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

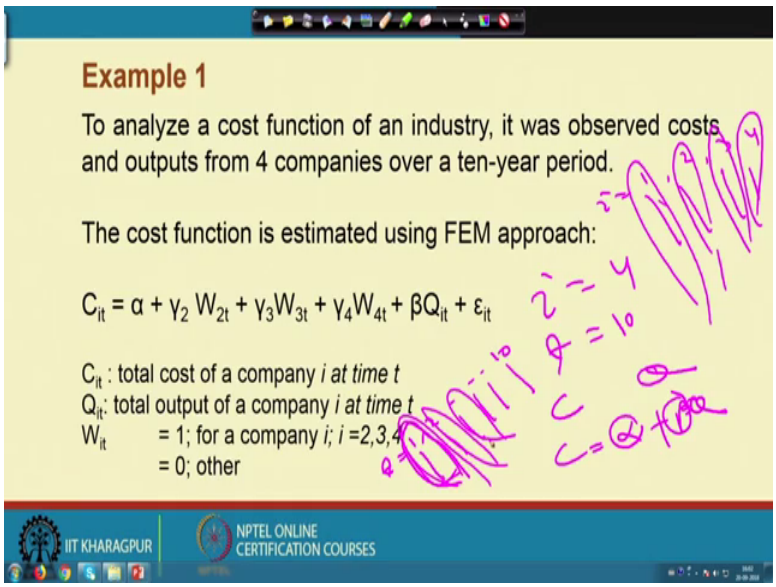
Lecture – 56
Panel Data Modelling (Contd.)

Hello, everybody. This is Rudra Pradhan here. Welcome to Engineering Econometrics. Today, we will continue with you know Panel Data Models and in the last lecture we have already highlighted the fixed effect model, the random effect model the estimation procedures of both random effect model and fixed effect model and get to know the difference between fixed effect model and random effect model.

And, we also discussed: what are the ways you can actually make a choice to use random effect models or fixed effect model sometimes we may use actually both the models simultaneously. But, ultimately in this lecture we like to check how the estimation process and how the panel data models can be used to bring some kind of you know difference corresponding to time series modelling you know modelling and the kind of you know cross sectional modelling.

So, in order to justify the importance of the panel data you know models for any kind of you know engineering you know problems. So, let us take an you know take an example and here the example is like this.

(Refer Slide Time: 01:36)



So, let it is a industry problems and generally the industrial problem is like that most of the instances it is the game between input and output relationship and everytimes we are we are actually dealing with you know operations that is mostly operational problems and here we are keen to know the cost aspects. So, that means, you know how output can be optimised with respect to minimum cost. So, that means, every times we should have a more and more output within a less and less cost that is what the kind of you know requirement.

So, now what is happening since we are using actually both cross sectional units and time series units you know that too in a panel data setup, so, if you look into the real life problems you will find you know when there are you know many firms. And, you know that is what the cross sectional type of you know situation and then how these forms are behaving you know uniformly or differently you know with respect to different time frame.

By default the immediate problems which you can actually highlight is that you know for you know means technically it may be two different ways. The first one is a you know for a particular time periods how the cross sectional units are you know you know they behave are they performing uniformly or they are you know you know they behave differently for a particular time periods. So, we like to check whether they their behaviour is same or you know different.

Again, with a particular time you know cross sectional units we like to check whether the particular you know cross sectional units say industry or a particular firm is maintaining kind of you know stability or you know sustainability; that means, they maintain a similar kind of you know status. So, either they maintain a similar kind of you know status or they may have a different kind of you know status; that means, there is ups and downs. So, far as a production is concerned cost is concerned like that.

So, that means, these are the types of you know problems if you understand the situation is like this then you know the immediate choice is to use panel data model because here we have a plenty of you know cross sectional elements and plenty of you know time series element and keeping time component constant and if you allow cross sectional you know variations then there is a kind of you know implications. Again keeping you know

cross sectional unit constant and allow time element will vary again there is a different implications.

So, now, we like to bring that because we have already discussed all these you know concept in the form of you know fixed effect models and random effect model. So, now, we like to bring into the reality, how is that and how they can differentiate and what are the best outcomes which you can actually bring here to analyse the situation. To analyse this particular structures let us start with you know a cost and output for you know for this particular you know you know structures what we can like to do we will we will have a problems.

So, the problem is a problem is like this. We have we have actually a cost function of an industry it was observed that you know cost and outputs from different companies of a 10 different you know time you know time periods technically. So, here if it is 4 different companies means I equal to 4 and 10 different time periods means t equal to 10. So, that means, if you go by actually you know analysing or to understand the situation the situation will be it is game between C and Q. So, simple the function will be C equal to alpha plus beta Q, ok. So, alpha is the intersect beta is the coefficient that too represent the output with respect to cost.

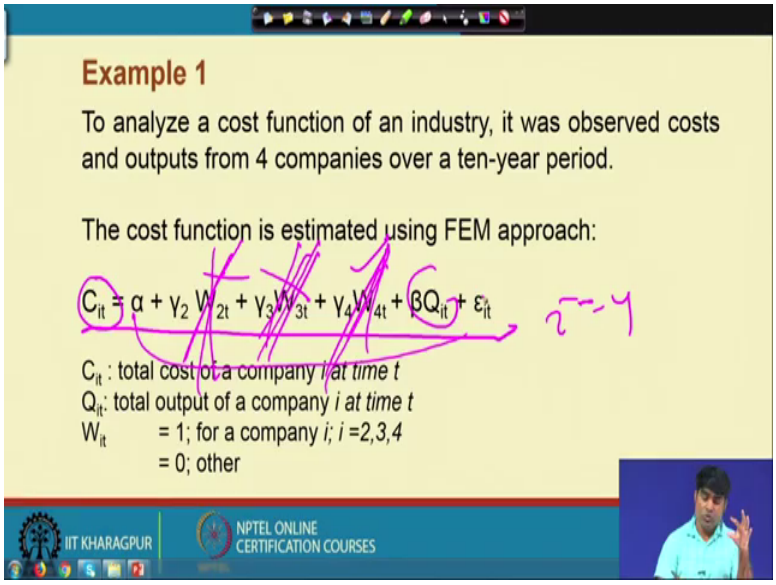
So, now the game is actually with respect to i and t that is the sample frame work through which you can understand the reality and we can use this particular structure for the estimation now you know if you start with let us say I equal to 1 2 3 and 4 then time will be vary 10 10 10 10. So, that means, we have a 10 different outputs all together like this and again if you apply t equal to 1 1, 2, up to 10 then every case we have a four different sample points of course, the sample size is very low, but still with two you know variables we can we can actually estimate the kind of you know process.

Of course, for you know having more and more i and more and more t the problem will be more and more interesting. So, that means, technically if you allow t will start with 1, 2, up to 10 separately then; that means, technically we have a 10 different outputs, and every output will represent the a functional relationship between cost and an output. So, that means, technically we like to know in that case you know if you allow you know t constant then you know i will vary. So, that means, the clear cut indication is that you

know for a particular you know time period how these forms are actually you know bringing the cost and output relationship.

So, again if you go to you know another side where i will be you know remain fixed; that means, when i equal to 1, then we have a one output then when i equal to 2 we have another output. So, in that case so, we like to check actually how you know time variations will bring the cost and output difference. So, this is how the theoretical understanding about this problem. So, now, what we will like to do, so, to differentiate this particular structures let us start with a simple model like this.

