Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology - Kharagpur

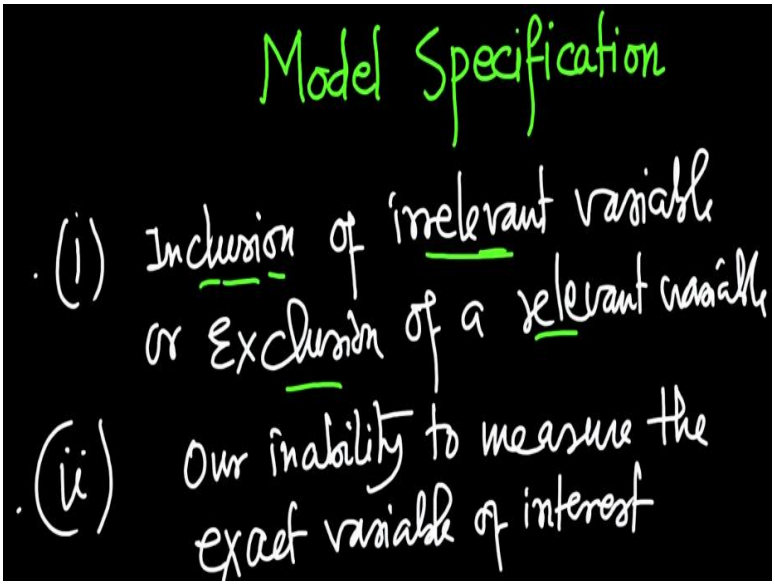
Lecture – 83
Model Specification

Hello and welcome back to the lecture on applied econometrics. So, we have been talking about different tools and techniques to ensure that our econometric model is as correct as possible. We previously have seen the functional specification, we want to specify the exact or the right function for our model. We have seen the problem of multicollinearity. We have seen the problem heteroscedasticity or autocorrelation.

So, all of these different techniques that we have learned is basically to ensure that my econometric model is as correct as possible. Now, one essential problem that we cannot get away with is the problem of model specification. So, what it means is that we need to ensure the variables that we are including in the model are exactly the right variable. So, that means that we do not include something that is irrelevant or we do not exclude something that is relevant, so that is one aspect of the model specification problem.

And the second aspect of model specification problem is that we are able to measure the variable correctly. So, what I mean by that is that the variable let us say if I am talking about a particular variable, let us say the impact of education on wage. So, now in education there are so many other factors perhaps which are needed to be measured along with education, for example your ability or talent, but you cannot have a proper variable to actually identify your say let us creativity or talent.

So, all you can do is you use the variable called education and that you kind of consider as a proxy for the variable talent. So, you cannot measure the variable talent. So, this is another problem where we actually face the problem of model specification. So, let me write it down.
(Refer Slide Time: 02:05)

