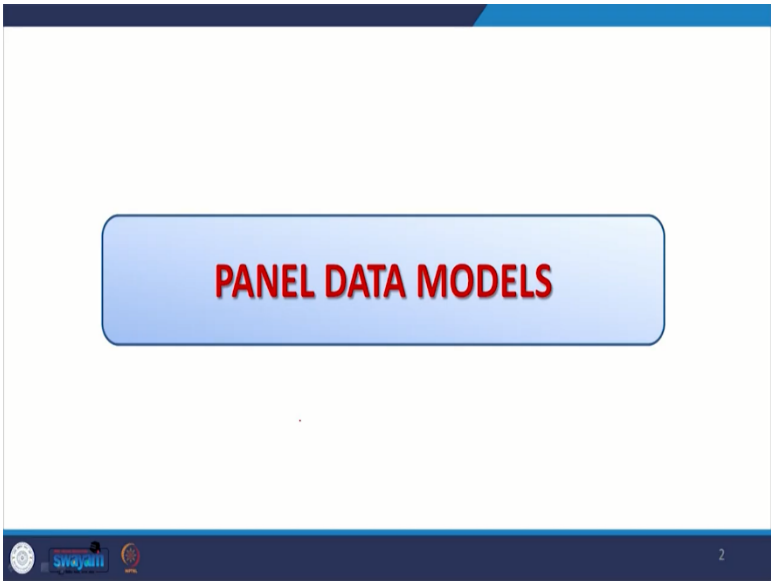
Handling Large-Scale Unit Level Data Using STATA

Professor Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee

Lecture 37
Panel Data Models

So, welcome once again to the NPTEL MOOC module on handling large scale data. We are at the verge of the final week of our module. We already started discussion of the panel data in the last lecture and last lecture was purely conceptual of panel data. I always suggest everyone to not to miss the previous lecture because that contains the background of the panel data. Any word we are using, we have already clarified in the previous lecture. So, that is my suggestions for the participant to certainly go through just to have a look. You will certainly be motivated to go for panel data analysis further.

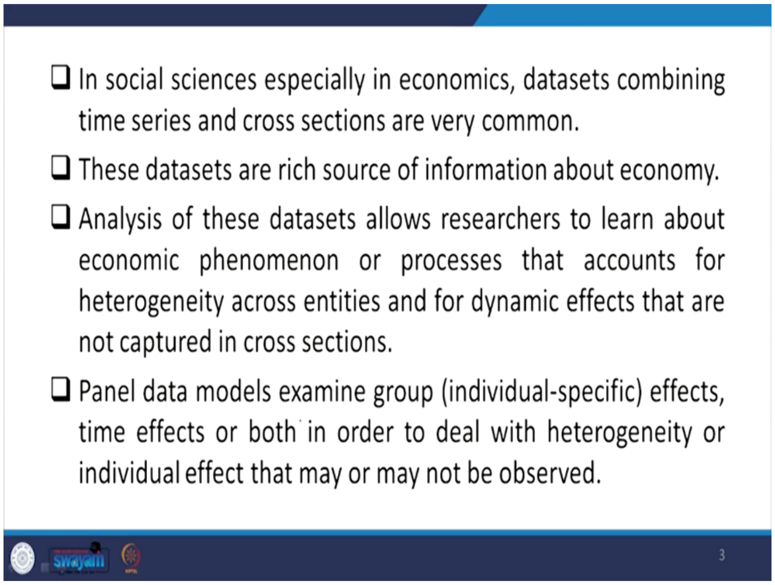
(Refer Slide Time: 01:13)



PANEL DATA MODELS

In this lecture we will have clarification related to panel data models which type of models are used and why they are used, .

(Refer Slide Time: 01:24)



❑ In social sciences especially in economics, datasets combining time series and cross sections are very common.

❑ These datasets are rich source of information about economy.

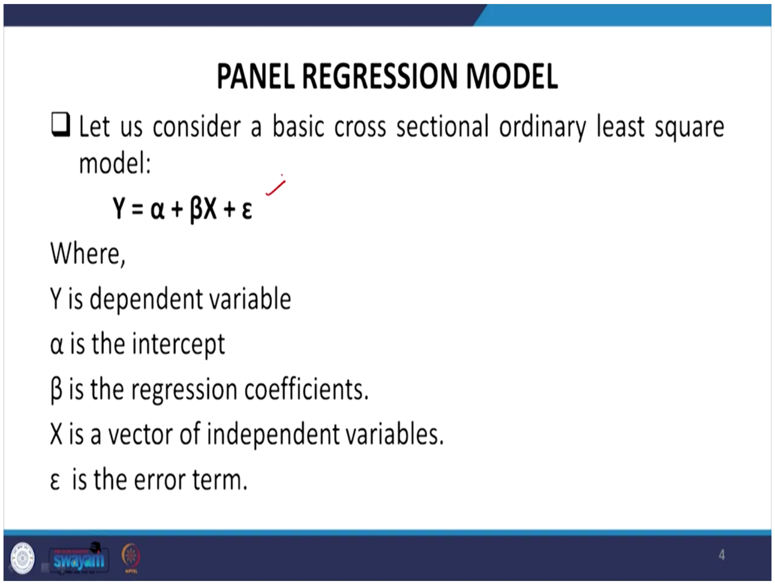
❑ Analysis of these datasets allows researchers to learn about economic phenomenon or processes that accounts for heterogeneity across entities and for dynamic effects that are not captured in cross sections.

❑ Panel data models examine group (individual-specific) effects, time effects or both in order to deal with heterogeneity or individual effect that may or may not be observed.

So, here in social science, especially in economics, datasets combining time series and cross sections are very common. These data sets are rich source of information about economic analysis of these data sets, always allows us to research, to learn about economic phenomenon or processes that accounts for heterogeneity across entities and their dynamics, or the effects of the changes that are not captured in cross-section analysis.

Panel Data Model examines group effects that is individual specific. Or time effects or both individual specific. Basically, a group effects that is called group effects in panel and time effects or both together in order to deal with the heterogeneity or individual effect that may or may not be observed. Coming to the model. Let us start with a simple ordinary least square model.

(Refer Slide Time: 02:27)



PANEL REGRESSION MODEL

❑ Let us consider a basic cross sectional ordinary least square model:

$$Y = \alpha + \beta X + \varepsilon$$

Where,

Y is dependent variable

α is the intercept

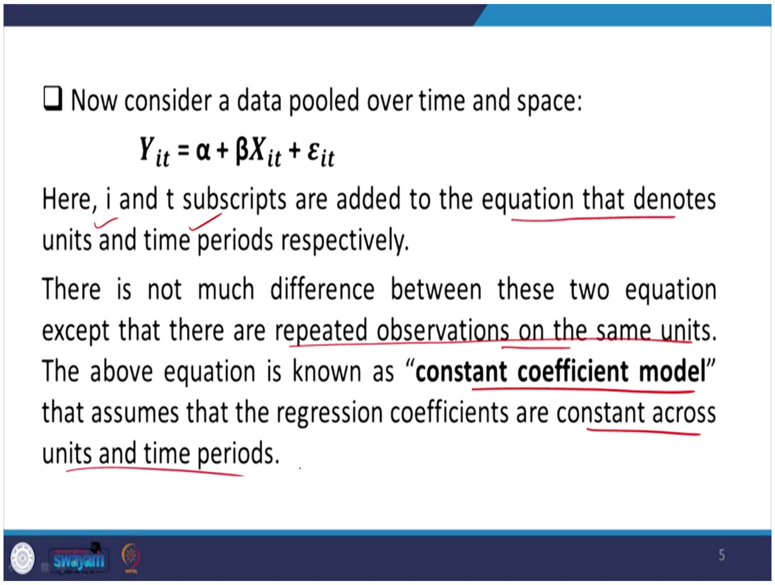
β is the regression coefficients.

X is a vector of independent variables.

ε is the error term.

Y is equal to Alpha Plus beta x plus Epsilon? . We have already clarified Y is dependent variable, alpha is the intercept then beta is the regression coefficient, X is the vector of independent variables and Epsilon is the error term.

(Refer Slide Time: 02:44)



❑ Now consider a data pooled over time and space:

$$Y_{it} = \alpha + \beta X_{it} + \varepsilon_{it}$$

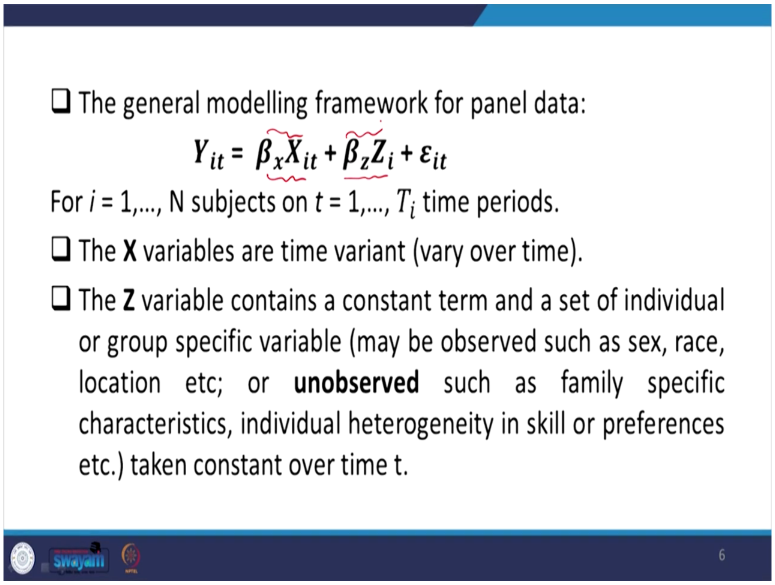
Here, i and t subscripts are added to the equation that denotes units and time periods respectively.

