Econometric Modelling
Prof. Rudra P. Pradhan
Department of Management
Indian Institute of Technology, Kharagpur

Lecture No. # 20
Multicollinearity Problem

Good afternoon, this is Rudra Pradhan here, welcome to NPTEL project on econometric modelling. Today we will discuss multicollinearity problems. In the last couple of lectures we have discussed the details about the econometric modelling that too bivariate analysis, trivariate analysis and multivariate analysis. And everywhere we visualize techniques to estimate the model and we consider for the fitness of the model so far as a best fit is concerned.

Every time we start with o l s assumptions and and with checking all these assumptions. But, perfectly we have not discussed anything in any particular problem regarding assumption till now. So, today we will highlight one of such problems after the estimation through this o l s technique. So one of the interesting problem is called as a multicollinearity; multicollinearity, itself is a multivariate problem.

So, we started with bivariate modelling, trivariate modelling and multivariate modelling. So for as a bivariate concerned there is one dependent variable and one independent variable and that is this starting point of econometric modelling. But, when we start with bivariate setup, where there is one dependent variable and another independent variable, then obviously there is no question of multicollinearity. But, when we move to trivariate model or multivariate models, then the problem of multicollinearity can be generated.

So now, we like to know how is that particular concept? What are the problems associated with this particular multicollinearity? And how we have to detect and what are the solution for this problems? Is it genuine or is it mandatory that you have to remove or you have to go ahead with these problems that we have to discuss in detail in today's lecture. So the thing is that, multicollinearity is multivariate problem and that too it is in the right side of the problems.

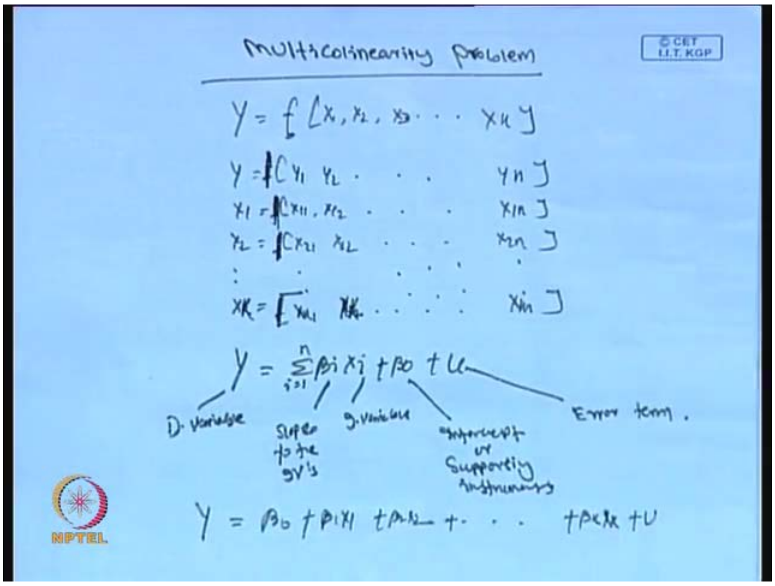
So that means, when we fit a econometric model then obviously, whether it is bivariate set up or multivariate set up, every time Y is in the left side; that is considered as a dependent variables and X(s) in the right side that is called as independent variables. If there is 1 X and there is Y then it is called bivariate. So, when 1 Y then 2 X then it is called as a trivariate, then 1 Y with more X means more than two then we called as a multivariate analysis. But, if you start with a bivariate to multivariate in every stage you find Y is in the left side that is always consider as a dependent variable and that too one every time.

So, that means we are discussing here the entire problem with respect to one dependent variable. And right now, we will discuss so many problems again with respect to one dependent variable with the one independent or many independent variables .But, there is another system called as a structural equation modelling or simultaneous equation modelling, where the dependent variables cannot be one, it can be many also.

That means in the left side there is a series of variables Y 1, Y 2 up to say Y n and in the right side there is series of independent variables X 1, X 2 up to X n. So then, we like to know how each variable are interdependent to each other. But, when you go for a pure multivariate analysis, then the classification of dependent and independent may not be very handy. Because when they are considering as interdependent to each other, then it is very difficult to analyze which one is this specific dependent variable and which are specific independent variables.

So that means, even if in the X sides if two variables are correlated each each other then obviously, within the two within the two variables; one may be dependent and another is independent. So, that is how the problem is more and more complex. So, with this basic back ground we like to start, what is all about this multicollinearity problem? So multicollinearity the term itself will give a indication that, it is a multivariate problem. Now, we start with a multivariate system then, we will discuss what is all about multicollinearity problem and how you have to sort out its solutions.

