Longitudinal Data: MLMs and Fixed Effects

> Outline
>
> 1. Basic Regression
> 2. Multilevel Models (MLM)
>   a. Conceptual
>   b. Pooling
>   c. Shrinkage Factor
>   d. Model Specification
>   e. Bayesian Specification with Priors
> 3. Fixed Effects
>   a. Conceptual

Basic Regression

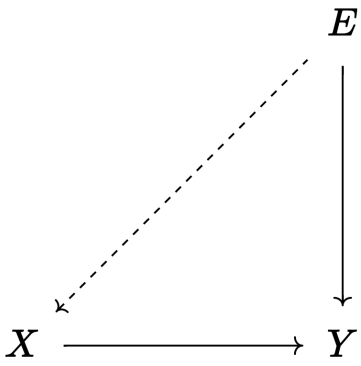
The vanilla regression equation and its corresponding DAG is the following:

$$y_i = \alpha + \beta x_i + \epsilon_i$$

- An essential assumption of this regression equation is that the predictor x_i is uncorrelated with the error term ϵ_i
  - $\text{cor}(x_i, \epsilon_i) = 0$

- What does this mean?

- If the predictor and error term are correlated, then there is an open backdoor between the predictor and the outcome through whatever is contained in the error term.
- We can think of the error term as all the other influences on the outcome other than the predictors in the model. We want to make sure that the other influences are unrelated to the model predictors because they will otherwise open some backdoors shown in the dashed line in the DAG to the right.



- Closing backdoors is the essense of causal identification. If there is an open backdoor that isn't closed (or a closed backdoor which adjustment or sampling bias opens), then the estimate of the effect of X on Y will be biased.

> What do we do with this information?
>
> The concern with the correlation between x_i and ϵ_i is not merely a matter of theoretical possibility. The real world is immensely complicated and interrelated. What does this mean? This means that it is pretty unlikely that any set of predictor variables are going to be uncorrelated with things outside of the model which are also related to the outcome.
>
> So the reality is that we have to think carefully about possible backdoors (both those which contain variables we can measure and can't - we'll address the issue of unobserved confounds with fixed effects)!

Multilevel Model (MLM)

Conceptual

Before getting into the more technical nodes about MLMs and their different specifications, we'll look at a conceptual introduction.

1. Standard regression (as we have done above) makes a number of assumptions. As we have discussed above, predictors are assumed to be uncorrelated with the error term (i.e., everything not in the model that predicts the outcome). But this is often violated and we have to address open backdoors. The assumption which concerns our attention here is the i.i.d. assumption. This stands for independent and identically distributed and means that any one observation is not predictive of another.
2. Data often have some kind of natural clustering - either where observations represent people or things actually clustered in space or contact somehow (e.g., classroom, household) or where observations represent repeated measurements for the same people or things (people are more likely to be similar to themselves over time than others!). When data are clustered, subsets of data are predictive of other subsets of data and the i.i.d. assumption is violated. If you are curious to see what the implications result from ignoring the i.i.d. assumption, you can run a simulation to see what happens where subsets of a variable are correlated with itself!
    a. Just to make this concrete, we can think about an example where a factory is interested in taking measurements of every employees rate of some task completion over time. This introduces two sources of variability - across measurements of the same person and across measurements between people at any given time (differences in difference (DiD) gets at the matter of differences between people across time!).
        i. Clustered data is the very nature of longitudinal data - repeated measurements of the same people or things (compared to repeated cross sectional data where the separate samples are deliberately independent).
        ii. Other common scenarios where such clustering emerges is where people or things are in contact with each other somehow such that they are more likely to be similar on variables of interest relative to other observations who they're not in contact with (e.g., students in a classroom influencing each others' performances or growth of animals in difference areas of a forest with different levels of resources).

Now that we have a concept sense of the motivations for multilevel models (different levels at which we have variability), let's proceed into seeing what we can actually do about this issue of clustering in data.

