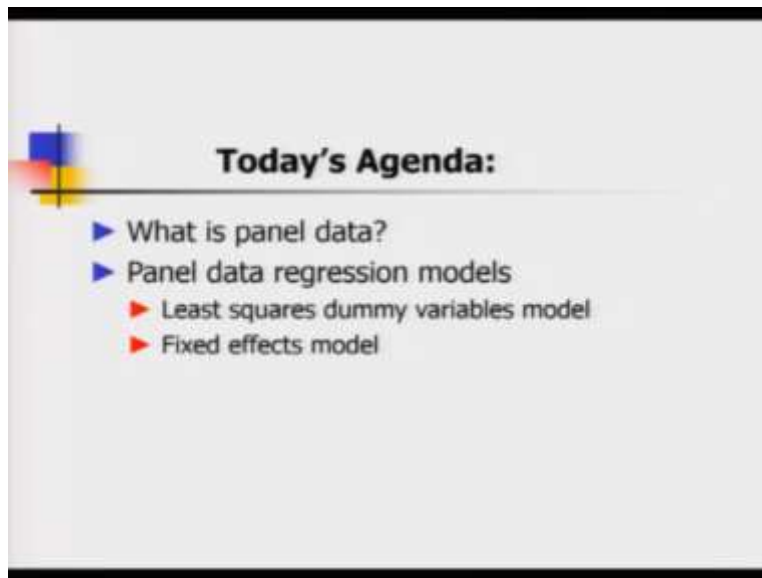


Applied Statistics and Econometrics
Professor. Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology, Kanpur
Lecture No. 38
Panel Data Regression

Hello friends. Welcome back to the lecture series on Applied Statistics and econometrics. So, today we are going to talk about another advanced topic in applied econometrics and econometric theory and that is called Panel data econometrics. So, before we go to formal models and all, let us have a look at today's agenda items.

(Refer Slide Time: 00:37)

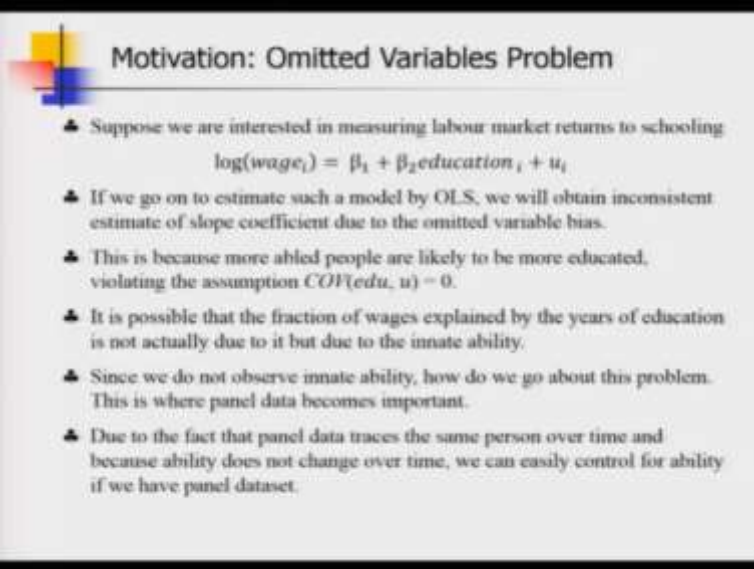


So, I will start today's lecture with introduction to panel data, this is a new kind of data that we have not dealt with so far in this course. So, it requires some bit of introduction. And in this lecture, we are also going to cover regression models, when you have panel data and we are going to cover two different techniques for it and they are called least squares dummy variables model and the fixed effects model.

There are many other types of models available for panel data regulation, but we are not going to cover in this lecture, because we are dedicating only one lecture in this course and I just want to cover the simple concepts in panel data econometrics. But before we go to introduction to panel data, let us start by motivating you, why we need to have a new kind of data which is called panel data or often in statistics language it is called longitudinal data?

So, I will take you back to one of the misspecification errors that we have studied in this course, and there you will see why availability of these kind of special data will help us. So, we are going to take you to the problem of omitted variable bias.

(Refer Slide Time: 02:03)



Motivation: Omitted Variables Problem

- Suppose we are interested in measuring labour market returns to schooling
$$\log(\text{wage}_i) = \beta_1 + \beta_2 \text{education}_i + u_i$$
- If we go on to estimate such a model by OLS, we will obtain inconsistent estimate of slope coefficient due to the omitted variable bias.
- This is because more abled people are likely to be more educated, violating the assumption $\text{Cov}(\text{edu}, u) = 0$.
- It is possible that the fraction of wages explained by the years of education is not actually due to it but due to the innate ability.
- Since we do not observe innate ability, how do we go about this problem. This is where panel data becomes important.
- Due to the fact that panel data traces the same person over time and because ability does not change over time, we can easily control for ability if we have panel dataset.

