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

Lecture – 25
Modelling Diagnostics (Contd.)

Hello everybody, this is Rudra Pradhan here, welcome to Engineering Econometrics. Today we will continue with Modelling Diagnostics and that too the coverage is on multicollinearity problem. In the last couple of lectures we have already discussed this particular component that is the multicollinearity problem. In the simple you know language multicollinearity is a multivariate problems, where the issue is the to know the existence of linear relationship among the regressors. As per the regression modelling requirement or the requirement of ordinary least square mechanism, there should not be any correlation among the regressors.

But in reality when we will be dealing with the a real life problems, any kind of you know engineering problems you will find most of the variables are you know more or less correlated. Our job is to check how is the degree of correlation among the regressors and how what extent we can tolerate, and what extent we like to change it you know level of modeling. So, that the problem of multicollinearity can be minimized and the estimated model can be used for the prediction and forecasting.

So, the problem will be more in more and more interesting, when the number of regressors are very high. So, we can start with a simple structure and then we will analyze how is this particular complexity and ultimately, how we can get the kind of you know solution. In the last lecture specifically we have discussed you know the nature of the multicollinearity the kind of you know detection criteria, and the kind of you know solution. So, we have a you know you know by the way we know what is exactly multicollinearity, and we know you know how to detect the kind of you know multicollinearity.

And also we have gone through certain you know structure through which you can also solve the multicollinearity, not completely you know 0 relationship among the regressors, but somehow we have discussed something how to minimize this particular

you know problems in order, to in order to justify that the estimated model can be used for the prediction and forecasting.

So, the solutions which we have discussed earlier like you know are like this, a first you can increase the sample size, you can drop the collinear variable, you can change the functional form, you can change the structural of the structure of the problem we can use different techniques like factor analysis principle component analysis. So, like you know we have couple of you know mechanisms through which you can actually solve the multicollinearity problem.

But you know these are all you know kind of you know very a supportive kind of you know structure through which without any change of the modeling structure, still we like to solve the multicollinearity problem that too minimize the problems in such a way that the model can be used for the prediction and forecasting but when the problem is a too complex. So, increase in sample size decrease in sample size or change of functional form, change of the structure of the problems. So, we will we will not you know solve the problem, even you know use of you know some other technique also may not solve the problem.

So, ultimately we have 2 different you know extremes what we can call, one is called as you know the use of the techniques like principal component and factor analysis, which is very specifically used to you know to drop the variables means say to reduce the variable in some extent. Because the literally meaning of factor analysis is that it brings you know new cluster, which is a smaller than the original cluster and the new cluster will be the linear combination of the original ones. So that means, like you know instead of you know using you know 20 variables, we can have a 4 to 5 factors and these 5 factors are linear combination of these 20 variables.

Somehow you know we have to fix in such a way which factors can you know you know club you know or connect to these 20 variables. So, that the problem of multicollinearity by default will be solved and usually factor analysis applied to solve the multicollinearity problem and in another extent, when your sample size is a you know lesser enough than the number of variables. So, we can use also factor analysis. And besides factor analysis and that too the use of principal component analysis, one of the strongest mechanism to solve the multicollinearity problem is to use the stepwise regression.

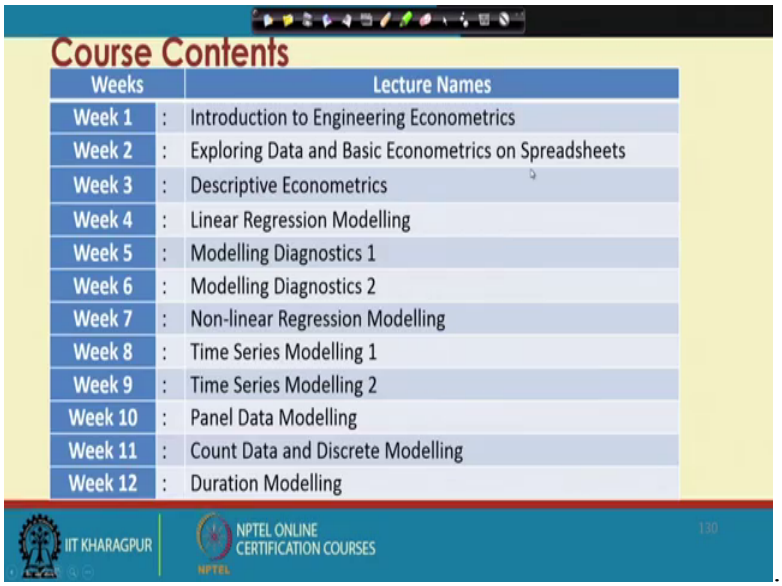
So, usually we use stepwise regression where that is the last step of the particular process, when other options are not actually it can be a you know helpful to solve this particular you know problem. So, ultimately last step is the step use of you know stepwise regression, and this lectures specifically it talk you know will talking about the stepwise regression, because it is one of the fantastic kind of you know structure through which we can solve the multicollinearity problem and we can bring a particular you know estimated model, which is exclusively very useful for the prediction and you know forecasting.

