Fixed Effects Models
for causal inference
(and a brief comparison to random effects models)

Data structure

Suppose we have data for grade school students with the following structure

- First a random sample of schools were chosen
- Then about 20 students in each school were randomly sampled
- (assume full compliance with the survey and no missing data)
- We have the following measurements on the student/family
  - From Kindergarten
    - social, demographic measures
    - test scores
    - teacher and parent reports of behavior
    - measures of all child care arrangements until this point
  - From 1st grade
    - Whether the child was held back in school (retained) at the end of this year
  - From 3rd and 5th grade (2nd and 4th for those retained after 1st grade)
    - reading and math test scores

Research question leading to a "fixed effect" model

- What is the *causal effect* of 1st grade retention and test scores 2 years after the retention decision is made
- What if we can make the following argument…?
  - controlling for all the variables in $\mathbf{X} = (X_1,\ldots,X_K)$ isn't quite sufficient to interpret τ causally because we haven't controlled for school characteristics that could impact both the probability the child is retained and subsequent test scores
  - while we have access to some school characteristics we don't think these are sufficiently rich to fully capture this confounding
  - what if "everything" about each school (or everything that stays constant across students) could be captured by a new parameter α_i
- We could fit this model to identify the causal effect, τ

$$TEST_{ij} = \beta_0 + \beta_1 X_{ij1} + \ldots + \beta_K X_{ijK} + \tau RETAIN_{ij} + \alpha_i + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$

i indexes school
j indexes child

Assumptions

- The assumptions required for these models to identify a causal effect vary substantially depending on the structure of the data
- In a simple setting with none of the timing issues we'll discuss later in the lecture we can frame the required identifiability assumption simply as
- $Y(1), Y(0) \perp Z \mid X, \alpha$
- This requires appropriate interpretation of α as representing a characteristic of the "group" (twin pair, school, person with multiple measurements) that doesn't vary over "members" of the group (twins in the pair, children in the school, patients who see the same doctor)

Fitting fixed effects models

- Let's look at what happens when we fit the fixed effects model
- One way to fit is to include a separate indicator variable for each school…. what does this do? This allows for a different intercept for each school
- Another way to achieve the same result is to de-mean the data at the group level

Fitting fixed effects models

- Another way to achieve the same result is to *de-mean* the data at the group level
- What does it mean to *de-mean* the data?
- First consider the school-level (aggregated) version of the model

$$\overline{TEST_i} = \beta_0 + \beta_1 \overline{X}_{i1} + \ldots + \beta_K \overline{X}_{iK} + \tau \overline{RETAIN_i} + \alpha_i + \overline{\varepsilon}_i$$

- The fixed effects model in effect subtracts out this mean model from the individual model

$$TEST_{ij} - \overline{TEST_i} = (\beta_0 - \beta_0) + \beta_1 (X_{ij1} - \overline{X}_{i1}) + \ldots + \beta_K (X_{ijK} - \overline{X}_{iK}) +$$
$$\tau (RETAIN - \overline{RETAIN_i}) + (\alpha_i - \overline{\alpha}_i) + (\varepsilon_{ij} - \overline{\varepsilon}_i)$$

Fitting fixed effects models

- Another way to achieve the same result is to *de-mean* the data at the group level
- What does it mean to *de-mean* the data?
- First consider the school-level (aggregated) version of the model

$$\overline{TEST_i} = \beta_0 + \beta_1 \overline{X}_{i1} + \ldots + \beta_K \overline{X}_{iK} + \tau \overline{RETAIN_i} + \alpha_i + \bar{\varepsilon}_i$$

- The fixed effects model in effect subtracts out this mean model from the individual model

$$TEST_{ij} - \overline{TEST_i} = (\beta_0 - \beta_0) + \beta_1 (X_{ij1} - \overline{X}_{i1}) + \ldots + \beta_K (X_{ijK} - \overline{X}_{iK}) +$$
$$\tau (RETAIN - \overline{RETAIN_i}) + (\alpha_i - \overline{\alpha}_i) + (\varepsilon_{ij} - \bar{\varepsilon}_i)$$

the "fixed effects" disappear! now more clear why it is necessary for $\alpha_{ij}=\alpha_i$

