Engineering Econometrics
Prof. Rudra P. Pradhan
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

Lecture – 54
Panel Data Modelling (Contd.)

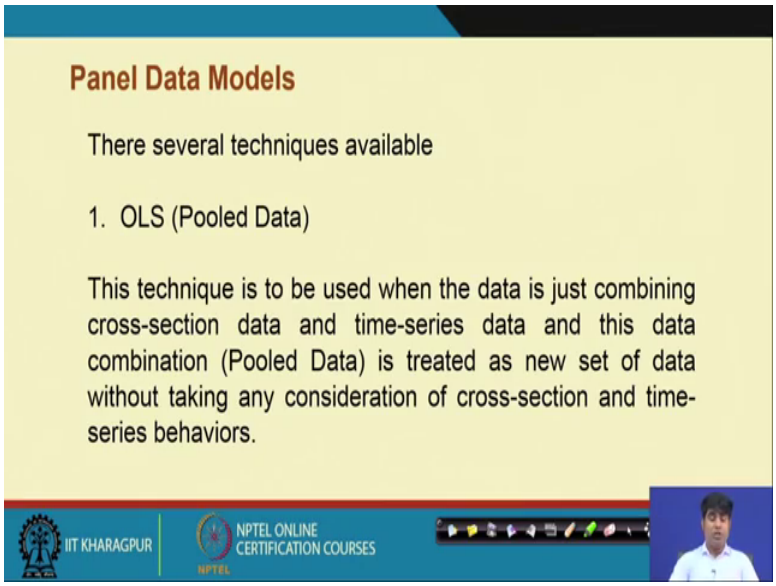
Hello everybody, this is Rudra Pradhan here. Welcome to Engineering Econometrics, today we will continue with Panel Data Models. In the last lecture we have discussed this panel data model that too understand the concept of you know panel data and the importance of the panel data to brief about this particular you know setups.

So, we have discussed the time series type data, cross sectional type data and then we have a data structure called in a pooled data and panel data by allowing time to vary and cross sectional to vary simultaneously. Because in the time series data only time factor will be allowed to vary well, cross sectional is the constant and in the cross sectional type data we allow cross sectional we need to vary while time factor will be remain constant.

As a result so, we may not get you know more and more variations in the samples while addressing some of the engineering problems, while linking dependent variable with the independent variables. So, now, the beauty of the pooled data and panel data is that. So, we are bringing lots of variation to the samples while addressing the same engineering problems and by the way the process will remove some of the bias in the system. So, that is why understanding panel data and use panel data for any kind of you know engineering predictions and engineering forecasting, my opinion is you know it is very you know excellent.

So, let us see how is the kind of you know structure about the panel data models. After knowing the kind of you know data structure about the panel data we like to move what are the models available through which we can use panel data and that too address some of the engineering problems as per the decision making is concerned ok. So, let us see how is this kind of you know models.

(Refer Slide Time: 02:45)



So, starting with actually some of the models we can called as you know pooled data and panel data. So, when some estimation will be start with respect to pooled data, then it will look like you know either time series or cross sectional, only advantage is actually having more sample points. And then in that case simple values can be apply; that means, technically while a clubbing actually pool and pool and time.

So, we have a different dataset. So, the same dataset you can apply OLS and then study the impact and that particular structure is called as a pooled data analysis and OLS can be directly used there without any issue and you know biasness.

But, when you when we are interested to know the time series impact and cross sectional impact to this pooled data, then we have to bring actually panel data models. So, technically with respect to pooling time series data and cross sectional data by default we have for different models all together which we would like to highlight here.

First one is the pooled data model and the second one is the panel data model; here we like to cover 3 things because they are very important and that too they are very common and you know very easy to understand and address the problem more appropriately and more accurately.

So, that is the concept of fixed effect model, the second one is the random effect model and the third one is the generalized method of movements.

So, let us start with the first OLS structure that too the concept called as you know pooled data and the other 3 are technically called as you know panel data. But, if you go to the axial so, the same data structure will be read as a pooled data; same data structure can read as a you know panel data. Just we have to bring some kind of you know extra specification or extra features to this you know pooled data, then by default the outlook and the kind of you know understanding is completely different. And there is a high chance that the expected outcome will be also different.

