Business Analytics for Management Decision
Prof. Rudra P Pradhan
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

Lecture – 33
Predictive Analytics: Panel Data Model

Hello everybody this is Rudra Pradhan here; welcome you all to BMD lecture series. Today, we will continue with predictive analytics and the topic of discussion is on panel data modeling. The problem a problem starts with understanding the data and then we will connect with a model. As a result, we will get to know more details about the panel data modeling in the last couple of lectures we have solved. So, many problems by using predictive analytics and. So, many tools are there is and a every times we are solving problems with respect to data and that to in a time series format or in a kind of in a cross sectional format.

But, in this particular you know lectures we will go little bit you know more about this sometimes you know by using cross sectional data or time series data. So, the sample size may not be strong enough so, as a result. So, we like to pool both time and a cross sectional data. So, that we can have more sample size and then we can predict the kind of you know problem in a much better way and in the same times using panel data models we can solve some of the additional you know business issues which we cannot actually solve through either time series data or through cross sectional data.

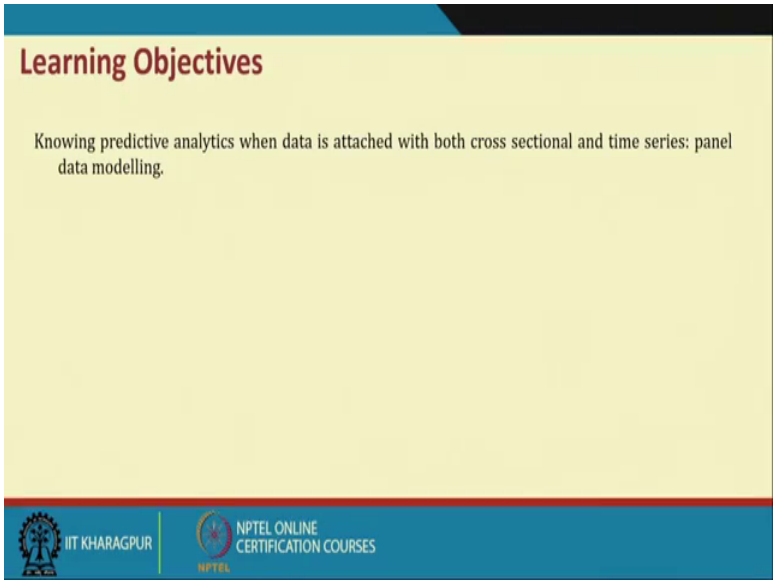
For instance, we have already discussed dummy modeling technique and a when we are actually dealing with the more number of you know business forms with more number of time periods. So, sometimes you know while a predicting dependent variable with independent variables. We can we can check the individual impact that is the business form impact on the relationship between dependent variable and independent variable, and against we can also check the particular you know time impact like in the last lectures which you have already discuss about the dummy modeling.

So, having the you know relationship between y and a x we can actually study the impact of you know genders we can study the impact of you know different at a you know religions we can study the impact of you know education and so, on. So, similarly we can also study the quarterly impact between y and a x and monthly impact on y and x.

So, now, these are all very valid kind of you know structure when we you know when we will connect with the dummy modeling and panel data modeling.

So in fact, panel data modeling is a kind of you knows extension to dummy modeling. So, the understanding of this particular you know structure requires the clear cut or you know null you know understanding of the regression simple regression structure and the a dummy structure. So, after knowing all these details then you can you know solve some of the business problem dealing with you know you know panel data structure. So, let us start with the kind of you know discussion.

(Refer Slide Time: 03:39)



So, here the idea is a to go for the predictive analytics by using both cross sectional data and time series data are together.

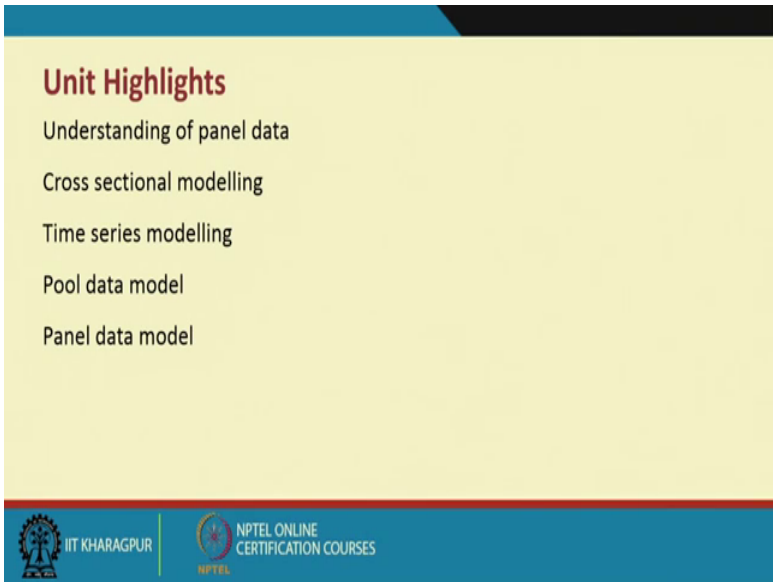
In the earlier problems we have already dealt some. So, many problems either through you know time series data only or through cross sectional data only even you know using the time series data. We can apply the dummy modeling and then we can check the structural break, we can check the you know a before and a after impact of a particular crisis or a wire something like that and against within a kind of you know cross sectional data, we can you know we can try to find out the individual impact.

So, now, you know we are just clubbing a both time series and cross sectional data ; that means, technically we are now in a position to study both time impact and cross sectional

impact together in the earlier case. When we have a time series data we can only check the time series impact for instance like you know monthly impact weekly impact quarterly impact now in the cross sectional side up in a data we can have a individual impact in organizational impact, now, since we are clubbing both cross sectional data and time series data.

So, we have a extra you know we have extra advantage to you know study the impact of you know individual business forms and at the same times we can go through you know individual time impact. So, how you have to you know below all these cases. So, we can you know better you know we can study through panel data modeling.

(Refer Slide Time: 05:25)



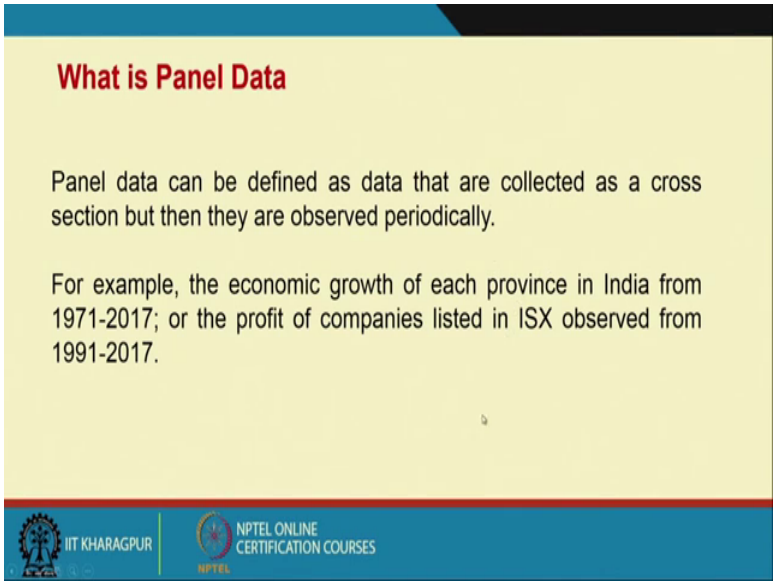
So, now, so, the fastened requirement of this particular you know lecture is to understand the panel data structure and then the same problem we can connect with the cross sectional modeling. Then with the time series modeling and we will be go through pool data structure and then finally, panel data structure.