Fitting fixed effects models

Note that if any of these X variables are constant across students in the school (for instance if they are school-level variables) the associated terms will also drop from the model. This isn't a big deal if we are only interested in τ.

$$\overline{TEST_i} = \beta_0 + \beta_1 \overline{X}_{i1} + \ldots + \beta_K \overline{X}_{iK} + \tau \overline{RETAIN_i} + \alpha_i + \overline{\varepsilon}_i$$

- The fixed effects model in effect subtracts out this mean model from the individual model

$$TEST_{ij} - \overline{TEST_i} = (\beta_0 - \beta_0) + \beta_1 (X_{ij1} - \overline{X}_{i1}) + \ldots + \beta_K (X_{ijK} - \overline{X}_{iK}) +$$
$$\tau (RETAIN - \overline{RETAIN_i}) + (\alpha_i - \overline{\alpha}_i) + (\varepsilon_{ij} - \overline{\varepsilon}_i)$$

the "fixed effects" disappear! now more clear why it is necessary for $\alpha_{ij}=\alpha_i$

Points worth noting about fixed effects

$$TEST_{ij} - \overline{TEST_i} = (\beta_0 - \beta_0) + \beta_1(X_{ij1} - \overline{X}_{i1}) + \ldots + \beta_K(X_{ijK} - \overline{X}_{iK}) + \tau(RETAIN - \overline{RETAIN_i}) + (\alpha_i - \overline{\alpha}_i) + (\varepsilon_{ij} - \overline{\varepsilon}_i)$$

- If there is no variation on $RETAIN_{ij}$ within a school, that school does not contribute to the estimation of τ. Conceptually we can think of this treatment effect as a (weighted) average of within-school treatment effects.
- Explicitly adding in indicator variables is a statistically and computationally inefficient estimation procedure
- De-meaning by group (here school) is more computationally efficient but if done manually requires "fixing up" the standard errors (`xtreg or xtmixed` will do this for you, just use the fe option)

Fixed effects models in Stata

- If you want to estimate by adding indicator variables, you can just turn the corresponding variable into indicators by using xi (or in Stata 12 or above, simply by adding the "i." prefix)

```
. xi: regress outcome treat conf1 conf2 i.group_var
```

- Since this is computationally inefficient a better option is often to use the xtreg command with the "fe" option

```
. xtreg outcome treat conf1 conf2, i(group_var) fe
```

Fixed effects models in R

- If you want to estimate by adding indicator variables, you can just turn the corresponding variable into a factor variable

```
. lm(outcome ~ treat + conf1 + conf2 + factor(group_var)
```

- Since this is computationally inefficient a better option may be to use the felm function in the lfe package

```
. lmer(outcome ~ treat + conf1 + conf2 | group_var)
```

- A random effects version of this model could be fit using the lmer function (part of the lme4 package) as

```
. lmer(outcome ~ treat + conf1 + conf2 + (1 | group_var)
```

For those of you who know something about random effects…

- It may not be immediately obvious why the random effects formulation doesn't accomplish the same goal as the fixed effects formulation. The random effects model would look like:

$$TEST_{ij} = \beta_0 + \beta_1 X_{ij1} + \ldots + \beta_K X_{ijK} + \tau RETAIN_{ij} + \alpha_i + \varepsilon_{ij}$$
$$\alpha_i \sim \mathrm{N}(0, \sigma_\alpha^2)$$
$$\varepsilon_{ij} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$

- From the perspective of this inferential goal the key difference is the requirement that $\mathrm{E}[\alpha_i \mid \boldsymbol{X}_i, RETAIN_i] = \mathrm{E}[\alpha_i] = 0$
- For this to be satisfied the α_i must be uncorrelated with the all of the predictors in the model *including the treatment variable*
- However we want to control for α_i precisely because we think it *captures something about the schools that is related to the retention rate in that school as well as subsequent outcomes*!
- So our motivation for using the fixed effects formulation automatically invalidates the random effects assumptions!