(Refer Slide Time: 08:06)



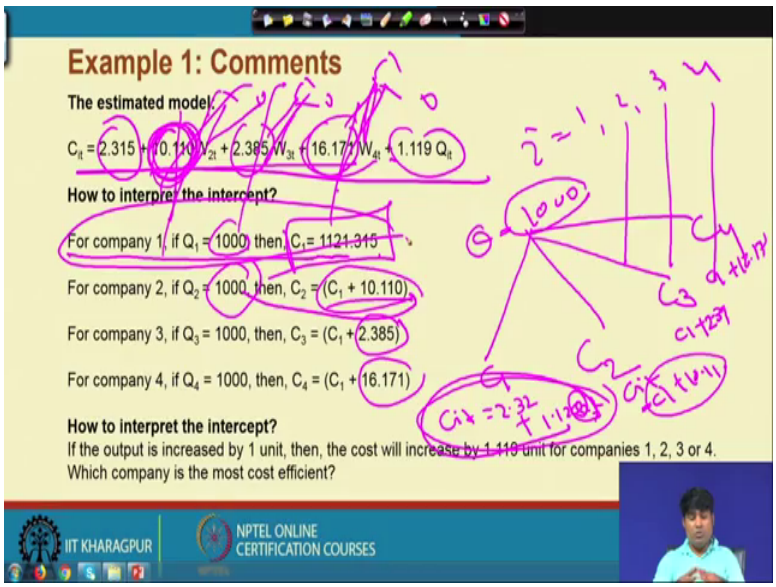
So, this model is actually this model is like this C i t equal to you know beta Q i t. So, that is how the pool data used and to represent the panel then you have to bring the panel features. So, now, what is happening here; so, we start with a first you know connecting to connecting to cross sectional units so; that means, since we have i equal to 4. So, that means, we are using three different domains ok. So, this is for company 2, this is for company 3, this is for company 4.

Now, we have i equal to 4; so, that means, when we put you know i equal to 2 so, that times this will be active and the these two will be inactive. So, when i equal to 3 then this will be active then these two will be inactive when i equal to 4 this will be active and this two will be inactive. So, now, when all this three will be inactive then by default the particular impart will go to the i equal to 1 that is the form one.

So, now obviously, so, we like to check you know how cost will be differentiate you know you know among these four forms. So, far as you know output is concerned so, now, theoretically what we can do. So, theoretically the observation is that we asked it typically you know for forms that you know they will be go for you know uniform productions. Let us say Q equal to one thousand numbers or something like that and then we like to check what is the cost difference.

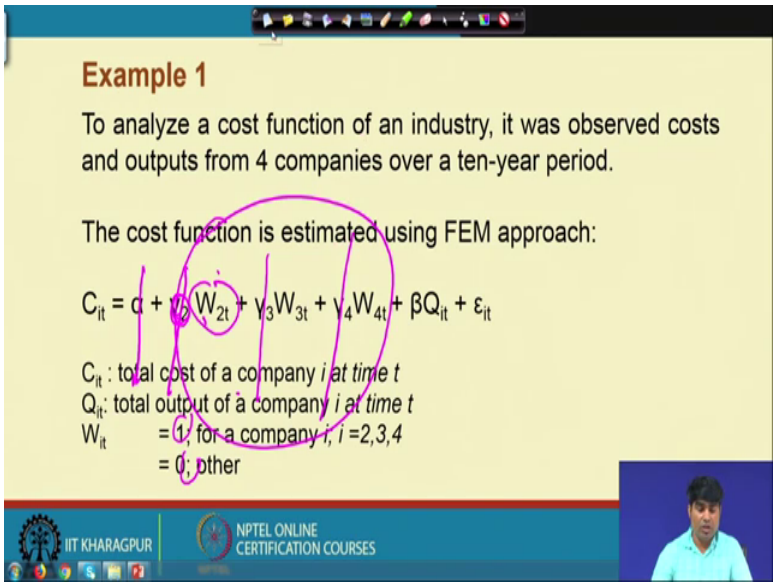
Then finally, after checking the cost difference with you know same level of production, then there are two possibilities either the cost you know cost for these you know one thousand output will be remain same or cost will be different for you know different forms if they are same then you know a econometrics will not you know give better kind of you know explanation. But, if they are different then this particular model or this particular you know structure can give better kind of you know interpretation in order to know what is exactly happening and what is the reality.

(Refer Slide Time: 10:20)



So, let us see the model outcome. So, now, having the structures so, we can go to the model outcome. So, now, the model outcome will be like this see here. So, means this is this is actually with the help of you know data you have to just estimate the model then ultimately after the estimation. So, we will have here alpha alpha component.

(Refer Slide Time: 10:34)



Then dummy, first dummy, second dummy and third dummy. So, these are all these three are you know actually dummy. So, it will behave either you know 1 and 0. So, it will be 1 then others will be differ 0, this is 1 then others will be 0 like this. So, ultimately so, the impact will be actually restricted with respect to 1 and 0. So, ultimately the final impact will be multiplied by the coefficient with respect to 1 or with respect to 0. So, when we like to allow this active then this will be 1 then by default others will be 0. So, this is what the kind of you know structure.

So, now what will we do here so, we will we will check the outcome. So, now, this is what the estimated outcome and in this estimated outcomes. So, we have actually a alpha coefficients then the first dummy coefficient, second dummy coefficient, third dummy coefficient and by default W 2 will be moving for actually 1 0 then this will be moving with 1 0 this will be also moving with 1 0. So, that means, W 3 and W 4 ultimately. So, now, what is happening here. So, we have actually i equal to 1 2 3 and 4. So, but we have taken actually three dummies. So, that too for company 2, company 3 and company 4.

So, now, what is happening we like to check a particular level of output let us say Q equal to 1000, right. So, now, we like to check what is the cost coefficient cost components for you know 10000 for company 1 that will be represented by C 1 then cost coefficient for the company 2 represented by C 2 then C 3 and then C 4, right.

So, now what is happening since this is the dummy coefficients for you know company 2, this is dummy coefficient for company 3 and this is dummy coefficient for company 4 then by default we assume that this is equal to 0 then by default the coefficient the particular you know impact will be restricted to 2.32 plus 16.17 plus 1.112 Q. So, that is how the impact. So, that means, if all are in actually 0 then for C 1 the cost coefficient will be C i t equal to 2.32 plus 1.12 Q Q i t. So, that is what the actually cost functions, ok.

So, now for company 2 so, then this will be a this will be a you know net addition with respect to net addition of you know 10.11. So, that means, technically so, it will be it will be here C i t equal to so, C C 1 plus 10 plus 1 1. So, similarly in the case of you know C 3 it will be C 1 plus 2.39 and then for C 4 it will be C 1 plus 16.17. So, that is what the difference. So, that means, this is the best model which represent the cost structure of the first company that too a uniform output and again for company 2 the cost structure will be C 1 plus 10.11 and then for company 3 C 1 plus 2.39 and C 4 C 1 plus you know 16.17. So, that is the cost difference.