So; that means, so, we have a kind of you know problem that ah; that means, to predict a dependent variable with respect to independent variables it may be a bivariate structure it may be a multivariate structure, but by the way. So, we have more number of time you know time series data and at the same times we have more number of you know cross sectional data so; that means, assuming that you know there are 10 company and 10

different time periods. So, a every company is having you know 10 different data points and similarly in every time periods we have actually 10 different company data.

So, then altogether you know there may be a sample size of you know 10 into 10 you know a is something called as you know 100 data points. So, within the particular structures we can check, what is the individual farming part; while studying the dependent the impact of independent variable on dependent variable and against, what is the time series impact; while studying the impact of you know independent variable to dependent variables so; that means, when actually we have actually big problems. So, your sample size should be also big enough otherwise you panel data handling is a kind of in a difficult task or a something like you know a very complex process.

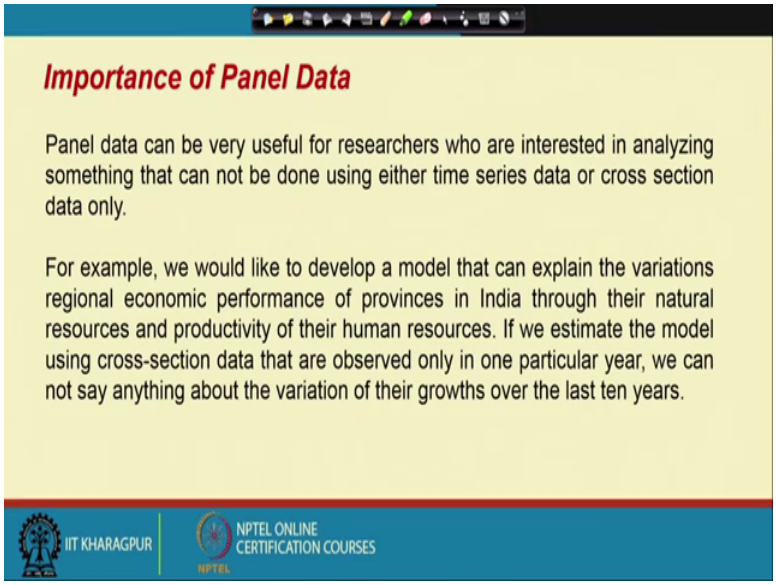
(Refer Slide Time: 07:06)



So, let us see how is the kind of you know structure altogether and so, what I told you earlier. So, that you know first of all you understand the panel data structure. So, panel data can be defined as the data that are collected as a cross sectional a, but then they are observed you know with you know kind of you know periodically maybe annually maybe weekly maybe quarterly maybe device whatever may be the structures, but it has both you know cross sectional kind of you know structure and time series kind of you know structure. For example, the economic growth of each parts of you know India from 1971 to 2017; or profit of you know various companies listed in a stock market and that to from 1991 to 2017.

So, the this is a classic example through which you can actually just check the kind of you know a panel data structures.

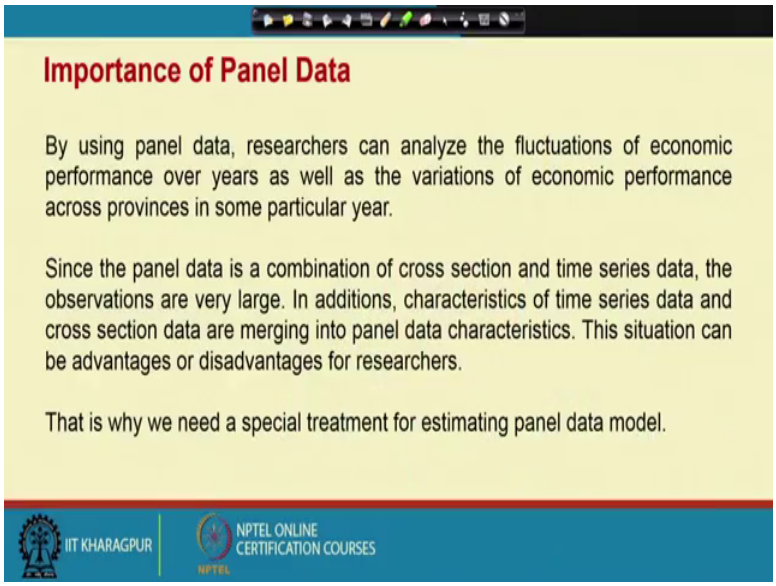
(Refer Slide Time: 08:03)



And then or something like you know individual impact of a particular country or you know individual impact of a particular you know time period and you know for examples we may like to develop a model that can explain the variations of in a regional economic performance you know in India the that is there you know with respect to there you know natural resources and productivity of their human resources.

So; that means, when we are in a dealing with the dealing with the all these you know problems. So, the panel data structure will help you lot you know to you know investigate the particular know process and predict.

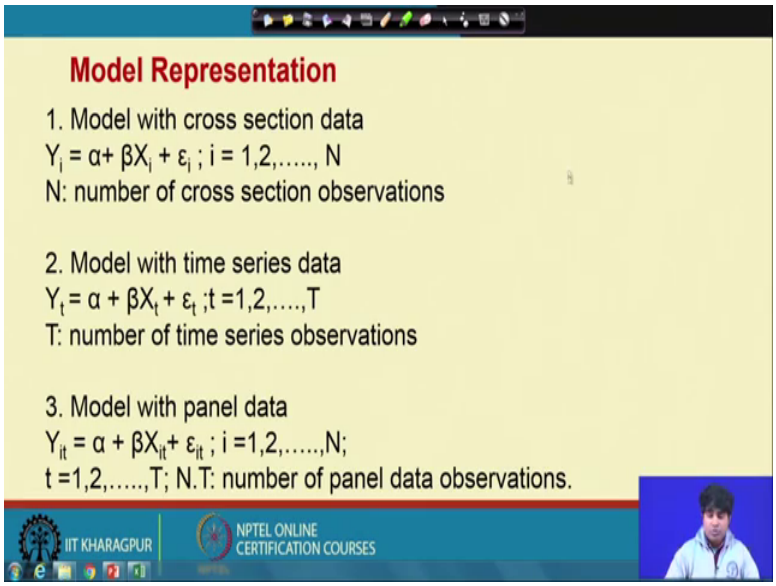
(Refer Slide Time: 08:43)



The particular you know requirement and similarly a various you know economic fluctuations and the kind of you know socks are the kind of you know crises the kind of you know up swings down swings. So, with the help of you know panel data. So, these are all things you can we can easily detect for instance you know while you know collecting the data with over the time. So, there may be actually somewhere may be up swings somewhere maybe down swigs.

