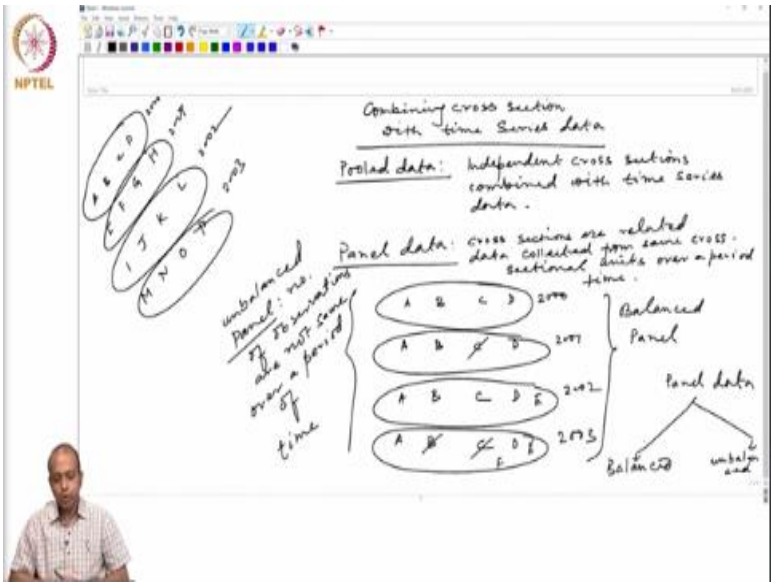
Applied Econometrics
Prof. Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras

Lecture - 23
Pooled Data and Panel Data Model Estimation – Part III

(Refer Slide Time: 00:16)



Welcome once again to our discussion on pool data. Yesterday we are talking about combining cross-sectional data with time series 1 and we said that when we combine independent randomly collected independent cross sections collected from a given population over a period of time, then we said that that is basically a pool data. Pooling cross section with time series that is called pool data when the cross sections are independent.

So, we are discussing combining cross-section with time series data. So, yesterday we talked about pool data and we say that this is basically combining independent cross section. So, here the when we combine the cross section, we say that independent cross sections combined with time series data. So, for example we can say that let us say this is in year first year we have collected data from these four firms.

Let us say this is year 2000 then in 2001 we are again collecting data from another four firms from the same industry let us say sorry E F G H so, another one another cross section in 2001

then I J K L in 2002 M N O P 2003. Now if you look at these cross sections are actually independent because there is no commonality among the elements of these cross sections in year 2000 it is A B C D form in 2001 it is E F G H in 2002 it is I J K L and in 2003 it is M N O P.

So, that means we have collected samples from different forms we have collected some information from different forms at different point of time. Now today we are going to discuss about another type of specific pool data wherein we are collecting information from the same individual and if that is the case, we say that is called panel data. So, in panel data what you are doing?

It is A B C D in 2000 then in 2001 also it is A B C D then in 2002 also it is A B C D and it is A B C D in 2003. This is called panel data. So, that means here the cross sections are actually not independent because I am collecting data from the same individuals or same country or same states or same forms over a period of four years. So, here cross sections are not independent sections are related data collected from same cross-sectional units over a period of time.

Now what is the need of this type of panel data? Yesterday while talking about pool data we said that sometimes we work with combining cross-sectional data with time series and make it as a pool because of its advent, what is the advantage we get in pool data? Instead of working on a single distribution for the dependent and independent variables in pool data we get multiple distributions of dependent and independent variables.

So, information contained in pool data is much higher compared to the cross sectional data and that makes the estimates derived from a pool data much more efficient and reliable compared to the cross-sectional estimates. Today we will try to understand another advantage of working with this specific panel data specific pool data that is called panel. So, you have to keep in mind that panel data is also a pool data it is a specific kind of pooling.

Instead of pulling independent cross section in case of panel we are actually pooling same cross sections over a period of time. Sometimes this is also again a specific type of panel where the number of observations are same and it is the same individuals or firms. So, A B C D, A B C D,

A B C D and A B C D. So, that is why when in each year you have same number of observation this is called balanced panel.

But there is no guarantee that in each year you would be able to observe same individual or same firm because collecting a panel data is not a very easy task. So, for example, in 2000 we have data from A B C D it may so happen that in 2001 actually this form C exited from the market or somehow the data for the C firm is not available. So, in 2001 it may so happen that we have data on A B and D.

