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

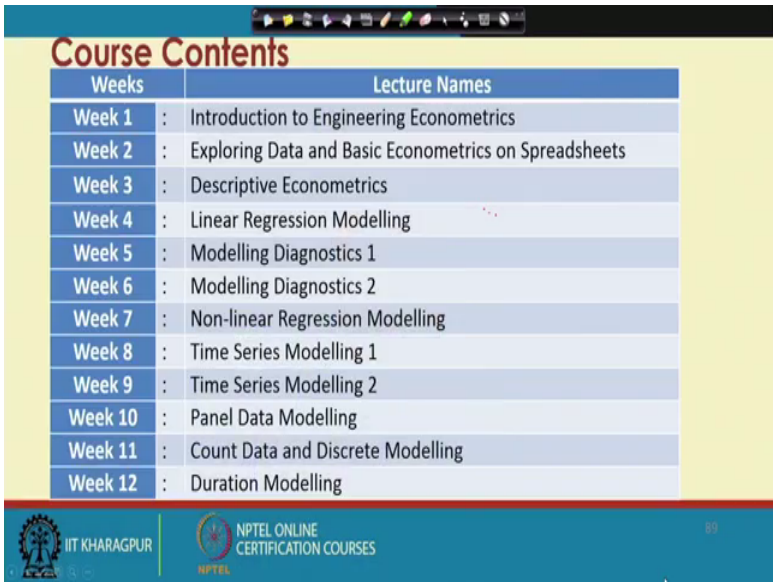
Lecture – 23
Modelling Diagnostics

Hello, everybody. This is Rudra Pradhan here. Welcome to Engineering Econometrics. Today, we will continue with model Modelling Diagnostics. In the last couple of lectures, we have discuss something about linear regression modeling. In that way in a bivariate framework and multivariate framework and we have discussed a kind of you know engineering problems where the problem may be with two variables, where one is dependent variable. And one is independent variables and we have also discuss a scenario where we have one dependent variable with multiple independent variables.

While doing the kind of you know predictions and forecastings that is with respect to dependent variable and independent variables so, the step by step process is very simple. So, we have to fit the model, then estimate the model and after a after the estimations we have to go for the reliability check. Once the estimated model will pass through reliability check, then we will go for the prediction and forecastings, but by the way in the process of discussions we have we have gone through various test procedures starting with specification test, goodness of fit test, diagnostic test and out of sample prediction test. These are all for you know main test through which we can apply in the estimated models to check the reliability before you do the prediction and forecasting.

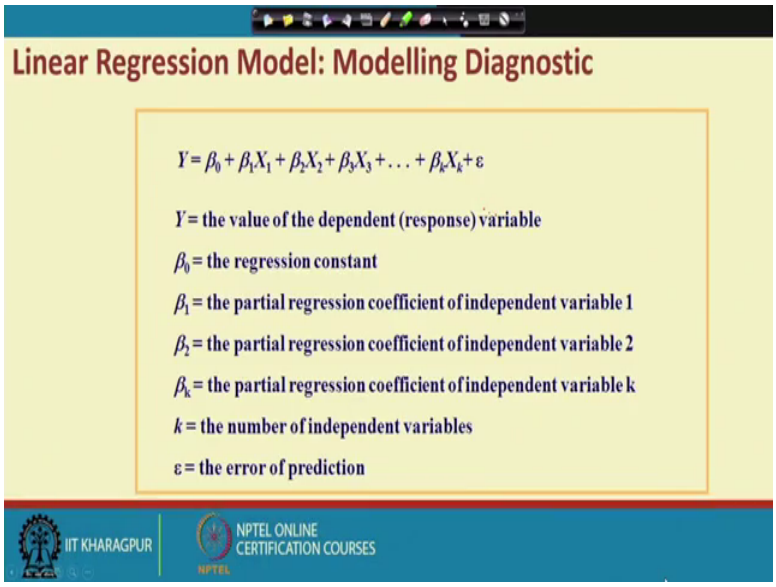
In the last class we have discussed something you know problems relating to the specification test and the kind of you know goodness of fit test, and also slightly we have discussed about the out of sample prediction test. The diagnostic test part is not yet covers. So, we will discuss the details of that particular test in this lecture and it is not a simple component. It is a kind of, you know it is like you know complex kind of you know structures. We have to we have to pass through various you types of you know test under the diagnostic test and then we like to check and we finally, declare that the estimated model is for the prediction and forecastings. So, let us see how is the kind of you know structure.