Pooling

When we are looking at differences between groups (setting aside clustering for the moment), there are three approaches that we can generally take. These approaches reflect different practices of a concept called "pooling." This refers to the extent to which the estimated parameters of any given subgroup are used to estimate other subgroups' parameters. That sounds weird but we can also think about this concept in terms of information. The question of pooling subgroups amounts to asking if we learn anything from subgroups when we estimate parameters for other groups, or do we estimate parameters for a group and forget everything from that group.

In his book *Statistical Rethinking*, Richard McElreath uses the example of going to coffee shops. Across one's experience of going to coffee shops throughout life, applying complete pooling would mean that a person has the exact same expectations about what will happen at *any* coffee shop (the time waited in line, the time spent waiting on one's order). Applying no pooling would mean that a person would have different expectations about how long things might take at different shops, but one's expectations at different shops are completely independent from their experience at other shops (i.e., reforming expectations anew with each subgroup).

Partial pooling introduces the idea of developing expectations about what will happen at any subgroup (or coffeeshop) based on what happens at other subgroups. For example, if one goes to a particularly crowded coffeeshop at their peak hour and they have a long wait time, this will inform their expectation of wait time at other coffee shops. However, it would be less informative because it may have been an unusually busy time compared to the experience of a long wait time when the staff seem not to be busy (this would likely change ones expectations more).

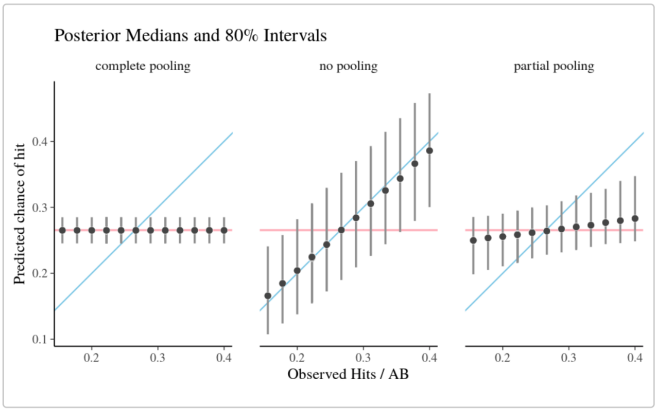
- Complete pooling
    - This approach refers to the disregard of subgroups entirely. If we have data with subgroups, completely pooling the data means that we are estimating a single aggregate parameter. For instance, national unemployment is a completely pooled measure aggregated from lower-level clusters (like states, job sectors, etc.).
- No pooling
    - This approach doesn't disregard differences between groups or clusters when estimating parameters, but estimates the parameters for groups independently. Each subgroup's parameter is estimated without any information from other subgroups. This approach allows us to get parameter estimates for every group (with no pooling).
- Partial pooling
    - Please note: This *is* multilevel modeling.
    - This approach sits between complete pooling (disregarding group differences) and no pooling (acknowledging group differences and estimating parameters independently). When we partially pool, each groups estimated parameters are pulled towards the overall mean across all groups

(called grand mean) by some amount based on something called the shrinkage factor (based on the within-group variability and sample size)

Here is a visual example of these different pooling approaches.

- With complete pooling, between-group variability is discarded
- With no pooling, between-group variability is fully expressed
- With partial pooling, between-group variability is constrained by the aggregate distribution (as a function of the shrinkage factor, which we'll discuss just below!)

Example of pooling with baseball data across player clusters.



- Keep in mind the idea of between-group variability! In MLMs, between-group variability is modeled but this has the downside of requiring the assumption that there is no backdoor between the predictor and outcome (that is, differences between subgroups other than the predictor itself are not related to the outcome!).
- When we move into fixed effects, we'll find that fixed effects handles this issue of the correlation between between-group variability and the error term not by assuming it is zero but by eliminating it by only baking within-group comparisons over time.