(Refer Slide Time: 05:50)



Multicolinearity problem

$$Y = f[x_1, x_2, x_3 \cdots x_k]$$

$$Y = f[Y_1 \; Y_2 \; \cdots \; Y_n]$$

$$X_1 = f[X_{11}, X_{12} \; \cdots \; X_{1n}]$$

$$X_2 = f[X_{21} \; X_{22} \; \cdots \; X_{2n}]$$

$$\vdots$$

$$X_K = f[X_{K1} \; X_{K2} \cdots \; X_{Kn}]$$

$$Y = \sum_{i=1}^{n} \beta_i X_i + \beta_0 + u$$

D. variable
Slope to the I.V's
I. variable
Intercept or Supporting instruments
Error term.
© CET I.I.T. KGP
NPTEL

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + U$$

So, let us start here with a multivariate problem. So, for a multivariate problems let us start with Y equal to function of function of X 1, X 2, X 3 up to X k, where Y is considered as a Y 1, Y 2 up to Y n and X 1 is consider as a X 1 1, X 1 2 up to X 1 n. Then, X 2 consider as a X 2 1, X 2 2 up to X 2 n and continue. Then X k is equal to that means every time there is obviously function function retained or you can say sorry Y is this much then X is this much is set of series, then this is X 1 X n equal to X. We can call it as X K X K 1, X K 2, X K 2 then continue X K n. So, this is how the system all about.

So that means, we are considering a multivariate problems where there are k number of independent variables and one dependent variable that is Y. So, Y has a sample Y 1, Y 2 up to Y n and X 1 has a sample X 1 1, X 1 2 up to X 1 k. Similarly, X 2 has a sample X 2 1, X 2 2 up to X 2 K; similarly, continue for k-th series X K 1, X K 2 up to X K 1 K n. So, that means we are considering k number of independent variables and where n represents number of observations.

So, by default the starting of this problem, the standard assumption is that all the variables have same sample of observations, that means for Y there is n sample, X 1 there is n samples, X 2 there is n samples and X k there is n samples. If the sample size is different for each variable or various variables then obviously there is a serious problem, that means the problem itself is inconsistent.

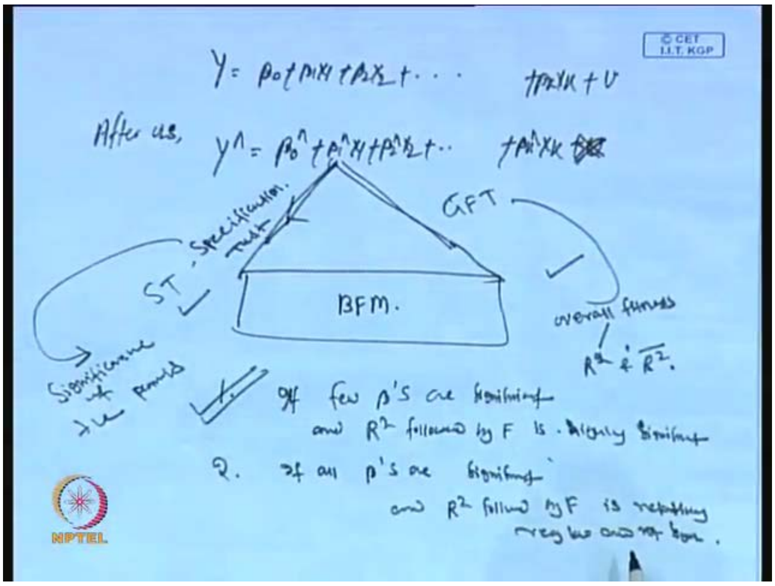
So, to make the consistent you have to apply plus minus means you decrease the sample or increase the sample, then you make the structure where all the variables has similar sample size or same sample size, so if it is so then you have to proceed. With this particular background, what is our model? Now, we will fit a multivariate model. So, Y equal to summation beta i X i, i equal to 1 to n equal plus beta 0 plus U.

So, this is the multivariate model, where Y is dependent variable, then X i is independent variable, beta is the slope slope to the independent variables, then beta 0 is the intercept that is otherwise called as a supporting component to this particular Y, supporting instruments or this U stands for error terms.

So, this is the pure multivariate systems, where Y is summation beta i X i plus beta 0 plus U, where Y is dependent variable, beta is this slope slope factor, slope components, X i is independent variable, beta 0 is intercept and U is the error term. And this model is purely in implicit format.

So, let us put it in a explicit format; so it will be coming beta 0 plus beta 1 X 1 plus beta 2 X 2 plus continue plus beta k X k plus U. So, this is the this is the pure multivariate models with k number of independent variables. Now, we have to see what is the structure of this econometric model with k independent variables and Y dependent variables? If it is so, then how is the proper structure?

(Refer Slide Time: 10:34)


© CET I.I.T. KGP
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + U$
After us,
$Y^\wedge = \beta_0^\wedge + \beta_1^\wedge X_1 + \beta_2^\wedge X_2 + \dots + \beta_k^\wedge X_k$
Specification test
ST
GFT
BFM.
overall fitness
R^2 & $\bar{R}^2$.
Significance of the params
1. If few β's are significant
and R^2 followed by F is highly significant
2. If all β's are significant
and R^2 followed by F is [illegible]
[illegible]
NPTEL

So that means, the starting point is here Y equal to beta 0 plus beta 1 X 1 plus beta 2 X 2 plus beta k X k plus U error term; error term obviously will be always there. Now this is the pure to regression equation. Now, the next step is to have the estimated models for that we have to apply so many techniques like O L S, C L S, W L S, M L E etcetera.