There is not much difference between these two equation except that there are repeated observations on the same units. The above equation is known as "**constant coefficient model**" that assumes that the regression coefficients are constant across units and time periods.

In this case we are trying to pooled this data over time and space to make it panel format, . Here we are using two subscripts. I mentioned from the beginning that panel comes with two subscripts, subscript i and t. What we are mentioning carefully that i and t subscripts in this model like Y_{it} Alpha plus Beta X_{it} plus Epsilon it, here i and t are added to the equation that denotes units and time period respectively. There is not much difference between these two equations, except that there are repeated observations on the same units . Because of the time variable.

The above equation is known as constant coefficient models, . Constant coefficient model because we have included a constant coefficient also in this model that assumes that the regression coefficients are constant across units and time period, . So, that is basically called constant coefficient model. The general modelling framework for the panel data is like this.

(Refer Slide Time: 04:00)



□ The general modelling framework for panel data:

$$Y_{it} = \beta_x X_{it} + \beta_z Z_i + \varepsilon_{it}$$

For $i = 1,\ldots,$ N subjects on $t = 1,\ldots, T_i$ time periods.

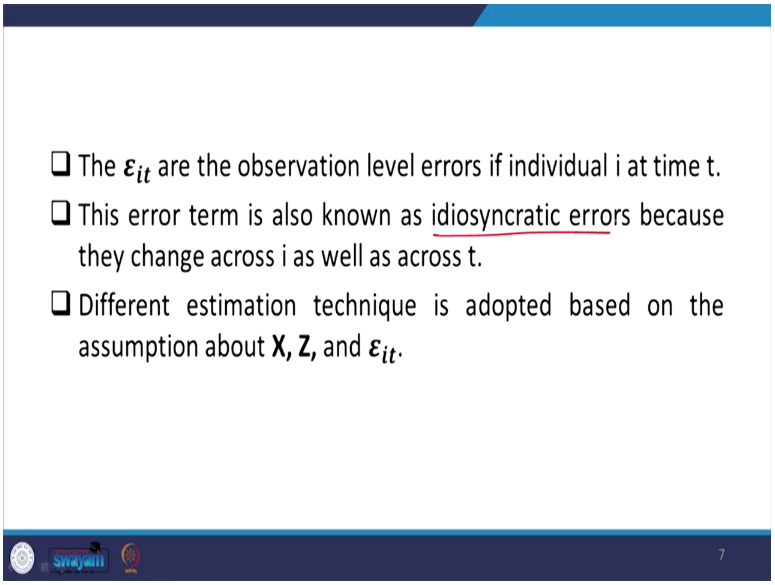
□ The **X** variables are time variant (vary over time).

□ The **Z** variable contains a constant term and a set of individual or group specific variable (may be observed such as sex, race, location etc; or **unobserved** such as family specific characteristics, individual heterogeneity in skill or preferences etc.) taken constant over time t.

Like Y_{it}, the constant coefficient is not mentioned. Rather, it is mentioned with this term and without the t component, the time component term like here X variables are time variant and the Z variable are the individual or space group specific variable usually like sex, race, location or on observed variables such as family specific characteristics, individual heterogeneity in skill preferences, et cetera.

Those usually taken constant over time. Those does not vary over time. So broadly, there are two core component, time component and cross sectional. The cross section component having no time variance. With the epsilon term these are called panel model.

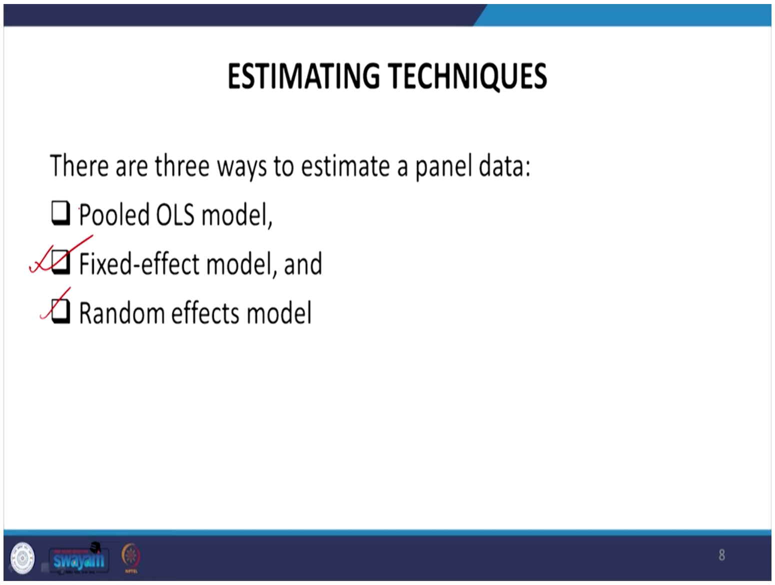
(Refer Slide Time: 05:00)



- □ The ε_{it} are the observation level errors if individual i at time t.
- □ This error term is also known as idiosyncratic errors because they change across i as well as across t.
- □ Different estimation technique is adopted based on the assumption about **X, Z,** and ε_{it}.

Here the error term that varies over the individual as well as over time. These error term is also known as idiosyncratic error. So, there are randomness attached with the error term. Stochasticity attached with across individual across observation, also across time variable they also the famously used concept called idiosyncratic errors, because they change across i as well as across t, not just i across t also. So, these are called idiosyncratic errors, alright. So, different estimation techniques adopted based on the assumption about X, Z and the error term.

(Refer Slide Time: 05:55)



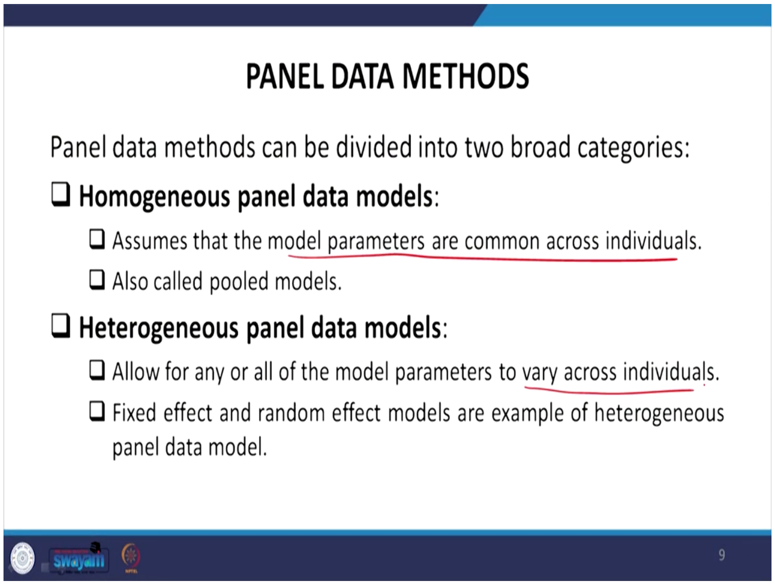
ESTIMATING TECHNIQUES

There are three ways to estimate a panel data:

- □ Pooled OLS model,
- ☑ Fixed-effect model, and
- ☑ Random effects model

Let us come to the estimating techniques. There are broadly three estimation technique discussed in panel. One is called pooled OLS model, fixed effect and random effect model. You might have heard about this majorly these two. But most importantly, fixed effect is used largely, but we will also discuss pooled OLS model.

(Refer Slide Time: 06:14)



PANEL DATA METHODS

Panel data methods can be divided into two broad categories:

- ❑ **Homogeneous panel data models**:
  - ❑ Assumes that the model parameters are common across individuals.
  - ❑ Also called pooled models.
- ❑ **Heterogeneous panel data models**:
  - ❑ Allow for any or all of the model parameters to vary across individuals.
  - ❑ Fixed effect and random effect models are example of heterogeneous panel data model.

