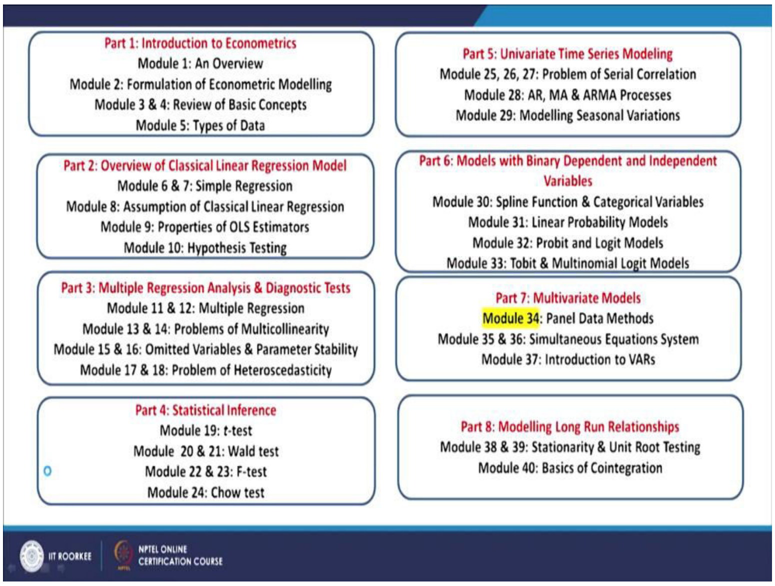
Econometric Modelling
Professor. Sujata Kar
Department of Management Studies
Indian Institute of Technology, Roorkee
Lecture No. 34
Panel Data Methods

Hello, this is Module 34 of the course on Econometric Modeling. This is basically Part 7, where we will focuss on multivariate models.

(Refer Slide Time: 0:38)



Part 1: Introduction to Econometrics
Module 1: An Overview
Module 2: Formulation of Econometric Modelling
Module 3 & 4: Review of Basic Concepts
Module 5: Types of Data

Part 2: Overview of Classical Linear Regression Model
Module 6 & 7: Simple Regression
Module 8: Assumption of Classical Linear Regression
Module 9: Properties of OLS Estimators
Module 10: Hypothesis Testing

Part 3: Multiple Regression Analysis & Diagnostic Tests
Module 11 & 12: Multiple Regression
Module 13 & 14: Problems of Multicollinearity
Module 15 & 16: Omitted Variables & Parameter Stability
Module 17 & 18: Problem of Heteroscedasticity

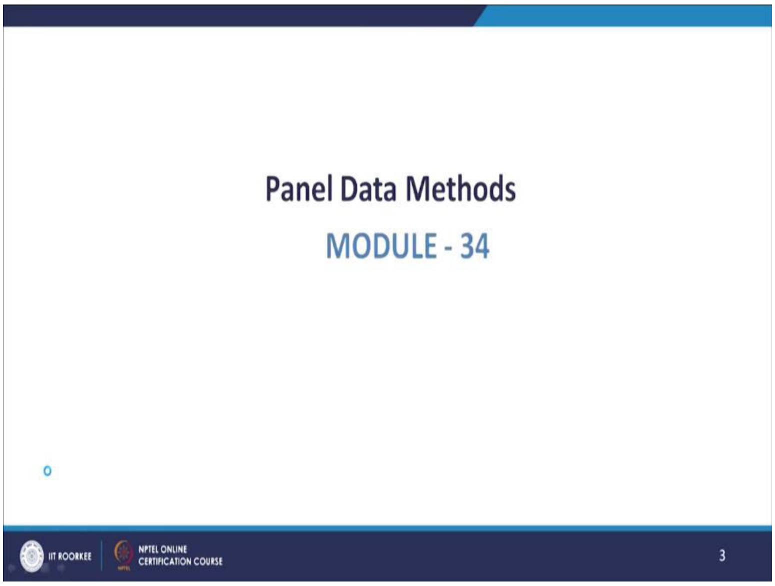
Part 4: Statistical Inference
Module 19: *t*-test
Module 20 & 21: Wald test
Module 22 & 23: F-test
Module 24: Chow test

Part 5: Univariate Time Series Modeling
Module 25, 26, 27: Problem of Serial Correlation
Module 28: AR, MA & ARMA Processes
Module 29: Modelling Seasonal Variations

Part 6: Models with Binary Dependent and Independent Variables
Module 30: Spline Function & Categorical Variables
Module 31: Linear Probability Models
Module 32: Probit and Logit Models
Module 33: Tobit & Multinomial Logit Models

Part 7: Multivariate Models
Module 34: Panel Data Methods
Module 35 & 36: Simultaneous Equations System
Module 37: Introduction to VARs

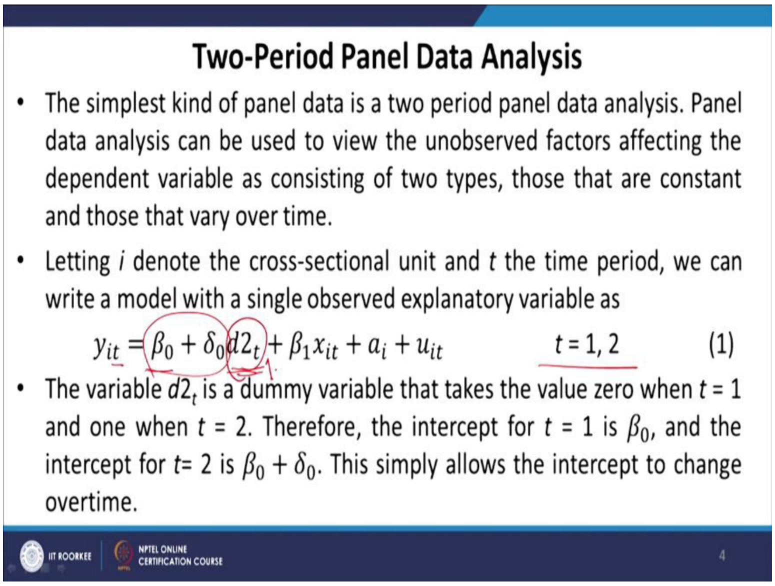
Part 8: Modelling Long Run Relationships
Module 38 & 39: Stationarity & Unit Root Testing
Module 40: Basics of Cointegration



Panel Data Methods
MODULE - 34

We would be primarily discussing three types of modeling techniques. The first one is the panel data methods. Panel data method itself is basically a huge area, and this is very difficult to contain the methods in a single module of only for 30 minutes or so. So, what I have done, is that try to give you a basic exposure to the panel data methods, it's understanding its very basic techniques, and so on.

(Refer Slide Time: 1:08)



Two-Period Panel Data Analysis

- The simplest kind of panel data is a two period panel data analysis. Panel data analysis can be used to view the unobserved factors affecting the dependent variable as consisting of two types, those that are constant and those that vary over time.
- Letting i denote the cross-sectional unit and t the time period, we can write a model with a single observed explanatory variable as

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + a_i + u_{it} \qquad t = 1, 2 \qquad (1)$$

- The variable $d2_t$ is a dummy variable that takes the value zero when $t = 1$ and one when $t = 2$. Therefore, the intercept for $t = 1$ is β_0, and the intercept for $t = 2$ is $\beta_0 + \delta_0$. This simply allows the intercept to change overtime.

So, we begin the discussion on panel data methods. So, the simplest kind of panel data is a Two-Period Panel Data Analysis. Panel data analysis can be used to view the unobserved factors affecting the dependent variable, as consisting of two types, those that are constant and those that vary over time. So, before I discussed this further, I just recapitulate that panel data is a kind of data where the cross-section and the time-series observations are basically brought together.

