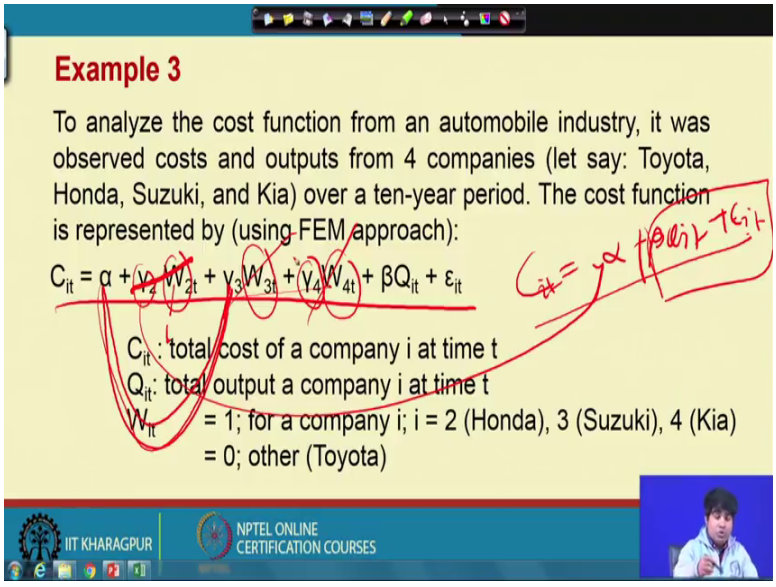
Business Analytics for Management Decision
Prof. Rudra P Pradhan
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

Lecture – 34
Predictive Analytics: Panel Data Model (Contd.)

Hello everybody this is Rudra Pradhan here. Welcome to BMD lecture series. We will continue with the panel data modelling and that too in the last class we have discussed about this particular concept.

(Refer Slide Time: 00:30)



That to understanding the panel data structure and the kind of you know requirement, the kind of you know objectives and then we have already highlighted the kind of you know types of you know model, and through which we can actually a you know handle the panel data and the kind of you know business kind of you know investigation.

So, now you know as per the kind of you know models discussion in the last lecture. So, what will do will connect with a particular problem and we like to highlight, how panel data can you know you know can be used for the you know predictions and the kind of you know the kind of you know management requirement.

So, now in this problem what we are you know going to do. So, we like to address the you know issue between cost and output that too we like to check what is the relationship

between cost and output and that to using the data of you know 4 companies and 10 times 10 you know different time periods. So that means, it is the annual you know time period and then we have a 4 different companies and we like to check which particular company is more effective, while storing the impact of you know output to cost and a particular level of you know output.

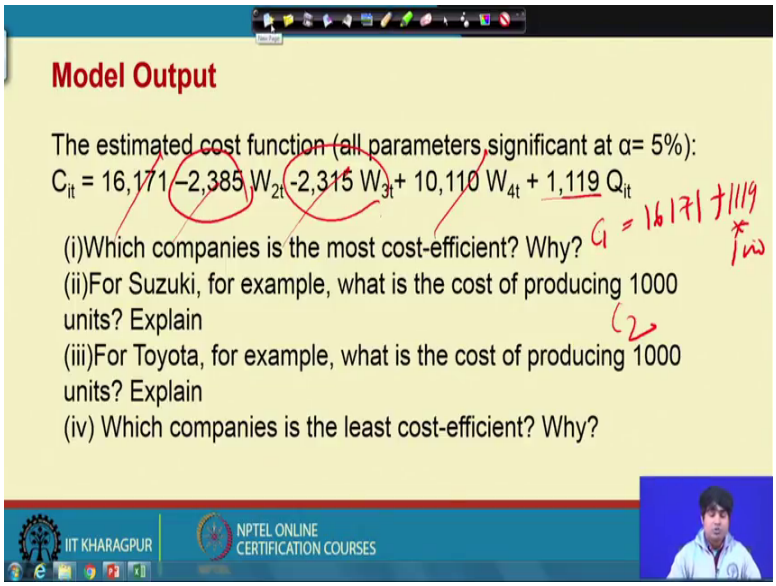
So, accordingly our model will be like this. So, our model will be like this. So, this is actually this standard model. So, since there are you know 4 companies by default we can actually start with you know fixed effect model and you know as a results we can develop a model here with respect to 3 dummies. So, W 2 is the dummy of you know second company, W 3 is the dummy of you know third company and W 4 is the dummy of you know fourth company.

So, by default W 1 is not there because of you know dummy a dummy variable trap and while when actually W 4 equal to 0 or W 3 equal to 0 and W 2 equal to 0 by default the model will take care you know with respect to the company 1. So that means, technically. So, when this this 3 will be removed. So, then the particular model will be restricted to C equal to C i t equal to alpha plus beta Q i t and is it is long I you know it. So, this is the you know company one you know structure.

So, when will actually allow sec you know second company to enter then this this becomes you know one and others will be 0 so; that means, this particular coefficient will be added to the intercept and then the impact of you know cost can be calculated. Similarly when we will go to the third company then and this will be equal to one rest of the items will be 0. So, then alpha or you know alpha plus this this this value will be coming as you know intercept and then rest of the items will be connected accordingly.

Similarly, for company so the impact will be alpha plus the particular you know you know coefficient. So, as a results we will have actually 4 different you know cost structure with respect to a particular level of output. Now in order to elaborate more you know attractive way. So, so what will you do? So, C is the total cost function of a company, you know company i at a time t and Q is the total output of a company i at time t and W i t is the kind of you know dummy variables, which can be represented as a one for a company and that too i will be vary from 2 to 4 and then 0 otherwise by default that will be take care by the first company.