So, the if you do not even study the individual impact then you are not in a position to check which year is very effective and which year is not. So, effective; so likewise you know panel data structure is very useful for you know a problems to investigate the kind of you know a business performance in a more attractive way.

(Refer Slide Time: 09:39)



So, this is how we are very you know you know keen to know this particular you know technique and we like to use this particular technique in some kind of you know business environment.

So, in the say in the in the simple structures, the usual framework of you know modeling is with respect to cross sectional then with respect to time and then with respect to panel so; that means, technically. So, here we have actually two variables Y and a X Y and a X and that will be connected with the i and t so; that means, technically I can vary and then t can vary and similarly Y and X can also vary.

So; that means, at a particular point of times you have to keeping i equal to 1, then t can vary in us up to a certain point of time. And similarly when i equal to 2, then against the data point can vary some extent, when i equal to 3 the data point can vary again like this. So, it will be actually it is a kind of you know chain and what will they do keeping t constant, then I allow I vary and then do the modeling and if that is the case then the particular model is called as you know cross sectional modeling. In this case at a particular point of time t will be constant then allow i vary with respect to Y and a X.

Then when we model this is called as you know cross sectional modeling against allow i to be constant and you know t should vary ; that means, for a particular company a. So, you know some observations are there time series observations are there that is with

respect to Y and X and when we do the modeling then this is called as you know model with time series.

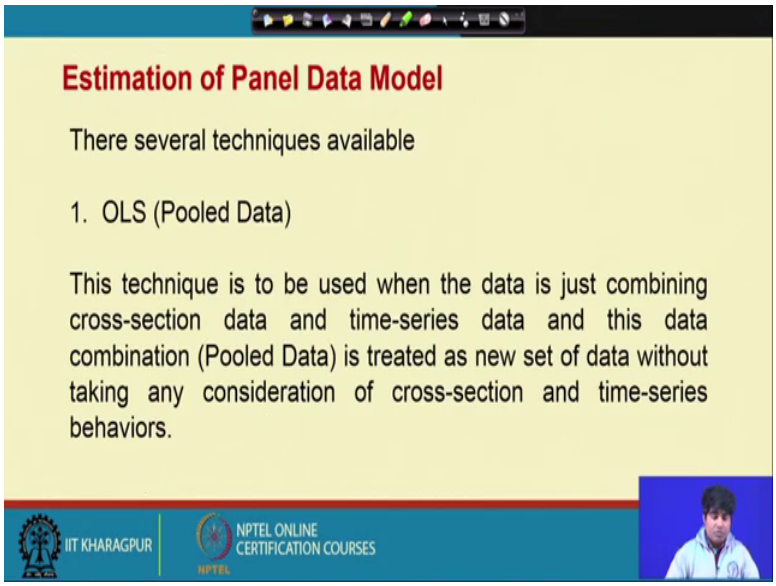
So, now we are doing here. So, we are just clubbing time with the cross and doing the same modeling with respect to Y and X. So, then it is actually with respect to you know panel data modeling. So, initially we you know we pooled the kind of you know structure, then this is called as you know pooled data. And then when we study the impact with the time series impact and cross sectional impact within the particular structure, then it becomes a kind of you know concept called as you know panel data modeling.

So, now in order to know much better about this particular you know process. So, let us see a how we can actually deal with such kind of you know problems right.

(Refer Slide Time: 12:04)

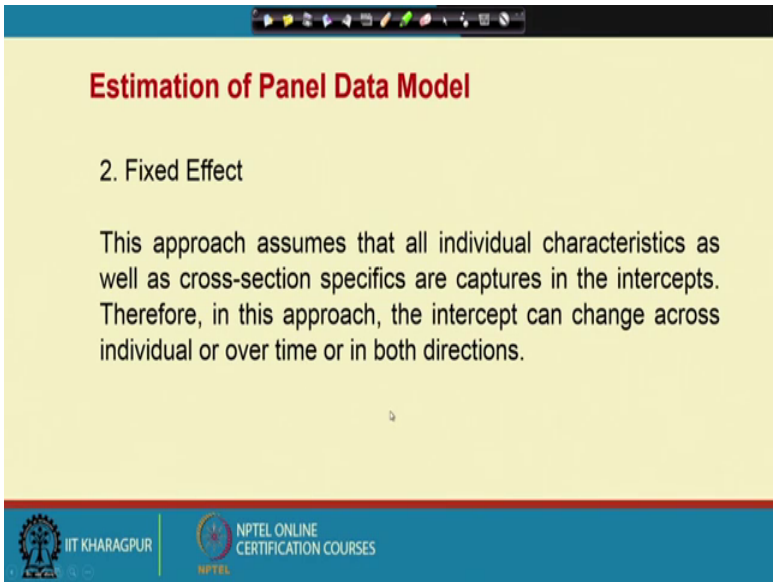
So, these are all things which we have already highlighted ok.

(Refer Slide Time: 12:09)



So, now, when we are dealing with actually time and cross. So, the first model you know we can actually in the you know deal with the with this particular you know series is called as you know pooled data and then we will go for the panel data.

(Refer Slide Time: 12:24)



So, now you know at the same problems we can actually simply go through cross sectional and then time series and finally, we can go with the pooled or you know panel. So, the these two are already discuss and a. So, what we like to you know do here. So, we are just pooling and do the modeling and against try to study the impact of you know

individual impact of you know cross sectional unit and individual impact of you know time series unit then it will becomes you know panel data structure.

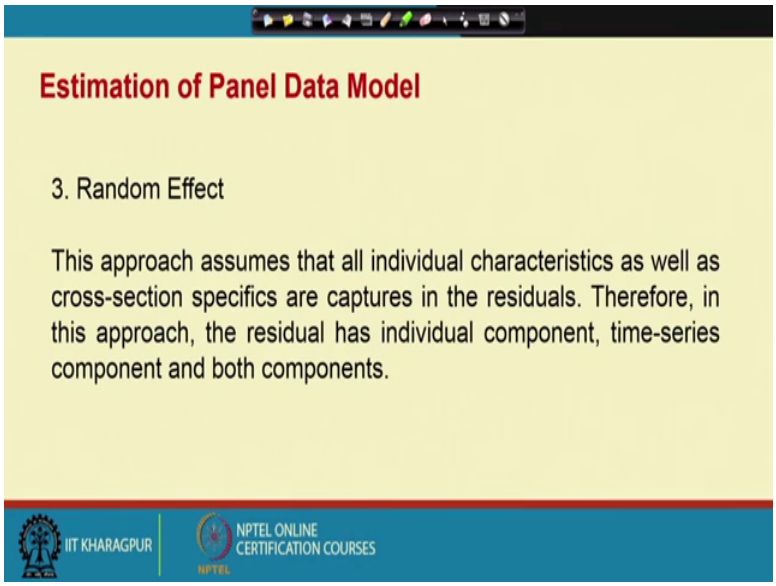
So; obviously, so we like to know now what is the pooled data structure and panel data structure and in what context we will use the pooled data analysis and what context we will use the panel data analysis. So, there you know within the panel data you know structure we have a couple of you know techniques and against we have to find out the best technique through which you through which the pro you know problem can be addressed more effectively like we have discuss earlier diagnostic you know kind of you know checks.