But, till now what we have discussed is the application of O L S technique. Again here also we have to discuss O L S technique and we have to see by using O L S technique how quickly we can have the best fitted model. So the moment we will apply O L S technique, then the technique itself based on certain assumptions and that assumptions will lead to lots of serious problems.

So, this is one of the assumptions we are going to verify with this particular multicollinearity problem. So, I will highlight what are the assumptions and how will come to these particular multicollinearity issues? In the mean times our standard multivariate model is like this, Y equal to beta 0 beta 1 X 1, beta 2 X 2 up to beta k X k all right.

So now after O L S technique after O L S technique, then the model will be transferred to transferred to estimated model. So then accordingly, the estimated model will be Y hat we call it Y hat, where Y hat equal to beta 0 hat plus beta 1 hat X 1 plus beta 2 hat X 2 plus continue beta k hat X k plus U. So, this is the standard estimated model sorry U is not here. Because the moment you will estimate, then obviously U will be removed. The way we, minimize the error sum squares then we will get the estimated model.

So obviously, in the estimated model there is no such U component; U component is only to regression line. We like to know, what is the percentage of U involved in this particular model? If the percent is very high, then the model cannot be considered as a best; so if the percentage is less means U impact is very less, then the model can be considered as the best provided other things can be on the right track.

So, in the mean times here our basic objective is to check what is the error sum and every time we have the objective to minimize the error sum squares and the way we will minimize the error sum square we will automatically get the best fitted model; that is the process or you can say systematic approach to econometric modelling.

So now, the moment you will get this particular issue then obviously, there are two steps for you. It will give you two different paths, to go for the best fitted models. So means once you get the estimated model, you are not sure whether this model is perfectly ok for you are not. So, to know this this particular task, you have two different paths and ultimately, you have to go two different paths and finally you will come to a conclusion.

That means, you have to start from this side and you have start from this side ultimately, you will join some place then ultimately you will get the conclusions. So this particular side is called as specification test. So, in this side you have to go for specification test, in this side you have to go for G F T goodness fit test. It is otherwise called as a S T specification test.

So, once you have the estimated model. So the moment, the moment you get the estimated model then of course, you have to test it. Whether it is consider as the best models and can be used for policy purpose? So for that, you have to go for specification test and goodness fit test. So, what is mean by specification test? That means, this specification test is a nothing but to check the significance of the parameters significance of the parameters. The specification test objective is to check whether the parameters are statistically significant.

G F T test is done to know the significance of the overall fitness of the model, so it will give you the significance of overall fitness of the model so, that too it does by R square and adjusted R square. So R square is coefficient of determination and accordingly, we have to see whether the R square is statistically significant or not? So, in the specification test we usually use t statistic to check the significance of the parameters and in the other side G F T we use F statistics to check the overall fitness of the model.

So that means, R square should be statistically significant in the right side for G F T goodness fit test. And in the other side, we have to apply t statistics to know the significance of each parameter at individual rate. So that means, we like to know beta 1, beta 2, beta 3 like this individual coefficients should be statistically significant and for that we have to use t statistics and for overall fitness of the test we have to use the R square statistics.

So, together we have to see how quickly we will have the solution, if both will go in the same way or as per your expectation then obviously, the model can be considered as the best. So that means, if it is true if it is true means if the condition is satisfied and if this condition is satisfied, then we will get the best fitted model best fitted model.

So, this particular model is called as a best fitted model. However, in reality the situation is something other way round. That means, your expectation is to make the estimated model or to have the best fitted models as per the significance of the parameters and overall fitness of the models. This is your expectation, but it may not be possible sometimes because some of the components can divert from the objectives. So what are these situations? So that situation may lead to serious issues or serious problems in the econometric modelling.

So that means, now we have two different specifications. Means all together two different problems; so what is the problem? First problem is so that means when we will call the model best fitted model, then it must it must fulfill 2 conditions. First condition is all the parameters should be statistically significant and R square must be statistically significant at a higher rate.

So that means, that means F statistic F statistic should be absolutely very high and here t of all these beta coefficients should be absolutely high. And at the end all the beta hats should be significant at the high rate and at the end the R square itself will be highly significant at a higher rate. If not, then there is a problem and that means that model cannot be considered as the best fitted model. So that means, now the question is what are such problems?

So then, we have to check what are the possibilities? From these possibilities you will face the problem. So that means, we have two different components all together; one is significance of parameter and overall fitness of the model. Now, there are few chances that all parameters cannot be statistically significant or say R square cannot be statistically significant. So, in the first case if few beta(s) are significant few beta(s) are significant and R squares followed by and R square followed by F is highly significant is highly significant. That means here we are we are going together.

So, that means so we are going overall fitness of the model and we are going to the specification test. That means in the first, if all these betas are significant and R square is significant, then we have the best fitted model. So, we have to go we have to go ahead with this property means forecasting issue etcetera. But if few beta(s) are significant and

R square is satisfied here that is highly significant, then the problem is very complex. You cannot say that this model will be considered as the best, so this is one such issue, where model cannot be considered as the best. Second issue: if all beta(s) all beta(s) are significant and R square followed by F is relatively very low relatively very low and not significant and not significant respectively

So, that means all the parameter should be significant and R square should be significant. Now, if not then there are two different possibilities. So, first possibility is let us say out of all beta(s) few beta(s) are significant and other side R square is substantially very high, that means the value of R square will be very close to 1so, that means it is extreme end.