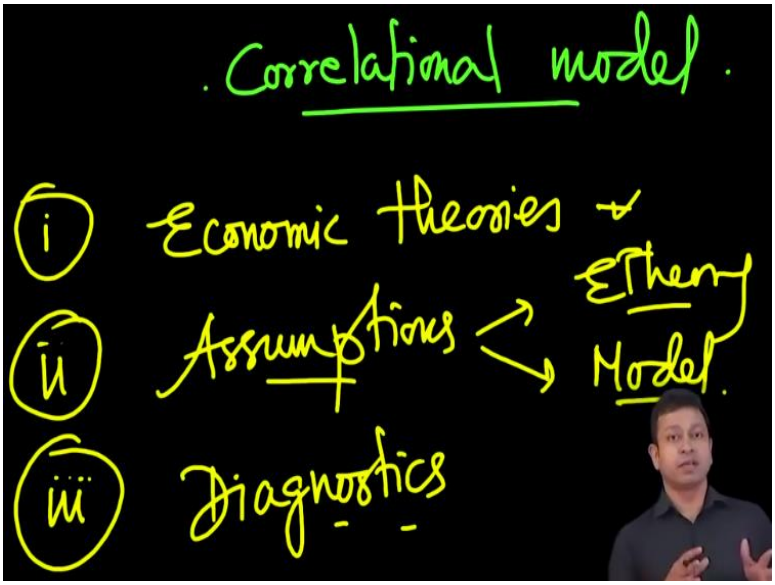


So, the first problem is inclusion of irrelevant variable or exclusion of a relevant variable. Whereas second problem lies in our ability, our inability I would rather say to measure; it is a measurement problem, inability to measure the exact variable of interest. So, if we cannot measure let us say talent, so then my model, the impact of talent would be captured somewhere else and then my model will not be correctly specified.

So, these are the two problems that we are going to talk about when we are going to discuss model specification. So, essentially, what we do is we can actually use all these different tools and techniques like your functional specification, we can check for multicollinearity, we can check for problem heteroscedasticity, we can check for autocorrelation but then that does not ensure my model is correctly specified. So, we have to take care of this two things.

Now, why exactly this model specification is a problem or why exactly inclusion of an irrelevant variable or exclusion of relevant variable is a problem? Now, the problem is that why cannot we identify that so that is even like we should ask this question, why cannot we identify or why the identification of irrelevant variable in the model or exclusion of a relevant variable in the model is a difficulty?

(Refer Slide Time: 03:59)



The problem is this, their correlational model. So, this is something you have to understand that correlational model. So, the relation between y and x we look at the correlation. Now, one let us say a variable which is let us say irrelevant, but you might actually see some correlation of the variable which is not relevant with the y variable. So, you might feel that well if I see some correlations, perhaps there is some the x variable is actually influencing my y variable, but actually in reality it is not.

Because it is a correlational model, so there might be some aspects which are picked up by the the specific variable which I am calling an irrelevant variable. Now, so this is the reason, we can never be sure of if my variable of interest is included or if my irrelevant variables are excluded. And this is specifically the reason why I can never be sure if my model is correctly specified.

In econometrics, so that is something we have to remember, does not matter how much we try we can never be sure that exactly those variables which are actually influencing a y variable are included in a model or if something irrelevant by mistake we are actually excluded in our model. So, in this lecture we are actually going to see the ways to identify if an irrelevant variable is included or relevant variable is excluded and what is the impact of such exclusion or inclusion.

So, how do we actually know the fact there is a correlation model? And the fact that it is a problem, how do I really identify if my model is correct? So, now, I said that I can never be sure if my model is correct or not. Now, given that as an econometrician, as an economist I

would want my model to be correctly specified, as correctly specified as possible. Now, I would say there are 3 different ways to actually identify your correct model or there are 3 different angles you can take to address it.

One is the economic theories. So, this is the first thing, so you have your economic theory. So, your theory actually talks about certain facts. So, your theory is at least partly talking about the truth if not though full truth. So, your economic theory always helps you as a as a guiding principle to approach an economic problem and then you also know from the theory what would be the variables of interest.

So, that is first thing. The second is the assumptions that you make. So the assumptions you make, both in economic theory as well as your models. So what do I mean by that? So, let us say in economic theory you are talking about the impact of let us say education on wage, so you keep certain variables constant, you say given these variables are constant. So you are making an assumption that these variables are constant and then you are trying to see the impact of one variable on the another.

So that is where you are making an assumption similarly in your model when you are talking about, when you are building your model and you are basically doing different tests, so you are assuming normality of the variables or the distribution or like the properties of standard error and so forth. So, all these different assumptions you are making when you are building the model, so that actually creates a boundary as in how within which ambit your model is correct.