9

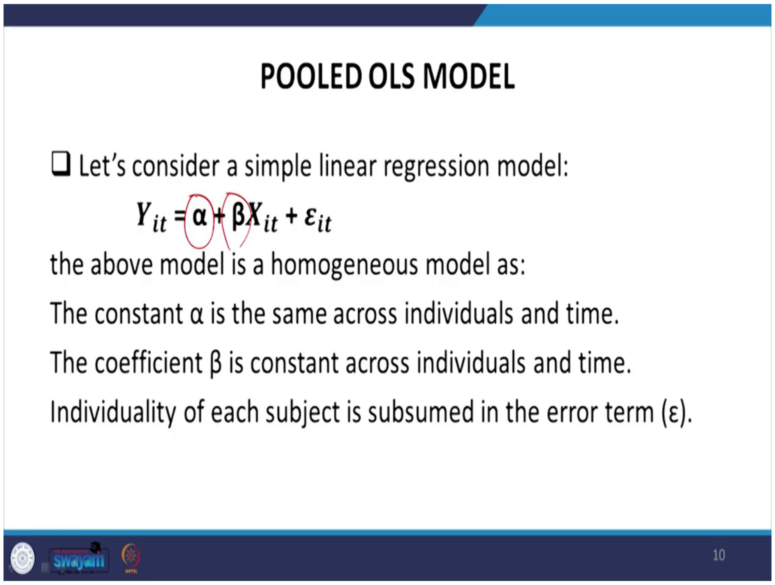
Coming to the panel data methods. This can be divided into two broad categories. One is called homogeneous panel data models and heterogeneous panel model. What do you mean by homogeneous panel data model? This assumes that the model parameters are common across individuals. The parameters we consider are common across individual, across i term we are taking.

The parameters are considered to be common, the estimated parameters to be common where in case of heterogeneous panel data model, this allow for any or all of the model parameters to vary across individuals1. The parameters that vary across individuals are heterogeneous panel data. homogenous panel data is usually called pooled models, .

In case a pooled we have to make the variables name to be same, the estimator for across individual or the parameter to be estimated across the individual to be same, isn't it? Whereas in case of heterogeneous panel data model we do not have the same parameter across individuals. We need to take fixed effect and random effect models in this particular case, especially for the case of heterogeneous panel data.

Let us compare pooled OLS model. Let us understand pooled OLS model is the first category of the panel format. In that case, we said that the parameters does not vary over time.

(Refer Slide Time: 07:58)



POOLED OLS MODEL

❑ Let's consider a simple linear regression model:

$$Y_{it} = \alpha + \beta X_{it} + \varepsilon_{it}$$

the above model is a homogeneous model as:

The constant α is the same across individuals and time.

The coefficient β is constant across individuals and time.

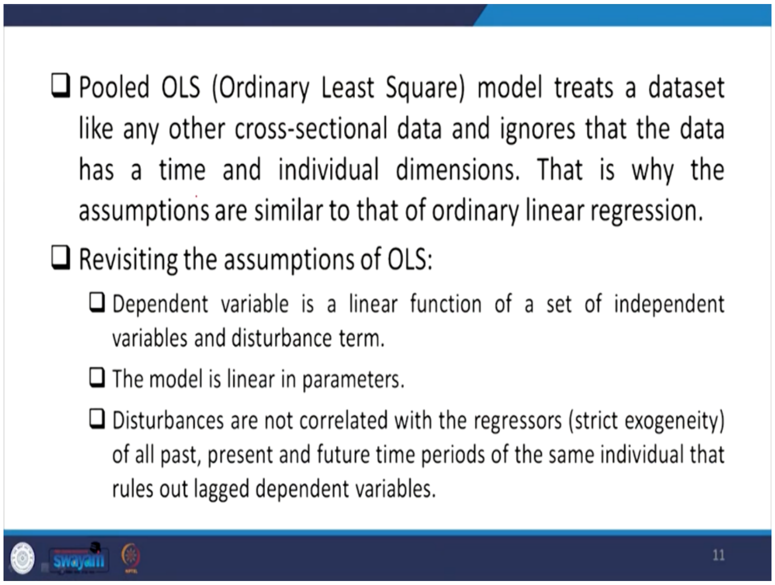
Individuality of each subject is subsumed in the error term (ε).

10

But that parameters are going to be constant for the observations. So, that is the reason why parameters are not added with subscript i or t. All right. So, here the alpha and beta you look at, we are not adding this subscript. So, the above model is also called a homogeneous model, not the heterogeneous model. So, the constant alpha is the same across individuals and time. The coefficient beta is constant across individual and time as well. Individuality of each subject is subsumed in the error term.

So, the error term subsumes the individuality of the data or information. Pooled OLS model treats the database like any other cross-sectional data ignores that the data has a time and an individual dimension is simply considering cross section units. But there are some problems we are going to discuss. That is the reason why assumption is similar to that of the ordinary least square model.

(Refer Slide Time: 09:11)



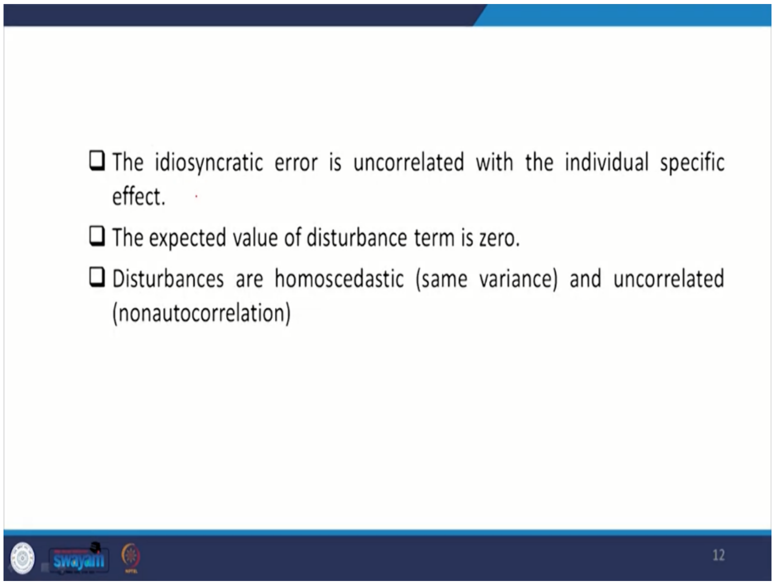
❑ Pooled OLS (Ordinary Least Square) model treats a dataset like any other cross-sectional data and ignores that the data has a time and individual dimensions. That is why the assumptions are similar to that of ordinary linear regression.

❑ Revisiting the assumptions of OLS:

- ❑ Dependent variable is a linear function of a set of independent variables and disturbance term.
- ❑ The model is linear in parameters.
- ❑ Disturbances are not correlated with the regressors (strict exogeneity) of all past, present and future time periods of the same individual that rules out lagged dependent variables.

Revisiting the assumption of OLS especially in the case of pooled data, there are some challenges. Dependent variable is a linear function of a set of independent variables and disturbance term. So, the standard assumption and the model is linear in parameters. Disturbances are not correlated with the regression or regressors. The error terms should not be correlated with the regressors that is called strict exogeneity of all past, present and future time period that has been pooled together in the data. It rules out lagged dependent variable.

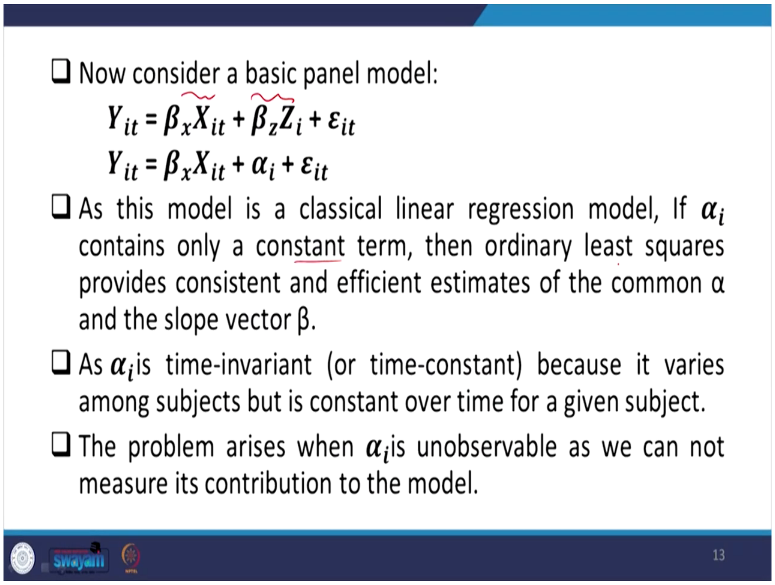
(Refer Slide Time: 09:44)



- ❑ The idiosyncratic error is uncorrelated with the individual specific effect.
- ❑ The expected value of disturbance term is zero.
- ❑ Disturbances are homoscedastic (same variance) and uncorrelated (nonautocorrelation)

The idiosyncratic error is uncorrelated with the individual specific effects. These are the assumption; the expected value of disturbance term is therefore 0. So, disturbances are homoscedastic or that is having same variance and uncorrelated in nature. These are called having non autocorrelation.

(Refer Slide Time: 10:09)



□ Now consider a basic panel model:

$$Y_{it} = \beta_x X_{it} + \beta_z Z_i + \varepsilon_{it}$$

$$Y_{it} = \beta_x X_{it} + \alpha_i + \varepsilon_{it}$$

□ As this model is a classical linear regression model, If α_i contains only a constant term, then ordinary least squares provides consistent and efficient estimates of the common α and the slope vector β.

