# Methodological Comparisons

This document has been created to briefly describe the Linear Mixed Effects Model and Generalized Estimating Equations methods that we will be using to model the individual-specific, single-cell data. The objectives of this documents are to:

- Communicate the general form of the model being used
- Compare the general model against standard linear regression techniques
- If & when possible, compare each modeling technique against the other
- Address unique aspects pertaining to model mean, covariance, error, testing & inference, and assumptions

It should be noted that each general form will correspond to a data structure which will have the following pre-defined notation, and assumptions:

Assumptions:

- Each irreducible structural component of the data (whether outcome observation or explanatory observation) is directly associated with at least two indices: $i = 1, \dots, N$ -individual (Subject), $j = 1, \dots, n_i$ -occasion (cell)
- Each irreducible structural component of the data may further be associated with a "time" at which the observation was recorded. Eg, the explanatory variable: $x_{ij}$ was observed at time $t_{ij}$
- Observations may be correlated across repeated measurements (within subject), but it is assumed that observations are independent between subject
- If random effects are present, they will be assumed to be "normally distributed" (appropriate to dimensionality–$b_i \sim N(\mathbf{0}, G)$) with covariance matrix $G$
- Errors are still assumed to be normally distributed ($\epsilon_i \sim N(\mathbf{0}, R_i)$) where $R_i = \sigma^2 I$
- The errors are assumed to be independent of the random effects

Notation:

- Linear Mixed Effects Model (LMEM)
  - $\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$ p X 1 vector of fixed coefficients
  - $i = \begin{bmatrix} 1 \\ \vdots \\ q \end{bmatrix}$ q X 1 vector of random coefficients (note: $q \leq p$)

- $X_i = \begin{bmatrix} X_{i11} & X_{i12} & \cdots & X_{i1p} \\ X_{i21} & X_{i22} & \cdots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in_i1} & X_{in_i2} & \cdots & X_{in_ip} \end{bmatrix}$ $n_i\ X\ p$ Covariate Matrix
- $Z_i = \begin{bmatrix} Z_{i11} & Z_{i12} & \cdots & Z_{i1q} \\ Z_{i21} & Z_{i22} & \cdots & Z_{i2q} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{in_i1} & Z_{in_i2} & \cdots & Z_{in_iq} \end{bmatrix}$ $n_i\ X\ q$ Design Matrix
- $\epsilon_i = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_{n_i} \end{bmatrix}$ $n_i$ X 1 vector of error terms
- $Y_i = \begin{bmatrix} Y_1 \\ \vdots \\ Y_{n_i} \end{bmatrix}$ $n_i$ X 1 vector of Response terms

Details of Linear Mixed Effects Models

"The underlying premise of the model is that some subset of the regression coefficients vary randomly from one individual [subject] to another"

The general form of a longitudinal-data, LMEM is:

$$Y_i = X_i\beta + Z_ib_i + \epsilon_i$$

This is no different than a regular least squares model for logngitudinal-data except for the term $Z_ib_i$ (note that the existence of this term also changes the definition of $\epsilon_i$). This term (called random effect(s)) is used to account for between-subject variation in the response using a subset of the explanatory variables.

The random effect is assumed to have a normal distribution, i.e: $b_i \sim N(\mathbf{0}, G)$, and therefore the model now has two random variables incorporated into it, leading to the marginal/population-averaged mean:

$$E\left[Y_i\right] = X_i\beta$$

and the conditional/subject-specific mean:

$$E\left[Y_i|b_i\right] = X_i\beta + z_ib_i$$

Since the coefficients $b_i$ are assumed to have a covariance structure that is directly related to between-subject variation, the mixed effects model makes it possible to distinguish between the observed sources of variation in the data.