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

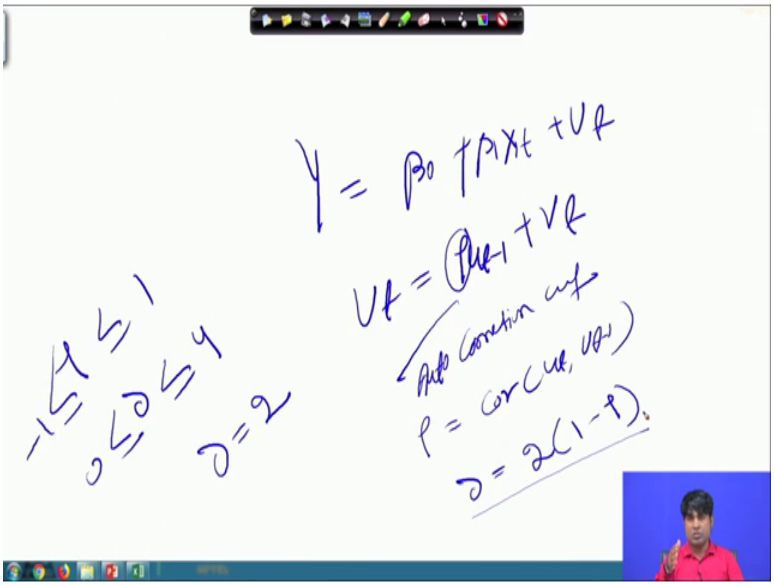
Lecture – 28
Autocorrelation problem (Contd.)

Hello everybody. This is Rudra Pradhan here. Welcome to Engineering Econometrics. Today we will continue with modelling diagnostics and that too autocorrelation problem. In the last lecture we have discussed this particular component, basically it is the problem among the error terms. As we have already discussed in a kind of you know econometric modelling we have dependent variable and we have independent variable explained and independent variable unexplained. So, autocorrelation is a game of unexplained independent variable that too among the error terms.

So, like we have discussed the issue of multi collinearity which the which is the game of the relationship between independent variables explained and here it is the similar kind of you know structure. But we like to establish here the relationship among the unexplained independent variable that to with respect to error variables only. And in fact, we have discussed in details how we can derive this particular you know auto correlations and we have discussed various regions, various you know mechanisms through which you can detect the autocorrelation. And we have also discussed the kind of you know tolerance reliables and the kind of you know consequence. And the then the thing which we like to discuss today is the a solution part and in before that we can discuss a particular problem and we like to check how the particular you know autocorrelation problem and what is the degree of involvement.

There are there are lots of you know techniques through which we can actually get to know the a degree of a autocorrelation in a system. For instance simply in graphical plotting or simply a you know correlating between 2 independent variable we can get to know the involvement of you know autocorrelation problem in a kind of you know estimated model. But the standard techniques through which we can actually you know detect the process is called as Durbin Watson d statistic that is otherwise called as a DWD statistics. Most of the statistical software and econometrics softwares used to provide this particular statistic. And we have already discussed in details in the last lecture.

(Refer Slide Time: 03:14)



And the usual structure is that we if you start with like this a the particular structure is a Y equal to beta 0 beta 1 X 1 t plus U t and then u t equal to rho u t minus 1 rho u t minus 1 plus b t. So, rho is the autocorrelation coefficient; autocorrelation coefficient and then rho is nothing but actually a correlation between u t and U t minus 1.

So, this is what we have discussed in the last lecture. Now, if you connect with Durbin Watson d statistics then it is nothing but d equal to 2 into 1 minus rho. Of course, there is a mathematical procedure how to obtain this d component, but ultimately after the simplification you can get d equal to 2 into 1 minus rho and rho is the autocorrelation coefficient. Since a autocorrelation coefficient lying between minus 1 to 1, so d will be ranging from 0 to 4 so that means, a rho will be varying from minus 1 to 1 and a Durbin Watson d statistic will be ranging from 0 to 4.

So, ideal range is the for autocorrelation and that too good modelling for any kind of you know decision making process where d must be equal to 2. So that means, technically if you have a rho equal to 0 then there is no autocorrelation and in that context by default d equal to 2, and if rho not equal to 0 then it may have a negative value it may have a positive value and 1 extreme is minus 1 another extreme is a plus 1. So, as a result, so

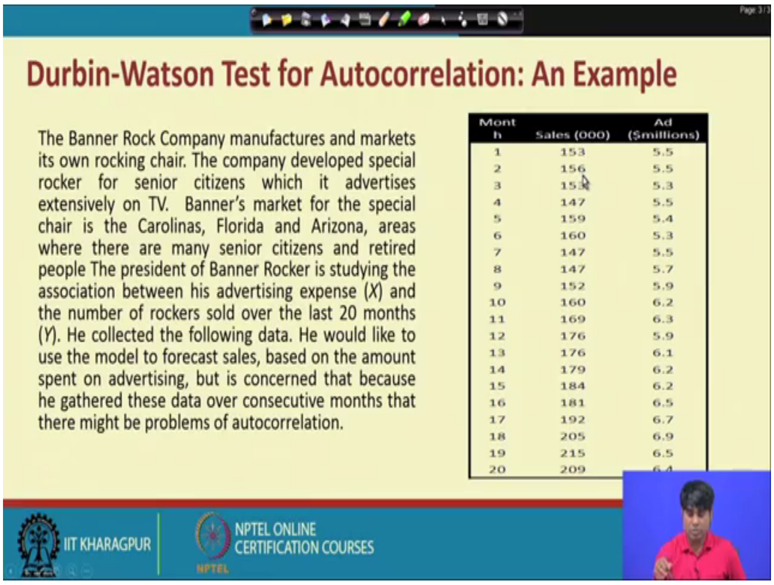
DW statistic the extreme ranges are 0 and 4 by putting rho equal to minus 1 and plus 1 respectively.

So, ultimately we like to check what is the value of d in a kind of you know estimated model, whether it is a ideal in nature or it is a kind of you know problematic in nature. So, we have given the confidence interval and that means, technically a d must be ranging between 1.5 to 2.5 that is the acceptable kind of you know criteria through which you can go ahead with the a modelling and the kind of you know decision making process.

Otherwise to getting the ideal value d equal to 2 here rho equal to 0 is very difficult 1 and as a results it may be lying in between 1.5 to 2.5. And if it is he exceedings you know less than 1.5 or more than say 2.5 then slowly slowly it will be actually you know dangerous or any kind of you know modelling process. So, if it is crossing you know 2.5 or lesser than 1.5 in that case we need actually restructuring of the model then we will go for you know freshman you know estimations. Again we will have a estimated model, then we will again follow the same procedure of checking the autocorrelation and then we can compare the previous autocorrelation value and the later autocorrelation value. And we can actually move to the second models or second iterative process if that particular DW statistic with a lesser or in a kind of you know ideal range compared to the previous one.

So, that is how it is a iterative process and in fact, it is a continuous process until unless you get a model which is free from auto correlations or which must be in a kind you know ideal range between 1.5 to 2.5. So, now, in this lectures we will like to highlight two things we first solve a problems and like to know how to get this a DW statistic you know through some kind of you know software and then we like to check the solution part of this autocorrelation problems and other major issues relating to the autocorrelation problem. But it is a big virus until unless you take care this particular part we should not use the estimated model or any kind of you know decision making process and that to solve any kind of you know engineering problems.