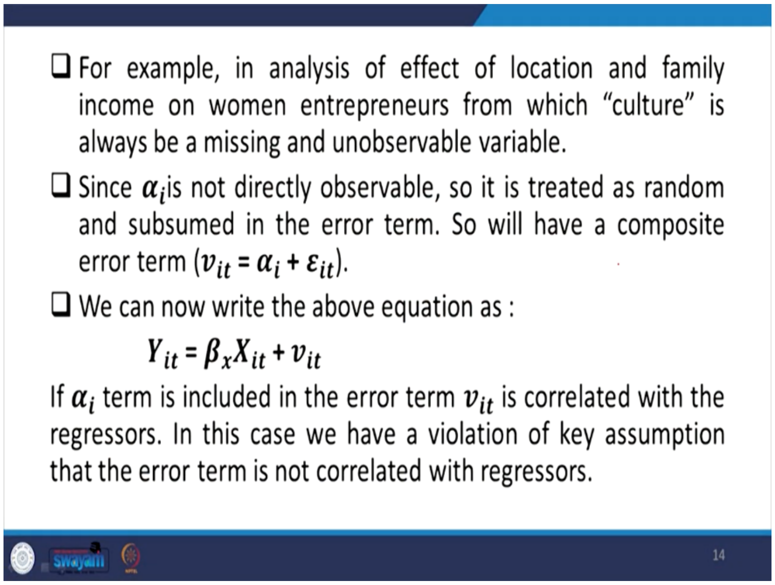
□ As $\alpha_i$is time-invariant (or time-constant) because it varies among subjects but is constant over time for a given subject.

□ The problem arises when $\alpha_i$is unobservable as we can not measure its contribution to the model.

Consider a basic panel model. Start with a basic panel model then. The basic panel model comes with beta X, the coefficient. Then the time variant variable and then time invariant variable. That is the basic panel model. Alright. Here we are writing another equation, giving information about the constant term varies across individuals. So, in that case, what is going to happen? This model is a classical linear regression model. If Alpha contains only a constant term B because beta X is explained with the change in the the time variant observations.

Alpha contains a Constant term and then ordinarily least square provides consistent and efficient estimates of the common alpha and the slope vector beta. Alright. As Alpha i is time invariant. We have mentioned because t component is not attached in Alpha i. Because it varies across objects and it is constant over time for a given subject. The problem arises when Alpha is unobservable and as we cannot measure its contribution to the model. Since Alpha i is mentioned and that is not observable. Then that is really a problem to the model how it is. Let us explain it.

(Refer Slide Time: 11:50)



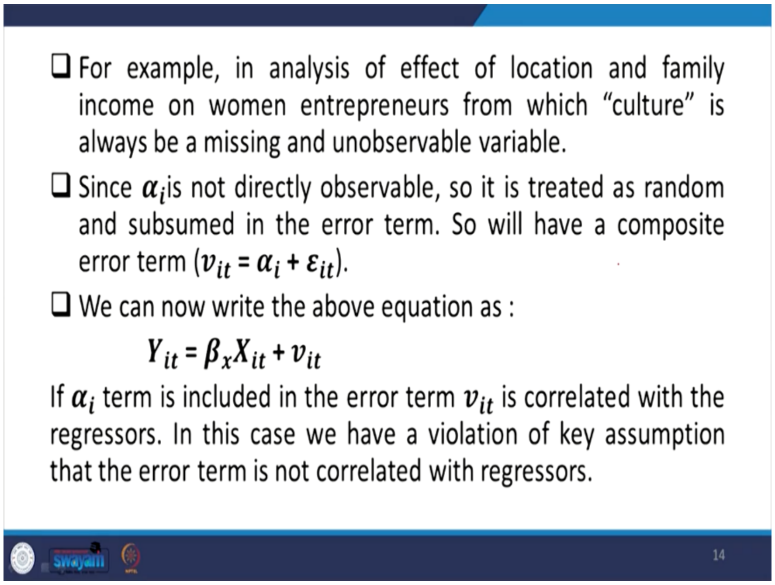
- ❑ For example, in analysis of effect of location and family income on women entrepreneurs from which "culture" is always be a missing and unobservable variable.
- ❑ Since $\alpha_i$is not directly observable, so it is treated as random and subsumed in the error term. So will have a composite error term ($v_{it} = \alpha_i + \varepsilon_{it}$).
- ❑ We can now write the above equation as :

$$Y_{it} = \beta_x X_{it} + v_{it}$$

If α_i term is included in the error term v_{it} is correlated with the regressors. In this case we have a violation of key assumption that the error term is not correlated with regressors.

For example, in analysis of effect of location and family income on women entrepreneurs from which culture is always be a missing and unobservable variable, which is not observed. So, how to capture it? That is basically representing your alpha in this case. Since Alpha is not directly observable, so it is treated as a random term and usually subsumed in the error term, error distribution. So, the error distribution in this case is a composite error term. That is v_{it}. We are defining Alpha i plus Epsilon it. So, alpha is subsumed in the error term. So, we are carrying with a v_{it} term.

(Refer Slide Time: 12:36)



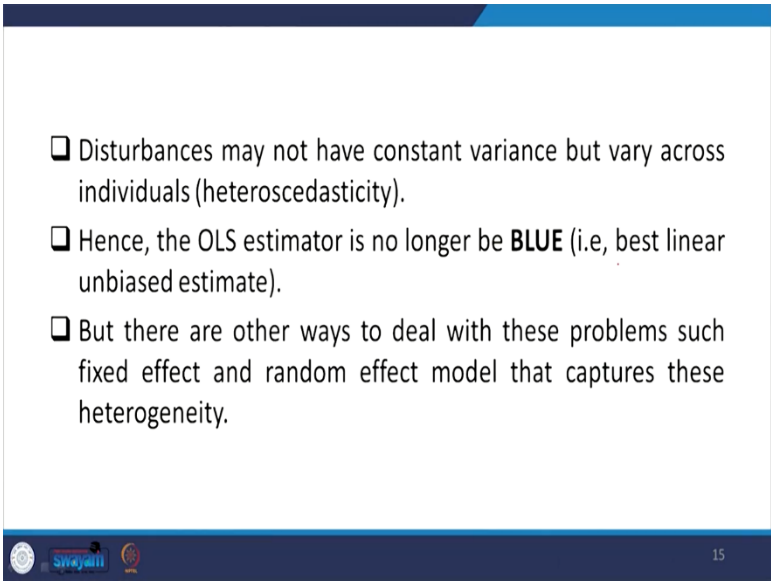
- ❑ For example, in analysis of effect of location and family income on women entrepreneurs from which "culture" is always be a missing and unobservable variable.
- ❑ Since $\alpha_i$is not directly observable, so it is treated as random and subsumed in the error term. So will have a composite error term ($v_{it} = \alpha_i + \varepsilon_{it}$).
- ❑ We can now write the above equation as :

$$Y_{it} = \beta_x X_{it} + v_{it}$$

If α_i term is included in the error term v_{it} is correlated with the regressors. In this case we have a violation of key assumption that the error term is not correlated with regressors.

So, alpha i term is included in the error term that is $v_{it,}$ correlated with the regressors. Since it is a subsumed in the error term that is obviously correlated with the regressors in this case, we have a violation of key assumption of the standard OLS model or the classical linear regression model. So, the error term is not correlated with the regressors.

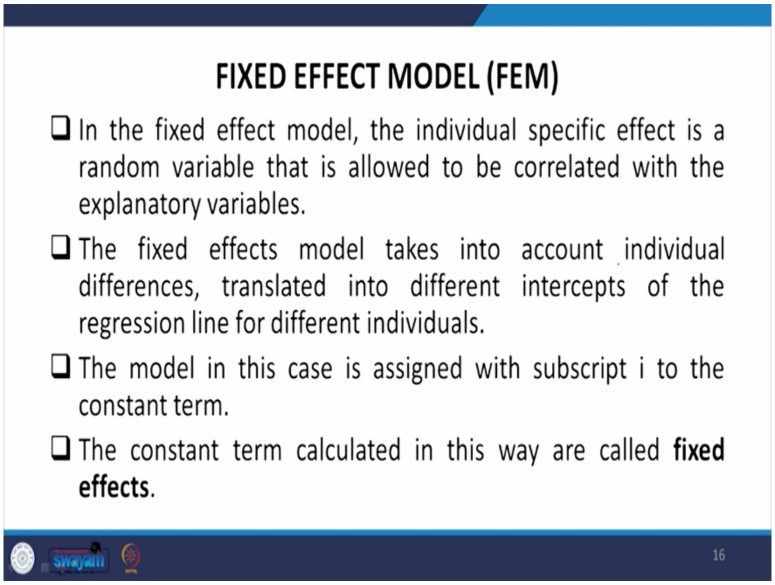
(Refer Slide Time: 13:05)



- □ Disturbances may not have constant variance but vary across individuals (heteroscedasticity).
- □ Hence, the OLS estimator is no longer be **BLUE** (i.e, best linear unbiased estimate).
- □ But there are other ways to deal with these problems such fixed effect and random effect model that captures these heterogeneity.

