

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

Lecture - 04
Instrumental Variable Estimation – Part IV

(Refer Slide Time: 00:15)



*if $z_2, z_3 \rightarrow$ use z_2 as instrument
 if $z_2, z_3 \rightarrow$ use z_3 as instrument
 - choose one which is actually
 more correlated with y_1
 if we need to use a linear
 combination of z_2 & z_3 as
 instrument*

IV technique for more than one instruments

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

z_2 & z_3 are excluded from the model

such that,

$$\text{Cov}(z_2, u_1) = 0; \text{Cov}(z_3, u_1) = 0$$

$$\text{Cov}(z_2, z_3) \neq 0; \text{Cov}(z_3, z_2) \neq 0$$

$$E(u_1) = 0; \text{Cov}(z_1, u_1) = 0$$

ln (2) eq) = $\beta_0 + \beta_1 \text{educ}_m + \beta_2 \text{educ}_f + u_1$

*z_2 : father's education
 z_3 : mother's education*

which instrument to be used? z_2 or z_3 or both?

$$y_2 = \pi_0 + \pi_1 z_2 + \pi_2 z_3 + v_2 \rightarrow R_2^2$$

$$y_2 = \gamma_0 + \gamma_1 z_2 + \gamma_2 z_3 + v_3 \rightarrow R_3^2$$

Now what we will do? We will bring back the estimation technique what we discussed in our previous class in a different context. What is this context? In this context we have now two instruments.

Here we will discuss IV technique for more than one instrument. What is our model?

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

this is our model. And we assume that z_2 and z_3 are excluded from the model such that

$$\text{Cov}(y_2, z_2) \neq 0 ; \text{Cov}(z_2, u_1) = 0$$

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

that means z_2 and z_3 both of them are correlated with the endogenous variable but they are not correlated with the error term.

$$E(u_1) = 0 ; \text{Cov}(z_1, u_1) = 0$$

z_1 is an exogenous variable included in the model.

Now as an example, what is the example of this model? Let us say that this is

$$\ln(\text{wage})_i = \beta_0 + \beta_1 \text{education}_{1i} + \beta_2 \text{experience}_{2i} + u_1$$

Here let us assume that z_2 is basically father's education and z_3 is mother's education. So, obviously father's education and mother's education both are correlated with the individual's education let us assume. Now if that is the case the question that comes to our mind which instrument to be used?

So, this is the question, which instrument to be used? z_2 or z_3 or both that is the question. Now given this is our situation apparently what we can think of let us regress the endogenous variable on z_2 and z_3 separately that means let me form the reduced form equation in this way,

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2$$

Then from this model we will collect R square let us say this is R_1 square.

Then what we will do we will regress y_2 on z_3 separately let us say this is

$$y_2 = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_3 + \epsilon_2$$

from this model we will get R square. Depending on which model gives higher R square I will use that model as an instrument, is that correct? So, I will run both the model and then what I will do? I will first use that means what I am doing, I am first using z_2 as an instrument from model one. Then I am recording the R square which is denoted as R_1 square. So, that means what I am saying basically if R_1 square is greater than R_2 square then use z_2 as instrument. Let us say z_2 means father's education. If R_2 square is greater than R_1 square then use z_3 as instrument. So, this type of solution apparently it may come to our mind while selecting whether to use z_2 or z_3 as instrument.

But the point is this type of decision making is actually incorrect. Why this is incorrect? Because if both the instruments are valid that means both z_2 and z_3 , they are correlated with education using one at a time that means the reduced form equations are mis-specified. Here z_2 and z_3 both are correlated with y_2 but I am using only one at a time that means this reduced form equations are mis-specified because reduced form equations are mis-specified.

So, what we need to do then? So, we need to use a linear combination of z_2 and z_3 as instrument. So, we cannot use only one as instrument either z_2 or z_3 rather a linear combination of z_2 and z_3 should be used as instrument and how to get that linear combination? That will be very simple.