(Refer Slide Time: 02:54)



And, then we can proceed.

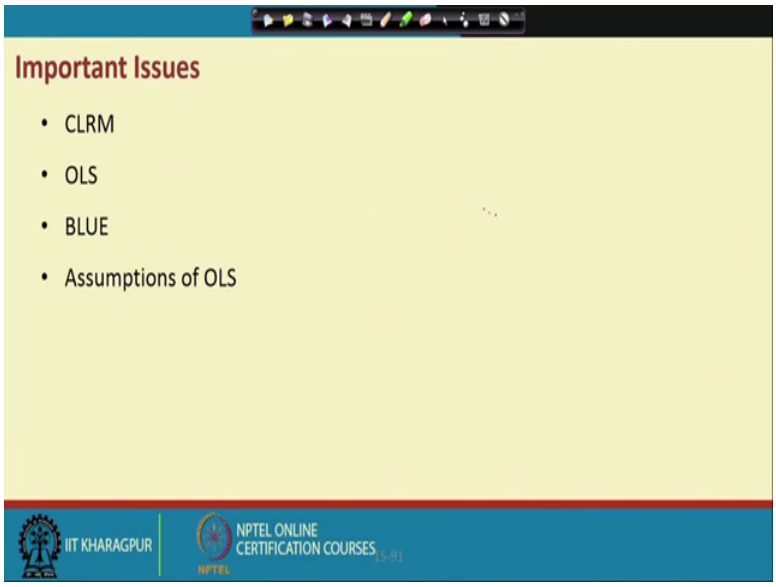
(Refer Slide Time: 02:55)



So, this is what a what is all about the a simple linear regression modelling and that too in a kind of you know multivariate structures where we have we have a dependent variables and we have couple of independent variables these are all called as you know k independent variable. So, now, dealing independent you know multivariate problems that too multiple regression modeling, so, we can actually express many things and our you know testing structure testing procedure will be more effective and you know to discuss

the kind of you know engineering problems and that too predict the engineering requirements and forecasting the engineering requirements. So, let us see how is the kind of you know structure.

(Refer Slide Time: 03:41)

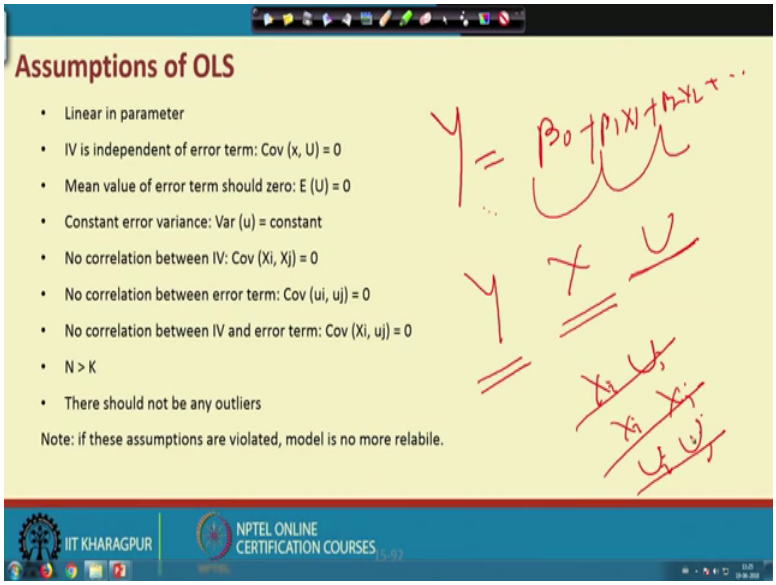


And, what I have already mentioned that you know once you pick up the particular problem, so, the standard procedure is to fit the model, estimate the model and then go for the reliability. Once it will pass through reliability then it can go for the prediction and forecasting. So, whatever we have discussed till now so, this is you know something like this, the first part of this particular discussion is the CLRM that is classical linear regression modelling and then the technique which we apply to estimate the model is called as a ordinary least square methods, that is OLS and then and then the reliability of the estimated model will typically called you know BLUE, that is the best linear unbiased estimator and it is followed by a theorem called as a Gauss-Markov theorem.