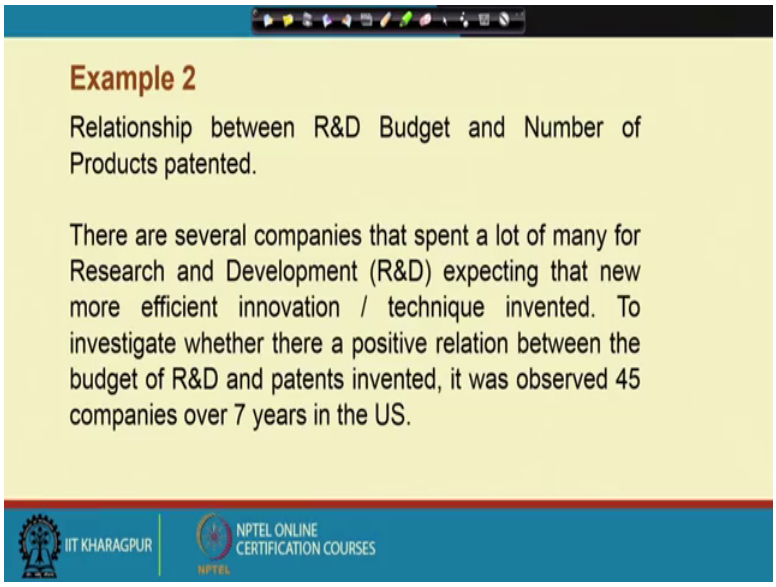
Now, to simplify this one; so, for company 1 put Q equal to 1000 then if you put Q equal to 1000 here and you simplify then the total cost will be a will be this much that that too let us say 1121.32 and that is the cost you know having for company 1 to produce 1000 output. So, now, for coming to company 2 to produce the same level of output 1000, the cost component will be C 1 plus 10.11, that is what we have already obtained. So, now, so, 1121.32 plus 10.11 that is the cost component for company 2 to produce the same level of output 1000 and again for company 3 the net addition will be 2.385 and for company 4 the net addition will be 16.17.

So, now so far as a efficiency is concerned since we are allowing the companies to produce the same level of output. So, now, for a particular company or a particular form can be declared as efficient where the cost factor will be substantially low. So, that means, technically in this case company 1 is declared as you know most efficient compared to company 2, company 3 and company 4 and that is how the beauty of this kind of you know panel data model.

So, that means, you know you know using the panel data you bring the kind of you know estimated output now through these estimated output and bringing the kind of you know

panel component that too cross sectional dummies and the time dummies you can actually analyse differently and then bring the problem in a more attractive way to highlight the some of the critical issues of the you know engineering problems.

(Refer Slide Time: 16:26)



So, likewise we can have actually different kind of you know structure to you know analyse through panel data model. And, we can consider you know means we can you know discuss the same kind of you know structure through second examples that too relationship between R and D expenditure and number of you know patents. So, there are several companies that spend you know you know more revenue towards you know R and D R and D activities and then you know they are expecting you know more and more innovations.

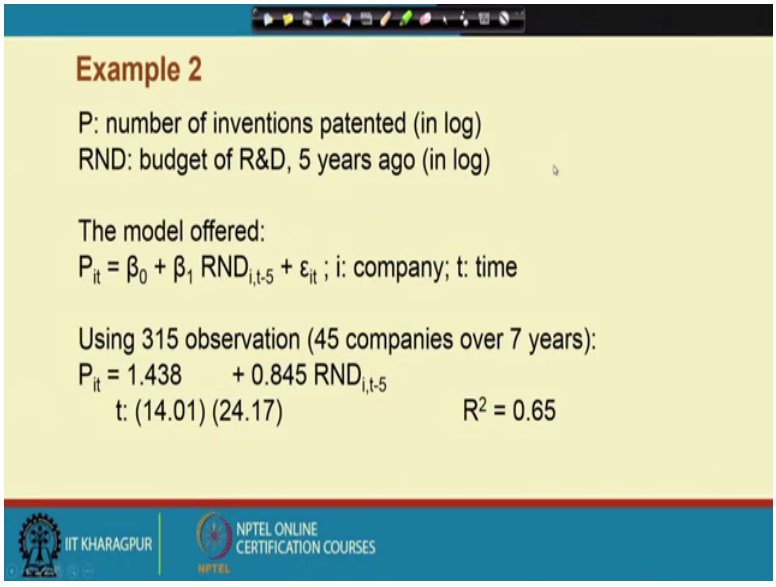
So, obviously, we are expecting the R and D activities and you know innovations are positively related, then again it is kind of you know similar problems. We can have a situation like that you know as you know companies to put you know similar kind of you know expenditure and check what is the you know outcomes that is the innovations and we like to check whether you know a particular form is more innovative a compared to the other forms so far as you know uniform expenditure is concerned. Like the like the previous case where you know we are checking the cost difference with respect to same level of output here we like to analyse a having same level of you know R and D

expenditure what is the level of you know innovation with different you know companies.

So, that means, these are the classic examples where you know panel data can be used to you know analyse the problems and come with you know excellent kind of you know implication through which you can generalize the structure and can get more inference and which is not exactly possible in the case of you know either cross sectional units and you know time series. Even it can be possible through cross sectional unit and time series unit the kind of you know findings and the kind of you know outlook is a completely different if you analyse the same problem through panel data modelling.

Because we are actually pooling the data and getting the reality; that means, we are bringing in a kind you know competitive situation then you are comparing. Otherwise you know you have a just individually comparing without any kind of you know kind of you know competitiveness or if the kind of you know competition. So, a panel data by default will bring such kind of you know environment which can bring the beauty as per the requirement of some of the engineering problems.

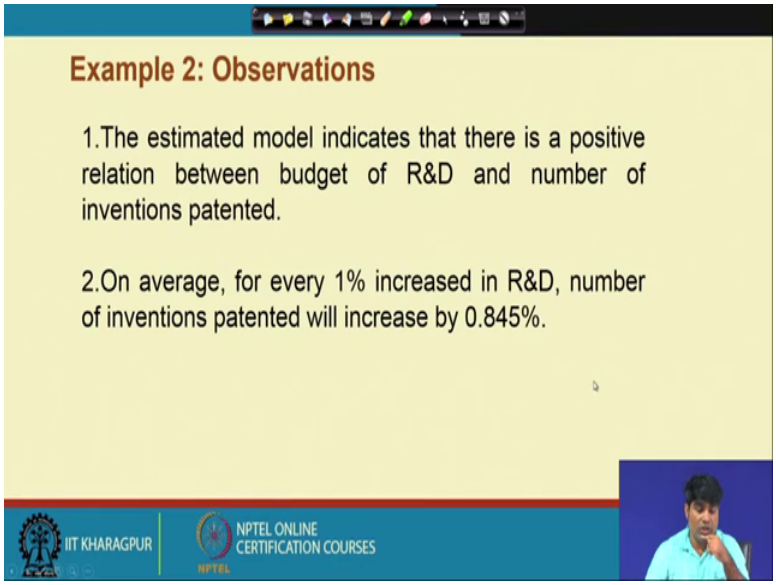
(Refer Slide Time: 18:52)



So, now corresponding to this you know problems. So, we have actually models. So, because patent is the kind of you know components here RND, if you fix this one is the constant then how they behave each other and you know differentiate each other same way. So, here the estimated model is like this. So, now, you can fix a particular

component and then you can examine the situation and then the model outcome is like this.

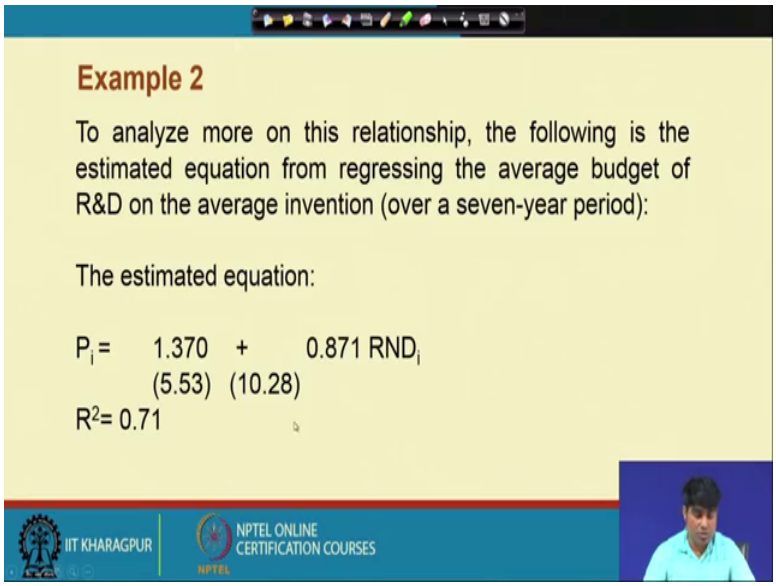
(Refer Slide Time: 19:20)



And, so, that means, technically the estimated model indicates that you know there is a positive relationship between budget you know R and D budget and number of you know patents that is what you know our main target and the secondary target is to know whether there is a kind of you know difference which form can be declared as you know more efficient here the interpretation will be with same level of R and D expenditure who is bringing more you know innovation to the system that that can be declared as you know more efficient compared to the previous one where giving same level of output to which form is having you know less cost.