So, ultimately at the end of the day we can check the things both you know as a you know structure of pooled data and panel data. And finally, we have to pickup which is having actually you know good ones and can produce good results. So, we have no issue to use the specification of pooled data or panel data.

But, you know the problem and the kind of you know output by default when I give you some kind of you know indications where to go with that too stick with you know pooled data or we have to stick with you know panel data that too with fixed effect model, this random effect models and the GMM.

So, let us start with you know OLS first that too pooled data structure. This technique is to be used when the data is just combining cross sectional data with time series data. And this data combination is called as you know pooled data and it is just stated as you know new set of data without taking any consideration of cross sectional variation and time series variations.

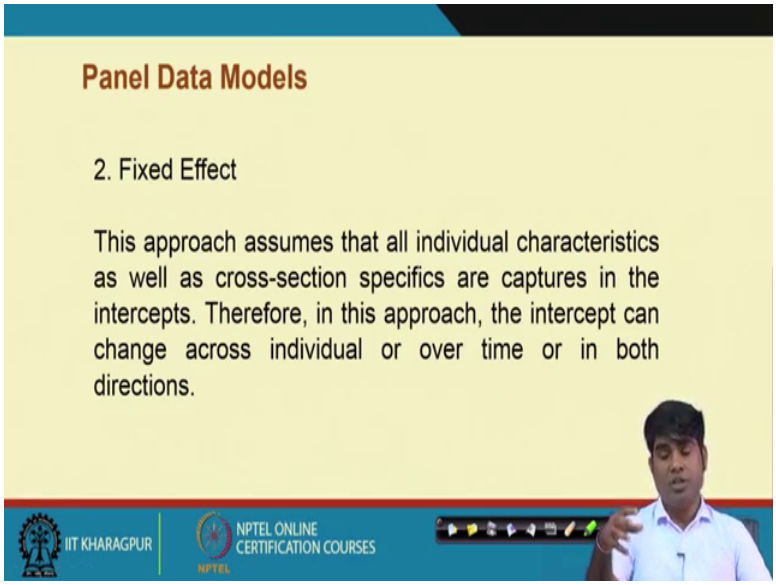
That means, cross sectional behaviour or cross sectional important time series behaviour and time series importance that is simply you know that means, as usual you know think here like you know cross sectional type and time series type. But, technically when we are in linking y and time we need some you know sample points, data points with respect to Y and X and then we maybe process it.

So, need some data points to linking Y and X to analyse Y and X. So, we have no botheration's whether it is a time series type, cross sectional type and pooled data type. We need some observation to test that is what the simple understanding about the pooled data.

But in reality, but in some situation or scenario, we need to know how is the time series you know impact or you know cross sectional impact within the pooled data setup. That is little bit in depth kind of you know study and in depth kind of you know requirement. Of course, the outlook and the kind of you know policy implicational for you know using the panel data model will be different all together. But, still we cannot you know just blindly pick panel data without you know understanding pooled data and without comparing the pooled data.

So, we have different kind of you know tests and this test will tell you whether to stick in OLS type of you know structure that too pooled data structure or whether to stick to panel data structure and the kind of you know panel data models.

(Refer Slide Time: 08:14)



So, over the times we like to change both the aspect and then finally, you have to fix as per the best and as per the requirement of some of the engineering problems.

So, as the second model; that means, in the second basket that the pooled data model here we like to cover fixed effect, random effect and GMM. So, the first one is the fixed effect this approach assumes that all individual features as well as the cross sectional specifications are capture in the intercept only. So; that means, that is the concept which we use called as you know delta i.

So, for instance, suppose there are you know 5 c unit so; that means, technically we can create a new dummy variables. So, we can start with you know 0 to 4 representing 0 you know 1 is indication of you know 1 cross sectional unit, two is the cross second cross sectional unit, 3 is the third cross sectional unit and fourth is the fourth cross sectional unit.

So, when 1, 2, 3, 4 is not there by default by represented as you know fifth cross sectional unit. So, likewise you know you have to classify the data; that means, by panel that has a beauty actually means that is one extra advantage that you know by default you can bring classification some kind of you know classification. That means, same pooled data can be can give you know different levels of you know classification provided he must have you know more and more observations with more professional units and more time series units.