(Refer Slide Time: 10:30)



1st step:
2.55 is actually found for estimation

The reduced form equation now becomes

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2$$

$$= \underbrace{y_2^*}_{\text{systematic component}} + \underbrace{v_2}_{\text{non-systematic component}}$$

Second step:
Since OLS applied for v_2 and y_2 are correlated, v_2 OLS becomes

Replaces \hat{y}_2 (or \hat{y}_2^*) in the structural equation and apply OLS. So, our structural equation becomes

$$\hat{y}_1 = \beta_0 + \beta_1 (\hat{y}_2^* + v_2) + \beta_2 z_1 + u_1$$

$$= \beta_0 + \beta_1 \hat{y}_2^* + \beta_2 z_1 + (\beta_1 v_2 + u_1)$$

composite error term $\text{Cov}[(\beta_1 v_2 + u_1), \hat{y}_2^] = 0$*

So, the reduced form equation now becomes

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2$$

$$= y_{2^*} + v_2$$

that means from this equation what we can understand y_2 is basically a sum of two components let me say that this is y_{2^*} and v_2 .

where this y_{2^*} is the systematic component and this v_2 is the non-systematic component. The systematic component is actually not correlated with the error term rather this non-systematic component is correlated with the error term. It means from this reduced form equation what we actually get is

$$\hat{y}_2 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3$$

this \hat{y}_2 is actually the estimated value of y_2 .

So, that means when you have more than two instruments what you are doing actually, in a first step of our estimation where we are writing the reduced form equation for the endogenous variable y_2 . In the second step, we are doing we are replacing the estimated value of y_2 that means $y_2 = \hat{y}_2$. and apply OLS. Here also I am applying OLS.

So, if you replace that then our structural equation becomes

$$y_1 = \beta_0 + \beta_1 (\hat{y}_2 + v_2) + \beta_2 z_1 + u_1$$

$$y_1 = \beta_0 + \beta_1 \hat{y}_2 + \beta_2 z_1 + (\beta_1 v_2 + u_1)$$

Here, in the structural equation the error term is now a composite error term that we have discussed.

Such that $Cov[(\beta_1 v_2 + u_1), \hat{y}_2] = 0$. So, there is no correlation between the estimated value and this composite error term and expectation of this error term is also zero. So, that means if we look at closely what we are doing here since y_2 in the original structural equation y_2 was endogenous we are trying to get a predicted value of the y_2 which is \hat{y}_2 to solve the endogeneity problem.

And how we are doing it? We are doing it in two step. In the first step we are writing the reduced form equation for y_2 and then we are applying OLS there and get the predicted value. And in the second step we are replacing that predicted value from the reduced form equation. So, that means OLS is applied two times, in the first step as well as in the second step. Since the OLS is applied 2 times the IV technique becomes 2 SLS two stage least square.

when we have one instrument then IV and 2 SLS are same. That means we can say that 2 SLS is actually a special IV estimation. This is also IV only because the idea is same. Why this is also IV? Because we are trying to find out the instruments for the endogenous variable and that was there in the first case also when we have one instrument.

Here also we are trying to find out the instrument. But only thing is that here we are trying to find out the instrument applying OLS in two stages and then inserting the estimated value of the endogenous variable in the structural equation. That is why I say that 2 SLS is different from IV. When you have two instruments for one instrument case both are actually same that is what I wanted to mention. Now we will discuss one more case.

(Refer Slide Time: 21:16)



Identification condn
 $H_0: \pi_2 = \pi_3 = 0$
 is also known as order condn.
 - We need a sufficient rank condn.
 known as IVC condition.