So, far as stepwise regression is concerned it is a step by step process and what is that step by step process? We have 2 different structural together one particular structure is called as a forward integration mechanism and the other one is called as a backward integration mechanism. In the forward integration mechanism we start with you know dependent variable to independent variable, one independent variable then subsequently you can add one after another independent variables and then in the same time you have to check the model there is a improvement of the model and finally, you have to stop where you know the model you know the addition of you know new variable will not improve the estimated model.

So, that is how the process of you know forward mechanism. In the backward mechanism we start the full model then you try to drop one after another variable and in the same times you check whether the estimated model will be coming something you know good or you know something best compared to the previous one. Then you will stop at a particular point of time where you know the model will be considered as the best to you know go for the kind of you know prediction and forecasting.

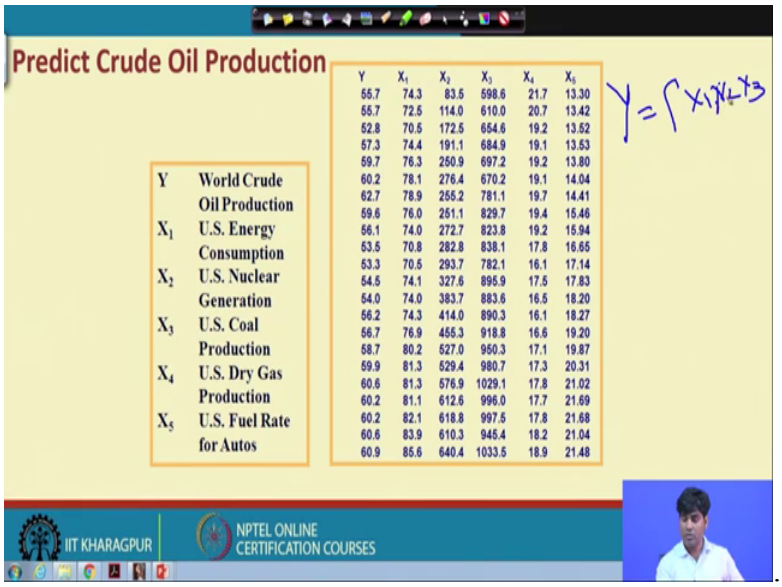
So, what will you do? So, let me give you the kind of you know snapshots, how is the kind of you know backward process and how is the kind of you know forward process and how you can you know proceed step by step for instance. So, we can start with you know 2 5 variables.

(Refer Slide Time: 08:08)



So, this is what the 5 variable case which you have already discussed in the last lecture.

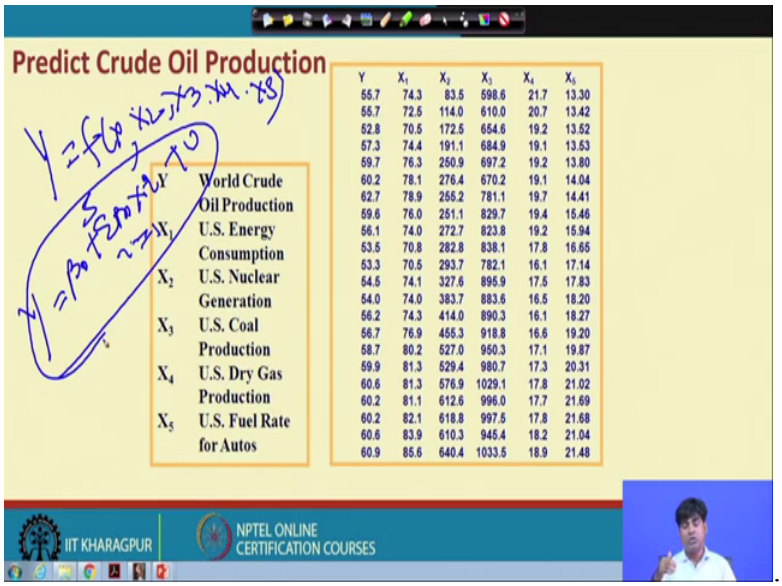
(Refer Slide Time: 08:11)



So, you can you have a 5 variables. So, without any you know hint you just you know allow the software to have the estimated model, that too with you know full setups.

So, if you say full setup then it is nothing, but called as you know Y equal to simply a function of X 2 X 3, then this is this is something coming wrong ok.

