UNIVERSITY OF ILLINOIS

AT URBANA-CHAMPAIGN

Using Panel Data Estimators to Identify Causal Effects

illinois.edu

Outline for the Session

1. The Omitted Variables Problem (OVP)
2. The Unobserved Effects Model (UEM)
3. How to choose the right model
4. Operationalizing panel data models
5. Other panel data estimators

The Omitted Variables Problem

The Omitted Variable Problem (OVP)

- As we’ve seen, causal inference is a missing variables or omitted variables problem
  - We don’t know what happened to those treated in the absence of the treatment
- RCTs solves the OVP by ensuring treatment and control groups are equivalent through randomization
  - We then assume the control group is representative of what would have happened to the treatment group had they not been treated

The Omitted Variable Problem (OVP)

- Matching solves the OVP by constructing a control group based on observable characteristics
  - Conditional on observables the matched group is representative of what would have happened to the treatment group had they not been treated
  - But this does not control for unobservables

The Omitted Variable Problem (OVP)

- IVs solve the OVP by assuming that there are unobservable differences between treatment and control and finding an instrument to break the correlation between the treatment and the unobservable differences
  - Conditional on a set of Identifying Assumptions the IV allows us to control for unobserved characteristics that make the treatment and control groups different and affect the outcome

The Omitted Variable Problem (OVP)

- Panel data techniques provide an additional way to try and establish causal inference
  - When we have multiple observations of plots/households/firms over time we can control for unobserved heterogeneity and obtain consistent and unbiased estimates of the treatment effect

The Unobserved Effects Model

Some Preliminary Assumptions

- Assume a large population of cross-sectional units (plot, household, firm) that we can observe over time
- We randomly sample from the cross-section, so observations are necessarily independent in the cross-section
- We have a large cross-section (N) and relatively few time periods (O)

Some Preliminary Assumptions

- The unobserved heterogeneity, c_i, is drawn along with the observed data
  - View the c_i as random draws. The "fixed" versus "random" debate is counterproductive. The key is what we assume about the relationship between the unobserved c_i and the observed covariates, X_{it} and T_{it}
- c_i is also called the unobserved component or the latent variable

Some Preliminary Assumptions

- Then the basic linear model with additive heterogeneity can be written as

$$Y_{it} = \alpha X_{it} + \beta T_{it} + c_i + \epsilon_{it}$$

- c_i is an unobserved effect
  - In our case it is the unobserved characteristics that cause one person to adopt the treatment and another person to refuse the treatment

Some Preliminary Assumptions

$$Y_{it} = \alpha X_{it} + \beta T_{it} + c_i + \epsilon_{it}$$

- X_{it} is a set of observed variables
  - Exactly what types of variables are in X_{it} will affect our choice of what are traditionally called Fixed Effects, Random Effects, and Correlated Random Effects
- ϵ_{it} are the idiosyncratic errors
  - The composite error term is $v_{it} = c_i + \epsilon_{it}$
  - v_{it} is almost certainly serially correlated and definitely is if ϵ_{it} is serially uncorrelated. This will be because the value of c_i is the same for all t

Rewriting the Regression Model

$$Y_{it} = \theta G_t + \delta R_i + \gamma W_{it} + c_i + \epsilon_{it}$$

- G_t is a set of time effects that do not vary over individuals
- R_i is a set of observed individual effects that are time-constant
- W_{it} is a set of variables that change across individual and time

Discussion: Irrigation Project Example

$$\log(income) = G_t + \beta_1 Irrig_{it} + \beta_2 \log(dist_i) + c_i + \epsilon_{it}$$

- $Irrig_{it}$ is the treatment, if the households had received the irrigation project
- G_t are year effects capturing secular changes in price index
- $dist_i$ is household distance to market and does not change over time

Discussion: Irrigation Project Example

$$\log(income) = G_t + \beta_1 Irrig_{it} + \beta_2 \log(dist_i) + c_i + \epsilon_{it}$$

- We are interested in effects of irrigation. Distance is just a control for cost of transporting the good
  - Are there time constant differences between households not captured by distance?

Discussion: Irrigation Project Example

$$\log(income) = G_t + \beta_1 Irrig_{it} + \beta_2 \log(dist_i) + c_i + \epsilon_{it}$$

- We are interested in effects of irrigation. Distance is just a control for cost of transporting the good
  - Are there time constant differences between households not captured by distance?