(Refer Slide Time: 07:48)



So, let us see how is this particular you know look. So, we have a problem here, and the problem is for a manufacturing company and a in the manufacturing company the process is actually there will be a there will be having pores process and they will go with you know kind of you know output.

And ultimately having output they will be they will be send it to the market for selling and a when there is a question of selling of their product in the market then management come into the picture. And obviously, there will be have a different kind of you know planning or the kind of you know management so that these sales should be at the highest level. And for that there are many factors and the particular company has to take care all these factors you know in order to achieve their goals to increase these sales and then increase their you know production process.

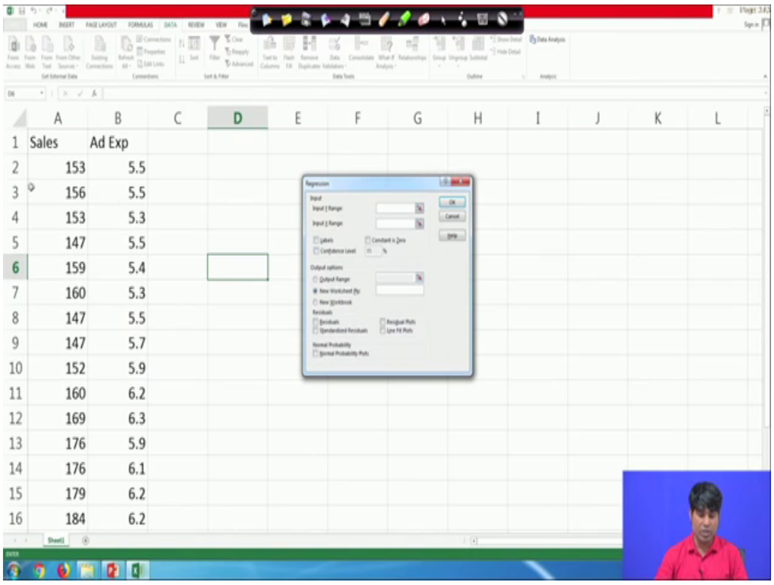
So, number of factors will be involved a in this process that to how to increase the sales of a particular company, a starting with you know price of the particular you know product and then the advertising expenditure, infrastructure of availability like so many factors will be there. So, you can go through theory of this particular you know problem then find out what are the factors main factors responsible for this particular you know product and that too for that company, because sometimes factors varying from product to product and factors varying from company to company with a particular product.

So, there are many different issues all together there. So, what will you do? We will you just to have this particular you know data and check whether the particular factor is very effective for this you know sales forecasting. And one of the factor which can usually helpful for self forecasting or production forecasting is the advertising expenditure with the theoretical you know understanding is there. So, if you put you know more and more advertising then people will get you know get to know about the product and a about the value of this particular product and then there is a high chance there will be more and more sales.

So that means, technically in the theoretical hypothesis that more is the advertising you know expenditure more is the kind of inner sales and less is the kind of you know advertising expenditure the less is the kind of you know sales. So, we like to check whether this particular statement is a true or false and for that we are using econometric modelling. And then we simply start with the a you know modelling structure is that sales is a function of you know advertising expenditure and they are linearly related to each other and there is a you know positive relationship between the two and as a result the beta coefficient should be positive and that should be statistically also significant.

So, having these problems for this company and the particular you know items like you know this side, this side, this side. So, we have actually sales value and then we have sales value and then will you will they have a actual advertising expenditure then we like to correlate and regress get to know how is the particular.

(Refer Slide Time: 11:07)



Furthest, for this we will go to this you know problem set. So, the same data set I have here actually you know loaded and to I will just increase the view. So, this is what the or this is what this sales data and the first row belongs to sales data, second row belongs to advertising expenditure.

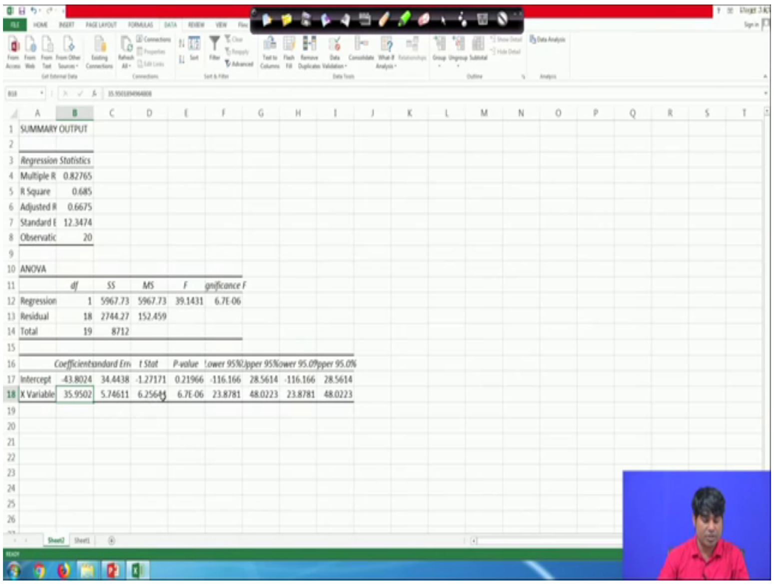
We can solve this problem in any software, but I am now using the excel software and because I would like to know how we can derive this you know Durbin Watson d statistics. So, this is obviously, faster the structure is that you know we like to regress sales with advertising expenditure, for that we can pick up the data analysis package and since it is a regression. So, just you click the regression and then you can give first a Y range that to starting with you know first item on 153, 153 to last item 209.

And then read the input range that is first item is the 5.5 and that to continue with the last item will be 6.4, and that to what both the variables have same samples that say 20 data points and a both having actually some kind of you know fluctuations. As a result, but the both variables informations are reliables and consistent to you know regress the particular you know equation and that to fit the model as for the particular you know problem is concerned.

So, now, giving the dependent variable range an independent variable range, now your software ready to operate and process and that to it can give you the kind of you know estimated output. But before that I would like to highlight that you know autocorrelation is a problem it may be a means if the model may be bivariate, may be trivariate, may be more you know multiple or multivariate the this problem will be always there. So, this is a genuine problem. And why I have highlighted this, because in the case of multicolinearity and there should be actually a you know multiple independent variables and then we can apply this particular structure.

But here even if you have a two variable case Y and single independent variable X still we can derive error component and create further error term and to regress and then go ahead with the auto correlations. Now, with this particular problem we will have this kind of you know structure.

