Lecture 21: Panel Data

Pierre Biscaye

Fall 2022

Agenda

1. Panel data
2. First differences estimation
3. Fixed effects estimation

Casual inference: overcoming MLR4

- The goal of econometrics is causal inference.
- All the assumptions are necessary for causal interpretations of results, but we focus mainly on MLR4.
- What can we do to get a handle on MLR4?
  1. Randomization
  2. Controlling for observables (did in first half); includes matching estimators - will not cover in this course
  3. Regression Discontinuity Design: matching on eligibility for treatment
  4. **Today - Panel data techniques**: control for broader set of potential omitted variables
  5. Instrumental variable techniques: use a third variable to isolate quasi-random variation in independent variable of interest
- Call 2-5 ‘quasi-random’ because we attempt to identify "as good as random" variation in the independent variable to generate a causal estimate.

Using repeated data over time to overcome MLR4

- In many contexts, $E[u|X] \neq 0$.
- Can we weaken this assumption if we observe the same cross-sectional units over time?
- Pooled cross-sections and panel data approaches allow this.
- With *panel data*, we observe the same cross-sectional observations (people, firms, states, or countries) over time. For example:
  - Observe test scores for the same students in several different grades
  - Observe the crime rate in the same cities in different years
  - Observe the pollution levels in the same states on many days
- With *pooled cross-sections*, we see new random samples at different points in time. For example:
  - Observe test scores in 10th grade for many different cohorts of students
  - Observe audits of random businesses across years

Example: crime and unemployment

- How does panel data help with MLR4?
- Suppose we wanted to estimate the effect of rising unemployment on crime in cities.
- We start with the basic statistical model

$$Crime_i = \beta_0 + \beta_1 unemp_i + e_i \tag{1}$$

- What concerns do we have with MLR4?

An alternate specification

- Maybe we don't have a great concept of *what* omitted variables are correlated with unemployment and crime in cities.
- *But* maybe we think that the most important omitted variables are things about cities that don't change very much over time (at least over a particular time scale).
- For example: geographic location, population density, types of industry present, etc.
- This would suggest that part of e_i are time-invariant city characteristics α_i. So we can write

$$Crime_i = \beta_0 + \beta_1 unemp_i + \alpha_i + u_i \tag{2}$$

- Then MLR4 requires: $E[\alpha_i + u_i | unemp_i] = 0$.
- Can panel data help with this?

Panel data structure

With panel data, we observe the same units at multiple points in time:

$$\begin{bmatrix} \textit{city} & \textit{year} & \textit{crimes} & \textit{unemp} \\ \textit{Albuquerque} & 1982 & 17136 & 8.2 \\ \textit{Albuquerque} & 1987 & 17306 & 3.7 \\ \textit{Baltimore} & 1982 & 75654 & 8.1 \\ \textit{Baltimore} & 1987 & 83960 & 5.4 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

Modeling panel data

- Suppose we have data in year 0 and year 1.

$$crime_{i0} = \beta_0 + \beta_1 unemp_{i0} + \alpha_i + u_{i0} \tag{3}$$
$$crime_{i1} = \beta_0 + \beta_1 unemp_{i1} + \alpha_i + \delta_1 + u_{i1} \tag{4}$$

- Note that all observations are indexed by *it* instead of just *i*.
  - α_i is an exception: this represents characteristics of cities that don't change over time.
- β_1 is the constant effect of unemployment on crime over time.
  - Could think of "stacking" annual regressions on top of each other: $crime_{it} = \beta_0 + \beta_1 unemp_{it} + \alpha_i + \delta_t + u_{it}$
- δ_1 reflects how year 1 is different from year 0.
  - With panel data, we will virtually always need to allow for *secular trends*: how unobserved factors affecting the outcome could be changing over time.

How do panel data help?

- Consider first a *first-differenced* specification.
- In a first-differenced specification we subtract the previous time period's data from the current time period

$$crime_{i0} = \beta_0 + \beta_1 unemp_{i0} + \alpha_i + u_{i0} \quad (5)$$

$$crime_{i1} = \beta_0 + \beta_1 unemp_{i1} + \alpha_i + \delta_1 + u_{i1} \quad (6)$$

$$\Delta Crime_i = \beta_1 \Delta unemp_i + \delta_1 + \Delta u_i \quad (7)$$

- We have eliminated α_i! MLR4 now requires $E[\Delta u_i | \Delta unemp_i] = 0$.
- If we think α_i was the main source of bias, then MLR4 is now more likely to hold.
- Even if $E[\Delta u_i | \Delta unemp_i] \neq 0$ we would still have greatly reduced the bias in $\hat{\beta_1}$.
- Can extrapolate this to more than two time periods.

First Differences as a linear regression

- Define
  - $y_i = \Delta Crime_i$
  - $x_i = \Delta unemp_i$
  - $v_i = \Delta u_i$
- Then

$$\Delta Crime_i = \beta_1 \Delta unemp_i + \delta_1 + \Delta u_i \tag{8}$$

$$y_i = \delta_1 + \beta_1 x_i + v_i \tag{9}$$

- Thus, we can use all of our linear regression tools.
- To Jupyter!