Is a compromise possible?

- This conclusion is a little unsatisfying because it forces us to use such an inefficient model. Is there a compromise?
- Yes! Several compromises have been proposed in the form of augmented models that include block specific means of the treatment variable in any of several different ways. The two most prominent approaches are
  - Correlated random effects (Mundlak, 1978, Chamberlain, 1982, 1984; Bafumi & Gelman, 2011)
  - Adaptive centering (Raudenbush, 2009)

Correlated random effects

- Recall that the problem is that our model assumed that
  $E[\alpha_i \mid \boldsymbol{X}_i, RETAIN_i] = E[\alpha_i] = 0$
- Why not change the model to explicitly model this correlation?
- For example one could build a model that looks like

$$TEST_{ij} = \beta_0 + \beta_1 X_{ij1} + \ldots + \beta_K X_{ijK} + \alpha_i + \varepsilon_{ij}$$

$$\alpha_i \sim \mathrm{N}(\tau \overline{RETAIN}_i, \sigma_\alpha^2)$$

$$\varepsilon_{ij} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$

- or the correlation with all the predictors could be explicitly modeled as in

$$\alpha_i \sim \mathrm{N}(\gamma_1 \overline{X}_{i1} + \ldots + \gamma_K \overline{X}_{iK} + \tau \overline{RETAIN}_i, \sigma_\alpha^2)$$

Adaptive centering

- A related proposal also implicitly allows for correlation between α_i and $RETAIN_i$, by essentially replicating the fixed effects specification in the random effects framework
- For example one could build a model that looks like

$$TEST_{ij} = \beta_0 + \beta_1(X_{ij1} - \bar{X}_{i1}) + \ldots + \beta_K(X_{ijK} - \bar{X}_{iK}) +$$
$$\tau(RETAIN_{ij} - \overline{RETAIN_i}) + \alpha_i + \varepsilon_{ij}$$
$$\alpha_i \sim \mathrm{N}(,\sigma_\alpha^2),\ \varepsilon_{ij} \sim \mathrm{N}(0,\sigma_\varepsilon^2)$$

- The goal is to get the same effect estimate and standard error as the fixed effects formulation, however additionally the approach should
  - accommodate heterogeneity of impact of school-level mobility across schools; and
  - correctly incorporate school-level clustering within standard errors.

Fixed Effects Examples

Example 1: Using twins to estimate the effect of birth weight on developmental outcomes (Almond, Chay, and Lee, 2004)

- Suppose we have a sample of i =1, …, n twins, and we index the twins by j=1, 2, i.e., Y_{ij} is the jth twin in the ith pair.
- For each of the twins, we observe the covariates X_{ij1}, ..., X_{ijK}, as well as birth weight Z_{ij}.
- There is reason to believe that birth weight is randomly assigned with pairs of twins
- We do not directly observe "genetic potential," $\gamma_{i1} = \gamma_{i2} = \gamma_i$, so we will treat these as parameters to be estimated. The regression we consider is thus:

$$IQ_{ij} = \beta_0 + \beta_1 X_{ij1} + \ldots + \beta_K X_{ijK} + \tau BW_{ij} + \gamma_i + \varepsilon_{ij}$$

- There are at most n+K+1 non-redundant parameters to estimate and we have $2n$ observations. So as long as $n \geq K + 1$, we can estimate all the model parameters.