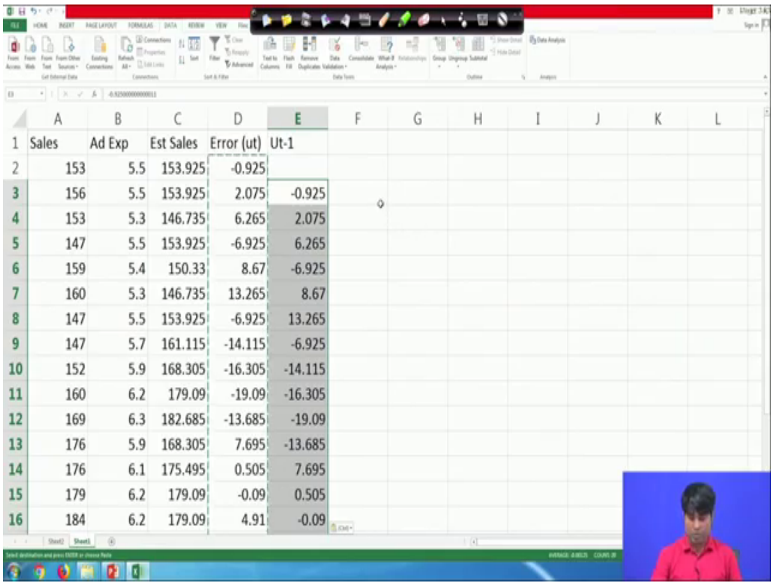
(Refer Slide Time: 13:51)



Let us put you know and then we will have actually estimated output this way. And then here as usual the slope coefficient is coming positive and statistically significant and as a result it is going in favour of our hypothesis; and what will it do. So, we like to check the a you know level of you know auto correlations for that the process is not actually ready. That means, technically we have an estimated coefficient standard errors then t statistic probability value and then r square value, then adjusted r square value, f value everything is actually you know in favour of our modelling.

So that means, this data and this model is very much you know effective so far as s of t and g, of t g concerned. So, now, I will means will go to digitive part where will it take care of the autocorrelation issue in that case we can go to the sheet 1. Here in the sheet 2, so the coefficient is actually minus 43.8 then 35.95.

(Refer Slide Time: 15:10)



	Sales	Ad Exp	Est Sales	Error (ut)	Ut-1
2	153	5.5	153.925	-0.925	
3	156	5.5	153.925	2.075	-0.925
4	153	5.3	146.735	6.265	2.075
5	147	5.5	153.925	-6.925	6.265
6	159	5.4	150.33	8.67	-6.925
7	160	5.3	146.735	13.265	8.67
8	147	5.5	153.925	-6.925	13.265
9	147	5.7	161.115	-14.115	-6.925
10	152	5.9	168.305	-16.305	-14.115
11	160	6.2	179.09	-19.09	-16.305
12	169	6.3	182.685	-13.685	-19.09
13	176	5.9	168.305	7.695	-13.685
14	176	6.1	175.495	0.505	7.695
15	179	6.2	179.09	-0.09	0.505
16	184	6.2	179.09	4.91	-0.09

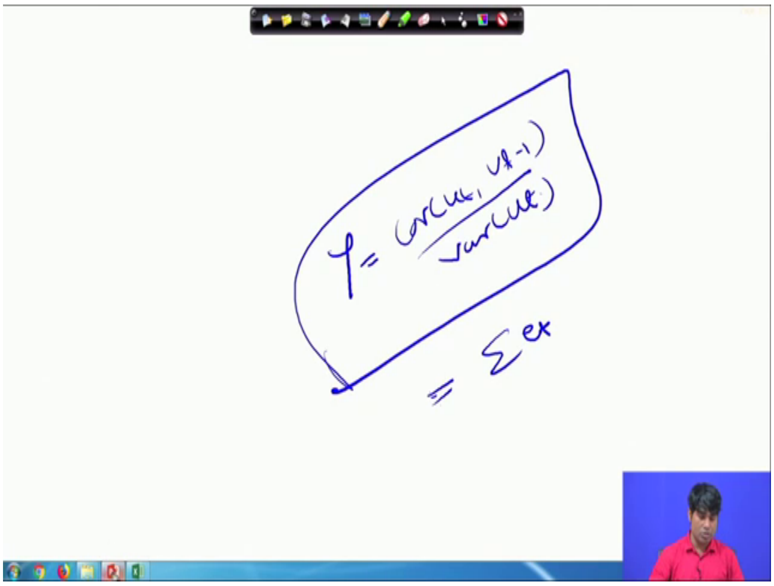
So, again so what will you do you recreate here you know estimated you know sales; let us say estimated itself now what is happening. So, this is equal to simply a minus 43.8 so that means, technically we move to here minus 43.8. So, this is equal to minus 43.8 a plus plus 35.9, 35.9 into s this is 35.95 into s. So, this will give you some kind of you know forecasted figure.

Now, what we can do you just actually scroll it you just scroll it then you will get the kind of you know. So, this is what the estimated output. And then after having the estimated output what we can do. So, we like to check the error component, then we will derive the error variable here. So, error is the difference between the actual sales and the forecasted sale that is the a predicted sales and this way. So, you will get the error component. So, this is what the error variable behaviours now, this is what is called as actually u t.

So, next in the next instance you can create actually U t minus 1 and for that. So, you just copy this particular you know variable you can copy this particular variable. So, this is all right. So, you (Refer Time: 17:43), this is all right.

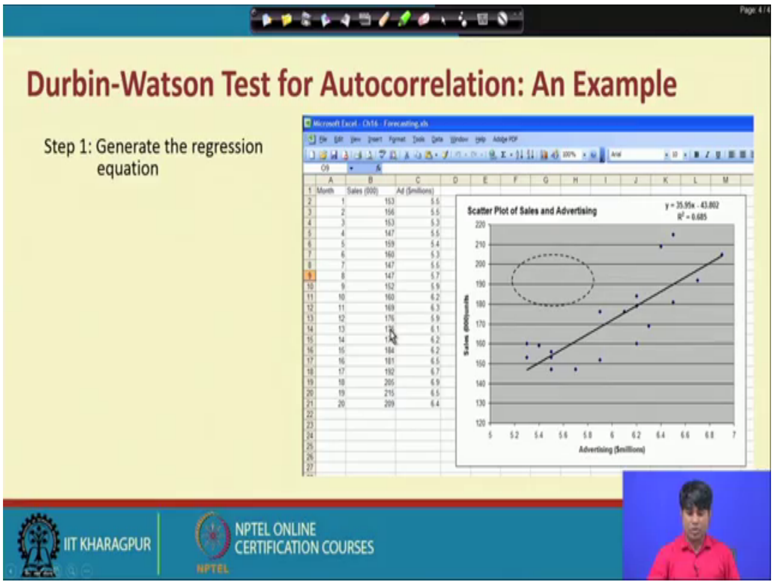
So, now we can copy this particularly you know variable then we can just paste 1 point behind and a that too we can have with you know like this it will give you this is. So, now, this should be like this. So, now, this is U t minus 1 and then our autocorrelation coefficient will be it will be autocorrelation coefficient will be a in between actually so the autocorrelation coefficient here it will be autocorrelation variable here will be covariance of a u t upon u t minus 1 divided by variance of you know u t.