So, disturbances may not have constant variance, but vary across individuals. And that is basically committing with an error called heteroscedasticity. Alright. Hence OLS estimator is no longer defined to be BLUE, best linear unbiased estimator. But there are other ways to deal with this problem such as fixed effect and random effect model that captures this heterogeneity.

(Refer Slide Time: 13:29)



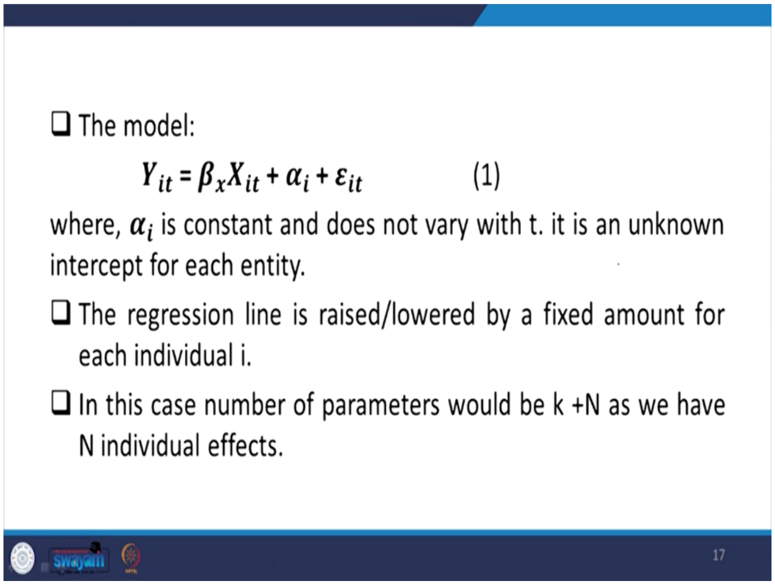
FIXED EFFECT MODEL (FEM)

- ❑ In the fixed effect model, the individual specific effect is a random variable that is allowed to be correlated with the explanatory variables.
- ❑ The fixed effects model takes into account individual differences, translated into different intercepts of the regression line for different individuals.
- ❑ The model in this case is assigned with subscript i to the constant term.
- ❑ The constant term calculated in this way are called **fixed effects**.

Coming to the classification of fixed effect model. Then last one will be on the random effect model. Fixed effect model is the model. The individual specific that is a random variable. The individual specific effect, individual effect is random variable that is allowed to be correlated with the explanatory variable. So, the individual specific effect is correlated with the explanatory variable that we are considering.

The fixed effects model takes into account individual differences. Translated into different intercept of the regression line for different individuals, like basically the individual effect that is captured through the fixed effect model considered to be the intercept in the regression line. And the model in this case is assigned with the subscript called i to the Constant term. We are going to clarify now. The constant term is calculated in this way, are called fixed effect like here.

(Refer Slide Time: 14:39)



❑ The model:

$$Y_{it} = \beta_x X_{it} + \alpha_i + \varepsilon_{it} \qquad (1)$$

where, α_i is constant and does not vary with t. it is an unknown intercept for each entity.

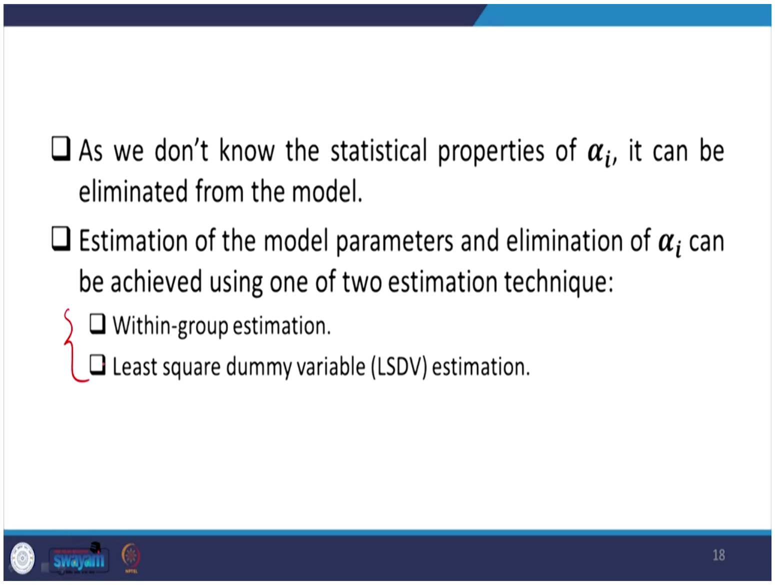
❑ The regression line is raised/lowered by a fixed amount for each individual i.

❑ In this case number of parameters would be k +N as we have N individual effects.

When we mention the same model where we have already committed with heteroscedasticity through the OLS approach. We are converting into the alpha i that is constant over time, but vary across individual. The individual heterogeneity is captured here. This does not vary with t but it is unknown intercept for each entity.

The regression line is raised or lowered by a fixed amount for individual, that is i. The regression line with the constant term will be adjusted either higher or lower, depending upon the alpha i. In this case, a number of parameters would be then k plus N, as we have N individual effects. So, N individual effects are there and accordingly, we can find out. Alright.

(Refer Slide Time: 15:38)

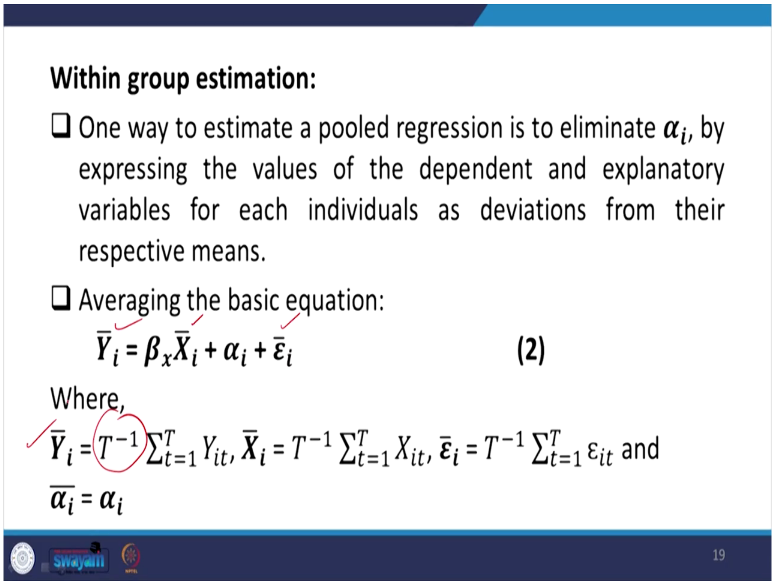


- □ As we don't know the statistical properties of α_i, it can be eliminated from the model.
- □ Estimation of the model parameters and elimination of α_i can be achieved using one of two estimation technique:
  - □ Within-group estimation.
  - □ Least square dummy variable (LSDV) estimation.

As we do not know the statistical properties of Alpha i, it can be eliminated from the model. The question here, since it violates the OLS properties or the assumptions especially the problem with heteroscedasticity was there. So, alpha i was problematic. So, if you can eliminate Alpha i from the model, then we are very much assured with the standard calculation estimation, the model parameter an elimination of Alpha i can be achieved using one of two estimation techniques.

There are two approaches, largely we are going to discuss. So, within estimation or then least square dummy variable estimation LSDV, famously known. So, within group we can eliminate, since it varies with the group so we can eliminate with the average impact within the group, then that is called within estimator. Then second one is through dummy variable within group estimation. One way to estimate a full regression is to eliminate an alpha i by expressing the value of the dependent and explanatory variables for each individuals as deviation from their respective means.

(Refer Slide Time: 16:48)



Within group estimation:

- ❑ One way to estimate a pooled regression is to eliminate α_i, by expressing the values of the dependent and explanatory variables for each individuals as deviations from their respective means.
- ❑ Averaging the basic equation:

$$\bar{Y}_i = \beta_x \bar{X}_i + \alpha_i + \bar{\varepsilon}_i \qquad \textbf{(2)}$$

Where,

$\bar{Y}_i = T^{-1} \sum_{t=1}^{T} Y_{it}$, $\bar{X}_i = T^{-1} \sum_{t=1}^{T} X_{it}$, $\bar{\varepsilon}_i = T^{-1} \sum_{t=1}^{T} \varepsilon_{it}$ and

$\bar{\alpha}_i = \alpha_i$

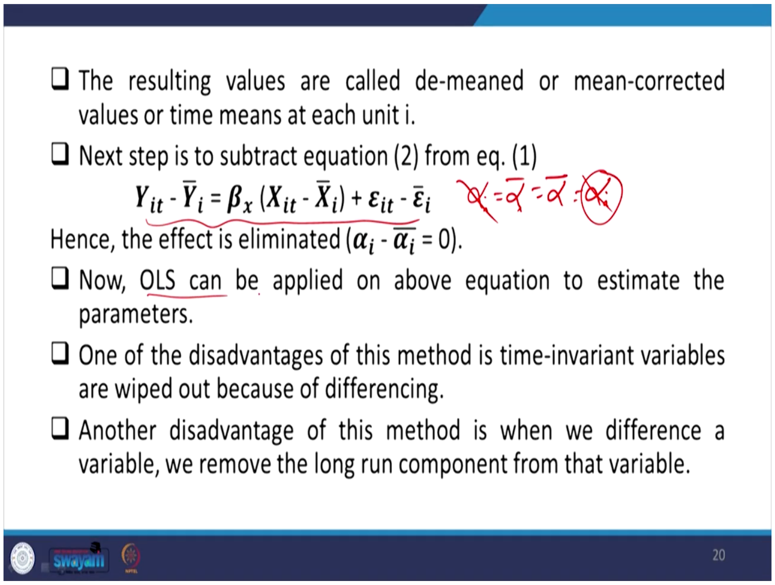
19

Since it does not vary over time. If you in our data, simply take the average of that alpha over time, then that will carry a constant term. And so if you subtract that average that will eliminate the constant individual effect. Like we are going to show it right now averaging the basic equation, the equation we have already taken is Y_{it}. We are going to take the average of that.