(Refer Slide Time: 08:50)



So, I will write here Y equal to function of Y equal to function of X 1 X 2, then X 3, X 4 and X 5 ok. So, this is the full model and as usual so, the model can be structured like this, you can formulate the model like this Y equal to beta 0 summation beta i X i, i equal to 1 to 5 and this is plus error term.

So, this is how the you know multivariate model or multiple regression model with 5 variables, but ultimately this is what actually in the form of you know mathematical structure and we need actually the estimated structure or the kind of you know model, which we can use for the prediction and forecasting. So, when you talk about the first model, then by default a means the original estimated models. So, then you have the structure like Y equal to beta 0 plus beta 1 X 1 beta 2, X 2 beta 3 X 3 beta 4 X 4 beta 5 X 5.

Now you know once you have the full you know estimated model with you know full setup; that means, Y with you know all these 5 independent variables, then you first check whether the model is or not in the sense that is with respect to multicollinearity. So, the I inspection is that if the model is ok, then all most of the variables or all the variables should be statistically significant, and the goodness of fit of the model will be also very high, that is that is a just through the component called as you know coefficient of determination that is that is R square.

So, ultimately in the first instance model can be considered as the best, if all these parameters or all the variables are statistically significant, and overall fitness of the model will be also very good that is indicated by a R square capital R square that is the coefficient of determination. So, now, the chance of multicollinearity will arise when R square will be high, but most of the variables are not significant, the vice versa is also true when most of the variables are statistically significant R square is it low.

But what is the best or what is what should be the kind of you know you know typical structure? Now when R square will be high most of the variables should be statistically significant, when R square will be low then there is no means its not necessary that you know all the variables will be statistically significant, but at least few variables should be statistically significant. If no variables are statistically significant, then we should not be use even R square is also high; that means, goodness of fit will be very high.

So, we will we will choose a situation, where most of the variables are statistically significant and R square will be also very high. But if other way around R square high and few t ratios are you know significant and all t ratios are significant R square are low. So, in that case or in those 2 situations, we can you know expect that there is a kind of you know multicollinearity problem. But if all the variables are statistically significant, R square is you know (Refer Time: 12:37) high and most of the variables are not significant or very few variables are significant and in the same times R square is low, then the even the model is not good, but still the chance of multicollinearity is not arise.

So, but there in between the situation when R square is high and most of the variables are not significant and most of the variables are significant when R square is low, so, these are the 2 you know instances without any you know check you know there in fact, there are criteria how to detect the multicollinearity, which we have already discussed like a variance in protein factors or something like you know tolerance factors. So, many components we have discussed through which you can detect the multicollinearity problem or the degree of multicollinearity.

For instance VIF; if the VI value will cross 10 then by default there is a multicollinearity problem and that too severe multicollinearity. If VI factor is less than 10 there is a kind of you know multicollinearity, but it is a less severe and it is close to 0 means the you

know very low multicollinearity and 10 above means degree of multicollinearity will start you know increasing.

So, now assuming that this problem is having actually multicollinearity problem, then we have tried here and there by increasing by using the sample increasing mechanism or sample decreasing mechanism or some of the other components data transformations change of the functional form in a you know by using all these you know criteria, we are not in a position to get a estimated models which is good and used for the prediction and forecasting in that context what will you do? So, we will go by stepwise regressions, and then to find out which one is the best model and what are the variables actually should be included in the system so, that the model can be used for the prediction and forecasting.

Ultimately the prediction is above to the dependent variable, whether we are you know including all the independent variables or we are including few independent variables that is not big deal, but ultimately while predicting dependent variables your model estimated model should be reliables, and is good for the estimation and predictions. Even if you are including actually 5 variables and variables are not significant, then that model should not be used for the prediction and forecasting. Instead of using 5 variables if your model is having only 2 variables or 3 variables, then you know and all these variables are statistically significant still that model can be used for the prediction and forecasting.

Without looking the statistical result or econometric results the theoretical understanding is that you know if more number of variables are involved in the process, then the prediction accuracy will be very high, but that should be supported by the data and the kind of you know estimated models. It is not always true that you know more the variables good is the you know consistency of the prediction, it is the prime you know conditions or necessary condition, but the second is sufficient condition or secondary condition is that. So, you have to fix a model and that model should be used for the prediction and forecasting, where the variable should be statistically significant.