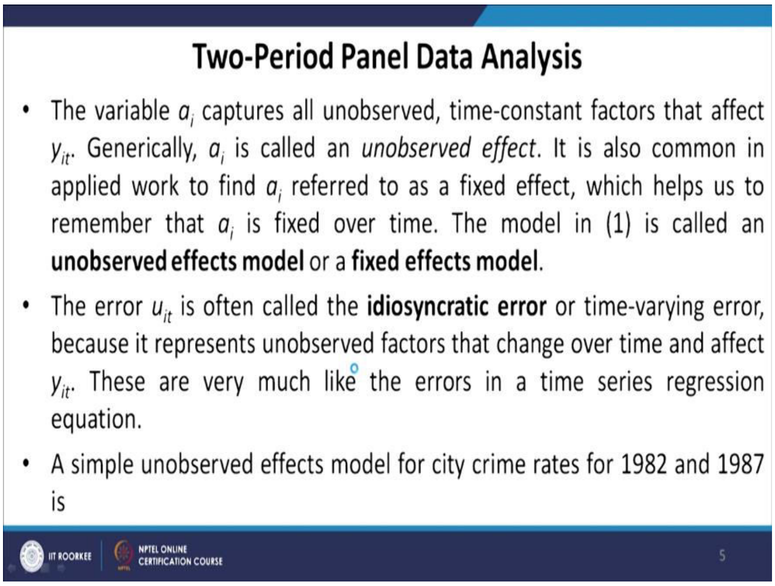
But this is different from a pooled cross-section from the fact that in parallel studies, the same units of the cross-section are observed or surveyed for several periods. So, at least for two periods, it can be several periods. So, we have the same cross-section entities for each and every period, which is not the case with pooled data wherein different time periods, the entities or the cross-section units can very well be different.

So, that is why panel data is characterized by two components. One is the cross-sectional variations and the other is the time series variations. So, what I first discuss, is that, we probably would try to take care of that time variations, how it is done *(refer to slide time 01:08).* So, letting i denote the cross-sectional unit and t the time period, we can write a model with a single observed explanatory variable, as Yit equals beta naught plus delta naught d2t plus beta 1 Xit plus ai plus Uit for t equals 1 to 2.

So, this shows that for each and every individual there will be two equations, one pertaining to the first period, and the second pertaining to the second period. So, t here takes only two values, 1 and 2. So, for an individual i one would be yi1 and so, on the other one will be an equation for Yi2. The variable d2t is a dummy variable that takes the value 0 when t equals 1 and when t equals 2. So, this is basically a time dummy specifically for the second period.

Therefore, the intercept for t equals 1 is beta naught because then at that point of time d2t takes a value 0 and the intercept for t equals 2 is beta naught plus delta naught because, when it is time period 2, then this takes a value equals to 1. This simply allows the intercept to change over time.

(Refer Slide Time: 3:58)



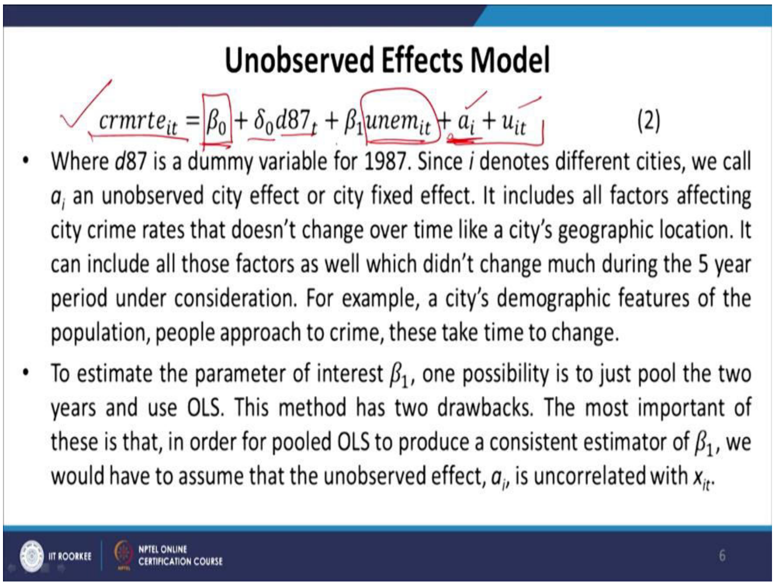
Two-Period Panel Data Analysis

- The variable a_i captures all unobserved, time-constant factors that affect y_{it}. Generically, a_i is called an *unobserved effect*. It is also common in applied work to find a_i referred to as a fixed effect, which helps us to remember that a_i is fixed over time. The model in (1) is called an **unobserved effects model** or a **fixed effects model**.
- The error u_{it} is often called the **idiosyncratic error** or time-varying error, because it represents unobserved factors that change over time and affect y_{it}. These are very much like the errors in a time series regression equation.
- A simple unobserved effects model for city crime rates for 1982 and 1987 is

The variable ai captures all unobserved time-constant factors that affect Yit. Now generally ai is called an unobserved effect, the thing that we are not able to observe. It is also common in applied work to find ai referred to as a fixed effect, which helps us to remember that ai is basically fixed over time. You can see that ai does not have any time subscript. So, this is basically up an entity or a component, which is free of time, which remains constant over a period of time over a long period of time or at least over the period of analysis.

The model in 1 is called an unobserved effects model or fixed-effects model. The error Uit is often called the idiosyncratic error or time-varying error because it represents an unobserved factor that changes over time and effect, Yit. So, ideally here, we have actually two components that are unobserved, one component varies with time the other component does not vary with time. So, these are very much like the errors in a time series regression equation the Uits.

(Refer Slide Time: 5:27)



Unobserved Effects Model

$$crmrte_{it} = \boxed{\beta_0} + \delta_0 d87_t + \beta_1 \boxed{unem_{it}} + a_i + u_{it} \qquad (2)$$

- Where $d87$ is a dummy variable for 1987. Since i denotes different cities, we call a_i an unobserved city effect or city fixed effect. It includes all factors affecting city crime rates that doesn't change over time like a city's geographic location. It can include all those factors as well which didn't change much during the 5 year period under consideration. For example, a city's demographic features of the population, people approach to crime, these take time to change.
- To estimate the parameter of interest β_1, one possibility is to just pool the two years and use OLS. This method has two drawbacks. The most important of these is that, in order for pooled OLS to produce a consistent estimator of β_1, we would have to assume that the unobserved effect, a_i, is uncorrelated with x_{it}.

IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 6

A simple unobserved effects model for city crime rates for 1982 and 1987 is or the crime rate refers to basically for i reference to here a particular city and t refer to either of the time period, either 1982 or 1987 *(refer to slide time 05:27).* Then we have beta naught the intercept. This intercept actually is the parameter of the reference period, that is, the first period.

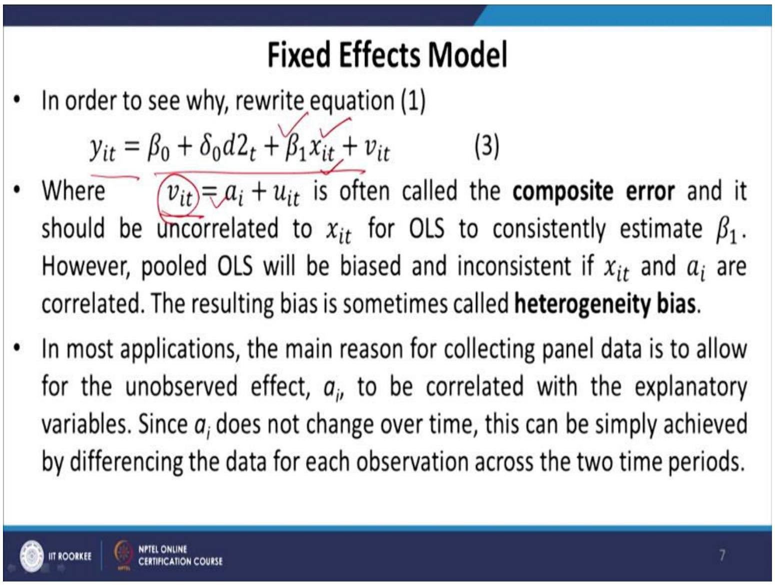
Then delta naught multiplied by the dummy for the second period d87 and we have plus beta 1 unemployment *it* plus the two unobserved factors, one is time-independent and other is linked to time. So, here we are trying to find out whether unemployment impacts the crime rates in a particular city or not, and that is the data collected over a span of two periods 1982 and 1987. So, where d87 is a dummy variable for 1987 since i denote different cities we call ai and unobserved city effect or city fixed effect.