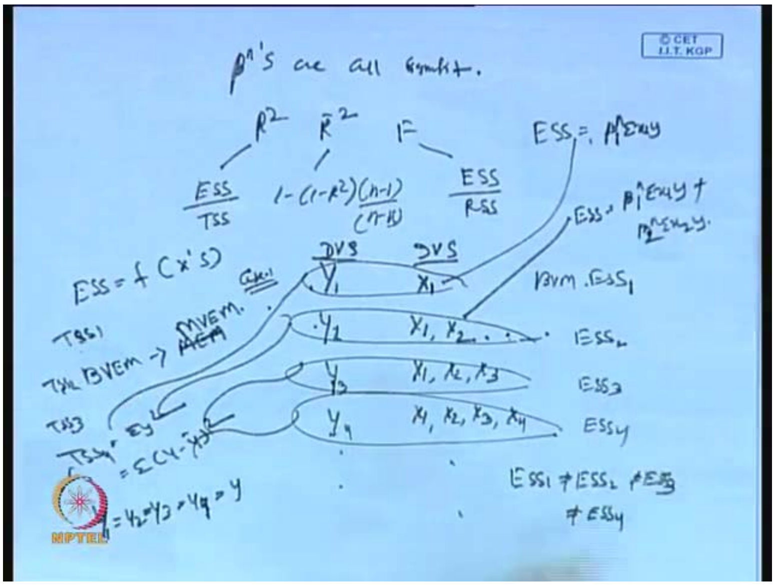
So, if it is coming extreme end then obviously, F will be accordingly very high. Because F is the function of R square so, that means F statistic depend upon the values of R square. If R square is very high then F will be by default very high. If R square is very low then by default F is very low. So, that means R square is considered as a strong component to check this particular issue.

So that means, if by any chance if few beta(s) are not significant and in the same time R square is very high and F is a very high means at R square is highly significant then, there is a serious problem. So in that, in that case model cannot be considered as the best this is one absence or one such problems. And second such problem is let us assume that all beta(s) are statistical significant, but on the other side R square is very low and followed by F is also very low and also not statistically significant or statistically significant at very low rate.

So, that means in that case, the problem cannot be considered as the best. So, what is the final conclusion? So final conclusion is this estimated model has to be restimated again to get the model again best fitted. So this is how the process is all about. But now, for this particular set up. So far as a first part is concerned, it is very frequent in the econometric models. Means there is a possibility that, few beta(s) are significant and few beta(s) are not significant.

So, that means if say out of 5, if 3 beta(s) are significant, 2 beta(s) are not significant or vice versa and in the same times R square will be relatively very high and followed by F is relatively very high. But in the second case if all beta(s) are significant then R square is very low and followed by F statistic is very low. Then in that case, this means it is very rare situation; it is very occasional and very exceptional say totally very associational that means, why it is so exceptional? You see here.

(Refer Slide Time: 24:24)


$\hat{\beta}$'s are all significant.
R^2
$\bar{R}^2$
F
$\frac{ESS}{TSS}$
$1-(1-R^2)\frac{(n-1)}{(n-k)}$
$\frac{ESS}{RSS}$
$ESS_1 = \hat{\beta}\Sigma xy$
$ESS_2 = \hat{\beta}_1\Sigma x_1y + \hat{\beta}_2\Sigma x_2y$.
$ESS = f(x's)$
DVS
IVS
Y_1
X_1
BVM. ESS_1
TSS_1
TSS_2 BVEM → MVEM.
Y_2
X_1, X_2
ESS_2
Y_3
X_1, X_2, X_3
ESS_3
TSS_3
$TSS_4 = \Sigma y^2 = \Sigma(Y-\bar{Y})^2$
Y_4
X_1, X_2, X_3, X_4
ESS_4
$Y_1 = Y_2 = Y_3 = Y_4 = Y$
$ESS_1 \neq ESS_2 \neq ESS_3 \neq ESS_4$
© CET I.I.T. KGP
NPTEL

When we will go for R square statistics because here it why exceptional? So in the first case beta hats are all significant are all significant. This is first instant, then in the second case in the second case R square and F or adjusted R square adjusted R square F. So R square is calculated by the ratio between ESS by T S S. So adjusted R square is 1 minus 1 minus R square into n minus 1 by n minus k. Similarly, F is calculated by the ratio between ESS by R S S. But you remember ESS is basically the function of X(s) only so, that means the value of ESS depends upon the depends upon the structure of X only means the status of X only. What is E S S? The ESS means explain sum squares.

So now, explain sum square and total sum squares, so when we will calculate R square then obviously, we have the ratio ESS by T S S. But by default T S S is always one every time, because we have already checked it here. Let us take a case here, so we start with Y and X this is dependent variable structure, this is independent variable structures. Now, my intension is how quickly we can extend the status of the model that is from bivariate to multivariate.

