Lecture 12: Panel Data (Part II)

Jaakko Meriläinen

5304 Econometrics @ Stockholm School of Economics

Extending Panel Data Models

- In the previous lecture, we saw the following models looking at variation over time:
  - Difference-in-differences
  - First-differences
  - Fixed effects

- We have focused on recovering causal effects and at how panel variation may help us significantly in removing various sources of potential bias

Extending Panel Data Models

- In this lecture, we will continue looking at panel data analyses

- Focus on specific cases that arise frequently in the analysis of causal effects:
  - Models with lagged dependent variables
  - Attrition in panel data
- For the most part, the purpose of this is to show the use of such models in practice/illustrate issues that arise and the questions they can help solve

Plan for Today

1. Introduction
2. FE recap
3. Lagged dependent variables
4. Attrition in panel data
5. References

FE Reprise

- We have introduced the FE estimator in the context of panel data
  - The estimator "de-means" the outcome and all regressors over time
  - This is a useful way of removing the unobserved, additive, levels-effect
- Note that there is nothing in the set-up of the model that requires the variation to be over time
- The method is generally applicable to settings where there is some group variable for which we are worried about unobserved fixed effects!

An Example

- For instance, imagine a cross-sectional household survey data set which collects information on
  - Household income (Y_{hv}) where h indexes households, v indexes villages
  - Some household characteristics (X_{hv}), such as demographics or number of adults with higher education
- The model we have in mind is: $Y_{hv} = \alpha_v + X_{hv}\beta + \epsilon_{hv}$
- You can use the within transformation to get rid of the village fixed effect α_v
- Instead of de-meaning over time for each individual, we will be de-meaning across individuals within a village
- In various settings you would e.g. include state fixed effects and/or time fixed effects

Multiple Sets of Fixed Effects

- There is no specific need to stick with just one set of fixed effects!
- In fact, often we would want to include multiple sets of fixed effects
- In a panel dataset, for example, the variable(s) of interest may vary both over groups in the cross-section and over time
  - E.g. whether a policy was implemented varies across states/counties
  - And over time (when, within-state, it was implemented in some periods but not others)
  - The key in FE specifications is to think carefully at which levels the variable of interest varies
- Thus, in regression difference-in-differences models, we often include state and year fixed effects

Generalized Difference-in-Differences: An Example

- Historically, it has been speculated that extending voting rights to women has contributed to the expansion of the public sector in many countries
- In the United States, different states did this in different points of time
- We could estimate the following regression to understand the effects of franchise extension on government spending:

$$y_{st} = \lambda_s + \lambda_t + \beta \textit{Female voting rights}_{st} + \epsilon_{st}$$

where λ_s are the state fixed effects, λ_t are the year fixed effects, and $\textit{Female voting}_{st}$ is an indicator telling whether state s allowed women to vote in year t (NB. can change at different times for different states!)

- This is sometimes called a "generalized difference-in-differences" or a "two-way fixed effects" specification

"Generalized Difference-in-Differences": Parallel Trends Assumption?

- We still need to make the parallel trends assumption

- Two easy ways to provide supporting evidence:
  1. Control for state-specific time trends—if the results do not change by much, this is good
  2. Estimate an event-study specification:

$$y_{st} = \lambda_s + \lambda_t + \sum_{\tau=T_0}^{-1} \beta_\tau D_{\tau s} + \sum_{\tau=-1}^{T_1} \beta_\tau D_{\tau s} + \epsilon_{st}$$

- Good to know: if treatment timing varies across units and/or if the treatment switches on and off, the basic two-way fixed effects estimation approach may have problems
  - These are too complicated to be discussed in this course
  - But important to keep in mind: also examine event-study-type specifications (can be implemented in various ways—explore robustness!)

Models with Lagged Dependent Variables

- Up until now, we have only used the panel variation in the data to remove time-invariant unobserved effects (FD and FE models)
- But panel data are important for more than just identification
- In particular, panel data allow us to study dynamics
- One particular class of models which we have not spoken about is where outcomes depend on the level of outcomes in the previous period, i.e. where current outcomes depend on past levels
- As we will see, this leads to further econometric issues

Models with Lagged Dependent Variables

- Imagine the following specification:

$$Y_{it} = \alpha_i + \lambda_t + \delta D_{it} + \beta X_{it} + \epsilon_{it}$$

  where D_{it} is a binary treatment indicator and δ is parameter of interest

- The individual effect α_i is unobserved
- We suspect D is correlated with the unobserved α_i
- The fixed effects model can recover δ consistently under the assumption that conditional on the fixed effects and X_{it}, D_{it} is uncorrelated with the error term
- So far, so good...

