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

Lecture – 26
Multicolinearity Problem

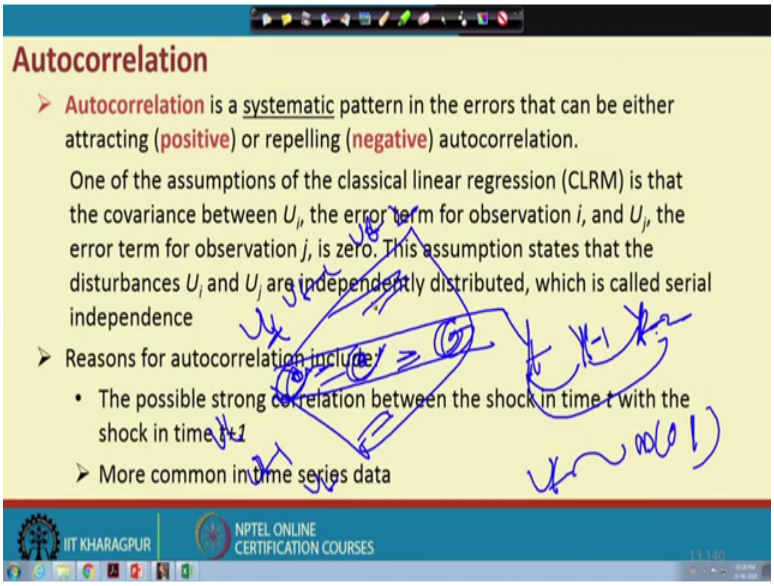
Hello, everybody. This is Rudra Pradhan here. Welcome to Engineering Econometrics. Today, we will continue with model diagnostics and that too the coverage is on autocorrelation. We have in the last lecture we have discussed something about the multicolinearity problem, the kind of you know regions, the kind of you know detection criteria, the kind of you know solution criteria then ultimately the kind of you know the choice of the particular you know models or you know fixing the particular model which is actually a will be the best to predict the engineering problems as per the particular you know requirement.

But, once you have the estimated model passes through a specification test, goodness of fit test and then the kind of you know multicolinearity problem still in this model may not be considered as the best until unless you check other diagnostics and one of such diagnostic is called as you know autocorrelation problem. So, we have already discussed the component called as you know correlation and autocorrelation is a kind of in a special kind of in a component and mostly in a in the case of you know time series data, it very frequently you know occurs and that is how to check the autocorrelation is a kind of you know mandatory component while you know dealing time series data while doing the estimation the kind of you know the prediction.

Even the auto correlation also arise in the case of you know cross sectional data setup, but you know it is a in fact, you know very frequent or very typical issue or you know or we can call as you know drastic issue in the context of you know time series data while you know doing the kind of you know prediction and forecasting. So, in this lectures we like to know what is exactly this autocorrelation and the way we have discussed the multicolinearity problems we like to know what are the regions through which autocorrelation is appearing and how to detect and then how to you know solve this problems and is there any kind of intolerance or something like that.

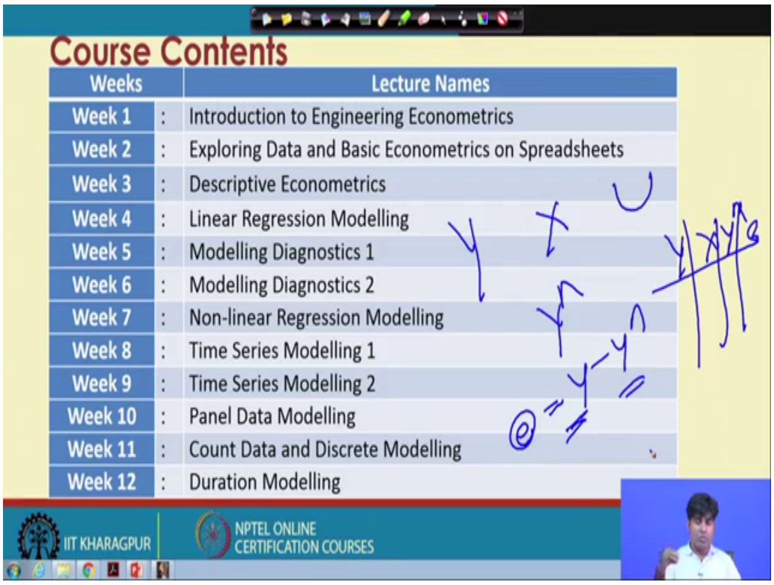
Like you know in the case of you know multicolinearity if you are model is actually for you know just prediction only and forecasting then with multicolinearity we can go ahead because more number of variables will be involve and that to it will give you high in fact, that you know high predictions. But, when there is question of you know reliability of the models or the kind of you know quality of the kind of you know judgment in that times so, you cannot just go ahead like this you have to check thoroughly and pick up the model which is good best and then go for the prediction and forecasting. Likewise, this is another way to search for the best pick models separates a prediction and forecasting is concerned to serve some of the engineering problem. So, let us see that you know: what is exactly autocorrelation.