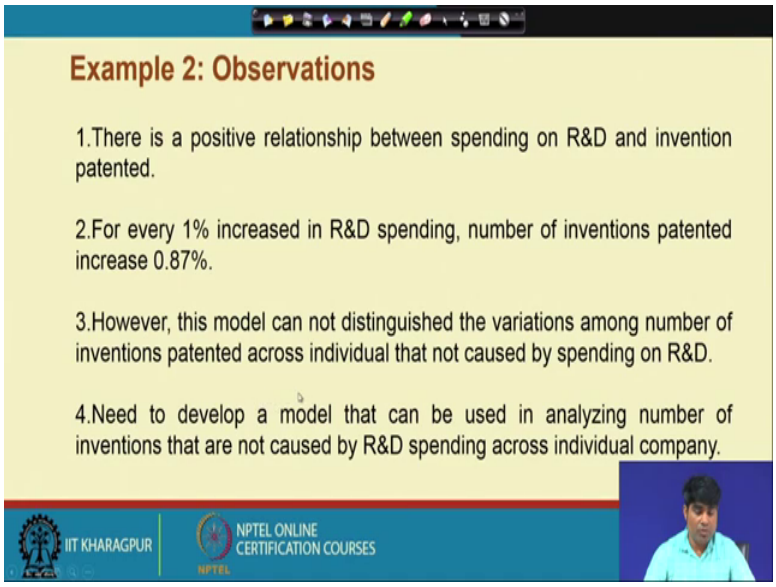
So, you know if these are you know these are the problems then by default you can use panel data model and then you can analyse in a more attractive way as per the particular you know engineering requirement. So, on an average for every 1 percent increase in R and D number of patents will increase by 0 point you know; that means, 80 the this is.

(Refer Slide Time: 20:18)



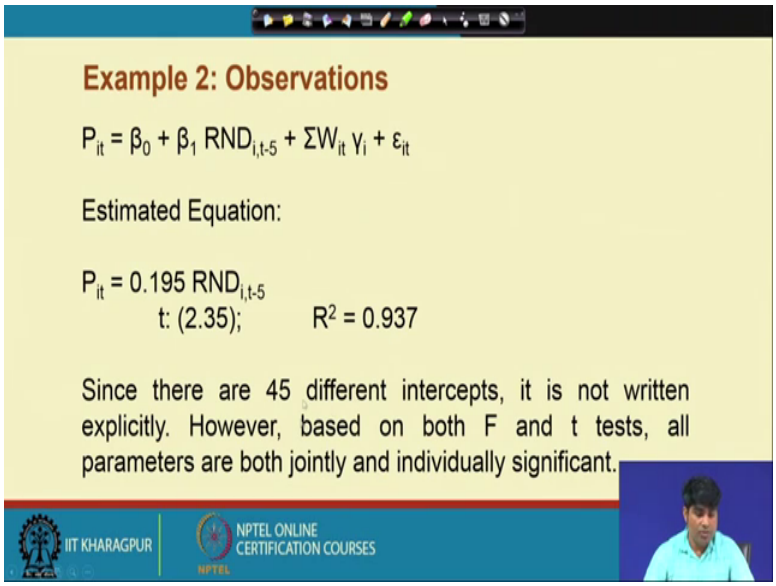
This is what the a kind of you know structure.

(Refer Slide Time: 20:26)



So, now what is happening actually we have actually the kind of you know model output which is which is actually satisfying as per the particular you know you know requirement, but on the top of that we are actually bringing different kind of you know inference which can you know produce much better results and much better inference compared to you know simply cross sectional modelling and the time series modelling.

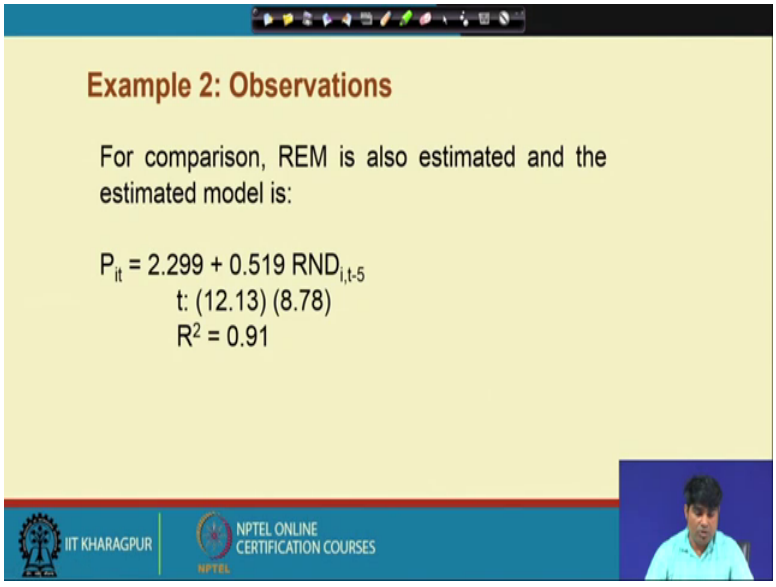
(Refer Slide Time: 20:55)



So, ultimately so, now, in the second observation you know models again you can bring actually dummy impact and then we try to check which form is more efficient compared to previous one.

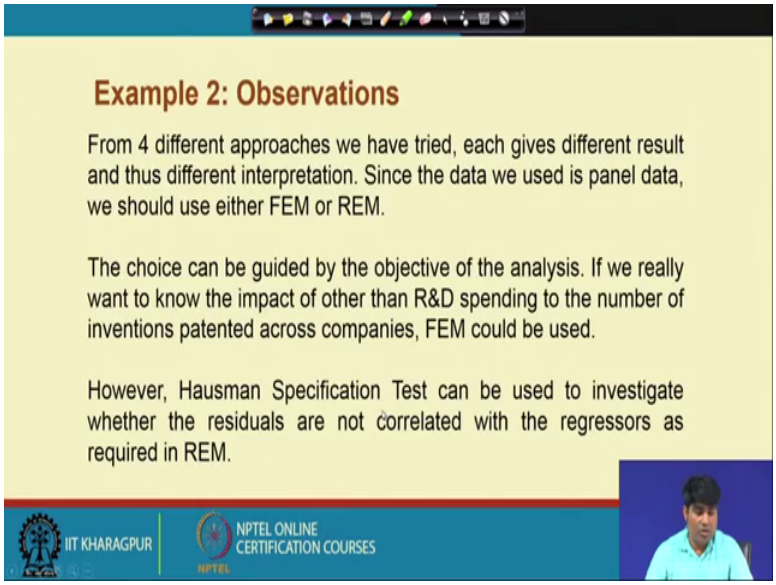
So, likewise you know you can use both you know a fixed effect models and random effect models and then bring the situations where we can actually in a position to analyse which form is more efficient so far as you know R and D expenditure is concerned and the output innovation is concerned, ok.

