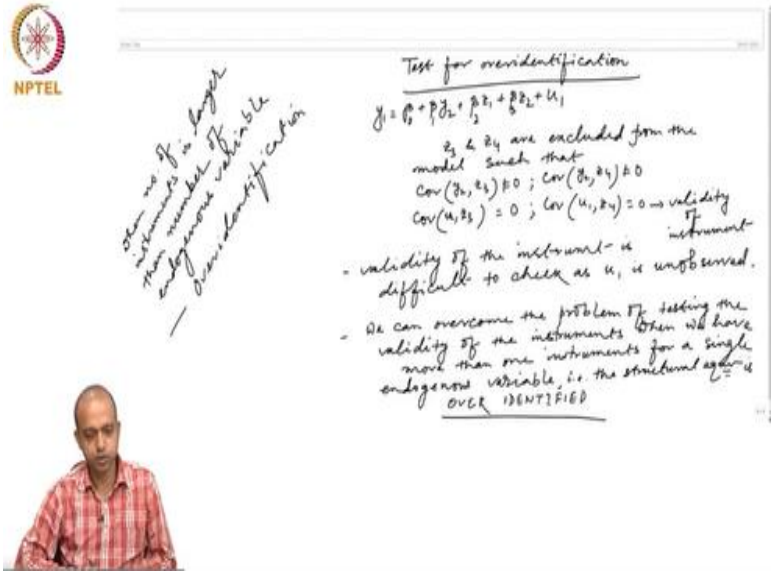
Applied Econometrics
Prof. Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras

Lecture - 07
Instrumental Variable Estimation – Part VII

(Refer Slide Time: 00:16)



Welcome once again to our discussion of instrumental variable estimation technique and in our last class we were discussing about Hausman test of endogeneity. We said that we must ensure that there is endogeneity in one or more than one explanatory variables before we actually implement or 2 SLS technique for estimation. Why this is so? Because we said that 2 SLS estimates actually they show larger standard error than the OLS estimates.

So, that means efficiency beta hat 2 SLS less efficient compared to OLS. So, that is the cost of using IV estimation and that is the reason we must ensure that there is endogeneity problem before implementing this solution. And two things we have discussed, firstly testing endogeneity when you have only one endogenous variable and secondly, we also discussed about testing endogeneity when we have more than one endogenous variables.

And we said that when we have more than one endogenous variable ah the problem is that the test says only one among these two suspected variables is endogenous but we are not sure that which

among these two variables are actually endogenous. Now today we will discuss another interesting feature of this 2 SLS estimation technique and that is called test for over identification.

Now let us once again assume that this is our model $Y_1 = \beta_0 + \beta_1 Y_2 + \beta_2 Z_1 + U_1$. And we assume that there are two excluded variables Z_3 and Z_4 are excluded from the model such that both of them are highly correlated with y_2 that means covariance between y_2 and since we have Z_1 only let us say I will write beta3 Z_2 also here plus u 1 so that Z_3 and Z_4 are excluded. So, y_2 and Z_3 they are highly correlated.

Similarly, covariance between y_2 and Z_4 also correlated. Both of them are highly correlated with y_2 but they are not correlated with the u 1. Covariance between u 1 and Z_3 = 0 and covariance between u 1 and Z_4 also equals to 0. Now the second condition we said that these are the two conditions that an instrument a true instrument must exhibit. The first condition that means the strength of the relationship shows whether the instrument what we identified is actually a strong instrument or a weak instrument.

And the second condition where we are showing that whether instruments are actually uncorrelated with the error term or not. That basically this is u 1 so that basically shows whether the instruments are valid or not. So, the second condition is validity this is called validity of instrument. Now if you recall initially when we started our discussion on this instrumental variable estimation technique, we said that the first condition is easy to check.

Because we can simply run a reduced form equation for the endogenous variable and we can test the individual or joint significance of these two instruments that is very simple. So, how will you check the first condition? That means strength of the instrument simply running the reduced from equation for y_2 where y_2 is actually a function of all exogenous variables included as well as excluded.