So, the beginning is Y and X that is the initial set up of this econometric model. If it is only one variable, then econometric model does not work at all. So, that means the starting point is obviously two variables and that too dependent classification, independent classification. So, here we are assuming that Y are the series of dependent classification and X are the series of independent classification.

Now, in the meantime in particularly today(s) discussion or last couple of lectures discussion, every time we are assuming that one Y with series or you can say one order series of independent variables. So, if we will start with one dependent with one independent, then one dependent with many independent. Then, how is the set up? And how quickly you can extend? And what are the problems? and what are these serious issues we have to detect there? So, that means if the if the structure is Y with X then you will call it bivariate modelling.

So in fact, instead of X, I will put it here X 1 because all together we are discussing here multivariate problem and this is one of the partition or this is one upon the deductive part of this particular issue. Now, Y equal to X 1 this is this starting process let us say this is case 1 this is case 1 or step 1 applied. Then, next step so Y equal to X 1 and X 2, so that means we try to increase one after another variables in the system.

So then, in the third case it is X 1, X 2 and X 3; in the forth case Y is X 1, X 2, X 3, then X 4 so it will be continue like this. So, this is bivariate system, then this is trivariate system then accordingly it will increase and we will call it as a multivariate system.

So now, if this this can be one model, this can be another model this can be another model this can be another model. Now, our approach is bivariate to multivariate that is how the entire discussion is the movement from bivariate econometric modelling to multivariate econometric modelling. So, actually it is better to write multivariate econometric modelling multivariate econometric modelling.

So, ultimately what is the problem here? So once you move here Y X 1 to Y X 1 and X 2 then Y X 1, X 2, X 3 then obviously this system is very accurate and very perfect, how quickly you have to generate this solution and various problems? Now our issue is here ESS and T S S; TSS is simply summation Y square. So, that means it is nothing but, summation Y minus Y bar whole square. So, this is this is applicable for this model, this is applicable for this model, this is a applicable for this model, this is applicable for this model.

Now whatever, your problem size all together whether it is bivariate or trivariate or multivariate in every case, in this similar problems we are just adding one after another variables. For instance, I am just going to highlight one problem here, stock price with its determinants means this particular problem we will discuss in the multicollinearity issue. Now when when we will start with stock price then, you have to find out which one is the most important determinant and obviously, we have to consider first variables say X 1.

So, later on we will check it with respect to their significance level of significance. So, the status of a particular variable can be examined or vectors of particular variable can be examined by the estimated coefficients and its significance levels only. So, by the way we can judge the relative weightage of this particular factor to the dependent variable.

Now, in this particular set up we like to have a problem here stock price, then one of the most important determinant is called as index of industrial production I I P. Then, there are another variable say money supply; then there are another variable say called as a rate of inflation, then exchange rate. These are the factors, which can affect the stock price; this is how we have borrowed from the lots of theoretical knowledge.

So now, if I will plot all these variables in this particular sequence this, if it is a stock price then obviously, this should be you can say I I P index of industrial production, then the same stock price. Then, we are integrating this I I P with another variable money supply. Then similarly, in Y case it is stock price already designated so obviously, X 1 is already designation I I P, X 2 is already designated money supply, then X 3 is variable say inflation.

So, this is how we have to add one after another variable but you remember once you proceeded from first to this forth head then obviously, every time b a c b a c b a c is there, means stock price stock price stock price are there. That means what we will go mathematically every time the left side component is Y Y Y. So, that means whether you are here or whether you here or whether you are here, every times you have the same summation Y square. So, it will not vary.

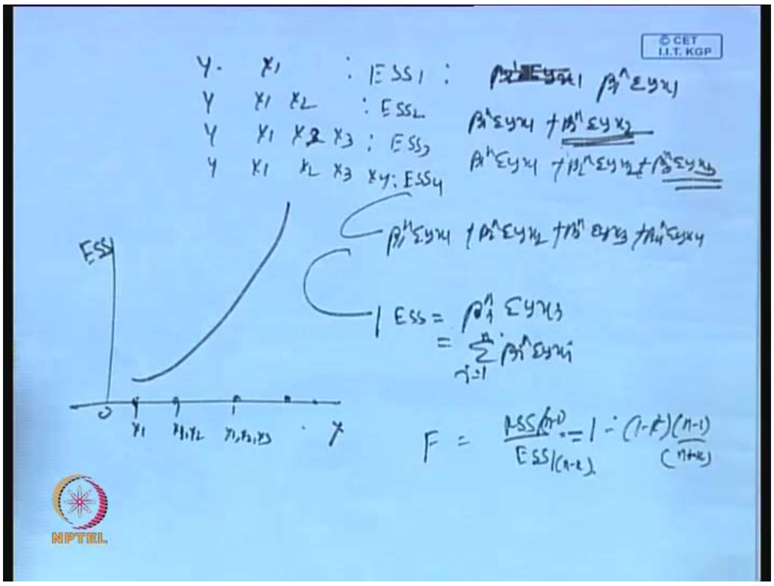
However, in the other side ESS will vary. So, that means for this ESS will be something different for this ESS will be something different. Let it be called as ESS 1, ESS 2, then this is ESS 3, this is called as a ESS 4. Completely different problem and in the same times we will call it here T S S 1, T S S 1, T S S 2, T S S 3, T S S 4. But T S S 1 is this one, T S S 2 this one, that means here every time Y Y Y Y. If you will call it here Y 1 Y 2 Y 3 Y 4 in that sequence then, every time we are assuming that or it is default Y 1 equal to Y 2 equal to Y 3 equal to Y 4 is equal to say Y.

