

**Applied Econometrics**  
**Prof. Sabuj Kumar Mandal**  
**Department of Humanities and Social Sciences**  
**Indian Institute of Technology, Madras**

**Lecture - 26**  
**Pooled Data and Panel Data Model Estimation – Part VI**

(Refer Slide Time: 00:15)

Panel data  
 Macro Panel:  $N \gg T$   
 Micro Panel:  $T \gg N$   
 FE why?

LSDV:  $y_{it} = \beta x_{it} + \alpha_i + u_{it} \rightarrow$  One-way error component model  
 $\alpha_i$ : a parameter to be estimated  
 $\hat{\beta}_{LSDV} = \hat{\beta}_{FE}$

① LSDV is practically very difficult to estimate when  $n \rightarrow \infty$   
 ② The idea by which  $\alpha_i$  is considered to be a parameter to be estimate is also very weak. For a finite fixed  $T$  if we increase  $N$ , we actually do not accumulate additional information for  $\alpha_i$  to be consistent.

$y_{it} = \beta x_{it} + \alpha_i + u_i + u_{it} \rightarrow$  Two way error component model  
 $\alpha_i$ : individual fixed effect  
 $u_i$ : Time fixed effect

Now this model

$$y_{it} = \beta x_{it} + \alpha_i + u_{it}$$

in this model we assume that there is unobserved effect and which is individual specific. So, that means individuals are different, so this is called one way error component model. Suppose, we want to accommodate a fact that it is not only the individuals are heterogeneous, but the years are also heterogeneous. For example, you are estimating the model in a specific context where the years are also very different.

For example, some major macroeconomic policies might have taken place in that place, let us say that I am estimating firm performance in a period after demonetization. So, obviously 2016 and 2017, 18, 19 they had be quite different or 2010, 11, 12, 13 and 14 would be quite different from 16, 17, 18, 19. So, in that context when you feel that there might be some kind of year specific or time heterogeneity also.

In the situation then it is better to include onetime specific factor also in the model and make the model two-way error component model. So, what would be the two error component model

$$y_{it} = \beta x_{it} + \alpha_i + \mu_t + u_{it}$$

So, this is called individual specific unobserved effect, this is called time specific unobserved effect, so this is called time fixed effect, this is called individual fixed effect.

So, how will you estimate that type of model? To estimate a two error component model you have to keep in mind this is called two-way error component model. So, again the application of LSDV only will help us estimating two-way error component model.

**(Video Starts: 03:26)**

So, you have to use this command reg then after that i dot year, so that means I am basically adding another set of dummies, one intercept for the year as well different years. So, we have three years data 17, 18, 19, 17 has been sorry 87, 88, 89, 87 is actually used as the base year. So, this model is called two-way error component model. Now once we estimate either one way or two-way error component model the assumption what we make that there is individual specific heterogeneity individuals are different.

So,  $\alpha_i$  factor  $\mu_t$  factor or we are including in our model, they are jointly significant that is the assumption. So, if the individual there is no significant unobserved heterogeneity across individual then, instead of applying this fixed effect model either one way or two way it is better to apply pooled OLS is not it? Please try to understand we have combined a cross section and time series then there are two options either to use pooled OLS or fixed effect.

**(Video Ends: 05:33)**

**(Refer Slide Time: 05:35)**

There are two options basically if we think. When you have combined data on cross section and time series, there are basically two options, either we use pooled OLS or we use fixed effect model fixed effect or LSDV the older version. Now, when to use FE or LSDV? When  $a_i$  is significant. when,  $a_i$  is insignificant.

That means how do you check whether  $a_i$  is significant or not that means we need to apply a test to check whether  $a_1 = a_2 = \dots = a_{n-1} = 0$  or not or if it is a 2 error component model then  $t_1 = t_2 = \dots = t_{n-1} = 0$  or not. So, we need to check whether there is a presence of significant individual specifications or your specific heterogeneity. If we do not check this then it is not reasonable to apply fixed effect transformation because I have told you repeatedly that OLS model is the most perfect and powerful model in econometrics.

Unless it is required, we should never deviate from the OLS model. If there is no significant  $a_i$  factored in the model then we should actually use a pooled OLS model, we combine the data and then simply run a pooled OLS model on that.

**(Video Starts: 08:27)**

So, when you estimate the model fixed effect model if you go back our fixed effect estimate, this is our fixed effect estimate. Now, stata is reporting that f statistic also to check whether there is a presence of significant  $a_i$ , look at this this is the f test. What is the f test? This is the f test where it is showing f test that all  $u_i$ 's are 0. That is what we are testing so this is 53 and 106 and 24.59.