IV estimation for more than one endogenous variable
 $y_1 = \beta_0 + \beta_1 z_2 + \beta_2 z_3 + \beta_3 y_1 + u_1$ → structural equation
 $Cov(y_1, u_1) \neq 0, Cov(y_2, u_1) = 0$
 $Cov(y_3, u_1) = 0$
 Let's assume we found z_2, z_3 such that
 $Cov(z_2, z_2) \neq 0, Cov(z_2, z_3) \neq 0$
 $Cov(z_3, z_2) \neq 0, Cov(z_3, z_3) \neq 0$
 $Cov(z_2, u_1) = 0, Cov(z_3, u_1) = 0$
 z_2, z_3 are instruments.
 Reduced form eqn for $z_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2 = 0$
 $z_3 = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3 + v_3 = 0$
 Identification condn from one reduced form eqn of z_2, z_3
 Test: $\begin{cases} H_0: \pi_2 = \pi_3 = 0 \\ H_1: \text{at least one among } \pi_2, \pi_3 \text{ is not equal to zero.} \end{cases}$

So far in our model we assume there is only one endogenous variable. But it may so happen that it in your model there are actually two three or more number of endogenous variable, what to do then? So, we will now discuss a situation, IV estimation for more than one indigenous variable.

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

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

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

$Cov(z_1, u_1) = 0$

Let us now assume we have z_2 and z_3 such that

$Cov(y_2, z_2) \neq 0$; $Cov(y_2, z_3) \neq 0$

$Cov(y_3, z_2) \neq 0$; $Cov(y_3, z_3) \neq 0$

both of them are correlated. But,

$Cov(z_2, u_1) = 0$

$Cov(z_3, u_1) = 0$

these are the condition. So, z_2 and z_3 are instruments. What would be the estimation technique? Here since we have two endogenous variable, we have to fit two reduce form equation. This is our structural equation.

So, the reduced form equation for y_2 would be

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2 = 0$$

and reduced from equation for y_3 would be

$$y_3 = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3 + \epsilon_2 = 0$$

Here we will get two identification conditions.

Identification condition let us say this is from the reduced from equation of y_2 is null hypothesis

$$H_0: \pi_2 = \pi_3 = 0$$

against the alternative

H_1 : At least one among π_2 and π_3 is not equals to zero by F test. What is the test you have to do?

We have to test this by the standard F test.

So, that means we have to check the joint significance of z_2 and z_3 .

Similarly reduced form equation for y_3 would be so for y_3 what would be the identification condition?

The identification condition is

$$H_0: \gamma_2 = \gamma_3 = 0$$

But these identification condition which is also known as order condition is necessary but not sufficient. It is necessary but not sufficient. Why this is not sufficient? It may so happen that from this reduced form of equation y_2 let us say π_2 is significant because what I need only one among them, so that means z_2 is significant.

Similarly, from the reduced from equation of y_3 also we found that z_2 is significant. So, that means while z_2 is actually significant in both z_3 is not significant that is why even though this is a necessary condition, this is not sufficient. What we need? We need a sufficient condition for identification known as rank condition. What is rank condition? That we will discussed in the context of simultaneous equation model. But in this context at least we understand that when we have two endogenous variable we need two valid instruments. So, checking that condition is very simple. We can simply count y_2 , y_3 and two excluded variables z_2 and z_3 . So, we will say that our order condition of identification is satisfied. But that is not sufficient because it may so happen that they are excluded.

But only one among those excluded variable is actually significant. That means we could identify only one instrument while we knew it actually for two endogenous variables. In other words if we have two endogenous variables in the model y_2 and y_3 . So, to estimate this model efficiently and to get the unbiased estimates we need two instruments. Let us assume that z_2 and z_3 we have identified.

Order condition is satisfied, two endogenous variables, two instruments. So, that means two exogenous variables excluded from the model. But what we need to check whether both of them are significant or not. Otherwise from the identification condition we are only checking this condition. From this F test we can find out that whether if π_2 and π_3 significant. That means it may so happen that let us say z_2 is significant here, z_2 is significant in the other equation as well.

Even though that means we have specified two equations with two exclusion restrictions only one exclusion restriction is valid. That is why we said while order condition is necessary and very easy to check whether it is satisfied or not; the rank condition is not. What is rank condition? That is actually the sufficient condition for this model when we have more than one endogenous variable that we will discuss in the context of simultaneous equation model.