So, keeping you know I constant you check you know how many I variation is there. So, you know if you arrange data in ascending to descending order, allowing I will be where we keeping other things remain constant then you will find you know different pools of you know result. In same way you can apply to the time series to vary and keeping other things remain constant again you will find different pool of you know data.

So; that means, with you know master data sheet you can find some of the sub clusters and that sub clusters can be used for that too analyse this same engineering problem. And on the top of that we can go as individual kind of you know you know estimation then you can go for you know pool kind of you know estimation and then we can go for you know panel kind of you know information.

So; that means, when you have more cross sectional type of you know data and more time series you know data within the X and Y. So, we have lots of lots and lots flexibility to address the engineering problem as per the requirement. And then the validations and the effectiveness of the model will be much you know higher and more effective as per the particular you know requirement. That is why we need to actually you know highlight this issue and give more importance to analyse this problem.

So, here so, individual cross sectional units need to be need to be a targeted and that will be that will be the vary you know variable factor through which you can you know a linking Y and X. So; that means, a all togethers you have Y information and the X

information and ultimately you have to create a dummy variable that too cross sectional type a situation. And even if it can be apply to time series also that is called as a dummy effect that to dummy with cross sectional effect, dummy with time series effect.

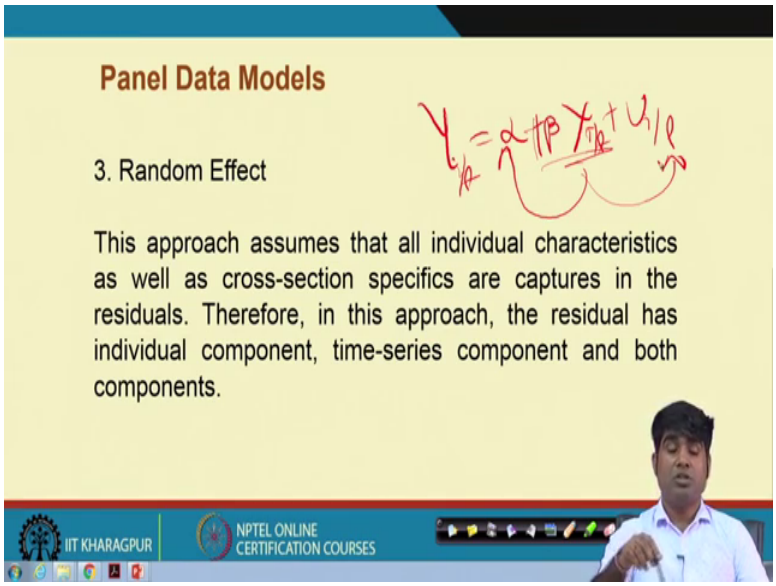
For instance, instead of you know delta you can put you know delta t or mu t. So, here corresponding to the particular data you know that too time series so, we have to give the signal. For instance, if the variation is actually with respect to 2 different time periods with you different cross sectional unit.

So, you can for the know previous data set which we have discussed that too with the availability of you know the year 1990 and 1993. So, what we can have actually we can create a dummy and if that particular data is for 90 then you can put simply 0, if not then you can put you know 1. So, putting 0, 1 so, that by default will give you know different kind of you know specification.

So, the new classification will be with respect to 0 and 1. Again the particular dummy can be again separated here you know d will be 0 and the other d will be simply one ok. So, you know if you have a 0, 1 only then my summation is that you know you can use only you know 1 dummy. But, if you have actually more kind of you know such dummy is starting with you know let us say 0, 1, 2, 3, 4, like this we can have a different dummy.

So, we can know the individual you know time series impact and individual cross sectional impact while addressing the link between the dependent variable and independent variables.

(Refer Slide Time: 13:49)



So, likewise so, we have another concept called as you know random effect model and in this approach we assume that all individual features as well as the cross sectional specifications are captured in the residuals instead of you know mean; that means, technically. So, we are you know technically we are actually linking Y and X and that too allowing intercept and then coefficient for X and then the error term.

So, ultimately so, the issue is actually with i or t and here with i or t and again with the i or t ok, either i or t or i, t to a simultaneously. So, then now in between if you allow i 2 you know the effect or t to be effect or both i and t to be affect or ultimately that that is actually numerically coded and quantified. And we know when you know put some you know some items 0 and the other item will allow to vary for instance 0, 1.

