

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

Lecture – 101

Problem of Endogeneity

Hello and welcome back to the lecture on applied econometrics.

(Refer Slide Time: 00:28)

Problem of Endogeneity
(Rsk of Instrumental Variable)

$$\boxed{\text{Cov}(x_i, u_i) \neq 0}$$
$$E(u_i | x_i) = 0$$

(i) OVB - $Y = \alpha_1 + \alpha_2 x_2 + u_i$

$\nwarrow \nearrow$
 (x_3)

We have been talking about instrumental variable and in this lecture we are going to talk about a particular type of problem that we often come across quantitative economics research problem and that we call problem of endogeneity. Now, before I begin let me tell you that sometimes endogeneity is construed as something like reverse causality but that is not true. And we will see endogeneity can come from various sources.

And there is under the broad rubric endogeneity, we can club all these problems together. Now, to explain endogeneity we will say that whenever we see a condition something like this. Let us say expectation or covariance of $X_i U_i$ is not equal to 0, we say there is a problem of endogenous. Now, we have seen this before particularly for a stochastic regressor that covariance of $X_i U_i$ may not be 0.

If the X_i and U_i are correlated and that is why we had to have specific condition for stochastic regressor, specific Gauss Markov assumption. We had to satisfy for stochastic regression that was expectation of U_i given X_i is equal to 0. Now, whenever to simplify our

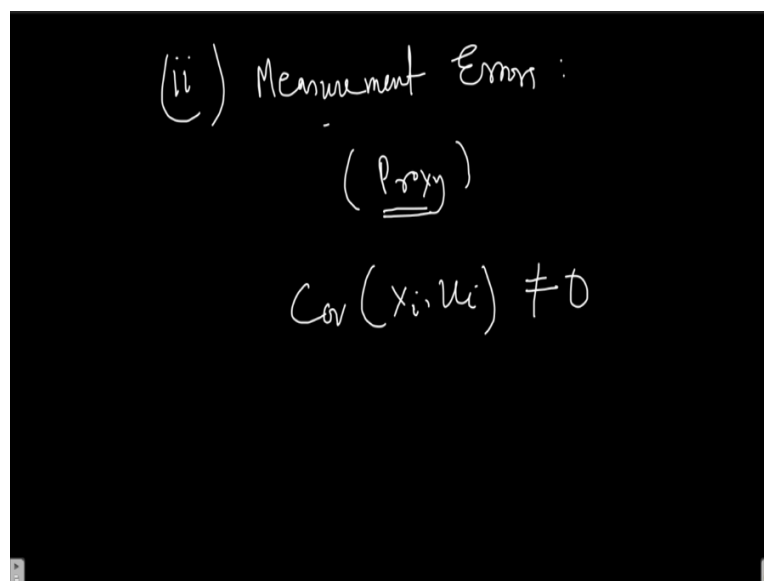
definition, whenever we see of this condition to exist, we call there is a problem of endogenic. Whenever we have this problem covariance $X_i U_i$ is not equal to 0.

We say that there is a problem of endogeneity. Now, we have already seen this problem to be existing in different occasions and one such occasion is omitted variable bias and I am just going to outline the different sources where from endogeneity may come. And in all these different sources we will have this particular criteria visible. So, we will see covariance of $X_i U_i$ is not equal to 0 and one such occasion is omitted variable bias.

And what happens in case of omitted variable bias? We know that let us say there is a one important variable which is let us say my regression equation is this. Its $\alpha_1 \alpha_2 X_2$ and U . Let us say some you know important regressor is missed out and that regressor X_3 is influencing a part of it is influencing X_2 and also U . Now, what happens is that because it is omitted and a part of it is actually influencing X_2 and a part of it is influencing U .

So, what happens is that we will see that X_2 and U are not independent? So, we will see this particular formula, the particular relation to hold in case of omitted variable bias. And whenever we have such criteria fulfilled, we call that there is a problem of endogeneity which is do not want or if you do not want, we need to actually get rid of this problem. But we will see how we can get rid of problem but before that let me actually outline all the different sources of this endogeneity.

(Refer Slide Time: 03:23)



