Applied Panel Data Analysis – Lecture 2

Christopher F. Parmeter

AGRODEP
September 9-13th, 2013
Dakar, Senegal

- In the previous lecture we saw how the pooled OLS estimator can be used to model panel data
- This estimator had well established statistical properties
- This estimator does not exploit the panel structure

- In this lecture we will discuss unobserved heterogeneity
- We will learn about the fixed and random effects frameworks
- Attention will be paid to the underlying assumptions necessary for these models to be conceptually plausible from an economic point of view

- The pooled estimator is easy to work with and has desirable statistical properties
- However, there are some unfortunate consequences associated with the assumptions underlying this estimator
- The main issues concern the variance-covariance structure of the error terms and the plausibility of the exogeneity assumption regarding the errors and the covariates

- Lets think about an unobserved variable c_i that enters into our basic panel model as

$$y_{it} = x'_{it}\beta + c_i + \varepsilon_{it} \tag{1}$$

- Our primary concern is the conditional mean of y on x
- However, with the presence of c_i, it is not clear how to interpret β with a *ceteris paribus* effect when we do not control for c_i
- You may see a Greek letter used for c_i in applied papers (such as α); I think using this notation is clearer because this unobserved variable is a random variable and not a parameter

- Why is this important?
- With our assumed structure in (1) interest clearly hinges on β, the $K \times 1$ vector of response effects
- If c is uncorrelated with each x then it makes up another component of ε and we do not have much to worry about regarding estimation
- If $Cov(x, c) \neq 0$ for some covariate, then failing to control for c can lead to serious estimation problems

- How might we control for c when $Cov(x, c) \neq 0$?
- We could find a proxy for c
- We could find instrumental variables for those elements of x that are correlated with c
- Neither of these approaches is appealing in a panel data context
- When we observe the same cross-sectional units at different time periods we have alternative options beyond the standard approaches available for cross-sectional datasets

- Suppose for the moment that c was time constant but varied across individuals in our panel
- Further, suppose that we have two time periods
- Our model for each time period is

$$\begin{aligned} y_{i1} = & x_{i1}\beta_1 + c_i + \varepsilon_{i1} \\ y_{i2} = & x_{i2}\beta_1 + c_i + \varepsilon_{i2} \end{aligned}$$

- We will also assume that $E\left[\varepsilon_{i1}|x_{i1}, c\right] = 0$ and $E\left[\varepsilon_{i2}|x_{i2}, c\right] = 0$

- If we subtract period 1 from period 2 then we have

$$\triangle y_i = \triangle x_i' \beta + \triangle \varepsilon_i \tag{2}$$

  which is a cross-section model and the presence of c has been eliminated
- Do we need additional assumptions to consistently estimate β?

- Generic OLS estimation of this first differenced model requires that $E\left[\triangle\varepsilon|\triangle x\right]=0$
- This condition is equivalent to

$$E\left[\varepsilon_2|x_1,x_2\right]-E\left[\varepsilon_1|x_1,x_2\right] \tag{3}$$

- For this expectation to be 0 we need our covariates to be strictly exogenous, a much stronger condition than we needed in the pooled panel data model or in our initial setup for our linear panel data model

- This stronger condition is a necessary tradeoff for allowing unobserved, time constant heterogeneity into the model
- This added flexibility comes at the cost of a more restrictive assumption between the observable variables and the error component in our model
- Note that in this setup we did not have to specify how c and the elements of x were correlated
- Notice that with the time differencing, any variable that is constant over time is eliminated from the model (we cannot recover a β for this variable)

- A key question when constructing the linear panel data model is whether we should think of the unobservable variables as fixed or random
- While it might seem odd to think of the random variable c as fixed, this terminology is heavily entrenched in econometric parlance and would be counterproductive to deviate

- Lets assume for this discussion that c is constant over time, but can differ across individuals
- The unobserved effects panel data model is

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it} \tag{4}$$

- x_{it} can contain variables that vary over i and t (GDP per capita), variables that vary over t but not i (a shock to the oil supply in a particular year) and variables that vary over i but not t (the latitude of a country)
- Given that c_i varies over individuals it is commonly referred to as individual heterogeneity or as an individual effect

- You must be judicious in reading panel data papers of a particular vintage
- When an author says that c_i is a random effect they are treating c_i as a random variable
- When an author says that c_i is a fixed effect they are treating c_i as a parameter to be estimated
- The literature has evolved in our understanding of how to appropriately treat c_i
- c_i will always be a random variable whether it is treated as a fixed or random effect

- Modern econometrics uses the terminology random effect framework to mean

$$Cov(x_{it}, c_i) = 0 \text{ or } E[c_i | x_{i1}, x_{i2}, \ldots, x_{iT}] = E[c_i] \qquad (5)$$

- Modern econometrics uses the terminology fixed effect framework to mean

$$Cov(x_{it}, c_i) \neq 0 \qquad (6)$$

- Proper use of these terms will help you in conceptualizing the appropriate model, in interpreting your estimates, and in staying current with the terminology when you write technical papers

- We need to discuss how the strict exogeneity assumption plays out for the unobserved effects panel data model
- Our condition is

$$E[y_{it}|x_{i1}, x_{i2}, \ldots, x_{iT}, c_i] = E[y_{it}|x_{it}, c_i] \tag{7}$$

- Once we control for x_{it} and c_i, x_{is} for $s \neq t$ plays no role in explaining y_{it}
- We term this condition strict exogeneity conditional on the unobserved effect

- Compare strict exogeneity to strict exogeneity conditional on the unobserved effect

$$E[y_{it}|x_{i1}, x_{i2}, \ldots, x_{iT}] = E[y_{it}|x_{it}] \tag{8}$$

- What this means is that strict exogeneity would fail if $E[c_i|x_{i1}, x_{i2}, \ldots, x_{iT}] \neq E[c_i]$, i.e. c_i is a fixed effect

- Strict exogeneity conditional on the unobserved effect also means

$$E[\varepsilon_{it}|x_{i1}, x_{i2}, \ldots, x_{iT}, c_i] = 0 \tag{9}$$

- This implies that

$$E[x'_{is}\varepsilon_{it}] = 0 \tag{10}$$

which is much stronger than just assuming contemporaneous exogeneity

- But if we have contemporaneous exogeneity then we cannot have a fixed effect framework, so this is our statistical tradeoff

- Suppose output is tons of soybeans produced by farms and our covariates contain capital, labor, materials and rainfall
- We can think of the unobserved effect as capturing land quality and the farmer's innate ability
- Strict exogeneity conditional on the unobserved effect is a more plausible assumption than strict exogeneity because we can think of the farm's inputs being contingent on both land quality and the farmer's ability
- We expect that if we do not condition on c_i then input use in one period will be correlated with output in a different time period

- When considering a panel data application your initial focus should be on two questions:
  - Is the unobserved effect correlated with x_{it}?
  - Is the strict exogeneity conditional on the unobserved effect condition plausible?

- Panel data offers additional modeling flexibility to the practitioner, allow for unobserved heterogeneity
- Controlling unobserved time constant or individual constant heterogeneity is possible with panel data
- Important to distinguish between 'fixed' and 'random' effects in the standard linear panel data model
- Plausibility of strict exogeneity conditional on the unobserved effect