Shrinkage Factor

- The idea of the shrinkage factor is that the within-group statistic is contributed towards the sample predicted parameter by a factor between 0 and 1 (0 meaning the predicted sample parameter is equivalent to the grand parameter [e.g., grand mean] and 1 meaning the predicted sample parameter is equivalent to the sample statistic).
  - $\hat{\mu}_j^{PP}$ is the "partially pooled" estimate of the mean for cluster *j*

- $\hat{\mu}_j^{ML}$ is the “maximum likelihood” estimate of the mean for cluster *j*. This sounds fancy but is just the cluster average
- $\bar{\mu}$ is grand mean (average across all clusters)
- τ^2 is the between-group variance (don’t overthink, this is just the variance [like standard deviation] where the data used are the averages of the groups [rather than the data within groups])
- σ^2 is the within-group variance (I am thinking the average of within group variances since this is not subscripted for each group…)
- n_j is the sample size for cluster *j*
- $\hat{R}_j$ is the “shrinkage factor” for cluster *j*. This refers to how much the estimates mean of the cluster will be pulled towards the cluster mean and the grand mean (higher shrinkage factor means more of the contribution of the cluster mean is preserved).
  - The shrinkage factor is a variant on intraclass correlation (ICC)
    - $ICC = \frac{\text{Between-group variance}}{\text{Within-group} + \text{Between-group variance} = \text{Total variance}} = \frac{\tau^2}{\tau^2+\sigma^2}$

$$\hat{\mu}_j^{PP} = \hat{R}_j(\hat{\mu}_j^{ML} - \bar{\mu}) + \bar{\mu} \tag{1}$$

$$\tag{2}$$

$$\hat{R}_j = \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n_j}} \tag{3}$$

Model Specification

- Because the issue we have with clustered data is an issue of the distribution of errors (really residuals), the MLM splits the variance into individual level and group level variability. To this end, we have two terms representing variability - μ_j which represents cluster-level baseline variability in the outcome (e.g., students in different classrooms tend to be similar to eachother by a classroom-level amount) and ϵ_{it} which represents variability in within-group values on the outcome (e.g., within any given classroom, there is still differences between students in addition to their baseline tendency towards the classroom average).
- Here is a common way the multilevel model is represented. The equation is shown below and all the terms are explains in the following.
  - α is the grand mean across all clusters
  - βx_{ij} is the estimated effect of X on Y (a slope for each cluster)

        - I'll introduce the concept of varying-intercepts and varying-slopes here. They sound weird but I promise you already have an intuition for them! Varying-intercepts simply refer to the allowance of subgroups to have different intercepts. This is like the no pooling approach where different groups have different estimates averages. Here we are getting partially polled varying-intercepts by defining the intercept for any cluster *j* as the sum of the grand mean and partially pooled cluster-level mean: $\alpha + \mu_j$.
    - μ_j, as implied just above, is the partially pooled estimate of the mean of the cluster.
        - Please note here that the difference between no pooling and partial pooling is whether we treat μ_j as μ_j^{ML} (standard estimation of predicted value of cluster mean) or μ_j^{PP} (partially pooled estimate of cluster mean based on shrinkage factor)!
    - ϵ_{ij} is the within-cluster variability at the lowest level (e.g., among individuals within clusters)

$$y_{ij} = \alpha + \beta x_{ij} + \mu_j + \epsilon_{ij}$$

Bayesian Specification with Priors

Note that the form of this equation varies across and within disciplines! Here is how McElreath shows multilevel models where *i* refers to the individual (lowest) level and *j* refers to the group level (or higher-level cluster).