So, hopefully, you remember that return to schooling story through which I explained the omitted variable bias problem. So, let me remind you again, if you have forgotten. So, some researcher is willing to measure the labor market returns to years of education or years of schooling. So, we have a very simple linear regression model to start with log of wage or salary of i th individual, and we also have data on that particular person's number of years in schools and colleges.

So, basically, it is an education variable and of course, the stochastic term is there. So, now, this looks a bit simplified model, because we had earlier in the discussion on omitted variable bias and you probably remember that there are many other variables which affect a person's wage or salary and some of these variables are observed and some of these variables are unobserved. So, if some variable is unobserved you cannot collect data on that, and one of such variables is a person's inherent stability and motivation.

And we had the discussion in omitted variable bias problem, how to take care of this kind of problem, but note one interesting thing, these are unobserved variables, if they are not varying over time, then can we handle them much easily without looking at instrumental variables and

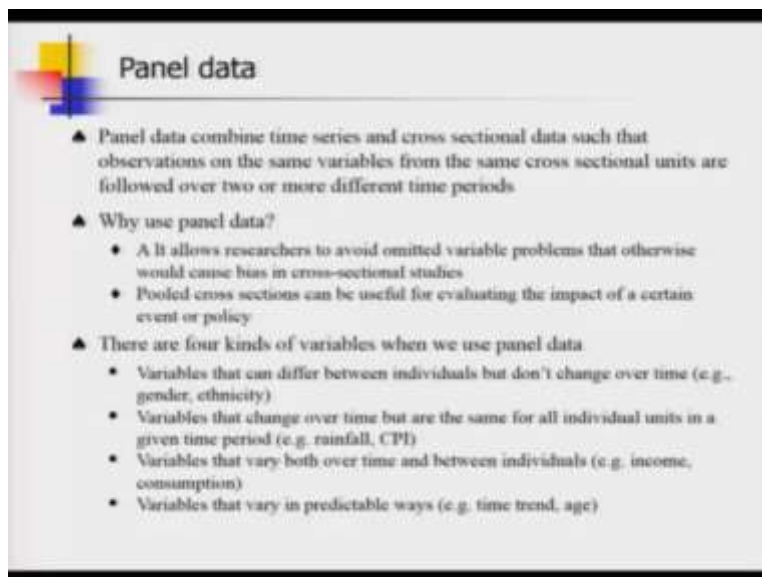
all? So, here a very basic assumption is that some of the unobserved or omitted variables are actually not varying with time. So, if we now collect data on the same individual units over a period of time, then these variables value is not going to change.

So, let me tell you about that crux of that omitted variable bias problem. So, if we estimate such a misspecified model by OLS, then we will obtain inconsistent estimate of the slope coefficient, and we are not going to get the true measure of return to schooling. And this is because, there is some omitted variable which is embedded in the error term u now and that causes a correlation to take place between these explanatory variable education, and random term u . So, the exogeneity breaks down.

So, now, we are saying that, if you have any data on the individuals who are in a period of time, and ability does not change over time, then you can take difference between two equations, for individuals over time and these inherent omitted variable which is unobserved but not varying over time will fall off.

So, suppose you have a data on individuals over a period of time. So, at least two time periods and now, we know if you have two different time points data, then you can always take difference and these are unobserved variable as it is not changing with respect to time it can actually fall-out from the equation and that will help you because you can avoid these omitted variable bias problems. So, this thing is going to be clear as we know, move on in this lecture.

(Refer Slide Time: 05:43)



Panel data

- ▲ Panel data combine time series and cross sectional data such that observations on the same variables from the same cross sectional units are followed over two or more different time periods
- ▲ Why use panel data?
 - It allows researchers to avoid omitted variable problems that otherwise would cause bias in cross-sectional studies
 - Pooled cross sections can be useful for evaluating the impact of a certain event or policy
- ▲ There are four kinds of variables when we use panel data
 - Variables that can differ between individuals but don't change over time (e.g., gender, ethnicity)
 - Variables that change over time but are the same for all individual units in a given time period (e.g. rainfall, CPI)
 - Variables that vary both over time and between individuals (e.g. income, consumption)
 - Variables that vary in predictable ways (e.g. time trend, age)

So, now, it is time that we formally introduce panel data. So, panel data combine time series and cross sectional data points such that observations on the same variables from the same cross sectional units are followed over 2 or more different time periods. So, you see, so far we have dealt with data sets with one dimension. So, when there is a population and you have drawn a sample from it to conduct regression analyses already know correlation analyses then basically that is a cross sectional data, because, time is fixed.