It includes all factors affecting city crime rates that do not change over time, like a city's geographic location, it can include all those factors, as well, which did not change much during the 5 year period under consideration. So, for example, the city's demographic features of the population people's approach to crime, these take time to change and accordingly all these can be considered fixed over a period of time.

To estimate the parameter of interest beta 1, one possibility is to just pool the two yours and use OLS, we call it pooled OLS. This method has two drawbacks. The most important of this, is that, in order to pooled OLS to produce a consistent estimator of beta 1, we would have to assume that the unobserved effect ai is uncorrelated with Xit So, Xit is here unemployment and this is the unobserved effect.

So, this actually tries to imply that the factors, which are remaining constant over a period of time should be uncorrelated with the independent variables included in order to apply a pooled OLS or you have to apply OLS on this data which consists of all the observations for 2 years. So, this is what is pooled OLS, and we need this assumption. But this actually need not be the case always. In order to see why let us rewrite the equation once.

(Refer Slide Time: 8:16)



Fixed Effects Model

- In order to see why, rewrite equation (1)

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + v_{it} \qquad (3)$$

- Where $v_{it} = a_i + u_{it}$ is often called the **composite error** and it should be uncorrelated to x_{it} for OLS to consistently estimate β_1. However, pooled OLS will be biased and inconsistent if x_{it} and a_i are correlated. The resulting bias is sometimes called **heterogeneity bias**.
- In most applications, the main reason for collecting panel data is to allow for the unobserved effect, a_i, to be correlated with the explanatory variables. Since a_i does not change over time, this can be simply achieved by differencing the data for each observation across the two time periods.

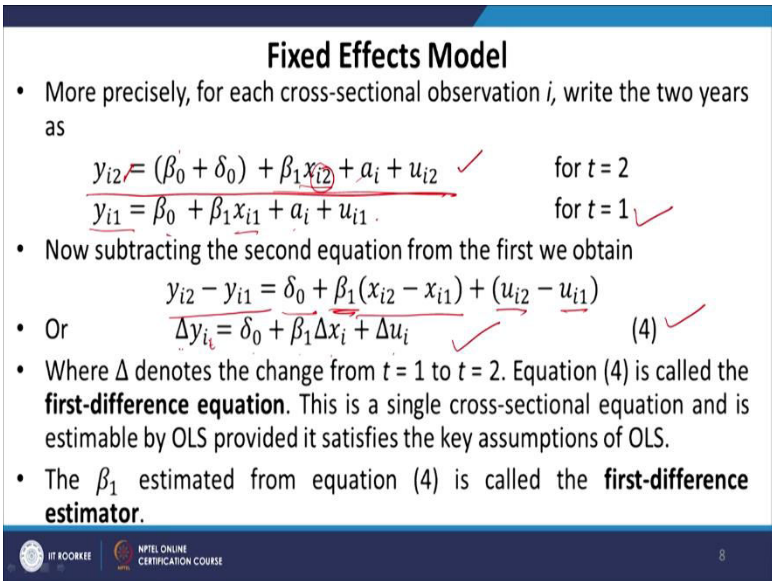
IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 7

This was my initial equation 1 *(refer to slide time 08:16),* we are rewriting it as to why Yit equals beta naught plus delta naught d2t plus beta 1 Xit plus Vit. So, what we are simply doing is that clubbing the two unobserved factors that are named as Vit and this is equal to ai plus Uit. And it is often called the composite error, and it should be uncorrelated to Xit for OLS to consistently estimate beta 1.

However, pooled OLS will be biased and inconsistent if Xit and ai are correlated, the resulting bias is sometimes called heterogeneity bias. In most applications, the main reason for collecting panel data is to allow for the unobserved effect ai to be correlated with explanatory variables. Because, if they are uncorrelated ai and Xit, then a lot of problems are solved.

But if they are correlated, and then we use an OLS estimation technique, then what is happening is that the error term, the composite error is correlated with the independent variable. And this actually violates one of the Gauss Markov or CLRM assumptions. So, OLS is actually not applicable in that context. So, since ai does not change over time, this can be simply achieved by differencing the data for each observation across the two time periods. So, how do we take care of the problem? There are two alternative ways of taking care of the problem. So, the first is to basically go for a difference.

(Refer Slide Time: 9:55)



Fixed Effects Model

- More precisely, for each cross-sectional observation *i*, write the two years as

$$y_{i2} = (\beta_0 + \delta_0) + \beta_1 x_{i2} + a_i + u_{i2} \qquad \text{for } t = 2$$

$$y_{i1} = \beta_0 + \beta_1 x_{i1} + a_i + u_{i1} \qquad \text{for } t = 1$$

- Now subtracting the second equation from the first we obtain

$$y_{i2} - y_{i1} = \delta_0 + \beta_1(x_{i2} - x_{i1}) + (u_{i2} - u_{i1})$$

- Or $\Delta y_{i_t} = \delta_0 + \beta_1 \Delta x_i + \Delta u_i \qquad (4)$
- Where Δ denotes the change from *t* = 1 to *t* = 2. Equation (4) is called the **first-difference equation**. This is a single cross-sectional equation and is estimable by OLS provided it satisfies the key assumptions of OLS.
- The β_1 estimated from equation (4) is called the **first-difference estimator**.

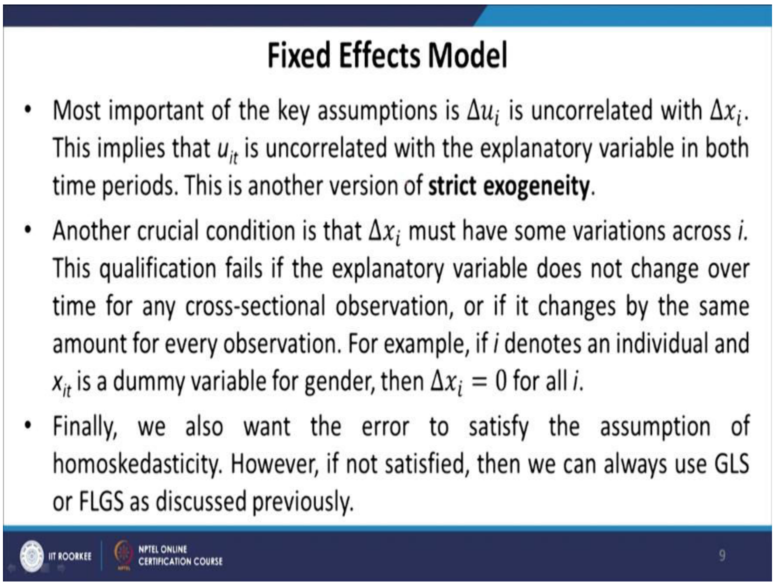
Now, more precisely for each cross-sectional observation, i write the two years separately. I write them separately, as Yi2 equals beta naught plus delta naught, as that is the interceptor, and we have the rest of the things same simply the subscripts changes and similarly, Yi1 equals beta naught plus beta 1 Xt1 plus ai plus Ut1. This is for period 1 *(refer to slide time 09:55).*

Now, subtracting the second equation from the first equation, what we obtain is that Yi2 minus Yi1 and that is equal to beta naught cancels out, I have delta naught plus beta 1 we take common out, and then we have Xi2 minus Xi1 plus ai, ai actually cancels out we have Ui2 minus Ui1. Now, we denote this difference as delta, and then delta naught plus beta 1 delta Xi plus delta Ui. So, delta denotes the change from t equals 1 to t equals 2. This is equation 4, and it is called the first difference equation.