And, all these you know items can be covers and that too with the help of the assumptions of ordinary least squares mechanisms. So, now, you know the reliability of the estimated models by the OLS mechanism should you know satisfy all the assumptions, otherwise it will go against the reliability part. So, what will you do right now, we will discuss all these you know assumptions behind OLS and a particular means one if you go one by one of this particular you know assumption then you will find a

couple of issues will arise and this issues can be actually checked and then finally, cleaned before you go for the estimation and forecasting.

(Refer Slide Time: 05:27)



So, now what we will do here, let see where means the assumption of OLS. So, first assumption behind this OLS mechanism is the model must be linear in parameters we start with various you know structure of the models simple you can write like this beta 0 plus beta 1 X 1 plus beta 2 X 2 and so on and every times this parameters are linears, but that does not mean that variables will be linear. So, we may have a function and that functions may be in any form. But, the parameters which will be connected to this particular you know modeling, that is a classical linear regression modeling, which is actually estimated by OLS mechanism so, where parameters should be technically you know linear in nature.

Then, second point is the independent variable of error mean say a independent variable is independent of error term. So, that means, technically in the process of you know estimations we have three kinds of you know variables. So, one is the dependent variables, then the second one is independent variables that is known to you that is called as explained and then independent unexplained variable which will called as you know U. So, now, the second point you know talks about the relationship between x and U. So, that means, technically there should not be any covariance between x and y.

Then, third one is the mean value of the error term should be 0. So, which we have already discussed because the line of the best fit or the estimated line will pass through all the points where you know 50 points of the observation will be above and 50 percent observation will be below that is the difference between actual y and estimated y. And the points which are above than estimated lines where the error term will be positive and the points which are below than estimated lines, where the error will be negative. So, sum of these positive errors and sum of the negative error should be equal and as a result a sum of the error term should be equal to 0. If that is the case the particular line can be called as the best fit line and that too OLS is valid for this particular you know estimation.

Then, we have constant error variance. So, where you know once we will find out the error component then we have to find out the error variance. So, the error variance should be constant over the time or you know any cross sectional unit. If that is the case, then this good for the OLS and good for the reliability. If the variants will differ from time to time and different cross sectional units then this will go against the OLS and as a result the particular model cannot be reliable one.

Then no correlation between independent variables; so, that means, technically we have two sets a variable typically multiple regression case or multivariate case, where particularly in the case of you know multiple regression we have a dependent variable with the series of independent variables. So, where we have the declaration that this variables are independent. So, as a result technically there should not be any linear relationship among these independent variables if that is the case then it goes against the OLS mechanism and by default the model will not be reliable. So, it is against the reliability. So, means as a result the particular model cannot be used for prediction and forecasting.

Then, you know correlation between error terms typically. Like you know two independent variable will not have any relationship. So, similarly error term is also independent variable here and like you know x 1 and x 2 so, there will be you know u 1, u 2 mean cluster of error terms which may be in the form of you know different cross sectional structures or different times of the structure by using the law concept. And as a result the requirement is the correlation among these error terms should not be something positive or negative it should be exactly equal to 0. If that is the case then the model will

be reliable and it is in favor of OLS, if it is actually not equal to 0 it goes against the OLS mechanism.

Then you know correlation between independent variable and error terms. So, technically there are three kind of you know structure. So, within X and U then XX and UU, ok. So, you just put you know i j, i j and this is let say i j. So, like this these are all three different clusters we have to check and in all the cases the covariance or the correlation should be equal to 0. If that is the case it is in favor of OLS if that is not the case then it is it goes against the OLS mechanisms. As a result the estimated model which is derive through a less mechanism is not at all reliable and as a result the particular model cannot be used for forecasting.

And, then finally, the sampling structure where the sample size should be substantially greater than to K and then there should not be any outlier in the system; that means, outline is a kind of you know data point which is a highly distance from other data points. If there is a outlier then typically the points which we have discussed, like you know correlation between a independent variables, then correlation with any error terms, then correlation between independent variable and error terms and the kind of you know constant variants or the kind of you know a mean sum of the error, whatever maybe the items which we have discussed so, it will get affected.

