Engineering Econometrics

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

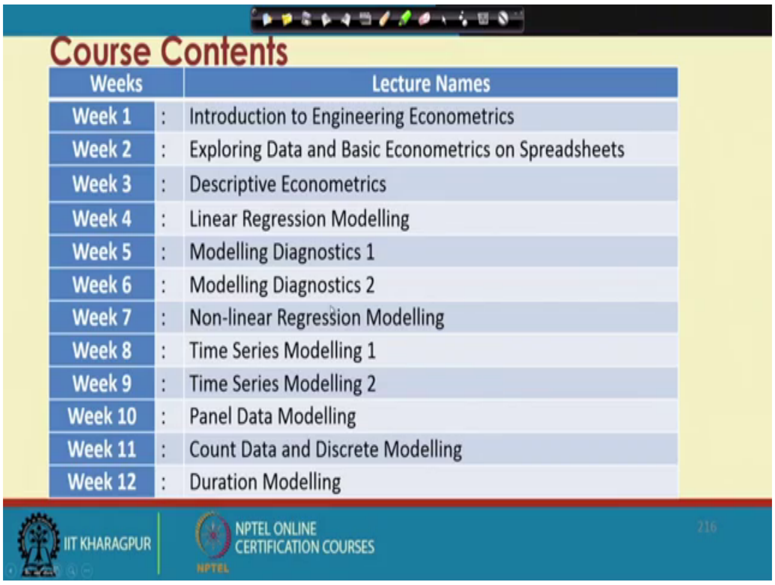
Lecture – 31
Model Specification- Choosing the Independent Variables

Hello everybody. This is Rudra Pradhan here. Welcome to Engineering Econometrics. Today we will discuss Model Diagnostics and that too specific highlights on model specification and that too how to choose the set of independent variables. We have gone through so many lectures and the base, the idea behind this particular modeling is to have input choice and then, with the help of inputs choice, we will build a model and have the estimated outcomes and with the help of estimated outcomes, we will go for management kind of you know a requirement that to how to go for you know good management decision.

We have gone through various process starting with the kind of you know input requirement, the kind of you know processing, then and the kind of you know delivery that is the output and before you use the kind of you know a estimated output for the decision making process. So, we would like to check the diagnostics and we have gone through various diagnostics in the last couple of lectures starting with the specification test, goodness fit test, then the issue of multicollinearity, issue of autocorrelation, the issue of heteroskedasticity.

Likewise we have gone through various you know which or various components through which we can we can declare that the particular outcome or particular you know model is good for the decision making process, but still there are couple of things we are supposed to highlight again or we like to address again before you declared that this model is a good for the decision making process. So, that is we are here for you know lectures.

(Refer Slide Time: 02:24)

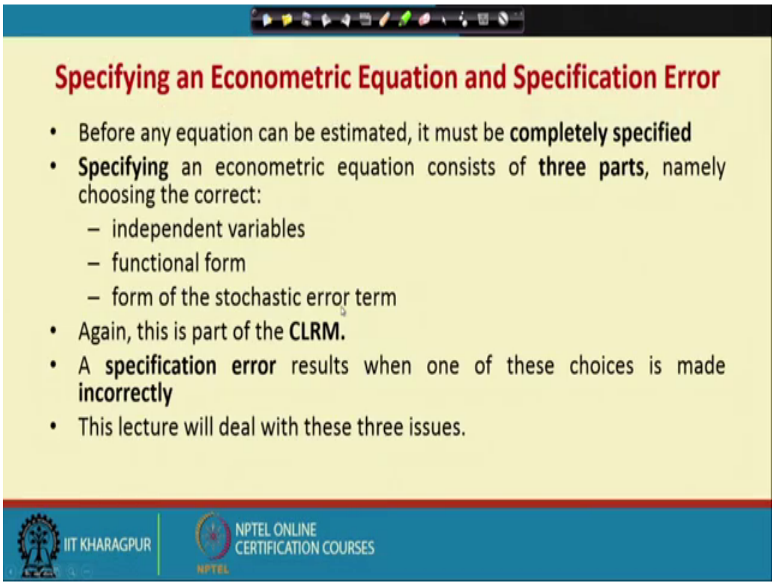


Course Contents

Weeks		Lecture Names
Week 1	:	Introduction to Engineering Econometrics
Week 2	:	Exploring Data and Basic Econometrics on Spreadsheets
Week 3	:	Descriptive Econometrics
Week 4	:	Linear Regression Modelling
Week 5	:	Modelling Diagnostics 1
Week 6	:	Modelling Diagnostics 2
Week 7	:	Non-linear Regression Modelling
Week 8	:	Time Series Modelling 1
Week 9	:	Time Series Modelling 2
Week 10	:	Panel Data Modelling
Week 11	:	Count Data and Discrete Modelling
Week 12	:	Duration Modelling

We discussed model specification specifically.

(Refer Slide Time: 02:27)



Specifying an Econometric Equation and Specification Error

- Before any equation can be estimated, it must be **completely specified**
- **Specifying** an econometric equation consists of **three parts**, namely choosing the correct:
  - independent variables
  - functional form
  - form of the stochastic error term
- Again, this is part of the **CLRM.**
- A **specification error** results when one of these choices is made **incorrectly**
- This lecture will deal with these three issues.

What about model specification altogether? If it is model specification means the literary meaning is model should be correctly specified. If not, then there is a kind of you know misspecifications. So, now the question is what are the ways we can get this kind of you know misspecification?

If the model is misspecified, then that cannot be used for the kind of you know model building or the kind of you know decision making even it passes through all these diagnostics, till we cannot you know use this model for the decision making process by the way if model is correct model is not correctly specified. That means, it is misspecified, then the model cannot pass through all these you know checks. It may have some kind of you know problem in betweens because the misspecified model has lots of you know bias. As a result there is a high chance the standard error of the estimates will be very high and in the first instance the parameters will not be statistically significant. So, it will it will affect you know many ways before use this model for the prediction and the kind of you know forecasting. So, what we like to you know discuss here that you know we like to find out what are the ways we can declare that this model is misspecified. Then, again we like to check what are the ways we can actually correct this misspecification and then finally, declare that this model is good for the you know decision making process.

So, the first instance you know on the kind of you know number of independent variables since we are dealing with you know multivariate structures, every times you know means till date we are discussing you know one dependent variable with the series of independent variables.