Now, you can see that this equation actually does not have the unobserved component, which is ai. This is a single cross-sectional equation, and is estimable by OLS provided, it satisfied the key assumptions of OLS. So, if the key assumptions of OLS that is the Gauss Markov assumption or the CLRM assumptions are satisfied then this equation can be estimated simply using OLS.

And this actually you can see that, once we difference it out then we do not have the time subscripts here, so this is simply a cross-section. The beta 1 estimated from equation 4 is called the first difference estimator. Essentially, this estimator does not change.

(Refer Slide Time: 12:04)



Fixed Effects Model

- Most important of the key assumptions is Δu_i is uncorrelated with Δx_i. This implies that u_{it} is uncorrelated with the explanatory variable in both time periods. This is another version of **strict exogeneity**.
- Another crucial condition is that Δx_i must have some variations across *i*. This qualification fails if the explanatory variable does not change over time for any cross-sectional observation, or if it changes by the same amount for every observation. For example, if *i* denotes an individual and x_{it} is a dummy variable for gender, then $\Delta x_i = 0$ for all *i*.
- Finally, we also want the error to satisfy the assumption of homoskedasticity. However, if not satisfied, then we can always use GLS or FLGS as discussed previously.

The most important of the key assumptions is, delta Ui is uncorrelated with delta Xi. This implies that Uit is uncorrelated with the explanatory variable in both time periods, and this is another version of strict exogeneity. Strict exogeneity, we had actually introduced in an earlier module, which implied that the excise or the independent variables would be uncorrelated to all Uis, the current one as well as the previous one or they across all observations.

So, that is strict exogeneity. And here also we need strict exogeneity in the sense that, Uit is not correlated with the explanatory variable in both periods. Another crucial condition is that delta Xi must have some variations across i this qualification false if the explanatory variable does not change over time for any cross-sectional observation or if it changes by the same amount for each and every observation.

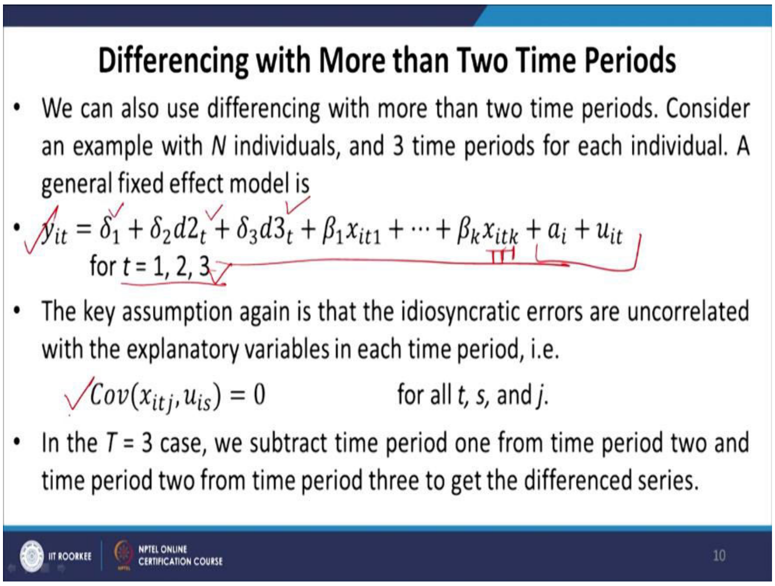
For example, if i denotes an individual and Xit is a dummy variable for gender, then delta Xi equals 0 for all i. This would also be the case, for instance, we are considering population. We are trying to find out the impact on the income of education, and all the people we are considering individuals we are considering they are currently working, which implies that broadly they are finished with their education.

So, all of them would be having the same level of education, and education would remain basically fixed in both periods. Education is not going to change because the time I am considering both the times or in multiple times they are working, and they are finished with their education. So, in that case, delta Xi value would all be the same.

Similarly, age is another thing. As time passes age increases at a constant, in constant numbers for each and every individual. So, if I consider age, then delta Xi would be the same for all individuals. And so, these are the problems and that is why these kinds of variables cannot be included as independent variables in fixed-effect models because then Xi will not have or delta Xi will not have any variations.

Finally, we also want the error to satisfy the assumption of homoscedasticity. However, if not satisfied, then we can always use GLS or FGLS, as we have discussed while discussing heteroscedasticity.

(Refer Slide Time: 14:48)



Differencing with More than Two Time Periods

- We can also use differencing with more than two time periods. Consider an example with N individuals, and 3 time periods for each individual. A general fixed effect model is
- $y_{it} = \delta_1 + \delta_2 d2_t + \delta_3 d3_t + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it}$
  for $t = 1, 2, 3$
- The key assumption again is that the idiosyncratic errors are uncorrelated with the explanatory variables in each time period, i.e.

  $Cov(x_{itj}, u_{is}) = 0$ for all t, s, and j.
- In the $T = 3$ case, we subtract time period one from time period two and time period two from time period three to get the differenced series.

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 10

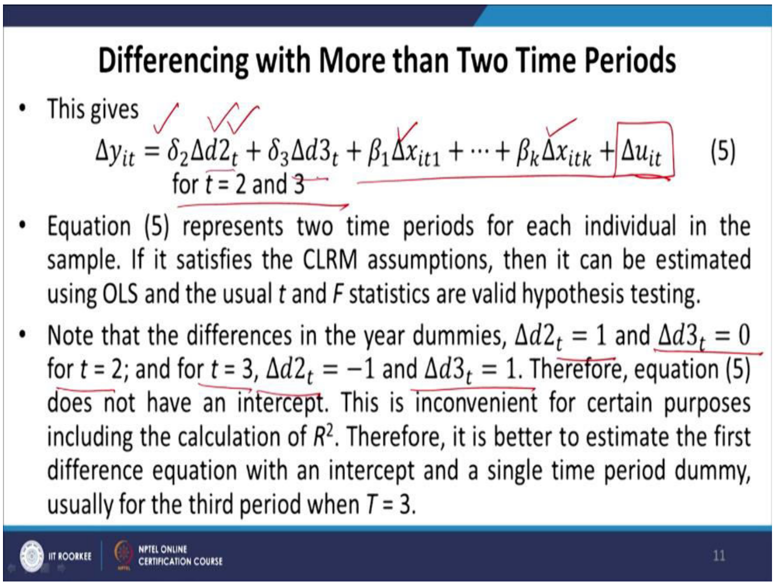
Now, what happens if we have more than two time periods? So, we can also use differencing with more than two time periods. Consider an example with n individuals and 3 time periods for each individual. A general fixed effect model is Yit equals delta 1 plus delta 2 d2t plus delta 3 d3t plus beta 1 Xit 1 plus beta k Xit k plus ai plus Uit. So, these are the unobserved factors.

Now, the equation looks slightly complicated because now X has three components in its subscript, but actually, this is pretty simple because, as we know that, there are three time periods, so there refers to time, it can be 1, 2 or 3, k refers to the dependent variable how many dependent variables are there and which dependent variable it is, and i refers to basically, the individual.

So, this equation is for an individual, and that individual's data will be collected for all three time periods and all the k independent variables. And also to be noted that we are having two-time dummies here, and this is the intercept, which is actually the parameter for the reference period.

Now, the key assumption again is the same here that the idiosyncratic errors are uncorrelated with the explanatory variables in each time period, that is, the covariance between Xitj, and Uis equals 0 for all ts and j. Now, T equals 3 three cases we subtract time period from time period 1 from time period 2, and time period 2 from time period 3 to get the difference series. So, here we will be having two equations. The first equation will be having period 2 minus period 1 and the second equation will be period 3 minus period 2.