So, there is only one dimension of the data and that is basically the individual units which is varying, but when you have one particular variable measured for one particular unit or individual or entity like nations or regions, over a period of time, then you have time series data. So, there time is varying, but the observation is varying with respect to time only it is not varying from unit to unit, there is only one unit on which you have collected data for different time periods. Now, panel actually brings both parties together.

So, here you are observing one particular unit, it can be an individual, it can be a farm, it can be a country, it does not matter, you observe that particular unit for a period of time and you are observing the values of different variables on these particular unit over a period of time. So, of course, now we can ask this question to ourselves that is this the only reason that we want to get rid of omitted variable bias and that is why we are interested in panel data?

The answer is no, not really. Sometimes you just want to have panel data to increase the sample size. So, let me give you one example suppose, you are interested to estimate a production function for an agricultural crop for India or US and you have data only at the state level, I mean at the total production of that agricultural crop and some of the inputs like fertilizer land, etcetera. Suppose, you have data on only 4 states.

So, in the case of India we do not have even 40 such observations, right, if you have data from one particular time point. So, it is a good idea to collect the data on the states for as many as years, so that you can actually generate a large data set. So, you see that panel data actually helps you to increase the sample size that you are going to use for your statistical inference analysis. Now, there is another reason why panel data is becoming more and more important and that is basically the case of program evaluation.

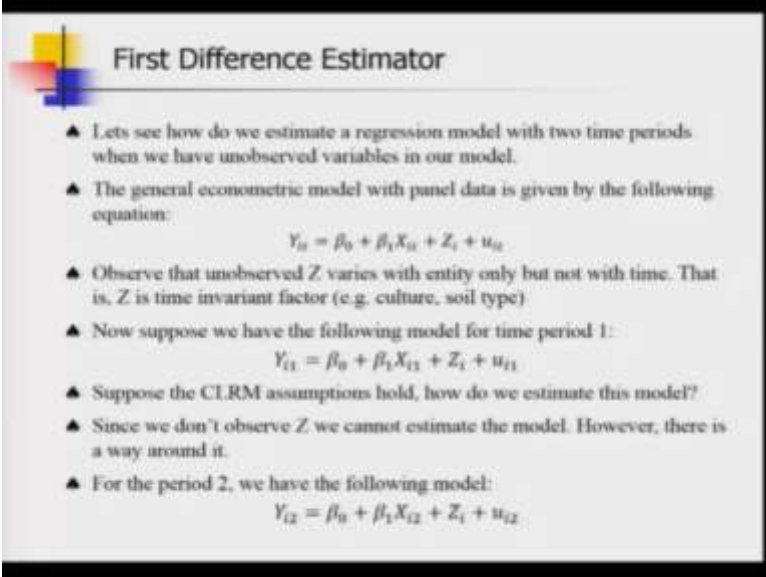
So, we are not going to cover these topics in today's lecture, but, next lecture I plan to talk about program evaluation at length, but in a nutshell, I can say that if you have pooled data on cross sectional units, by pooled data, I mean that you are observing one cross sectional unit for multiple time periods. So, if you have pooled cross sectional data, then that helps you to evaluate the impact of a particular policy or an event. And, I have dedicated the next lecture to this particular issue. There you will see how panel data is going to be very useful.

Now, when we encounter panel data, then actually, there could be four types of variables in there. And here I am going to talk about briefly, about these four types of variables. And first one is, of course, the variables that can differ between individuals but do not change over time. And the examples could be gender and ethnicity. And the second type of variables could be those which change over time, but are the same for all individual units in a given time period and examples could be rainfall or consumer price index number.

And the third type of variables are those which vary both over time and individuals. So, here the examples could be income and consumption and finally, we have a fourth type of variable and that varies in very predictable manner. And here the example could be time trained and age. Now, we are going to talk about a very special form of panel data where we have a set of cross sectional units and we have observed the same cross sectional units for only two time periods.

So, if that is the case, then how we are going to handle these kinds of data and conduct some regression analysis, that is what we are going to be studying in the next slide.