So, in the panel data structures we have a lots you know diagnostics kind of you know structure through which we have a you know variety of you know models through thorough which we can investigate the problem more effectively and come with a come out with a particular you know model which can be very useful for the kind of you know prediction and the kind of you know importance and in the case of you know well as pooled data.

The technique is to be use when the data is just combining the cross sectional data and the time series data and a then a do the analysis like you know simple regression structures. So, without any a you know giving the you know kind of you know importance to cross sectional hint and the time series hint.

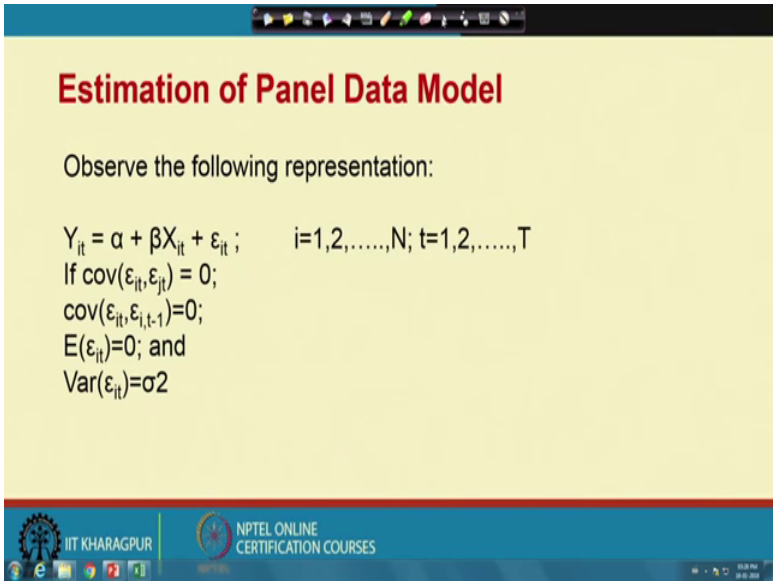
So, this is what actually pooled data structures; then when we will go for the panel data structures. So, in the first hand the panel data can be of you know two types one is called as the fixed effect models and the other one is called as a you know random effect model. In the fixed effect models; so, the impact will go to the intercept and a in the case of you know random effect the impact will be go to the error terms.

(Refer Slide Time: 14:40)



So, now what you know; you know what we will be like to know that; you know we first you know these prepare the structure and then we will like to highlight a how the impact will come to the intercept and a how the impact will go to the random term as a result we can understand what is the fixed effect model and what is the random effect model.

(Refer Slide Time: 15:01)



So, now in order to understand these; so, let us start with a simple panel data structure here Yit equal to alpha plus beta Xits. So, this is the model and assuming that you know like you know covariance of you know two error term should be equal to 0. So, that is a

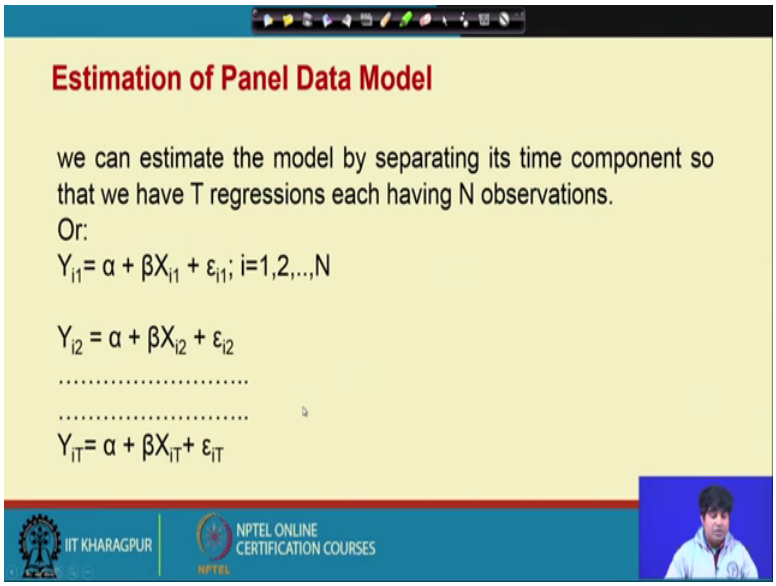
how what we have already discuss about the autocorrelation issue and then you know covariance of you know two error terms while i equal to j that is called as you know variance you know that is called a vari variance which we you know required which is the requirement we know how much capacity and a then against ah. So, the some of the error terms should be equal to 0.

So; that means, these are the things against you know we have to validate here in the case of panel data structure. So, the; that means, these are the common things a whether it is a time series or whether it is a cross sectional or whether it is a pooled or panel. So, these things are every times needs to be checked and you know you know clarified accordingly.

So, otherwise you know the model prediction would be getting affected and the management decision will be also accordingly getting affected in this context. So, we have a i is the cross sectional in indications and then t is the time series indication. So, usually the panel data structure can be two types it may be a balanced panels data it can be a unbalanced panel data in the balanced panel data, if i equal to phi and a every i having actually 10 data a 10 times rich data points and a in the case of you know balanced panel data.

So, not necessarily alt all units having actually balanced time series format or a with a every time frame every cross sectional units should not be also balanced. So, this is you have the difference between the balanced panel data and unbalanced panel data.

(Refer Slide Time: 17:00)



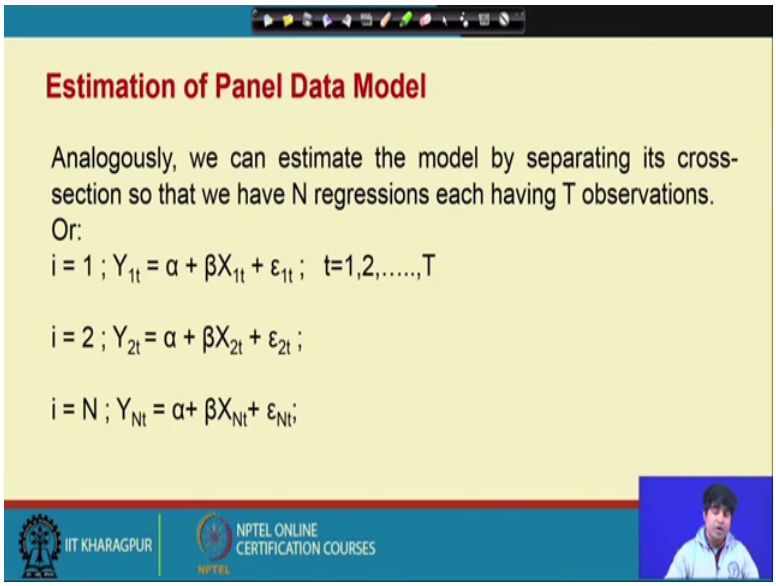
But by the way a this a structure will be as usual actually you know whether it is a balanced panel data or unbalanced panel data we like to understand and then we will work accordingly.