Similarly, in 2002 it may so happen that this C firm came back once again additionally we may have data on E form as well. And in 2003 we may not get again data on B and C rather we may have E and F this may also happen. So, that means in this type of pooling at least some individuals or some entities are common and there is some new element added in that distribution in that cross-sectional units.

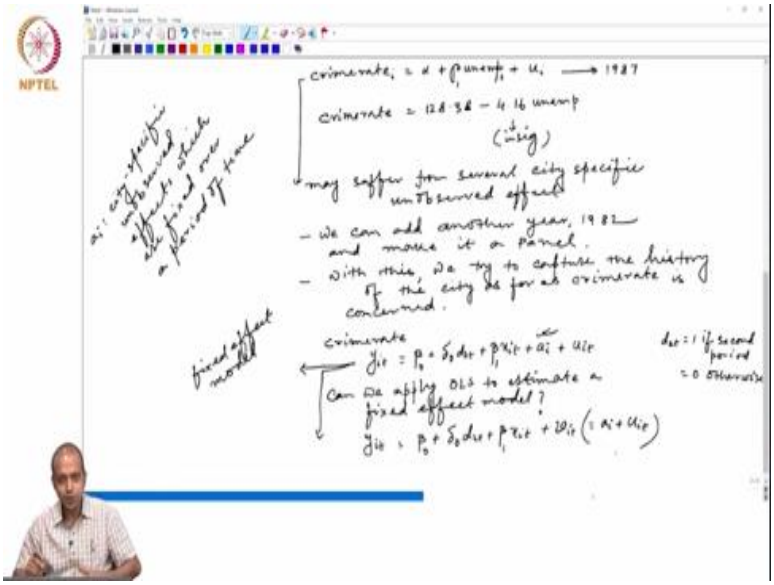
So, what will happen in that case when number of observations are not same this is called unbalanced panel. So, number of observations are not same over a period of time. So, that means now we can understand the panel data. So, panel data is of two type one is called balanced panel so this is called balanced and this is unbalanced panel, two specific type of panel data we have. So, when we have unbalanced panel there is a tendency generally.

What we do? We remove the uncommon elements and try to make it a balanced panel drawn right for that because you have to keep in mind in econometrics one observation means one piece of information. And information is very, very valuable in econometric estimation higher is the better. Higher the information contain better would be the value of those estimates. So, that is why no need of throwing out any piece of information no need of unnecessarily making an unbalanced panel a balanced one.

So, we have to take information as it is and then we will work with either balance panel or unbalanced panel depending on the situation because we have econometric literature for both type of panels no need of making the unbalanced a balanced one, that is what I wanted to say.

Then what we will do? We will try to understand another important objectives of this pool data or panel data.

(Refer Slide Time: 11:33)



So, let us take an example, suppose we are interested in estimating the relationship between crime rate equals to let us say we have estimated a relationship between crime rate and unemployment rate so alpha plus beta 1 let us say this is unemployment rate. So, let us say we have data only on 1987. So, for one period this is for 1987 so we have data on crime rate we have data on unemployment rate and this is only available city wise data.

We have data on several cities about their unemployment rate and crime rate only for a specific year and what we hypothesize that unemployment rate and crime rate they are negatively related, which is quite logical because as unemployment rate increases so what will happen is sorry they are positively correlated higher is the unemployment rate higher would be the crime rate as well because if people are not employed then they will be involved in several criminal activities.

That is what the literature says and that is what is our assumption is. But when we estimate this type of equation using a single cross section it may so happen that we may get this type of relationship. Crime rate equals to let us say 128.38 - 4.16 unemployment. It may so happen, so that means here we are getting a negative relationship between unemployment and the crime rate and this may not be insignificant this may be insignificant also.

So, what is happening when we are working with a single cross section first of all the coefficient of unemployment it is giving a wrong sign it is not as we are expecting because as unemployment increasing crime rate is decreasing which is difficult for us to believe. So, this is against the theory and secondly this variable this coefficient is not significant also. Then the question is why is it happening?

It is happening because in this simple relationship crime rate and unemployment rate this equation may suffer from several cities specific unobserved effect. This is a simple relationship so you can include many variable city-specific variables which are observed so you can make it as multiple linear regression model like what is the distribution of education, average education in that city, what is the religion, what is the income status so on and so forth.