Models with Lagged Dependent Variables

- But suppose the model we have in mind is instead:

$$Y_{it} = \alpha + \lambda_t + \rho Y_{i,t-1} + \delta D_{it} + \beta X_{it} + \epsilon_{it}$$

- δ can be recovered if D_{it} is uncorrelated with the error term conditional on our covariates

- Unfortunately, however, this is unlikely to be true
  - Any persistent effect in ϵ_{it} (imagine an α_i) will be correlated with $Y_{i,t-1}$
  - So $\widehat{\rho}$ is almost certainly biased

Models with Lagged Dependent Variables

- It is not just $\widehat{\rho}$ we need to be worried about
- If the estimator for one variable is biased, that bias can transmit to all other variables...
- Whether $\widehat{\delta}$ will biased as well depends on the correlation between $Y_{i,t-1}$ and D_{it}
- In the special case that D_{it} is conditionally exogenous given X_{it} alone, the bias in $\widehat{\rho}$ does not carry over to $\widehat{\delta}$
  - This is why it is okay to control for lagged dependent variables in an RCT and not worry about it
  - But this is a rare occurrence!

Models with Lagged Dependent Variables

- If $T \geq 3$, we could have fixed effects and a lagged dependent variable
- Imagine the following model:

$$Y_{it} = \alpha_i + \lambda_t + \rho Y_{i,t-1} + \delta D_{it} + \beta X_{it} + \epsilon_{it}$$

- Difference to remove the α_i

$$\Delta Y_{it} = \Delta\lambda_t + \rho \Delta Y_{i,t-1} + \delta \Delta D_{it} + \beta \Delta X_{it} + \Delta\epsilon_{it}$$

- The unfortunate problem is that $\Delta Y_{i,t-1}$ is necessarily correlated with $\Delta\epsilon_{it}$
  - Why? They are both functions of $\epsilon_{i,t-1}$!

Models with Lagged Dependent Variables

- What does the above imply?

- If we are genuinely interested in ρ, we need an appropriate IV
  - I.e. if the dynamics are what we are interested in, we need to deal with the endogeneity issue in all lagged dependent variables
  - This could come from specific shocks etc. in the past
  - A class of models further use more distant lags or lagged differences as instruments (dynamic panel models, GMM)—more advanced than we will cover here

- If we are only interested in δ and think of $Y_{i,t-1}$ as a control...
  - We are still not off the hook in general: bias in one coefficient affects others
  - But it may be possible to recover a consistent estimator for δ without needing a consistent estimator for ρ

Attrition in Panel Data

- Note we had assumed that we had a random subsample of the population in the first wave of the panel
  - And, implicitly, we have assumed that our panel is balanced
  - I.e., all units seen in Wave 1 are seen in subsequent waves
- A common issue in panel data analyses is the **attrition** of individual units
- Of your initial sample, you may only observe a subset in subsequent rounds
- For instance:
  - In a household survey, perhaps some households move away and are untraceable
  - In a firm survey, some firms go out of business
  - In a student survey, not everyone gives the final test

Attrition in Panel Data

- Attrition can be a **huge** problem in panel data analyses
- Let us focus on the case where we are interested in the treatment effects from a binary treatment
- The problem is essentially one of being left with a selected population
- This is best illustrated with an example

Attrition in Panel Data: An Example

- Imagine we have a population of 1000 households
- In Period 1, we randomly select 500 households to be given a large cash transfer; let us assume no spillovers
- We are interested in the effects on labor supply of adults in this household in Period 2 (after 1 year)
- If we can survey all 1000 households, we can just compare the treatment and control groups
- Randomization deals with the standard selection problem

Attrition in Panel Data: An Example

- But suppose we go back to the village after 1 year...
- ...and find 500 of our control group but only 300 of our treatment group
- This is no longer a random sample!
- The initial randomization does not ensure we have dealt with the selection problem!
- Similarly, suppose we had given management training to 50 firms out of 100
  - We want to see the effects on firm profits after 3 years
  - But after 3 years, 20 out of 50 control firms have gone out of business!
  - Comparing the profits of the firms that survive is not a consistent estimator of the treatment effect!
- In general, whether attrition poses a problem, and how severe it is, depends on how the ones who leave are selected

Attrition in Panel Data: How Are Leavers Selected?

- If observations are **missing at random**, this is not a problem
  - Restrict the sample to individuals in both rounds
  - It is still a random sample of the population, the randomization of the treatment still holds, the treatment effect is still unbiased

- If **missing not at random but based on observables**, then it is possible to deal with this
  - Estimate a parametric model that predicts the probability of being observed in the second round
  - Typically with a probit/logit model
  - Re-weight the regressions by this probability

Attrition in Panel Data: How Are Leavers Selected?

- If **missing not at random and based on unobservables**, this is much more complicated
- It is possible, with further assumptions, to bound the treatment effects
- Or we can try to explicitly model selection
  - The classic example is the "Heckman selection model"
  - For convincing identification, requires variables that affect selection but do not affect the outcome
  - These are hard to find (just like IVs!)
- Both these approaches are more advanced than we will cover in this course

Recap of This Lecture

1. Introduction
2. FE recap
3. Lagged dependent variables
4. Attrition in panel data
5. References

Recommended Readings

- Chapters 13 and 14 in Wooldridge (2013)
- Chapter 5 in Angrist and Pischke (2009)—or Chapters 8 and 9 in Cunningham (2021)