So, we can simply implement the apps test to check the joint significance of Z_3 and Z_4 and we can test their individual significance as well by t Test but how to test the second condition? Second

condition that means validity of the instrument is difficult to check as u 1 is unobserved. But this problem we can overcome when we have more than one instrument for the endogenous variable.

So, here we have only one endogenous explanatory variable y_2 but we are having two instruments Z_3 and Z_4. That is why we say that the system the structural equation is actually over identified. So, that is why we say we can overcome the problem foreign of testing the validity of the instruments when we have more than one instruments for a single endogenous variable, that means the structural equation is over identified.

So, now we have understood what is over identification. It is a very simple condition when we have one endogenous variable. So, that means number of excluded exogenous variable number of instruments should be more than number of endogenous variables. So, when number of instruments is larger than number of endogenous variable we call the over identification. Now when there is over identification that means when the; structural equation is over identified.

We have more than one instruments for the single endogenous variable then it is easy to implement it is easy to check the validity of those instruments as well. Now we will discuss how to check validity of instrument.

(Refer Slide Time: 12:01)

So, this is our model $Y_1 = \beta_0 + \beta_1 Y_2 + \beta_2 Z_1 + \beta_3 Z_2 + U_1$ and Z_3 and Z_4 are instruments. So, first step, step one let us say this is equation 1. So, in step one we estimate equation 1 by 2 SLS method using only Z_3 and get $\widehat{u_1}$. And in the second step we examine whether $\widehat{u_1}$ is correlated with Z_4 or not. So, that means if covariance between $\widehat{u_1}$ and Z_4 = 0 then we conclude that Z_4 is a valid instrument.

Now here the question is what is the implicit assumption in this procedure? If you think closely what we are saying that we estimate equation one by 2 SLS method using only Z_3 as instrument and get $\widehat{u_1}$ that means predicted value of the residual. And then we are testing the covariance between $\widehat{u_1}$ and Z_4. If that is 0, we said that Z_4 is a valid instrument. Similarly, if we reverse this process that means in step one instead of using Z_3 as an instrument.

If we use Z_4 as a valid instrument and get $\widehat{u_1}$ then actually what will happen, we will examine the covariance between $\widehat{u_1}$ and Z_3. And if that becomes zero then we will say Z_3 is a valid instrument. So, the implicit assumption here when we are using one excluded exogenous variable as instrument in the first stage that means we are assuming that there is at least one valid instrument, is not it?

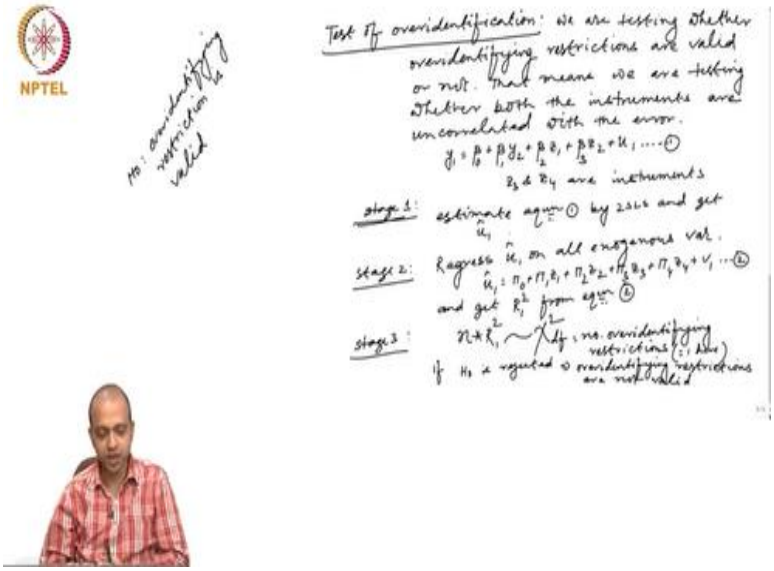
When I am using Z_3 in the step one, I am assuming that actually Z_3 is a valid instrument. But what is the guarantee? That means there is no guarantee that Z_3 is actually a valid instrument and if that is a valid instrument then we can use that instrument only to get the efficient estimates of beta 0, beta 1, beta 3 hat. So, even though this test is applicable, we can implement this test to check the validity of the instrument.