(Refer Slide Time: 16:54)



Differencing with More than Two Time Periods

- This gives

$$\Delta y_{it} = \delta_2 \Delta d2_t + \delta_3 \Delta d3_t + \beta_1 \Delta x_{it1} + \cdots + \beta_k \Delta x_{itk} + \boxed{\Delta u_{it}} \qquad (5)$$

for $t = 2$ and 3

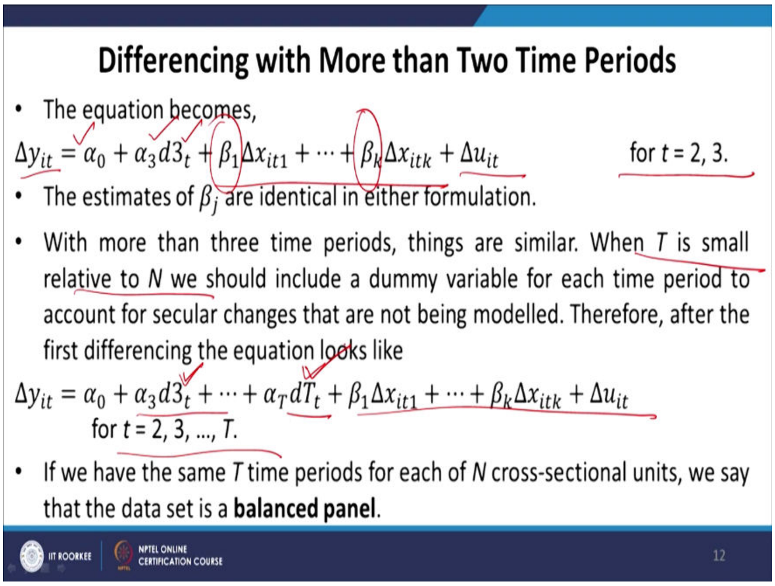
- Equation (5) represents two time periods for each individual in the sample. If it satisfies the CLRM assumptions, then it can be estimated using OLS and the usual *t* and *F* statistics are valid hypothesis testing.
- Note that the differences in the year dummies, $\Delta d2_t = 1$ and $\Delta d3_t = 0$ for $t = 2$; and for $t = 3$, $\Delta d2_t = -1$ and $\Delta d3_t = 1$. Therefore, equation (5) does not have an intercept. This is inconvenient for certain purposes including the calculation of R^2. Therefore, it is better to estimate the first difference equation with an intercept and a single time period dummy, usually for the third period when $T = 3$.

So, this gives delta Yit equals delta 2 and delta d2t and delta 3 delta d3t and the rest of the things are the same having delta in front of the independent variables. The unobserved fixed factor is gone, the time-invariant factor is gone, and we only have delta Uit here, and this is valid for t equals 2 and 3. We do not have any equation for t equals 1.

Now, equation 5 represents two time periods for each individual in the sample. If it satisfies the CLRM assumption, then it can be estimated using OLS, and the usual t and F statistics are also valid for hypothesis testing. Note that, the differences in the year dummies, that is, delta 2t is equal to 1and delta 3t equals 0 for t equals 1. And similarly, for t equals 3 delta 2t would be equals to minus 1, and delta 3t will be equal to 1.

So, therefore, we also note that we do not have any intercept term, instead, we have these expressions here. So, equation 5 does not have an intercept, this is inconvenient for certain purposes including the calculation of r square. Therefore, it is better to estimate the first difference equation with an intercept and a single time period dummy usually for the third period when t equals to 3.

(Refer Slide Time: 18:32)



Differencing with More than Two Time Periods

- The equation becomes,

$\Delta y_{it} = \alpha_0 + \alpha_3 d3_t + \beta_1 \Delta x_{it1} + \cdots + \beta_k \Delta x_{itk} + \Delta u_{it}$ for $t = 2, 3$.

- The estimates of β_j are identical in either formulation.
- With more than three time periods, things are similar. When T is small relative to N we should include a dummy variable for each time period to account for secular changes that are not being modelled. Therefore, after the first differencing the equation looks like

$\Delta y_{it} = \alpha_0 + \alpha_3 d3_t + \cdots + \alpha_T dT_t + \beta_1 \Delta x_{it1} + \cdots + \beta_k \Delta x_{itk} + \Delta u_{it}$

for $t = 2, 3, \ldots, T$.

- If we have the same T time periods for each of N cross-sectional units, we say that the data set is a **balanced panel**.

So, the equation becomes delta Yit, of course, this is the usual thing and this is valid for t equals 2 to 3 and so, are all the independent variables as well as the error term. What changes here, is that instead of having the difference in time dummies we simply have one intercept which is alpha naught and, and one time dummy. So, if we are having three periods, then I am having time dummy only for the third period, which is alpha 3 d3t.

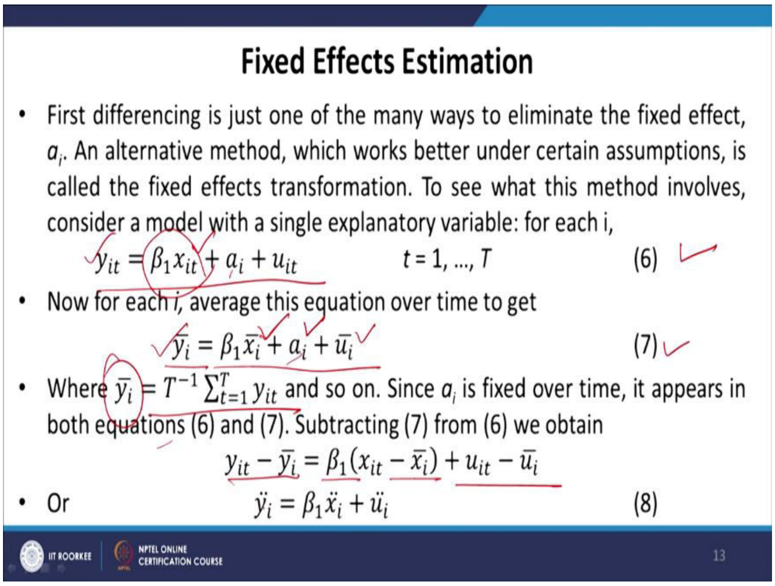
The estimates of beta t are identical in either formulation. So, these are our actually estimates or parameters of interest, and they remain the same. With more than three time periods things are similar, when T is small relative to N, we should include a dummy variable for each time period to account for secular changes that are not being modeled. So, ideally looking for information in the time in itself.

So, capturing information which is basically secular changes, and not there in the explanatory variables. Therefore, after the first differencing the equation looks like d3t to dTt, so we are actually incorporating time dummies except for the first two time dummies, and then the rest of the things are the same. And this equation is valid for t equals capital T that is the number of the time periods we have considered.

But here it is mentioned that when T is small relative to N, because if T is large relative to N, then incorporating so many time dummies would actually make the number of degrees of freedom very small. And that has its own problem. So, that is why we should preferably have a relatively small t compared to the number of observations. And in case we have a small t, then only it is recommended to include time dummies for each and every time period barring the first two.

Now, if we have the same T time periods for each N cross-sectional unit, we say that the data set is a balanced panel. For example, if I am considering data from individuals from the same individuals for say 5 year period 1990 to 95, and I actually obtain data for all these 5 periods for all say 1000 individuals I have considered then this is a balanced panel. As opposed to an unbalanced panel, well probably for some individuals, I would not have data for all 5 years. So, when I do not have data for all the 5 years under consideration from all 1000 individuals, then we call it an unbalanced panel. Now, we will talk about it a little later.