(Refer Slide Time: 21:31)



So, ultimately the it is same model can be used to random effect model then by default intercept term will not be there, but ultimately the model outcome will be adjusted through the error component. Of course, the effectiveness of the model will be depends upon the significance of the parameter and the R square value which is actually coming here very high.

(Refer Slide Time: 21:55)

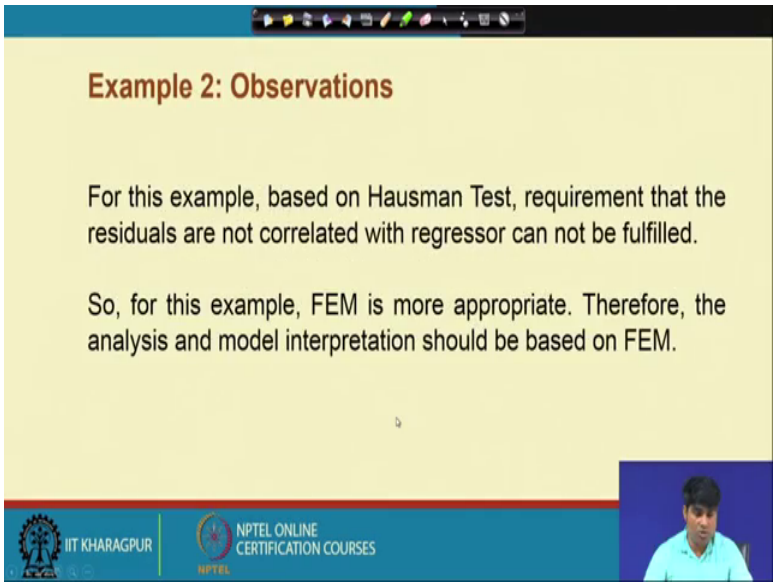


And, now what will we what will we do here from four different approaches that we have tried here is gives different results and different kind of you know interpretation. Since the data we use is actually panel data we should use either you know fixed effect model or you know random effect model. Now, what is happening here the choice can be guided by the objective of this analysis. If we really want to know the impact of you know other than R and D is spending to the number of you know innovation that too patents so, fixed effect model can be used.

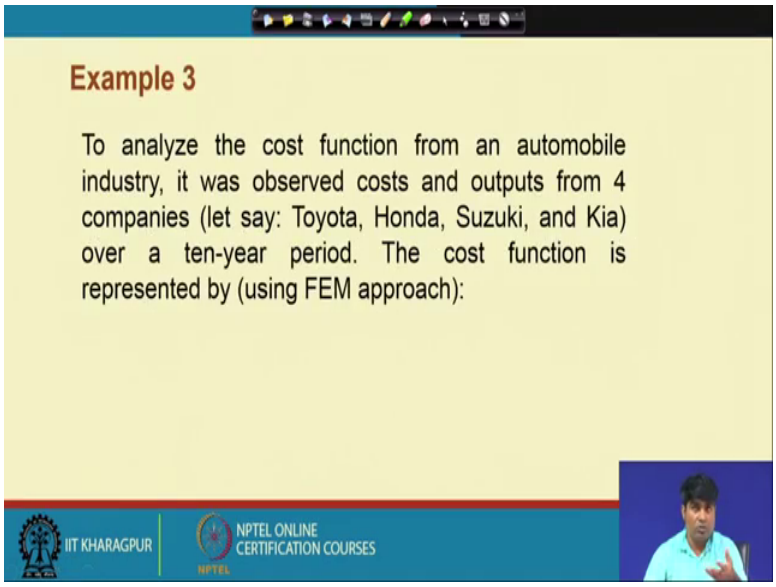
Moreover, the Hausman test can be used to investigate whether the residuals are not correlated with regressors and means as long as you know you like to check you know the use of fixed effect model and random effect model. These are all called as you know means two things; one is the technical requirements the other one is actually the problem requirements. If the problem requirement is you know kind of you know comparative analysis then it is better to bring the fixed effect model. If the problem requirement is not the comparative analysis if it is only to check whether there is a strong association

between you know dependent variable and independent variable then random effect models may be a better choice. So, that is how the a big deal between fixed effect model and random effect model.

(Refer Slide Time: 23:20)



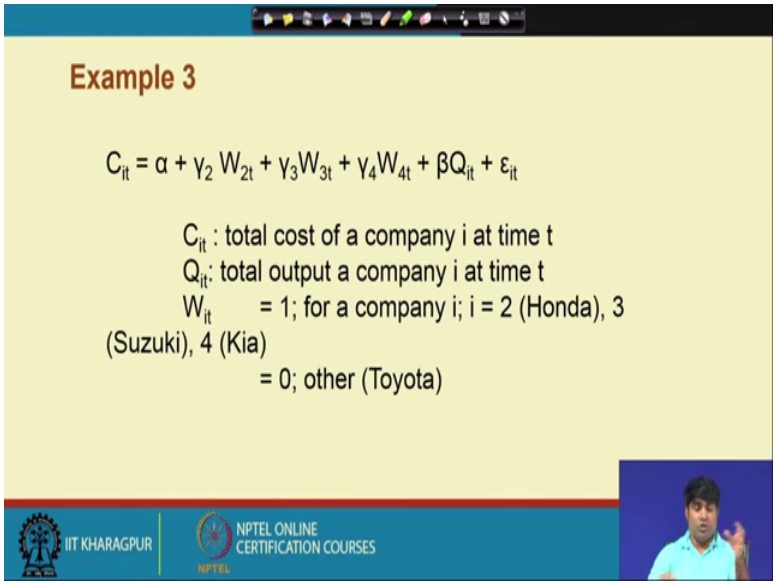
(Refer Slide Time: 23:22)



So, likewise we have actually a different kind of you know approaches. So, I can bring another examples you know where actually we have four different you know five different companies and where ten different time periods like the past examples. Again

we like to check the cost difference you know so far as you know output and you know cost is concerned.

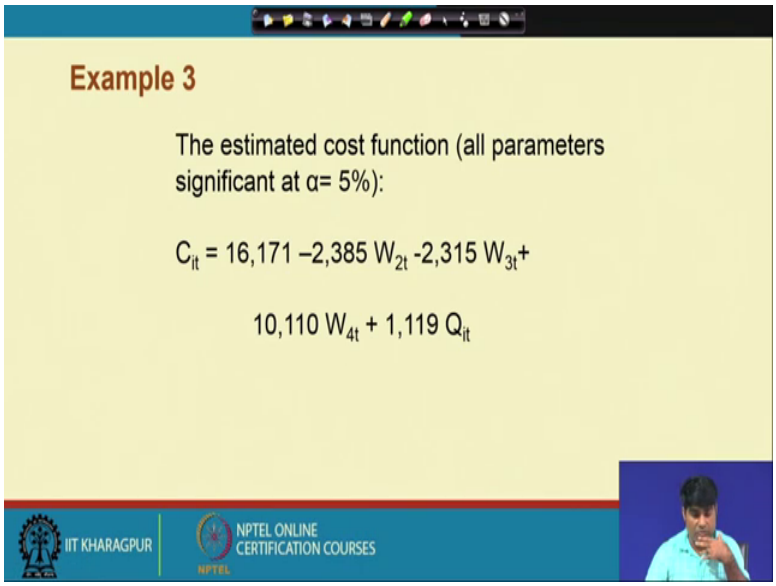
(Refer Slide Time: 23:43)



And, same way you can bring actually a three dummies and when these three dummies are will be actually 0, then the impact will go to the first company. And, the second company means the first company by default will be the best model and the impact of second company where you know first dummy will be come into the picture and second company the second dummy will come into the picture means third company will be with respect to second dummy and then finally, fourth company with respect you know third dummy.