So, this means one variable so, every time summation Y square is more appropriate to use there. But in the other side ESS 1, ESS 2, ESS 3, ESS 4 these are not at all equal. So, that means here ESS 1 not equal to ESS 2 not equal to ESS 3 not equal to ESS 4. So that means, what is the set up of E S S? That is what is explain sum squares? By basic formula explains sum square is beta hat beta hat summation X 1 Y summation X 1 Y for this particular model for this particular models.

So, for this particular model beta beta 1 equal to summation summation X 1 Y or Y X 1. So, this particular setup this particular setup is for bivariate model. When you will go for trivariate model then obviously, you have to add another variable. So, for instance for the second variable this one then ESS equal to beta in fact this will be beta 1 hat beta 1 hat summation X 1 Y. So similarly, in this case, beta 1 hat summation X 1 Y plus beta 2 hat summation X 2 Y.

So, this is how the factor has to be added. So, that means once you move one after another then something will be added into this particular E S S. So, the trend of ESS E S S will start increasing when we will involve one after another independent variables.

(Refer Slide Time: 34:31)



Y. X_1 : ESS_1 : $\hat{\beta}_1 \Sigma y x_1$

Y X_1 X_2 : ESS_2 $\hat{\beta}_1 \Sigma y x_1 + \hat{\beta}_2 \Sigma y x_2$

Y X_1 X_2 X_3 : ESS_3 $\hat{\beta}_1 \Sigma y x_1 + \hat{\beta}_2 \Sigma y x_2 + \hat{\beta}_3 \Sigma y x_3$

Y X_1 X_2 X_3 X_4 : ESS_4

$\hat{\beta}_1 \Sigma y x_1 + \hat{\beta}_2 \Sigma y x_2 + \hat{\beta}_3 \Sigma y x_3 + \hat{\beta}_4 \Sigma y x_4$

$$ESS = \hat{\beta}_1 \Sigma y x_1$$
$$= \sum_{i=1}^{n} \hat{\beta}_i \Sigma y x_i$$

$$F = \frac{ESS/(k-1)}{ESS/(n-k)} = 1 - (1-R^2)\frac{(n-1)}{(n-k)}$$

ESS
0
X_1
X_1, X_2
X_1, X_2, X_3
X
© CET
I.I.T. KGP
NPTEL

Let us see here is once again so, the structure is like this, Y X 1 so Y X 1 and X 2, then Y X 2 X 3 then Y X 1 X 2 X 3 X 4. So, this is I will call it explain sum square ESS 1; this I will call it as ESS 2, then this I will call it as ESS 3, then this we will call ESS 4. So there are two specific objectives for writing all these components. So, first thing is what is the ESS value? How is the trend? So, the trend will give you whether the trend is constant so, that means this value this value are same or it is increasing trade or decreasing trade or you can say there is some up and downs.

So that means, increasing decreasing or decreasing but out of all these possibilities the most feasible is that it will be in increasing trade. Because by default we will see here, it is how that structure for ESS 1 it is beta 1 hat summation Y X 1. So this one will be beta 1 hat summation Y X 1 plus beta 2 hat summation Y X 2 then in that case ESS 3 equal to beta 1 hat summation Y X 1 plus beta 2 hat summation Y X 2 plus beta 3 hat summation Y X 3.

So, this is how you have to continue, so that means you see here every time there is involvement of something extra. Now, in that case there is something extra. So similarly, in the case of ESS 4 it will be coming like beta 1 hat summation Y X 1 plus beta 2 hat summation Y X 2 plus beta 3 hat summation Y X 3 plus beta 4 hat summation Y X 4 so like this.

So now in generalize ESS equal to beta hat i summation Y X i. So, of course, it is better you can write like this way, beta i hat summation Y X i, i equal to 1 to n. So this is the ESS format, so that means if we will plot like this way, this side is X and this size is E S S. Then obviously it will increase like this it will increase like this. So that means, if you plot like this let us say this is a X 1 this is X 1 means number of variables introduced in the system this is X 1, X 2 this is X 1, X 2, X 3 and so on like this way so, that means this is 0.1 0.2 0.3 0.4 0.5 like this 0.1 one variable 0.2 two variables 0.3 three variables.

So that means, when there is one variable what is E S S? When there is two variable what is E S S? When there is a three variable what is the E S S? So, like that if you will plot all these thing, then it will be increasing sequence. So, that means what is our conclusion is that, so when we will add one after another variable ESS cannot be constant it will increase at a increasing rate, in the same time T S S will remain same T S S remain same.