Now the model converts to beta X. The average of X_i that is X bar i, Y_i bar is equal to X_i bar plus alpha i since it is constant over time. So, it is not going to be defined as Alpha bar. Simply it is alpha i because it varies from person to person not over time. Alright. But the epsilon term added with a time because you are dividing the time average.

So, all these are in average, this is an average, this is an average, but this will be constant. Once we take the average term of it. So here what is the average? Because 1 upon t sum of the i varies from t to run total time period one to t time period of Y_{it} that will give the average of this. Similarly, for other terms also.

(Refer Slide Time: 18:22)



- ❑ The resulting values are called de-meaned or mean-corrected values or time means at each unit i.
- ❑ Next step is to subtract equation (2) from eq. (1)

$$\underbrace{Y_{it} - \bar{Y}_i = \beta_x (X_{it} - \bar{X}_i) + \varepsilon_{it} - \bar{\varepsilon}_i} \qquad \cancel{\alpha_i} = \bar{\alpha} = \bar{\alpha} = \boxed{\cancel{\alpha_i}}$$

Hence, the effect is eliminated ($\alpha_i - \overline{\alpha_i} = 0$).

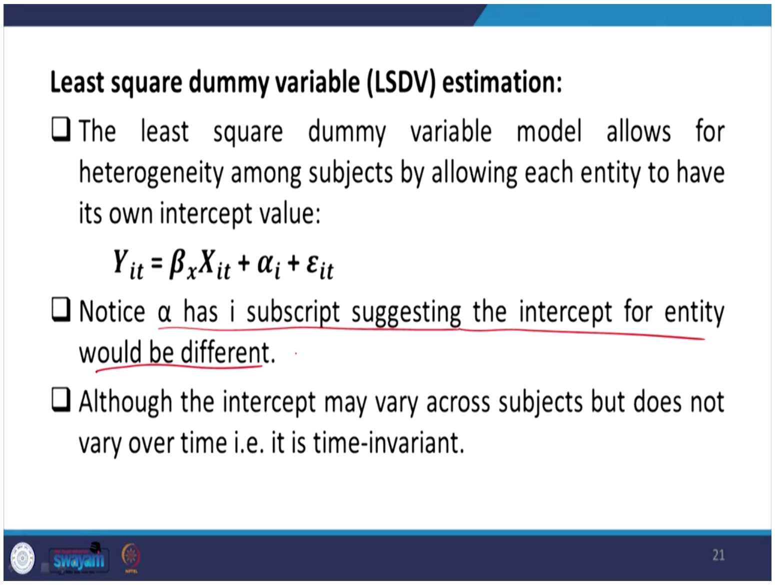
- ❑ Now, OLS can be applied on above equation to estimate the parameters.
- ❑ One of the disadvantages of this method is time-invariant variables are wiped out because of differencing.
- ❑ Another disadvantage of this method is when we difference a variable, we remove the long run component from that variable.

The resulting values are called de-meaned or mean corrected values or time means at each unit i time means basically have been taken. Next step is to subtract the original model to the average transformed model. So, what we have taken the Y_{it} minus the Y_i bar because t has been eliminated in this model. So, what is left here? Since Alpha i or Alpha is nothing but Alpha bar is equal to alpha bar every time.

So, basically Alpha is every time it is alpha i so even average we have taken alpha i remains. So, if you subtract, both the equations so Alpha i cancelled in both the case. What is left here? is simply the net change from its average. So, it seems as if we are left with a standard regression equation. Hence, the average effect is eliminated. That is 0. OLS can be applied. This is what I just said, OLS can be applied on above equation to estimate the parameter.

One of the disadvantages of this method is time invariant variables. Because time we have already eliminated. So, we are not going to capture the time invariant impact or effects, alright. Since these are wiped out because of the differencing technique. Another disadvantage of this method is when we difference a variable, we remove the long run component also of that particular variable.

(Refer Slide Time: 20:09)



Least square dummy variable (LSDV) estimation:

- ❑ The least square dummy variable model allows for heterogeneity among subjects by allowing each entity to have its own intercept value:

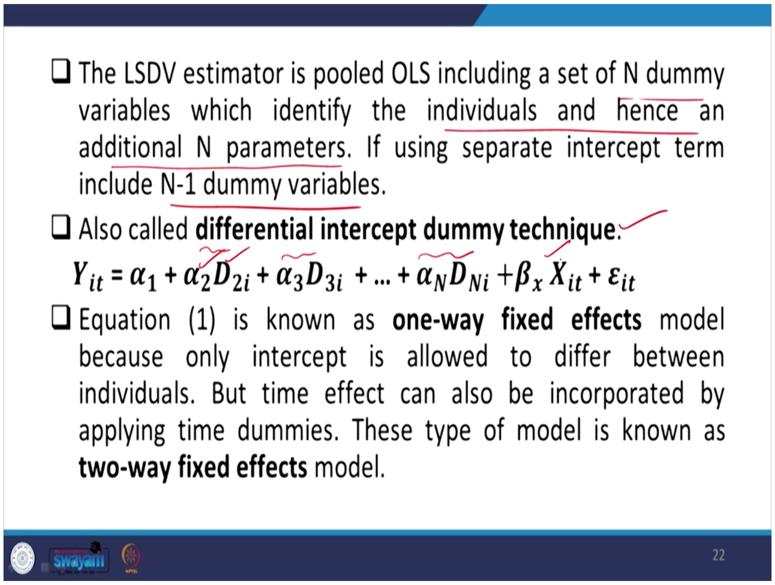
$$Y_{it} = \beta_x X_{it} + \alpha_i + \varepsilon_{it}$$

- ❑ Notice α has i subscript suggesting the intercept for entity would be different.
- ❑ Although the intercept may vary across subjects but does not vary over time i.e. it is time-invariant.

The second approach of the fixed effect model is called the least square dummy variable estimation, this allows heterogeneity. Allows for heterogeneity among subjects also by allowing each entity to have its own intercept value.

Like we have eliminated the individual impact. Alpha was simply eliminated in the previous within group model. In the LSDV we are not eliminating it. Rather we are putting them in different categories, different dummies. Like in the same model framework. Alpha has i subscript suggesting the intercept for entity would be different. Although the intercept may vary across subject but does not vary over time it is time invariant. So, it has value but it is time invariant.

(Refer Slide Time: 21:13)



❑ The LSDV estimator is pooled OLS including a set of N dummy variables which identify the individuals and hence an additional N parameters. If using separate intercept term include N-1 dummy variables.

❑ Also called **differential intercept dummy technique**.

$$Y_{it} = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \ldots + \alpha_N D_{Ni} + \beta_x X_{it} + \varepsilon_{it}$$

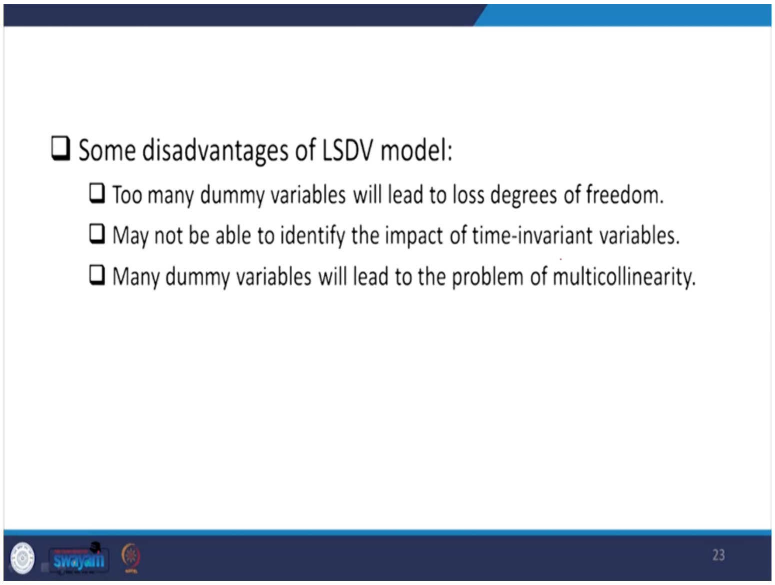
❑ Equation (1) is known as **one-way fixed effects** model because only intercept is allowed to differ between individuals. But time effect can also be incorporated by applying time dummies. These type of model is known as **two-way fixed effects** model.