So, how this f statistic is calculated? Basically, if you recall from our basic econometrics stata is actually estimating 2 versions of this model one is called restricted another one is called unrestricted. So, unrestricted means you have no restriction on  $\alpha_i$ , so you estimate a model using the one intercept for one individual and the restricted model is basically you put the restriction a  $\alpha_1 = \alpha_2 = \dots n - 1 = 0$

Then, you calculate this f statistic based on the 2  $R^2$ , so  $R^2$  restricted -  $R^2$  unrestricted that model that formula I am not again going back to the formula the idea is this. So, we need to check actually the validity of the restriction, the restriction is  $\alpha_1 = \alpha_2 = \dots n - 1 = 0$ , that is the restriction. We will first estimate the unrestricted one then we will estimate the restricted one will calculate the  $R^2$ .

We will compute the statistic and this is the F statistics data is showing you. The value is 20, 4.59, 24.59 and corresponding if value is  $f = 0.000$ , so which is quite significant.

**(Video Ends: 10:33)**

So, that means in this case f statistic = 24.59 and P evaluate is actually 0.000, so that means this is significant  $\alpha_i$  and that is why the situation calls for fixed effect model to be estimated not pulled OLS. So, stata is showing that also, using your data.


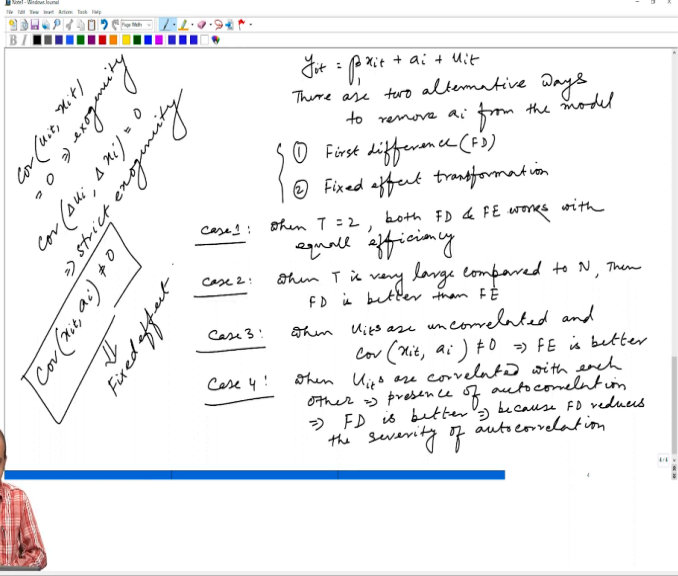
**(Video Starts: 11:08)**

So, that means from the output we have now learned many things we have learned within between overall  $R^2$ , we have learned how to estimate this model, we have learned what is the interpretation of this constant term which is basically the average heterogeneity in the sample. We have also learned the significance of this f statistic which is basically testing the presence of significant amount of unobserved heterogeneity.

There are still  $\sigma_u$   $\sigma_e$  and  $\rho$  which is which we have not yet explained which we will discuss in our next class.

**(Video Ends: 11:59)**

**(Refer Slide Time: 12:00)**

$y_{it} = \beta x_{it} + a_i + u_{it}$   
 There are two alternative ways to remove  $a_i$  from the model  
 ① First difference (FD)  
 ② Fixed effect transformation  
 Case 1: when  $T = 2$ , both FD & FE works with equal efficiency  
 Case 2: when  $T$  is very large compared to  $N$ , then FD is better than FE  
 Case 3: when  $u_{it}$  are uncorrelated and  $Cov(x_{it}, a_i) \neq 0 \Rightarrow$  FE is better  
 Case 4: when  $u_{it}$  are correlated with each other  $\Rightarrow$  presence of autocorrelation  $\Rightarrow$  FD is better  $\Rightarrow$  because FD reduces the severity of autocorrelation  
 $Cov(u_{it}, x_{it}) = 0 \Rightarrow$  exogeneity  
 $Cov(a_i, \Delta x_{it}) = 0 \Rightarrow$  strict exogeneity  
 $Cov(x_{it}, a_i) \neq 0$   
 Fixed effect

Before we wind up that means so, that means if our model is

$$y_{it} = \beta x_{it} + a_i + u_{it}$$

there are two alternative ways to remove  $a_i$  from the model. And based on which particular way we adopt we will get, if you recall one is called first difference and another one is called fixed effect transformation. Now the question is which particular model to take, first difference or the fixed effect?

Now there are several cases case one when  $T =$  just 2, that means we have only two periods' data both FD (First Differencing) and FE(Fixed Effect) works with equal efficiency. So, you can use either first difference or fixed effect when your time period is only 2. Case 2, when  $T$  is very large compared to  $n$  then FD is better, why FD is better? Because when  $T$  is large as I already mentioned the assumption that  $a_i$  is basically fixed or constant occur over a period of time that assumption is very difficult to maintain.