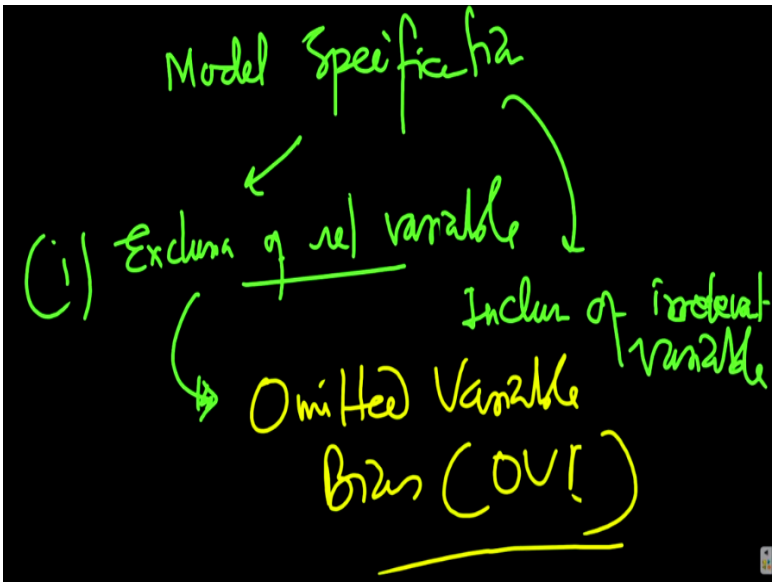
So, that is another way we say that we look at the econometric model that we try to ensure economic model is correct by making assumptions. The third part is the diagnostics test. So, all the different techniques that you have learned, the different problems that you can see in a model that which we have explained previously, so all those problems we try to identify in this diagnostic stays, for example the functional specificity or a multicollinearity or heteroscedasticity.

So, all these different problems that we see in a model, we want to ensure that the problems are not there. So, we essentially do the different tests to actually get remedy of the problems. Then of course once we identify the problems, we try to remedy those problems. And when

we try to remedy those problems, we actually also get a sense of which variable might be creating the problem if you exclude some variable.

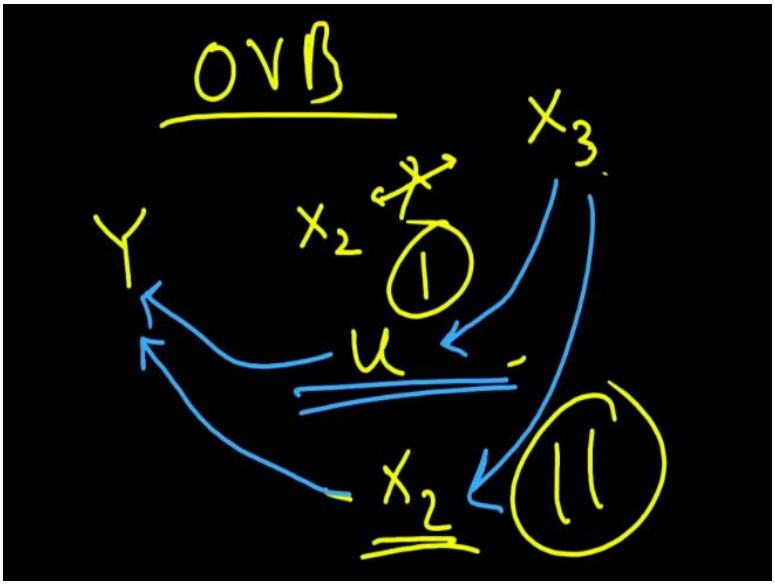
For example in omitted variable bias if you have heteroscedasticity or autocorrelation, so you might actually think that perhaps you have omitted some variable and try to see if you have actually omitted some variable and you try to include that. So, we just going to talk about omitted variable bias problem in a while, alright. Now, there are two ways.

(Refer Slide Time: 09:09)



Now, as I said model specification problem, we have the first thing, the first problem we said let us say exclusion of relevant variable. So this is the first thing we are going to discuss and then we will talk about inclusion of irrelevant variables, sometimes you do not remember the correct spelling. So exclusion of relevant I say. So this is something we call as omitted variable bias problem or OVB.