(Refer Slide Time: 03:33)



So, it is a kind of you know game among the error terms.

(Refer Slide Time: 03:40)



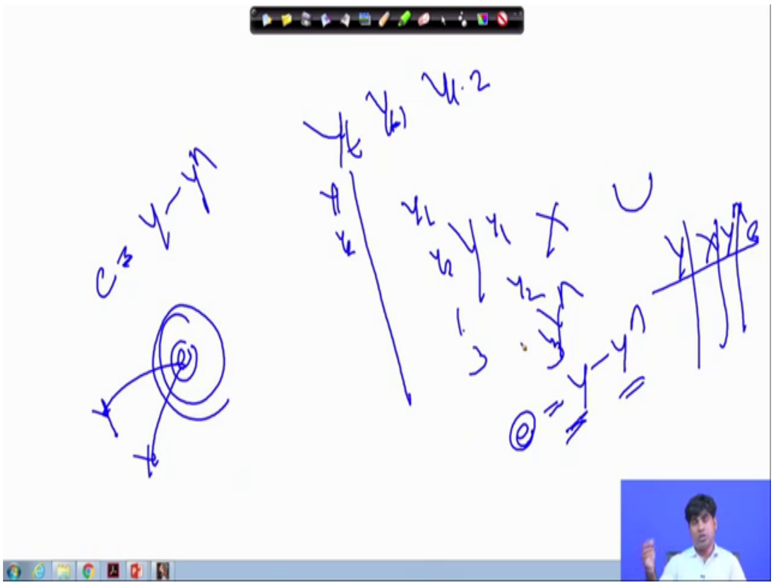
Generally, what we have already discussed till now it is like this a that you know every times we have the game like you know different variables then independence variables for instance let us say Y, X and U. Then, ultimately with the process we will get the estimated dependent variable let us call as you know Y cap. And, then we will find the error term which is the difference between Y minus Y head, that is the Y actual and Y estimated, ok.

So, once you get the error terms then ultimately you have actually Y series X series Y header series and error series. So, that means, after the estimation process your game plan will be Y X Y head and then error. So, we have actually four different column altogether. So, now, we have lots of you know understanding about the Y series we have lots of understanding about the X series, we have also understanding about the Y head series, then the component called as you know error component which is actually the difference between the a actual dependent variables and the estimated dependent variables.

In the deviation or the kind of you know difference you can check from you know one cross sectional unit to another cross sectional unit at a particular point of time or you know at a particular you know cross sectional unit with different points of time. So, that means, there are many different ways you can actually explore the kind of you know you know issues and then detect as per your you know requirement. So, now once you get the error terms that is a after the estimated models you find out the difference between the actual Y and the estimated Y.

So, once you get the estimate in error term so, then there is a you know different kind of you know game all together while a while you are searching for the best models which is actually very liable as per the kind of you know engineering predictions where you know forecasting is concerned.

(Refer Slide Time: 06:02)



So, ultimately you have a error component that is e and what will you do here, so, the error component which is actually the difference between Y minus Y head and then you know so, you know you have the structure called as you know Y, X and then the kind of you know error term.

So, you may have the link between e and Y, e and X, so, because this is another kind of you know variables ultimately our game is how is the actually you know error variable environmental all together will find plenty of you know models are there which can which can do the prediction and forecasting that to with the help of only error term. So, like you know we have you know in including this autocorrelation issue. So, we have actually lots of time series modelling like you know autoregression modeling, moving average modeling, ARIMA modeling, ARCH modeling, GARCH modeling so many advanced modellings are there where you know you know, the game plans will be dones with the help of error terminology.