(Refer Slide Time: 04:47)



So, as a result so with see how is the particular you know structure and then we will connect with the kind of you know discussions right. So, the typical structure will be a like this. So, now, corresponding to this model corresponding to this model so this is the output slide and you know we have estimated output. So, now, here you know starts with the alpha and this estimated model alpha is will coming actually a 16 171 and this is the second dummy variable vertex this is third variable dummy vertex and then this is fourth variable dummy vertex and this is the output you know slope coefficient.

So, now with the particular you know structures we like to know, which company is the most effective and why now for that we can we can fix a particular output level so; that means, you know the interpretation will be like this if we fix a particular level of output let us say 1000, then which company is having high cost and which company's having low cost and as a result we can you know declare, which one is the most efficient and which one is the less efficient and that to with respect to high cost and you know low cost respectively.

So, what will it do here? So, we like to check here's what we what is the kind of you know form one you know cost structure and what is the form 2 cost structure and form 3 cost structure and form 4 cost structure that 2 at output 1000. So, now in order to simplify so we can go to the you know next slide and then we can get to know the exact structure. Now what is happening here? So, you can just put here you know 1000 then

the first company impact will be the first company impact will be a. So, company wants that will be company wants 16 171 plus this will be 1000. So, technically this much of you know output. So, 1 1 1 9 multiplied by 1000. So, this will be the first one company impact.

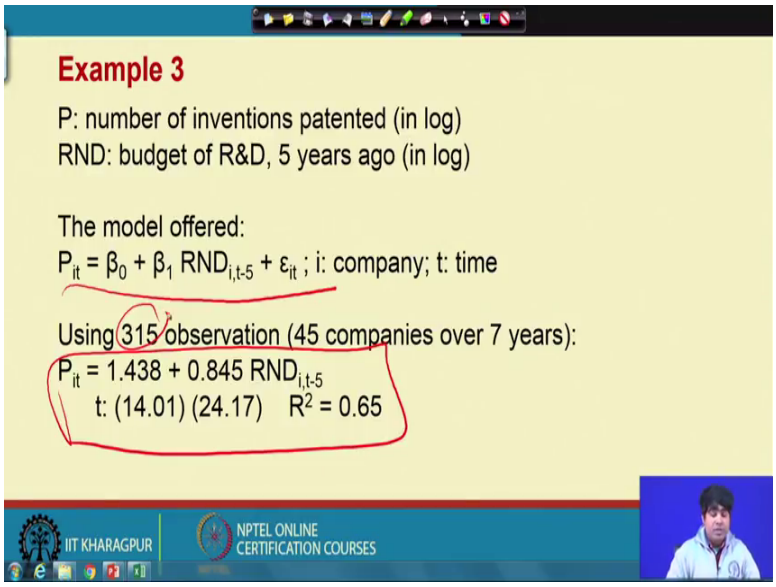
And the second second second company impact will be. So, this much of you know 2,385 will be deducted and in the case of you know company 3 2,315 will be deducted. So, as a result as a result so the for you know for the first or the company which is more effective will be definitely this ones and then followed by this one then this one and then this one. So, this is how the kind of you know you know structure which you can actually have in this analysis to justify that you know at a particular level of output which particular company is more efficient compared to you know other company.

(Refer Slide Time: 07:56)

So, as a result you know you know panel data can help you lot to find out the solution not only to you know taking the advantage of you know sample size, but also it is a effective to address some of the side objectives like he here we have studied the, you know you know kind of you know cost efficiency similarly there is another example. So, the relationship between R and D expenditure and the number of products patented. So, again so this is actually a panel data structure and here to address these problems we have taken 45 companies and that to 7 different years.

So, here actually companies very high. So, that is why if you use actually fixed effect model then you know we allow time you know should be kind of you know factor which time period the relationship is more effective, but form wise impart will be little bit complicated.

(Refer Slide Time: 08:59)

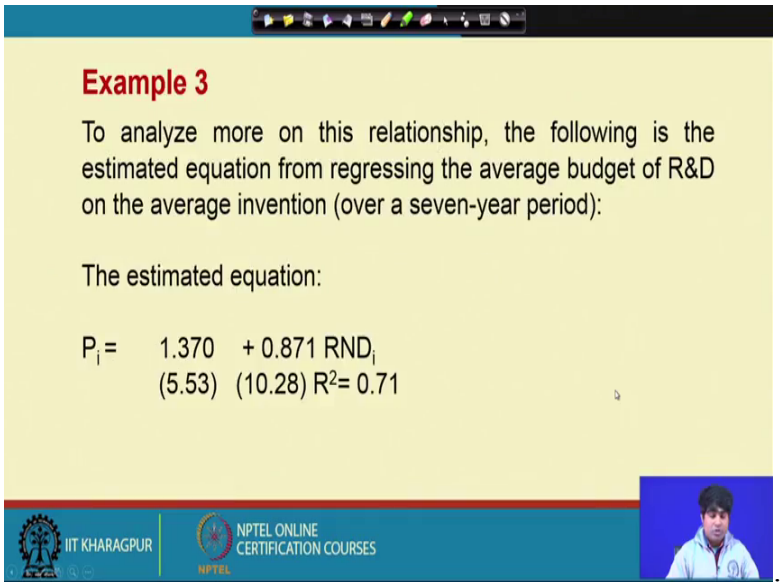


So, as a result in this case we can actually go for you know random effect model, because you here the you know the kind of you know times in the cross sectional units are you know very high.

So, as a result we will see how is the kind of you know output. So, again you know with the estimation process the a output to the first MO the first SD model a kind of you know representation and the estimated output will be like this and this is what the total observations again. So, the output is showing that you know there is a significant positive impact and that to it is the kind of you knows structure called as you know random effect models.