(Refer Slide Time: 10:11)

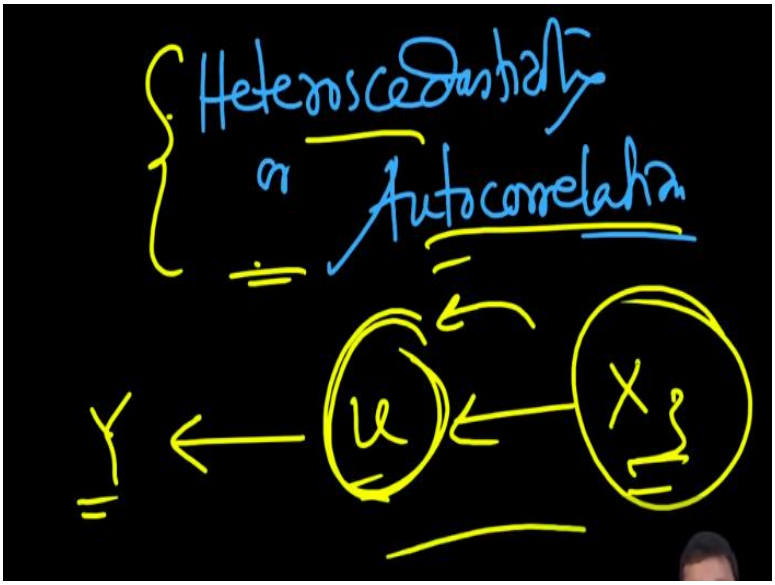


So when we actually omit a particular variable, so let me elaborate this omitted variable bias problem, so when you actually exclude a particular variable of interest. Now, there are two ways that variable of interest can actually influence your model. So, one it can let us say I have a model, let us say I have Y and then let us say I have included a variable X 2 and there is a variable which you omitted which is actually present in your model which is X 3.

Now, if you have not included X 3 in your model, so what will happen is that there are two ways your X 3 can actually influence the Y. So, I will use a different color, it will influence either through the error term. So, it will basically take the error route, so where X 3 and X 2 are not related. So, then it will influence the Y through error term or it may also include Y through X 2. So, instead of error term it can actually influence through, alright.

So, there are two ways the omitted variable can actually influence your model and we will explain how the two paths can actually show us different results. So, let me explain the first part in this lecture and we have already explained that in fact previously.

(Refer Slide Time: 11:39)



So, that first part is actually you remember the problem of heteroscedasticity or autocorrelation. So, first part is essentially that we have already talked about when I have omitted a variable and that variable is actually influencing my Y through my error term so that is where my problem of heteroscedasticity or error error term is the autocorrelation is coming. Now, in this case what is happening?

We see some pattern in the error term because some variable is actually excluded from your model. So, what is happening the entire pattern due to that X 3 variable is being captured by u and when it is captured by u what you see there are some either u is correlated with Y in cross sectional data when you see heteroscedasticity or you can see correlation between u error in different time period.

So, that is in autocorrelation in both cases what is happening is that because u was getting inflated, because the error is getting inflated, the variance of the model is getting inflated and when you basically estimate the t statistic, so the t statistic is going to be; so you are dividing the coefficient with a higher variance and the t statistic you get is lower and as a consequence what is happening is that you are finding certain variables to be significant whereas they might not be significant.

So, that is a problem that is seen in the case of heteroscedasticity or autocorrelation. So, this omitted variable, so this is the first thing let us say the assumption is that omitted variable is actually influencing the Y through u term, the error term and when that is happening I am

getting either heteroscedasticity or autocorrelation. So, that is through the case 1. Now, we will talk about the case 2 when X 3 is influencing Y through X 2.

And there we will see the extent of bias and extent of inefficiency that has been created. In this case, we have already seen in heteroscedasticity or autocorrelation there is no bias problem, but the problem remains in the fact that the standard deviation is a problem, the the model is inefficient because it is not the best model we have this in the model that we get here is having a higher variance, so we do not want that.

So, with this we will end the lecture and in the next lecture we are going to talk about the second part of the omitted variable bias problem where X 3 is actually influencing Y through X 2 route. So with this, we end this lecture here. Thank you.