So, if it is 0, then you know the impact will go to the intercept and again if it is 1, then the impact will go to the intercept because there is no variable there. So, it is the dummy which classifies the data into 2 different cluster all together.

So, that means technically so, the time effect and cross sectional effect either it will come to the intercepts side either it come to the intercept or it will go to the error term. If that particular impact will be studied through intercept that is what the technical called as you know fixed effect models. And again that particular impact will go to the error term keeping intercept constant then the particular model is called as you know random effect model.

So; that means, when you like to capture the time series impact and cross sectional impact there are only 2 is which is you can actually highlight. Because, ultimately the impact either will go to the intercept or the impact will go to you know error term.

So, if you allow the impact will go to the intercept not to the actually error term then that particular structure is called as you know fixed effect model. So, ultimately in this case the a particular alpha value will go up or you know down depending upon the kind of you know coefficients value ok. So, whether the coefficients are coming minus or plus because ultimately it is a delta i. So, i will vary and corresponding to i then delta is the coefficient actually the.

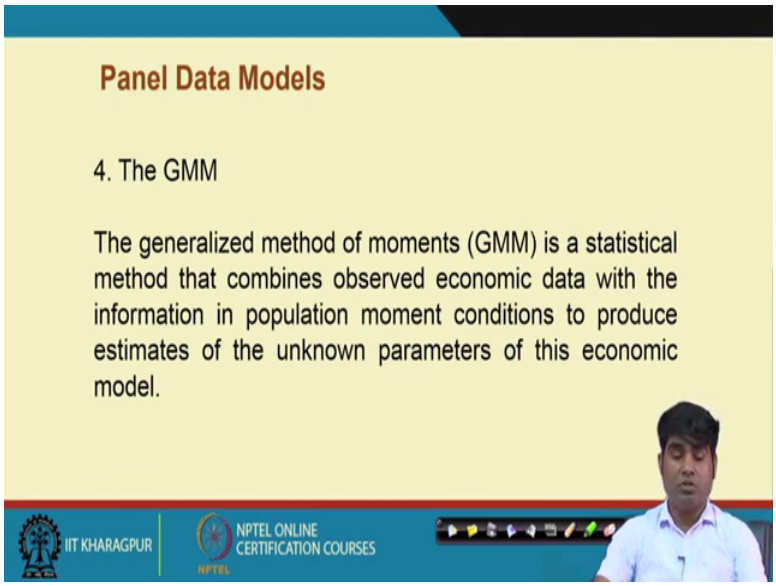
So, the coefficient value if the coefficient value is not coming 0 then technically either it will have a minus you know coefficient or it will have a plus coefficient. If it is plus coefficient then the alpha value will go up and; that means, intercept you know have more weightage. And if th coefficient of delta is coming negative then the alpha will value will be down. So, that is how the difference we will define.

Similarly, the case of you know t so, if you put you know mu t. So, then ultimately, after estimation mu will be the differentiating factors. So, either equal to 0 which is our null hypothesis to be tested, if not 0 then either mu coefficient will be positive one or negative one. If it is positive one and then again the impact will go up and if it is negative one then the impact will go down that is what the kind of structure.

So, in any case either it will be you know you know just through you know intercept or it will be just through you know error terms. So, technically so, the issue of you know fixed effect model and the random effect model. The fixed effect model is the impact of cross sectional and time series units to the intercept and the impact of cross sectional units to units and the time series units to error term is nothing, but called as you know random effect model.

So, that is what the dealing and in addition to that. So, what we have actually the third ones.

(Refer Slide Time: 18:07)



Under the panel data basket is called as you know GMM that is you know little bit more you know complicated and more powerful compared to a simple pooled data, then fixed effect, then the random effect.

So, usually GMM is statistical method that combines observed economic and engineering data with the information in population movement conditioning to produce estimate of the unknown parameters of this you know econometric models or any kind of you know engineering models as per the particular you know estimation process.

So, that is means technically this is another form of you know panel data model, where we will you know address same engineering problems that too linkage between dependent variable and independent variable. But, in a kind of you know different you know flavour or you know different outlook.