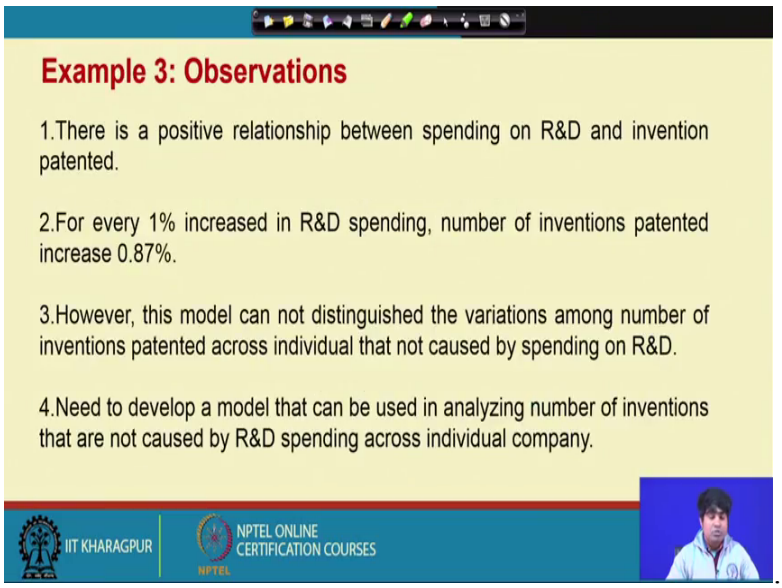
As a result so here the a there is no kind of you know individual you know organizational impact.

(Refer Slide Time: 09:56)



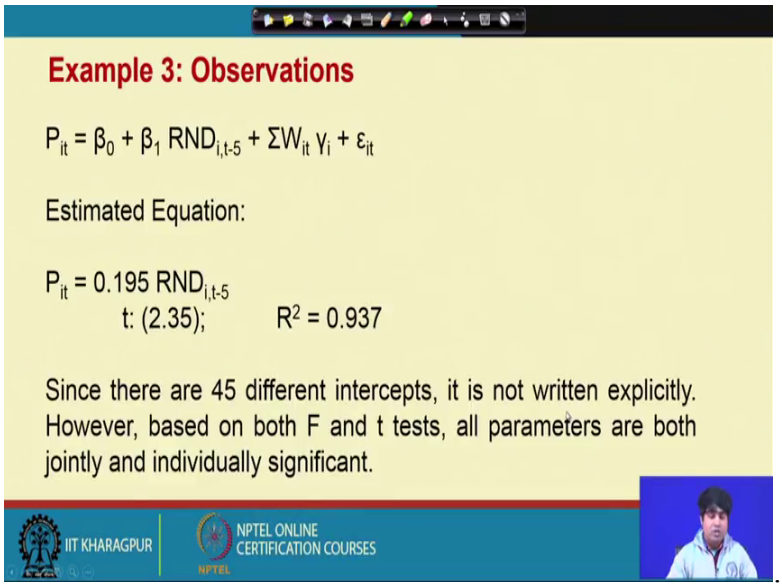
So, likewise you can also actually extend this model and you will find the particular structure again to analyze the relationship between these 2 R and D expenditure and the kind of you know patents. So, the particular model can be represented in a different format and again. So, that is actually with respect to kind of you know kind of you know individual time periods and again. So, will go for you know averaging the kind of you know structure and then with the estimated estimate again. So, this is also saying actually significant impact and the impact is also coming positive.

(Refer Slide Time: 10:28)



So, now; that means, technically if you know you know analyze. So, then you know it will be show that you know in a both the case there is actually positive relationship between you know R and D expenditure and the kind of you know innovation.

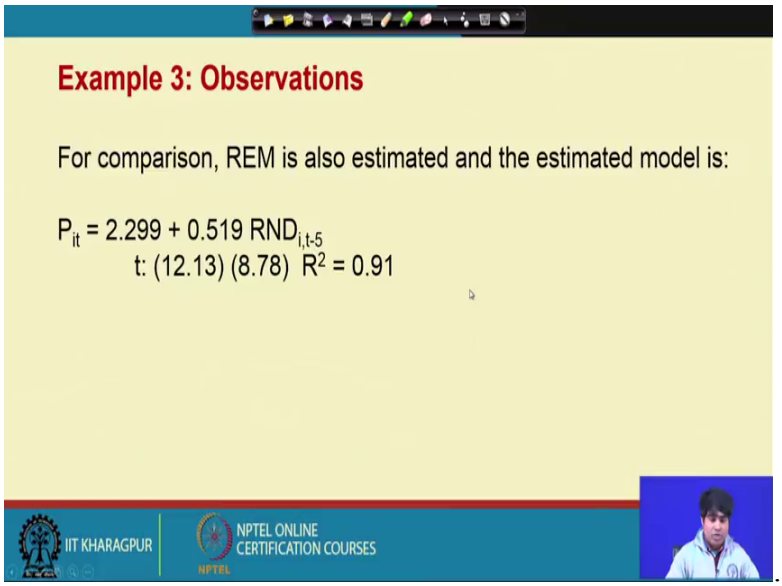
(Refer Slide Time: 10:47)



So, now I mean say it means it gives the signal that you know this is a kind of you know standard technique through which you can actually study the particular you know impact. And again if you go by dummy kind of you know structure then again. So, this is the dummy part we can you know integrate again re estimate the process and then you will check whether there is actually significant impact or not right.