So, as a result so, the outlier problems would be drastically you know checked and then you know go ahead with the kind of you know prediction and forecasting. So, we have to see whether there is outlier. So, if there is a outlier so, there are two way you can solve the issue, either we remove the outlier without affecting the sampling or else you keep the outlier, but do the kind of you know transformation very you know different ways of you know set up, so that the influence of outlier particular outliers may not affect the modelling structure drastically.

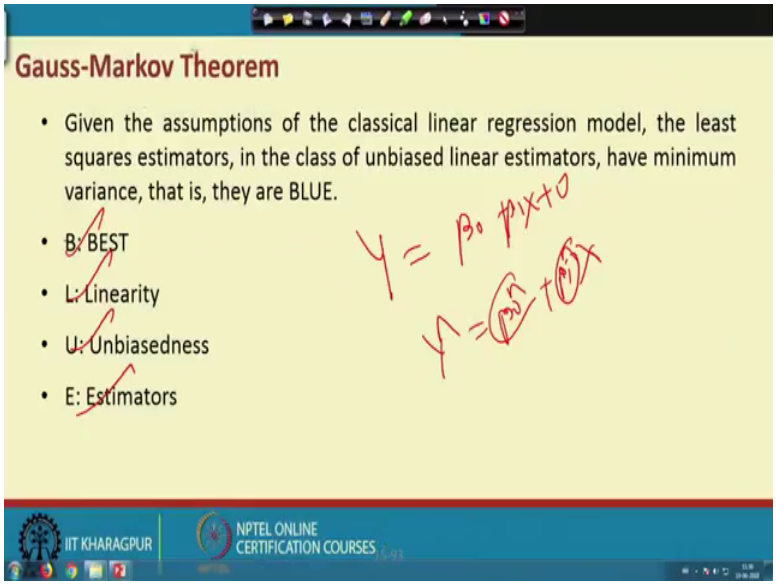
And, the simple logic is that you know you know means after going through all these assumptions. So, the simple understanding is like that if the assumptions are violated then the model cannot be is reliable. So, that means, the estimated model will pass through all these assumptions before we declare that the model is a reliable one for the for the best one for the prediction and forecasting. So, that means, the particular

assumptions or all these assumptions should be satisfied before the prediction and forecasting.

So, if not if that is not the case then the OLS mechanism may not be applied. So, we can look for some different you know mechanisms like general square methods (Refer Time: 12:57) square method to estimate the particular process and then go for the kind of you know prediction and forecasting. Since OLS mechanism is very simple, very easy to understand, very easy to pick up so, in the first instance we have to apply OLS and then do the estimations then check the reliability and go ahead with the prediction and forecasting.

If it goes again a kind of you know diagnostics or the OLS assumptions then by default we have to think about the restructuring or you know re-estimation process. Otherwise this is the simple way to estimate check and then proceed for the prediction and forecasting. So, likewise we have to be very careful all this you know scenario then we have to discuss the details about the scenario.

(Refer Slide Time: 13:50)



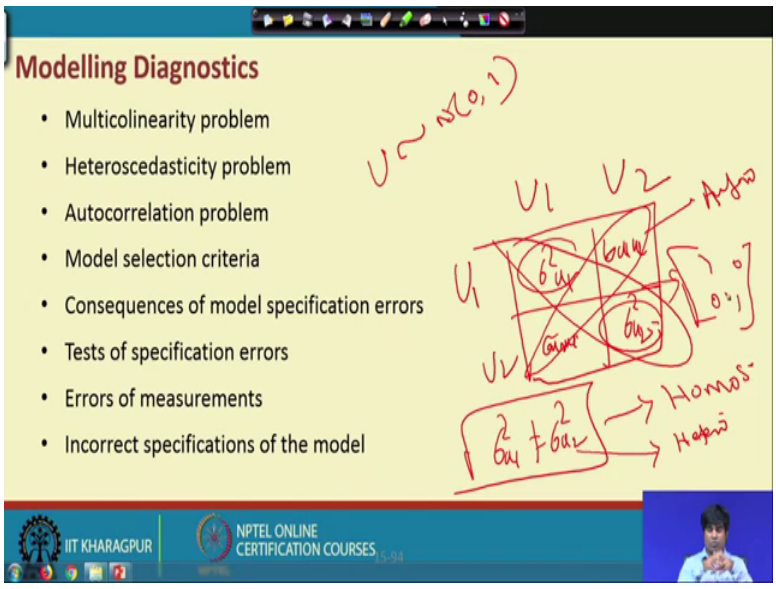
So, that means, technically whatever we have discussed till now, it is the problem identifications then model building, then model estimation and that too with the help of typically OLS techniques. So, once you use OLS technique or apply OLS technique to have the estimated model, then the estimated model should pass through all these assumptions if that is the case then the particular model is called as you know BLUE so;