So, that means, once you get the estimated models and get the error term then this time series modelling behavior will be used for the prediction and forecasting that too with the help of you know error term in most of the instances. That is why you know just you guess that you know how is the beauty of this particular you know error component, because this is one of the component which can give you signal about the best beauty of the prediction and forecasting for instance after getting the estimated models that is Y cap and then you will get the error in error component that is the difference between Y minus Y head.

And, you will find if you in the simple one you know way if you find actually some of the error term is coming exactly equal to 0 by default the model in a most of the instance the model will be actually good for the prediction and forecasting, but you know if the if the model is a not perfect once or the line up the relation line is not the best one then by default some of the error term will not close to 0. That is first instance or you know first indications how to justify that you know your model is not good you know something best for the prediction and forecasting.

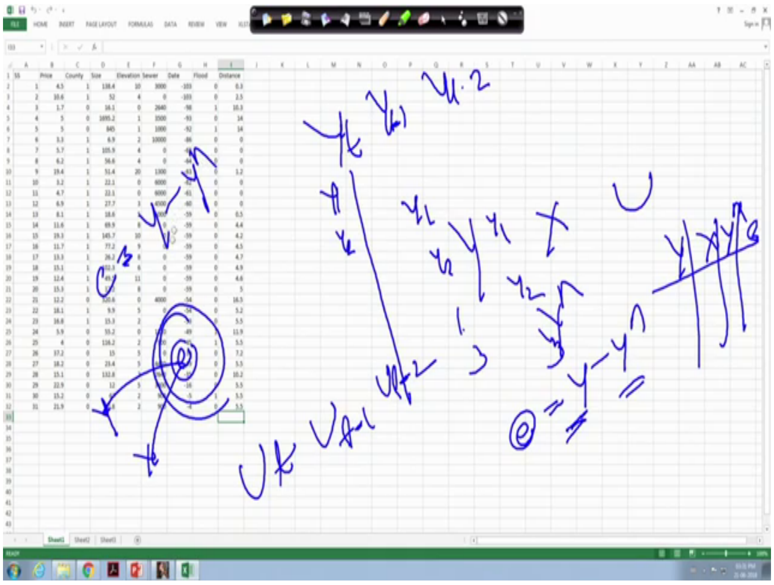
So, now, what is happening so, once you get the error term so, the you now you refine a there is a you know there is lots of you know linkage with you know other component. So, in order to understand the auto correlation issue so, you have to understand the concept called as you know lag variables. So, for instance let us assume that there is a you know variables Y and the data which is having with respect to Y is it time specific. So, as a result we can write the variables Y t and this Y t.

So, we have a couple of you know series. Then from the Y t you can create another variable called as you know Y t minus 1 that is the actually previous on the basis of previous variables. For instance if you start with Y 1, Y 2 like this so, Y t minus 1 means it will restart here only Y Y 1, Y 2, Y 3 and so on. Similarly, you can start Y t minus 2. So, one two lags behind. So, this will start with Y two Y three and so on. So, this is a the sequence about the you know lag variables. So, that will apply to the dependent variable Y t. So, like you know Y t, Y t minus 1 Y t minus 2 up to you know Y t minus k so, n number variables can be created.

Similarly, X t that is if with respect to you know X. So, for the original data which you can call as you know X t and then we can create a new variable X t minus 1 which is the lag variable of original X t. X t minus 2 two lag variable and then X X t minus three. So, likewise you can create you know n number of you know lag variables with respect to Y and X.

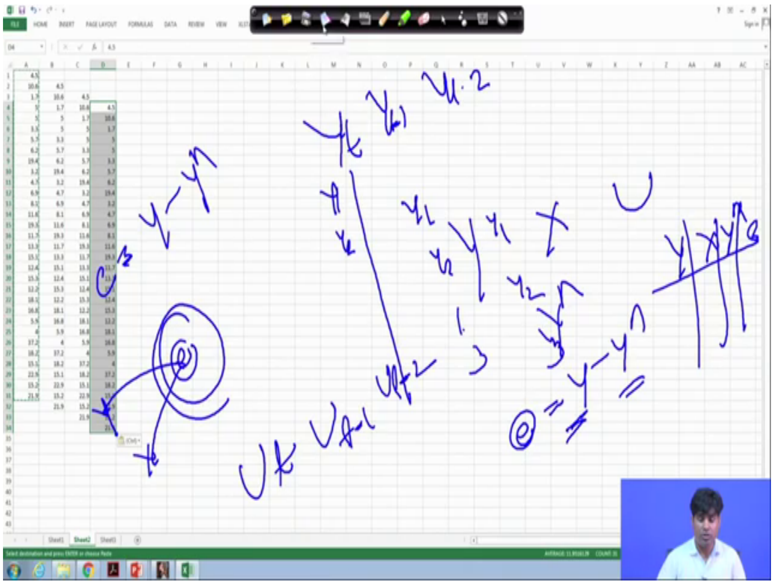
So, now, you will find there is also a error terms. So, the error term can be also connected with you know. So, for instance let us say error term which you have already specified and there is U t then you can create U t minus 1 you can create U t minus 2 U t minus 2 and so on like you know Y t, Y t minus 1, Y t minus 2, X t, X t minus 1, X t minus 2. So, you can have a structure called as you know U t, U t minus 1, U t minus 2 and so on. So, that means, ultimately what is the beauty of this understanding is that you know very easy to you know settle the kind of you know or create the variable. For instance, I will show you here in the excel spreadsheet where you know you can actually see the data set, ok.