(Refer Slide Time: 11:15)



First Difference Estimator

- ▲ Lets see how do we estimate a regression model with two time periods when we have unobserved variables in our model.
- ▲ The general econometric model with panel data is given by the following equation:
$$Y_{it} = \beta_0 + \beta_1 X_{it} + Z_i + u_{it}$$
- ▲ Observe that unobserved Z varies with entity only but not with time. That is, Z is time invariant factor (e.g. culture, soil type)
- ▲ Now suppose we have the following model for time period 1:
$$Y_{i1} = \beta_0 + \beta_1 X_{i1} + Z_i + u_{i1}$$
- ▲ Suppose the CLRM assumptions hold, how do we estimate this model?
- ▲ Since we don't observe Z we cannot estimate the model. However, there is a way around it.
- ▲ For the period 2, we have the following model:
$$Y_{i2} = \beta_0 + \beta_1 X_{i2} + Z_i + u_{i2}$$

So, the problem is that we want to estimate a regression model with two time periods data on a set of individuals and suppose that there are some unobserved variables in our model. So, let us propose a general model with panel data and this is very important, concentrate on this equation that I am discussing now because this equation is somewhat different from the regression equations that I have shown you previously in this course.

So, here you see that all the variables have two subscripts i and t and what do they mean? So, here the i subscript refers to the cross sectional units and these cross sectional units are taking values from 1 to say a capital N . So, there are no capital N number of cross sectional units at one time point and then this small t subscript actually talks about the time point. So, basically if there are capital T time periods of data available on these cross sectional units, then the t value small t value can take values 1 to capital T .

So, now, let us revisit these equations, so, Y_{it} basically says that this is the value of the dependent variable for i th individual in time period t and similarly X_{it} means that this is the value of the explanatory variable for i th cross sectional unit for T th time period and note that there is this variable Z_i so, which is not varying over time, but it varies over individual. So, that is why it has the subscript i .

Now, this Z_i variable could be an unobserved variable on which you don't have the data. So, suppose let us go back to that old example of return to schooling and this Z could be these

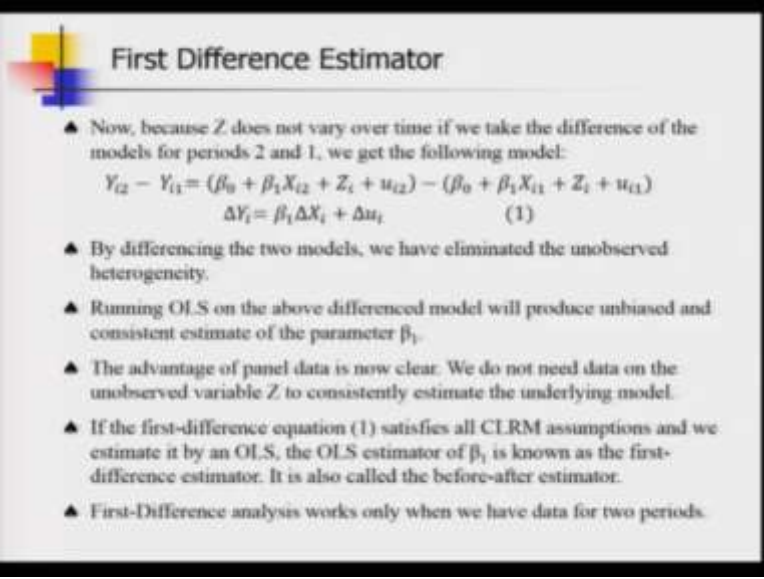
inherent motivation factor or the ability of a person for which you have collected data on salary and your number of years spent in school. So, let us see how we can actually now take care of these unobserved variable Z_i . But note that U_{it} the stochastic random term now, we know it actually varies over both the cross sectional units and the time units. So, that is why we are writing here it as U_{it} .

So, what could be no other examples of these time invariant factors? We have discussed about the inherent ability or motivation of a person in the labor economics context, but from agricultural context you can say that it could be the nature of the soil, because nature of the soil does not change drastically in short run, but when you are estimating agricultural production function or you are trying to explain yield of a particular crop then soil type actually plays a role and this is generally unobserved because the econometricians many times do not find data on soil type.

So, now, let us focus on that regression equation for time period 1. So, you see, the first expression is now replaced by putting a value for t there and that is why you see all these symbols like Y_{i1} and X_{i1} . And now suppose the classical linear regression assumptions are holding, so, how do you estimate this model? So, here is a problem, although that this is the true model but you do not have any data on Z . So, Z is basically an unobserved variable, it is just sitting there in the regression equation.

So, then what to do? So, you write down the same regression equation for time period 2, and so, the first equation that I showed you the mother regression equation, the panel regression equation, now, we there you see I put the value of t equals to 2 and I get a new equation. So, that equation is at the bottom of the slide.