(Refer Slide Time: 21:26)



Fixed Effects Estimation

- First differencing is just one of the many ways to eliminate the fixed effect, a_i. An alternative method, which works better under certain assumptions, is called the fixed effects transformation. To see what this method involves, consider a model with a single explanatory variable: for each i,

$$y_{it} = \beta_1 x_{it} + a_i + u_{it} \qquad t = 1, \ldots, T \qquad (6)$$

- Now for each *i*, average this equation over time to get

$$\bar{y}_i = \beta_1 \bar{x}_i + a_i + \bar{u}_i \qquad (7)$$

- Where $\bar{y}_i = T^{-1} \sum_{t=1}^{T} y_{it}$ and so on. Since a_i is fixed over time, it appears in both equations (6) and (7). Subtracting (7) from (6) we obtain

$$y_{it} - \bar{y}_i = \beta_1 (x_{it} - \bar{x}_i) + u_{it} - \bar{u}_i$$

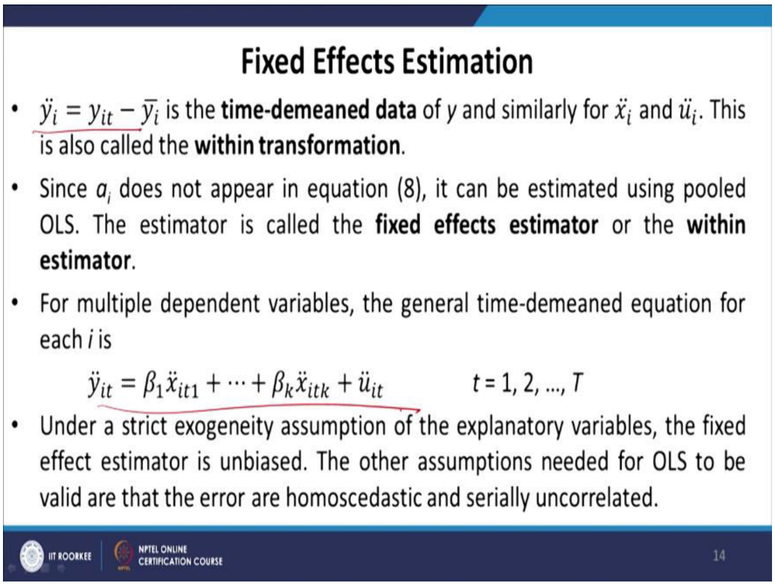
- Or $$\ddot{y}_i = \beta_1 \ddot{x}_i + \ddot{u}_i \qquad (8)$$

So, first differencing is just one of the many ways to eliminate the fixed effect. We are now going to talk about the alternative way of eliminating the fixed effect. An alternative method that works better under certain assumptions is called the fixed effects transformation. To see, what this method involves consider a model with a single explanatory variable like before for each i.

And I have also not mentioned the intercept separately or rather we are not including an intercept here, straight away the independent variable. Now, for each, i average this equation overtime to get Yi bar equals beta 1 Xi bar plus ai plus Ui bar. Now, where Yi bar is simply an average of the dependent variable. Similarly, I will be having an average of the independent variables and then the average of the error terms.

ai being constant or fixed over time, it appears in both equations 6 and 7 and no averaging is possible because it has not changed over a period of time. So, subtracting equation 7 from equation 6, what do we obtain? Yit minus Yi bar, then ai, ai actually cancels out, we have beta 1 Xit minus Xi bar and then you Uit minus Ui bar. Suppose, we denoted by Y double dot i equal to beta 1 X double dot i plus u double dot i.

(Refer Slide Time: 22:58)



Fixed Effects Estimation

- $\ddot{y}_i = y_{it} - \bar{y}_i$ is the **time-demeaned data** of *y* and similarly for $\ddot{x}_i$ and $\ddot{u}_i$. This is also called the **within transformation**.
- Since a_i does not appear in equation (8), it can be estimated using pooled OLS. The estimator is called the **fixed effects estimator** or the **within estimator**.
- For multiple dependent variables, the general time-demeaned equation for each *i* is

$$\ddot{y}_{it} = \beta_1\ddot{x}_{it1} + \cdots + \beta_k\ddot{x}_{itk} + \ddot{u}_{it} \qquad t = 1, 2, \ldots, T$$

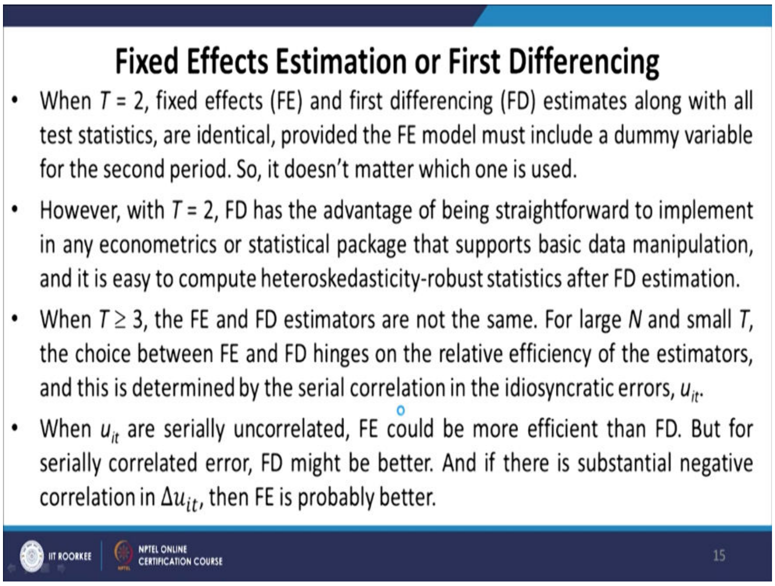
- Under a strict exogeneity assumption of the explanatory variables, the fixed effect estimator is unbiased. The other assumptions needed for OLS to be valid are that the error are homoscedastic and serially uncorrelated.

So, this is how we define it. Ideally, Yi double dot is equal to Yit minus Y bar, and this is called the time demeaned data of y. Similarly, we have the time domain data for x and u. This is also called the within the transformation. Since ai does not appear in equation 8, it can again be estimated using pooled OLS, the estimator is called the fixed effects estimator or the within estimator.

We mentioned here pooled OLS here because, in order to estimate, the model would also include time dummies, once we have considered the time-demeaned observation. for multiple dependent variables the general time-demeaned equation for each i. So, it is very similar to therefore 1 period or 2-period model, we simply consider the mean and the means are subtracted from the actual values in order to arrive at the time-demeaned observations.

Under a strict exogeneity assumption of the explanatory variables, the fixed effect estimator is unbiased. So, if we have these strict exogenously assumptions valid or satisfied, then, of course, OLS estimators are unbiased. The other assumptions needed for OLS to be valid, are that the errors are homoscedastic and serially uncorrelated.

(Refer Slide Time: 24:29)



Fixed Effects Estimation or First Differencing

- When $T = 2$, fixed effects (FE) and first differencing (FD) estimates along with all test statistics, are identical, provided the FE model must include a dummy variable for the second period. So, it doesn't matter which one is used.
- However, with $T = 2$, FD has the advantage of being straightforward to implement in any econometrics or statistical package that supports basic data manipulation, and it is easy to compute heteroskedasticity-robust statistics after FD estimation.
- When $T \geq 3$, the FE and FD estimators are not the same. For large N and small T, the choice between FE and FD hinges on the relative efficiency of the estimators, and this is determined by the serial correlation in the idiosyncratic errors, u_{it}.
- When u_{it} are serially uncorrelated, FE could be more efficient than FD. But for serially correlated error, FD might be better. And if there is substantial negative correlation in Δu_{it}, then FE is probably better.

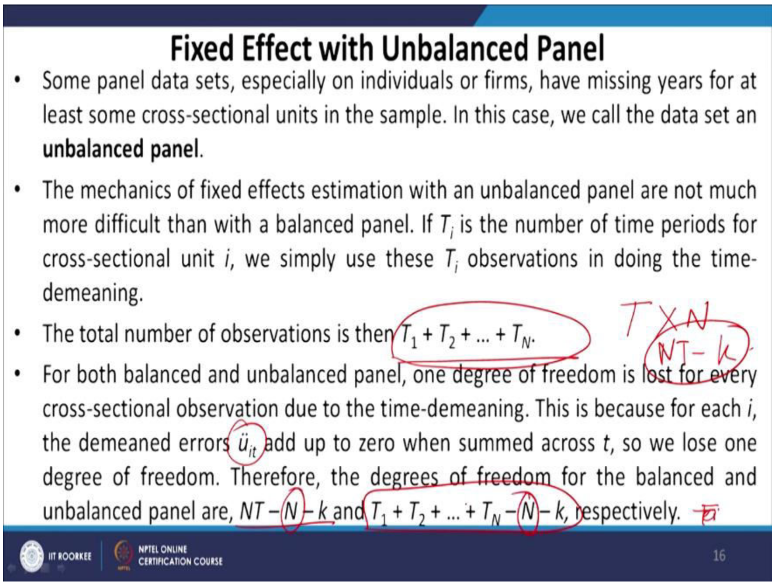
Now, we compare fixed effects with the first differencing. When t equals 2 fixed effects and first differencing estimates along with all test statistics, are identical, provided the fixed effect model must include a dummy variable for the second period, that is, we include a time dummy. So, it does not matter which one is used.