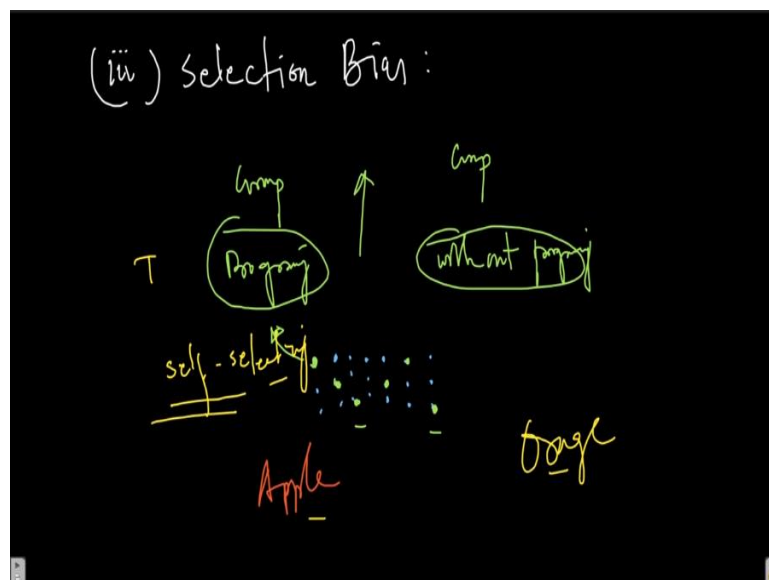
(ii) Omitted Error :
(Proxy)
 $Cov(X_i, U_i) \neq 0$

Now, the second problem we already spoke about is the measurement error. What is the measurement error? Measurement error occurs when we do not have the right variable for our model. So, let us say we want to measure $i q$ but we do not have any data on $i q$ or talent. So, what we do? We actually use education as a proxy. Now, the proxy often is not the perfect proxy and if we do not have a perfect proxy we have seen that it might be a problem.

Because, if we have an ideal proxy which is really good, we do not have any problem, we can use an ideal proxy but in most of the cases, it is very very difficult to get an ideal proxy. And we use an imperfect proxy and when you use an imperfect proxy we actually see this condition to be there. This relation to exist. So, covariance $X_i U_i$ is not equal to 0 and we call that the endogeneity problem occurs when we have measurement error.

Now, in this lecture so, these two problems we have already you know already sort of we are familiar with. I will talk about the third and the fourth source of endogeneity.

(Refer Slide Time: 04:37)



And third source, if you use a new page, the third source is the selection bias, the third source of endogenic is selection bias. Now, what is the selection bias? I will actually explain that with an example. Let us say in our, let us say this is a school of management and all the students have come here for their MBA and they do their regular MBA course and then they go and go out in the job market and then they bag all these consulting and investment banking jobs and so forth.

Now, Let us say you know realizing that programming language, knowledge in programming language is really, really important. So, perhaps let us see in our business school we have actually started a programming language course. Let us say we are offering python, C, C++ or something. Now, let us say we have started that back in you know 2018 and today we want to actually see, if our program has actually yielded any result.

Because, we have to take a call whether we want to continue that program or not. Now, what do you do? We see that the cohort, the batch which has gone to the labour market and they are basically there for last 3 years and Let us say we see the wage difference between the group who took that programming course and which have been the the group who did not. Now, if I compare that.

Let us say I actually have found that the group which took the programming course actually have gotten a a better return from the labour market. So, their wages are higher. Let us say, the group with the programming language so, this is a treatment variable programming has got a better wedge with a group without programming. Now, is this comparison fair? That is the first question.

This comparison is not actually fair and the reason is in all likelihood, it is not fair and the reason is that when we had let us say. These is our group students, all our students and some students will select the programming course, some students will not select the programming course. Let us say, it is given voluntarily, students can choose which one they want to attend.

Let us say they want to attend the programming and let us say this group, this green group let us say they have actually gone for the programming course. And the remaining blue they did not go for the programming course. Now, it can so, happen that and it is in all likelihood that might be the case that all the students who actually took that programming course despite the rigor and hard work that is needed for to complete their regular MBA course.

These guys can actually be really hard working and they are actually like perhaps they are more dry even they are more sort of you know eager to get success or they are actually you know they are smart enough to get all these things done and they can actually learn programming language alongside their regular MBA curricula. So, they will inherently, invariably they are going to do better in the labour market.

Because they have so and so criteria then there is so and so crates. So, essentially when we so, this group, this green group actually self-selecting themselves into the treatment. This treatment here and this green group is actually self-selecting, they are selecting themselves to this treatment. Now, because of this the comparison here is not really with apple and orange is not really apple and apple.

But it is basically a comparison of apple and orange you can say apple and orange. We do not want that and we will see when we actually use some technical called randomization to actually get rid of this self-selection. But this can be a problem in you know, this is a problem in economics wherever will be experimental research or even with observational data the self-selection problem could be there.

And we can either randomize and also instrumental variable could be an important tool to actually get rid of the self-selection problem and we will talk about it.

(Refer Slide Time: 08:54)

(iv) Reverse Causality:

$$\text{Child Labor} = \alpha_1 + \alpha_2 \text{ family Income} + \alpha_3 \text{ no of children} + \alpha_4 \text{ mother's Edu} + u$$

$$\text{Family Income} = \beta_1 + \beta_2 \text{ father's Edu} + \beta_3 \text{ father's employment} + \beta_4 \text{ Child Labor} + v$$

Now, the last source of endogeneity is reverse causality. And that is something again very close to the heart of an econometrician reverse causality. So, let us talk about reverse causality. Reverse causality so, basically when you see the relationship between Y and X. So, essentially what we see? Let us say there is a Y and there is X. So, in a regression equation we always try to understand the impact of X on Y.