that means, it is called as a best linear and unbiased estimator. So, this is the case here. So, BLUE stands for best linear unbiased estimator. So, that means, we start with the model simply y equal to beta 0 plus beta 1 X plus U and then estimated will be beta 0 plus beta 1 hat X and so, these parameters or these estimators should be actually free from all errors and finally, declared as you know best. That is you know or you know Gauss-Markov theorem that is best linear unbiased estimator.

So, it is a popularly known as a Gauss-Markov theorem. So, that means, if you go if you just connect with the assumption and the kind of you know requirement then the theorem you know tells like this giving the assumption of the classical linear regression model the least square estimators in the class of unbiased estimators have minimum variance that is there in a BLUE. So, that means, a so, whatever you know values of the parameters we have received through the estimation process it should be called as you know best and followed by linearity unbiasness and a minimum variance.

So, that means, the parameters should be technically free from all kinds of you know errors, then we will called as you know best. So, once the parameters should be declared as the best then by default the estimated model will be declared as the best and as a result you can use this model for prediction and forecasting. So, this is what the Gauss-Markov theorem.

(Refer Slide Time: 16:04)



And, now if the Gauss-Markov theorem will not a you know valid, so, that means, the particular model is going against the OLS assumptions either few or you know all or something like that then in any case model cannot be declared as the best model or you know the line cannot be declared as a line of the best fit.

So, now let us start with since means this is a class of you know diagnostic checks or diagnostic structure so, what will you do? So, we like to point out you know what are the diagnostic we are supposed to do, before you declare the model is good fitted and the model should be used for the prediction and forecasting. So, combining all these assumptions so, we are supposed to discuss all these problems. So, the first problem dealing with this particular process is called as a multicolinearity problem. So, here the issue is about the relationship among the regression, that is the independent variables.

Then, heteroscedasticity problem: so, that is the correlation variance of the error terms if the variance of the error term is the constant then this is the declaration of the homoscedasticity and good for the OLS mechanism and good for the model fit and then you can use for forecasting and predictions. If the error term is not error variance is a not constant over the time or correlation unit then this particular problem is called as a heteroscedasticity problem and it is typical virus [laughing] in the modelling process or estimation process. So, before you use that particular model for prediction and forecasting so, it should be free from heteroscedasticity problem; that means, it is simply called as you know heterogeneous variance so, which is against the OLS mechanisms.

So, because OLS mechanism requirement is a error variance should be constant and if that is the case then that is called you know homoscedasticity; that means, technically homoscedasticity is good for the OLS and heteroscedasticity against you know with the OLS. So, we have to see whether we are in a heteroscedasticity structure or homoscedasticity structure, but there may be the problem may be in between so, but we try to restructure re-estimated in redesign you know until unless you get the model in such a way that it is free from all kinds you know error.

Then, the third problem is the autocorrelation problem where we like to check whether there is a correlation among the error terms, not the variance components it is the covariance component. For instance suppose there are two error terms, then the error term then the then we can write a covariance matrix like this. V 1 V 2 and U U 1 U 2. So,

then we will have 2 into 2 matrix this is sigma square U 1 and this is sigma square U 2 this is sigma U 1, U 2, then this is sigma U 2 U 1. So, now, this part is the homoscedasticity part; that means, sigma square U 1 should be sigma square U 2. So, if that is the case declaration is a homoscedasticity and if that is not equal then it is the signal is heteroscedasticity.