(Refer Slide Time: 15:52)



First Difference Estimator

- ▲ Now, because Z does not vary over time if we take the difference of the models for periods 2 and 1, we get the following model:
$$Y_{i2} - Y_{i1} = (\beta_0 + \beta_1 X_{i2} + Z_i + u_{i2}) - (\beta_0 + \beta_1 X_{i1} + Z_i + u_{i1})$$
$$\Delta Y_i = \beta_1 \Delta X_i + \Delta u_i \quad (1)$$
- ▲ By differencing the two models, we have eliminated the unobserved heterogeneity.
- ▲ Running OLS on the above differenced model will produce unbiased and consistent estimate of the parameter β_1 .
- ▲ The advantage of panel data is now clear. We do not need data on the unobserved variable Z to consistently estimate the underlying model.
- ▲ If the first-difference equation (1) satisfies all CLRM assumptions and we estimate it by an OLS, the OLS estimator of β_1 is known as the first-difference estimator. It is also called the before-after estimator.
- ▲ First-Difference analysis works only when we have data for two periods.

So, now, we know what to do, because Z does not vary over time. If we take the difference of these two regression equations for periods 1 and 2, then actually we get a new model. So, here let us take the difference between Y_{i2} and Y_{i1} . So, here I am showing you the calculations and you see I get a revised equation in forms of the delta operator.

So, if you remember our time series discussion, I hope that you remember that delta operator. delta operator actually talks about time difference. So, here you look at the equation 1 which is expressed in the time difference manner. So, basically, delta y_i means that for i th individual, what is the difference between the value of the dependent variable in time period 2 and time period 1. Similar explanation can be provided for delta X_i as well.

Note that error is also going to change and that is why we are denoting it by delta U_i , but very interestingly you will see that that Z_i variable which was unobserved and not varying with time, that has dropped out from this revised or redefined equation, because when you take the time difference as these particular variables Z is not reading with respect to time, it will fall off. So, this is a very good news, because, we have just eliminated the unobserved heterogeneity that was present in your regression model.

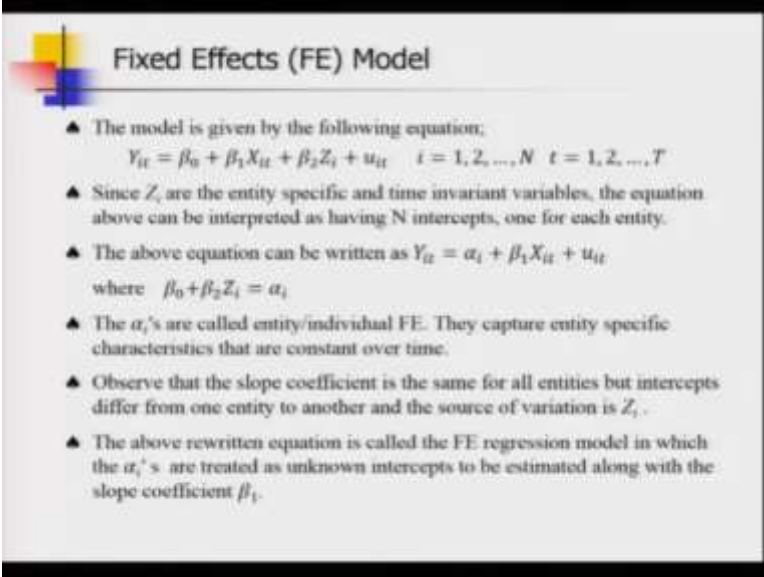
Now, you can run OLS on this equation 1 and this equation one note can be called a difference model or first difference model and if you run OLS on this particular equation, then that will give you the unbiased and consistent estimate of the parameter beta 1. So, this is the way you can

actually handle the problem of omitted variable say a person's inherent ability or motivation, when you were trying to get a measure for return for schooling on the salary or wage whatever in the labor market.

So, if the first difference equation number 1 satisfies all the classical linear regression model assumptions, and if we actually apply OLS, then this OLS estimator of beta 1 is known as the first difference estimator. So, there is a new jargon that I am introducing here and this is also sometimes called before-after estimator. But note that there is a problem, this particular estimator will only work when you have data for 2 periods only. If you have data for 3 periods or more number of time periods, then you cannot apply this particular estimator that I just have shown you, although it has some merit.

So, we have to now learn some more sophisticated regression models and techniques so that we can handle the situation where we are lucky to have multiple time points observation for the same set of individuals.