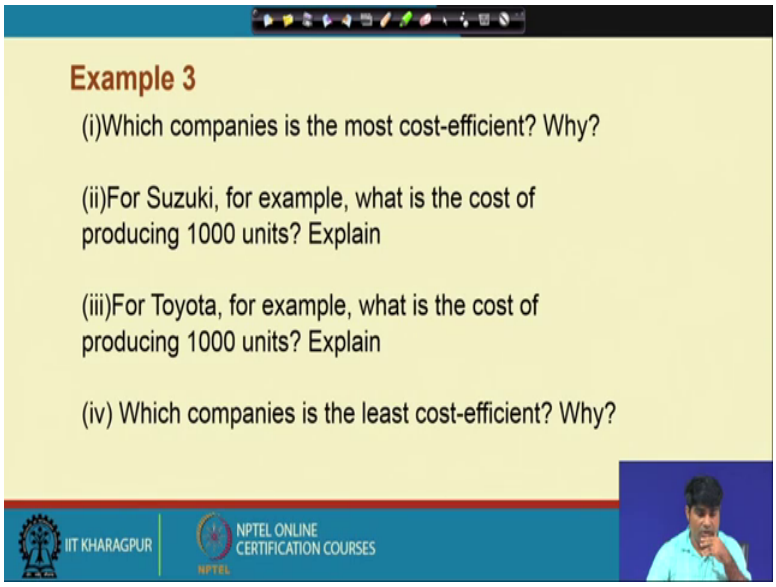
Because, we use ultimately three dummies that too the representative of you know company 2 company 3 and company 4. So, when company 2 company 3 company 4 will be inactive then the impact will go to the intercept and the kind of you know independent variable coefficients only. So, ultimately that is the best model. So, now, the first dummy if you introduce then that will be the impact of the second company and then second dummy will be impact of the third company and you know third variable third dummy impact will go to the company 4. So, likewise we have a different kind of you know examples.

(Refer Slide Time: 24:57)



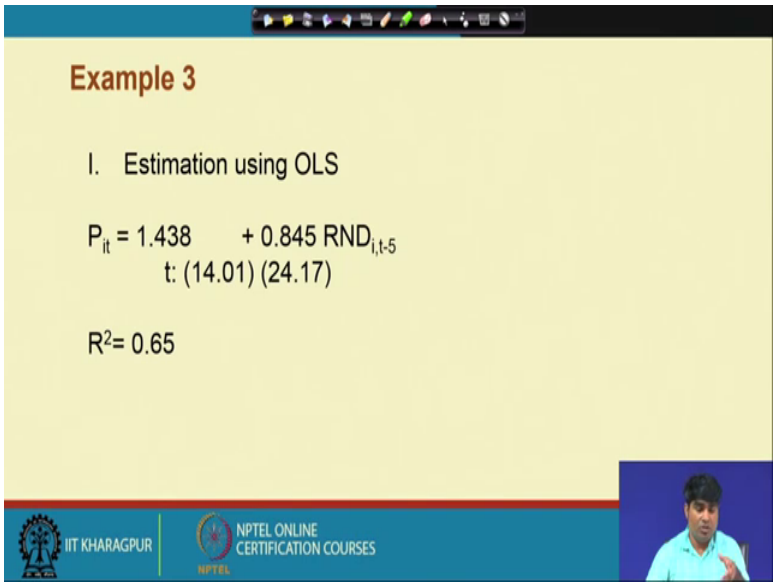
And this is how the kind of you know structure.

(Refer Slide Time: 25:01)



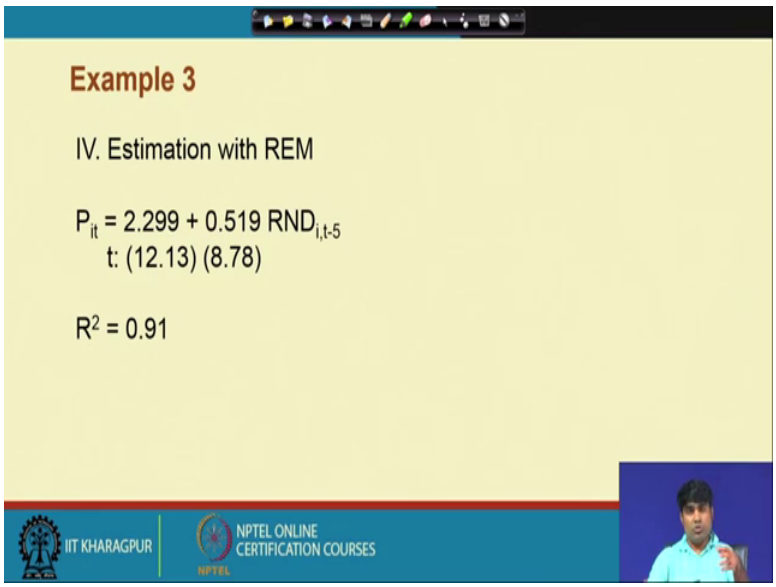
And, ultimately like the previous case we can fix actually output constant and then check the difference which company is more effective compared to you know same level of output, ok.

(Refer Slide Time: 25:08)



So, that means, technically we have discussed you know various kind of you know examples.

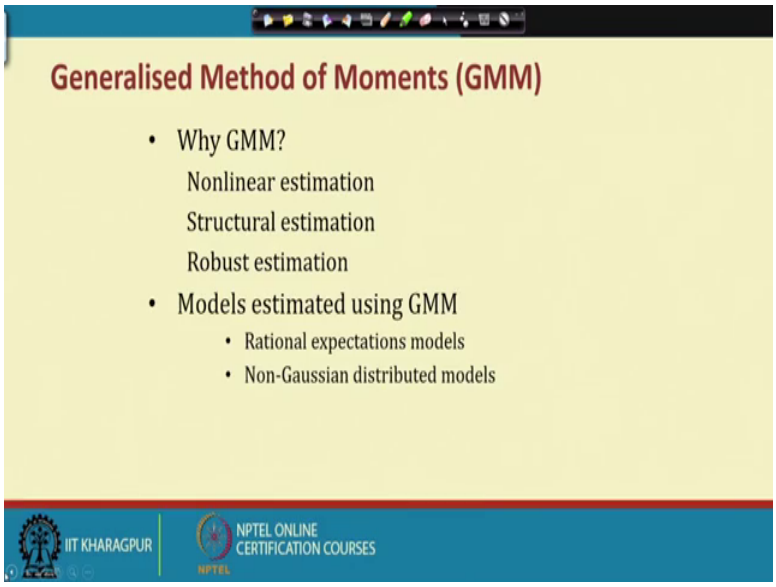
(Refer Slide Time: 25:18)



Through which you can justify that you know panel data model or panel data modelling is actually good choice compared to cross sectional modelling and time series modelling because it will bring much better inference and some the outlook is actually much you know broader compared to the you know simple cross sectional modelling and you know time series modelling.

So, ultimately till now we have discussed the pool data models and in the panel data model that too fixed effect concept random effect concept and then we have another concept called as you know GMM concept.

