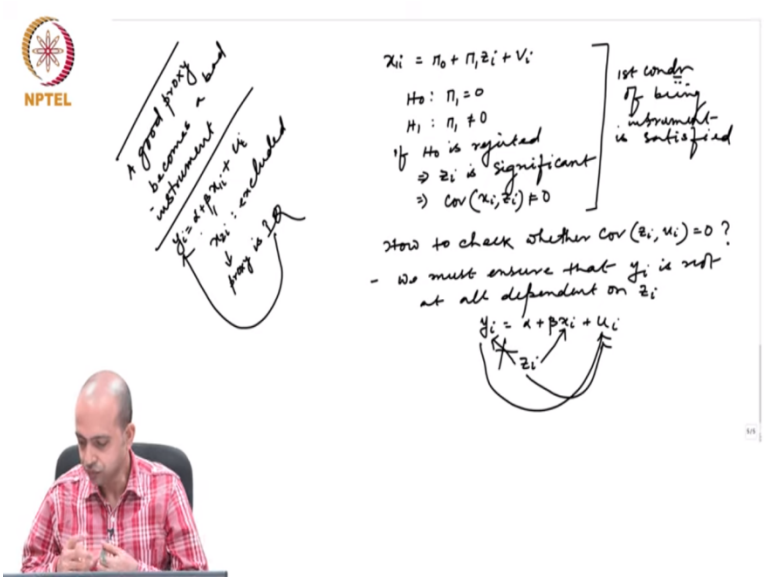
Applied Econometrics
Prof. Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras

Lecture - 02
Instrumental Variable Estimation – Part II

(Refer Slide Time: 00:15)



What about the second condition? How to check the second condition? How to check whether covariance between z_i and μ_i not equals to 0.

$$Cov(z_i, \mu_i) = 0$$

Now this is something which is really difficult. There is no direct mechanism by which we can actually check this condition whether z_i is correlated to the error term because error term is something which is unobservable.

But there is some indirect way, what is the indirect way? We must ensure that y_i is not at all dependent on z_i, if that we can ensure that means

$$y_i = \alpha + \beta_1 x_{1i} + \mu_i$$

and we are thinking z_i as an instrument. So, that means z_i is highly correlated with the endogenous variable but this correlation should not be there, z_i should not have any impact of y_i because if zz_i has some impact on y_i, obviously z_i will qualify to be included in the model itself.

And the moment we include this model this variable in the model we can no longer ensure that z_i and u_i they are actually uncorrelated. At the same time u_i should not capture any variable which is actually correlated with the z_i. So, that means what we are doing indirectly we are saying y_i is related tou_i because u_i captures all those variables which has impact on y_i. So, the moment I say that z_i is uncorrelated with y_i.

That means we are saying z_i is actually uncorrelated with u_i as well, z_i is not correlated with y_i ensures that z_i is not correlated with the u_i. This is how indirectly we can ensure that z_i the second condition is also satisfied but there is no direct mechanism to test this. So, before selecting the instrument we must be very careful and think whether is there any connection between this father's education and somebody's wage.

If at all some connection is there then we cannot include, this is how we have to identify an instrument. Now suppose somebody find a proxy for this IQ variable, a proxy or I would say that a good proxy becomes a bad instrument. Why this is so? Suppose

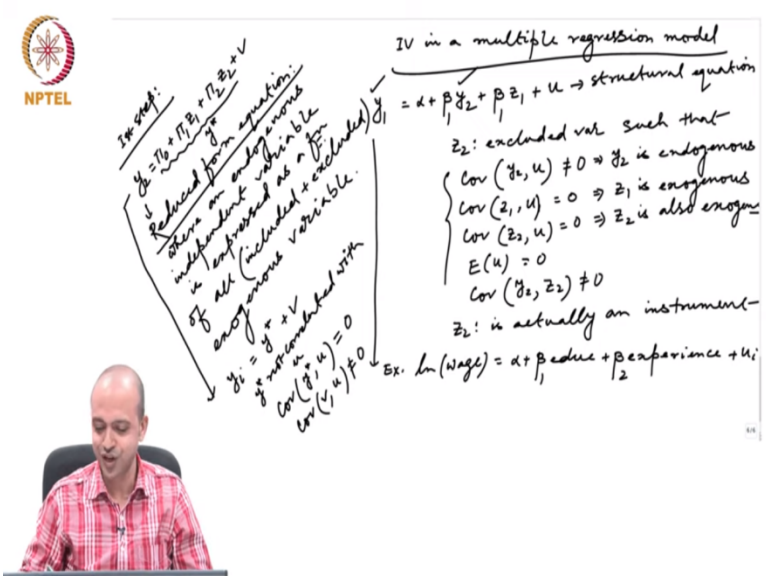
$$y_i = \alpha + \beta_1 x_{1i} + \mu_i$$

this is the model, u_i so this is the model and x_{2i} is actually excluded from this model. What is x_{2i}? x_{2i} is basically ability and we are using a proxy is let us say IQ.