YES

Discussion: Irrigation Project Example

$$\log(income) = G_t + \beta_1 Irrig_{it} + \beta_2 \log(dist_i) + c_i + \epsilon_{it}$$

- We are interested in effects of irrigation. Distance is just a control for cost of transporting the good
  - Are there time constant differences between households not captured by distance?

YES

  - Are those factors, in c_i, correlated with $Irrig_{it}$?

Discussion: Irrigation Project Example

$$\log(income) = G_t + \beta_1 Irrig_{it} + \beta_2 \log(dist_i) + c_i + \epsilon_{it}$$

- We are interested in effects of irrigation. Distance is just a control for cost of transporting the good
  - Are there time constant differences between households not captured by distance?

YES

  - Are those factors, in c_i, correlated with $Irrig_{it}$?

Probably

How to choose the right model

Panel Data Model Options

- Primary focus will be on the following
  - Pooled Ordinary Least Squares (OLS)
  - Random Effects (RE)
  - Fixed Effects (FE)
  - Correlated Random Effects (CRE)
- Alternative models
  - First Differencing (FD)
  - Multilevel Model (MLM)

Pooled OLS

- Assumes $Cov(v_{it}, v_{is}) = 0$
  - In words: the composite error term is uncorrelated across time (no serial correlation).
  - This will clearly not be true if there are unobserved effects in our model
- How likely is it that there are no unobserved effects in our model?
  - Isn't the whole point of impact assessment that we can't perfectly observe all the characteristics that affect treatment?

Random Effects

- Assumes $Cov(X_{it}, c_i) = 0$
  - Alternatively, $E[c_i|X_{it}] = E[c_i]$ – conditional mean independence
  - In words: the unobserved effect is uncorrelated with the observed explanatory variables
- How likely is it that unobserved individual characteristics are uncorrelated with observed characteristics?
  - Isn't the whole point of using panel data to allow for c_i to be arbitrarily correlated with X_{it}?

Fixed Effects

- Allows for $Cov(X_{it}, c_i) \neq 0$
  - Alternatively, $E[c_i|X_{it}]$ is allowed to be any value
  - In words: allows for arbitrary correlation between unobserved effect and the observed explanatory variables
- But, FE does not allow us to estimate time-constant variables
  - Can back them out however

Correlated Random Effects

- Assumes $E[c_i|X_{it}] = E[c_i|\bar{X}_i] = \psi + \xi\bar{X}_i$
  - In words: we model the dependence between unobserved effect and the observed explanatory variables as

$$c_i = \psi + \xi\bar{X}_i + a_i$$

- Allows us to unify FE and RE estimation approaches

Operationalizing panel data models

Pooled OLS

- Using OLS estimate

$$Y_{it} = \theta G_t + \delta R_i + \gamma W_{it} + c_i + \epsilon_{it}$$

- To test if the errors are serially uncorrelated, save $\hat{\epsilon}_{it}$ and then regress
  - $\hat{\epsilon}_{it} = \rho \hat{\epsilon}_{it-1} + u_t$
  - If $\rho = 0$ then errors are serially uncorrelated and Pooled OLS is BLUE
  - If $\rho \neq 0$ then errors are serially correlated and you need a panel data estimator

Random Effects

- Using GLS estimate

$$Y_{it} = \theta G_t + \delta R_i + \gamma W_{it} + v_{it}$$

- Several tests for validity of REs
  - To test if $c_i = 0$ you can use the Breusch-Pagan Lagrangian multiplier test for RE
  - To test if the unobserved effect is uncorrelated with the observed explanatory variables we can use a Hausman Test

Fixed Effects

- First, take the time average of our estimation equation

$$\bar{Y}_i = \theta\bar{G} + \delta R_i + \gamma\bar{W}_i + c_i + \bar{\epsilon}_i$$

- Second, subtract the time averages from standard equation

$$Y_{it} - \bar{Y}_i = \theta(G_t - \bar{G}) + \delta(R_i - R_i) + \gamma(W_{it} - \bar{W}_i) + (c_i - c_i) + (\epsilon_{it} - \bar{\epsilon}_i)$$

Fixed Effects

- Third use OLS to estimate the simplified equation

$$\ddot{Y}_{it} = \theta \ddot{G}_t + \gamma \ddot{W}_{it} + \ddot{\epsilon}_{it}$$

- Note that this time demeaning removes the time-constant unobserved effect but also removes the time-constant observed effects