So, in the same you know in the similar way. So, the extension of the this model is like this. So, now, this is the first hand model then what we will do; we keep i constant and allow t will vary. So, here Yit and alpha plus beta Xit, when i is a constant t can vary from one to t and against when t is a constant I can vary from 1 to N.

So, we which we in that couple of slides we can highlight this issue in this case we you know Yi and beta Xi and here instead of t. So, we are putting t equal to 1. So, as a result the first equation will be Y Yi 1 equal to alpha beta Xi 1 and then error term similarly when I am when we are putting i a t equal to 2, then the model will be Yi 2 equal to alpha plus beta Xit then error second error term.

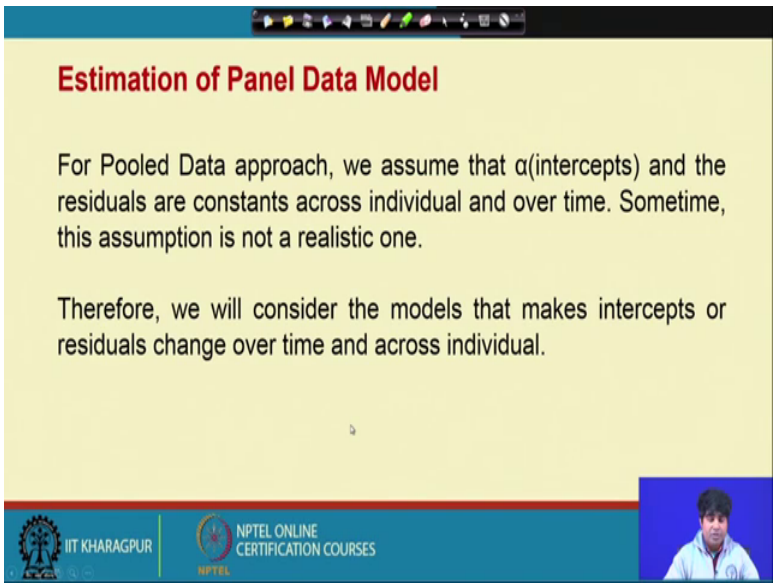
Similarly, when a it will be continue up to tth time period and again.

(Refer Slide Time: 18:13)



If you proceed for you know further where i equal to 1, then this becomes the model and i equal to 2, then this becomes the model so; that means, compare to previous 1 where i is a constant we are allowing t vary in this case we are actually putting t constant and allowing I equal to vary. So, when i equal to 1, then this will be the model when i equal to 2 this will be the model again it will be continue when i equal to N and this is the model so; that means, your a keep I constant allow t vary or keep you know keep t constant and allow i i will vary.

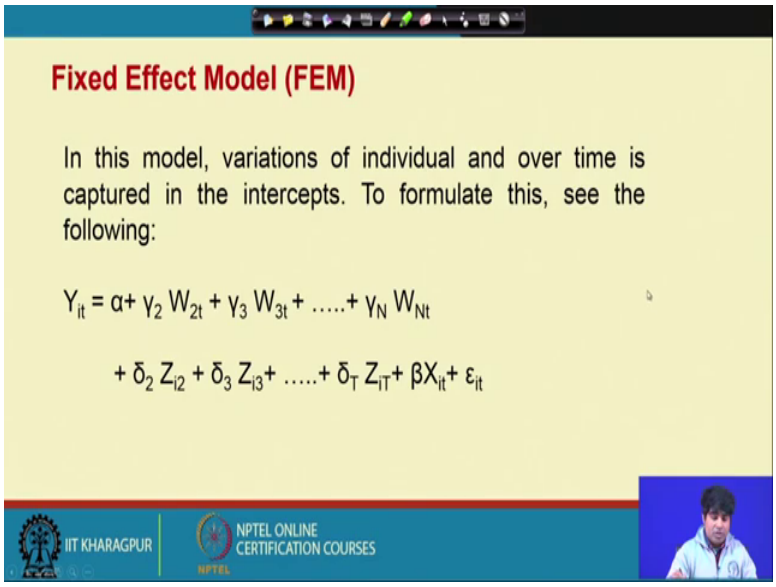
(Refer Slide Time: 18:54)



So, likewise, there will be a n number of you know equation which need to be actually integrate it is while doing the kind of you know panel data analysis. So, every times, the individual impact will be either going you know intercept or it will go to the residuals.

So, if it will go to the intercept then the particular structure will be represented as a you know kind of you know fixed effect model otherwise it will be you know decolorized as you know random effect model.

(Refer Slide Time: 18:29)



We will go in details about this. So, let us starts with in a fixed effect model in the fixed effect model and you see here the full model Yit equal to alpha and; that means, technically since it is actually cross sectional units and a transit unit structure is there. So, we can write the model like this.

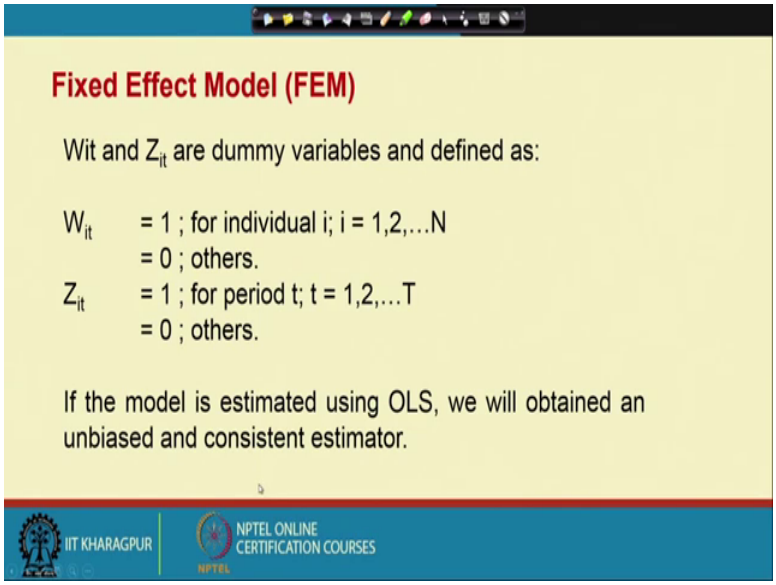
So, Yit equal to alpha plus summations delta I delta i di plus summations mu j you know t a tj and plus summatio plus summations beta Xi beta Xit plus uit so; that means, this is the part called as you know cross sectional domain and this is the part called as you know time series domain.

So, now accordingly we have actually classified here. So, now, this is actually dependent variable and this is the original intercepts and then a these are the Y 2, W 2, Y 3 W 3 these are all called as you know cross sectional domain and then these are a delta 2 Z 2; delta 3 Z 3 these are all called as you know time series domain.

So, now the, with the help of this data we can create actually dummy a dummy structure first the we would like to check actually a; what is the individual impact; you know you know individual firm impact and individual time impact. In fact, the way have already a solve a problem. In the case of you know in the dummy independent modeling while you know studying the quarterly effect between you know sales and the expenditure.

So, here also in this you know similar fashions you have to highlight and you can actually address this particular you know problem.