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma) && && (4)\\
\mu_i &= \alpha_{ji} + \beta x_i && && (5)\\
\alpha_j &\sim \text{Normal}(\bar{\alpha}, \sigma_j) && \text{prior for each group based on overall} && (6)\\
\bar{\alpha} &\sim \text{Normal}(0, 1) && \text{prior for average group} && (7)\\
\sigma &\sim \text{Exponential}(1) && \text{sd between groups} && (8)\\
\sigma_j &\sim \text{Exponential}(1) && \text{sd within groups} && (9)
\end{aligned}
$$

To make this more comparable with the above formulation of the multilevel model, we can also express it as follows (noting again that *i* refers to the lower-level and *j* refers to the group level.

$$
\begin{aligned}
y_{ij} &= \alpha_{ij} + \beta x_i + \epsilon \\
\text{where: } \alpha_{ij} &= \alpha_i + \mu_j
\end{aligned}
$$

We can note that this is essentially the same as the above formulation of the multilevel model (noting that in the previous specification, *i* referred to the group and *t* referred to the individual).

- The difference McElreath offers is the bayesian version where parameters are modeled as probability distributions with priors such that rather than a point estimate, posterior distributions of all parameters are estimated.

Fixed Effects (FE)

Conceptual

There are three main distinctions which set the stage for fixed effects:

- Observed vs. unobserved confounders
- Between-case variability vs. within-case variability
- Time-constant vs. time-variant confounders

So what are fixed effects and what do these distinctions have to do with them?

In causal systems, there are inevitable confounders between any set of relationships. With the goal of causal identification, we want to close open backdoors. But right away we have the issue that not all variables which are in the *true* causal system can be practically and validly measured. This presents the distinction between observed and unobserved confounders. With observed confounders, we have the tools to match/weight and/or statistically adjust for these variables to close the backdoor. But what happens when we have unobserved confounders that we can't adjust for directly? Similar to matching where we compare cases which are similar on observed covariates, in longitudinal data we can compare cases to themselves over time such that we are focusing on within-case variability rather than between-case variability. And just as with matching, making these kind of within-strata comparisons closes the backdoors for the strata within which we are comparing (e.g., comparing persons *within* the same occupations closes the backdoor of occupation).

What does this mean? When we compare cases to themselves, this removes the influence (i.e., closes the backdoor) of any variable which is the same over the period of measurement. This is the same idea as comparing different cases which are the same or similar on covariates of interest. This requirement that different cases be about the same on covariates translates to looking at the same cases over time in that looking at within-case variability only closes the backdoor on confounders which are constant over time. If we were to compare a person to themself at a different point in time and aspects about them have changed beyond the predictor of interest, then isolating on within-case variability does nothing to address those changing features (i.e., possible remaining open backdoors)!

> This is the essence of fixed effects. When we have unobserved confounders that we wish to adjust for (to close the open backdoor), we can do so by isolating within-case variability. This comes with the condition that isolating on within-case variability only works for time-constant confounders.

Model Specification

Now with a conceptual sense of fixed effects, how can we implement these ideas?

$$y_i = \alpha + \beta x_{ij} + \mu_j + \epsilon_{ij}$$

Consider the scenario where we are interested in the effect of dieting on weight change over time:

In a random representative sample, the overall effect of diet on weight change is confounded by differences between participants' baseline tendencies towards weight change.