So, FD is better than FE. Case 3, when  $u_{it}$  are uncorrelated and

$$Cov(x_{it}, a_{it}) \neq 0$$

that means we assume that when  $a_{it}$ 's are uncorrelated and this  $x_{it}$  the explanatory variable is correlated with the unobserved heterogeneity then FE is better. So, that means for applying fixed effect model we need two things.

First of all, we assume that this unobserved effect is actually correlated with the  $x_{it}$  and the error term the idiosyncratic error  $u_{it}$  they are also uncorrelated with each other. But what happens when these  $u_{it}$ 's are actually correlated with each other? That means presence of autocorrelation then FD is better first difference. Why, because if you recall from our basic econometrics that one of the solution by which you can actually reduce the severity of autocorrelation is first difference.

So, the moment we take first difference the severity of autocorrelation goes down FD is better because FD, first difference reduces the severity of autocorrelation. So, in general we need to then understand that when time period is small, we have large or macro panel in is very large unobserved effect is correlated with one or more explanatory variable,  $u_{it}$ 's are actually uncorrelated.

When we say that  $u_{it}$ 's are actually uncorrelated? That means we are basically talking about strict exogeneity. So, that means we are ensuring this is  $a_i$  or you can say that  $a_i$  is also another component, so sometimes we say that this is  $u_{it}$ . So, that means strict exogeneity

$Cov(u_{it}, x_{it}) = 0$  then this is simple and exogeneity but when

$Cov(\Delta u_{it}, \Delta x_{it}) = 0$  then that is called strict exogeneity.

So, for FE to be implemented we must ensure there is strict exogeneity, so that means in the set of explanatory variables it should not include any lag dependent variable like  $y_{it-1}$ . So, these are the situation, conditions that must be satisfied for the fixed effect model to be employed. Model should be large panel  $n$  is quite larger than  $t$ ,  $u_{it}$ 's are uncorrelated unobserved effect is correlated with  $x_{it}$  there is strict exogeneity.

Otherwise, we need to include we need to estimate the model using first difference when  $T$  is large. That means, when  $T$  is large basically, we are adding more of time series component in the panel. And if you add inject mode of time series component in the model, we need to additionally check for time series properties in the panel setup that will discuss later. So, in that case FD works better because that reduces at least the severity of autocorrelation.

When  $u_{it}$ 's are actually correlated with each other. So, far then we have discussed about mainly fixed effect model and the major emphasis main assumption that we make is the unobserved, this is the main assumption that  $x_{it}$  is actually correlated with the unobserved effect model. So, if this is the case when you are unobserved effect is actually correlated with some of your explanatory variable that is the idealistic situation for fixed effect model to be employed.

And we need to remove this  $a_i$  from the model because if you do not remove that will lead to the endogenous problem. But sometimes it may so happen that in your model there is unobserved heterogeneity. But that unobserved heterogeneity is actually not correlated with the explanatory variable. What do you do in that context? Do we really need to eliminate  $a_i$ ? There is unobserved heterogeneity we cannot actually measure that  $a_i$  that is there in the model.

So, if we do not eliminate this  $a_i$ ,  $a_i$  will be there in this  $u_{it}$  and for each and every  $u_{it}$  that means  $u_{it,1}$ ,  $u_{it,2}$ ,  $u_{it,3}$  for each period this  $a_i$  component would be come. So,  $u_{it}$  will show obviously autocorrelation problem. So, from that logic apparently, we may think of removing this  $a_i$  but if we when we  $a_i$  is actually not correlated with  $x_{it}$ , a removing  $a_i$  is actually not efficient.

Because the idea was this would be correlated with  $a_i$  that is why to solve endogeneity, we remove this. So, endogeneity problem is not there, so an endogeneity problem is not there why do you take the fixed effect transformation. So, that means in a situation where there is a presence of significant  $a_i$  but that is uncorrelated with  $x_{it}$  we are in a dilemma whether to remove  $a_i$  or not to remove  $a_i$ .

If we remove  $a_i$  there is a problem because it is actually not correlated and you are going for that fixed effect transformation. So, that will call for inefficiency in the mod estimates, if you do not remove  $a_i$  then what will happen it will be there and sitting in  $u_{it}$  and obviously that

$u_{it}$ 's will be correlated with other period  $u_{it}$  so, it will call for autocorrelation. So, that means the solution what we talked about so far is not applicable in that context when this is actually 0. We need to discuss a different type of panel data model which is called random effect and a data model that we will discuss in our next class.