(Refer Slide Time: 11:05)



So, let us assume that you know this is a data set and something you know I will I will give you new dataset here. So, this is let us say this is let us say you can create actually a you just copy here some data. Let us say this is actually starts with you know Y t and t equal to 1 up to t 30 31. So, that means, let us say annual data.

(Refer Slide Time: 11:26)



And, how to create actually we you can say you know Y t minus 1. So, this is Y t and then to create a Y t minus 1 just you copy again and then you know paste one lag behind, ok. So, if you Y t minus 2 then its two lag behind Y t minus 3, it is a three lag behind. So, likewise you can create actually n number of you know lag variable similarly for X X and also similarly for actually a U. So, that means, technically what is happening here what is happening here so, you can actually you can actually go to this particular you know component and then and discuss as per the particular you know requirement.

For instance, ok. So, that means, technically we have the kind of you know dependent variable Y t and then independent variable X t and that is an explained then the error term U t that is unexplained. So, now, with the help of Y t we can create Y t minus 1, Y t minus 2 and so on. Similarly, for X you can create X t X t minus X t minus 2 which I already highlighted in the excel sheet how we can easily create new variables on the basis of you know the original information. Similarly, once you get the U t so, then you will find there maybe actually U t minus 1 U t minus 2 U t minus 3 and so on.

So, now, what is the beauty of this particular sequence that you know your current if I say Y t. So, that that is what actually it is like this when I when I called actually Y t when you call Y t then you know there is Y t minus 1 Y t minus 1 and Y t minus 2 ok. So, this is Y t minus 2. So, that means, technically since Y t minus 1 is derived from Y t and Y t minus 2 is derived from Y t or Y t minus 1 the by default there will be some relationship here there will be some relationships here. So, you cannot just ignore. So, that means, when will you actually a you know regress Y t upon Y t minus 1 Y t minus 2 by default there will be multicollinearity problem because all these Y t minus 1 Y t minus 2 are so you know explicitly depends upon its you know original data. So, that is why there is a severe multicollinearity in this kind of you know situation; that means, any lag modelling you will use so, multicollinearity is a genuine problem.

So like so, as a result so, autocorrelation will be also genuine problem. So, because when you are you know connecting Y t Y t minus 1 Y t minus; for instance Y t is the dependent variable which you can predict through the help of lag variable let us say Y t minus 1 and Y t minus 2. Similarly, for X t; X t that is the independent variable which can be also predicted with respect to X t minus 1 X t minus 2 and X t minus 2 we have actually different kind of you know modelling like you know lag modelling can be autoregressive type, autoregressive distributed type.

So, we have different schemes and we will discuss in details in the time series modelling, but in the mean times we are discussing actually the issue of the autocorrelation which is actually game of the error terms and the issue of the autocorrelation that you know existence of linear relationship among the error terms on the that means, technically. So, when we have Y t U t and U t minus 1. So, then we like to check what is the correlation between U t and U t minus 1. So, obviously, there will be a correlation, but that correlation coefficient should not be statistically significant.

If it is statistically significant then this will be this will give you actually a you know wrong message to the reliability or the you know separates the use of the estimated model is concerned to predict and to forecast the particular you know engineering problem. So, ultimately so, there should not be any autocorrelation problem; that means, technical if you say that you know what is autocorrelation and what is the extremes. Autocorrelation means so, it is a simply serial correlation and the kind of you know observation about the systematic pattern in the errors.