Using twins to estimate the effect of birth weight on developmental outcomes

- Estimating the regression on the previous slide (i.e. adding in indicator variables for each pair of twins) is very wasteful and can lead to computational inaccuracies. In general, we are not interested in the values of the "fixed effects" (the γ_i) – they are "nuisance parameters". So we consider instead the new differenced equation

$$Y_{i2} - Y_{i1} = \beta_1(X_{i21} - X_{i11}) + \ldots + \beta_K(X_{i2K} - X_{i1K}) + \tau(Z_{i2} - Z_{i1}) + (\varepsilon_{i2} - \varepsilon_{i1})$$

- Notice the γ_i no longer appears – it has been "differenced out"
- This is a simpler version of the "de-meaned" fixed effects model presented previously
- The results showed that for the most part birth weight did not seem to effect future developmental outcomes, with the possible exception of children born with extremely low birth weight

Example 2– Estimating the Effect of Neighborhood Quality on Educational Outcomes – Aaronson 1998

- What is the effect of neighborhood quality on educational outcomes?
- Problem: parents often choose neighborhoods in which to live based on their perceptions of the quality of schools in the neighborhood, how much they care about their children's education, how much they can afford in the way of housing, etc. It seems very difficult to think of and/or to measure all the types of variables that could influence both their choice of neighborhood and their children's educational outcomes.
- To get around this problem, Aaronson uses data on the educational outcomes of siblings. The thinking is similar to the case of the twins, but we need to make sure that there will be some variation in neighborhood quality (the treatment) among siblings within families.

Data and Model

- Data from the PSID (Panel Survey in Income Dynamics)
- Used all families with more than one child such that there is at least a 3 year gap between children.
- Each row in our dataset is a different child
- The author focuses on 2 neighborhood quality measures
  1. The dropout rate: % of young adults in a census tract aged 16-19 in 1980 (16-21 in 1970) who a) did not graduate from high school and b) are not currently in school
  2. The poverty rate

  He uses two different ways of summarizing this info across years.
  1. average of the measures across communities that the person lived in from ages 10 to 18
  2. a single measure from the community the family lived in when the child was 14

Data and Model (cont)

- Outcomes are
  - high school graduation
  - college attendance
  - grades completed)
- Measured confounding covariates: gender, race, parental education, household income, parents' marital status, number of children in household, whether the teenage worked, year born region of country (same kind of averaging as with measures)

Results

Under various specifications and with modifications of the sample, Aaronson typically finds, using either the log poverty rate in the neighborhood or the log(dropout rate) in the neighborhood, that poorer quality neighborhoods adversely affect the probability of graduating from high school, the number of grades completed, as well as the probability of attending college.

Concerns

The estimates may be biased if

1. Families may choose neighborhoods based on an individual child's (perceived) ability
2. Move to a different neighborhood is prompted by unmeasured changes in family circumstances (e.g. mental health) that also affect child outcomes
3. Covariates for a younger child are affected by the neighborhood quality of an older child (controlling for post-treatment variables)

The authors hope that variation in choice is based on reasons like:

1. Commuting preferences
2. Change in house size or house size to price preferences
3. Desire to be close to friends or relatives
4. Change in urban/rural preferences

- Generalizability? Not a tremendous amount of neighborhood quality variation within each family. Who are we making inferences about?

Example 3: Longitudinal example

- That last example seems more aligned with the types of data you've been discussing in this class
- In fact it's also probably the most popular setting for fixed effects analyses, so why didn't I start with this kind of example
- Answer: Because it is MUCH MUCH harder when we have longitudinal (panel) data to fit either fixed or random effects models to answer causal questions
- Why??

Longitudinal example: marriage and men's earnings

- Suppose we have panel data on continuously employed men (aged 18-50) spanning 10 years where in each year we have measured
  - marital status
  - age
  - years of education
  - industry and occupation
  - number of children
  - health, mental health

Let's think of what (a simple version) of one person's data might look like…

Year	Hourly wage	Married	Age	# of children	Years of Educ.	Occup.	Health
2000	10	no	20	0	12	manuf.	bad
2001	10	no	21	0	12	manuf.	good
2002	12	yes	22	1	12	manuf.	bad
2003	12	yes	23	1	12	manuf.	good
…		…	…	…	…	…	…
2009	20	yes	29	3	12	manage.	good

Fitting a model to these panel data…

- Now consider fitting any kind of panel data model to these data (i.e. any model where a row in the data represents a person/year combination)
- First just think of a regression model (which would ignore the within-person dependencies and potential trends over time)

$$WAGE_{it} = \beta_0 + \beta_1 age_{it} + \ldots + \beta_K health_{it} + \tau MARRIED_{it} + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$

- From a purely causal point of view, and even assuming that the observed covariates represented all the confounders, what would be the problem with fitting this model?