And stepwise is a kind of you know mechanisms by default it will operate in such a way, only significant variables should be in the systems and remaining variables by default will be out in the system, that is the beauty of this particular you know technique. So; that means, how to start this particular stepwise process and how to end this stepwise process. Yes, the term stepwise itself indicates that you know it is a step by step process,

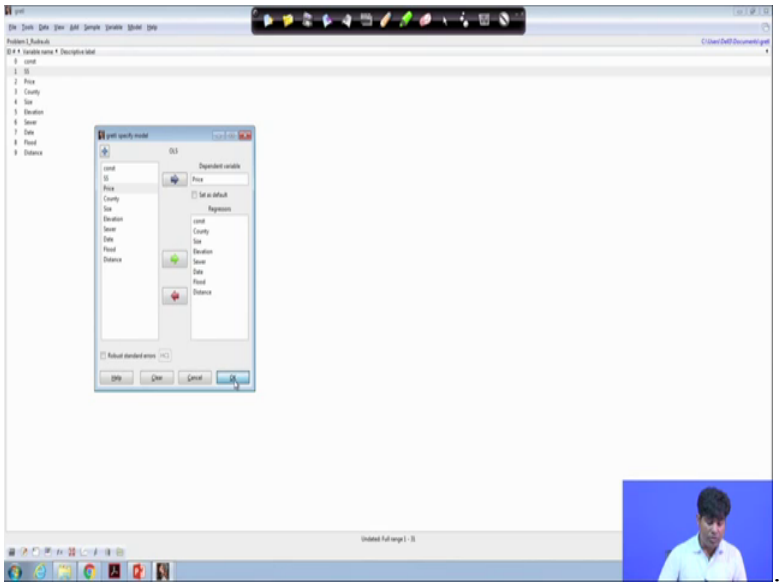
but how is the step by step process and how we can actually move one point to another point? It is in management language it is like you know you know it is a question of you know opportunity cost. So, it is the comparison means the stepwise regression will give you indication that you know, there will be alternatives and you have to choose the best then by default the other one will be kicked out.

So, that is how we have to actually follow the particular mechanism and then finally, we choose a model, which is good for the prediction and forecasting. So, its theoretically again very clear to very clears about the kind of you know structure the understanding, but practically how you have to deal with this kind of you know problems and how you have to use this stepwise and come to the kind of you know conclusion that is that is very interesting and that is that is the key actually to follow ok.

. So, what we will do in the stepwise process. So, I will I will start with first forward and then I will go to the backward in the forward process. So, you start running the full model and then check, which one is the most you know most significant and most important variable and that will be decided through the significance of the t ratio, and the value of the t ratio. For instance beta 1 beta 1 X 1 beta 2 X 2, I can say X 1 is more important than X 2 provided both are significant, but the value of you know beta 1 is much higher compared to you know beta 2.

So, then beta X 1 can be considered as the most important variable so; that means, technically we like to see which one is the most important variables to predict the Y. So, when we are running the multiple regression, then we will find estimated coefficients we will find standard errors we will find the kind of you know t coefficients. So, let me give an example. So, this is what actually the software's till now which we have used is called as you know excels spreadsheet and that too there is a data analysis package.

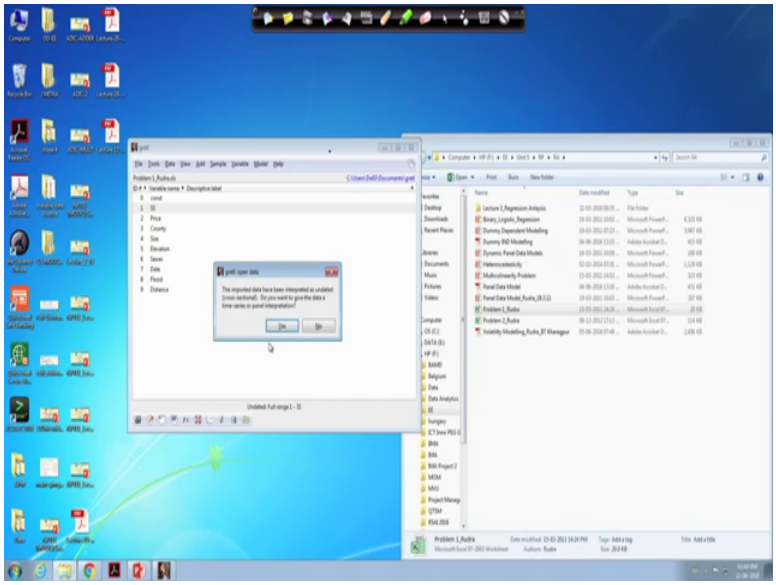
(Refer Slide Time: 18:52)



And from the data analysis package, we can actually go to the regression and then we are building the models and getting the kind of you know solution.

But now I will be solving the same problem with different software's. So, what will you do? So, you simply actually I am closing this particular you know software now I will start a fresh ok.

(Refer Slide Time: 19:12).



So, this is a software econometric software called as a gretl, and its a free softwares and you just download and you know install and you know like r, it is very easy and you know here you need not required to write programming, its a window based software just you know install and it will helpful for the entire you know the particular you know subject.