What does first differencing do?

- By *first differencing* we have removed any features in our data which are constant over time.
- This includes many potential omitted variables - anything that is correlated with both crime and unemployment that doesn't change within cities over time.
  - Critically, we don't even *have* to know what the omitted variables are.
- What does it not do? Deal with potential omitted variables that *do* change within cities over time.
- How useful this strategy is depends on whether you think the most important omitted variables are time-varying or not.

Another interpretation with panel data: Fixed Effects

Can think of α_i terms as qualitative data.

$$crime_{it} = \beta_0 + \beta_1 unemp_{it} + \delta_t + \alpha_i + u_{it} \tag{10}$$

$$\begin{aligned} crime_{it} = \beta_0 + \beta_1 unemp_{it} + \delta_t + \alpha_1 city1_i \\ + \alpha_2 city2_i + ... + \alpha_k cityk_i + u_{it} \end{aligned} \tag{11}$$

- We are concerned about time-invariant city characteristics α_i: controlling for what city you are in deals with that.
- Qualitative data interpretation: for each city j, we hold constant *any* ways that city is different from other cities by including a dummy variable for that city.
- Interpretation: holding those city effects constant, how does crime change when unemployment increases in a city?
- We call the method of including a series of dummy variables that capture fixed characteristics *Fixed Effects*.
  - Two main types: *unit* fixed effects (e.g., city) and *time* fixed effects (e.g., year).

Fixed effects (FE) as a linear regression

Think of these two specifications as equivalent when talking about FE (the first is shorthand for the second)

$$crime_{it} = \beta_0 + \beta_1 unemp_{it} + \delta_t + \alpha_i + u_{it} \quad (12)$$

$$\begin{aligned} crime_{it} = \beta_0 + \beta_1 unemp_{it} + \delta_1 year1_t + \delta_2 year2_t + \ldots + \delta_j yearj_t \\ + \alpha_1 city1_i + \alpha_2 city2_i + \ldots + \alpha_k cityk_i + u_{it} \quad (13) \end{aligned}$$

- Implement linear regression by including dummies for each unit and for each time period as controls.
- Note that as with first differences (FD), can't include any time-invariant variables as controls.
- With two time periods, FE and FD give identical results.
- With more than two time periods, results will differ somewhat between FD and FE but both will give consistent estimates.
    - FE usually preferred with $n > 2$ time periods.
- To Jupyter!

Interpretation with fixed effects

- Unit fixed effects control for all time-invariant characteristics within units.
- Time fixed effects control for all unit-invariant characteristics within time periods.
- But it's not just fixed variables that are captured by these fixed effects: variable *means* (within units or time periods) are also fixed.
  - For example, unemployment rates will vary across cities within a year, but a year fixed effect will control for *mean* unemployment in that year: this is fixed across cities.
  - Further, unemployment rates will vary within cities over time, but a unit fixed effect will control for *mean* unemployment in that city: this is fixed over time.
- Interpretation of X variables is then about the effect of changes (or "deviations") in X relative to means within city and time period.
  - E.g. what is the effect of unemployment being higher than usual in a given city and time period?

Example: effects of crop pests on agricultural profits

- Suppose we want to test how destruction from crop pests affects agricultural profits among poor farm households.
- We could estimate $profit_i = \beta_0 + \beta_1 pest_i + u_i$, but we are concerned about MLR4.
  - For example, farmers that experience pests might be those that don't invest in pesticides, or plant very different types of crops.
- Suppose we have data on the same panel of farm households across multiple years. What specification could we use to leverage this panel structure to reduce OVB?

Example: effects of crop pests on agricultural profits

- Suppose we want to test how destruction from crop pests affects agricultural profits among poor farm households.
- We could estimate $profit_i = \beta_0 + \beta_1 pest_i + u_i$, but we are concerned about MLR4.
  - For example, farmers that experience pests might be those that don't invest in pesticides, or plant very different types of crops.
- Suppose we have data on the same panel of farm households across multiple years. What specification could we use to leverage this panel structure to reduce OVB?
- We could estimate a fixed effects regression!
- $profit_{it} = \beta_0 + \beta_1 pest_{it} + \alpha_i + \delta_t + u_{it}$
- What are the fixed effects here, and what do they represent?
- How do we interpret β_1?
- Are there potentially still some concerns about MLR4?

Panel data and MLR4

- Clearly, panel data techniques are powerful.
- In Wage equations, we worry that people with different levels of education are different in some ways.
- In CO_2 and GDP regression, worry that richer countries are different from poorer countries in some ways.
- We brainstormed *a lot* of these potential explanations, and worry that we were not exhaustive.
- With panel data, we can hold constant *all* omitted variables which *do not change over time*.

