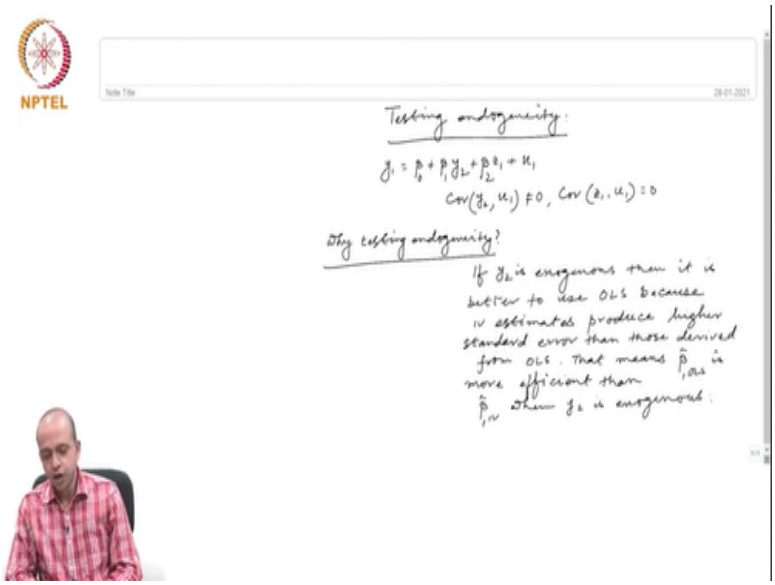
Applied Econometrics
Prof. Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras

Lecture - 05
Instrumental Variable Estimation – Part V

(Refer Slide Time: 00:16)



Previously we were discussing about the 2 SLS mechanism when we have multiple explanatory variable in the model and multiple instruments as well. Now after discussing all those one natural question that comes to our mind is that instrumental variable estimation technique is suggested when there is endogeneity.

That means unless endogeneity is confirmed in the model should we use a instrumental variable estimation technique or should we use OLS. So, that means is it essential to check endogeneity first and then implementing its solution as IV technique. Now to answer that question today we are going to discuss about testing endogeneity. let us say that this is our model,

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$
$$Cov(y_2, u_1) \neq 0$$
$$Cov(z_1, u_1) = 0$$

we are suspecting y_2 to be endogenous it is still not conformed but z_1 is an exogenous variable. We are not suspecting anything about z_1 because of the theoretical reasons or our own logical

reasoning. So, at this stage what I am saying we are only suspecting y_2 to be endogenous in the model.

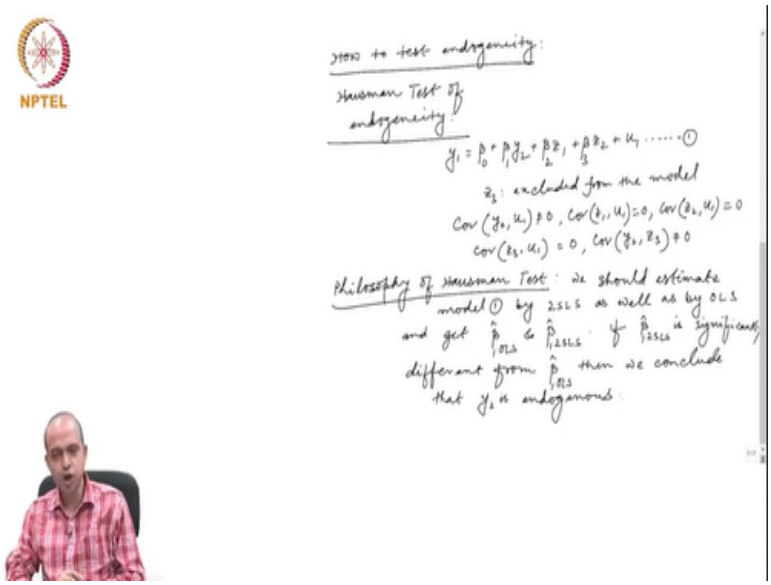
For example, when we are estimating the wedge function, we are suspecting that education to be endogenous in the model but experience is not. Now the question is why endogeneity to be tested? The answer is if y_2 is exogenous then it is better to use OLS because IV estimates produce higher standard error than those derived from OLS.