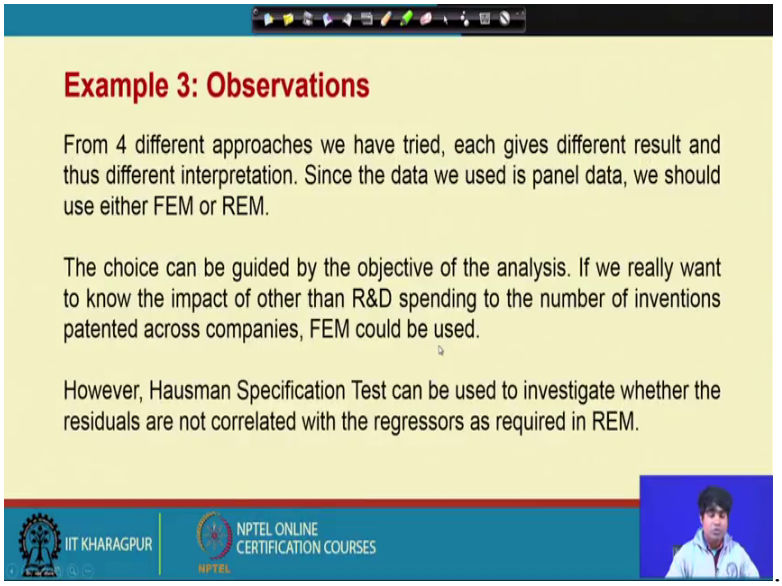
So, in this case we are not actually ah writing in details, but this will be actually subsequently come; that means, actually you have to format the dummy structure again then you have to re estimate the particular you know process all together.

(Refer Slide Time: 11:21)



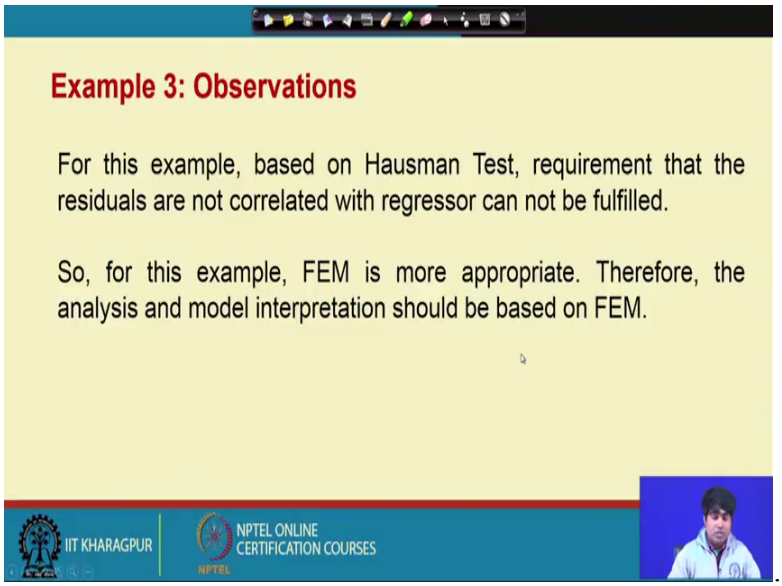
So, again actually this is another way we have to represent and then again we are finding you know the particular impact is coming significant and R square is also improving.

(Refer Slide Time: 11:35)



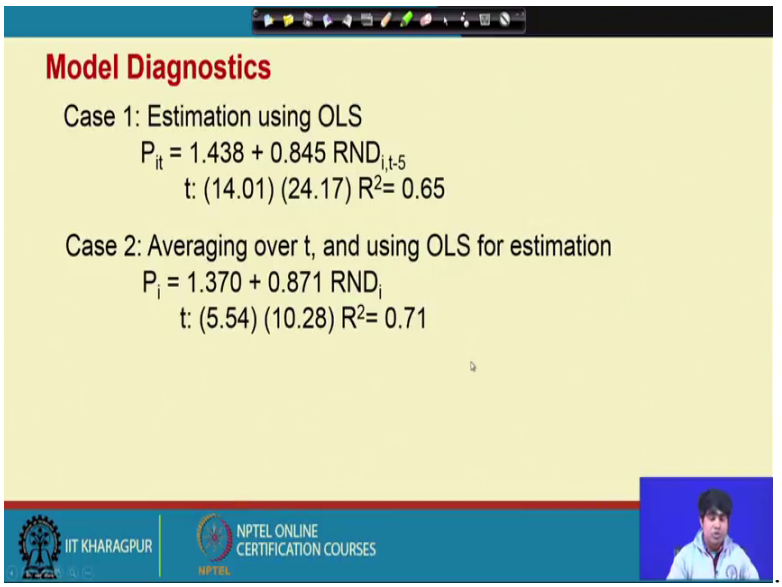
So, likewise you know there are many different you know ways you can actually check the kind of you know linkage and you know then we will see a which particular you know linkage is more effective, while addressing the link between r and d expenditure and the kind of you know innovations right.

(Refer Slide Time: 11:54)



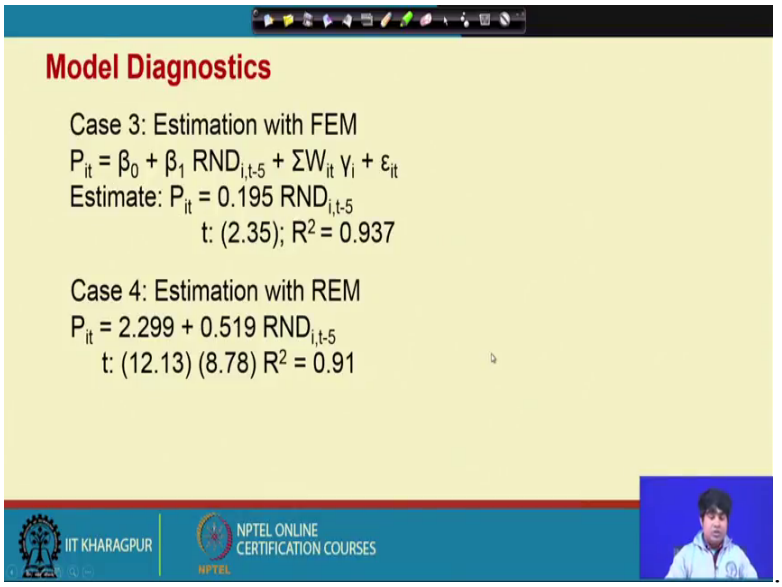
So, likewise there are so many examples. So, typically so what I mentioned earlier. So, out of you know several models which particular model will be effective, whether you will go for you know means in this case whether you will go for you know random effect model or fixed effect model. So, obviously, the houseman test can help you to find out which one is the best fit as per the kind of you know problem requirement right.