(Refer Slide Time: 19:17)



Fixed Effects (FE) Model

- ▲ The model is given by the following equation,
$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it} \quad i = 1, 2, \dots, N \quad t = 1, 2, \dots, T$$
- ▲ Since Z_i are the entity specific and time invariant variables, the equation above can be interpreted as having N intercepts, one for each entity.
- ▲ The above equation can be written as $Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it}$
where $\beta_0 + \beta_2 Z_i = \alpha_i$
- ▲ The α_i 's are called entity/individual FE. They capture entity specific characteristics that are constant over time.
- ▲ Observe that the slope coefficient is the same for all entities but intercepts differ from one entity to another and the source of variation is Z_i .
- ▲ The above rewritten equation is called the FE regression model in which the α_i 's are treated as unknown intercepts to be estimated along with the slope coefficient β_1 .

So, in this slide, I am going to talk about the basic fixed effects model as proposed by statisticians and economic pressures. So, this model is given by the following equation and this is not new to you but the only difference here you see that I have changed the values for t. So, here the T can take any values from 1, 2 to capital T. capital T is definitely greater than 2.

So, now as my Z_i these unobserved individual heterogeneity are the entity specific or individual specific variables, but they are time invariant variables, then the above equation can be seen, as one regression equation having a intercept terms, one for each entity. So, if I buy these arguments, then the above regression equation can be written as Y_{it} equals to α_i plus $\beta_1 X_{it}$ plus u_{it} where α_i is defined as $\beta_0 + \beta_2 Z_i$, the intercept term in that original regression equation plus β_2 times Z_i .

So, these α_i are called the entity or individual specific fixed effects and they capture the entity specific idiosyncratic characteristics that are constant over time. Now, note that the slope coefficient is the same for all the individuals, but intercepts are now differing from one individual to the other and that is basically the source of variation in these variables Z_i .

So, the regression equation that I have shown here at the top of the slide if you now insert these α_i variable there in place of $\beta_0 + \beta_2 Z_i$, then that regression equation is called the fixed effects regression model and there these alphas are treated as unknown intercepts to be estimated along with the slope coefficient β_1 .

Now, note that here in this regression equation, I am keeping only one explanatory variable. So, does it mean that our fixed effects regression model cannot have more than one continuous explanatory variable or some explanatory variable for which you have data? No, do not have that impression. Just for simplicity's sake, I am keeping one explanatory variable here.

(Refer Slide Time: 22:02)

Fixed Effects Model: LSDV Estimator

- ▲ In order to estimate FE model we need to quantify the entity specific intercepts, α_i 's.
- ▲ One way to do so is to introduce N dummies, one for each entity in the model such as the follows:

$$Y_{it} = \gamma_1 D1_i + \gamma_2 D2_i + \dots + \gamma_N DN_i + \beta_1 X_{it} + u_{it}$$

Where $D1_i = \begin{cases} 1 & \text{if } i = 1 \\ 0 & \text{if otherwise} \end{cases}$, $D2_i = \begin{cases} 1 & \text{if } i = 2 \\ 0 & \text{if otherwise} \end{cases}$, $DN_i = \begin{cases} 1 & \text{if } i = N \\ 0 & \text{if otherwise} \end{cases}$
- ▲ Observe that a dummy is included for each entity, because we do not have an intercept included. If we include intercept then introduce $N-1$ dummies only. So, alternatively you can estimate the following model:

$$Y_{it} = \beta_0 + \gamma_2 D2_i + \dots + \gamma_N DN_i + \beta_1 X_{it} + u_{it}$$

where $\beta_0 = \alpha_1$ and $\alpha_i = \beta_0 + \gamma_i$, $i \geq 2$
- ▲ OLS estimation of the fixed effects model with entity dummies is called the *least squares dummy variables regression (LSDV)*.

So, now, we are going to talk about the estimation of fixed effects model. So, here at this moment, before we start talking about fixed effects model estimation, I should tell you that there are 3 different types of fixed effects modeling estimation procedures and one we have already seen that is basically the first approach, now, we are going to talk about the second output, which is known as the least squares dummy variable.

So, we have to estimate the fixed effects model, but there are these parameters which are coming from the individual specific intercept terms. So, how do we model? So, we have to quantify these individual specific idiosyncratic factors, which are shown as α_i , these are the terms which are going to change my grand intercept or overall intercept of the model. So, basically the way to deal with this problem is to make use of the dummy variable technique.

So, if you have n number of cross sectional units for which you have data, for our t time periods, then you can throw capital N number of dummies for, one for each of the cross sectional units, but note that do not fall in the trap called dummy variable trap, I have spoken about it couple of lectures back. So, you must remember it, so, to avoid the dummy variable trap, you can actually now exclude the overall or grand intercept term from the regression equation, and just keep N number of dummies in regression.