Fitting a model to these panel data…

- Now consider fitting any kind of panel data model to these data (i.e. any model where a row in the data represents a person/year combination)
- First just think of a regression model (which would ignore the within-person dependencies and potential trends over time)

$$WAGE_{it} = \beta_0 + \beta_1 age_{it} + \ldots + \beta_K health_{it} + \tau MARRIED_{it} + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$

- From a purely causal point of view, and even assuming that the observed covariates represented all the confounders, what would be the problem with fitting this model?
- We're controlling for post treatment variables!!

Fitting a model to these panel data…

- Can we solve the problem by fitting fixed effects model?

$$WAGE_{it} = \beta_0 + \beta_1 age_{it} + \ldots + \beta_K health_{it} + \tau MARRIED_{it} + \alpha_i + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$

- Nope.

- How about by fitting a random effects model?
- You guessed it.

What can we do?

- If you want to keep the panel data structure and exploit the complexity of the longitudinal treatment assignment you can use models such as
  - Marginal Structural Mean models
  - Structural Nested Means models
- These tend to rely on strong assumptions such as "sequential ignorability" which basically says that each time point ignorability is satisfied conditional on all the previous data.

Other situations where we might have to choose fixed vs. random effects?

Other settings where this decision might come up?

- Suppose who have data from a randomized experiment where children were randomly assigned to treatments within "blocks" or "strata" defined by pre-treatment covariates
- Examples:
  - Lotteries used to decide what children win a scholarship to attend private school. Different lotteries are run in schools with lower average test scores versus higher average test scores
  - Children from families who have agreed to participate in a study of an early childhood intervention for high-risk infants and toddlers are randomized separately within each of 10 cities
  - Patients are assigned to a new versus more traditional surgical procedure within 20 different hospitals
  - The twins natural experiment might be framed in this way since randomization to birth weight took place within each pair

Randomized block experiments

- These are all examples of what is known as a randomized block experiment
- The discussion about whether to fixed effect or random effects models in this setting has traditionally been framed as being based on a decision about whether we consider the blocks to be "fixed" or whether we conceive of them as a random sample from a larger population
- However our discussion today highlights the fact that a different distinction may be more pertinent
- What if unknown characteristics associated with the blocks (α_i) are also associated with the probability of being assigned to the treatment? In this case a random effects specification may be an inappropriate choice

Any additional concerns with fixed effect if ignorability is not satisfied?

Problems when ignorability isn't satisfied: Bias Amplification

- Generally we are driven to everything we can to include enough confounders so that we can feel more confident about ignorability
- However, it's not true that every covariate we add leads to less bias
- In fact, if ignorability isn't satisfied certain classes of covariates are known to amplify any bias that exists
- The classic example is an instrumental variable
- However the danger falls along a continuum so, in general, variables that are highly predictive of the treatment but not the outcome (directly) may also be worrisome

See e.g. Pearl 2009

Why might fixed effects be problematic?

- Most instruction on fixed effects for causal inferences gives the message "They can't hurt and they might help".
- NOT TRUE
- When can they get us in trouble wrt bias?
  - First if each fixed effect is itself an instrument.
  - Second if fixed effects act as an instrument in aggregate. This former happens when the covariances between the vector of fixed effects coefficients in the assignment model is the vector of fixed effects coefficients in the outcome model is 0.
  - Third, when we are close to either of these circumstances
- Middleton, Scott, Diakow, and Hill (2016) describe and also provide a sensitivity analysis strategy for helping researchers define when they may be in trouble