The implicit assumption here is that there is at least one valid instrument. If that assumption is violated that means in the first stage itself if Z_3 is not a valid instrument, then we cannot test the validity of Z_4. Similarly, if Z_4 is not valid instrument we cannot test the validity of Z_3. So, when you are using one instrument in the first stage itself, we are making an assumption that particular instrument is actually a valid instrument then we proceed further.

Now this test is also known as a test of over identification. Why? Because this test involves two instruments. When you have only one endogenous variable that means actually, we are testing

whether the over identification is valid or not, that means whether both the instruments are actually uncorrelated with the error term or not.

(Refer Slide Time: 18:55)



So, this test is also known as test of over identification. That means we are testing whether over identifying restrictions are valid or not that means we are testing whether both the instrument are uncorrelated with the error. So, if this is your equation $Y_1 = \beta_0 + \beta_1 Y_2 + \beta_2 Z_1 + \beta_3 Z_2 + U_1$ and we have two excluded variables Z_3 and Z_4 let us say equation one Z_3 and Z_4 are instruments. So, in stage one what we do estimate equation 1 by 2 SLS and get $\widehat{u_1}$.

Then in stage two, regress $\widehat{u_1}$ on all exogenous variables. So, that means this type of equation $\widehat{u_1}$ equals to some $y_2 = \pi_0 + \pi_1 Z_1 + \pi_2 Z_2 + \pi_3 Z_3 + \pi_4 Z_4 + v_1$ and this is called an auxiliary regression. And get let us say this is equation 2, R_1^2 from equation 2. In stage 3 what we do? We multiply R_1^2 by number of observations. If we multiply into R_1^2 then that basically follows a chi square distribution with degrees of freedom equals to number of over identifying restrictions.

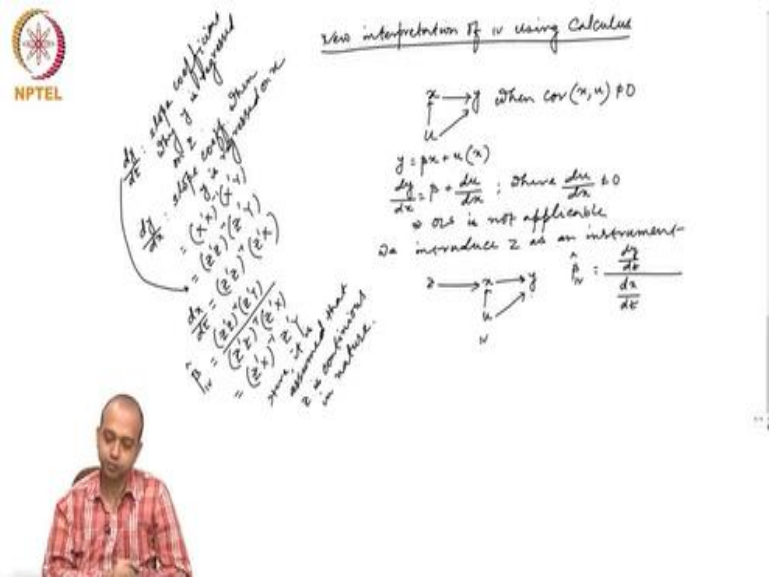
What is the number of over identifying restrictions here? We have one endogenous variable we have two instruments so that means actually if there is only one endogenous variable if and there is only one excluded exogenous variable that means we will say the system is exactly identified. But when we have more than one that means we have two we will say that there is only one over identifying restriction.

So, number of over identifying restriction in this case equals to one here. And what is the null hypothesis we have to keep clearly in mind. Null hypothesis here over identifying restriction is valid. So, that means if this is called calculated R square calculated chi square if that calculated chi square is actually greater than the tabulated R square at 1 percent or 5 percent level of significance then we will reject the null.