The LSDV estimator is pooled including a set of N dummy variable, n dummy which i just mentioned, N dummy variables which identify the individuals and hence an additional N parameter since we are defining dummies to interpret it N minus parameters, dummies are defined. If using separate intercept term, we are left with N minus 1 dummy variables.

So, N minus one dummy variables has been captured. So this is also called differential intercept dummy technique. So, in simple equation, Y_{it} is equal to Alpha one the time invariant value we are just mentioning in terms of dummy, right, it is a constant intercept with first dummy, the second dummy till n^{th} dummy.

So, in total we are including N minus one dummy because we start with alpha two dummy variables. $D2_i$ is the first dummy variables with its time variant explanatory variables and this error term. So, equation one is known as one way fixed effects. So, this is one way fixed effect because only intercept is allowed to differ between individuals. But time effect can also be incorporated by applying time dummies. Alright. This type of model is called two way fixed effects models. If we also find out the time effect as well time dummies could also be mentioned.

(Refer Slide Time: 23:03)



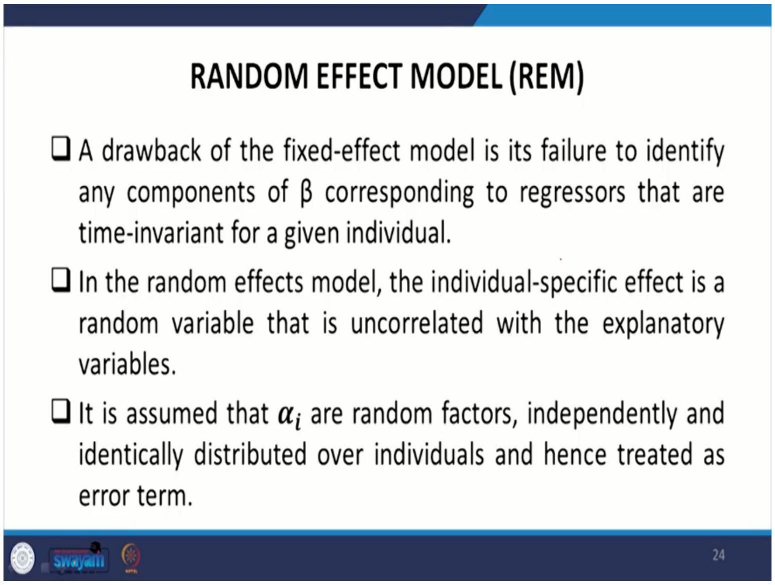
□ Some disadvantages of LSDV model:

- □ Too many dummy variables will lead to loss degrees of freedom.
- □ May not be able to identify the impact of time-invariant variables.
- □ Many dummy variables will lead to the problem of multicollinearity.

There are some disadvantages of this dummy variable technique that, we are including so many dummies. There are for based on the nth observation, if you are including N minus one dummy is the number of estimator, the parameter becoming very high, which reduces the degrees of freedom.

So, this may not be able to identify the impact of time invariant variables, many dummy variables will lead to the problem of multicollinearity also because so many dummy within a model. We have already mentioned using our quality variable models that it may confront with multicollinearity problem.

(Refer Slide Time: 23:41)



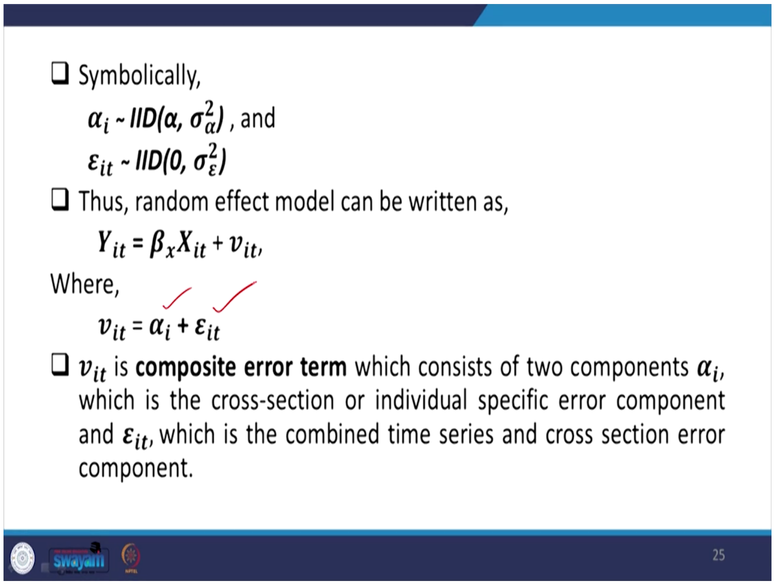
RANDOM EFFECT MODEL (REM)

- ❑ A drawback of the fixed-effect model is its failure to identify any components of β corresponding to regressors that are time-invariant for a given individual.
- ❑ In the random effects model, the individual-specific effect is a random variable that is uncorrelated with the explanatory variables.
- ❑ It is assumed that α_i are random factors, independently and identically distributed over individuals and hence treated as error term.

Coming to the understanding of random effect. The third the last model we are discussing and its clarification are giving is that a drawback of the fixed effect model is its failure to identify any component of beta corresponding to regressor that are time invariant for a given individual. Beta which are time invariant are also not properly captured.

In the random effects model. The individual specific effect is a random variable. Even the individual specific effect is also a random variable that is uncorrelated with the explanatory variable. So, that is the alpha i, which you are saying that is also a time variant. So, that is captured through considering the assumption of its distribution identically distributed over individuals and over time as well.

(Refer Slide Time: 24:41)



- Symbolically,

  $\alpha_i \sim IID(\alpha, \sigma_\alpha^2)$, and

  $\varepsilon_{it} \sim IID(0, \sigma_\varepsilon^2)$

- Thus, random effect model can be written as,

  $Y_{it} = \beta_x X_{it} + v_{it},$

Where,

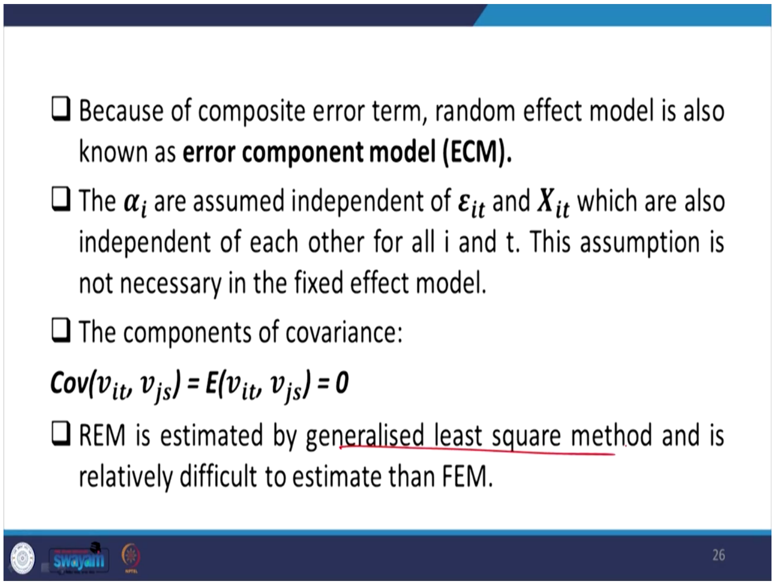
  $v_{it} = \alpha_i + \varepsilon_{it}$

- v_{it} is **composite error term** which consists of two components α_i, which is the cross-section or individual specific error component and ε_{it}, which is the combined time series and cross section error component.

How we are mentioning here symbolically, Alpha i in this case assumed to be identically independent and identically distributed with its mean alpha and standard deviation Sigma square Alpha of its distribution. Then its error term distributed with 0 and standard deviation of that particular distribution. Thus, the random effect model can be written as here, since we have captured that Alpha into the time variant error distribution that is in composite error term.

And here we have taken the name is v_{it}. So, v_{it} is nothing but alpha i plus e Epsilon it. So, v_{it} is a composite term that consists of two components that is Alpha i, which is cross-sectional or individual specific error component and error epsilon 'it' is the combined times series. And cross sectional error component.