(Refer Slide Time: 12:19)



So, these are the various model diagnostics. So, we have actually solved these problems or analyzed these problems with the 4 different cases and every case the these are the first 2 cases.

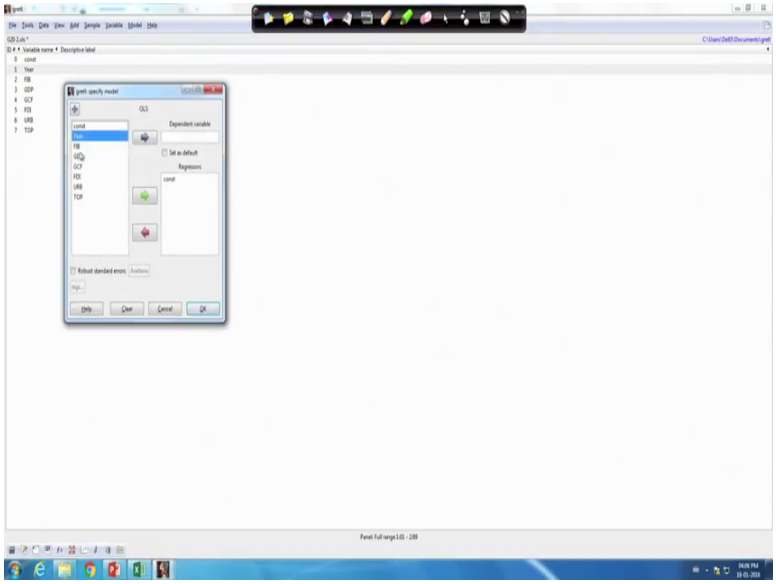
(Refer Slide Time: 12:34)



And then again these are the later 2 cases and in all the cases the main variable that is the kind of you know R and D expenditure, which is coming positive towards you know innovation. So that means, you know panel data structure is a very effective tool to highlight the particular you know issue and come with a kind of you know solution which is very effective as per the you know management requirement.

In order to highlight you know more attractive way and let us you know go to the kind of you know spreadsheet and we will serve the kind of you know problems right. So, what I will do I will take you to the software and then I will solve the problem.

(Refer Slide Time: 13:32)



So, this software and the data which we have here is actually like this. So, what will it do I will just you know collect this data to this particular, you know spreadsheet and then allow this is a we will take it to panel data and then proceed.

So, now this is actually spreadsheet and what we have here. So, this particular you know software is called as you know detail and this is actually software like r software and this is a software through which we can actually solve the a particular you know model. So, if you go to the model then we will find you know plenty of options are there you see here.

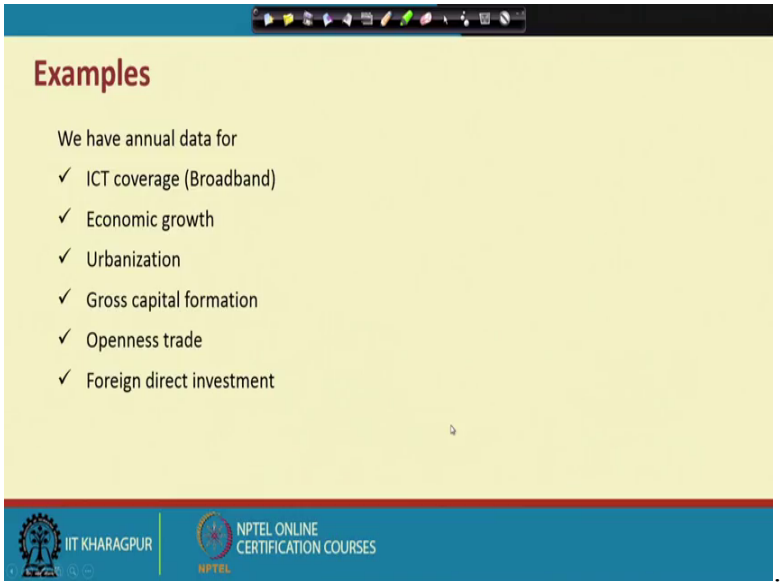
So, I am making it a bigger and till now whatever problems we have solved that is through excel spreadsheet, but typically the panel data modelling. So, excel spreadsheet will not directly help, but it will help if you actually design in the form of you know dummy modelling and then only fixed effect model can be estimated, but random effect models and other kind of you know models we which you can actually solve through excel spreadsheet.

So, that is why standard software statistical software is a highly required here in that case of you know panel data modeling, we have a vari you know various types of you know statistical software starting with R stata huge and this is actually particular software called as you know detail. So, this particular software is a you know freely available and

you can download and then you know work as per the particular you know model is concerned.

So, what will you do here in this in this you know in this particular you know problem. So, we will we will address the problem like this. So, let me give the background about this problem.