So, the series of when there is a question of series of independent variables starting with you know 1 independent variable, 2 independent variables and it may be 20 independent variables, it may be 30 independent variables. So, one of the particular requirement in this extension that your sample size should be substantially very high. So, if it is actually 1 dependent with 1 independent and the sample size may be small, then when a particular model you know involves 1 dependent with 20 independent variable, then in the similar way the sample size should increase you know as per the increase of you know in number of independent variables that is the first under requirement.

So, we can actually you know number of ways you can increase the number of independent variables and then, we can you know create a multivariate system because having more number of independent variables in the system, the model building or the kind of you know model forecasting is very mean. In many instances it will be very accurate, but in the same times the kind of you know, misspecification should not be there.

So, in the first and misspecification means either the model is actually under identified or the model will be over identified. If it is not misspecified, then this is called as you know correct specifications. Then, that in that case it is called as you know the model is exactly identified, otherwise there is a question of you know, under identifications or you know kind of not our identification.

So, in both the cases there will be an issue of bias and as a result model cannot pass through you know diagnostic check and as a result this model cannot be used for decision making process whatever may be the engineering problem, but if the estimated output will be having or the estimated model is having actually a misspecification, then we should not use this model for any kind of you know decision making process.

So, the first issue is actually the number of independent variables that should be very optimum and the way we will be fixed that should be correctly specified and that is how it is called as you know exactly identified, but it is very difficult sometimes to get the exact number of variables which can be used for the decision making process, but we will try our level best to find out the optimum numbers that is how the econometric modeling is all about.

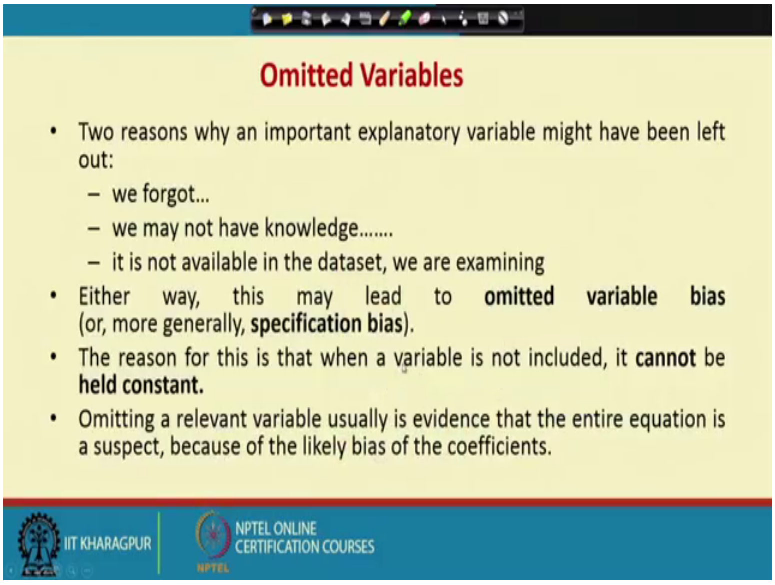
So, you we may have a different problems or different kind of natures. So, we like to learn how this particular issue can be addressed nicely, so that you know we can use this model for the decision making process and the second issue is the functional form. Sometimes whatever you know issue of you know independent variables we may sort out, but if you know the set of variables and the kind of you know functional form is not correct, then again there is a misspecification. That means, the requirement is the linear model you know if there is use of you know non-linear models or vice versa can be also true.

So, it is the requirement of you know a non-linear models and we are using linear format, then there is a kind of you know misspecification again in the if the model is not actually linear. So, it may be exponential form, it may be logarithmic form, it may be a quadratic form. So, there are many different ways. If the model requirement is accessed you know let u say exponential and we are using quadratic, then again there is a chance of you know a specification. That means, there are many different ways you know we can again declare misspecification that too with respect to functional form.

Then, the third is the form of the stochastic error term. So, the error term declaration should be you know as per the best you know kind of you know requirement and the identification. It should not be used for some kind of you know you know uneven kind of in a situation where the variables may be identified, but you know yeah we are you know unnecessarily we are saying that you know this can be taken care of by error term.

So, like that you know there are many different ways you can actually check the specification, that is with respect to you know independent variable, then the functional form, then that with respect to the error term. So, these are all you know, part of CRLM that is the Classical Linear Regression Modeling and the specification error results when one of these choices is made, you know incorrect in the simple that in the simple way of you know understanding.

(Refer Slide Time: 09:12)



Omitted Variables

- Two reasons why an important explanatory variable might have been left out:
  - we forgot...
  - we may not have knowledge.......
  - it is not available in the dataset, we are examining
- Either way, this may lead to **omitted variable bias** (or, more generally, **specification bias**).
- The reason for this is that when a variable is not included, it **cannot** be **held constant.**
- Omitting a relevant variable usually is evidence that the entire equation is a suspect, because of the likely bias of the coefficients.

So, now how will you deal with all these things? So, there are you know various ways we can declare a first we start with you know omitted variables so; that means, that is actually under identification situation there are lots of variables which are actually very useful, but are you know unfortunately that is not in the system and the reverse is also true the variable is not. So, important, but it is included into the system.

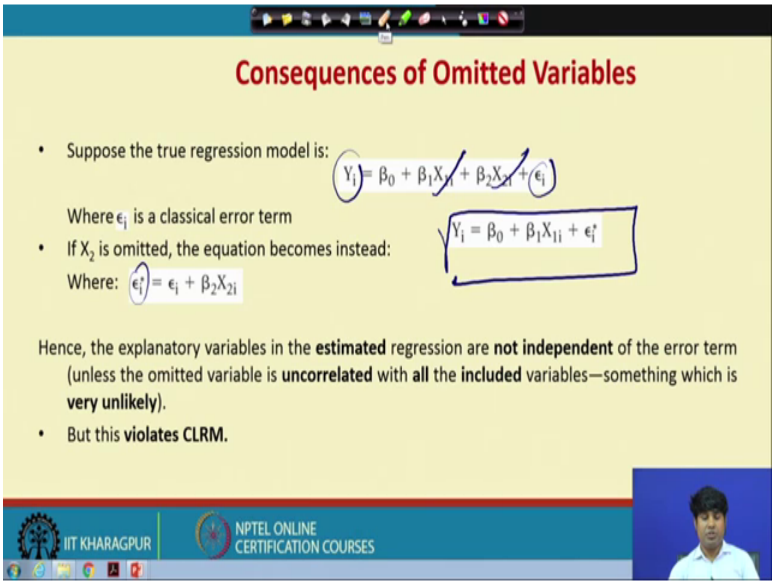
So, important variable which is excluded in the model and you know is not important, but included in the model. So, both the cases the situation is misspecification. So, we have to be very careful how you could declare and how you have to identify; all these things. So, two reasons why an important explanatory variable might have been left out sometimes we will forget, sometimes we may not have in knowledge, sometimes it is not available in as far the actually data set, sometimes we can identify the variables, but data is not available, sometimes data is available, but there is some kind of you know large missing observation or something like that. So, there are many ways we can actually skip a particular variable and in whatever may be the case, but ultimately the end result is the specification. So, what is you know requirement, ultimate requirement? The ultimate requirement is we try we try our level best to have you know less misspecifications or at best we can say that you know should be correctly specified.