(Refer Slide Time: 25:57)

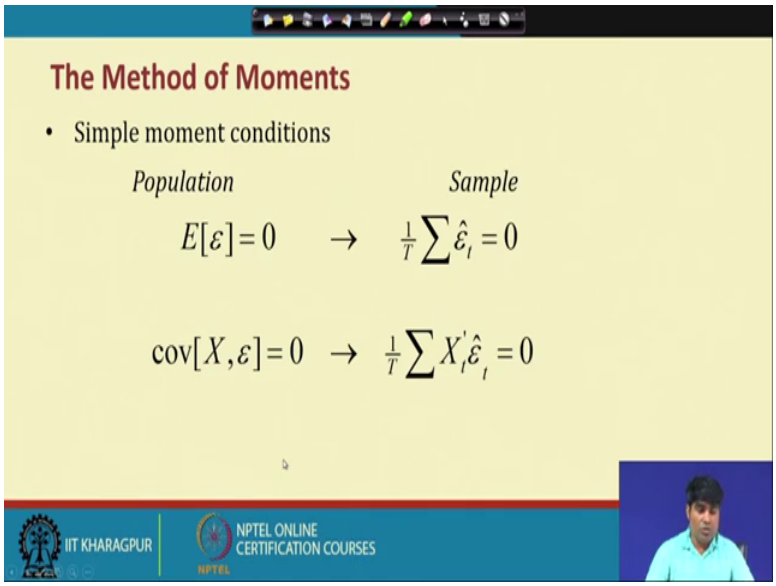


So, GMM basically generalized method of moments. Now, the question is why GMM, if we have a pool data then we have actually fixed effect model and we have a random effect model, but GMM is a different kind of you know approach which is actually more advanced compared to pool data fixed effect and random effect.

And, we have actually three important you know structure through which you can justify that the requirement of you know GMM and that will be high class model and it is you know much kind of you know efficiencies compared to you know means more better you know models or you know better efficient model compared to fixed effect model and random effect because some kind of you know endogeneity issue which you can address through GMM and here if you could if you ask why GMM then straight answer is a with respect to three scenarios.

So, first is the non-linear estimation then structural second one is the structural estimation then the robust robustness you know estimation. And, the models of estimated using GMMs can be of you know a rational expectations models or can be non-Gaussian distributed models, right.

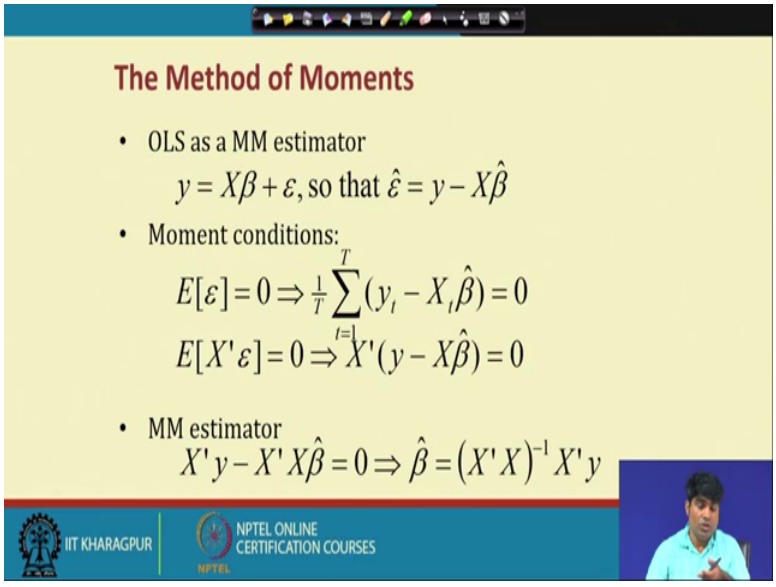
(Refer Slide Time: 27:20)



So, basically it starts with you know simple moment conditions population versus samples and this is what actually mean of the error terms and covariance of the error terms and the independent variables. And, by default sample mean we which we can derive by taking the you know total sample size.

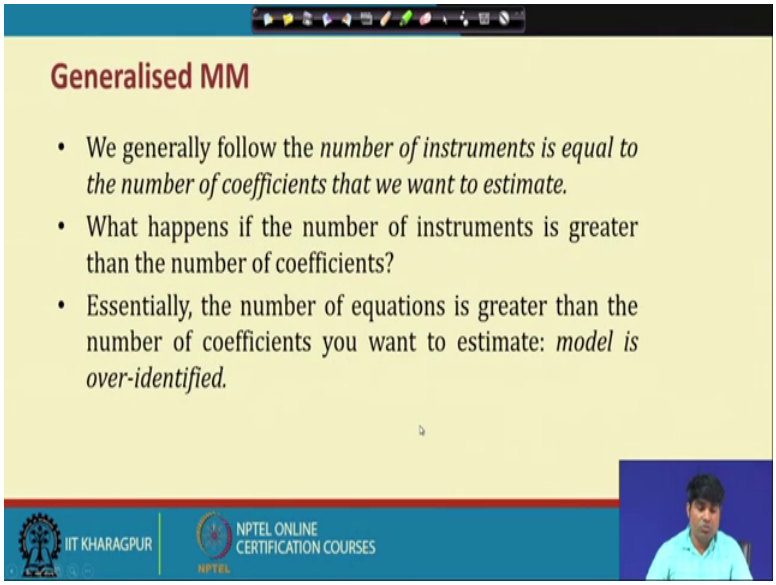
So, we are expecting that you know mean of the sample will be also equal to 0, and also the covariance between these two also equal to 0. So, that is actually the requirement of the you know regression modelling.

(Refer Slide Time: 27:55)



So, ultimately if you go by you know GMM so, OLS as a you know methods of moment estimator. So, simple y equal to X beta plus error terms so, that you know the we can obtain the error component. Then moment condition will follow like this and the mm estimator will be beta head equal to X transpose X inverse by X transpose y that is the same structure.

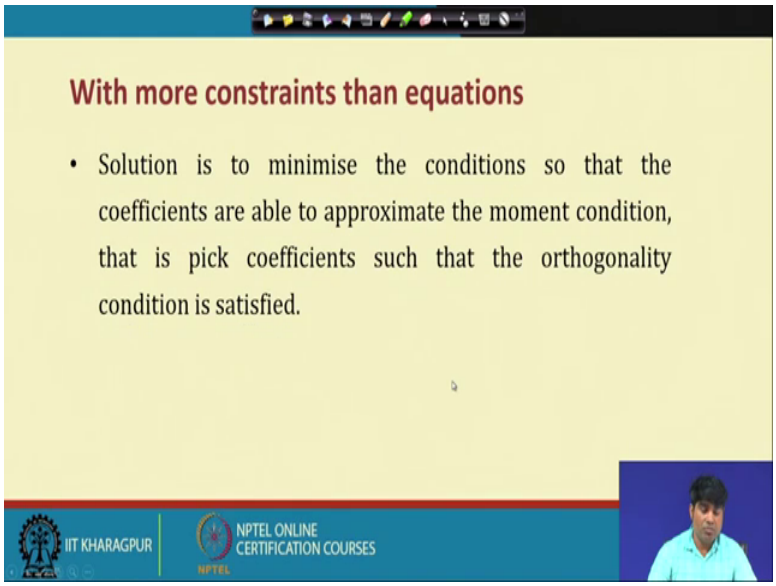
(Refer Slide Time: 28:20)



But, the you know the kind of you know adjustment will be is slightly different. So, now if you generalize the kind of you know method. So, we generally follow the condition where the number of instrument is equal to number of coefficient that we like to actually a process it. But, what is happening in some cases number of instruments is greater than to number of coefficient which is actually again this simultaneous equation system.