Now the question is whether IQ is a good instrument or a bad instrument? That is the question. Now since IQ is a good instrument good proxy for x_{2i}. That means we can easily include IQ in the model itself. That means what I am saying IQ has a direct connection with y_i itself and what I said if IQ is correlated with y_i that means IQ is also correlated with the error term as well because IQ qualifies to be an explanatory variable in the model.

And the moment the variable is included it cannot be an instrument. So, the variable must be excluded, z_i must be excluded from the model and it should be exogenous because z_i and u_i not equals to zero, the moment is it included we cannot ensure that it is not correlated with the error term. That is why we say that a good proxy makes a case for a bad instrument that we need to understand.

(Refer Slide Time: 07:04)



So, if we extend this model, let us discuss IV in a multiple regression model. Let us assume that this is our model

$$y_1 = \alpha + \beta_1 y_{2i} + \beta_1 z_{1i} + \mu_i$$

this is the model and z_2 is excluded variable such that

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

implies y 2 is endogenous.

$$Cov(z_{1i}, u_i) = 0$$

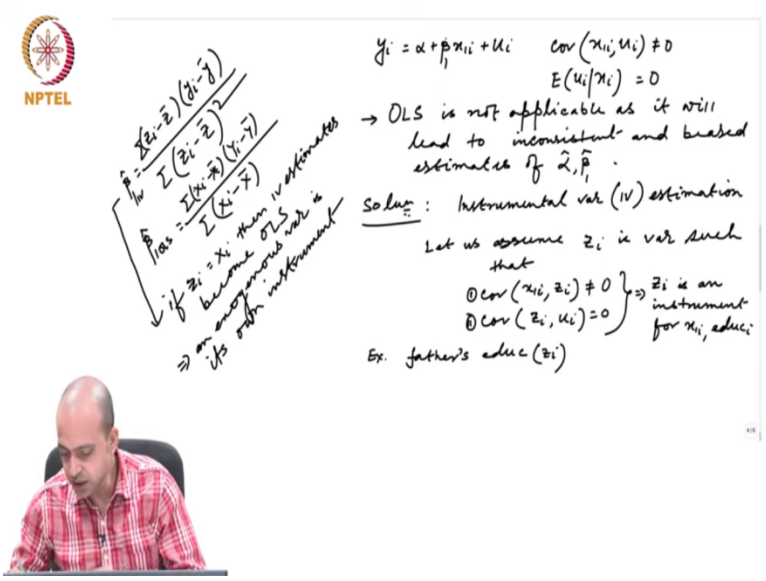
implies z_1 is exogenous.

$$Cov(z_{2i}, u_i) = 0$$

these are the assumptions to satisfy. So, if that is the case when y 2 becomes and covariance between y 2 and z 2 is actually not equals to 0, so this implies z 2 is actually an instrument.

Before we talk about these two variable models in this context when z_i.

(Refer Slide Time: 11:01)



This context we said that z_i is an instrument for x_{1i}. How to estimate β1 hat?

$$\hat{\beta}_{1IV} = \frac{\sum(z_i - \bar{z_\iota})(y_i - \overline{y_\iota})}{\sum(z_i - \bar{z_\iota})^2}$$

this is the β 1 hat this is IV. So, the IV estimates of β 1 hat is given by z i - z bar y i - y bar divided by z i - z bar whole square, that means our originally β 1 hat OLS was if you look

$$\hat{\beta}_{1OLS} = \frac{\sum(x_i - \bar{x_\iota})(y_i - \overline{y_\iota})}{\sum(x_i - \bar{x_\iota})^2}$$

So, if you compare $\hat{\beta}_{1OLS}$ and $\hat{\beta}_{1IV}$ you can easily understand we are just replacing x by z then we are getting the formula of $\hat{\beta}_{1IV}$. Now from this formula we can easily understand if z i = x i then IV estimates becomes OLS which implies an exogenous variable is its own instrument, that is the point I wanted to make.

So, if we have data on z and y, we can easily estimate this $\hat{\beta}_1$ what we are interested in coefficient please keep in mind. We are actually interested in coefficient of x_1 I that means in this which function we are interested in estimating returns to education and that returns to education we are actually estimating indirectly by using an instrument for x_{1i}. How to estimate with the data and software? That I will discuss in detail in a later part using the statistical software stata.

For the timing just keep in mind we are interested in $\widehat{\beta}_1$ but we are not able to use OLS because x_{1i} is correlated with the error term. That is why you are using this z_i, z_i is an instrument we are using and this is the formula simply replacing x by z in the OLS formula we will get the IV estimates. That means when z_i becomes x_i, IV estimates converge to OLS and that implies an exogenous variable is its own instrument that is the point.

With this now we are coming to the IV concept in a multiple regression model. Let us say

$$Ln(wage)_i = \alpha + \beta_1 education_{1i} + \beta_2 experience_{1i} + \mu_i$$

we not only have one endogenous variable in the model we are also having one another variable which is exogenous experience in the model.