So, in other case it is a covariance between in these two error term and that is the case of called as you know auto correlation. So, if you simplify with you know normality structure then you should be you know closely connected to a mean 0 and unit variance and as a result this particular matrix should transfer into unit matrix. So, if that is the case then that is good for the OLS estimations and good for the model building, model estimations and the kind of you know prediction and forecasting. So, that means, technically so, you are supposed to prepare two different matrix or two different kind of you know testing. So, one particular testing is with respect to all independent variables explained and all independent variables unexplained. So, two different matrix you have to prepare like this and then check what is the status of the particular you know estimated model.

Then finally, you can also establish the relationship between U and X jointly whether there is a kind of you know relationship and for that the requirement is there should not be any relationship between y U and X. So, to summarize so, there should not be any relationship between X i and X j many you know relationship between U i and U j and again there should not be any relationship between U i and X j. So, that means, X upon X, U upon U and X upon U or U upon X. So, these are the kind of you know possibilities. So, these relationships should not be there in the OLS estimation process while you know using all these technique to predict the particular you know dependent variable with respect to independent variable.

Then model selection criteria. So, even if you are fixing a particular you know models then the going through estimation process by OLS technique or other techniques. So, there are certain criteria is there to see that this whether the particular model is you know, for the prediction and forecasting. Sometimes you know what happens even the model is free from multicollinearity, heteroscedasticity, autocorrelation and the diagnostics, so, still you know you cannot say that the model is actually good for the prediction and forecasting. So, when we have actually couple of models by applying the robustness you

know structure then you have to finally, select a model which is very good and which is actually very much useful for prediction and forecasting.

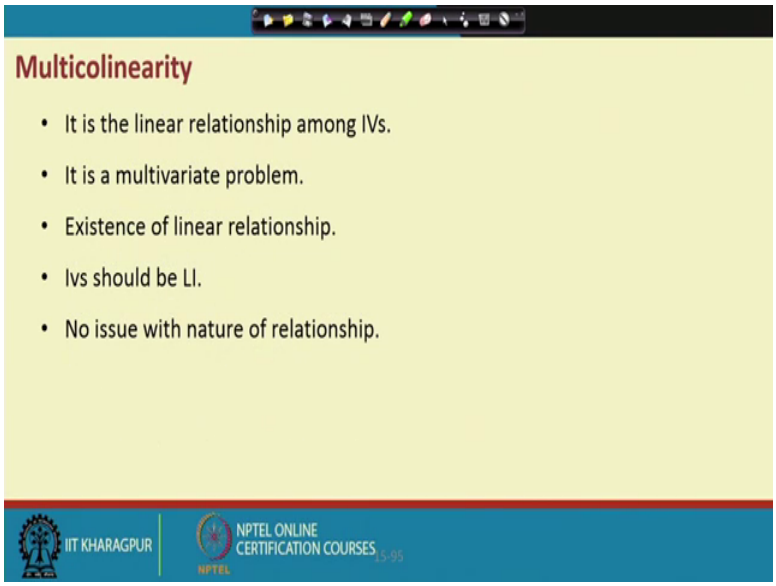
So, when we have a n number of you know models with in you know different structure and different kind of you know features with the same variables and the same problem then ultimately it is you to decide which particular model final will be picked up to go for the kind of you know prediction of protesting. Ultimately you are not in a position to use all the models to go for the prediction and forecasting. Out of several models you have to pickup particular model which is useful for the prediction and forecasting. So, which particular models you will pick up so, that needs actually model specification kind of you know structure.

So, it depends upon a many things. So, typically model selection criteria, consequence of the model specification errors when it is high so, how you to deal, test of you know specification error, error from measurements, incorrect specification of the model. So, in addition to the problems like multicollinearity, heteroscedasticity, autocorrelation so, these are all the additional kind of you know features you have to check. These are called as you know model diagnostics, before you declare that the model is completely free from all these you know virus then you can use this model for prediction and forecasting.

So, since a since we have two different units for model diagnostics so, then we will go one by one. So, starting with the you know first multicollinearity problem. So, we have to we have to go through in details about the multicollinearity, what is exactly the particular term you know items and how is it is features, how to detect, how to cure and whether it will go with you know model building or model estimations or we need to clean it completely. So, these are the things we are supposed to do before we go for the you know prediction and forecasting.

So, let us see how is this kind of you know structure so, that means, technically these are the major model diagnostics and, then we will start with the a multicollinearity problem first.