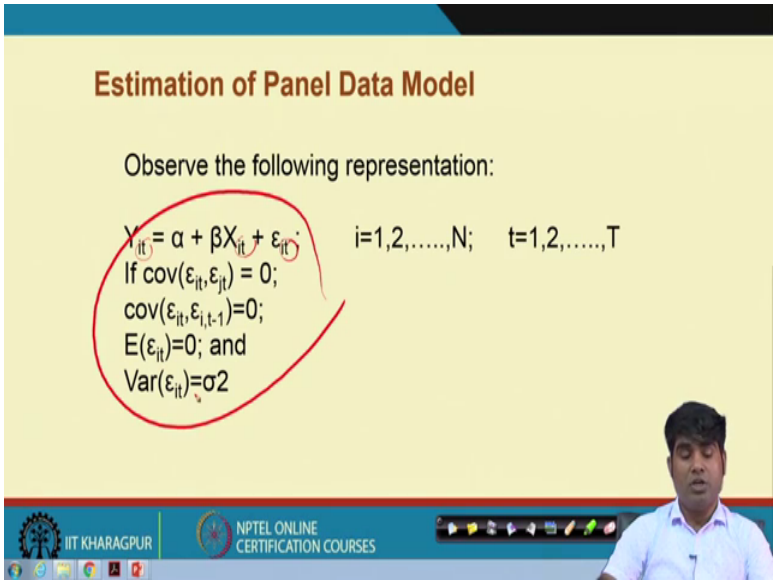
In fact, there is a concept called as you know dynamic panel data modelling where we sometimes you know allow the model to go through you know endogeneity the impact; that means, you know we can bring actually a lag variables a in the in the panel data set up.

So in fact, we use i, t so, now, it can go for you know i t minus j. So, that is the kind of you know specification which you can bring and these are all sometimes required where using panel data and that took the combination of both time series data and you know

cross sectional data. Because, of the cases the endogeneity which is stronger while you know linking time series data with you know cross sectional data.

And the particular flavour is different all together; that means, technically you can go for you know simple panel data you know set up or you can go for you know dynamic panel data set up. And GMM is a kind of you know concept where you know we can actually bring dynamic a concept in the panel data approach. which is actually not you know really highlighted in the case if you know simple fixed effect models and the it kind of you know random effect model.

(Refer Slide Time: 20:36)



Now, to understand the kind of you know estimation process. So, the particular you know a model is you know model is like this, where we have the link the Y equal to X. And the specification is it here and the other component also i t and if it is actually only i or only t then you know the particular you know covariance will be simply you know you u i u j or u t u t minus j like this.

And again here since i and t will together. So, we have a different kind of you know specification and first of all the covariance of these 2 error term should be equal to 0 that is as per the OLS requirement and as per the discussion you know in the earlier you know different models.

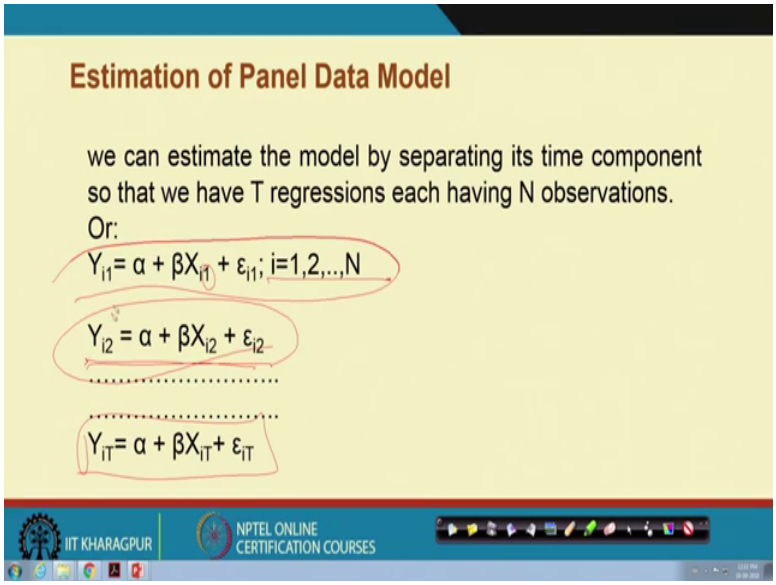
And then covariance of you know you know term with 2 different time periods keeping i remain constant. And in the previous case giving t remain constant to cross sectional unit variation should be actually not there. Because, of you know some of the requirement of you know estimation process that too the OLS requirement. Then some of the term should be equal to 0 and by default variance of the term should be you know you know approaching unit variance or you know homogeneous variance.

