Multi-level Models

In this module, you'll be familiarized with **multi-level models** as an approach for modeling *nested data*. You'll frequently encounter nested data structures, for example:

- Predicting **student** college admissions, where students are drawn from different *high schools*
- Modeling **patient** health outcomes, where patients are drawn from different *hospitals*
- Estimating **faculty** salaries, where the faculty are drawm from different *departments*

In each example above the **observations** (**students**, **patients**, **faculty**) belong to different *groups* (*schools*, *hospitals*, *departments*). These data can be described as **nested** because each observation comes from within a group (and we believe these groups to be important). As you can imagine, there could be some effect at the *group level*, as well as the individual level. For example, a student's college acceptance may depend on *individual predictors*, such as their GPA, number of volunteer activities, and other factors. However, their admissions status may also depend on which *school* they belong to, based on the reputation of that school, financial aid, or other factors. In this module, we'll explore various (introductory) ways to handle nested data.

Vocabulary

There are a variety of different terms that statisticians use to refer to modeling nested data. Confusingly, *many terms* may refer to the *same procedure*, and many people may use the *same term* to refer to *different procedures*. The vocabulary introduced here largely comes from this cannonical text.

- **Multi-level models**: this term refers to modeling strategies for working with nested data. Using one of many possible approaches, these appropriately handle the fact that variables may exist at multiple levels (i.e., a dataset may describe *individuals* as well as the *cities* they come from). These are also often refered to as **hierarchical models**, because the data exist in a hierarchical structure (i.e., people within cities)
- **Fixed effects**: One component of a multi-level model is the set of *fixed effects* that are estimated. These, in short, are the betas (coefficients) you are familiar with estimating (i.e., each β value in this formula):

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$$

.

These effects are considered *fixed* because they are *constant* across individuals (observations) in the dataset, and are commonly estimated through least-squares (or, more genearlly, maximum likelihood methods). they While this text prefers to refer to these as **constant slopes** (and intentionally avoids the term *fixed effects*), you will encounter it commonly in other literature.

- **Random effects**: This an an umbrella term referred to ways in which you can incorporate information about *group level variation* in your model (i.e., variation across the *cities* that *individuals* live within). Broadly speaking, one may expect the variation across groups to vary *randomly*, and multi-level models allow you to incorporate that information in various ways. In the section below, we'll describe the ways in which this variation can be built into your model

Group level variation

As described above, your outcome variable (y) may vary based on the *group* to which each observation belongs. Linear models are comprised of two components: the *intercept* and *slope*. Appropriately, you may expect an individual's group to determine their baseline *intercept*, or an associated *slope*.

Varying Intercept

One type of **random effect** is to allow each group to determine the *intercept* for each observation. Using faculty salary data as an example, it may be the case that each *department* has a different baseline salary for each faculty member, but the averge increase in salary for each year of experience is consistent across the *University.* This could be written as a *mixed effects* model as follows:

$$y_i = \alpha_{j[i]} + \beta x_i$$

In the above formula, the vector of **fixed effects** (constant slopes) is represented by the term β. The **random intercept**, for individual i group j is denoted as $\alpha_{j[i]}$. Applying this to some (hypothetical) faculty dataset, it could be written as:

$$salary_i = base_{j[i]} + 1500 * experience_i$$

In this example, a faculty member's salary depends on the **base** salary of department j that person i belongs to ($department_{j[i]}$) plus \$1500 times the amount of experience of individual i ($1500 * experience_i$).

Varying slope

Another type of **random effect** is to allow each group to determine the *slope* for each observation. Using faculty salary data as an example, it may be the case that the University has a constant baseline salary, and each *department* has a different average incrase in salary for each year of experience. This could be written as a *mixed effects* model as follows:

$$y_i = \alpha + \beta_{j[i]} x_i$$

In the above formula, the vector of **random effects** (varying slopes) is represented by the term $\beta_{j[i]} x_i$. This retrieves the *slope* for group j, of which individual i is a member. The **constant intercept** across individuals is denoted as α. Applying this to some (hypothetical) faculty dataset, it could be written as:

$$salary_i = 20000 + Raise_{j[i]} * experience_i$$

In this example, a faculty member's salary starts at \$20,000 (regardless of department). Estimating their salary requires that you retrieve the estimated slope for departmet j which individual i belongs to ($Raise_{j[i]}$).

Varying slope and intercept

As you can imagine, we can also specify a model in which the slope and intercept both vary. Continuing with our example, this would imply each department has a different starting salary (*varying intercept*), **and** each department has a different raise associated with each year of experience (*varying slope*). That model can be described as follows:

$$y_i = \alpha_{j[i]} + \beta_{j[i]} x_i$$

Or, using our salary data:

$$salary_i = base_{j[i]} + raise_{j[i]} x_i$$

Someone's salary thus depends on the *base salary of their department* ($base_{j[i]}$) as well as the annual raise associated with each year of experience *in their department* ($raise_{j[i]} x_i$).

Put graph here