So that means, whether it is Y X 1 or Y X 1 and X 2 or Y X 1, X 2, X 3. In every case your T S S will be same. So, the problem is only for E S S. So, since ESS is changing at a increasing rate and T S S is constant so, by default R square cannot be low and followed by F cannot be insignificant or F value cannot be low. The moment you will add one after another variable then obviously, by default or by natural process means by mathematical process it will increase the value of R square coefficient determinants or adjusted R square.

Similarly, f can also increase but when you will add one after another variable then obviously, you are introducing another coefficient into the system. So for instance, for a bivariate set ups we have only beta 0 hat and beta 1 hat, then when we will go for trivariate then beta 0 hat, beta 1 hat, beta 2 hat. So when we will add another variable then beta 0 hat, beta 1 hat, beta 2 hat, beta 3 hat, beta 4 hat, beta 5 hat like this.

So one one after another variable you are getting one after another beta coefficient. So that means, the requirement is that so every time whatever coefficients are involved at what point of time, all these coefficients should be highly statistically significant. If not, then there is serious problem, serious problem one of the one of such problem may be because of multicollinearity issue.

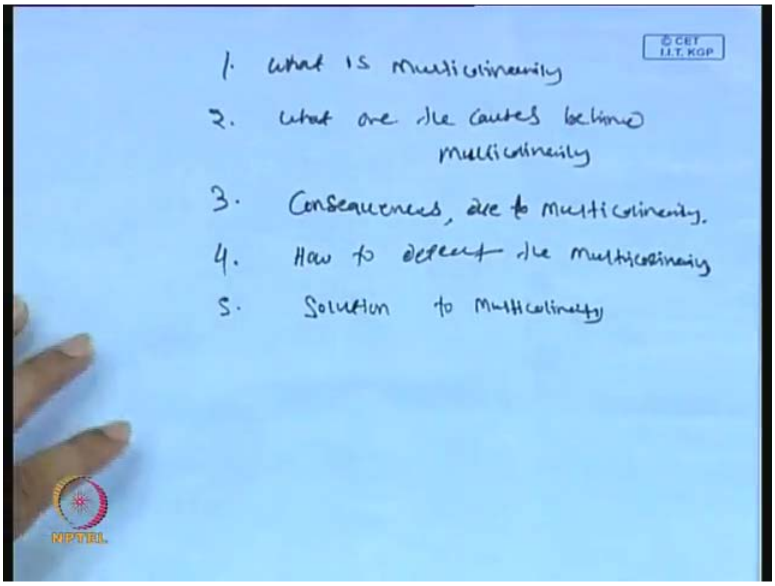
So now, the thing is that so when we will add one after another, then obviously T S S will remain constant only thing is ESS changing at the changing at the increasing rate. So obviously, there is the second possibility may not be occurred it is very very very very exceptional case, where beta coefficients are highly significant or all R significant where R is a very very low and F is insignificant. So this second possibility is very exceptional this is exclusively exceptional. But first cases it is very rare.

So, that means most of the multicollinearity problems will be because of this first case, that means it is the case where the few parameters may not be significant and on the other side R square will be very high followed by F statistic. Because, when R square is high then obviously, F will be high which we have highlighted already because you see here is F is a function of F is a function of R S S by R S S by E S S. So, if you will simplify then this is nothing but 1 by 1 minus R square divide by n minus 1 by n minus k of course, of course, it has a degrees of a freedom here, this is n minus 1 and this is n minus k.

So, this is how the structure all together all right. So now, within that particular frame work, so we like to highlight what is all about this multicollinearity problems. So let us see what is this structure of multicollinearity? So, with this basic background we have to intrude the issue of multicollinearity. It is a very very strong component and very very serious issue as far as multivariate model is concerned.

Until and unless you go to go to that particular problem, then the problem cannot be considered as a best fitted model. It is one of the it is one of the serious problem in the case of a multivariate model. So when when we will handle the multivariate problem then obviously one of such problem it is mandatory that you have to check the multicollinearity issue; otherwise this model cannot be considered as the best fitted model. So now, so what is this exact problem of multicollinearity so that means, we like to now discuss various structure and issues of multicollinearity.

(Refer Slide Time: 42:47)



1. what is multicolinearity
2. what are the causes behind multicolinearity
3. Consequences, due to multicolinearity.
4. How to detect the multicolinearity
5. Solution to Multicolinearity

So what is all about multicollinearity? Second is, what are the causes behind multicollinearity? Then then what are the consequences due to multicollinearity issue? due to multicolinearity Fourth is how to detect? how to detect So far as detection of multicollinearity problem is concerned there are various standard rules, some of the rules you know by inspection you can observe, some of the points you can highlight that there are multicollinearity problems. That means this is by inspection and everybody cannot judge or cannot put remarks or can identify until and unless he is a great statistician or econometrician.

So once a you have a sufficient knowledge about this statistics and econometric, then by look you can say that this problem and this is the models which has severe multicollinearity problem or minor multicollinearity problem and the fourth issue is how to detect. So, some of the detection is very easy, some of the detections are very complex. For that to you have to go for typical procedure procedure measures how to check the multicollinearity.

