \sigma_1^2 \boldsymbol{I}$$

# Within-unit Variation

- It's probably reasonable to assume that errors in measurement are uncorrelated over time. Thus,

$$var(\boldsymbol{e}_{i2}) = \sigma_2^2 \boldsymbol{I}.$$

- We then have

$$\boldsymbol{R}_i = var(\boldsymbol{e}_i) = var(e_{1i}) + var(e_{2i}) = \sigma_1^2 \boldsymbol{I}_{n_i} + \sigma_2^2 \boldsymbol{I}_{n_i} = \sigma^2 \boldsymbol{I}_{n_i},$$

  where $\sigma^2$ is the aggregate variance reflecting variation due to both within-unit sources.

- The assumption that $e_{ij2}$ and $e_{ij2}$ are independent is standard, as is the assumption that $\boldsymbol{e}_{i1}$ and $\boldsymbol{e}_{i2}$ (and hence $\boldsymbol{e}_i$) are independent of $\boldsymbol{b}_i$.

# Within-unit Variation

The two special cases of within-unit variation:

- If there is no (or very little) measurement error (e.g. height and weight), $\boldsymbol{e}_i = \boldsymbol{e}_{i1}$ (all within-unit variation is due to things like "fluctuations").
- Similarly, we may have a rather "noisy" measuring device such that, relative to errors in measurement, deviations due to within-unit subjects are virtually negligible. In this case, $\boldsymbol{e}_i = \boldsymbol{e}_{i2}$ (all within-unit variation is solely the measurement error variance).

# Among-unit Variation

- The random effects $\boldsymbol{b}_i$ have mean 0 and represent variation resulting from individual units' differences.
- Intercepts and slopes may tend to be large or small together, so subjects with steeper slopes tend to "start out" larger at the beginning.
- This suggests that it would not necessarily be smart to think of $var(\boldsymbol{b}_i)$ as a diagonal matrix (independence).

# Among-unit Variation

- For this reason, we can also specify a covariance matrix for the random effects.

$$var(\boldsymbol{b}_i) = \boldsymbol{G}$$

- For $\boldsymbol{b}_i = \{b_{0i}, b_{1i}\}'$, $\boldsymbol{G}$ is a $2 \times 2$ matrix

$$\boldsymbol{G} = \begin{pmatrix} G_{11} & G_{12} \\ G_{11} & G_{22} \end{pmatrix}$$

with

$$var(b_{0i}) = G_{11}, \quad var(b_{1i}) = G_{22}$$

$$cov(b_{0i}, b_{1i}) = G_{12}$$

## Among-unit Variation

- A standard assumption is that the $\boldsymbol{b}_i$ have a multivariate normal distribution

$$\boldsymbol{b}_i \sim MVN(\mathbf{0}, \boldsymbol{G})$$

- It is usually assumed that $\boldsymbol{e}_i$ and $\boldsymbol{b}_i$ are independent. This says that the magnitude of variation within a unit does not depend on the magnitude of $\boldsymbol{b}_i$ for that unit.

# Conditional vs marginal mean

- The **conditional** mean of $\boldsymbol{Y}_i$, given $\boldsymbol{b}_i$, is

$$E(\boldsymbol{Y}_i|\boldsymbol{b}_i) = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i$$

- The **marginal** or **population-averaged** mean of $\boldsymbol{Y}_i$ is

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

- In contrast to $\beta$, the vector $\boldsymbol{b}_i$ is comprised of subject-specific regression coefficients.
- All covariates in $\boldsymbol{Z}$ will be in $\boldsymbol{X}$, and it's rare to consider more than 2 variables in $\boldsymbol{Z}$.

# Conditional vs marginal variance

- In the mixed model
$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i$$
- We have the following conditional and marginal expectations
$$E(\boldsymbol{Y}_i|\boldsymbol{b}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, \quad E(\boldsymbol{Y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$$
along with the following conditional and marginal variances
$$\begin{aligned} var(\boldsymbol{Y}_i|\boldsymbol{b}_i) &= var(\boldsymbol{e}_i) = \boldsymbol{R}_i, \quad \text{and} \\ var(\boldsymbol{Y}_i) &= var(\boldsymbol{Z}_i\boldsymbol{b}_i) + var(\boldsymbol{e}_i) \\ &= \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i' + \boldsymbol{R}_i \end{aligned}$$

# Linear Mixed (Effects) Models

- Introducing random effects induces correlation among the $\boldsymbol{Y}_i$.
- $Var(\boldsymbol{Y}_i)$ is described in terms of a set of covariance parameters, some defining $\boldsymbol{G}$ and some defining $\boldsymbol{R}_i$.
- It is difficult to disentangle the variance for $\boldsymbol{G}$ and variance for $\boldsymbol{R}_i$, which is one reason why we usually assume $\boldsymbol{R}_i = \sigma^2 \boldsymbol{I}$
- Linear mixed models are just another type of covariance matrix, which can lead to strange results (as we'll see).

# Linear Mixed (Effects) Models Summary 😎

- LMMs account for correlation through random effects that are unique to each individual.
- LMMs offer flexibility in modeling different data types and can handle unbalanced designs much better than covariance pattern models.
- The interpretation of the fixed effects is similar to that in standard linear regression.
- LMMs come with assumptions such as normality of residuals, independence of errors, and homoscedasticity (constant variance of errors), which must be checked for valid inferences.
- LMMs are especially useful for hierarchical or multilevel data.