So, that is what we are doing here. So, here you see, I am showing you this complicated linear regression model Y_{it} equals to γ_1 times D_{1i} plus γ_2 times D_{2i} dot γ_1 times D_{Ni} , so, here D_1, D_2, D_n are basically the individual specific dummy variables. So, D_1 denotes the dummy variable indicating the individual 1 and then this takes value 1 for Y equal to 1 otherwise it takes value 0. So, other dummy variables will have similar interpretations. So, this is basically our model.

Now, note that a dummy is included for each entity because we do not have an intercept included. So, I have already spoken about this issue, but what if you decide that no-no, I want to keep my grand intercept or the overall intercept for the model. Then to avoid the dummy variable trap you have to introduce capital $N-1$ number of dummy variables. So, basically you can say that the individual 1 is my base and keeping him or her or that particular unit as base you define other $N-1$ dummy variables for the rest $N-1$ observational units.

(Refer Slide Time: 25:26)

Hypothesis testing for unobserved FE

- ▲ Suppose we are interested in testing the equality of FE in the model:
$$Y_{it} = \gamma_1 D1_i + \gamma_2 D2_i + \gamma_3 D3_i + \beta_1 X_{it} + u_{it}$$
- ▲ Set $H_0 : \gamma_1 = \gamma_2 = \gamma_3$ and $H_1 : H_0$ is not true
- ▲ Test statistic F^{obs} is
$$\frac{(SSE_R - SSE_U) / J}{SSE_U / (NT - K)} \sim F_{(NT-K)}^J$$
 - SSE_R is the restricted error sum of squares (1 intercept)
 - SSE_U is the unrestricted error sum of squares (3 intercepts)
 - N is the number of cross-sectional units ($N = 3$)
 - K is the number of parameters in the model ($K = 4$)
 - J is the number of restrictions being tested ($J = 3$)
 - T is the number of time periods
- ▲ Decision:
 - Reject H_0 if $F^{obs} > \text{Critical value}$
 - Reject H_0 if $\beta\text{-value (Area in the } F\text{-distribution to the right of } F^{obs}) < \alpha$.

So, we can estimate our least squares dummy variables regression easily and we can get r square and all the regression summary statistics, but, you can still be a bit worried by this question that, what if there is no difference in these individual dummy variables that I have thrown in the regression equation? So, by that, I mean to say that you could be a bit suspicious and you can actually suspect whether there is such statistical difference between these individual specific dummy variables or individual specific intercept terms.

So, in a nutshell, you may want to test whether jointly these terms, these individual intercept terms are different from each other or not. So, in other words, you can say that I am going to test for equality of these fixed effects. So, if I want to test whether these individual specific fixed effects are same or not, then basically the way to do it by setting a null hypothesis, we says that gamma 1 equals to gamma 2 equals to gamma 3 and alternative should say that H not is not true, okay.

Now, note that this particular test is very similar to one test that we have studied couple of lectures before and that was the case of our model selection and selection has to be made between a restricted model and an unrestricted model. So, here you can say that the regression model that you have used to estimate our least squares dummy variable regression equation, that is basically your full or unrestricted model and when you are imposing these restriction that my

individual specific dummies or the intercepts are actually equal, then you are putting a restriction and then that becomes a restricted model.

So, we are going to make use of the F test. So, as I said, we are going to use F test, as this is a kind of restricted versus unrestricted model problem. So, here we define our test statistic as F observed and that is defined as typically how we define these you know F statistics, we have to take the difference between the sum of squares error from the restricted model and the unrestricted model.

So, sum of squares of error from the restricted model is denoted by SSE_r and sum of squared errors from the unrestricted model is denoted by SSE_u and then that needs to be divided by the degrees of freedom and that is basically the number of restrictions that you are imposing on your model and this entire ratio has to be now divided by another ratio, and that is basically the sum of squared error from the unrestricted model divided by its degrees of freedom and that degrees of freedom is equal to Nt minus k .

So, this F observe test statistic will now follow an F distribution and what would be the degrees of freedom for these F distribution. So, these F statistic will follow F distribution with 2 degrees of freedom, which is the number of restrictions that is say J and the other one is NT minus K . So, here in this particular regression equation, which is very specific as illustration or as our case here you see that the capital N takes value 3 why?

Because, that is why we are throwing three individual dummy variables and then, we have K takes value 4 here, because, we are going to estimate four parameters in the unrestricted model and then T is the number of time periods, now, T can take any value here we are not specifying a particular value of t and we all know by now probably you remember how to conduct a hypothesis testing, so, I am not no repeating those steps again, but still for your here, I am putting the decision rule at the bottom of this slide.