Fixed Effects

- Alternatively, and potentially easier, is to estimate

$$Y_{it} = \theta G_t + \gamma W_{it} + \zeta c_i + \epsilon_{it}$$

- Include binary indicators for each individual
  - Note this controls for c_i but removes R_i due to perfect collinearity
  - It is a nice exercise in least squares mechanics to show these two "Fixed Effects" estimators are the same

Correlated Random Effects

- First, define the relationship between the unobserved effect and the observed covariates

$$c_i = \psi + \xi \bar{X}_i + a_i$$

- Second, estimate the equation with OLS

$$Y_{it} = \theta G_t + \delta R_i + \gamma W_{it} + \psi + \xi \bar{X}_i + a_i + \epsilon_{it}$$

Correlated Random Effects

- Note that we have controlled for the unobserved effect, allowed it to be correlated with our observed variables, AND kept the time constant variables
- Several interesting facts about CRE estimation
  - Pooled OLS estimators on the CRE equation gives the FE estimates of θ and γ
  - Pooled OLS estimators on the CRE equation when $\xi = 0$ gives the RE estimates of θ, δ and γ

Other panel data estimators

First Difference

- Recall the Fixed Effects equation

$$\ddot{Y}_{it} = \theta \ddot{G}_t + \gamma \ddot{W}_{it} + \ddot{\epsilon}_{it}$$

- $\ddot{Y}_{it} = Y_{it} - \bar{Y}$
- $\ddot{G}_t = G_t - \bar{G}$
- $\ddot{W}_{it} = W_{it} - \bar{W}_i$
- $\ddot{\epsilon}_{it} = \epsilon_{it} - \bar{\epsilon}_i$

First Difference

- Recall the Difference-in-Difference equation

$$\ddot{Y}_{it} = \theta \ddot{G}_t + \gamma \ddot{W}_{it} + \ddot{\epsilon}_{it}$$

- $\ddot{Y}_{it} = Y_{it} - \bar{Y}$
- $\ddot{G}_t = G_t - \bar{G}$
- $\ddot{W}_{it} = W_{it} - \bar{W}_i$
- $\ddot{\epsilon}_{it} = \epsilon_{it} - \bar{\epsilon}_i$

First Difference

- The First Difference equation is

$$\Delta Y_{it} = \theta \Delta G_t + \gamma \Delta W_{it} + \Delta \epsilon_{it}$$

- $\Delta Y_{it} = Y_{it} - Y_{it-1}$
- $\Delta G_t = G_t - G_{it-1}$
- $\Delta W_{it} = W_{it} - W_{it-1}$
- $\Delta \epsilon_{it} = \epsilon_{it} - \epsilon_{it-1}$

Hierarchical/Multilevel Models

- Multilevel Models provide a way to model grouped data
- Example: irrigation project
  - Some parcels receive irrigation some do not
  - Some farmers receive irrigation some do not
  - Some villages receive irrigation some do not
- How do we account for the different correlations within all of these groups?

Hierarchical/Multilevel Models

$$Y_{it} = \alpha X_{it} + \beta T_{it} + c_i + c_h + c_j + \epsilon_{it}$$

- c_i is a unobserved parcel level effect
- c_h is a unobserved household level effect
- c_j is a unobserved village level effect

Hierarchical/Multilevel Models

- Our typical panel data techniques can only control for one of these levels. As an example, Fixed Effects

$$Y_{it} = \alpha X_{it} + \beta T_{it} + \zeta c_i + v$$

  - Where $v = c_h + c_j + \epsilon_{it}$
- If $Cov(X_{it}, c_h) \neq 0$ or $Cov(X_{it}, c_j) \neq 0$ then our results will still remain biased
  - Even if we could control for c_h and c_j, the grouped nature of the data violates standard independence assumptions

The Multilevel Model

- Level 1: $Y_{it} = \alpha X_{it} + \beta T_{it} + \zeta c_i + \epsilon_{it}$
- Level 2: $c_i = \xi c_h + \epsilon_i$
- Level 3: $c_h = \varrho c_j + \epsilon_h$
- Level 4: $c_j = \mu + \epsilon_j$

- Includes a unique intercept term for each unique parcel, household, and village
  - Allows intercepts to vary based on which group the data comes from

Varying-Intercept and Vary-Slope Models

- The multilevel framework can accommodate a variety of data structures