So, the works of the gretl will look like this and what we will do? Whatever we have discussed right now starting with a bivariate multiple regressions and the problems associated with multiple regressions; that means, there are 2 step process all together till now, the problems found problems the kind of you know foundations that is what is the exact you know objective to you know predict or the kind of you know forecasting. So, it should be followed by variables identification data availability, then you have to build a functional form, then we use the software or you can go by the manual process.

Ultimately we need a estimated model, which can used for the prediction and forecasting of the dependent variable. So, we have gone through the bivariate process, we have gone through multiple process, multiple regression process, but every times whether it is a bivariate trivariate or multivariate. So, once the problem is clear data is available and by the help of you know various mechanism, you know get the estimated model then ultimately after getting the estimated model, we have to check whether the model is actually you are not.

So, for that we have gone through specification test, goodness of fit test and then the diagnostic test which we are going through the kind of you know multicollinearity problem and some of the other components, but right now we have discussed up to multicollinearity problem. So, bivariate and then trivariate and multivariate and then passes through all the you know tests like specification test, goodness fit test and diagnostic test. So, what I will do? I will pick up a particular problem and then I will handle how these estimations are coming and finally, how to detect the multicollinearity problem and again how to get the you know revised models, where the multicollinearity problem will not be there.

So, what will we do? So, like you know regression you know excel spreadsheet case you have to just go and install the data analysis package, then you know ask the menu bar to go for the regressions and then indicate the dependent independent variables and finally,

you will get the estimated output with all these you know specification parameter, goodness fit parameters and the diagnostic part of this particular problem specifically directly not visible. So, you have to understand first, then you know a slightly you have to check yourself then you known try to improve or try to minimize all these problems as much as possible.

But specification test goodness fit test by default estimated model will give you some outputs like you know, beta parameters then standard error of the beta parameter t statistic then R square adjusted R square coefficient of you know f statistics. So, there is a component called as a adjusted r square. So, when you add one after another variable into the system, then R square will start indefinite means increasing indefinitely and because it will not adjust the degree of freedom, but when we when we actually a use adjusted R square, then the variables involvement and sample size will be adjusted simultaneously, then finally, it will give you a particular component, which is called as a adjusted R square and that will be used as a goodness fit indicator or the judgment of the estimated models or the reliability of the model.

So ultimately how to go about this particular process? So, let us say this is a menu bar and we have here actually a some kind of you know dataset. So, I am just bringing this dataset to this particular box, and the moment I will be bringing, then it will come the transformation will be like this you just put and the menu bar will ask you the imported data have been interpreted as undated, and that is cross sectional do you want to give the data a time series or panel interpretation. What we have already discussed? This is how the message will come in the toolbar. So, ultimately when you allow the software to estimate, then software definitely will ask whether the particular dataset is the cross sectional type or time series type, full type or panel type.

In fact, this gretl will ask you only 3 things, cross sectional, time and panel. So, when you import the data by default it will be read as you know cross sectional data. And if your actual data is time series then you specify the time series, then you indicate the time series which range to which range because there is need of you know consistency. And again when there is a kind of you know panel data, again you have to give the command up to how many years and for how many county. So, you have to give the indication, then after that the software by default will read as a panel data or time series data or cross sectional data, then it will help you to solve the problem very easily.

So, now let us assume that we have this data and we are actually just imported and ask a software is asking whether it is a cross sectional data or you have to give time you know time indications or you know panel indications. So, let us say no so; that means, if you say no, then the software will read the data that you know it is a cross sectional type, then you will go for the estimation. And if you go for the estimation if you again go to the toolbar, then you will find there are many items file tool data view add samples variables and model ultimately this is a regression modelling and this package is exclusively for you know regression modeling, and you will find a model and you click the model you will find you will plenty of you know options are there starting with the ordinary least squares, instrumental variables other linear models, limited dependent variable models these are all things will be coming over the time.

But in the first instant till now whatever we have discussed, it is mostly with respect to OLS mechanism that is ordinary least square. And in this menu bar the first the first option is the ordinary least squares. So, you have to just click actually. And that whatever we have discussed till now it is only, on the basis of OLS mechanism only and OLS technique has its you know limitation the these are all called as you know assumption of OLS, and we are testing all these assumptions and these test these test will give you the diagnostics or something kind of reliability of the estimate if all these you know assumptions are not satisfied with you know OLS criteria then the OLS estimated results will not be used for the prediction and forecasting.