(Refer Slide Time: 18:42)



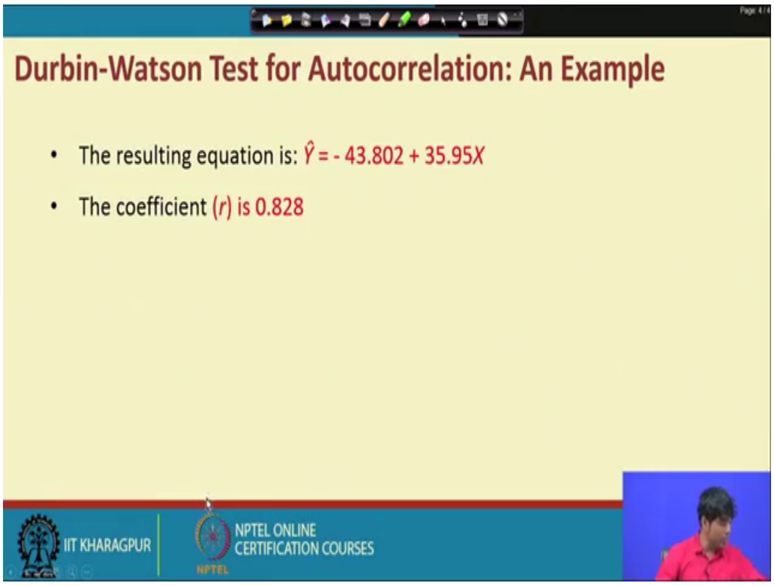
So, this is your d autocorrelation coefficient so that means, if you simplify then you know it will give you something like you know summation e t, if you simplify then it will give you some kind of you know the covariance between u t t minus 1 that is nothing but you know e t t minus 1 by a sigma square e t. So, it is better to come here and see the kind of you know structure. So, all right so, let me complete or this you know this particular you know problem. So, this is what the kind of you know case.

(Refer Slide Time: 19:31)



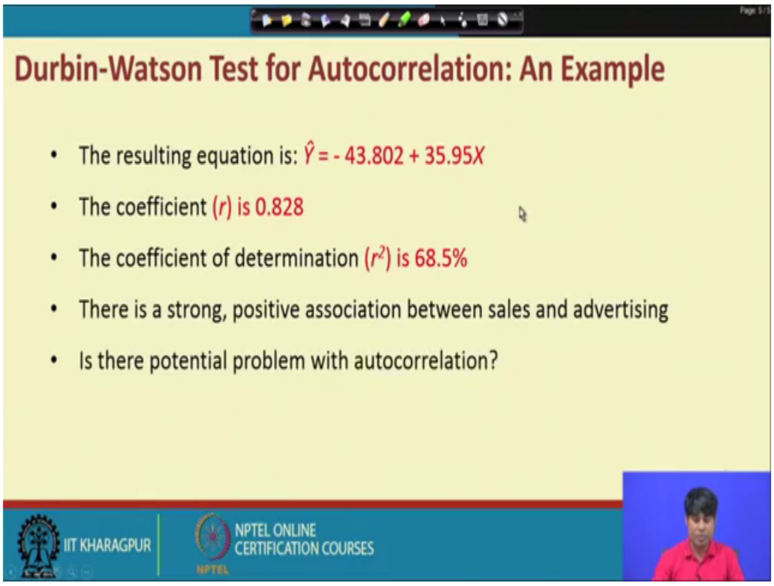
So, now we have actually monthly data up 20 data per month, 20 different months, then sales figures, then advertising expenditure and this is the plotting. And that to know the link between you know sales and advertising expenditure and after that we are looking for actually the Durbin Watson d statistics.

(Refer Slide Time: 19:49)



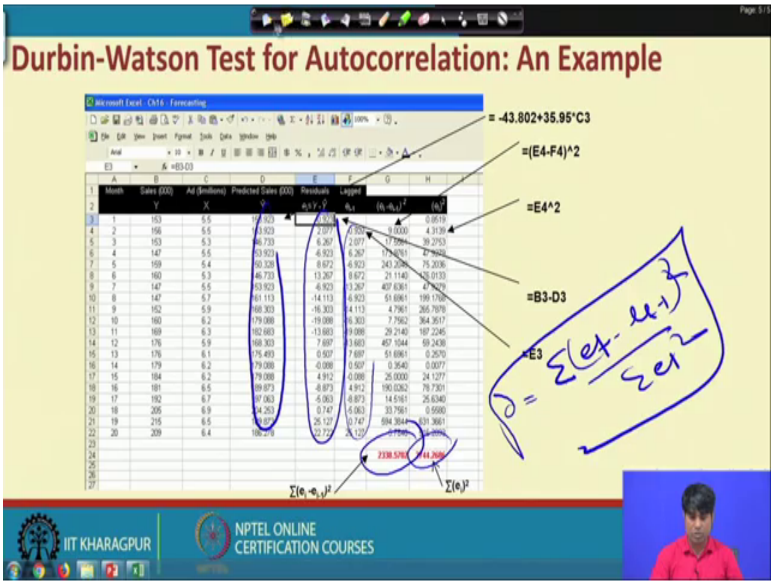
So, the estimated equation would be Y hat equal to minus 43.80 and plus 35.9 X which which we have already highlighted here from the software. So, that is the software result here 43.8 and 39.5. And then we have the error component and the lag of the error component, these are the typical inputs in this particular you know autocorrelation detection.

(Refer Slide Time: 20:22)



So, now so the estimated equation is the Y equal to minus 43.8 and this one, and then coefficient is r square 0.83. So, now, with this actually we like to proceed for the auto correlations, that is why we have a created actually the kind of you know error components that to here the Y hat which you have already derived.

(Refer Slide Time: 20:49)

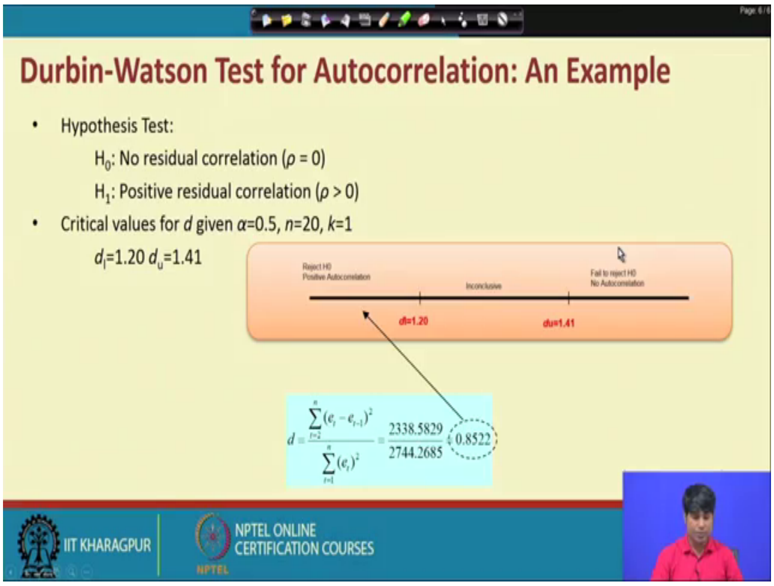


And that to error component is this much which you have also arrived, and then what we will do here. So, we will create a lag of the error terms this row and a this is what the estimated equation and this is how the error terms. Just now I have highlighted in the excel sheet and then we can create you know lag of the error term. And the Durbin Watson d statistic a value will be a there will be actually. So, the a the division between actually the square of this you know covariance and then and the variance of this you know error terms.

