Longitudinal Data: Fixed Effects

Created	@March 23, 2023 11:21 AM
Tags	

Before getting into fixed effects, let's take a brief detour into the standard regression equation and mixed model to see where fixed effects fits into the project of causal inference.

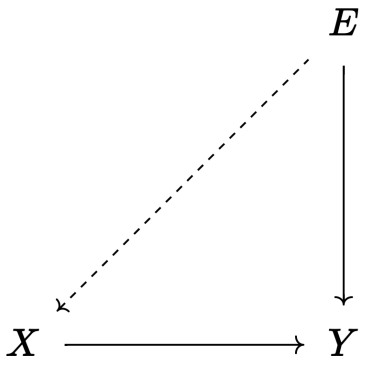
Basic Regression

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.

Multilevel Model (MLM)

Multilevel models (or hierarchical linear models, random effects models) become relevant when we have some type of clustering in out data which violates the assumption of the independence of error terms

- In the above regression model, we are also assuming that data are i.i.d (independent and identically distributed). This requires that $\text{cor}(\epsilon_i, \epsilon_j) = 0$.
- If there is clustering in data, then we can't assume that data are i.i.d. and should try and model this dependence of clustered observations. (most commonly through spatial clustering [e.g., classroom or household]
- Because the issue we have with clustered data is an issue of the distribution of errors (really residuals), the MLM splits the variance into case level and group level variability. To this end, we have two terms representing variability - μ_i which represents variability in the baseline values on the outcome (i.e., higher-level variability in the outcome [e.g., classroom rather than individual student]) and ϵ_{it} which represents variability in within-group values on the outcome (i.e., lower-level variability in the outcome [e.g., students within classrooms]).

$$y_{it} = \alpha + \beta x_{it} + \mu_i + \epsilon_{it}$$

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).

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma) && (1)\\
\mu_i &= \alpha_{ji} + \beta x_i && (2)\\
\alpha_j &\sim \text{Normal}(\bar{\alpha}, \sigma_j) && \text{prior for each group based on overall} \quad (3)\\
\bar{\alpha} &\sim \text{Normal}(0, 1) && \text{prior for average group} \quad (4)\\
\sigma &\sim \text{Exponential}(1) && \text{sd between groups} \quad (5)\\
\sigma_j &\sim \text{Exponential}(1) && \text{sd within groups} \quad (6)
\end{aligned}
$$

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.

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

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.

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.

$$weight_i = \beta diet_i + \mu_i + \epsilon_i \tag{7}$$

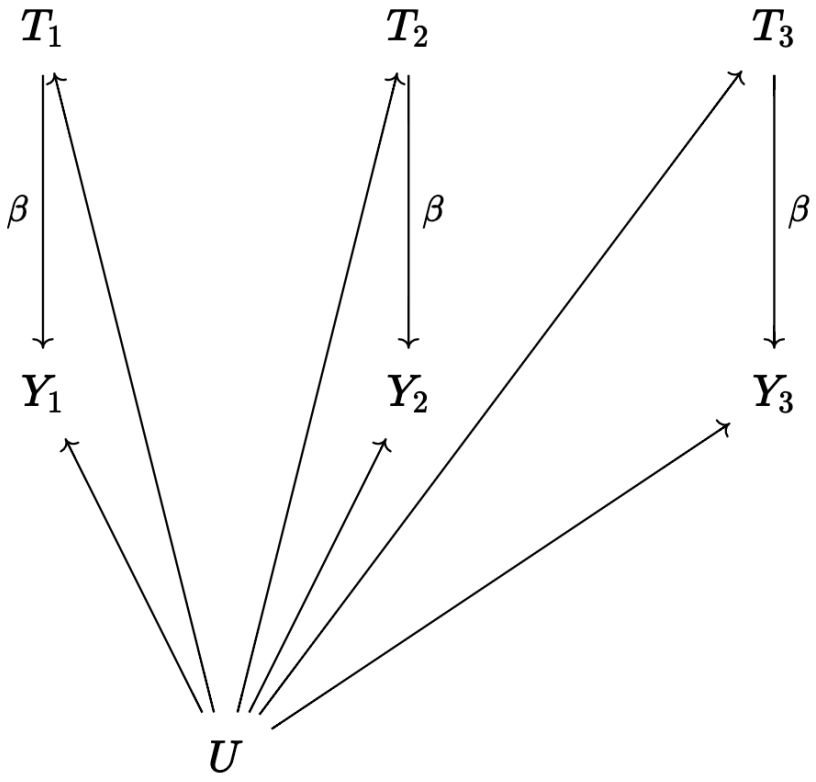
If we estimated the effect of diet over all members of the sample, this effect would be confounded by underlying differences in people (both time-constant things in people like genetics and education as an adult, and time-varying individual things like education among young people, eating habits throughout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

$$\begin{aligned}
\mathbb{E}[(weight_i|diet = 1) - (weight_i|diet = 0)] &= (\beta diet_1 + \mu_i + \epsilon_1) - (\beta diet_0 + \mu_i + \epsilon_2) && (8)\\
&= (\beta(1) - \beta(0)) + (\mu_i - \mu_i) + (\epsilon_1 - \epsilon_2) && (9)\\
&= \beta + (\epsilon_1 - \epsilon_2) && (10)
\end{aligned}$$

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).