(Refer Slide Time: 21:26)



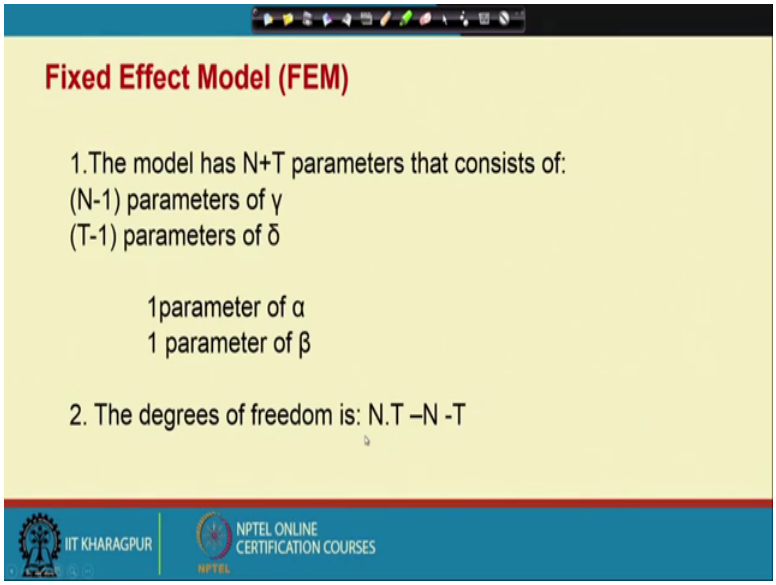
And then similarly if you move for those then you know the representation will be every time. So, you know when a particular company will be targeted that will be represented by 1; otherwise it will be 0.

For instance; so, when we are saying that you know W 2 equal to 1, then rest of the you know units will be by default to 0. So, the impact will be alpha plus you know delta and then a rest of the impact will be added subsequently. So, similarly in the case of you know time. So, when a particular period will be targeted then dummy will be represented as a 1 then in other case it will be represented as a 0.

So, likewise we have actually big you know multiple or multivariate to regression model through which we can study the individual firm impact and individual time impact to

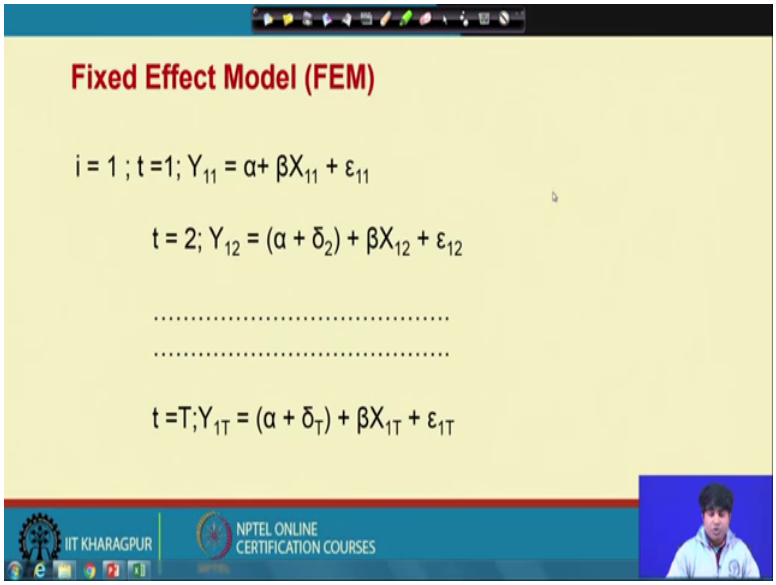
while you know investigating the impact of you know independent variable or dependent variable.

(Refer Slide Time: 22:22)



So, likewise you know. So, the process will be a extended and by the way; since it is actually I will start from 1 to N-T will starts from 1 to T then the total sample size will be finally, N-T a and then the degree of freedom will be N-T minus N minus T and this will be with the help of you know cross sectional you know integration in the kind of you know time series integrations.

(Refer Slide Time: 22:54)

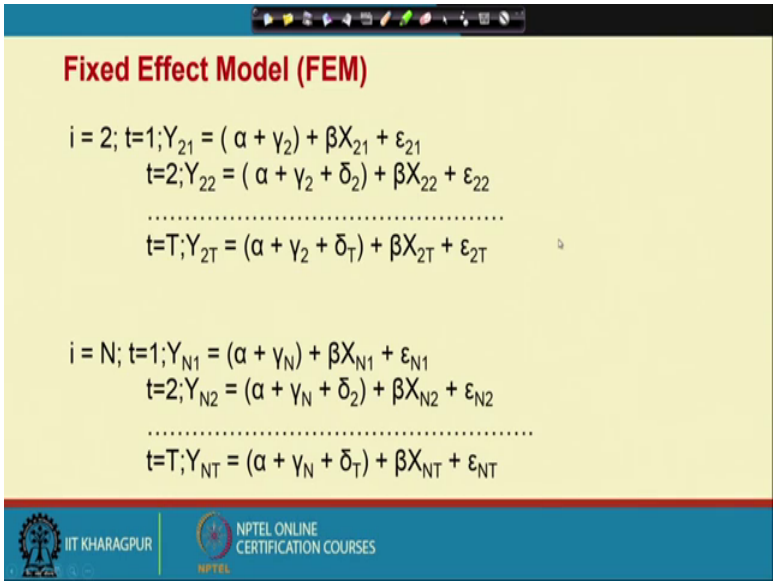


So, then as usual we will said the model and go for the kind of you know estimations. So, more elaborative where here we allow it simultaneously if they will be in our allow to vary, then when i equal to 1 t equal to 1, then the model will be represented by Y 1 once and this is this is what actually when i equal to 1 t equal to 1, then this is the model when i equal to 1 t equal to 2, then Y 1, 2 equal to this much of. So, this is actually. So, we start with dummy 2. So, so that you know when other dummies are 0. So, this will take care the first firm impact.

So, this will be the second firm impact similarly this will be go for you know third firm impact. So, here since tj you know varying. So, now, it is the first time period impact and this is actually second time period impact and this is third time period impact and that two for a cross sectional unit one. So, likewise you can go for you know allow time constant then I can vary then individual firm impacts as a second firm impact like this.

So, it is a kind of you know continuous process through which you can understand the particular problem and then we can you know address the it kind of you know in a issues in a more attractive way.

(Refer Slide Time: 24:10)

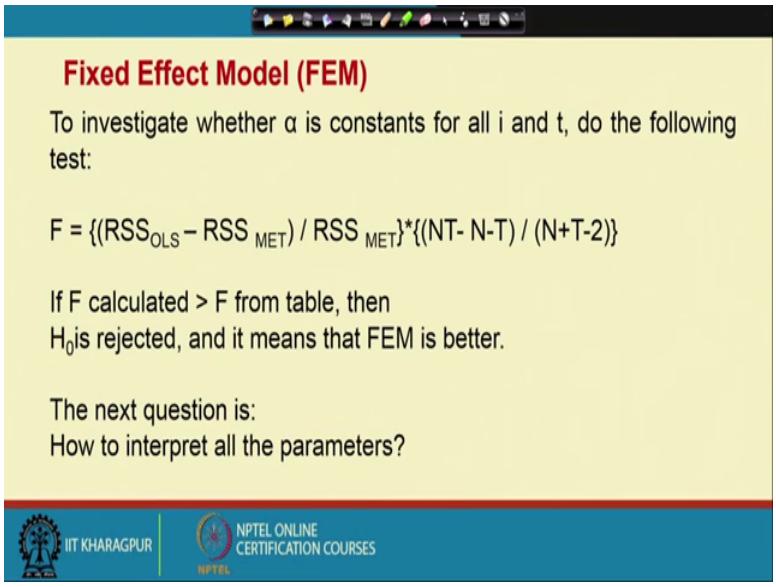


So, here when I putting i equal to 2, then again the model will be extend simultaneously. So, i equal to 2, then alpha, alpha then you know the second variable impact then accordingly a it will extend a subsequently.