So, now we have to you have to find out the you know square of this difference and then square of the error terms and then if we will divide. So, you will get the kind of you know component of you know Durbin Watson and d statistics. So, technically, so this is how the behaviour. So, we will see how is the kind of you know structure altogether.

So, ultimately in the equation is like this and this is what the Durbin Watson d statistic testing and whatever you do here. So, after getting this particular series so you just divide this into this. So, you can get the autocorrelation coefficient. So, means technically, so it is nothing but actually a summation e t minus u t minus 1 whole square divided by summation e t square. So, that is what the Durbin Watson d statistics. And after this process you can have this value and let us see what is that particular you know value from this a equation and a for that you have to actually move further.

(Refer Slide Time: 22:57)

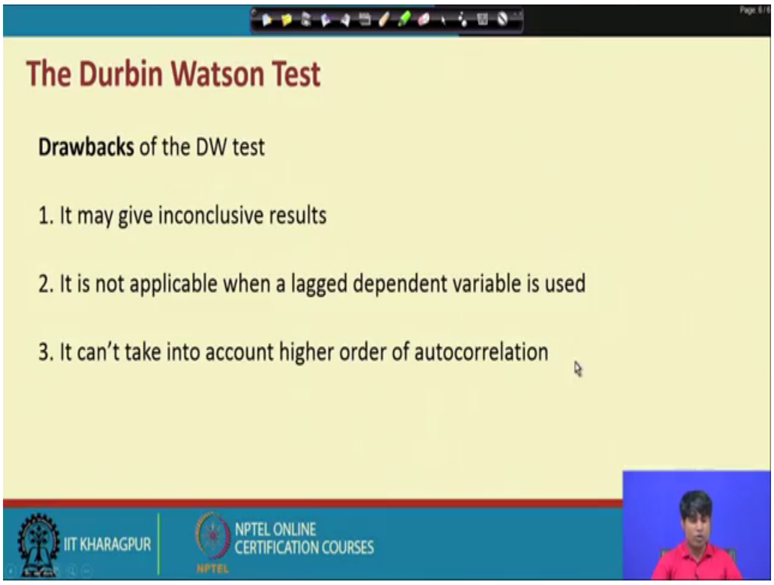


So, this is what the first of all the null hypothesis will be rho equal to 0 against rho not equal to 0, and then the critical value we will go for unit testing because we have also Durbin Watson d statistic tables. So, here we will calculate the Durbin Watson d value and then compare with you know tabulated value. If it will overtake this particular coefficient then we can say that you know auto correlation is the a present and up to it is significant, and if it is significant then it may not be used for the kind of you know modelling model decisions or you know managing decision.

So, this is what the calculating mechanisms which I have already highlighted. So, it is coming actually 0.85 so that means, if you divide these two values so it is coming 0.25 which is actually not exactly ideal because for ideal we should have actually d value equal to 2 and a against we have kept the confidence interval 1.5 to say 2.5 and here it is actually less than to 1.5 that means, say close to 1 let us say 0.85 means close to 1. And that to positive autocorrelation and as a result this particular value is not actually good for the modelling and that to for any kind of you know decision making process.

So, the value itself indicates that you know there is need of you know restructuring of the whole process and go for the reactivations. And again have the new estimated model and then repeat the process till you get the d value within the range between 1.5 to 2.5 that is what the usual procedure. And then I get accordingly we can move go ahead and a that means, a the whole process is actually iterative kind of you know things.

(Refer Slide Time: 24:45)

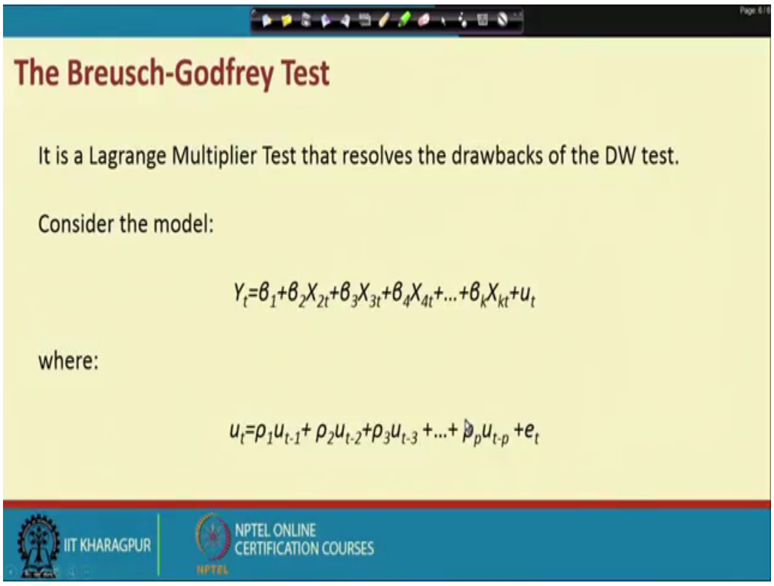


Every times you can actually check and then see the kind of you know range. If there is need of you know restructuring or the estimation go ahead with restructuring or re-estimation and repeat the process otherwise you can stop and go ahead with the prediction forecasting and the kind of you know decision making as per the requirement of the engineering problem.

And you know Durbin Watson d statistic itself has some drawbacks it may give some kind you know inconclusive results, and it is not applicable when lagged dependent variable is there. Because we whatever you know technique whichever it is still now with the intention that you know it is the dependent variable, and independent variable only that to explained, so no lag involvement.

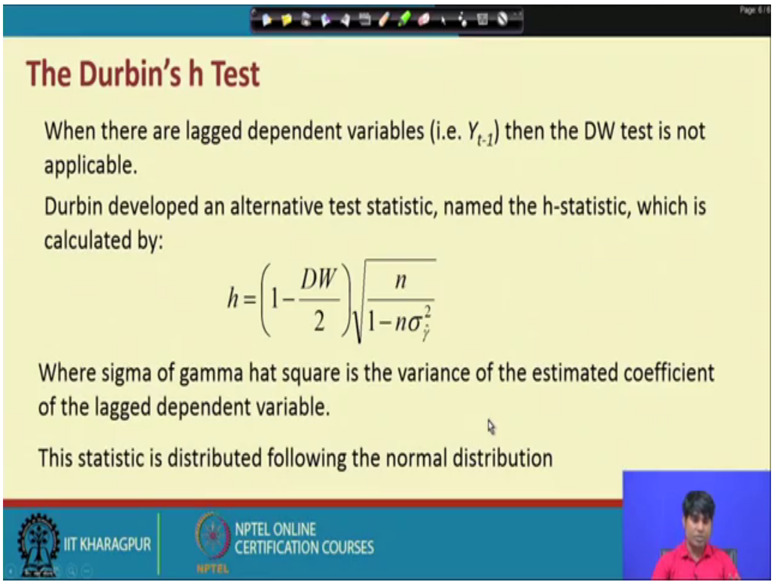
Of course, after getting the error terms we are connecting to the a lag error terms, but the starting process is that you know dependent variable with the only explained independent variable without any lag involvement. That is what is technically called as you know the endogeneity issue. But in a real life scenario and the environment of lag variable is also needed because of you know endogeneity check then in that case that is for the instead of using Durbin Watson d statistic will use Durbin Watson h statistics.