Of course, it is very difficult to say that you know these are the only group of variables which can influence the dependent variable or this is the only functional form which can have the best for the decision making. It is very difficult to say, but still we will try our level best by checking all the things. We can simply declare that you know yes this model is correctly specified that too with respect to number of variables involvement and the functional form and the kind of you know on this the involvement of the stochastic term.

So, either way you know whatever you know situation may be and these are all lead to omitted variables bias and in total it is called as specification bias. The reason for this is that you know a variable is not included. It cannot be you know held constant. So, that is how the kind of you know, issue in the system. So, omitting a real you know useful variable usually the evidence that the entire equation is a suspect because of the like device of the coefficients. So, that means if the important variable is not there in the system and that will you know have a multiple doubts whether the functional form is you know wrong or whether the sample size is actually not optimum. So, there are many different ways you can actually you know you can think, but ultimately the solution means everything can be correctly specified if the optimum number of variables in the system. That means, in most of the important variables should not be excluded whatever may be the reasons.

So, as a researcher, a journalist, you must ensure that you know all the important variables to do, to predict the dependent variables should be in the system whatever may be the ways or whatever may be the cost. So, that should be in the system, otherwise in the first instance the model will be you know may be specified and if the model is misspecified, there is no need to go through all the diagnostic altogether because it definitely will affect some error. If the model is in the first in trans correctly specified, then you can go through all the diagnostic. That is what the understanding is your level, but there is you know you just you know go through lots of iterative process and in a continuous process and then, ultimately when you get the good result as for you know a particular requirement and what again whatever may be the engineering problem.

(Refer Slide Time: 13:22)



Consequences of Omitted Variables

- Suppose the true regression model is: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$

  Where ϵ_i is a classical error term
- If X_2 is omitted, the equation becomes instead: $Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i^*$

  Where: $\epsilon_i^* = \epsilon_i + \beta_2 X_{2i}$

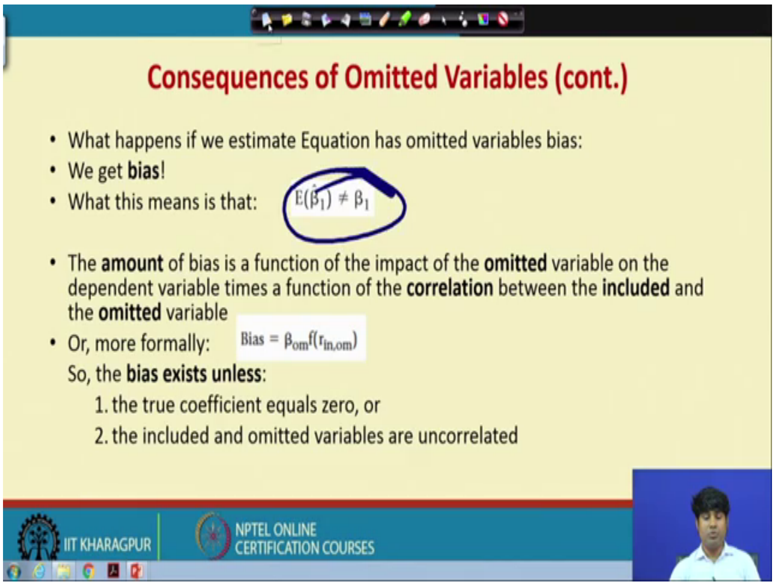
Hence, the explanatory variables in the **estimated** regression are **not independent** of the error term (unless the omitted variable is **uncorrelated** with **all** the **included** variables—something which is **very unlikely**).

- But this **violates CLRM.**

So, I will just let you know what is the consequence of you know omitted variable. Obviously the bias will be very high. So, let us say let us start with you know issue that you know the variable is a kind of you know correctly not there is no correct specification. So, we start with a simple regression. That is what we are actually two independent variable X1 and X2 and the dependent variable Y i and this is what the error term. Let us assume that X2 is omitted, then the equation will be reduced to bivariate only that is Y i equal to beta 0 and beta X1 and obviously, when we are omitting a particular variable, we assume that you know or you know we can say that you know this impart will be taken care by the error term.

So, explanatory variable in the estimated regression are not independent of the error term. So, that is how the kind of you know, issue. So, unless the omitted variable is uncorrelated with all the included variable something which is very unlikely and that is what actually the issue of you know classical linear regression modeling. That means, technically it violates. So, if it violates this kind of you know conditions, so this model cannot be used for the kind of you know decision making process. So, that is a kind of you know requirement. So, ultimately we need to we need to check all these details before you go for this.

(Refer Slide Time: 15:01)



Consequences of Omitted Variables (cont.)

- What happens if we estimate Equation has omitted variables bias:
- We get **bias**!
- What this means is that: $E(\hat{\beta}_1) \neq \beta_1$
- The **amount** of bias is a function of the impact of the **omitted** variable on the dependent variable times a function of the **correlation** between the **included** and the **omitted** variable
- Or, more formally: $\text{Bias} = \beta_{om} f(r_{in,om})$

  So, the **bias exists unless**:

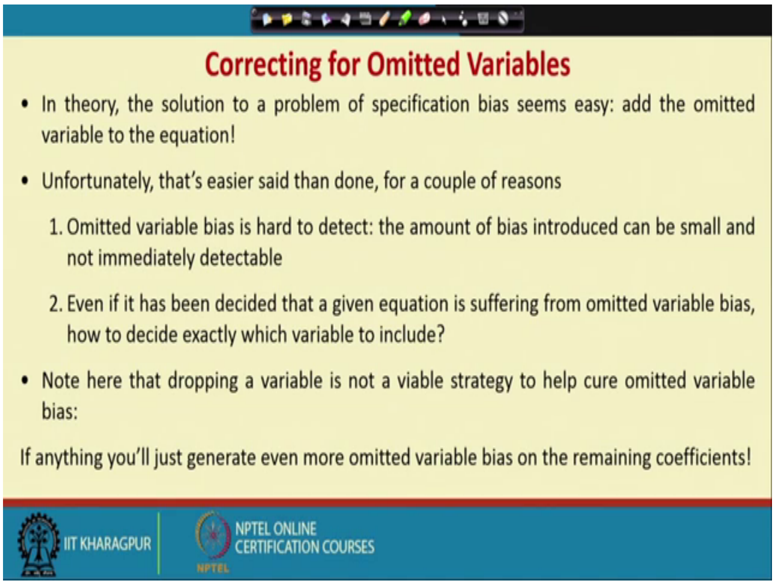
  1. the true coefficient equals zero, or
  2. the included and omitted variables are uncorrelated

You know use any kind of you know, requirement and again we would like to know; what is the ultimate consequence of omitted variables which I have already pointed out. So, when there is an omitted variable, then there is a high chance that you know the particular models means the estimated model will not pass through Guru Theorem. That is what called as you know best linear unbiased estimators.

In the first instance, the parameters will not be completely unbiased, that is what we have. So, if the case is like this, if the expected value of a parameter is not exactly equal to true value of the parameter, then that means technically there is a bias. If that is equal to true value, then there is a non-bias. So, the difference will be exactly equal to 0 and that itself will be created you know lots of you know issue in the model building and the kind of you know decision making process.

So, we have to be very careful how you have to deal with all these situation on and then, coming with a kind of you know right choice which can you know use for the model building.

(Refer Slide Time: 16:10)



Correcting for Omitted Variables

- In theory, the solution to a problem of specification bias seems easy: add the omitted variable to the equation!
- Unfortunately, that's easier said than done, for a couple of reasons
  1. Omitted variable bias is hard to detect: the amount of bias introduced can be small and not immediately detectable
  2. Even if it has been decided that a given equation is suffering from omitted variable bias, how to decide exactly which variable to include?
- Note here that dropping a variable is not a viable strategy to help cure omitted variable bias:

If anything you'll just generate even more omitted variable bias on the remaining coefficients!

So, correcting you know the omitted variable bias, so that means ultimately you know these are the issues always there. So, now you know how you can actually correct this particular error. So, you can go through theory because the theory itself will give you lots of you know hints.

So, what are the variables should be in the system before you use that model for you know decision making process because the problem identification, the variables you know identifications all are actually derived from the theory that is why in a kind of you know research investigations. So, the first requirement of this particular you know in the engineering econometrics that you should go through you know literature review.

So, literature review will give you enough exposures to understand this problem, to identify the variables and the kind of you know expectations, the kind of you know nature that is the impact of a particular variable. So, theory will give you sufficient you know clues you know by what we can say logical kind of you know, interconnections and then, the theoretical kind of you know end. So, that will be actually you know means you can say that is the best through which you can start the process and then, finally you can check the omitted variable issues and then, come into their particular you know set up where correct model is correctly specified.

So, what is happening, the omitted variable bias is hard to detect sometimes because you know it is very difficult to say that you know which variable you know ultimately very important until unless you check in the particular you know process. Ultimately we will you really say that you know this is the variable which is more significant or you know more important.

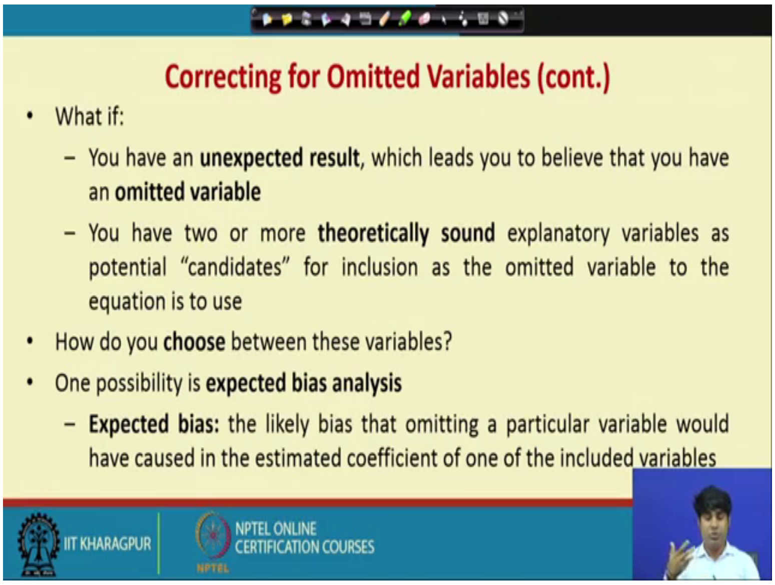
So, theoretically you may get some kind of you know guess, but in the statistical you know kind of you know output should also be very supportive. Sometimes there is a high chance and the theoretical you know expectation is something and let us speak something different. So, it may be due to actually omission of a particular you know important variables or it may be due to you know uneven sample sampling. So, that means the size of sample may not goals or it may be by sampling sometimes.

For instance, you know if it is not you know randomly selected and not actually in a kind of not different atmospheres, then that will give you some kind of you know bias and that really that will also you know question of you know omitted variable bias. So, we must be very careful how you have to deal with all these you know he says even if it has been decided that a given equation is suffering from omitted variable bias and then, question is how to decide exactly which variable to be included. Of course, we have already discussed this issue in the multicolor in the problem. So, when actually means multiple entry problem, usually these issue of you know our identification and sometimes also it may be due to under identification because when some variables are not significant, then there is a high chance that you know there is a chance of you know multicollinearity.

One way of you know solving multicollinearity, you drop the called you know the variable which is actually collinear in nature and sometimes if we you know add a particular new variable which may actually you know effect the particular you know link and then, finally there is a possibility that you know all variables finally can be significant. So, this is actually you know kind of you know again iterative process until unless you check and kind you know the kind of you know estimate, re-estimate widget.

So, you cannot get exactly you know situation and the kind of you know requirement. So, ultimately it is actually a kind of you know a trial mechanism. So, every time you have to check and you know report every time, you have to check and report like this. Then, finally you will come to a point where you know the model is correctly specified and that can be used for you know decision making process.