The you know these are all you know whatever we have actually earlier you know highlighted or assumed in the case of you know simple time series modelling and the kind of you know cross sectional modelling.

So, now what is happening here? So, I will vary from 1 to N and t will vary 1 to T. So, we ; that means, the first in the requirement of you know panel data is a almost a more number of i and more number of t then the game will be very and game will be very interesting. And if a means a is a if I will be more and more and t will be more and more. So, game will be more and more interesting. And then we will find we will find clue to use different kind of you know models to like similar kind of you know engineering problems

So, on the top of this so, the model can be actually analyse in a different way see heres. So, this is what the general models Y i t equal to alpha plus beta X i t plus error term.

(Refer Slide Time: 23:07)



And then see how we can actually generalize or you know analyse the case means we are just expanding the kind of you know situation. it is starting with Y i t and alpha plus beta X i t and error term. So, what will you do here? So, keeping i remain constant and allow t 2 vary. So, t will be vary starting with you know t equal to 1, t equal to 2, t equal to 3 and so, on.

So, when i remain constant and allow t vary. So, then first t equal to 1 and in that case the model will the just you know transfer into like this. So, here i will be vary and t actually fixed heres that too equal to 1. Agains in the second models where we are putting t equal to 2 and then i remain constant that is how the model is all about similarly it will continue a.

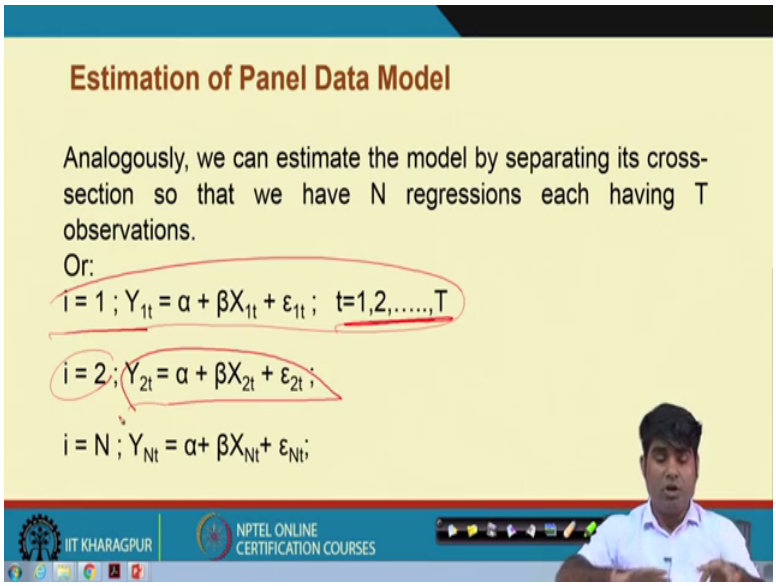
So, up to transpose tth time period so, in the last step. So, this is what the made on so; that means, every time we can have a different estimation and different model outputs keeping you know i remain constant. So; that means, let us say a company you know individual company and we have actually a time series data for a particular period and again we have time series data you know the time period that too 2 and then you can run the model separately ok.

So, of course, we have the option to pool and again go for you know single estimated output. But, in the meantime since we have different years data and different cross sectional data. So, we have you know so, like flexibility to do this.

So, and for that you should understand clearly how it is happening and what is the you know procedure through which actually you can bring this kind of you know flexibility to analyse the problem more accurately and more appropriately as per the need and the requirements.

So, ultimately so, this is how the case and what will you do? Agains.

(Refer Slide Time: 25:18)



If you go further so, here i constant and t will be vary. So, now, what will do actually in the next iteration we allow actually a i 2 variant t remain constant ok. So, compared to the previous ones so, there are where i constant and t vary. So, now, here if you put you know i constant let us say i equal to 1 and t will vary.