Now, note that although the least squares dummy variables technique is pretty useful in panel data regression, when you have more than two time periods of data, but it has big limitation and this limitation actually in terms of degrees of freedom. So, if you have data set where you have a large number of cross sectional unit say 100 plus and you have data on only three time periods then actually you have a data set of 300 observations.

But, suppose you have a 5 explanatory variables that you are interested in but now if you are going to follow these squares dummy variables regression technique, then you have to throw 100 dummies in the regression equation, so, it will eat up a lot of degrees of freedom and that is not good, because, if you lose out on degrees of freedom for hypothesis testing purpose, that is very bad news. So, we are still not done, we have to find out a more sophisticated estimation technique, when we have panel data for more than three time periods.

So, basically, now we come to the last technique that we are going to visit and that is probably the most popular panel data fixed effects estimator and that is known as within estimator. So, for that actually we have to first understand what is within transformation. So, the next slide is going to talk about that first and then we are going to propose within estimator.

(Refer Time Slide: 31:50)

Fixed Effects Model: Within Estimator

- ▲ The other way to estimate the FE model is a two step procedure:
 1. **Within transformation:** Demean Y_{it} and X_{it}
 2. **Within estimation:** Estimate the demeaned model by OLS
- ▲ We have the fixed effects model:

$$Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it} \quad (A)$$

$$\bar{Y}_i = \alpha_i + \beta_1 \bar{X}_i + \bar{u}_i \quad (B)$$

where $\bar{Y} = \frac{1}{T} \sum_{t=1}^T Y_{it}$, $\bar{X} = \frac{1}{T} \sum_{t=1}^T X_{it}$, $\bar{u}_i = \frac{1}{T} \sum_{t=1}^T u_{it}$
- ▲ Subtracting (B) from (A) we have the following entity-demeaned model:

$$(Y_{it} - \bar{Y}_i) = \beta_1 (X_{it} - \bar{X}_i) + (u_{it} - \bar{u}_i)$$
- ▲ Let $(Y_{it} - \bar{Y}_i) = \bar{Y}_{it}$, $(X_{it} - \bar{X}_i) = \bar{X}_{it}$ and $(u_{it} - \bar{u}_i) = \bar{u}_{it}$, all these are the entity demeaned variables.
- ▲ Then the FE model to be estimated becomes $\bar{Y}_{it} = \beta_1 \bar{X}_{it} + \bar{u}_{it}$

So, let us first explain, what do we mean by within transformation. So, by within transformation, we mean that we need to demean the dependent variable y and explanatory variable x and how this is done? That I am going to show you here in the second bullet point. So, start with the fixed effects model and that is given by this equation A. So, now what you do, you take the mean of this equation, on both sides, so that is why I am taking the mean and I am placing these bars on the top of these variables Y, X and U.

Now, note that alpha is not changing, because I am taking the mean over time periods and actually alpha is not changing with respect to time that is why there is no mean because it is

just a constant over time. So, here as I am taking the average over different time periods, you see the formula for \bar{Y} , \bar{x} and \bar{U} , they are shown here at the bottom of equation B and this is quite self-explanatory, I believe.

Now, you subtract the equation B from equation A and you have this demeaned model. So, in the demeaned model you see what we are doing basically, we are basically subtracting that individual specific mean from the individual specific observations. So, for the dependent variable y that is given as Y_{it} minus \bar{Y}_i . Now, let us introduce some new notations for these demeaned variables. So, the new variables are defined by placing a tilde sign on the top of the variable.

So, now, the demeaned variables are \tilde{Y}_{it} , \tilde{X}_{it} and \tilde{U}_{it} . Note that by virtue of these are demeaning exercise or within transformation, we have a got rid off this nuisance variable which is basically this unobserved individual specific heterogeneity. So, that fixed effect is gone. So, now, you can actually apply your good old friend OLS on these redefined or demeaned or within transformed regression equation.

Now, if you apply OLS technique to these fixed effects model, which is after within transformation, then you have to OLS on these within transform variables \tilde{Y}_{it} and \tilde{X}_{it} and as theoretically it can be shown that if you are dealing with this within transformed regression equation, the classical linear regression models assumptions they are holding and the coefficient estimate that you derive from this exercise is going to be unbiased and consistent. So, this technique is called within estimation technique.

So, we stop here. Next lecture, I am going to come back with discussion on program evaluation which is going to be an application of these panel data models that we have discussed here. So, join me for that. See you then. Thank you.