So, these are the kind of you know structure. So, this is i equal to 2 and t equal to 1. So, this is the first model when i equal to 2 t equal to 2, then this will be the a subsequent models. So, likewise so, this model can be extended these are all you know you know mathematics behind this particular you know panel data structure, but ultimately if you understand the particular you know data and you know represent in the form of a kind of you know dummy structure then you know estimations and these a in a design of the model will not be a you know actually a problematic.

So, like you know dummy structure it will be just you know you know integrated and then a estimated as per the particular you know requirement and the particular you know objectives. So, now, a here what will we do. So, we like to you know proceed again and then a against.

(Refer Slide Time: 25:23)

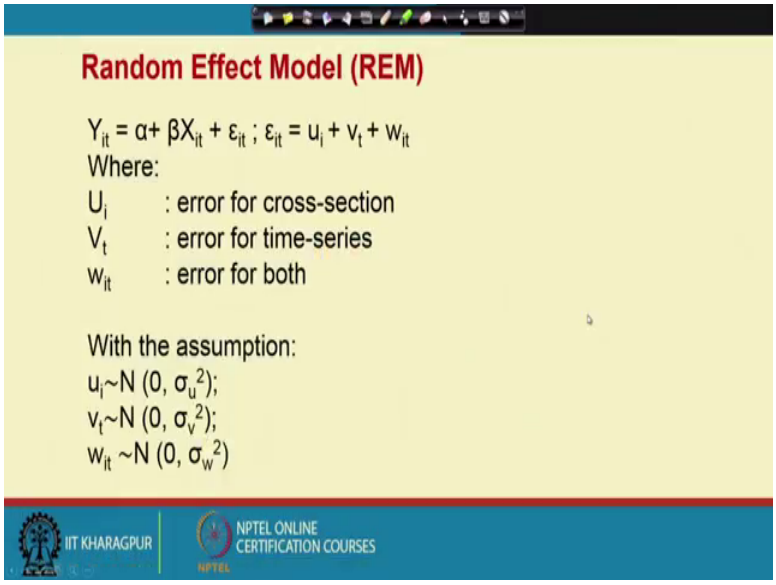


When will we a go for you know kind of your time impact. So, the particular you know structurally extend accordingly.

So, now what will we do we like to check whether you know we like to follow the kind of you know pooled data models or we like to follow the fixed effect models. So, the particular extension is nothing, but actually fixed effect models so, but when we will go for you know pooled data model you are we can just pooled the data and then investigate the investigate these to between y and x like this year.

So, here just y and the x then the rest of the components individual impart will be removed, but in the cased of you know fixed effect model. So, these are you know will be incorporated. So, now, the question is whether you know pooled modern will be effective or you know random effect model will be effective. So, in that context. So, we can go for you know test and then the f test will give you the indication whether the particular you know model is a effective for this particular you know problems and then accordingly we have to choose whether we can go ahead with the pooled data modeling or will we go with the panel data modeling that to fixed effect model and against.

(Refer Slide Time: 26:38)



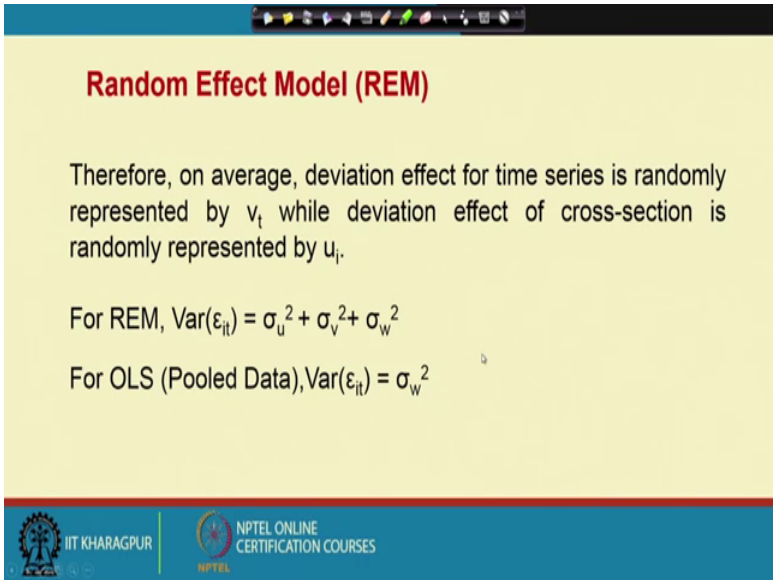
In the case of you know random effect model. So, the typical structure will be like this and here we are not allowing the impact to go to the intercept will be allow the impact will be go to the error terms as a result. So, the total error term will be compose of or you know Ui, Vt and then finally, Wit. So, Ui deals with you know cross-sectional the error term and dealing with the cross sectional you know effect then Vt a represents the a you know kind of an error term dealing with the times effect then Wit it is a kind of you know while pooling both time and cross.

Accordingly, it will be exactly followed by homoscedasticity and accordingly this should be unit variance with respect to u, v and w and the cross correlation coefficients means the autocorrelation coefficient should be equal to 0. So, this is the as usual again a requirement, whether it is actually fixed effect model or you know random effect models.

So, now, we get to know now three types of you know models pool models a fixed effect models and random effect model and then in the first hand we like to choose which model is the best first choice is whether actually pool or you know panel against in the panel whether choose actually fixed effect model or random effect model.

So, there is a procedure through which actually we have to check and then finally, freeze the particular model as for the you know requirement of you know predictional and this model will be free from all kinds of you know error and this will give you as per the a structure of the you know panel data modeling.