So, if H_0 is rejected then our conclusion is that over identifying restriction is actually not valid because it is rejected. It implies over identifying restrictions are not valid. And if H_0 is not rejected then we will say that over identifying restriction is actually valid so that means both the instruments are actually uncorrelated with the error term. So, we are achieving two things, we are by testing the validity of over identifying restriction, we are actually testing the validity of instrument itself.

So, this is all about over identification how to test over identification over identifying restriction and how to test whether the instruments are valid or not. Now once this is done what we will do?

(Refer Slide Time: 27:08)



We will get a new interpretation of IV using calculus. So, let us say that we have x which is actually the explanatory variable we have y here and we have u here. Now when we apply OLS that means when we assume that x and u they are actually not correlated. So, that means direction of position

is like this x crosses y then y is also was crossed by u that means u and y correlated x and y correlated but there is no correlation between x and y.

Now what happens? In presence of the covariance between x and u not equals to 0 we get this direction as well and we say that this OLS is not applicable. Because what happens here the model is then $y = \beta X + u$ where u is also a function of x. Then $\frac{dy}{dx} = \beta + du/dx$ where du/dx is actually not equals to 0. And that is the reason OLS is not applicable when this type of direction is there.

So, this implies OLS is not applicable. And what is the solution? We introduce z as an instrument. When z is introduced as an instrument this diagram this arrow diagram changed as x crosses y, u crosses y, u is also correlated with x. Then we have z where z is correlated with x but z is not directly correlated with y and z is also not correlated with this, this is also not there. So, that is why we can simply remove this type of direction to avoid confusion.

So, that means we will get the impact of z on y only through x, z will cross x and x will again cross y so that means there is some kind of indirect impact of z on y. For an example when we are discussing about which is a function of education, we said that father's education is correlated with education and education in turn will affect which so that means we will get some indirect impact of father's education on somebody's wage.

So, that means when you are doing beta hat IV which is actually nothing but $\frac{dy}{dz}$ impact of y on z, impact of z on y. How is it happening? Through dx dz, z will cross only x and then we will get some indirect impact of z on y. So, this is the IV estimates of beta hat and the interpretation is $\frac{dy}{dz}$ divided by dx dz. Because what happens here you will if you do a $\frac{dy}{dz}$ into dz dz so ultimately it becomes dy dx.

So, $\frac{dy}{dz}$ is the indirect impact of z on y, why it is happening? Because z is impacting x as well. Now what is $\frac{dy}{dz}$? $\frac{dy}{dz}$ is slope coefficient when y is actually regressed on z. Now when y is regressed on

x and what is dy dx? dy dx is when slope coefficient when y is regressed on x. And what is this value equals to we can say that x prime in matrix form x prime x inverse x prime Y. So, obviously $\frac{dy}{dz}$ is nothing but then what would be this?

This would be equals to z prime z inverse z prime Y. And dx dz equals to what will happen? z prime z inverse z prime X. So, beta hat IV then equals to what will happen? So, $\frac{dy}{dz}$ equals to what I have written z prime z inverse z prime Y divided by z prime z inverse z prime X. So, z prime z inverse and z prime z inverse will get cancelled out. So, we will get z prime X inverse z prime Y.

So, if you recall X prime x inverse x prime Y, we are just substituting x for z and we are getting beta hat IV using calculus. Here it is assumed that z is actually continuous in nature. So, this is how we can understand the interpretation of IV instrumental variable estimates using calculus. This is a very simple diagram but it gives new insights this arrow diagram it shows that X and Y are correlated, Y and u correlated.

And the moment so this becomes when covariance between x and u not equals to 0 this is the diagram what we get and this is basically the IV. So, we are introducing z as an instrument for x which is supposed to be endogenous here. So, that means impact of z on Y is actually an indirect impact. So, with this we are basically closing our discussion today and rest of the things we will discuss tomorrow.