But there are certain city specific features which are not directly observable maybe the geographical location may be the history of that city for example if we collect crime rate across several Indian cities then it may so happen let us say that let us assume that Mumbai shows the highest crime rate. Then you can include several variables like unemployment rate, religion, average education, average income so on and so forth but there are certain features of this Mumbai city which is unobservable.

It may so happen that historically the nature of the city itself is such that there is high crime rate which is difficult to explain. So, in that case what we can do instead of running the regression only on a particular year 1987 we can add another specific or let us say in 2002. So, what we can do? We can add another year let us say 1982 and make it a panel. Now when I am adding one previous year observation on crime rate unemployment rate basically what I am trying to do.

By this so, with this we try to capture the history of the city as far as crime rate is concerned. So, we have two periods' data 1982, 1987 and we have collected information from the same cities. So, as a result of which these will become a pooled a specific type of pool which is panel data and adding the previous year's information makes this data captioning the history of the cities as far as this crime rate is concerned.

So, this is one way we can this is another advantage of capturing another advantage of panel data to capture history of the city which are unobserved in terms of city specific information. Now what we will do? There is another way of looking at this there. Now there is another way of looking at this model let us say that in this equation we are modifying the equation as crime rate.

Or let us say the crime rate is denoted by now y it = beta 0 + delta 0 dT d or let us say d 2t + beta 1 x it + a i + u it, well d 2t = 1 if second period 0 otherwise. Now what is the additional feature of this model? Here look at this I have included a factor called a i, and what is a i? This a i is called city specific unobserved effect which are fixed over a period of time. So, that is why a i is called individual specific time fixed effect.

That means while modelling crime rate and unemployment rate we accommodate the fact that several cities they have certain city specific features which are difficult to observe but that does not change over a period of time. If that is the case then this would be our modelling. So, look at this here we have included one year dummy d 2t to represent the fact that at different point of time distribution of this y it and x it they are different.

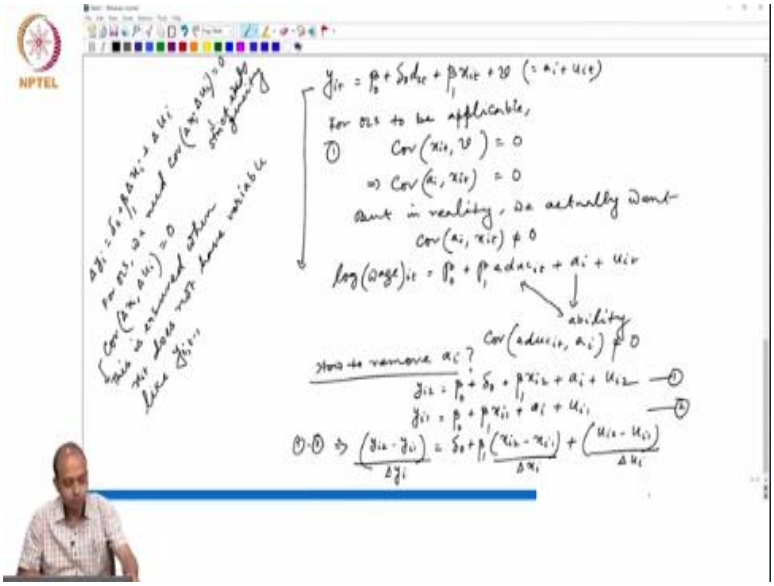
We also accommodate the fact that there are several cities they have city specific unobserved effect and that is captured by a i. So, this should be our modelling of a simple panel data which is called fixed effect model. Why this is called fixed effect? Because we have included this is called a fixed effect model. What is fixed here? The city specific features are fixed over a period of time that is why it is called fixed effect model.

Now the question is can we apply OLS to estimate a fixed effect model. Now to estimate a fixed a model using OLS what is required that so now we can modify this equation as y it = beta 0 + delta 0 d 2t + beta 1 x it plus let us say v it which is equals to a i + u it. So, v i is basically a composite error term consisting of the unobserved effect plus this idiosyncratic error. We have the idiosyncratic error u it and a i so, both are actually unobserved.

So, a i is also like an error term the only difference between a i and u it, a i varies across individuals across cross section it does not change over time u i t it varies across individual as well as over a period of time. But both are unobserved effect which has impact on the crime rate we need to understand the difference between these two a i and u it both are like error term both are unobserved effects that have impact on crime rate.