Assumptions for panel data: MLR1

$$\text{MLR1: In the population}$$
$$y = \beta_0 + \beta_1 x_1 + ...\beta_k x_k + u \tag{14}$$
$$\text{For FD, MLR1: In the population}$$
$$\Delta y = \delta_0 + \beta_1 \Delta x_1 + ...\beta_k \Delta x_k + \Delta u \tag{15}$$

- In other words, with FD MLR1 says that we've correctly modeled how *changes* in y relate to *changes* in x.

MLR2 and MLR3

- MLR2: we have a random sample.
- With panel data, our *cross-sectional units* (e.g., cities) must be sampled at random from the population.
- MLR3: None of the x_j are multicollinear in the other $x_1, .., x_{j-1}, x_{j+1}, ..., x_k$
- MLR3 for FD: none of the other Δx_j are multicollinear in the other $\Delta x_1, .., \Delta x_{j-1}, \Delta x_{j+1}, ..., \Delta x_k$
- For both FD and FE: all x_j must vary over time (for at least one unit i). We *cannot* use panel data methods to test relationships between x and y for x variables which do not change over time.

Example: returns to education

- Suppose we have panel data on adults' earnings and want to estimate

$$log(wage_{it}) = \beta_0 + \beta_1 Ed_{it} + \beta_2 exper_{it} + \alpha_i + u_{it} \tag{16}$$

- α_i might include many of the omitted variables we have discussed.
  - Innate ability, parental wealth, location, etc., are fixed within individuals over a given time period.
- We are fairly certain that $E[\alpha_i + u_{it}|Ed_{it}] \neq 0$

Using panel data for the returns to education

$$log(wage_{it}) = \beta_0 + \beta_1 Ed_{it} + \beta_2 exper_{it} + \alpha_i + u_{it} \quad (17)$$
$$\Delta log(wage_i) = \delta_0 + \beta_1 \Delta Ed_i + \beta_2 \Delta exper_i + \Delta u_i \quad (18)$$

- α_i is removed
- But, is MLR3 satisfied?

Using panel data for the returns to education

$$log(wage_{it}) = \beta_0 + \beta_1 Ed_{it} + \beta_2 exper_{it} + \alpha_i + u_{it} \quad (17)$$

$$\Delta log(wage_i) = \delta_0 + \beta_1 \Delta Ed_i + \beta_2 \Delta exper_i + \Delta u_i \quad (18)$$

- α_i is removed
- But, is MLR3 satisfied?
- Panel Data buys us *a lot* in terms of omitted variable bias
- But, using first differenced or fixed effects estimators prevent us from estimating some meaningful relationships: education is (typically) fixed for adults above a certain age.

MLR4

- Previously, MLR4: $E[u_i|x_i] = 0$.
- Now with FD, MLR4: $E[\Delta u_i|\Delta x_i] = 0$.
- And with FE, MLR4: $E[u_{it}|x_{it}, \alpha_i, \delta_t] = 0$.
- These are very similar equivalent, since the FE case rules out the time-invariant parts of u and x and the common trends over time in these variables, leaving only changes over time.
  - Tend of think of the FE assumption as stronger since it is conditioned on more controls.
- Interpretation is we need there to be no omitted variables whose *changes* are correlated with *changes* in x and *changes* in y.
  - If there is a variable in u that changes over time in a manner correlated with changes in some x_j, that bias will not be addressed with panel data methods.

Example: crime and unemployment

$$crime_{it} = \beta_0 + \beta_1 unemp_{it} + \delta_t + \alpha_i + u_{it}$$

- In our initial example, we need there to be no omitted variables correlated with both changes in unemployment and changes in crime.
- It is now ok if places that have low unemployment rates *always* have low or high crime for *any* reason.
- It is *not* ok if places where unemployment is increasing or decreasing have increasing or decreasing crime rates for other reasons.
- Examples?

Example: crime and unemployment

$$crime_{it} = \beta_0 + \beta_1 unemp_{it} + \delta_t + \alpha_i + u_{it}$$

- In our initial example, we need there to be no omitted variables correlated with both changes in unemployment and changes in crime.
- It is now ok if places that have low unemployment rates *always* have low or high crime for *any* reason.
- It is *not* ok if places where unemployment is increasing or decreasing have increasing or decreasing crime rates for other reasons.
- Examples?
- What if a low tax base leads to high unemployment and limited policing?
- It depends - is this a change in the tax base (and policing) or is it something about the place?

Policy analysis using panel data

- We've replaced one MLR 4 with another.
- Now, instead of needing *levels* of x variables to be uncorrelated with u we need *changes* in x variables to be uncorrelated with u.
- This will be less attractive if we don't know *why* x and y are changing.
    - Harder to argue in this case nothing else is changing simultaneously.
- One compelling case where we do know why x is changing: policy analysis.
    - When x is a policy that takes effect, we know why there was a (big) change in x.
    - Next lecture: another panel data approach used when some "treatment" changes over time.