(Refer Slide Time: 15:51)

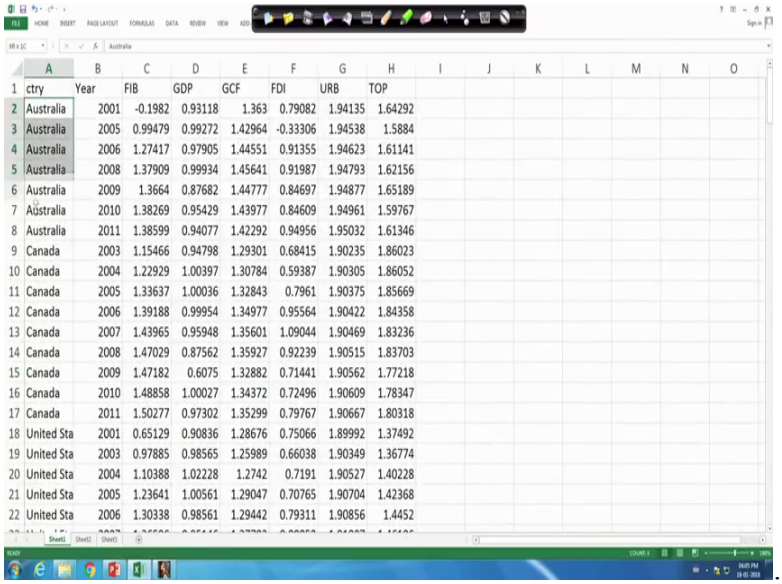


So, we have a altogether 6 variables and that to here in information and communication technology coverage that is with respect to fixed broadband. And then economic growth urbanization rate gross capital formations and trade openness and foreign direct investment this is actually macro-economic environment problem.

And here's our objective is to predict economic growth subject to all these you know independent variables like ICT coverage, urbanization, gross capital formation, trade openness, foreign direct investment. We actually since it is a macroeconomic data and the sample structure panel data structure will be with respect to time and you know country and we have taken you know country wise information's and then time wise information.

So, what is happening here? So, we will just check the view of the data. So, this is the typical view of this data.

(Refer Slide Time: 17:00)



	A	B	C	D	E	F	G	H
1	ctry	Year	FIB	GDP	GCF	FDI	URB	TOP
2	Australia	2001	-0.1982	0.93118	1.363	0.79082	1.94135	1.64292
3	Australia	2005	0.99479	0.99272	1.42964	-0.33306	1.94538	1.5884
4	Australia	2006	1.27417	0.97905	1.44551	0.91355	1.94623	1.61141
5	Australia	2008	1.37909	0.99934	1.45641	0.91987	1.94793	1.62156
6	Australia	2009	1.3664	0.87682	1.44777	0.84697	1.94877	1.65189
7	Australia	2010	1.38269	0.95429	1.43977	0.84609	1.94961	1.59767
8	Australia	2011	1.38599	0.94077	1.42292	0.94956	1.95032	1.61346
9	Canada	2003	1.15466	0.94798	1.29301	0.68415	1.90235	1.86023
10	Canada	2004	1.22929	1.00397	1.30784	0.59387	1.90305	1.86052
11	Canada	2005	1.33637	1.00036	1.32843	0.7961	1.90375	1.85669
12	Canada	2006	1.39188	0.99954	1.34977	0.95564	1.90422	1.84358
13	Canada	2007	1.43965	0.95948	1.35601	1.09044	1.90469	1.83236
14	Canada	2008	1.47029	0.87562	1.35927	0.92239	1.90515	1.83703
15	Canada	2009	1.47182	0.6075	1.32882	0.71441	1.90562	1.77218
16	Canada	2010	1.48858	1.00027	1.34372	0.72496	1.90609	1.78347
17	Canada	2011	1.50277	0.97302	1.35299	0.79767	1.90667	1.80318
18	United Sta	2001	0.65129	0.90836	1.28676	0.75066	1.89992	1.37492
19	United Sta	2003	0.97885	0.98565	1.25989	0.66038	1.90349	1.36774
20	United Sta	2004	1.10388	1.02228	1.2742	0.7191	1.90527	1.40228
21	United Sta	2005	1.23641	1.00561	1.29047	0.70765	1.90704	1.42368
22	United Sta	2006	1.30338	0.98561	1.29442	0.79311	1.90856	1.4452

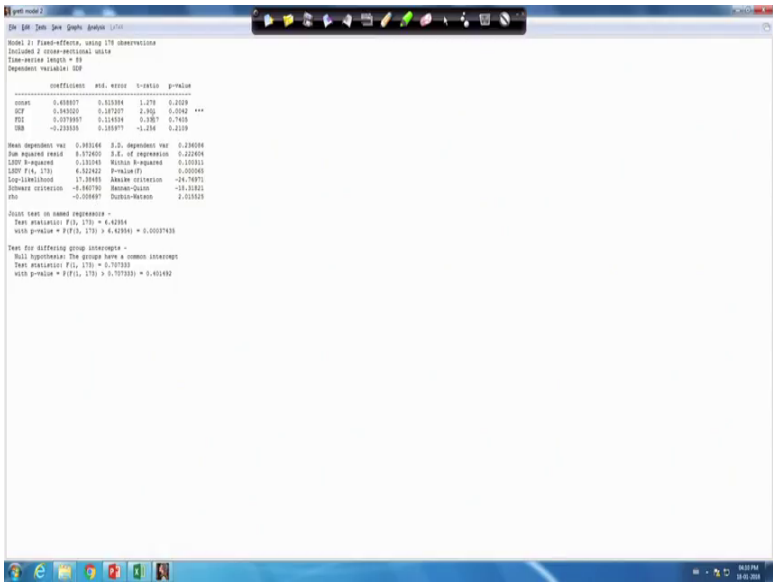
And you see here so the view of the data is actually this is the country wise information's and this is a time series information's like we have already highlighted the panel data structure where you know I can be constant T will be vary, then T can constant I will be vary. In this case you see here country; country is constant for a particular point of time, then a time is allowed to vary up to this point then again country is constant then time is you know varying one point to another point of time.