So, that is how we are in the process of testing and bring the things. So, now, after importing the data, allowing the software to read as a cross sectional data then you choose a models that is through OLS mechanism that; that means, up to whatever we have already discussed is through OLS whether it is a bivariate trivariate or multivariate, but every times we are discussing, with the OLS. Means up to till now and in this case there a couple of variables in the left hand side box. And if you see in the excel package and you know this gretl package you will find drastic difference, in the in the excel spreadsheet we have the regression package just click and indicate the variables. Again in the same box in the upper one is the dependent variables lower one is the independent variable.

But here the box is little slightly different in the left hand side it will give you the give you the indication about the input box, and in the right side you will see here in the right

side this is the bar and this is the input box, and this is the command box, where you have to give the command then automatically software will process the output.

. So, in the input box you have the variable indication, now in the right hand side of the this you know circles you will find there are 2 different box one box is dependent variable which is smaller one, then there is actually regressors box which is actual bigger one. So, the smaller one because it is the dependent variables and every times, whether it is a bivariate, trivariate and multiple. So, you have only one dependent variable. So, that is how the box is very small one.

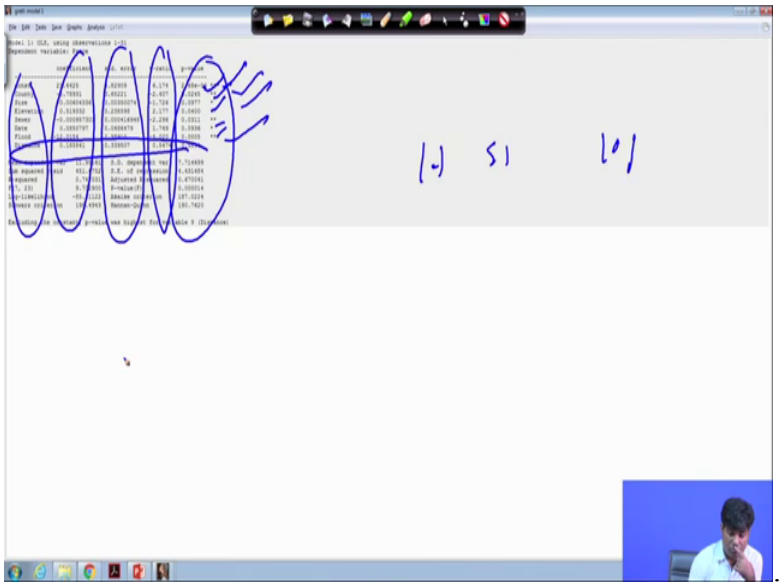
So, you just allow the particular dependent variable to that box, and then rest of the independent variables should come to this area. So, then you allow the software to go ahead with the estimation process. Because when we use any software's, so, it is you to actually you know give the command and you know process the kind of you know software, automatically will not read all these you know instruction. So, it is you to rotate actually the kind of you know process.

So, what will you do here? So, we will just clean and then you put actually let us say price is a kind of you know this is a civil engineering problem, and that too you know housing price determination, what are the factor which can actually affect the housing price. So, there are many a you know factors which can actually affect the housing price, that depends upon you know size of the house, the kind of you now facility of the house distance from the market and distance from the kind of you know sea area, then how is the kind of you know locations it is a (Refer Time: 29:23) location or something.

Likewise, we have actually different actually variables identified and that too on the basis of you know theory, and then you allow actually the kind of you know all these variable to the input box. So, you will find all these input box you know coming like this flood, distance ok. So, the all these independent variable now in the box, command box and then allow the software to run the models. So, you just put and by default ok. So, dependent variable is not allowed.

So, now, so, dependent variable is fixed and independent variables are fixed now you put ok.

(Refer Slide Time: 30:10)



So, now after putting the ok, you find you know this is what the regression result, you will find plenty of regressions where see here, so, as usual. So, these are all variable indication, these are all variable indication this is the coefficient beta coefficients and these are all standard error of the beta coefficient and these are all t statistics and this is the probability level of significance.