$\hat{\beta}_{1OLS}$ is more efficient than $\hat{\beta}_{1IV}$ when y_2 is exogenous. That is the problem, if we do not test endogeneity and simply implement IV estimation technique because IV can be applied even though y_2 is actually an exogenous variable.

What is the problem? Problem is that when we have y_2 is actually exogenous and if we still implement the two SLS method then what will happen, $\hat{\beta}_{12SLS}$ will have larger standard error than $\hat{\beta}_{1OLS}$. And larger standard error means the efficiency property would be lost. It means $\hat{\beta}_{1OLS}$ is more efficient than $\hat{\beta}_{12SLS}$ when y2 is actually exogenous.

That is why without conforming y_2 is actually an endogenous variable we cannot implement IV or 2 SLS technique to estimate this model, is this clear? So, now the next step is then how to test endogeneity?

(Refer Slide Time: 07:41)



So, the famous test available in econometric literature is suggested by Housman test of endogeneity. First of all, let us say that this is the model

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1$$

and our excluded variable is z_3, z_3 is excluded from the model such that

$$Cov(y_2, u_1) \neq 0$$

here y_2 is suspected to be endogenous in the model.

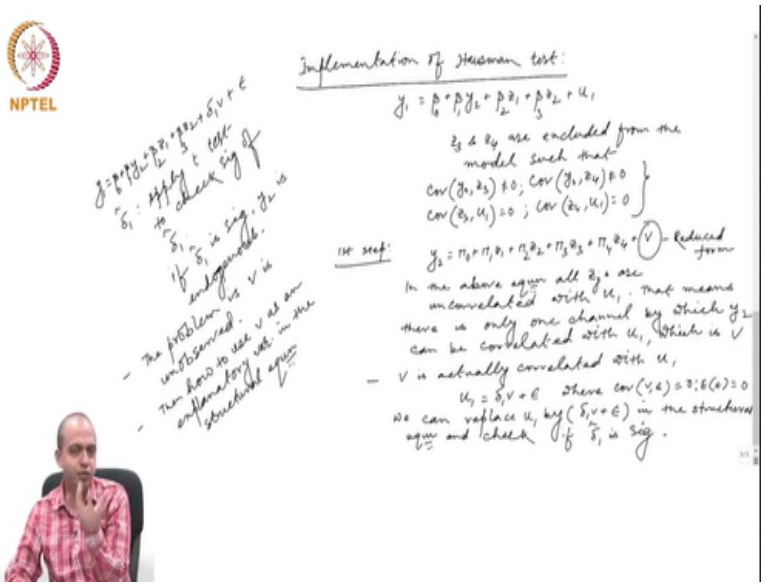
$Cov(z_1, u_1) = 0$; $Cov(z_2, u_1) = 0$

$Cov(z_3, u_1) = 0$; $Cov(y_2, z_3) \neq 0$

z_3 we can use as an instrument to estimate this model, provided that y_2 is actually an endogenous variable.

Philosophy of Housman test: Now the Housman tests philosophy is very simple he suggested we should estimate above model by 2 SLS as well as by OLS. And get $\hat{\beta}_{OLS}$ and beta $\hat{\beta}_{2SLS}$. If $\hat{\beta}_{2SLS}$ is significantly different from $\hat{\beta}_{1OLS}$. Then we can conclude that y_2 is endogenous. The philosophy is very simple to understand also because if there is no endogeneity then $\hat{\beta}_{OLS}$ and $\hat{\beta}_{2SLS}$ will produce the same estimates but if they are not then of course $\hat{\beta}_{OLS}$ would be significantly different from $\hat{\beta}_{2SLS}$ that is the philosophy. But how to implement this test?