So, now once you close this then this becomes actually called as you know panel you know full data and then when we try to extract the you know cross sectional impact and time impact then it becomes a panel data structure. So, this is the data overview now we can go to the model and then check how is the you know you know typical structure altogether. So, now, if you go to the model we will have a plenty of you know options starting with you know ordinarily it is clear.

So, now if you click the ordinarily square then you allow dependent variable like you know we have already discussed in the context of in the context of you know regression analysis that to excel spreadsheet, but here we can actually it you can check this particular structure and let us see how it can you know workable and then.

So, now this is appearing since it is actually free software. So, it is creating problem. So, now, it is all loaded. So, once you put. So, this will be show you the model outcome.

(Refer Slide Time: 18:41)

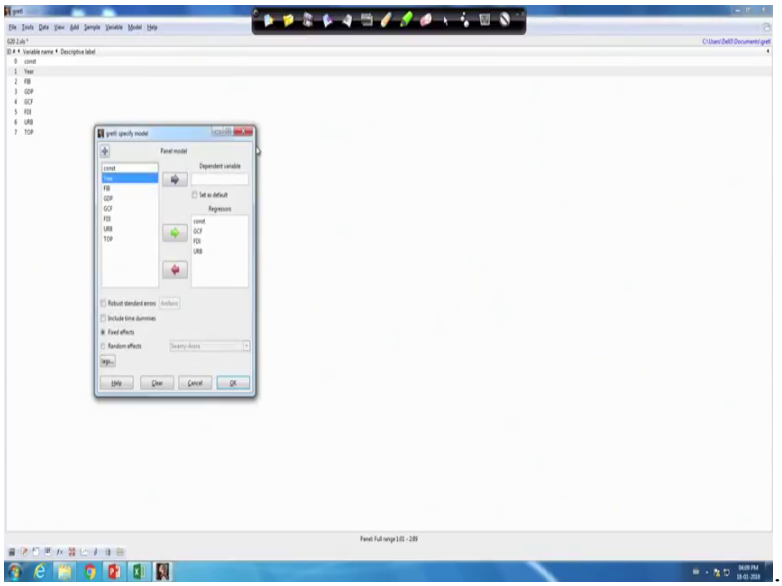


So, now growth is actually affected by these 3 variables and in both the cases these are all actually coefficients these are all coefficients. So, these are all coefficients and then this is the T coefficients and most of the cases variables are coming significant so; that means, out of 3 variables 2 variables are statistically significant and followed by a other diagnostics you know you know test statistics.

So, like you know R square adjusted R square then the sum of the diagnostics checks like you know AIC statistics SQ statistics so; that means, technically this model is showing something actually good for this particular you know predictions; however, our requirement is our requirement is it just to see what is the kind of you know you know output in the context of you know panel data.

So, now, you go to the panel data. So, you have actually plenty of you know option what I have already told you that you know we have a variety of you know model in the panel data structure. And the discussion which you have had earlier you know with respect to fixed model fixed effect model and random effect model.

(Refer Slide Time: 19:59)



So, you can click the fixed effect model and random effect model and again allow the dependent variable to enter to this particular you know process.

And then and then which a check whether the effect is coming actually statistically significant or not. So, now, putting actually this so now, we have actually option here fixed effect model and just put then the actually the output box is coming here and you see here so in the fixed effect model. So, this similar kind of you know results are coming and agains the first variable is coming significant while the second variable and third variable is not actually coming significant.

So, these are all actually a post occur test which can justify the validation of a particular model whether to accept this model or you know reject the model. So, that means, actually this particular technique will help you to you know to go for you know the kind of you know; penal data structures and then you would like to estimate the model as per the particular requirement. Again you go to the penal data structure and the kind of you know you know there is a structure called as you know (Refer Time: 21:42) squares methods through which you can also check the particular process.

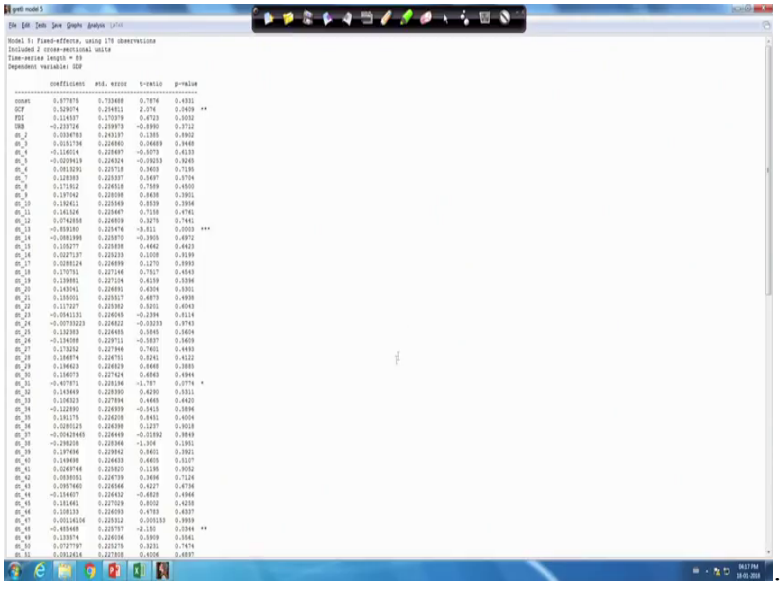
So, now, putting a case so you will get actually this is the outcome at now of weighted least square method. So, the same structure we are actually exploring and then we are studying the impact of you know economic growth by all these, you know independent variables like you likewise you we have actually different kind of you know structure

through which you can actually calculate the various procedures agains, what will you do. So, I will take you to this particular you know model then I will take I will be check you the kind of you know different dummy impact in in the process of this particular you know investigations .