So then five is a solution solution to multicollinearity, so that means here it means how to detect the multicollinearity? So far as a solution is concerned, once you detect the multicollinearity then obvious obvious obviously there are two questions in your mind. First question is is it a natural problem or is it technical difficult or artificially you are creating that problem. So this answer is very complex. Some of the problems by natural process there will be multicollinearity problem and some of the artificial it can be drawn either by knowingly or unknowingly. I will highlight what are the cases you can knowingly bring the multicollinearity problems? And sometimes there may be unknowingly the problem of multicollinearity is coming.

So means both are artificial means you are the culprit to generate the multicollinearity. You remember one thing why I am using the term culprit, because multicollinearity problem part of this econometric modelling, so it is just like a virus so that virus has to be either clean totally; if not then you have to clean at the certainly levels so that the model can be use for best fitted all rights.

So, let us assume that the multicollinearity is a serious problem. Then obviously by default you need to find out some solutions. So that means just like a virus, it will totally make your system inconsistent or stagnant you cannot run or you cannot work anything. So that means you need to clear it so that means you need to find out software(s) case that the virus can be totally removed.

So that is one of the way how you have to go for the solution. Another trick is some of the virus may be in a one particular way. So still the model at the broad level, it can need, it can work. So that means sometimes multicollinearity may be there but you have to go ahead with the forecasting or policy use.

So now, to make a judgment whether this is ok or that is ok? So it is very serious issues and for that you need to know what your objective is? What is your problem? and whether you are creating artificial or you can say that to knowingly or unknowingly. These are the factors should be in front of you, and then you can make a judgment. Whether you have to go ahead with a multicollinearity problem or you have to find out its complete solution. For instance, see here the standard definition of multicollinearity is that the linear relationship among independent variable that is to with respect to regressors.

We will discuss details what are the various reasons. Since we are discuss discussing one of the component about solution one of the standard region is that the involvement of various independent variable in the systems for a particular dependent variable. It is very very tuff for a researcher, he may not be statistician, may not be econometrician. To find a variable which exactly influence the dependent variable or to find out few variables which exactly determine by dependent variables.

If that is the case then obviously econometric has a zero rule, that means the way we are expecting that there should be error component and there should be explain components, then there will be total component. If error will be 0, then obviously the problem complexity will over there that means the game will over there. The game will be very interesting when there is some type of error component. So your objective every time is how to minimize the errors? So of course, if you will get error free component then your journey is end there. But you will get different type of satisfaction. If you have problem by the way you have to get it solution.

So, that is one way very interesting but it is very frustrating also. That is one type of very enjoyment and that is what is econometric modelling all about. So econometric modelling means the moment you will be entering, so you are expecting that everywhere there is a problem. So every times you have to find out solutions no where you will get the complete solutions. So if you will finds solution for something, then there will be additional problem person. So at very very last days you can get a solution by compromising so many things in your problems or with respect to various objectives.

So why I have mentioned this issue for instance, for a particular variable because you are not sure which variables are exactly information dependent variables? So you are adding one after another then, you will find the model itself. Estimated model is giving very nonsense results for instance, a very few significant R square, very few significant parameters so and very higher square. Then obviously this model is totally unfit and cannot be considered as a policy use or for forecasting etcetera.

So that means you are in shear that because of some variables this system is totally in the ruff side. So, you have to check or you have to detect, which particular variable (( )) to damage this particular environment? So you have to remove that one, that is how the objective all about for multicollinearity. This is what is artificial and that too sometimes unknowingly issue; unknowingly issue means because no idea previously, then by the process of statistical investigation you come to a conclusion.

Sometimes knowingly you can create such problem. Take a case of time series modelling. We have so various components under time series modelling, so when we will go for time series modelling, we start with one variable then it create series of various variables. For instance, let us say variable Y t. Y t is considered as a variable for time series modelling then you have to create various independent variables like Y t minus 1, Y t minus 2 like Y t minus k.

So now, there when will fit Y t as a function of Y t minus 1 Y t minus 2 up to Y t minus k, then obviously there may be possibility that Y t minus 1 as a function Y t minus 2, Y t minus 3 Y t minus k again Y t minus 3 may be also function of Y t format or vice versa. Y t minus 3 may be function of Y t minus 2 or t minus 1 or Y t. So there are various problems are there. We need to have such type of time series frame work or times series modelling is all about that structure. So that time you artificially create variable and you have to find a (( )) solution, because the game and problem is completely different angles and different setup.

So that is you know that there will be obviously multicolinearity problems, still you are going with that particular. So that time your objective is something different not for multicolinearity issue. But by default you will be getting the multicollinearity, but particularly in the cross sectional modeling set up, so here you cannot you cannot artificially create such multicolinearity.

Because there is no such way to create artificially, but most of the cases unknowingly you generate some multicollinearity problem. But over the over the time frame you have to find out solution so far as the details, structure of multicollinearity and solution are concerned we will discuss in the next class. With this, we have to conclude this session. Thank you very much; have a nice day.