(Refer Slide Time: 13:23)



Let us now talk about implementation of Housman test? Let us say that this is our model

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1$$

Here we assume that y_2 is endogenous z_1 and z_2 they all are exogenous and then there are two excluded exogenous variables z_3 and z_4 are excluded from the model such that both of them are correlated with y_2.

$Cov(y_2, z_3) \neq 0$; $Cov(y_2, z_4) \neq 0$

$Cov(z_3, u_1) = 0$; $Cov(z_4, u_1) = 0$

z_3 and z_4 can be treated as instruments. So, now how will you estimate this model using 2 SLS? We have already discussed the first step is we have to write a reduced from equation for the endogenous variable.

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v$$

this is the reduced form equation. In this reduced form equation of y_2 this z_j is are all uncorrelated with u_1 because they are all exogenous variable. So, in the above equation all are uncorrelated with u_1. Because of the assumption z_1 and z_2 I have already mentioned that they are exogenous variable z_3 and z_4 they are identified as instruments.

So, based on that assumption they are also uncorrelated with u_1. So, if that is the case there is only one channel by which y_2 can be correlated with u_1. Can you think of what is that channel? Look at

this equation. In this equation all these z_j 's are exogenous variable they are no way correlated with u_1. So, that means y_2 has only one component or indirectly I can say that there is only one channel through which this y_2 can be correlated with the error term in the structural equation u_1.

What is that channel? What is that component? which is actually V, this is the component otherwise y_2 cannot be correlated with u_1. Otherwise y_2 cannot be an endogenous variable. Now if V is the channel through which y_2 is correlated with u_1 that means we are assuming V is actually correlated with u_1. So, that means we are saying V is actually correlated with u_1. If that is the case, we can write a relationship between u_1 and V.

$$u_1 = \delta_1 v + \epsilon$$

$Cov(v, \epsilon) = 0$; $E(\epsilon) = 0$;

It means this v and this error term is actually not correlated and epsilon shows the assumption of classical linear regression model. Now what we can do? We can actually replace u_1 by $\delta_1 v + \epsilon$ in the original structural form equation and check $\hat{\delta}_1$ is significant or not. So, that means if we replace this in the original structural form equation our structural form equation will become

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 v + \epsilon$$

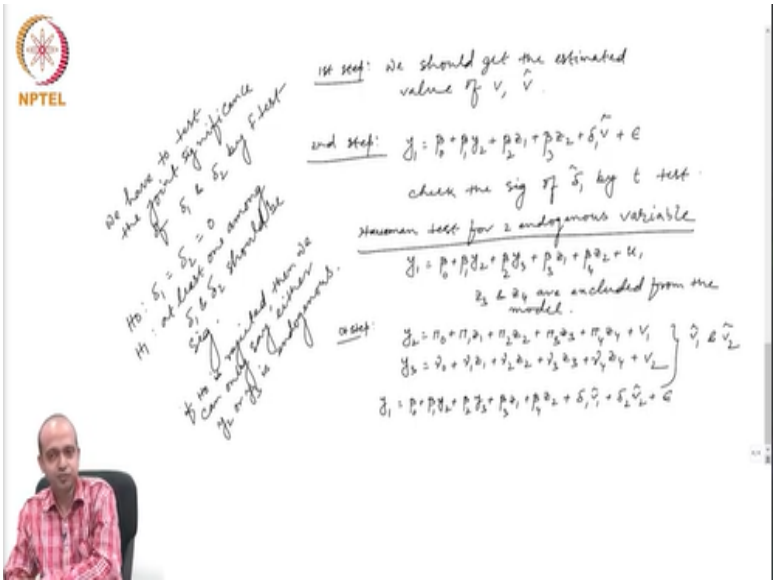
$\delta_1 v + \epsilon$ this is the new error term.