(Refer Slide Time: 20:23)



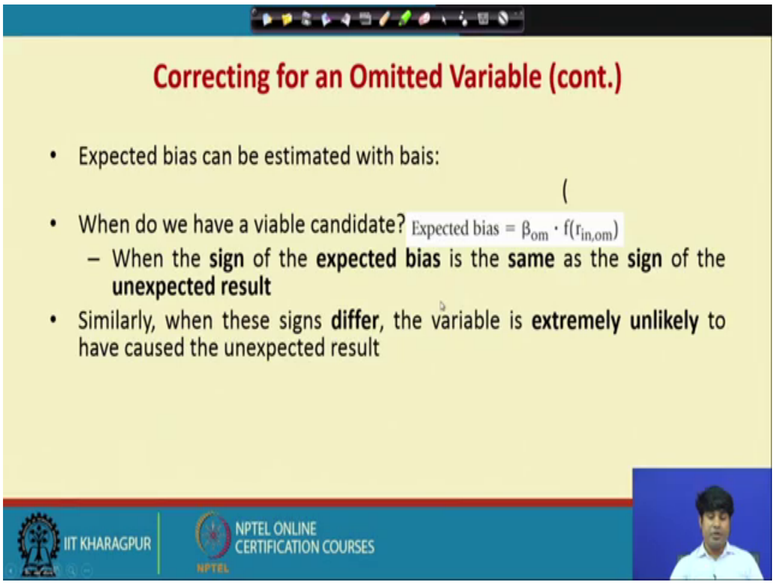
Correcting for Omitted Variables (cont.)

- What if:
  - You have an **unexpected result**, which leads you to believe that you have an **omitted variable**
  - You have two or more **theoretically sound** explanatory variables as potential "candidates" for inclusion as the omitted variable to the equation is to use
- How do you **choose** between these variables?
- One possibility is **expected bias analysis**
  - **Expected bias:** the likely bias that omitting a particular variable would have caused in the estimated coefficient of one of the included variables

So, sometimes you have an unexpected result. That is usually actually happens it may be due to actually very good sampling or you know very huge sampling or it may be due to you know law sample size.

Sometimes the law sample size may give you know better result, but, but the reality will be you know coming into the picture when you will increase you know more number of samples in the kind of you know set up. So, ultimately it is you know your job to check and you know see the kind of you know problem and then, think how we can actually remove these problems and come with the kind of know solution where the model is free from all kind of you know errors and that too specifically the misspecification bias.

(Refer Slide Time: 21:16)

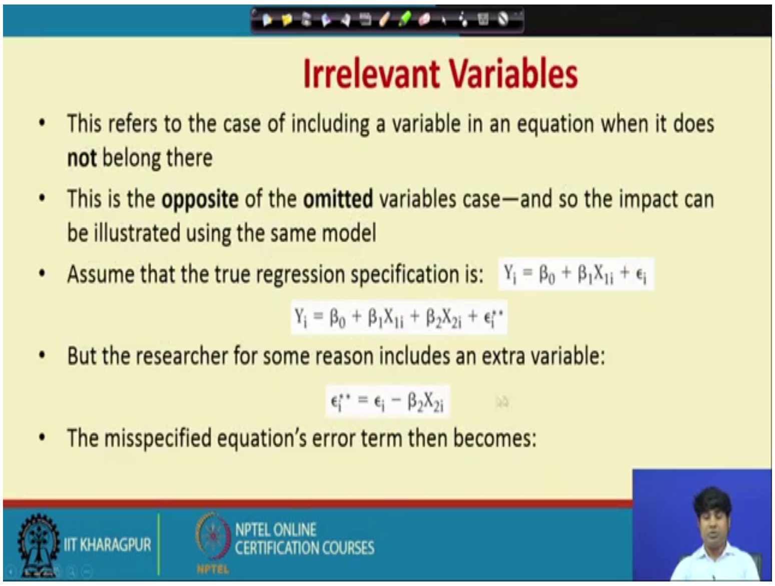


Correcting for an Omitted Variable (cont.)

- Expected bias can be estimated with bais:
- When do we have a viable candidate? Expected bias $= \beta_{om} \cdot f(r_{in,om})$
  - When the **sign** of the **expected bias** is the **same** as the **sign** of the **unexpected result**
- Similarly, when these signs **differ**, the variable is **extremely unlikely** to have caused the unexpected result

So, that is you know very much important. So, sometimes the other problem may be actually because of you know misspecification, the sign of a particular variables. So, that means the nature of a particular variable may be getting affected because once the important variable will be included, then the nature of the variables which can be very supportive as for the theoretical you know in a hint or the kind of you know logical kind of you know flow, otherwise if the model is correctly specified whatever you know theoretically you know hint is coming or logical issues are coming. So, here you know estimated models that mean the output which is derived from the data will be very much you know consistent. If that is not the case, then this may be a problem for again you know the kind of you know reliability and that too simply it affects the diagnostics kind of you know requirement and as a result you cannot use this model for the decision making process.

(Refer Slide Time: 22:20)



Irrelevant Variables

- This refers to the case of including a variable in an equation when it does **not** belong there
- This is the **opposite** of the **omitted** variables case—and so the impact can be illustrated using the same model
- Assume that the true regression specification is: $Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i^{**}$$

- But the researcher for some reason includes an extra variable:

$$\epsilon_i^{**} = \epsilon_i - \beta_2 X_{2i}$$

- The misspecified equation's error term then becomes:

IIT KHARAGPUR

NPTEL ONLINE CERTIFICATION COURSES

So, the other issue is actually, the earlier issue is under identification case. Now, sometimes the useless video I cannot say useless, but irrelevant variables may not be actually using the system. So, the variable which is very essential should be included and the variable which is not at all essential or not at all important should not be included into the system. So, ultimately because you know, yes in the one side we need actually more number of variables to predict the thing dependent variable which is actually the high requirement in the same times.