Yes now if you put they nearly find you know plenty of you know option here. In fact, we like to or not the kind of you know on the dummy impact. So, what will go or what will do you will go to this particular you know process and then allow to check the dummy.

So, now, you can put you know in including dummy then you find out the kind of you know dummy impact. So, these are the various results and you know we will have a plenty of you know options here these are all called as you know different a dummy impact and in the process of this investigation about the economic growth with a different variable.

(Refer Slide Time: 27:32)



So, that means, we have actually plenty of results now and that too oh you know with respect to pull data then with respect to panel data and with respect to various dummy impact and some you know. So, that means, technically what I like to say you know justify that you know. So, the kind of you knows panel data modelling is a kind of in a sophisticated technique which cannot be actually easily handled through you know excel spreadsheet. So, we need actually standard software's through which you can actually do

the processing of the data and then connect the problem as per the particular you know requirement.

So, typically what we have actually you know we have already seen how the results are actually coming through panel data modelling that too by using pull data structure, then you know fixed effect structures, random effect structure, we tell this square mechanisms. Then we have also the particular structure called as you know dynamic panel data modeling, then and then again there is a advance particular structure called as a generalized methods of you know moments.

So, you know the all these techniques are you know hardcore techniques and with the standard software that can be solved, but the structure is actually more or less same. So, we like to try to you know investigate a dependent variable with respect to independent variables, but here the data structure will be just you know the pool of time series and cross sectional together and then we are actually trying to explore their you know relationship and then we are thinking about the kind of you know prediction it is not something actually a difficult task this particular you know process, but here the idea is the idea is that you know when we are dealing the same problem with the only time series data or cross sectional data. So, sample size is not. So, you know strong enough.

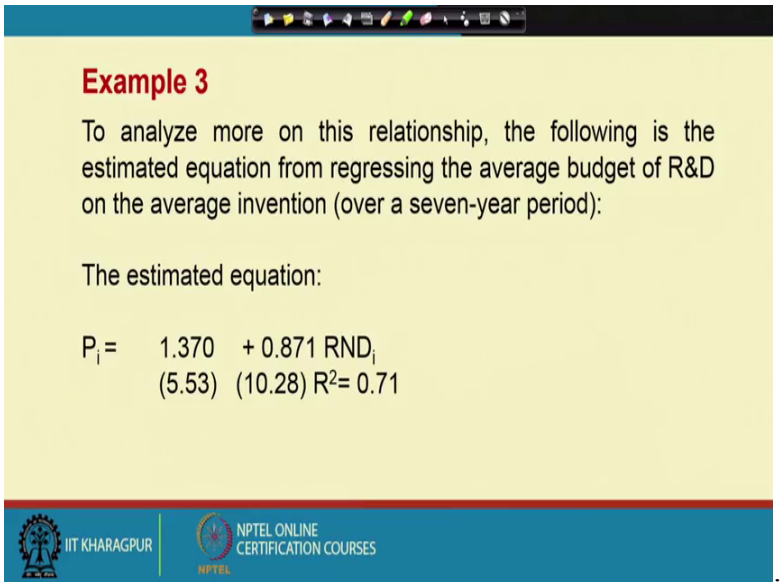
So, this is the first you know obstacles through which we are looking for the panel data structure and the panel data by default will give increase the sample size and agains while dealing with the particular you know panel data structure through addressing the particular, you know problem we can actually simultaneously study the individual cross sectional impact and individual time series impact, but when you have actually big data and you know be big cross sectional units and big time series unit. So, the individual cross sectional impact and individual time series impact is not so important.

So, what is actually important is the effect of you knows independent variables on dependent variable. So, the requirement is actually to understand the particular you know data structure and pick up the problems that need to be investigated and then you adjust the data as per the particular you know estimation requirement, whether it is actually with respect to pool data structure or panel data structure, but impact actually when we are dealing with you know excel spreadsheet. So, you just you know connect the data structure as per the panel requirement only then you take this data to what standard

software and then software will help you to you know give the kind of you know statistical output.

So, once you will get the statistical output the standard discuss energy as per the particular recurrence so; that means, the problem which you have already discussed here the particular case these are all actually estimated through kind of you know panel data process, but only thing actually if it is a small kind of you know structure then it is easily a you know handle through kind of an excel spreadsheet that to only you know fixed effect model, but for advanced kind of you know structure. So, you are bound to go to the kind of you know advanced software's.

(Refer Slide Time: 31:53)



And ultimately the idea we need actually statistical output through which you can you know address the problem and then come out with a kind of you know solution, which can help you a lot for the management requirement right. So, that means, technically till now whatever we have discussed you know starting with you know dummy structure then the kind of you know the use of you know cross sectional data the use of you know time series data then finally, the panel data structure.

So, these are all actually various kind of you know predictive analytics you know tools to tools through which you can actually solve some of the business problems where every time the idea is to predict the dependent variable with respect to independent variables. In fact, in between actually in the panel data structure so many additional tools are there,

what ultimately depending upon your you know problem and the kind of you know objectives we can actually connect and then we will go for the estimation.

So, ultimately end of the day the model should be a very effective and free from all kinds of you know error starting with you know multicollinearity autocorrelation heteroscedasticity and all other diagnostic checks sometimes you know, when we are dealing with actually a panel data more modelling that to the use of more cross sectional observations times these observations, then there is a lag variables involvement that is the concept called as you know dynamic panel data and in that case. So, this choice of lag length is also a also you know an issue while addressing the problems and for the kind of inner predictions.

So, what is the kind of you know structure of you know choice of you know lag length and how you have to optimize all these structures. So, these are already you know we can discussed in details and that to in the next lecture while we are you know addressing time series you know modelling in a kind of you know different approach with this we will stop here.

Thank you very much have a nice time.