(Refer Slide Time: 25:42)



- ❑ Because of composite error term, random effect model is also known as **error component model (ECM).**
- ❑ The α_i are assumed independent of ε_{it} and X_{it} which are also independent of each other for all i and t. This assumption is not necessary in the fixed effect model.
- ❑ The components of covariance:

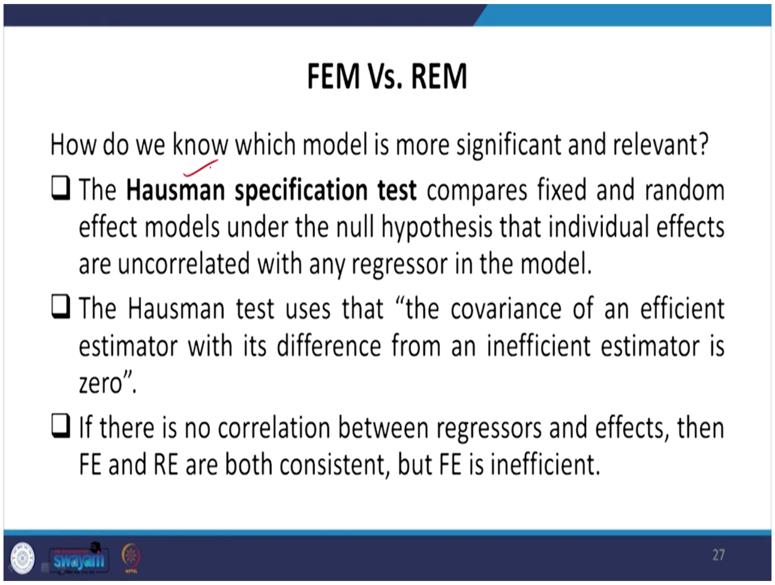
$$Cov(v_{it}, v_{js}) = E(v_{it}, v_{js}) = 0$$

- ❑ REM is estimated by generalised least square method and is relatively difficult to estimate than FEM.

Because of the composite error term, random effect model is also known as error component model in short it is written as ECM. The alpha i are assumed independent of the error term and the explanatory variables which are also independent of each other for all i and t this assumption is not necessarily in the fixed effect model.

So, we need to well understand whether it is a fixed effect or not, we have some technique mentioned in this slide, whether to differentiate whether the data is in fixed effect model or in the random model. The components of variants, is the covariance basically is assumed to be 0. That is in the random effect model estimated by generalized least square method. GLS we are going to apply and is relatively difficult to estimate then that of the fixed effect model.

(Refer Slide Time: 26:52)



FEM Vs. REM

How do we know which model is more significant and relevant?

- ❑ The **Hausman specification test** compares fixed and random effect models under the null hypothesis that individual effects are uncorrelated with any regressor in the model.
- ❑ The Hausman test uses that "the covariance of an efficient estimator with its difference from an inefficient estimator is zero".
- ❑ If there is no correlation between regressors and effects, then FE and RE are both consistent, but FE is inefficient.

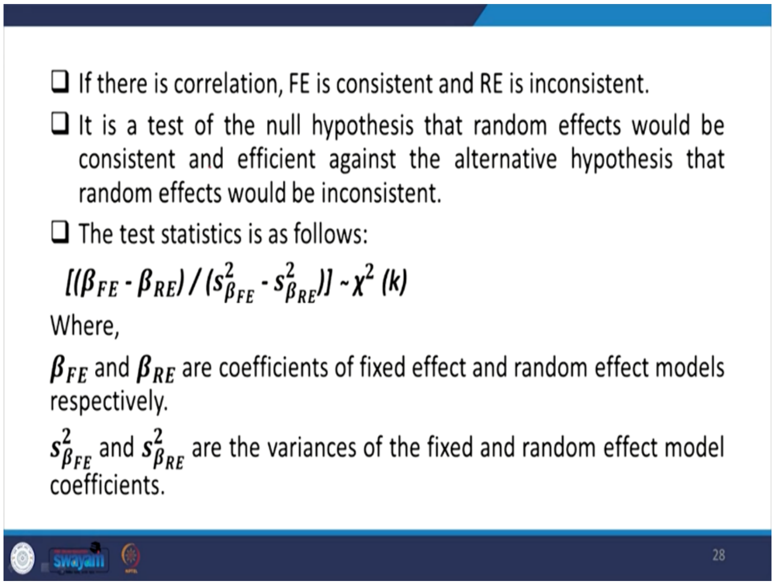
So, comparison to FEM that is fixed effect model and random effect model, we need to understand that, how do we know which model is more significant and relevant? The Hausman test is very very important to be noted. For your record, Hausman specification test compares fixed and random effect model under the null hypothesis that individual effects are uncorrelated with any regression in the model.

So, the individual effect is uncorrelated with the regressor that is the Alpha i we are mentioning it is uncorrelated with the regressors. Alright. That is the assumption. So, the Hausman test uses the covariance of an estimate of an efficient estimator with it, difference from an inefficient estimator is 0.

So, basically it uses the Hausman test, gives that the covariance of an efficient estimator, with its difference from an inefficient estimator zero. Covariance is expected to be zero. If there is no correlation between regressors and effects. Then fixed effect and random effect are both consistent. Alright. But fixed effect is inefficient when there is no correlation between.

Basically, that is the basic assumption of the random effect model that, it is expected to be having no correlation. It has to be stochastically distributed the individual effect which has been included in the composite error term has to be randomly distributed. And if we are getting from our test that there is no correlation between regression and effects, then these are the assumption violated. So, that is why the fixed effect is inconsistent. Alright.

(Refer Slide Time: 28:56)



- □ If there is correlation, FE is consistent and RE is inconsistent.
- □ It is a test of the null hypothesis that random effects would be consistent and efficient against the alternative hypothesis that random effects would be inconsistent.
- □ The test statistics is as follows:

$$[(\beta_{FE} - \beta_{RE}) / (s^2_{\beta_{FE}} - s^2_{\beta_{RE}})] \sim \chi^2\ (k)$$

Where,

β_{FE} and β_{RE} are coefficients of fixed effect and random effect models respectively.

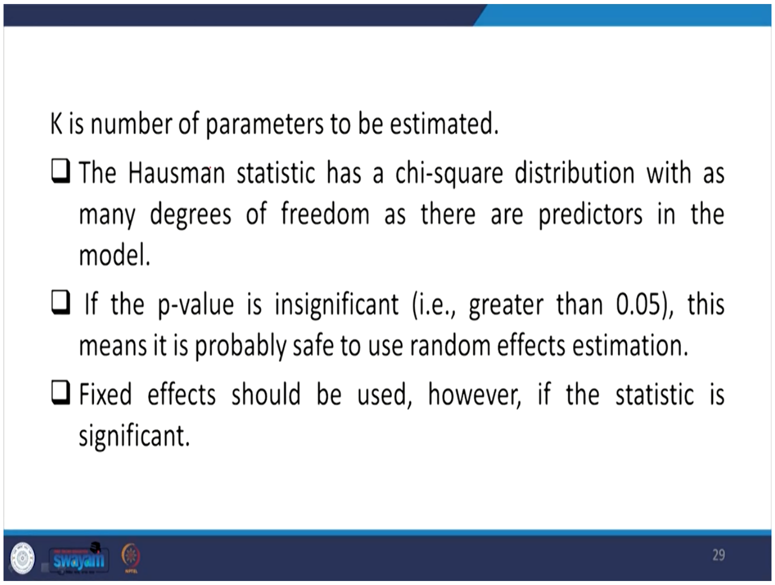
$s^2_{\beta_{FE}}$ and $s^2_{\beta_{RE}}$ are the variances of the fixed and random effect model coefficients.

So, if there is correlation, fixed effect is consistent, and so random effect is inconsistent in that case. Correlation means the individual effect is not randomly distributed. So, you have to nullify the individual effect in the model, you have to separate it out by either group effect or by LSDV. Alright.

So, it is a test of null hypothesis. That random effect would be consistent and efficient against the alternative hypothesis. That random effect would be inconsistent, alright. The test statistic is as follows. So, the test statistic is the beta fixed effect minus beta random effect divided by their net standard errors.

So, that is the standard deviation of fixed effect and in the standard deviation of the random effect. And if that is estimated through the chi square distribution with k degrees of freedom. So, in that case, beta fixed effect and beta random effect are coefficients of fixed effect and random effect model respectively S square FE. S square RE are variances, since we are taking square of it, variance of the fixed and effect model coefficients.

(Refer Slide Time: 30:49)



K is number of parameters to be estimated.

- ❑ The Hausman statistic has a chi-square distribution with as many degrees of freedom as there are predictors in the model.
- ❑ If the p-value is insignificant (i.e., greater than 0.05), this means it is probably safe to use random effects estimation.
- ❑ Fixed effects should be used, however, if the statistic is significant.

K is the number of parameters to be estimated. So, in case of chi square test, generally the k in the bracket denotes the number of parameters to be estimated. The Hausman statistic has a chi square distribution with as many degrees of freedom as there are predictors in the model depending upon the predictors accordingly. We also write down the degrees of freedom. If the P value is insignificant, that is greater than 0.05. This means it is probably safe to use random effects model.

So, once p value you get, it will show it in the next class, one it is exceeding 0.05. It is safe to use a random effect model and fixed effect should be used. However, if the statistic is significant. That means if your assumption is valid, then you can apply your fixed effect model otherwise if it is not significant, you may not use it.

So, these are the clarifications so far on fixed effect model, random effect model even pooled OLS model with group effect we discussed with LSDV model. So, these are the details for today and for this particular lecture, and we will apply the STATA package in the next class with the real data. So, thank you so much. We look forward to all of you in the next class. Thank you.