However, with T equals 2, first differencing has the advantage of being straightforward to implement in any econometrics or statistical packages that supports basic data manipulation, and it is easy to compute heteroskedasticity robust-statistic after FD estimation. When T is greater than equals 3, the FE and FD estimators are actually not the same. For large N and small T the choice between FE and FD hinges on the relative efficiency of estimators and this is determined by the serial correlation in the idiosyncratic errors Uig or Uit.

When Uit are serially uncorrelated, FE could be more efficient than FD, that is, the fixed effect is more efficient than first differencing. When Uit are serially uncorrelated and we are considering more than one or two time periods. But for serially correlated error FD might be better because then we are going for the first differencing. I have not talked about random walk models yet, but if Uit is a random walk model, then, it is, we can very easily prove that first differencing actually takes away the problem.

So, in those situations, FD is better. But if there is a substantial negative correlation in delta Uit, then probably fixed effect is better. So, when it comes to T greater than equal to 3, then possibly there is no straight away judgment about, which one is better, that depends on other properties of the data and the error term.

(Refer Slide Time: 26:32)



Fixed Effect with Unbalanced Panel

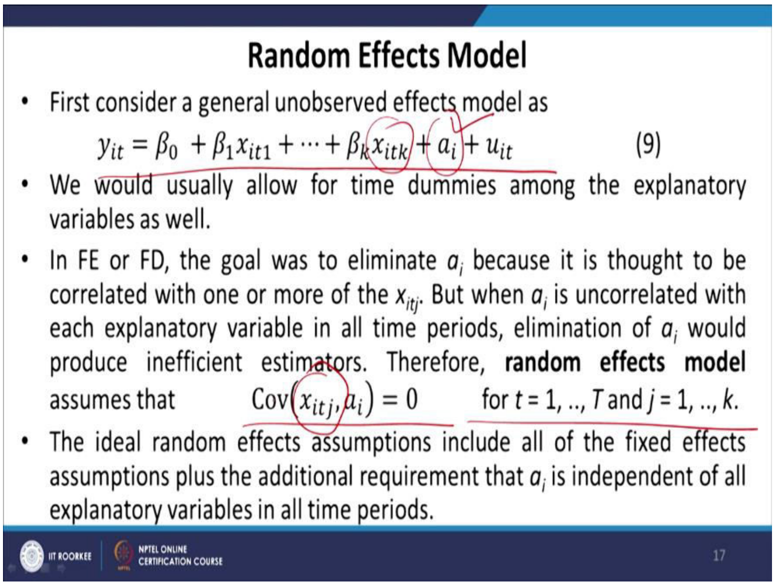
- Some panel data sets, especially on individuals or firms, have missing years for at least some cross-sectional units in the sample. In this case, we call the data set an **unbalanced panel**.
- The mechanics of fixed effects estimation with an unbalanced panel are not much more difficult than with a balanced panel. If T_i is the number of time periods for cross-sectional unit i, we simply use these T_i observations in doing the time-demeaning.
- The total number of observations is then $T_1 + T_2 + \dots + T_N$.
- For both balanced and unbalanced panel, one degree of freedom is lost for every cross-sectional observation due to the time-demeaning. This is because for each i, the demeaned errors $\ddot{u}_{it}$ add up to zero when summed across t, so we lose one degree of freedom. Therefore, the degrees of freedom for the balanced and unbalanced panel are, $NT - N - k$ and $T_1 + T_2 + \dots + T_N - N - k$, respectively.

Some panel data sets especially on individuals or firms of missing years for at least some cross-sectional units in this sample. In this case, we call the data set an unbalanced panel, as I have mentioned a few minutes ago. The mechanics of fixed effects estimation with an unbalanced panel are not much more difficult or different than with a balanced panel. If Ti is the number of time periods for the cross-sectional unit, we simply use these Ti observations in doing the time demeaning.

The total number of observations is then T1 plus T2 plus so on and up to TN. So, you can see that, if we have a balanced panel the total of observation is T multiplied by N, but in case we have an unbalanced panel, this is T1 plus T2 plus TN and so on. For both balanced and unbalanced panels, one degree of freedom is lost for every cross-sectional observation due to the time demeaning.

This is because for each i, the demeaned errors, that is, U double dot it add up to 0 when summed across T, so we will lose one degree of freedom. Therefore, the degrees of freedom for the balanced and unbalanced panel are NT minus N minus k and similarly, T1 plus T2 up to TN minus N minus k. So, usually, we will be having the degrees of freedom equals to two TN or NT minus k. But here, in this case, we are also having minus N here that is because of the time demeaning we are losing outer one degree of freedom.

(Refer Slide Time: 28:19)



Random Effects Model

- First consider a general unobserved effects model as

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it} \qquad (9)$$

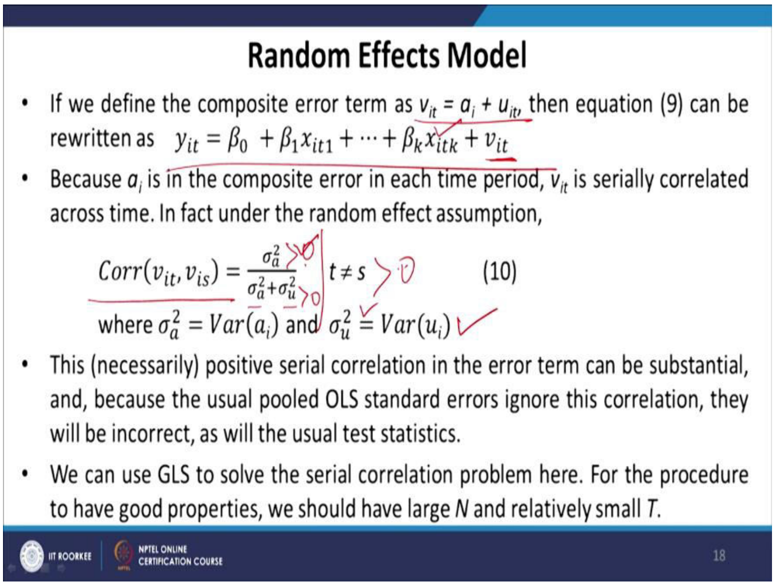
- We would usually allow for time dummies among the explanatory variables as well.
- In FE or FD, the goal was to eliminate a_i because it is thought to be correlated with one or more of the x_{itj}. But when a_i is uncorrelated with each explanatory variable in all time periods, elimination of a_i would produce inefficient estimators. Therefore, **random effects model** assumes that $\mathrm{Cov}(x_{itj}, a_i) = 0$ for $t = 1, .., T$ and $j = 1, .., k$.
- The ideal random effects assumptions include all of the fixed effects assumptions plus the additional requirement that a_i is independent of all explanatory variables in all time periods.

Now, we talk about the other possible model under panel data very briefly, which is the Random Effects Model. So, we first consider a general unobserved effects model, as it was observed earlier or mentioned earlier. We would usually allow for time dummies among the explanatory variables, as well. So, these are not separately mentioned here.

In fixed-effect or first differencing the goal was to eliminate ai because it is thought to be correlated with one or more of the Xit j that is independent variables. But when ai is uncorrelated with each explanatory variable in all time periods elimination of ai would produce inefficient estimators. Therefore, random-effects model assumes that covariance between Xitj and ai equals to 0 for t equals 1 to T, and j equals 1 to k.