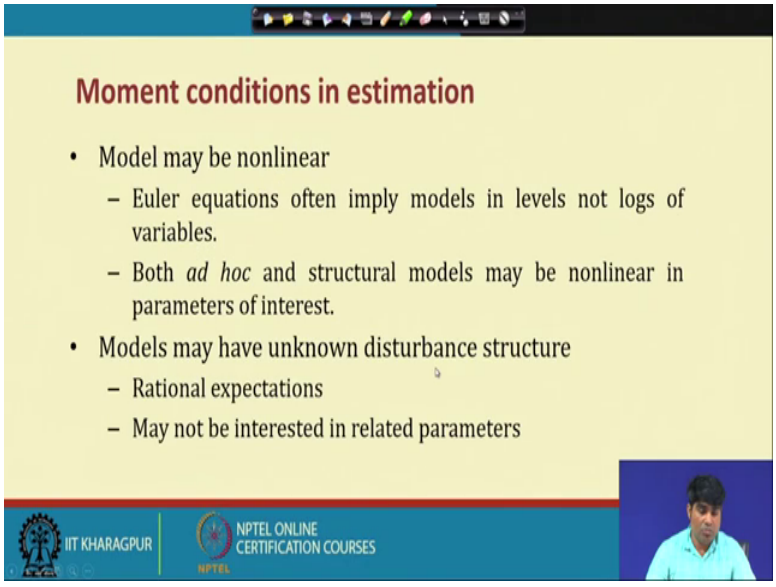
So, essentially the number of equations you know is greater than to number of coefficient then the model will be declared as actually over identified models. But, when number of you know equation equal to number of coefficient that is called as you know exactly identified and the simultaneous equation structure is very effective and in that case we may not have any problem, but when there is a question of you know over identified model then the GMM structure is very you know practically is very useful.

(Refer Slide Time: 29:15)



So, what is happening in the over identified case. So, the solution is to minimise the conditions. So, that the coefficients are able to approximate the moment condition and that is a pick coefficients such that the orthogonality condition can be satisfied, that is what is the kind of you know requirement.

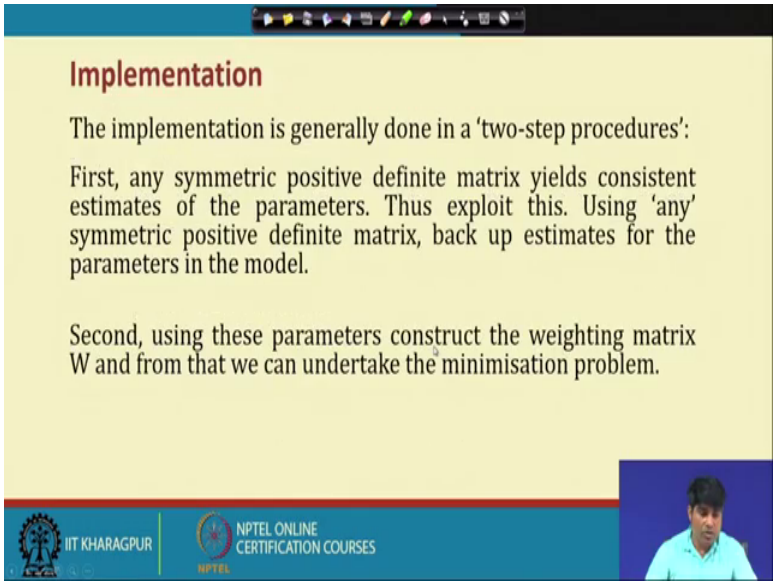
(Refer Slide Time: 29:36)



And, models may be non-linear, but some Euler equations often you know imply models in levels not logs of the variables; that means, some kind of you know adjustment you have to do you know when the situation is actually over identified. And, I am not

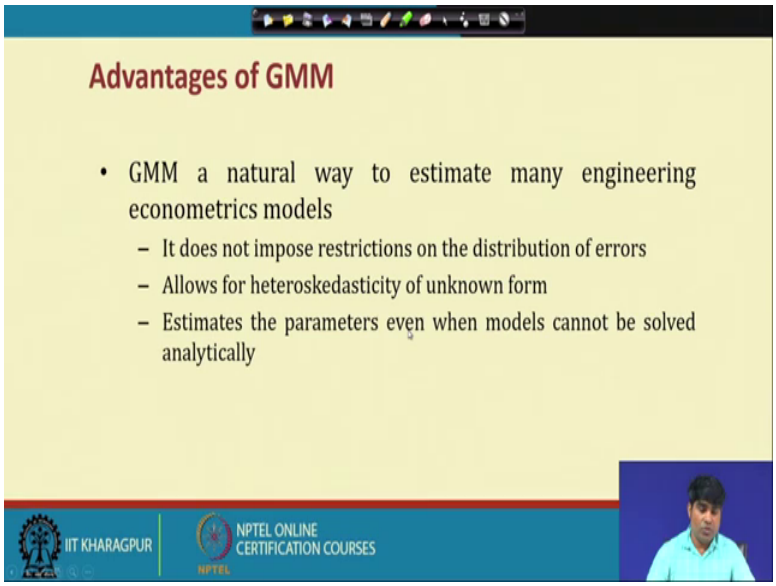
bringing actually too much mathematics here because the particular method is a too much critical and very complex compared to fixed effect models and random effect model. But, it is actually means the bigger advantage of this model is you know bringing the dynamics to the system and again addressing the kind of you know endogeneity component.

(Refer Slide Time: 30:18)



So, now the you know there the so far the implementation is concerned it is actually two step procedure. In the first steps any a symmetric positive definite matrix which yield consistent estimates of the parameters and second using these parameter construct the weighting matrix you know and then from that we can undertake the minimization problem. So, that is the technical you know means technical procedure you have to follow how to get the estimated outputs from the GMM.

(Refer Slide Time: 30:52)



Ultimately you know our you know problem kind of you know understanding and the kind of you know problem discussion ultimately depends upon the model outcome whether the outcomes are coming through fixed effect models or random effect models or a GMM. Of course, we have to use the models and the structure the model as per the specific objectives and the kind of you know requirements. If there is no specific objective and the kind of you know requirements ultimately the game will be ending with you know the impact of you know independent variable to dependent variable.

But, in between there are couple of side objectives which can actually explore and that too you know with the help of original panel data modelling starting with you know fixed effect models, random effect models and the kind of you know GMM. So, of course, it is actually kind of you know a complex models, but it has actually lots of you know advantage to solve actually a some of the you know engineering problems. And, the advantage of GMM is actually it does not impose restriction on the distribution of errors which is actually there in the case of you know fixed effect models and random effect model. And, it allow you know for heteroskedasticity of unknown form and then finally, estimate the parameters even when models cannot be solved analytically; that means, you know through the adjustment the over identified condition can be materialized. So, that is the actually big deal and big advantage of you know GMM.

So, to this you know you know end so, what we are you know actually summarizing that you know panel data is a kind of you know structure which can actually bring you know different kind of you know situations and different kind of you know kind of you know structure through which you can generalize the problem in a more attractive way and we can actually bring some of the issues or you know inference which is not actually possible through a simple cross sectional modelling and time series modelling.

That is why in the same problems it is possible it is better to analyse through panel data modelling and of course, it exclusives will be depends upon you know what is the kind of you know objectives ultimate objectives then you have to make a choice which kind of you know models you finally, use and whatever may be the kind of you know choice whether it is a fixed effect model or random effect model or you know GMM ultimately all these models are very reliable and very good for you know you know solving some of the big problems that too when the data structure is having both you know time series type and you know cross sectional type. And, we are interested to pool the data to analyse the situation as per the particular you know requirement with this we will stop here.

Thank you very much. Have a nice day.