While u it changes over a period of time that is why I have t subscript as well I have not given any t subscript here to represent the fact that a i is basically individual specific time constant it is not changing over a period of time.

(Refer Slide Time: 26:31)



Now the question is if we try to apply so our model is y it = beta 0 + delta 1 d 2t + beta 2 x it + v it = a i + u it. Now for OLS to be applicable what we need actually this covariance between x it and v i should actually be instead of v i let us say that this is simply V should be actually zero, covariance between x and the error term should be 0. And that is the standard assumption because if it is related to a i then that will lead to again endogenic problem.

When you explanatory variable is correlated with error term. So, if this condition has to be satisfied then what we actually need covariance between a i and x it should actually zero. So, that means the unobserved effect this is a i, the unobserved city-specific effect should be uncorrelated

with the x it. So, that assumption if it is fulfilled then only, we can apply OLS here but this assumption is very difficult is very difficult to be fulfilled.

Because in reality we actually want covariance between a i and x it to be related actually. So, unobserved city specific effect to be correlated with the unemployment rate so that is what we want and basically, that is one reason because of this correlation since we want this unobserved effect to be correlated with some of the explanatory variable, we actually apply panel data. So, that is what we need to understand.

This same example what we are discussing in our previous chapters when you are talking about or instrumental variable estimates and simultaneous equation model, we say that log of y it = beta 0 + beta 1 education it + some a i + u it and we say that in this case a i is basically ability. That is what we said that individual they have some kind of unobservability which also has some impact on their wage rate.

And we also what we wanted that education and ability they are actually correlated. So, we wanted covariance between education it and this a i to be not to have some kind of correlation. If that is what we want then we need to have this panel data instead of using a cross-sectional data. So, cross-sectional data cannot accommodate the fact that this type of fact that there are unobserved effect.

What is the problem? Because the moment we include this and a i is correlated with this, this model will be suffering from endogeneity bias we have already discussed. So, at the one hand we need to accommodate the fact that there is the presence of a i in our model and at the other hand we also want that the correlation between a i and this to be eliminated. So, that means we need to transform the model to eliminate a i from the model.

How to do that? So, how to eliminate how to remove a i? Because unless we remove a i from the model, will suffer from endogeneity we cannot apply OLS. So, what do we do? We write the equation for both the periods so for period 2 this would become y i 2 = beta 0 + delta 0 because d

2t will take the value 1 for the second period and I am writing for the equation for the second period plus beta 1 x i2 + a i + u i 2.

And for the second first period y i 1 = beta 0 plus this would become 0 because d 2t = 0 for the first period beta 1 x i1 + a i + u i1. So, let us say this is equation 1 this is equation 2. Now if I take 1 - 2 so that means if I take the first difference of these two then what will happen, we will get y i2 - y i1 = beta 0 will get cancelled delta 0 + beta 1 x i2 - x i1 this a i and a i will get cancelled will get u i2 - u i1.

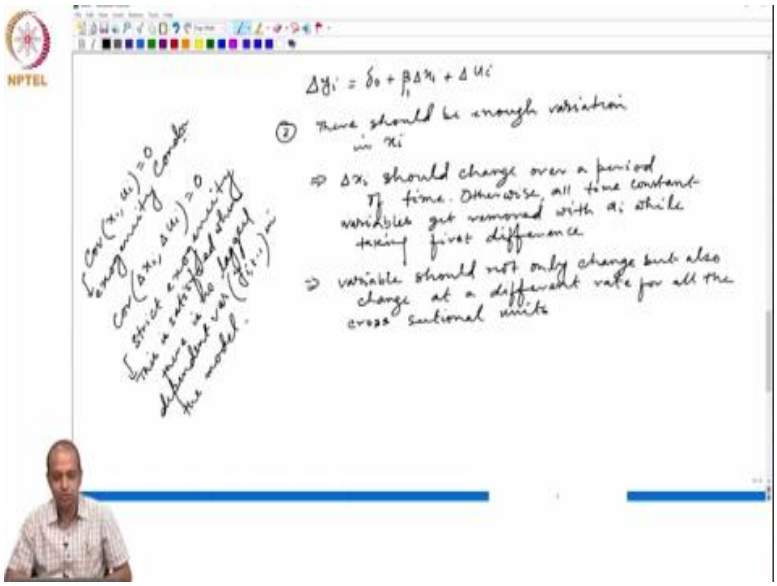
Now if we assume that this y 2 – y 1 is basically delta y i this x i2 - x i1 this becomes delta x i and this becomes delta u i then our equation will become delta y i = delta 0 + beta 1 delta x i + delta e y. This will be the equation. And if we want to apply OLS here then what we need to do we need to for OLS we need the covariance between delta x i and delta u i should be 0. And this condition is called strict endogeneity.