So, the random effect model most importantly deviates from the fixed effect model on the basis that the unobserved factor ai which is time-independent, is assumed not to be correlated with the independent variable. It is a completely random unobserved effect, it is not correlated with the independent variables. The ideal random effects assumptions include all of the fixed effect assumptions plus the additional requirement that ai is independent of all the explanatory variables in all time periods.

(Refer Slide Time: 29:53)



Random Effects Model

- If we define the composite error term as $v_{it} = a_i + u_{it}$, then equation (9) can be rewritten as $y_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + v_{it}$
- Because a_i is in the composite error in each time period, v_{it} is serially correlated across time. In fact under the random effect assumption,

$$Corr(v_{it}, v_{is}) = \frac{\sigma_a^2 > 0}{\sigma_a^2+\sigma_u^2 > 0}, \quad t \neq s > 0 \qquad (10)$$

where $\sigma_a^2 = Var(a_i)$ and $\sigma_u^2 = Var(u_i)$ ✓

- This (necessarily) positive serial correlation in the error term can be substantial, and, because the usual pooled OLS standard errors ignore this correlation, they will be incorrect, as will the usual test statistics.
- We can use GLS to solve the serial correlation problem here. For the procedure to have good properties, we should have large N and relatively small T.

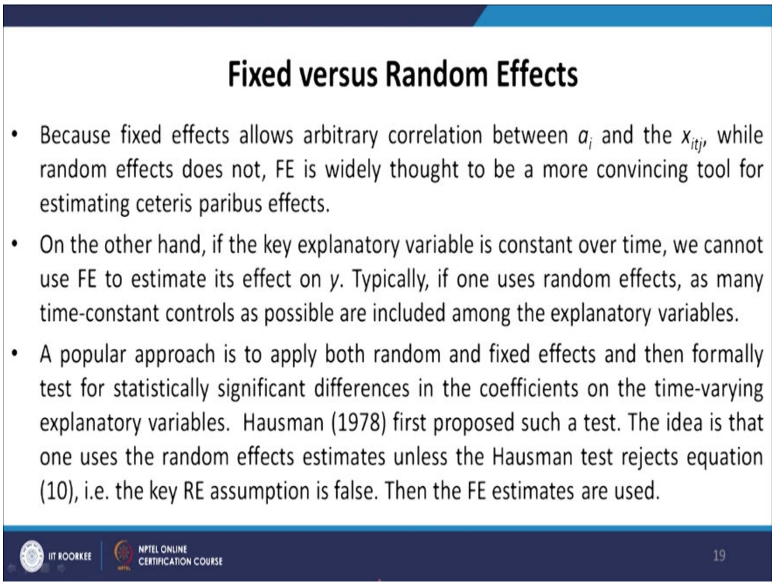
IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE 18

So, if we define the composite error term, Vit as ai plus Uit then equation 9, this our equation 9, can be rewritten as Yit equals the usual thing plus Vit, because ai is in the composite error in each time period, Vit is serially correlated across time. So, now the problem is that Xit is not correlated with Vit but it is really correlated because ai is actually a time-independent component.

In fact, under the random-effect assumption, the covariance, the correlation between Vit and Vis is sigma a squared divided by sigma a squared plus sigma u squared for all t naught equal to s, where sigma a squared is the variance of ai and sigma u square is the variance of Ui *(refer to slide time 29:53)*. This is necessarily positive because this should be positive individually all of them are positive, so the entire thing must be positive. So, this is necessarily a positive serial correlation in the error term that can be substantial.

And because the usual pooled OLS standard errors ignore this correlation, they will be incorrect, as with the usual test statistics. So, because of the presence of this serial correlation, we cannot actually apply OLS, pooled OLS on random-effects model. We can use GLS to solve this serial correlation problem here. For the procedure to have good property so we should have a large N and a relatively small T.

(Refer Slide Time: 31:28)



Fixed versus Random Effects

- Because fixed effects allows arbitrary correlation between a_i and the x_{itj}, while random effects does not, FE is widely thought to be a more convincing tool for estimating ceteris paribus effects.
- On the other hand, if the key explanatory variable is constant over time, we cannot use FE to estimate its effect on y. Typically, if one uses random effects, as many time-constant controls as possible are included among the explanatory variables.
- A popular approach is to apply both random and fixed effects and then formally test for statistically significant differences in the coefficients on the time-varying explanatory variables. Hausman (1978) first proposed such a test. The idea is that one uses the random effects estimates unless the Hausman test rejects equation (10), i.e. the key RE assumption is false. Then the FE estimates are used.

IIT ROORKEE NPTEL ONLINE CERTIFICATION COURSE 19

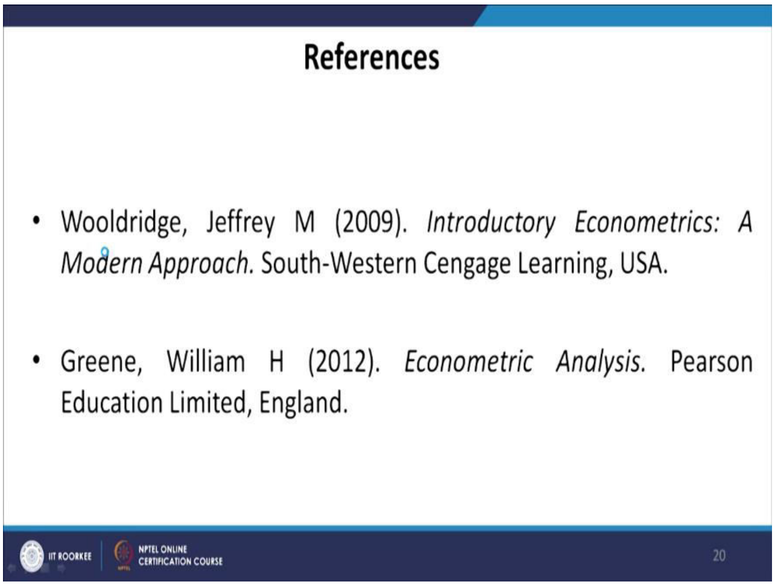
So, now, we just compare the fixed effects and the random-effects model. Because fixed effects allow arbitrary correlation between ai and the Xitj, while random effects do not. The fixed effect is widely thought to be a more convincing tool for estimating ceteris paribus effects. So, since we are not always certain about whether independent variables are correlated with the unobserved factors or not, it is better to always go for fixed effects under the assumption that, they could be correlated.

On the other hand, if the key explanatory variables are constant over time, we cannot use FE to estimate its effect on Y. For example, we have already mentioned that certain factors like education for the working population, age, and cross-gender do not change over a period of time, and as a result of which they cannot be used as independent variables in a fixed-effect model.

Typically, if one uses random effects, as many time constant controls, as possible are included among the explanatory variables. So, the random effect has this advantage. A popular approach is to apply both random and fixed effects, and then, formally test for statistically significant differences in the coefficients on the time-varying explanatory variables.

Hausman first proposed such a test in 1978. The idea is that one uses the estimates of the random effects unless the Hausman test rejects equation 10. So**,** this is my equation 10 or expression 10, *(refer to slide time 31:28),* which basically says that there is a serial correlation. So, if this assumption is actually rejected or if this assumption is false, then we can go for fixed effect otherwise, we will be estimating the models using random effects.

So, the idea is that one uses random effects estimates unless the Hausman test rejects equation 10, that is, the key RE assumption is false, Then the FE estimates are used, otherwise, we will go with random-effects model. So, that is broadly about the panel data methods.



References

- Wooldridge, Jeffrey M (2009). *Introductory Econometrics: A Modern Approach.* South-Western Cengage Learning, USA.
- Greene, William H (2012). *Econometric Analysis.* Pearson Education Limited, England.

These are the books I have followed in order to come up with today's discussion. Thank you.