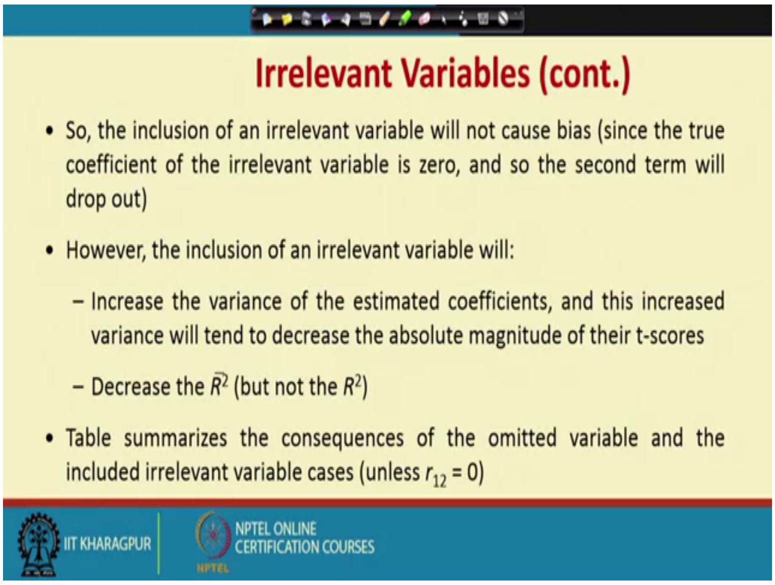
So, unnecessary variables should not be in the system, otherwise it will affect the entire system drastically. So, there is a high chance the particular you know you know unnecessarily variables may be affecting the entire unit system. So, the moment will too drop that particular variable which can affect the system perfectly. For instance, you know if the variable is having actually culinary teeth you know other variables, then simply you can drop that one's. So, that is how multicolor integer all about. So, multicolor will tell you the kind of knee requirement that you know yes this is the upper identify, identification case and this variable is actually unnecessary that may be dropped actually.

If you drop that particular variable, then by default model will be correctly specified and all the diagnostics checks can be passed through. So, if the unnecessary variable in the systems, then this may affect the entire you know model outcomes and as a result it will actually block somewhere else and as a result you cannot use again for the decision making process.

So, what is the question here like this? So, some of here the actual error term is a just you know the kind of you know in the earlier case, it was actually included. In later case, it is actually explorated. For instance, you see here is this error terms. So, we are having it to the plus signs. If you are saying that you know this is actually omitted variables, then we will add this one and that will you know in the error terms, but in this case the kind of you know unnecessary variable case. So, the particular question will be actually subtracted.

So, that means we declared that you know this is actually out of the error terms and then, we will go ahead with the estimation process and I am very sure this is the, this is the you know for the kind of you know model building in the kind of you know decision making process.

(Refer Slide Time: 25:0)



Irrelevant Variables (cont.)

- So, the inclusion of an irrelevant variable will not cause bias (since the true coefficient of the irrelevant variable is zero, and so the second term will drop out)
- However, the inclusion of an irrelevant variable will:
  - Increase the variance of the estimated coefficients, and this increased variance will tend to decrease the absolute magnitude of their t-scores
  - Decrease the $\bar{R}^2$ (but not the R^2)
- Table summarizes the consequences of the omitted variable and the included irrelevant variable cases (unless $r_{12} = 0$)

So, the issues is that you know if unnecessary variable will be included into the system, there is a high chance that the parameters, all the parameters will not be statistically significant. That means, all the variable should not be statistically significant and by the way the adjusted R square which will be getting affected, so the flow of R square is that you know when you add one after another independent variables, it will start increasing even if that variable is unnecessary. Still the R square will only start increasing because it is a mathematical way of you know connection, but the actual fact is that you know when you add one after another variables, then the best model can be choosed on the basis of you know adjusted R square comparisons, not the R square comparison.

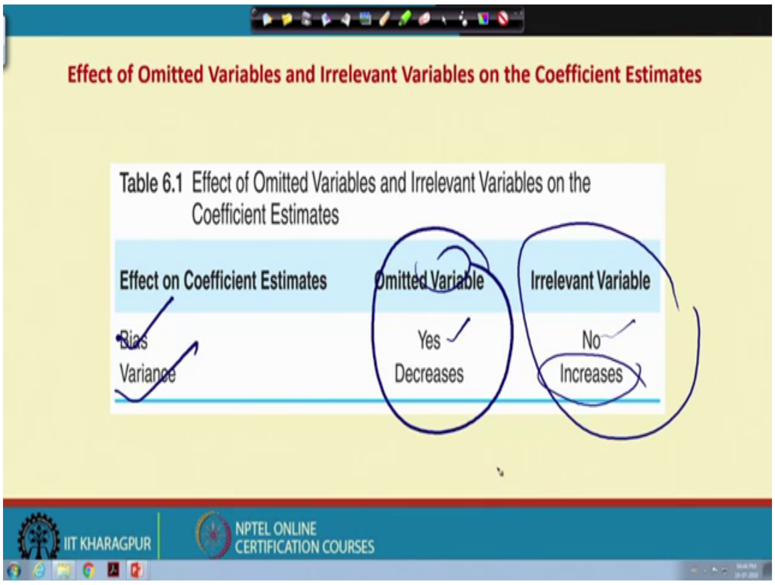
So, adjusted R square will exist the degree of freedom and the variables involvement and the sample size. So, as a result it can give an indication that you know this is the kind of you know omitted variable case or something like that as a result. So, the model is not correctly specified. So, ultimately either the coefficient should not be startled significant or R square really start you know register squarely start decreasing. So, it means deployed like that which we have already discussed in the multicolor indicates.

If you add one after another independent variables, one way we extend your you are boosting the models that too you know predicting the dependent variable that we will get more, more and more pillars to understand, then the kind of you know a flow, but ultimately if the particular you know pillar is not so important, then you know the strength of this particular you know pillar will get affected.

So, as a result you know the model cannot be used for the decision making process and that the simple you know flow to check is that you know the comparison between adjusted R squares. So, let us start like you know why with you know one independent variable, why with you know first independent variable, then you add another independent variable, then you check R square, adjust R square, adjust R square, adjust R squares, then if the variable will be actually very useful, then by default R square will definitely increase, but in the same time adjusted R square will be increasing, but if the addition of new variables, new independent variables is not you know a very relevant or you know useful, then by default you know R square will start increasing, but adjusted R square really started declining you know significantly, but what is actually requirement you how you they say that you know the particular variable is useful and in the same time model is correctly specified.

Then, in that case you know adjusted R square in every stage starting increasing in comparison with R square increase and at the same time there should be actually significance of that variables because if you are saying that you know useful one variable that should you know ultimately significantly that is not the actually ultimately you know or the kind of you know finances, perhaps you know the inclusion of that useful variables may not directly affect, but the presence of the involvement itself will increase the significance level of the other variables. Then, how do you know and that the you know observe this once and therefore, the observation will be simply through adjusted R square and the kind of you know F statistics if the variable is very useful is not directly through indirectly also, then adjusted R square by default only start increasing. That is the simple signal as you can have and then, we can goead the process.