Why it is called strict endogeneity? Because that means we are ensuring not only x i and u i are uncorrelated but the difference of that also is uncorrelated delta x i and delta u i should be uncorrelated. Now when I say that covariance between x i delta x i and delta u i they are not correlated. So, that means this is ensured when x it does not have variable like y it - 1. Suppose one of our explanatory variable, is lag dependent variable y it - 1.

If that is the case, we cannot ensure that delta x i and delta u i is uncorrelated, x i is actually a representation of this is not x i a representation of let us assume that this is a set of expanded variables. In that set of explanatory variables if one of the explanatory variable is the lag dependent variable then we cannot ensure this because in that case what will happen this would become y it - 1 - y it - 2 and these would become u i1 – u it – u it - 1.

So, since there is a component called y it - 1 that would be correlated with u it - 1. So, we cannot ensure strict endogeneity strict this is called strict exogeneity this is strict exogeneity. So, in presence of lag dependent variable we cannot ensure that this delta x i is actually correlated with delta y i.

(Refer Slide Time: 38:53)



Now second thing when you are estimating this type of equation delta y i so our equation is delta y i = delta 0, what is our equation? Our equation is delta 0 + beta 1 delta x i + delta u i. Now second condition for OLS to be applicable or to require that there should be enough variation in delta x i. So, that means basically what it says that delta x i should change over a period of time. If it does not change then what will happen?

If it does not change if any of our explanatory variable is not changing over a period of time. For example, let us say that in our model we have some qualitative variable like gender. So, gender also does not change over a period of time. So, when we take first difference the variable which are not changing over a period of time that will also get eliminated along with a i. Otherwise, all time constant variables get removed with a i while taking first difference.

So, while taking first difference we must ensure that in our model there is no variable which is time constant. First difference model cannot accommodate the fact that sum of our variables are actually not changing over a period of time. Secondly even if the variables are changing, they should change at a different rate because if the variables should change variable should not only change but also change at a different rate.

If the variables are changing at the same rate, they will also get eliminated along with a i, they will also get eliminated because when I am taking the first difference between second period and first period they will also get eliminated. So, change should not be the same for all cross sectional units. Variables should not only change but also change at a different rate for all the cross sectional units.

So, we must ensure while applying first difference transformation of this type of model that all the variables are changing for all the cross-sectional units. If some of the variables are not changing for some of the cross sectional units while taking first difference those units would be dropped. So, number of observation will go down. For example, if it is unemployment rate and let us assume that for Chennai and Delhi the unemployment rate are not changing then Chennai and Delhi would be dropped while taking first difference.

Similarly, if the variables are changing at the same rate for all the cities that is also is undesirable in this first differencing of the panel data model. So, if these conditions are satisfied that means x i does not contain any variable which is lag dependent one like y it - 1 if that is the case, we cannot ensure that this is uncorrelated with the error term even after first difference. And that condition while the condition that means when we are saying that covariance between delta x i and delta u i they are not correlated that is called strict exogeneity.

So, that means we are not only ensuring exogeneity at level but also at their first difference when covariance between x i and u i is 0 that is called exogeneity condition. And when covariance between delta x i and delta y i is 0 that is called strict exogeneity. Why it is called strict? Because this is the stricter version of the exogeneity condition. We are not only ensuring exogeneity at level but also at their first difference.

And this is satisfied when there is no lagged dependent variable like y it - 1 in the model. So, this is the first condition and in the second condition we say that delta x i should change so all the variables should change over a period of time. So, if that is the case then only so there should be enough variation in x i sorry here in instead of delta x i this is actually x i because then only delta x i would become some non-zero value.

So, if these conditions are satisfied then only will say that OLS is applicable to estimate this type of model. Of course, there are several limitations of this first difference model that will discuss in a word next session, thank you.