So, like you know there should not be any linear relationship among the regression that to X 1, X 2, X 3 like this. So, that should not be any relationship between U t U t minus 1 U t minus 2 and so on, but un unfortunately there will be some relationship between U t U t minus 1 U t minus because U t minus 1 U t minus 2 all are you know derived from the original U t. So, when you talk about the autocorrelation then the issue of you know the issue of you know collinearity will be coming in the picture that within error terms.

So, what is exactly this error terms just you know highlighted; that means, so, when you are connecting U t U t minus 1 U t minus 2 U t minus you know 3 and so on. So, first of all you have to fix the lag length, what extent the you know you can enlarge the model. So, there is a you know there is a lag selection criteria which will be discussed in the later stage when dealing the time series modelling and let us assume as you know the final lag selection should be up to 2. So, then technically you have the model U t, U t minus 1, U t minus 2 like you know this is actually U t, U t minus 1 and U t minus 2 and three variables.

Then you can prepare a covariance matrix U t U t then. Then U t minus 1 then U t minus 2 you can prepare a co you know correlation matrix and this correlation matrix you will find that the diagonal elements will give you the variance vectors that is the error you know error variance and at that error variance we have we have one assumption under OLS that the error variance should be uniform or you know that is what is called as you how much capacity; that means, all these are not diagonal elements, ok. So, in this 3 into 3, so, here three diagonals that is the variance vector. So, this variance vector should be actually uniform which will discuss actually in the in the next lecture that is the issue of you know how much capacity versus heterous capacity.

But, the other component that is the of diagonals you know on diagonals which are symmetric in nature; that means, U t, U t minus 1 and U t minus 2 again U t minus 1 U t minus 2 since there are three variable. So, we have a three cross correlations. So, now, our journey is to check whether these cross correlations are 0 or you know non zero. If it is 0 then there is a no autocorrelation, if it is actually non zero then there is a autocorrelation. So, that means, technically so, this particular you know matrix that is with respect to error terms. So, what is the assumption error terms should be normally distributed with main 0 and unit variance.

So, that means, a this is how the nature of the a U or a nature of the U must be like this then ultimately. So, gain between error terms will be here gain between error terms will be here then ultimately. So, this will give you the signal that you know ultimately once you find the you know autocorrelation matrix then in this will be actually transport into a form you know what is the best actually. So far as the best is concerned that sets a kind of you know unit matrix. So, then by default all the cross correlation will be a 0. So, that means, U t U t U t minus 1 and U t U t minus 2 and then finally, U t minus 1 verses U t minus 2.

So, whether when there is a there consistent then you know it is a U t U t minus 1 or U t minus 1 U t so, it is a kind of the just change of the position then ultimately this should be equal to 0. If that is the case then there is in no autocorrelation or there is no serial correlation, but if they are not 0 and by default so, it will indicates the indication is that you know there is a serial correlation or there is autocorrelation. So, that means, in the simple understanding the covariance between two error terms should not be equal to 0. So, that is the case of you know autocorrelation if the correlation between these two error terms are coming 0 then by default. So, there is no correlations if the correlation covariance between U t, U t minus 1 is actually not equal to 0, then the game will start that it is called as you know game of serial correlation or the game of you know autocorrelation.

And, what I have already mentioned it is very frequent and you know dominate kind of you know problems in the time series data. So, that is why it is mandatory that you know you have to check the level of you know autocorrelations before you go use this model for the prediction and forecasting. So, that means, you just see that you know how interesting is that just you know just putting the input and indicating the dependent and independent indicating the estimated out output that is not the enough actually. The whole work or the beauty of the particular process will start when you start checking or you know testing all these you know diagnostics that is very interesting and that is how the beauty of this engineering econometrics.

So, ultimately what will you do actually so, you just estimate the models have the you know Y bar ykf that is the then we will get the error term which is the difference between the actual Y and the predicted Y and then you know check what is the level of you know error terms and then you will check the kind of you know nature of correlation. So, whatever you know items that is you know the example is the multicolinearity which we have discuss in previous lecture in real life scenario any kind of you know real life engineering problem. So, you will not find a situation where all independent variables are independent. So, that that is how the multicolinearity problems sometimes you know genuines, but ultimately our requirement is how to minimize these degree of you know relationship among the regression.

Here also same case, since you know the variables are you know independent variables are not you know align, linearly independent and some extent there is linearly dependent and that linear independent should not be statistically significant after doing the processing then ultimately we will be find a situation where you know most of the variables are not actually highly correlated. So, in that in that contest these are the variables can be used for the prediction and forecasting.