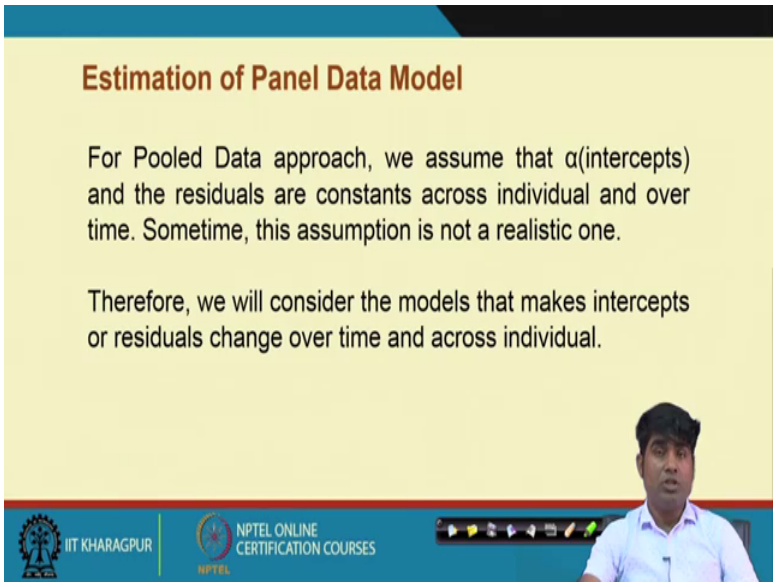
So, then by default the model will be like this and here t will be vary varying to 1 to t. So, agains again we put equal to 2 and allow t 2 vary. So, then the model will be like this so; that means, here what is happening? So, we have a data with respect to different cross sectional units. So, we can run the model for the cross sectional unit 1, then run the model cross sectional unit 2, run the model cross sectional unit 3 like this in the previous case.

So, same cross sectional unit, but you can run the model with different time frame different time frame, but here what is happening? So, time remain constant within a particular time frame you run the model with different cross sectional setup that is how the difference ok.

And that is what the beauty of this you know panel data structure if you are understanding is very clear. So, you have lots of you know; that means, technically you have plenty of you know flexibility how you have to generate the things and can bring a best output or you know best model through which you can do the better prediction and forecasting as per the particular you know engineering requirement.

So, ultimately so, this is what the structures then finally, if you if you goead like this ok.

(Refer Slide Time: 27:13)



So, then ultimately you can understand the entire you know panel data structures. So, understanding panel data and understanding the panel data models 2 things are a different. But, there is a huge and huge you know kind of you know chemistry between the panel data and panel data models.

So, you know you have a different kind of you know understanding you know on this before you link to a particular you know engineering problems and do the prediction as per the decision making you know requirement.

So, for panel data approach we assume that the intercept and the residuals are you know constant across individual and that too over time. Sometimes this is assumptions you know are not actually realistic one. So, that is how panel data come forward and you know you know solve the problems with you know less assumptions. And you know more in a kind of you know solid way to address the same problems with different kind of you know setup and you know different kind of you know structure.

Therefore, we can we can you know considered that the models that next intercept residuals change over time and across the individuals is nothing, but called you know panel data models.

So, if they are constant then; that means, technical that is called as you know simply pooled data structure and where you know the time specification, cross sectional specification, pooled specification is not. So, important because ultimately it is the data set simply used to run the models or estimate the models. Ultimately end of the day we have a different observations only, if you use only time have a less observation, if you know run the model with cross sectional again have a less observation and when you put you have just you know extra observation.

But, the process is more or less same actually what we have done either through time series or through cross sectional and then just linking that is brings the pool. But, on the top of pool if you bring some kind of you know specification or the kind of you know importance about the time element and cross sectional element then the entire setup will change and the particular you know framework will be also different.

Of course, the outcome will be more reliable and more solid a to address the engineering problem provided the model should give the signal that you know yes this model can be appropriate or this problem can be very appropriate to analyse through panel data models.

That means we have a different you know specification test through which we can proceed to either you know panel data or you know we can stick to the pooled data only. And then finally, we can actually analyse the problem as per the particular you know requirement.

So, in the next class we will discuss all this test and the kind of model estimation procedures with respect to fixed effect models, random effect models and GMM. And then we can connect with some of the engineering and industrial problems to highlight it the. So, if you know panel data models and to address the problem or effectively and more accurately with this we will stop here.

Thank you very much have a nice day.