(Refer Slide Time: 26:13)



So, it will be having like this. So, I will give you the Durbin Watson h statistic structures.

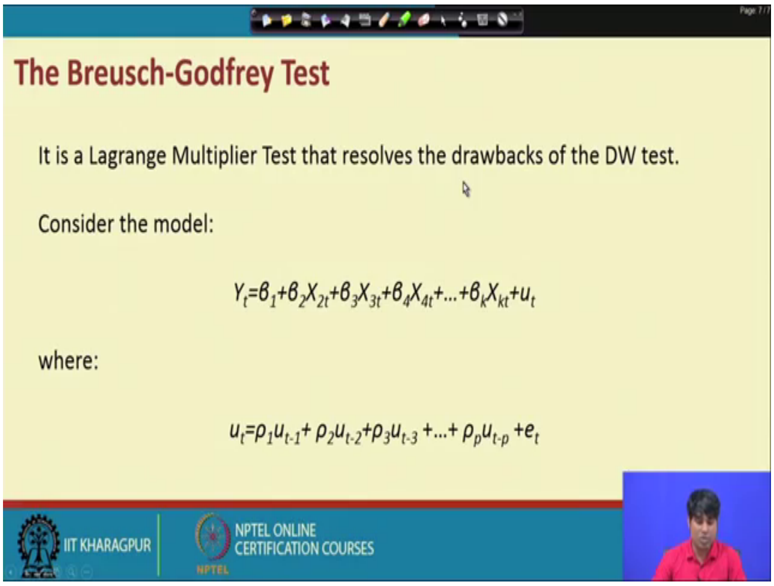
(Refer Slide Time: 26:18)



This is what actually Durbin Watson d statistic structure. So, instead of you know d we call it say h and a then; a so it is simply actually did first you negate the DW value. So, 1 minus DW by 2 then in square root of n by 1 minus n sigma square estimator.

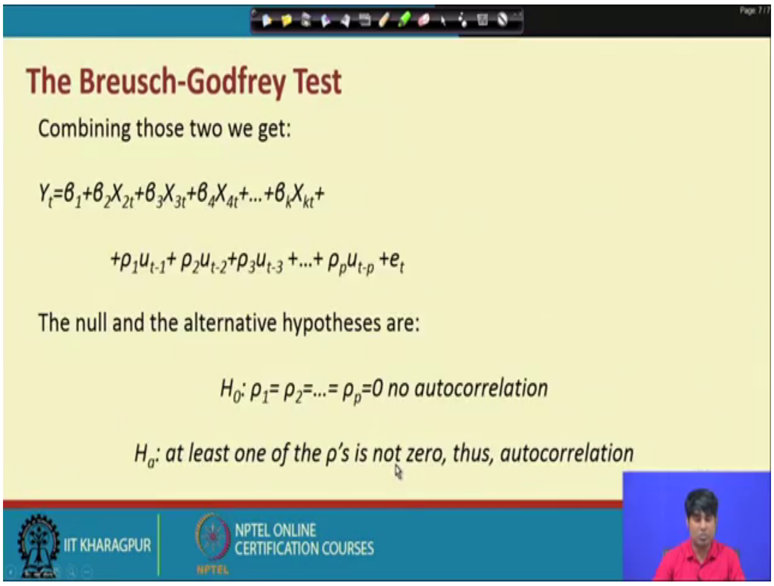
So, it will give you some kind of you know figures like you know Durbin Watson d statistic. So, we will use now Durbin Watson h statistics to represent the accuracy of the models and then use the model for the kind of you know decision making. So, this is actually normal procedure through which you can actually proceed and move.

(Refer Slide Time: 27:07)



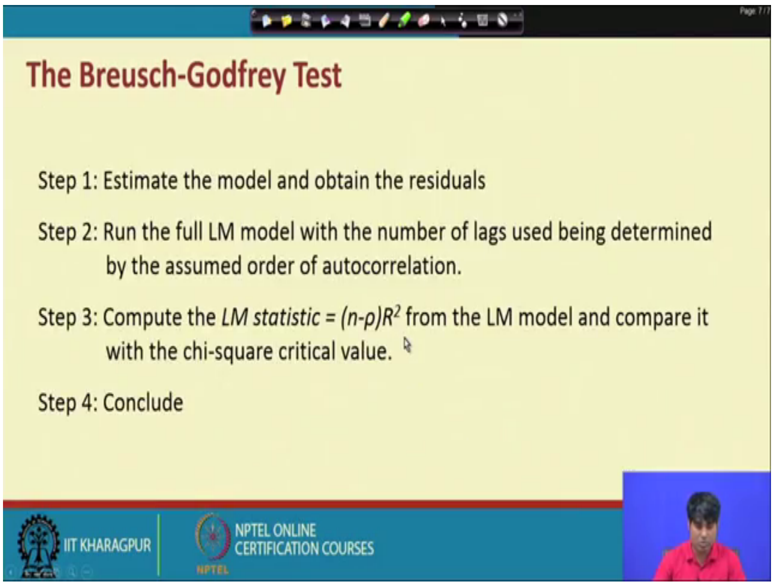
And besides we have lots of other tests like you know BG tests where you know we are regressing first you white with you know independent variables that is explained. And then we derive the error terms and then connect error term with lag of the a error terms.

(Refer Slide Time: 27:22)



And then check the significant error means just you have to derive the coefficients and check whether the coefficients are statistically significant or not. If they are structurally then there is a auto correlation, and if they are not statistically significant then there is you know autocorrelation. So, this is how the process is all about.

(Refer Slide Time: 27:36)

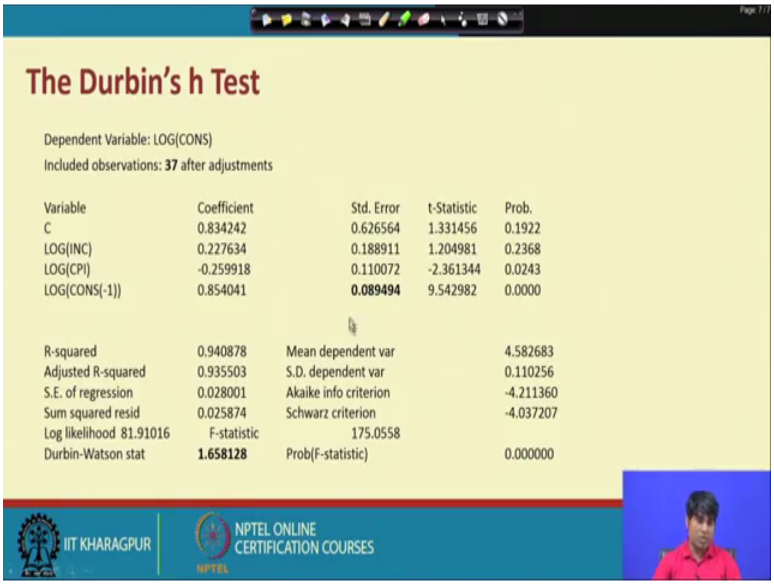


And then if you are just regressing u t upon you know log of the u t like you know U t minus 1, u t minus 2 and so on. So, in that case, so we can use actually Lagrangian multiplier test, where a the value of R square can be tested a and compare with the tabulated value. And in the final conclusion is that you know whether the particular you know h value is statistically significant or the means that is the a Durbin Watson value is statistically significant or not.