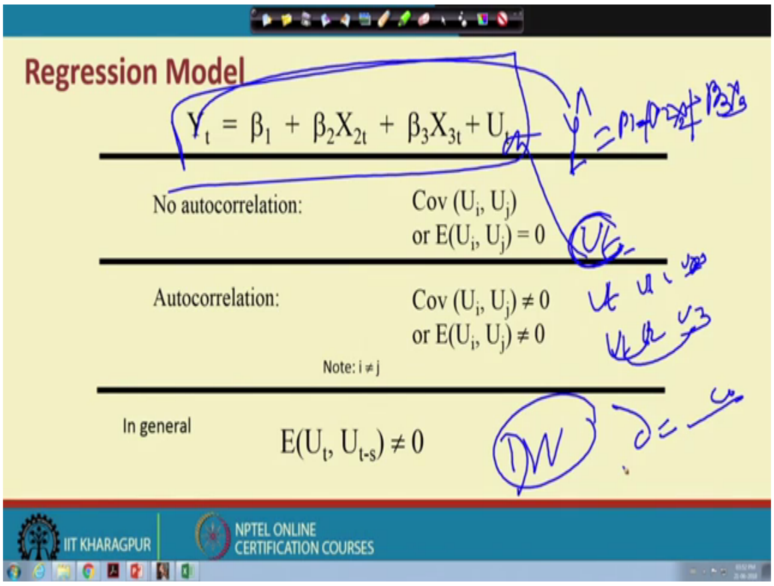
So, similarly in that case of you know serial correlation. So, so that means, it is step by step process. So, you your first step is Y with you know all the independent variable X 1, X 2, X 3 like this then in the second step while we getting the estimated models. So, the first check process is the multi colinearity of course, we have already gone through a specification test and you know the kind of you know goodness fit test after that you will come to the diagnostic test that too you know multi colinearity issue and you see actually whether the model can be change or restructured somehow re-estimate and in such a way that you know the impact of these you know or the kind of you know the relationship between the independent variable should not be actually so high.

So, ultimately simple message is that multicolinearity cannot be clean completely. So, in any extent it will be having a always it will be there in the system, but you have to go ahead with the prediction and foresting if your degree or multicolinearity is very low and that too the correlation coefficient among the having regressors are not statistically significant. This is also the case of you know serial correlation and autocorrelation. Here the relationship between U t U t minus 1 U t U t minus 2 or U t minus 1 U t minus 2, these are the kind of you know correlation coefficients which should not be actually a you know not equal to 0. If they the relationship between U t U t minus 1 U t U t minus 2 and U t minus 1 U t minus 2 are not equal to 0 so, that means, the sum amount of you know autocorrelation arise in the systems.

So, ultimately the job is now to look up to look this look into this problem and find out what is the degree of you know auto correlation and then think about the some kind of you know solution mechanism through which you can actually being a news a new kind of you know process or estimation estimated model, where we can actually you know you know you can do the way better prediction and better forecasting without any obstacles or the kind of you know errors. So, that is why autocorrelation is very interesting game and that too you know very interesting again for the time series data. And, that is one of the you know most mostly genuine problem in the time series data. So, the region of autocorrelation includes the possible strong correlation between the shock in time with the shock in time t t plus 1.

So, usually a its happening actually in the real life scenario and when will we looking for the macro level kind of you know environment most of the variables are you know interdependent as a result the correlation coefficient the error correlation among the terms will be also having some kind of you know independency. So, now we will find out: what is the level of dependency and how you to care the data. One way you can say that you know autocorrelation with here then you start you know estimating the process with you know lag variable. The moment you use lag very old autocorrelation by default will be these in problem. So, you must be very careful mean this is the first indication how you to you know start and understanding the kind of you know requirement.

(Refer Slide Time: 26:21)



So, this is my this is one of the regression models where you know we have two independent variables X 2 and X 3 and Y is the dependent variable and when we estimate the model ultimately so; that means, this is the actual model this is the actual model functional form, then the estimated model will be Y cap that is beta 1 beta 2 X plus beta 2 X and then beta 3 X X 3 these beta 2 X 2 and beta 3 X 3.