(Refer Slide Time: 24:15)



So, as we have already discussed you know multicollinearity is a problem of you know independent variable. Since a it is a kind of you know you know structure where we have one dependent variable with the many independent variable so, these variable should be technically independent. So, that means, technically multicollinearity is a kind of you know game where we like to check whether there is a linear relationship among the regressions or any relationship among the independent variable. So, the issue is very complicated and let us go through by first with if the features of this multicollinearity, then you think about the detections and the kind of you know tolerance and the kind of you know the solutions and then the kind of you know overall structure of the model validations.

So, first feature of the multicollinearity is a it is the linear relationship among the independent variables and then a it is a multivariate problem and; that means, technically we have two different setups which we have already discussed. One setup we have two variable one dependent and one independent, another setup we have one dependent within an independent. But, ultimately a in the second case the problem of multicollinearity can be detected and where we have a you know many independent variables; that means, can be you know checked when a model involves at least two independent variables. Otherwise a this problem may not be there in the system. For instance if we use bivariate modelling for any kind of you know predictional forecasting by default the kind of you know multicollinearity will not arise there.

So, third point is the existence of linear relationships. So, we are not looking for non-linearity part and we are checking the linear relationship among the regression by using either covariance or correlation or any other you know advance criteria through which you can get to know whether there is a relationship among the regression and how is this particular relationship and that is the kind of you know degree through which you can actually predict the environment.

Then independent variable should be you know you know linearly independent. So, that means, technically if you go by vector algebra. So, for as a multivariate problem is concerned so, the variables in the multivariate setups are you know two different clusters, one is the linear dependency and linear independency. So, if these variables are linearly dependent then this is having actually multicollinearity problem. If these variables are linearly independent then this is having non multicollinearity problem. So, we like to check whether the basket having independent variables are linearly dependent or linearly independent.

So, our requirement or OLS requirement is that you know these variables should be linearly independent, but in a reality you may not find that particular you know situation because reality something different. So, as a result so, you will not find that you know all these independent variables are actually linearly independent. So, there is a some relationship and we try to check whether the relationship is statistically significant or it at the tolerance levels. If it is going you know above the tolerance levels or statistically significant then it is a major issue. So, before you go for the prediction and forecasting you first actually each you know thoroughly check the particular items and try to clean the particular virus and then go ahead with the prediction and forecasting.

Then finally, no issue with the nature of the relationship; so, since multicollinearity existence of linear relationship among regression so, surprise a correlation is concerned covariance is concerned. So, the relationship can be tracked with you know correlation coefficient and covariance coefficient. So, we have now easiness whether the correlation coefficient is coming negative or covariance component is coming negative or correlation coefficient is positive so, if the particular item is actually statistically significance; that means, the correlation coefficient whether it is a positive or negative is coming statistical significant then this is a major issue.

So, whether it is a positive nature so, negative nature so, that that is a material, but alternately we like to see whether the depth of relationship or the existence of relationship, linear relationship among the regressions are statistically significant or not. If there statistically significant then this particular models cannot be used for production and forecasting. So, we try to minimize this particular you know 0. Even you cannot make it 0; so, it should be very low, so that we can actually validate the kind of you know you know models and then go ahead with the prediction and forecasting.

So, now, ultimate is actually how to you know go ahead with this kind of you know detection means; that means, technically we have two different environmental together. In one situation it is 0 and another situation it is 1 by probability. So, when the independent variables are linearly independent. So, then this is the case of you know mean this is the situation of 0 and when there in a linearly independent completely, then this is the case of you know 1. So, that means, say the game is a you know one extreme is the 0 and another extreme is the 1.

So, when the linear relationship among the independent variables are coming actually 0 in each case then you are in good track, but when it is coming actually 1 then, that means, technically it is a perfect relationship among the regression. So, and that is completely wrong term, but in reality you may not actually 0 situations or 1 situation you may be in between. So, in between means whether you are actually above you know 0.5 positively or negatively or below than 0.5. Even if it is below than 0.5 that is the correlation coefficients, still you have to check the significance level. Until and unless the particular component is statistical significant you should not use that particular model for the prediction and forecasting.

So, how you have to you know detect and how you have to proceed for the kind of you know solution we will discuss in details in the next lectures. So, in the mean time we will stop here and have a nice day.

Thank you very much.