(Refer Slide Time: 29:00)



Effect of Omitted Variables and Irrelevant Variables on the Coefficient Estimates

Table 6.1 Effect of Omitted Variables and Irrelevant Variables on the Coefficient Estimates

Effect on Coefficient Estimates	Omitted Variable	Irrelevant Variable
Bias	Yes	No
Variance	Decreases	Increases

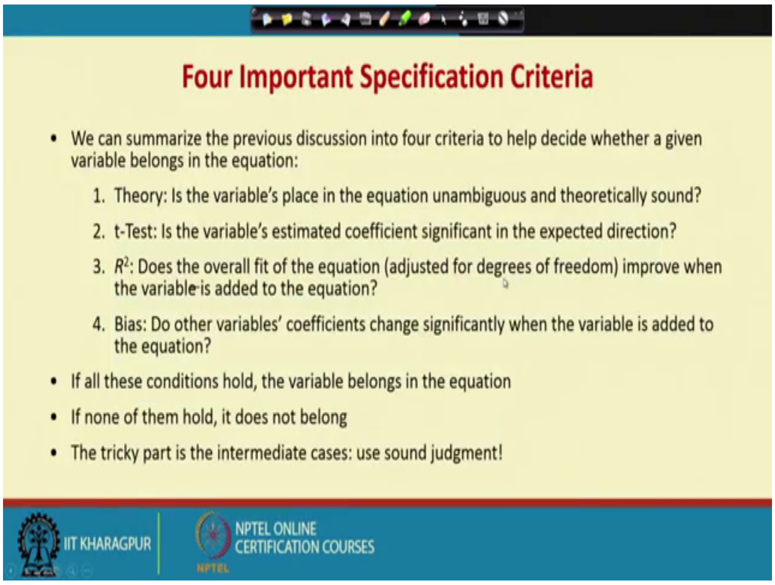
To summarize effect of omitted variables or you know relevant variables on the a coefficient estimates, so there are two issues here. So, first is the bias and the variance. So, when there is omitted variables, then the bias obviously it is you know bias will be obviously yes and there is an omitted variable case. So, that means here in that case most an import one, one or you know few important variables are you know omitted.

So, in that case bias will be definitely in the system and variance you know will decrease. So, ultimately both the, both the kind of you know inference look at affected the model, modeling process and as a result you may not use this model for the decision making process again. So, this, is the case of you know under identification case and this is the case of you know our identification case.

In this case, some unimportant variables in the systems as a result bias may not be there, but you know it will be increased actually a kind of you know variance and variance early in start increasing, then the significance of the parameters definitely get affected. So, that really again you know affect the adjusted R square and the kind of you know F statistic. So, ultimately the requirement is that you know models should not be misspecified. So, there should not be on you know kind of you know under identification case and it should not be our identification case all yes.

Of course, it is very difficult to find out such situation, optimum situation, but we try to find out or we try our level best to that particular you know situation where there is no question of you know omitted variable case and there is no question of you know the involvement of you know irrelevant variable in the system. So, that is how the importance of the economy modeling. So, you cannot remodeling or engineering econometrics which we help you a lot to check all these things and finally, committee kind of you know situation or you know declare a model which is actually free from all errors and then, finally it can be this SI for the decision making process and that to solve some of the engineering problems and as far as per the particularly no organizational requirement or in a corporate requirement is concerned. So, these are all the kind of you know case through which actually we can we can get to know what the omitted variable case is and all this.

(Refer Slide Time: 31:43)



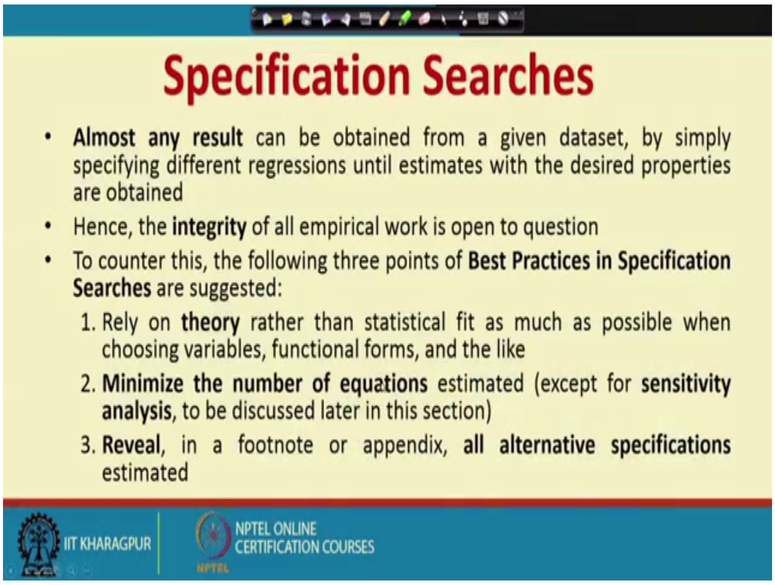
Four Important Specification Criteria

- We can summarize the previous discussion into four criteria to help decide whether a given variable belongs in the equation:
  1. Theory: Is the variable's place in the equation unambiguous and theoretically sound?
  2. t-Test: Is the variable's estimated coefficient significant in the expected direction?
  3. $\bar{R}^2$: Does the overall fit of the equation (adjusted for degrees of freedom) improve when the variable is added to the equation?
  4. Bias: Do other variables' coefficients change significantly when the variable is added to the equation?
- If all these conditions hold, the variable belongs in the equation
- If none of them hold, it does not belong
- The tricky part is the intermediate cases: use sound judgment!

So, 4 important specifications kind of you know criteria, so you know we just summarized the you know whatever we have discussed till now. So, how do you check actually my specification that too with respect to under identification case where the omitted variables you know issue and another case of an identification where you know not you know irrelevant variables will be involved in the process.

So, how in this case you have to go through theory check through T statistic, R squares, adjust R squares bias in the system and then, if all these you know conditions are you know you know having, then the variables belong in these equations or something like that will be getting affected drastically. So, what is the ultimate declaration that you know. So, you have to find out you know optimum scenario, where the particular model will be correctly specified without any issue of you know specification that too omission of the relevant variable and the inclusion of you know relevant variables. That is what the final you know kind of you know requirement which is you need to check and you know ultimately declares.