So, now the difference between Y t and Y Y hat t will give you the kind of you know error terms U t. So, that is what here, ok. Now, once you get the U t you define a you know the U U t U t minus 1 U t minus 2 U t minus 2 something like that and or else you can call as you know U 1, U 2, U 3 and so, on like this. So, that means, series of error terms will be error where will be there and if you say that you know autocorrelation the other name is called as a serial correlation then covariance of U i U j, so; that means, U U 1 U 2 or U 1 U 3 not equal to 0 or you know the expected value of you know U i U j is not equal to 0, if that is the case and that is called as you know not a there is a not correlates means there is autocorrelations if it is exactly equal to 0, then there is no autocorrelation.

So, but, that means, technically when your linking error terms like you know U 1, U 2, U 3, U 4 like this. So, autocorrelation is problem is a problem among the error terms that too only you know off-diagonals and on-diagonals not the diagonals, because the diagonal elements will be called as you know error variance. But, the autocorrelation is a problem of you know error covariance that is why every times U t U t minus 1 U t U t minus 2 and so on similarly, U 1 U 1 U 2 U 1 U 3 or U t U 3 where i not equal to j.

But, when the moment I equal to j; that means, we are in a diagonal elements only and that too at U 1 U 1 sigma U sigma square U 1 that is the error variance similarly sigma square U 2 sigma square U 3. So, these are all called as a error variance and that particular you know detection and the kind of you know exploration is nothing, but new problem called as you know heteroscedasticity which will be discussing in the later part.

But, in the mean time in this in this problems we like to check actually how is the relationship among these error terms; that means, technically whether the covariance among the error terms are 0 or not 0. If they are 0, then the simple declaration is that you know there is no autocorrelation. If they are not equal to 0 then there is a autocorrelations.

So, now if say more or less like you know multi colinearity problem. So, if not equal to 0, then the range will be starting with you know 1 minus 1 or something like you know indefinite. So, now, ultimately we how to check what is the degree of you know environment in the autocorrelation. Like degree of multi colinearity here also we can check the degree of you know autocorrelation. So, there are several you know test statistics are there which can help you to find out the level of you know autocorrelation and what is the what is that particular value where the model is actually a completely free from this autocorrelation problem and if not what is the tolerance levels through which you can go ahead with the prediction and forecasting or in what extent the model need actually restructuring completely.

So, that means, there three different instances after the check of the autocorrelation. First instance if there is a covariance among the error terms are coming 0 for every case and there is no autocorrelation. So, you can go ahead with original you know estimated model for the prediction and forecasting because the indication is that you know there is no serial correlation. Now, if there is a serial correlation that means, the covariance between two error terms are not equal to 0 so, that means, there is a problem.

So, now, what how much that problem so far as the prediction forecasting is concerned? Usually theoretically or you know as per the statistical rule some minimum level maybe actually not an issue, but ultimately if that particular you know in statistical significant then a definite it will be an issue. So, ultimately so, in order to know the degree of you know involvement. So, we have couple of statistic as I have already mentioned.

So, one of the standard test statistic which you have frequently used in the case of you know Euler's mechanism that too estimate the models go ahead with the prediction and forecasting that is called as a Durbin Watson d statistics. So, its otherwise called as you know DW, DW Durbin Watson d statistic two statistician you know they develop this particular you know technique that simply actually Durbin Watson d statistic that is small d. So, which is actually covariance of error terms by you know variance of like you know correlation coefficient, but that too here with respect to error term. So, I will I will give you the details about the calculation of Durbin Watson mechanism.

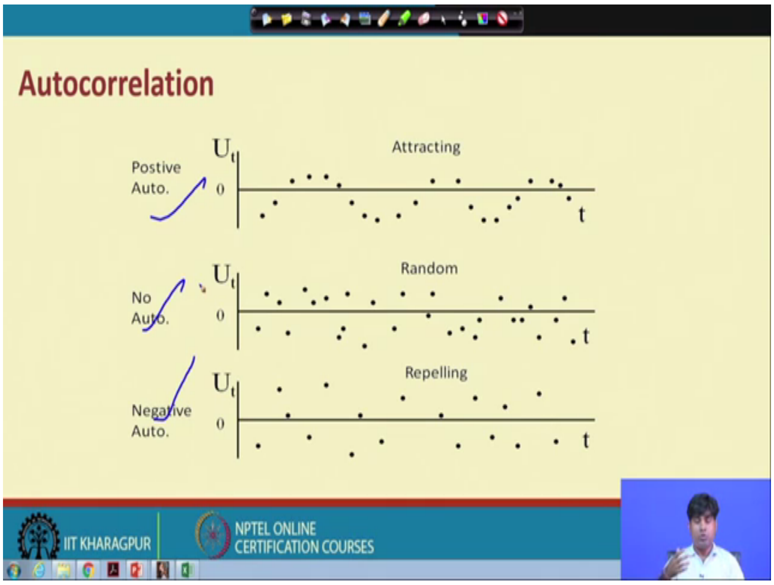
So, the simple understanding is that you know it will give the kind of you know standardized value through which you can check whether there is correlation or autocorrelation or not autocorrelation and what extent the model will be you know very reliable so far as you know presence and absence what auto correlation is concerned.