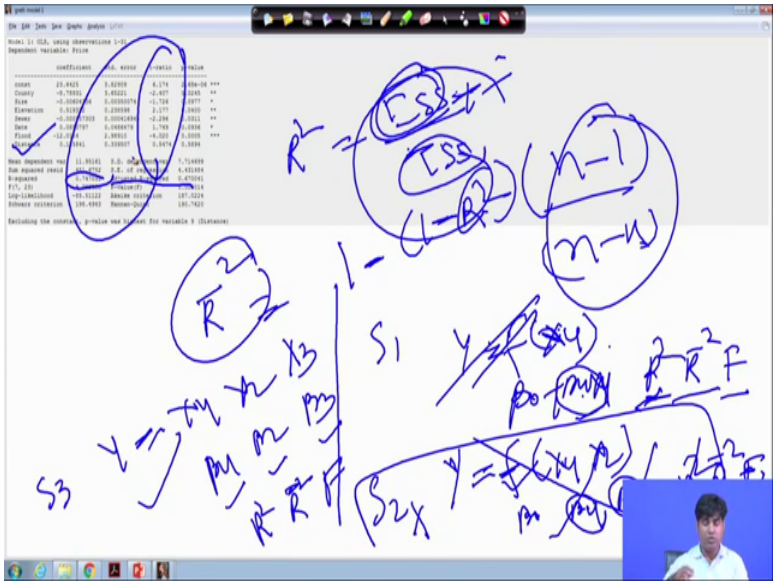
So, now most of the variables I mean say most of the occasions you like to check the significance of variables at 3 levels 1 percent, 5 percent and 10 percent 10 percent. So, what will you do ultimately? a 1 percent 2 5 percent and 10 percent any variable in statistically significant 1 percent means it is the best if not then that we can check it 5 percent, if you cannot then again you can go to the 10 percent, so that means, if variable is a significant at 1 percent it is the best, if not then 5 percent is the best if not then 10 percent will be the final choice. Then after that statistically we cannot go oh you know proceed further, but software will give you the exact probability level through which you can actually reject the true null hypothesis.

So, what we can do here? So, already software calculated everything. So, these are all beta coefficient for variables independent variables, and these are all standard errors and coefficients and standard errors and t statistic and the p value. So, now, you will find this model is actually very good ones in the sense most of the variables are actually

statistically significant except the last variable, which is actually distance according to the housing price.

So; that means, technically since and then this gives you now significance of the para parameters or significance of the variable. Since most of the variables are statistically significant, then R square should be very high and this will be followed by this will be followed by a you know this is this is why actually followed by 0.75 which is actually good enough and corresponding to R square adjusted R square is also 0.67.

(Refer Slide Time: 32:27).



So, adjusted R square actually we know R square R square is actually ratio between (Refer Time:32:43) sum of squares by total sum of the square TSS and then adjusted R square is nothing, but 1 minus 1 minus one minus R square 1 minus R square into n minus 1 by n minus k. So, this is this is what the you know on the basis of you know formula you can calculate.

And ultimately a software will help you to you know get this particular value. So, this is R square and this is the adjustment factor; otherwise R square will start increasing when we add one after another variable into the system. Because ESS is nothing, but actually weightage of you know independent variable for instance if it is bivariate, then the weightage will be summation Y X 1 by TSS when it is a trivariate then summation Y X 1 plus summation Y X 2 divided by TSS so; that means, TSS the lower part will be remain constant.

So, when you add one after another variable the first variable weight will be one, then add another variable then suppose additional variable weight will be you know added; that means, technically upper side will start you know increasing, because the component TSS is always actually positive one. So, as a result any new variable means, it will be positive support to the impact.

So, as a results the ratio will be very high, when you add one after another variable into the system; that means, technically R square will be very high. So, that is why when we have more number of variables or multiple regressions or multivariate regression R square is not used as a good indicator for the model judgment, in that case adjusted R square that is the a which is calculated like this will be used for the prediction and forecasting.

So, now ultimately in this models only one variable is not significant so, that is why a you may not for you know stepwise, to follow the particular you know process. But if you I will tell you the exact structure you know when you need actually stepwise, even if this model can be used stepwise because one way we can use that you know since a different variables have a different significance level. So, we like to fix a model you know where all the variables will be statistically at 1 percent, this is another way to use this stepwise.

Or else most of the variables are not statistically significant, still we use stepwise to find out the situation how to actually deal with a problem. So, in that context what will you do? So, you see here. So, this column you have to actually target. So, this column will give you the indication that you know, which variable is having very high impact to the dependent variable that depends upon you know value of you know t ratio, higher the value of t ratio higher is the impact lower the t value lower impact.

So; that means, without actually you know checking the sign, because whether positive negative that that is not a you know matters matter is a which one is a more you know impercolates the dependent variable. So, it depends it exclusively depends on the value of the t statistics so; that means, technically after the estimation you just you know check all the beta coefficient, and the t value of these beta coefficient; which one is the highest so, that model will come you know that variable will come first in the estimation process. So, this is full model.

So, now if you go by stepwise; so, in the step 1, so, the most important variable is here coming as you know flow right so; that means. So, let us in the step 1. So, there is a particular variable which is coming actually let us say fourth variable is coming as a most significant variable. So, X 4 is the first model. So, ultimately your model outcome will be beta 0, beta 0 plus beta 4 X 4 and then you will have R square, adjusted R square and f then I am very sure in the multiple context you know X 4 is coming very high very high impact and it is high statistically significant.