Because v is used as an additional explanatory variable in the; reduced form equation. From there will $\hat{\delta}_1$ and again we will put that estimate will put that as value in the original structural form equation and then we will get $\hat{\delta}_1$. And apply t test to check significance of $\hat{\delta}_1$. If $\hat{\delta}_1$ is significant y_2 is endogenous.

But there is one problem in this mechanism this v what we have hypothesized here in the reduced form equation this error term is unobserved error term is actually unabsorbed. So, the problem is V is unobserved. So, if V is unobserved how I will use V as an additional explanatory variable. Then how to use V as an explanatory variable in the structural equation, the way I have written I have inserted V in the structural equation that means we are using V as an additional explanatory variable.

For that we must have data on V. V we must be observable quantity but that is not the case because it is the error term as we have hypothesized in this equation. So, what is the solution. The solution is then this reduced form equation should be estimated by OLS and we should get the estimated value of this error term which is $\hat{v}$.

(Refer Slide Time: 25:43)



That is why from the first step we should get the estimated value of V which is $\hat{v}$. And then in the second step or second step what we have to do we have to say that

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{v} + \epsilon$$

Instead of V we are putting the estimated value of V which and then check the significance of $\hat{\delta}_1$. If the t test confirms that $\hat{\delta}_1$ is actually significant then we will say that y_2 is endogenous. This is actually the procedure of Housman test.

Now how can we extend this Housman test? Let us now say there are two endogenous variable in the model. Housman test for two endogenous variable, two or more than two. Our model is now

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 z_1 + \beta_4 z_2 + u_1$$

here there are two endogenous variable y_2 and y_3.

$Cov(y_2, u_1) \neq 0$; $Cov(y_3, u_1) \neq 0$

$Cov(z_3, u_1) = 0$; $Cov(z_4, u_1) = 0$

and we have two excluded variable z_3 and z_4 are excluded from the model.

Then we have to just extend the philosophy or procedure of Housman test in the context of one endogenous variable in this context when you have two endogenous variables.

So, that means in the context when we had one endogenous variable we were writing only one reduced form equation in this case we will get two reduced from equation.

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_1$$

$$y_3 = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3 + \gamma_4 z_4 + v_2$$

And after estimating these two reduced form equations, what we will get? We will get from these two we will get $\hat{v}_1$ and $\widehat{v_2}$ and then this estimated value of this error term we have to replace in the original structural form equation. So, if we replace that what will happen

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 z_1 + \beta_4 z_2 + \delta_1 \hat{v}_1 + \delta_2 \hat{v}_2 + \epsilon$$

And now what we have to do? We have to test the joint significance of this $\hat{v}_1$ and $\hat{v}_2$ so then next step. We need to we have to test the joint significance of δ_1 and δ_2 by F test. In that case what would be our null hypothesis in that case?

H_0: $\delta_1 = \delta_2 = 0$

H_1: at least one among δ_1 and δ_2 should be significant.

This is the way we can extend the mechanism of single endogenous variable to multiple endogenous variables for testing endogeneity using Housman test. But there is one problem when we have more than one endogenous variable and we apply Housman test. What is the shortcoming of this? See the alternative says at least one among this δ_1 and δ_2 should be significant.

So, if H_0 is rejected then we can only say either δ_1 and δ_2 significant. We can only say either y_2 or y_3 is endogenous. Of course, there is a possibility that both of them are endogenous but this test suggests that at least one. So, that means which among these two is actually endogenous that cannot be confirmed by this F test as suggested by Housman.

There is another mechanism to confirm that endogeneity that we will discuss later. So, we must understand the shortcoming of Housman test for testing endogeneity when we have more than one endogenous variable. What is the shortcoming? We are applying F test to test to check the joint significance. Since the alternative of F is at least one among them is significant so we can only say at least one is endogenous in this model.

But we do not know which among these two variables y_2 or y_3 is actually endogenous so either one whether y_2 or y_3 or both are endogenous that we do not know. To confirm about endogeneity for which particular variable we need to apply another test that we will discuss later.