This is what we try to understand. But in many instances we will see that actually there is a link or there is an influence that is operating between Y to X. So, X is not really independent here or exogenous here. I explain the term exogenous but rather, it is an endogenous variable we will just talk about endogenous and exogenous variable. But here, it is note that the causality, there is a circularity here.

So, let me give you an example to illustrate this. So, let us say that the so, let us say the child labour case. So, there are many families in India or any other poor countries where the you know children, if it is a poor family what happens is and if they have many kids. Let us say one kid actually joins in the labour market and they do some work to sort of add to the family income. So that is not completely unknown, it is rather common phenomena.

Now, if let us say I want to actually build a model around participation in labour force of the child labours. So, let us say I write child labour is my dependent variable and in my independent variable let us say I have now, why the children from families will join labour force. So, one reason of course is that the family income α_2 into family income. So, if the family income is less, it is likely that the kids will actually join the labour market.

The next point could be α_3 , number of children. So, if a family has a lot of children usually and they are poor so, some of the kids may actually go for the labour market. Particularly the elder one might actually go to the labour market. Then there could be something like mother's education. So, if the mother is educated, she might actually resist the children from joining the labour force. So, mother's education.

And let us say I have some error terms. Let me reduce the size a little bit and let us have some air return. Now, we can understand, if the relationship between all these values. So, let us particularly look at this particular two variables. So, child level participation and family income. So, if family income is less, I can convincingly argue that there is a possibility that one or two kids from the family may join in the labour market.

Now, could there be a case where participation in labour market you know child participating in the labour market may actually add to the family income. And you will see actually that is the case because, if a kid is actually joining the labour market and he or she is usually doing

that because he or she wants to actually add to family income. So, the moment a kid is you know participating in the labour market that might actually increase the family income.

So, I can write this equation as, family income is equal to left side is beta, beta2. Let us say we have all these different variables father's education then I have father's let us say father's income or father's employment status and then I will also have what is the child labour and there can be a you know whole lot of other factors. But child labour could also be a factor here. Now, essentially that is what we call the problem of endogeneity.

So, child labour is influencing family income whereas family income is influencing child labour. So, they are intrusions they are not really independent they are not really exogenous. And I will explain this term exogenous and endogenous. So, here what you see is that this variable, this child labour variable and the family income variable, they are not really, they are actually getting determined by the model.

(Refer Slide Time: 13:35)

Endogenous Variable / Exogenous Variable

$$b_{2,ols} \neq \beta_2$$

$n \rightarrow \infty$

$$b_{2,ols} \neq \beta_2$$

• Causality $\rightarrow [Cov(x_i, u_i) \neq 0]$

So that because they are determined by the model so, I will tell that them as endogenous variable. But there could be some other factor which are not determined by the model but let us say given from outside and that could be exogenous variable. We will talk more about this endogenous variable and exogenous variable going forward. But essentially this is the idea of reverse causality.

And when we have reverse causality the beta 2 b 2 ols estimate of beta 2 is actually wrong or in the sense, it is biased. So, when it is biased, it is not b 2 ols is not equal to beta 2 and it is

inconsistent. So, what inconsistency means? If n tends to infinity, b_2 OLS does not converge to β_2 . So which we do not want, we want as n tends to infinity b_2 OLS should converge to β_2 . But we will see that is not the case.

Now, essentially we talked about all these four problems now. So, omitted variable bias, measurement error selection bias and reverse causality. So, all these four problems comes under the rubric of endogeneity problem. And we will see that instrumental variable can actually play an instrumental role in actually addressing this endogeneity problem you know irrespective of whichever source they are coming from.

Of course there are lot of you know care that you need to take while choosing an instrument. But actually theoretically instrumental variable can actually address all these problems which we will talk about in the next lecture. So, essentially what happened? You know when we talked at the beginning of the lecture that usually people are shy when to talk about causality and statisticians and econometricians.

There is a little difference between two discipline. The econometricians also build their statistical tool over a period of time and instrumental variable is one of those monumental contribution that econometricians have made in this domain. And essentially econometricians are more vocal about causality because economics the we need to we just do not need to build tools.

But we rather need to solve problems, we rather need to understand the cause of linkages. And that is why econometricians have always been obsessed with causality and that is why we have actually come up with different you know weapons in our arsenal. And one such weapon is instrumental variable. Now, the reason I spoke about all these different problems is that actually using instrumental variable, we can address these problems basically.

The problem that we spoken at the beginning that is the covariance of $X_i U_i$ not equal to 0. So, this particular problem when it is there which is basically common for all these four cases. So, we will see that instrumental variable can actually help us to get rid of this particular condition. So, with this we end this lecture here. Thank you.