And if it is statistically significant then the particular problem is the problematic a and we have to find out the solutions that to where as either a restructuring of the entire model or the entire process or simply you know you know change the functional form and then go for re-estimations or change the variable structure again go for you know re-estimation. So that means we have a plenty of options. So, for as a solution to the Durbin Watson means autocorrelation is concerned.

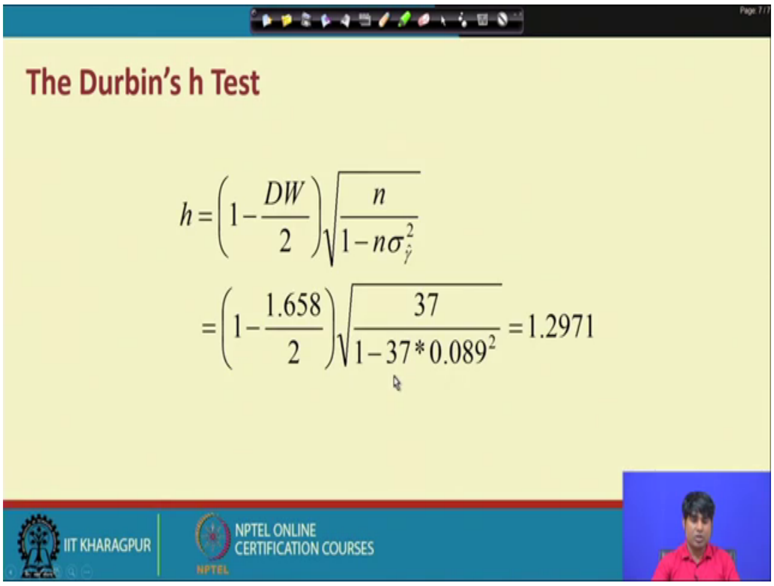
The best you know solution mechanisms are you know go for you know data transformation increase the sample size, change the functional form here and different variable sometimes you know due to miss specification of the model. So, we can have some kind of you know autocorrelation problem. So, we can use trial and error method by increasing sample size, decreasing sample size, change the functional for addition of you know new variables, then a removal of the existing variables. So, after doing all these you know kind of you know cross check. So, you define the or you know ultimately means the estimated model which is a pre from you know autocorrelation. Even if there is a auto correlation, but it will be in a ideal range that to in between 1.5 to 2.5.

(Refer Slide Time: 29:37)



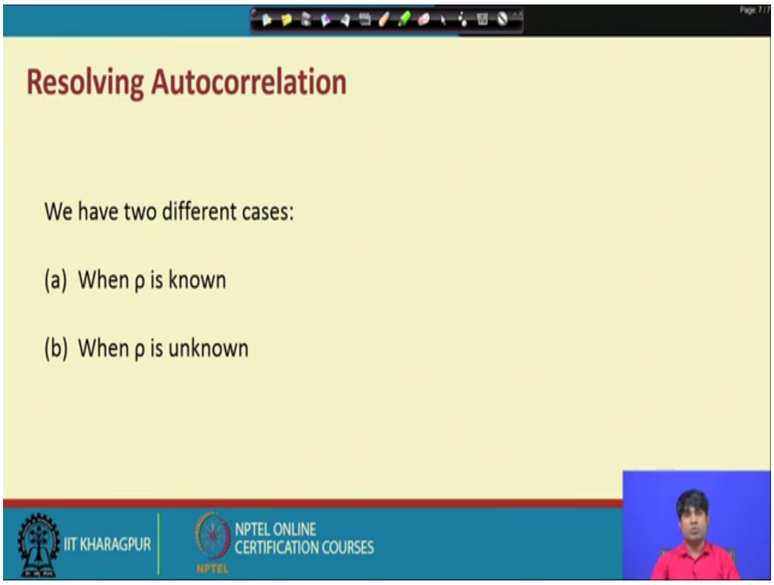
So, this is how the a process and this is the similar kind of you know problems through which actually means same problems can be connected with you know Durbin Watson h statistic instead of d statistics. So that means, we are regressing sales with you know advertising expenditure just you add you know sales of the previous years that is that is the kind of you know lag dependent variables. If you a enter to the model then the model will become a trivariate and in that case a you can use the Durbin Watson h statistics instead of you know Durbin Watson d statistic and then check the accuracy of the model.

(Refer Slide Time: 30:13)



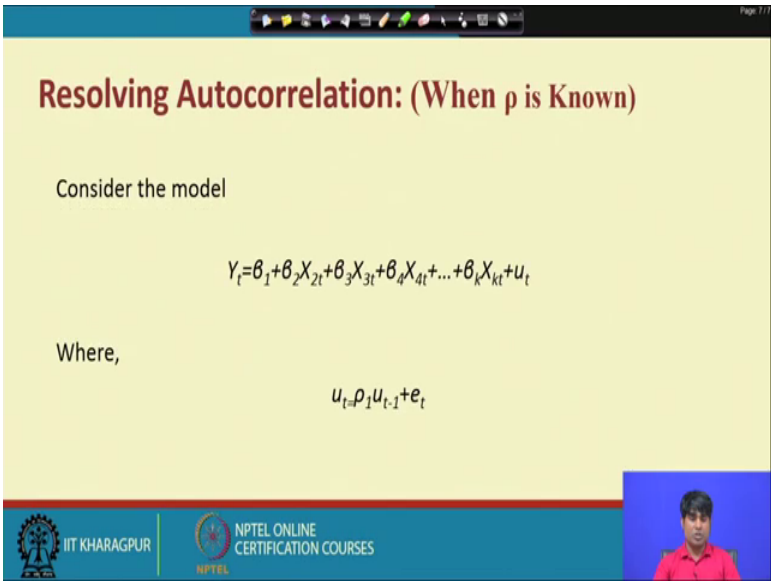
So, this is the usual procedure, but every time a the requirement is that you know you have to see a whether the model is a free from all kind of you know diagnostics and then come to a kind of you know conclusion that you know yes, this is the model through which you can go for decision making process.