So, it will be streamlined the process like you know r squares is in probability the DW value will have a some kind of you know standardizations through which you can actually give remarks or comments to the autocorrelation, so far as the estimated model is concerned to predict a particular you know engineering problems or to forecast a particular engineering problem.

So, ultimately so, this is how the you know basic understanding about the autocorrelation it is the it is simply the game among the error terms and as such you need actually covariance among the terms and then the target is on cross correlation summary that too you have to find out covariance of Y i Y j, where i not equal to j. So, means only off diagonal and on diagonals components need to be checked. If there is no serial correlation then these of diagonals and non diagonals will becoming 0, and the on diagonal element should come to 1. So, if that is the case that is called as you know homogeneity and in that case if the correlation cross correlation off diagonal and on diagonals are not zero and that is called as you know serial correlation.

Of course, is mathematically to understand is very easy or practically we need actually a clean environment where there the particular estimated models should not have any autocorrelation, but in reality to get that it is not so easy. So, that is why you have to go ahead with the prediction and forecasting with some you know tolerance level. So, what is the tolerance levels and how to calculate the tolerance levels and how the model will be you know means will not be affected with that particular tolerance will discuss in details in the later stage, but in the mean time this is what the kind of you know game about the autocorrelation.

(Refer Slide Time: 34:40)



And, ultimately in order to understand the kind of you know nature like you know multi colinearity we have no issue about the kind of you know positive relationship among the regressors or the negative relationship among the regressor, but ultimately we like to check whether there is a relationship.

The case is also same here in the case of in the you know occasion of you know autocorrelation. Here since we are you know establishing or you know checking the relationship between error terms like you know U 1 U 2 U 1 U 3 or U U t U t minus 1 U t U t minus 2 and so on. So, ultimately so, we are actually establishing the relationship the moment we will get the covariance between U t U t minus 1 or U t U t minus 2 or U 1 U 2 and U 3, then definitely will find some value,.

So, the variance of the error term should be or as you know positive and you know it should be uniform, but here these are all covariant. So, sometimes it may be negative sometimes it may be positive, but whether it is a positive or negative it is not an issue, whether there is a presence or not presence that is that is the major issue. If there is if there is no presence; that means, the you know coefficients will coming 0 and that is good sign for the a you know modeling, the relation modelling separates the estimates and predictions and the kind of you know forecasting is concerned.

But, what is happening in the reality. So, some kind of you know relationship will be there. Because a the there is a relationship among regressors. So, by default so, the same way the impact will come to the error terms. So, there will be some way there is a relationship among the error terms. So, now, we have to plot these error terms and check whether the relationship are coming positive or negative or there is a you know link it all. So, that is why we have here actually not a correlations and then negative auto and positive auto accordingly.

So, ultimately this will give you some kind of you know random check and you know how you can go ahead. That is called as you know just graphical inspections about you know the behavior of the error terms corresponding to you know corresponding to the estimation of you know Y t upon you know X t that is you know the game between dependent variable and independent variable.

So, ultimately how to obtain the error terms which we have already discussed, but now after getting the you know error term. So, how is the nature of these error terms or how is the diagnostics so far as you know you know autocorrelation and the other connected problem is heteroscedasticity. So, we like to know all these things how to you know dictate all these kind of you know situations and think about the kind of you know solution and how you to being a kind of you know green environment where the estimated models can be used for the prediction and forecasting of any kind of you know engineering problem without any substantial you know defective.

So, ultimately we need a kind you know situation where you know we have estimated models that is free from all kinds of you know errors or it will I mean in short it will pass through all the diagnostics before you use prediction and forecasting. We will discuss in details in the next class next lecture. With this we will stop here.

Thank you very much, have a nice day.