Notation for Longitudinal Data

General Notation

Just as you are likely used to, we denote random variables by capital Latin letters (X, Y, Z and so forth) and we denote realizations of these random variables with lower-case Latin letters (x, y, z and so forth). We use Greek letters (θ, β, α and so forth) to represent unknown parameters and will denote estimates of these parameters with a "hat" ($\widehat{\theta}$, $\widehat{\beta}$, $\widehat{\alpha}$ and so forth). I use the notation X' to mean the transpose of X, but you are welcome to use X^T as well.

Individual Notation

We denote an individual outcome as Y_{ij}. Here, we are using i to index over each individual in the study (typically taking $i = 1, \dots, n$) and we take j (generally $j = 1, \dots, k_i$) to index over the time points in the study. This is taken to be the outcome for individual i measured at time j. Occasionally we may use Y_{it_j} to denote the outcome for individual i measured at time t_j, where more complex times are used.

We denote an individual variate as X_{ijk} where i and j index over individuals and times respectively, and k indexes over the different variates that are of interest. That is, if for a specific patient we measure their age, treatment, and symptom status, we may take $k = 1, 2, 3$ for the three variables, respectively.

Often times factors X_{ijk} are not going to be changing over time (for instance, the treatment group that you belong to in a study). In this case, $X_{ijk} = X_{ij'k}$ for all j and j'. We will also often take $X_{ij1} = 1$ to be a constant, which enables us to include intercepts in our models (for instance). If a variate is time-varying we need to be more careful about their inclusion in our models.

Often it is convenient to represent some of these values as either matrices, or vectors, as is relevant. For an individual, we take $Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{ik_i})'$ to be a vector of the outcomes. For variates, we will often take $X_{ij} = (X_{ij1}, X_{ij2}, \dots, X_{ijp})$, where p different variates are measured. We can then go one step further and define

$$X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{ik_i} \end{pmatrix},$$

to be a matrix containing all of the variates. Note, we typically think of Y_i as a $k_i \times 1$ vector and X_i as a $p \times k_i$ matrix, but in certain contexts it will be useful to think of Y_i as a row vector or to take the transpose of X_i.

Notation and Considerations for Time

We typically use t_{ij} to represent the time for the i-th individual at the j-th measurement. Depending on the unit we are using, t_{ij} may often simply take the value j (where j serves as an index of visits, for instance). If the scale of time is related to calendar time (e.g., minutes, days, months, etc.) then this will be made clear and we may have items like $t_{i1} = 0$ and $t_{i2} = 14$, if for instance, the first visit and second visit are two weeks apart, and time is measured in days.

If, for all individuals, $t_{ij} = t_{i'j}$, then we say that the design is **balanced**. In this case, we drop the i subscript from times, and simply talk of $t_1, \ldots, t_k$. Most of the time we will be considering balanced designs in this course, but they are not strictly necessary.