(Refer Slide Time: 30:17)



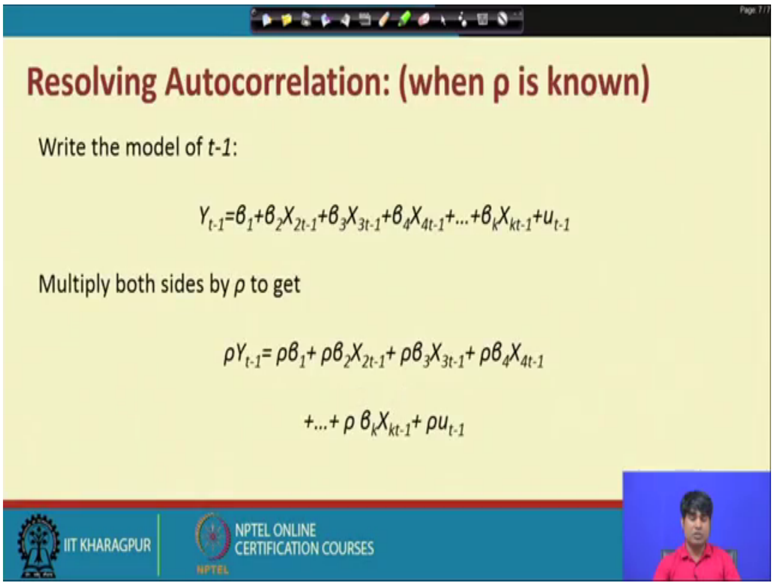
Ultimately here rho is the kind of you know component which can you know check the kind of you know autocorrelation behaviour. So, for as a solution is concerned so we have two different instances first when a rho is known and when rho is in unknown.

(Refer Slide Time: 30:47)



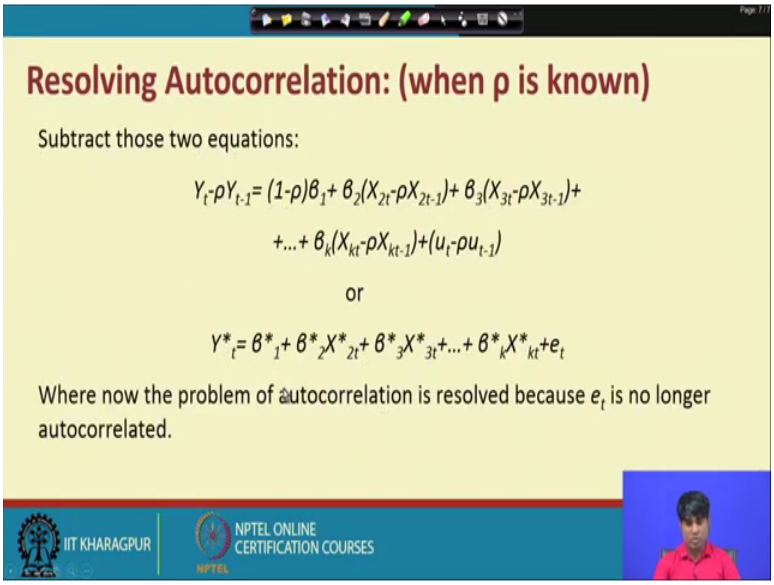
So, now looking for the kind of you know structure we have to first estimate, then first you go by you know simple estimation by dependent and independent variable. Then get the error component and again estimate the error component with, it say lag component then you get to know the rho value and against. So, the rho value can be a used for the kind of you know autocorrelation solutions. So that means, when rho value is known.

(Refer Slide Time: 31:00)



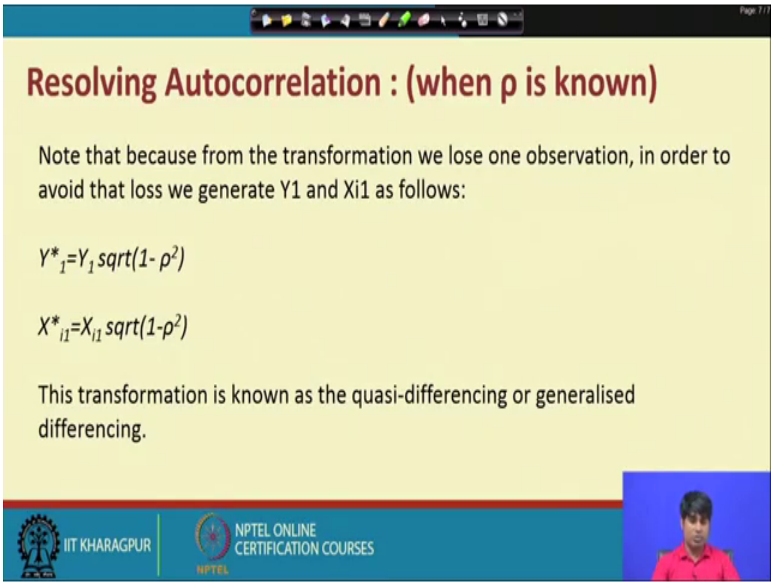
So, just to multiply the rho value and again restructure the entire process, I am sure the autocorrelation problem can be solved some extent. That means, like this way here after multiplying rho value. So, you will get actually some kind of you know additional tricks.

(Refer Slide Time: 31:32)



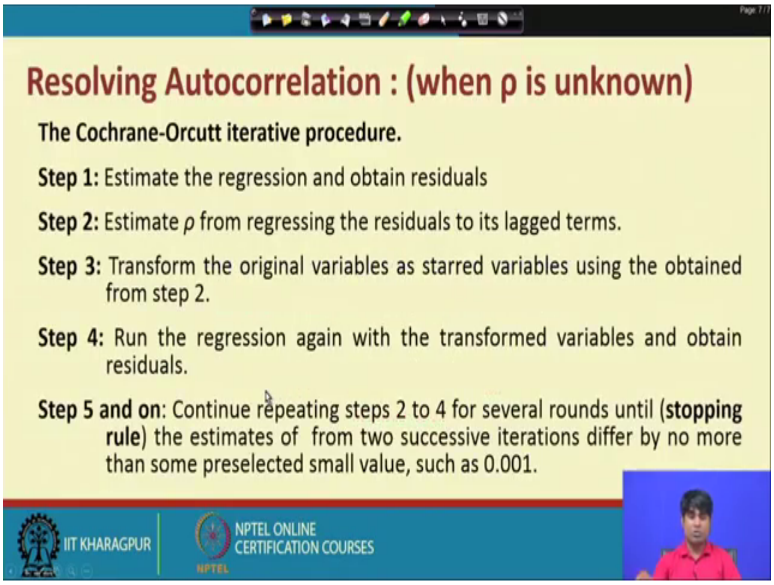
So, now just you know the first equation and the second equation if you simplify and you will get a new equation and that new case and if you go by estimation it would absolutely free from the autocorrelation problem. So, that is what the model is all about.

(Refer Slide Time: 31:48)



And when p is known, so we will follow the particular you know process.

(Refer Slide Time: 31:55)



Likewise you know there are many different you know mechanisms through which you can actually solve the autocorrelation problem. So, ultimately the summation is that you know while solving a particular engineering problem through engineering econometrics. So, you must have a sufficient theory, you must know the variables you know through which the theory can represent, then you can specify the models, have the data and a development model, then estimate the model have the estimated parameters then check the parameters are statistically significant the fitness of the model is the good enough.

Then you go through all the diagnostic starting with the multicolinearity, then autocorrelation and at end of the day we should have a you know estimated model which is actually passes through a specification test g of t and then non multicollinearity problem if it is more multiple regression case and no autocorrelation. Then finally, you can use this model for the a kind of you know management decision and that to solve the a engineering problem.

And till now we have discussed up the up to the autocorrelation, but there are a lots of other diagnostic which we need to take care before you declare that the model is the a fit for the decision making or the kind of you know solving the engineering problem. And in the next class we will discuss these other diagnostics. And in the meantime we will stop here.

Thank you very much. Have a nice day.