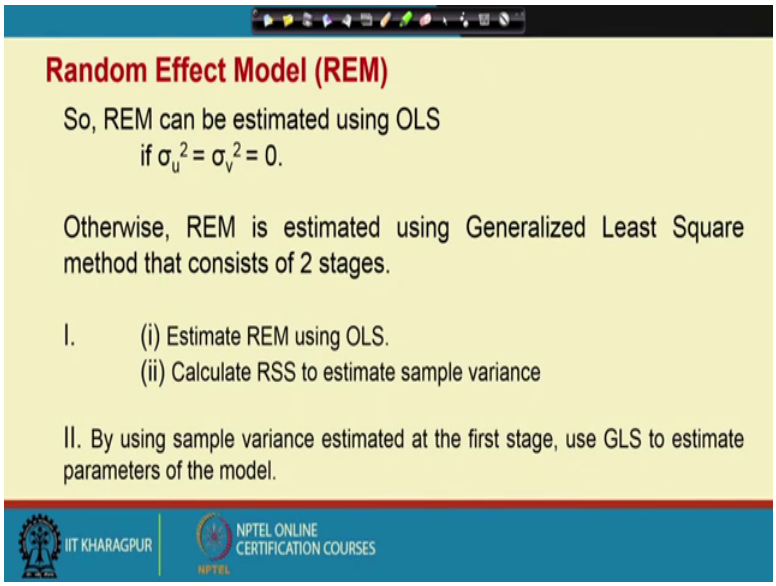
(Refer Slide Time: 28:23)



So, now you know going ahead you know about this particular issue. So, the random effect a model case. So, the total variance error variance will be a sigma square u sigma square b and sigma square w.

So; that means, it is the error variance with respect to cross sectional impact with an error begins with respect to time impact and then error begins with respect to pooled impact.

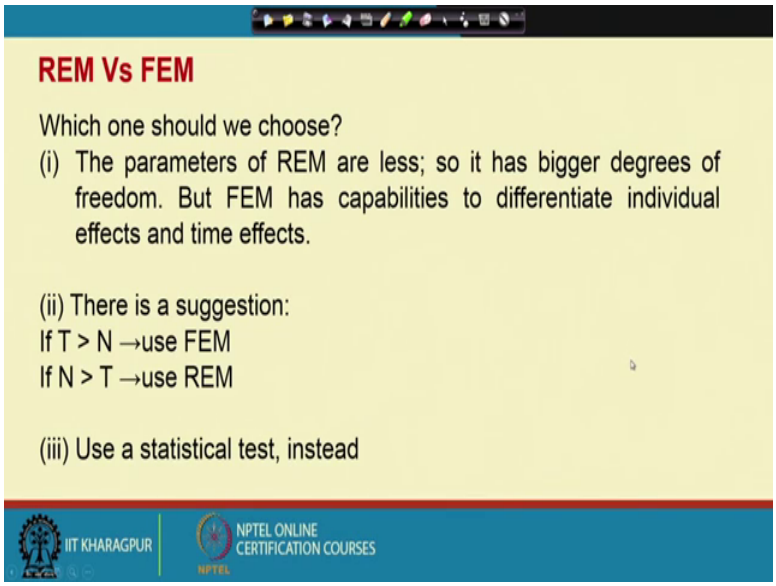
(Refer Slide Time: 28:48)



So as a result, it we have a three different you know error variance, then you know now the choice between actually random effect models or you know fixed effect model. So, we have a some kind of you know test. So, here we are assuming that you know in OLS mechanism sigma square u and you know sigma square v and then we like to actually a estimate the process you know as a kind of you know structure called as you know generalized GLS square methods because you know in that case of you know panel data modeling particularly not a random effect model.

So, the usual OLS structure will not work and then will go for you know generalizally square mechanism. So, what will we do first we take the problem and investigate with the pooled data and again will go ahead with a the fixed effect model and then come with you know a random effect model.

(Refer Slide Time: 29:44)

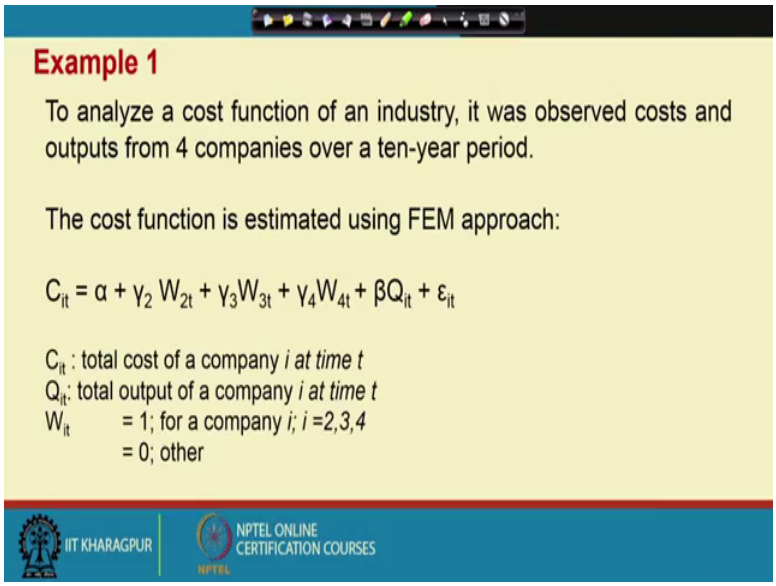


So, now, between the random effect models and fixed effect models there are lots of criteria which one is the best, but in the theoretical grounds it depends upon the choice you know depending upon the a size of you know T and N. So, T represents the a time length and N a represents the cross sectional length.

So, we time length exceeds you know a cross sectional length then a fixed effect model will be more effective to study the individual firm impact and when N greater than to T and in that contest you know theoretically random effect model will be more effective to justify the a kind of in model and then go ahead with the kind of you know prediction structure, but by the way. So, we have actually statistical structure through which we can check the random effect model and fixed effect model we have lots of you know fitness indicators and the kind of you know test like (Refer Time: 30:45) test a through which we can actually in a position to justify which particular model will be finally, taken into consideration for this you know prediction.

So; that means, it depends upon a particular problem the kind of you know size of you know data the kind of you know variables involved in the process and then a model with the help up in a model results and you know fitness indicators ray you know indications we have to finally, freeze which one is the good fit for the kind of you know requirement.

(Refer Slide Time: 31:17)



So, now in order to analyze the particular you know structure. So, we start with a simple example and here the example will be like this we have a we have here four companies and the time period which you have taken you know 10 ten different you know early structure.

And the problem which you like to address here is a cost versus output so; that means, we like to know what is the relationship between a cost and the output with the help of you know 4 different companies and 10 different time periods. So, technically it is a 4 companies information with 10 different time periods of course, 10 different time period is a small one, but we can have actually individual company analysis.

So; that means, for individual company analysis. So, using 10 different time periods you can just connect with the C and Q of course, it is a small data set even if you know four companies having instead of 10 if you have a 100 you know time and data points then; obviously, by default you have actually 4 different models for 4 different companies and that too using the time series data and then to investigate the link between output to cost, but by the way and now in the case of you know panel data modeling. So, we can just pool all the data the same impact we can actually observe and that to the relationship between a output and cost.

Now, since it is a industry problem, because we are actually clubbing all the company and then within the pool of you know all companies and we like to investigate the

relationship between output and cost and within the particular you know pooling we like to check which particular you know for company is more effective so; that means, you know the idea is to justify here you know which particular form is more efficient ah. So, far as you know cost is a concern.

So, technically the objective will be. So, if we fix actually output at the particular point of time or you know given a particular level of output then you know we like to check which particular you know company will be more efficient and that too with respect to you know less cost. So, panel data structure you know and the panel data structure or panel data modeling will be very effective to be here and to address this a, this particular you know in our problems right. So, you know how we have to actually you know deal with this problem and that will be discussed here after the break here.

Thank you very much have a nice time.