$$weight_i = \beta diet_i + \mu_i + \epsilon_i \tag{10}$$

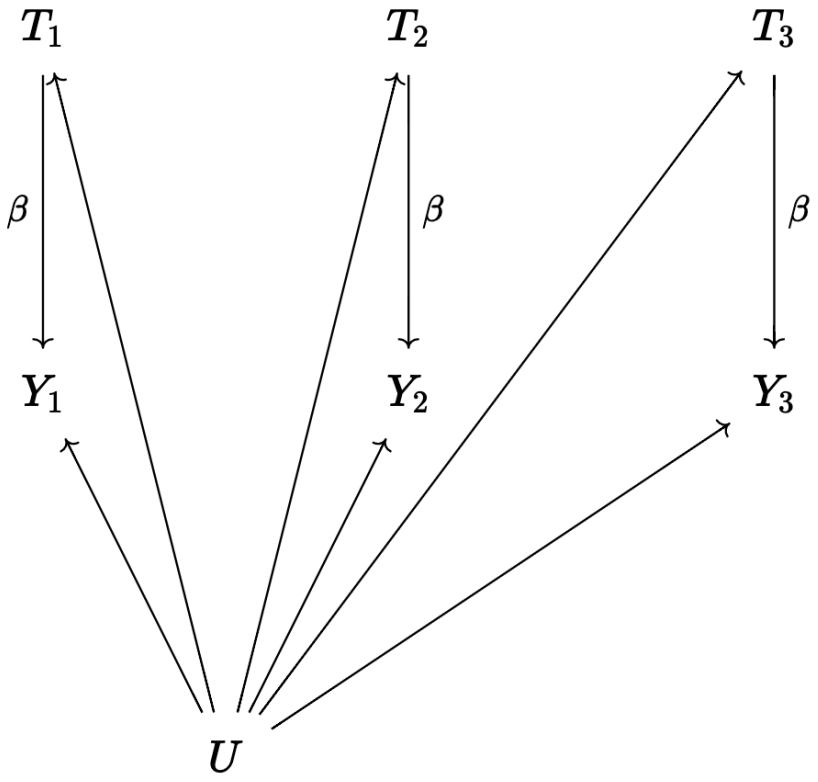
If we estimated the effect of diet over all members of the sample, this effect would be confounded by underlying differences in people (both time-constant things in people like genetics and education as an adult, and time-varying individual things like education among young people, eating habits throughout the year [holidays for instance]).

How *can* we address this?

Similar to the logic of matching and weighting, we want to close the backdoor on unobserved confounder in the form of (reasonably) time-constant variables. I am referring to time-constant variables because time-constant variables (in the context of longitudinal data) are easier to adjust for. Here is a visual why!

Consider the following DAG representing the causal model of a relationship between T and Y over time (over three waves). You'll notice that β is the same across these waves because we are modeling the overall effect of T on Y (we could decide to model change in T's effect over waves also!).

- Across all three waves, the confounder U has the same influence on both T and Y (time-constant confounder). The thing about U is that is represents ALL time-constant confounders.
- How we use fixed effects (at the most basic level) is by obliterating between-case variation and ONLY looking at differences over time over within-case measurements



In the context of diet and weight, we would be interested in the following average difference where x is some a

$$\mathbb{E}[(weight_i|diet = 1) - (weight_i|diet = 0)] = (\beta diet_1 + \mu_i + \epsilon_1) - (\beta diet_0 + \mu_i + \epsilon_2) \quad (11)$$
$$= (\beta(1) - \beta(0)) + (\mu_i - \mu_i) + (\epsilon_1 - \epsilon_2) \quad (12)$$
$$= \beta + (\epsilon_1 - \epsilon_2) \quad (13)$$

Whats going on here?

Because when we use fixed effects, we are looking at the expected difference in the outcome with and without some treatment for the same person (over the time that has transpired between before and after the treatment). Note here that we can also incorporate time as a fixed effect to remove any of the effects of time itself on the outcome (e.g., time varying differences in the outcome like seasonality).

The left and right side of the equations are equivalent as they represent the difference in the expected weight of a person at these two values of the treatment.

- Because we are doing fixed effects, the μ_i refers to individuals baseline tendency on the outcome. We are comparisons individuals to themselves in another state, so their baseline tendencies are the same (we are manually getting rid of between-case variability by comparing people to themselves).
- When remains with the difference between individuals before and after treatment is the effect of the treatment itself and potentially other time-varying confounds which the fixed effects can't address.
    - If, for instance, a person was measured pre-diet during March and post-diet in November, there may be some time-varying influences introduced by seasonal eating habits during holiday season that might likely conceal the effect of the diet (or inflate its effect in difference circumstances).

NOTE: fixed effects can be useful when we want to adjust for time-constant confounders. The question arises as to what we mean practically by time-constant. Variables don't need to be constant over *all* time, but really over the period in which the data are measured.

To cluster or not to cluster standard errors:

- Is there treatment effect heterogeneity? If so then consider clustering if the sample is non-randomly selected.