(Refer Slide Time: 30:01)



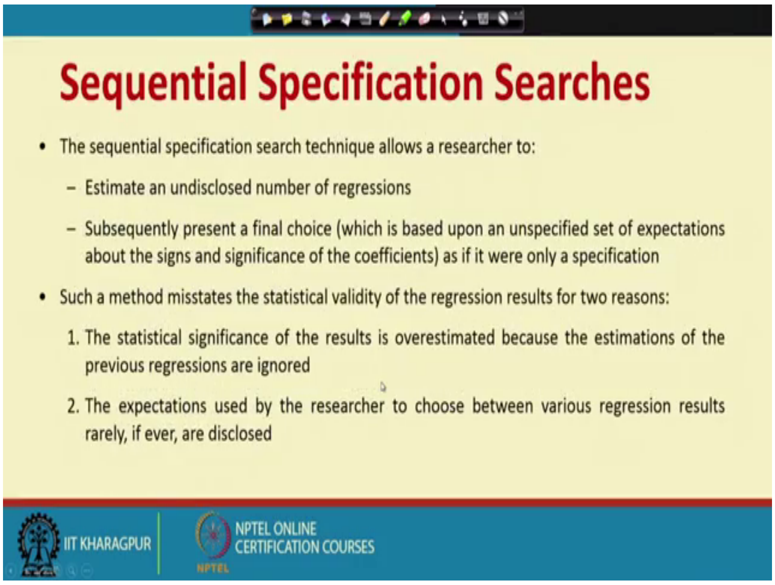
Specification Searches

- **Almost any result** can be obtained from a given dataset, by simply specifying different regressions until estimates with the desired properties are obtained
- Hence, the **integrity** of all empirical work is open to question
- To counter this, the following three points of **Best Practices in Specification Searches** are suggested:
  1. Rely on **theory** rather than statistical fit as much as possible when choosing variables, functional forms, and the like
  2. **Minimize the number of equations** estimated (except for **sensitivity analysis**, to be discussed later in this section)
  3. **Reveal**, in a footnote or appendix, **all alternative specifications** estimated

So, that means a couple of tests are there to check the misspecification. For instance, you know test is there which can declare that you know whether the model is correctly specified. Even it will be passing through all the specification test good, goodness fit test and the diagnostic test like you know multi-collinearity, autocorrelation, heteroskedasticity and something like that.

So, even after doing all these things tests still you know can be applied to declare whether the model is correctly specified as per the particular requirement and that model can be used for the decision making process.

(Refer Slide Time: 33:49)

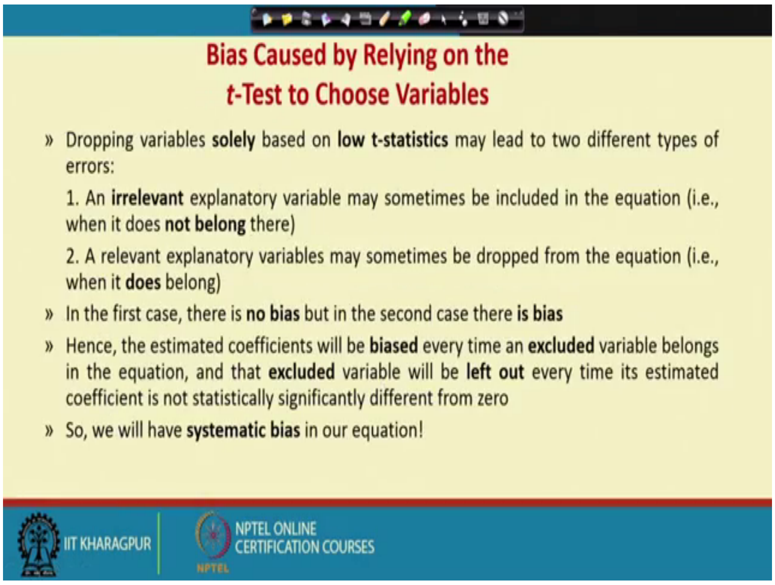


Sequential Specification Searches

- The sequential specification search technique allows a researcher to:
  - Estimate an undisclosed number of regressions
  - Subsequently present a final choice (which is based upon an unspecified set of expectations about the signs and significance of the coefficients) as if it were only a specification
- Such a method misstates the statistical validity of the regression results for two reasons:
  1. The statistical significance of the results is overestimated because the estimations of the previous regressions are ignored
  2. The expectations used by the researcher to choose between various regression results rarely, if ever, are disclosed

So, these are the issues which you can actually frequently use to you know deal with this you know problem and the kind of you know requirement.

(Refer Slide Time: 33:52)



Bias Caused by Relying on the *t*-Test to Choose Variables

» Dropping variables **solely** based on **low t-statistics** may lead to two different types of errors:

1. An **irrelevant** explanatory variable may sometimes be included in the equation (i.e., when it does **not belong** there)
2. A relevant explanatory variables may sometimes be dropped from the equation (i.e., when it **does** belong)

» In the first case, there is **no bias** but in the second case there **is bias**

» Hence, the estimated coefficients will be **biased** every time an **excluded** variable belongs in the equation, and that **excluded** variable will be **left out** every time its estimated coefficient is not statistically significantly different from zero

» So, we will have **systematic bias** in our equation!

So, ultimately there should not be you know any kind of you know systematic bias with respect to the exclusion of the important variable or kind of you know inclusion of the you know unimportant variables.

So, we ultimately actually a check and you know restructure the process, re-estimate the process till you get a kind of you know clean sheet that yes this is the final model and this is the final set of the variables and this is the final kind of you know requirement through which you can actually use this model you know for the kind of you know decision making process ultimately. So, we like to check all these kinds of you know misspecifications before you actually use this model for the management requirement or the kind of you know engineering requirement. That means, technically in the in the diagnostics kind of you know sides whatever you know diagnostic we have checked after doing all checks, the model misspecification need to be actually checked. This is one of the very important component authorized. There is a high chance that you know and the model passes through all these diagnostic by you know some kind of for instance you cannot clean completely multicollinearity, you cannot clean completely autocorrelation, you cannot clean completely clean the heteroskedasticity, but somehow we can compromise and even we will compromise and ultimately that compromise whether it is accepted or not accepted. So, the misspecification test can be you know finally declared. So, even we lift up you know tolerate with you know kind of you know tolerance level, but if the model will pass through specification test, then by default this model can be used for the decision making process, but on the other side if the model actually passes through all these kind of you know diagnostic and is again in the misspecification test, it is misspecified and then, this model should not be used for the decision making process.

So, that means there are various ways actually check, recheck, estimate, re-estimate, then finally come to a kind of you know final choice through which you can, this model can be used for the decision making process. So, with this we will stop here.

Thank you very much.