As a result when you estimate the dependent variable again with respect to that variable that dependent independent variable only then the impact will be much higher again. So, as a result by default this beta 4 will be statistically significant, and will give you R square value and also adjusted R square and f. In the first instance we need not required to compare, but now you go to the step 2; in the step 2 the next most important variable this is the on the basis of you know t statistics.

So; that means, technically you just put you know ascending to descending order the impact a highest t to lowest t. So, that automatically give you the impact of a particular variable to the dependent variable. So, now, let us say that you know S 2 is the next one which is having you know high impact to the dependent variable, after x four. So, then the next step your model will be Y equal to Y equal to f of f of X 4 and X 2 and then finally, your model estimated model will be beta 0 beta 4 and beta 2.

And then we have a adjusted R square R square and F. So, now, now since already beta 4 is statistically significant. So, here you check whether beta 4 is still statistically significant or not, and again beta 2 is coming statistically significant or not. If a beta 4 statistically significant and beta 2 statistically significant and R square is improving compared to the first one, because you are adding one another variable R square is improving by default it will improving, but what is important is to check adjusted R square whether it is increasing compared to the previous one, and f will be also increasing and that too f should be statistically significant.

If this is significant, this is significant, this is high, this is high, this is high then technically this model can be considered as the best compared to the previous one. So, now, between these 2 model this will be now rejected and this will be finally, accepted again this is not the ultimate model, then because there are more number of variables are

there. So, now, you check what is the next variable means which one is the next most very you know important variable to predict the y.

So, let us assume that you say S 3 again. So, that then you go to the step 3 then your regression will be Y versus X 4 X 2 and X 3 So, again, so, beta 4, beta 2 and beta 3 and again. So, what will you do? So, you have to check the significance levels, significance level and significance level and then R square adjusted R square and adjusted R square and then F. So, if you will find all these coefficients are statistically significant and R square is high, adjusted R square is high and f is coming statistically significant, then this model will be considered as the best compared to the previous one this one this one.

So, then in that case this will be rejected and this will be finally, accepted to predict Y and that too with respect to X 4 X 2 X 3. Then again in the step 4 you will check which one is the again most important variables as per the ordering and that variable again included into the system and estimate the process and then check whether the variables are still statistically significant, R square is improving, adjusted R square is improving and F is improving. And if that is the case then you keep that model and then the previous model will be by default rejected.

But after adding another variable new variable let us say X 1 and the impact of other variable start declining and not significant and then R square you know adjusted R square is declining, and f is declining in that case. So, you have to stop and then you can reject this one then previous model is already declared as the good models, and that can be used for the prediction and forecasting. So, this is how the you know forward process in the backward process you start with you know full model and then check which one is the least impact to the dependent variables for instance in this case distance is coming the least variable.

So, you have to re estimate the model by dropping the distance component, and then you know have the estimated output now this is the original output then after dropping the distance you will find another you know set up in the output. Now you compare in this case how many variables are significant and at what level again you have to see how many variables are statistically significant at what level.

So; that means, if dropping distance and your model still most of the variables are statistically significant and R square is a R square adjusted, R square and F is coming fair

enough, then in that case the second model will be best and then you reject this one. And again you can go ahead with a or you know dropping another variable, which is least impact and then again re estimate and then check the results to the previous ones and compare which one is the best. If the model is improving; obviously, if you know unnecessary variable, if you drop then adjusted R square will start increasing the impact of variables impact of variables will be very high and again the f statistical will be statistically significant and also very high.

So, now you will continue and continue in such a way that you know the next entry or next dropping will actually affect the estimated output compared to the previous one. So, that is the stopping point how to you know stop you know or how to you know get the best models, on the basis of backward operation and forward operations.

. So, you start with the full and drop you know remove one after another, till you get the optimum model which is actually good for the estimations prediction and forecasting, where all the variables are statistically significant supported by adjusted R square and in the in the in the on the other side. In the case of forward process you add one after another variable and same times you check whether the variables are coming statistically significant, and that too improve the model with respect to adjusted R square. And you will stop at a point where you know the variables impact will start declining and adjusted R square value will also start declining.

So, in that context, you know you have to fix you have to stop and then use the previous model for the prediction and forecasting. So that means, this very interesting technique to get the estimated model by the process of elimination. So, something like that. So, that the estimated model can be declared as the best models or line of at the particular line will be declared as the line of the best fit, and which is a means very easily or you can you can you know typically use for the prediction and forecasting as per the particular you know engineering problems requirement.

So, this all about the multicollinearity issue at. So, till now we have discussed what is that exact problem and how to how to detect this, how to solve this and what are the reasons through which multicollinearity is coming that itself will give you the indication about the kind of you know revisit the process, restructure the process and then finally,

you get estimated model, which is actually very useful for the prediction of the Y and that too you know within high accuracy, and high reliability this we will stop here.

Thank you very much, have a nice day.