How will you estimate this model? That means since y2 is correlated with u the error term we need to use an instrument which is z_2 here. So, how to estimate this model, the first step of estimating this model is we will write an equation like this

$$y_{2i} = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_i$$

This equation it has a specific name in the context of instrumental variable estimation and simultaneous equation model. This equation is called reduced form equation and the equation of our interest is called structural equation. Why this is called structural equation? Because this is the equation which is derived from an economic theory. What is the economic theory? Economic theory of wage determination, economic theory of labour market says that wage is actually a function of education as well as experience which is coming from a labour market theory.

This equation explains the structure of a model structure of a theory that is the name structural equation. But this reduced form equation is not coming from a theory rather this equation is formulated by following this definition. What is the definition of a reduced form equation? Reduced form equation is where an endogenous variable or I would say endogenous independent variable is expressed as a function of all exogenous and here I am saying all included plus excluded exogenous variables.

So, while specifying the reduced from equation we must keep one thing in mind that it is a function of an endogenous independent variable, please be very careful. Here we have in this equation we

have two endogenous variables y_1 is indigenous variable, y_2 is endogenous variable. We are not talking about y_1 because y_1 is the dependent endogenous variable, dependent variable is always endogenous.

So, we do not have to bother anything about the dependent variable. Our entire discussion is focused only the independent endogenous variable y_2. So, while formulating the reduced form equation, what is the definition? y_2 should be a function of all exogenous variable included plus excluded. So, when we write the reduced from equation firstly you specify your included exogenous variable which is z_1.

Then you come back and include the excluded exogenous variable z_2 also in this reduced form equation. first step should be to included exogenous variable which is already there in the model, otherwise you may forget in the process. That means from these; what we can say that

$$y_{2i} = y* = v$$

What is y star? This component $\pi_0 + \pi_1 z_1 + \pi_2 z_2$ we call that as y star.

This is the component which is y* is a systematic component and this v is non-systematic component.

$$Cov(y^*, \mu_i) = 0$$

That means I can say covariance between y star and u actually equals to 0. But this v this error term and u they are correlated and that is the reason y_2 is correlated with the error term.

$$Cov(v, \mu_i) \neq 0$$

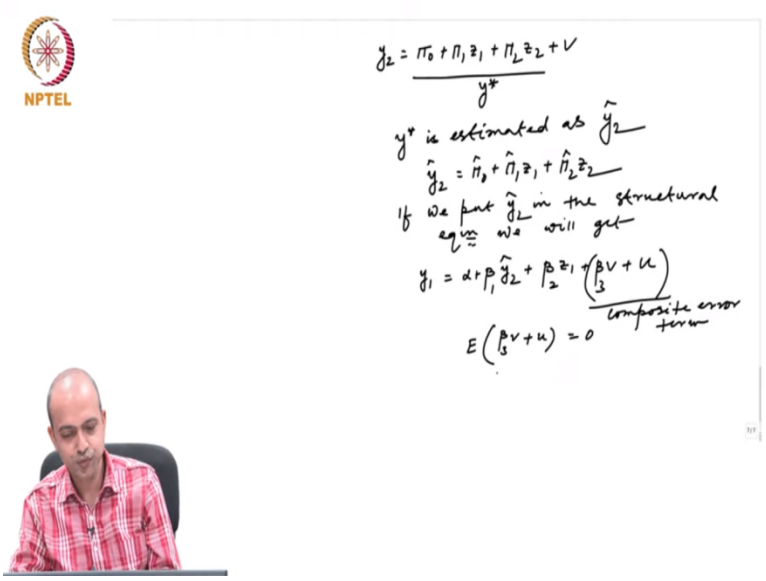
In my original model y_2 is actually correlated with the error term that is why y_2 is called an endogenous variable. So, to solve that endogeneity problem what we are trying to do, we are formulating a reduced form equation and then we are saying y_2 is decomposed into two components y* which is $\pi_0 + \pi_1 z_1 + \pi_2 z_2$.

The first component of the reduced form one and the error term while the first component y star is not at all correlated with the error term. It is not at all correlated with v also we assume this v is

correlated with the error term and that is why y_2 is correlated with the error term. So, the channel through which the endogeneity or relationship between y_2 and u runs is actually this v. So, what we need to do then?

We need to get an estimated value of this y* and that estimated value since that is not correlated with error term, we will put again in the structural form equation.

(Refer Slide Time: 23:28)



In the second step, y* is actually an estimated value of this which is let us say y* is estimated as let us say

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

y hat or this is let us say y 2 hat is estimated

$$\hat{y}_2 = \widehat{\pi_0} + \widehat{\pi_1} z_1 + \widehat{\pi_2} z_2 + v_i$$

Then if we put the model in the original equation, so here our original equation was this. So, in place of y_2 now I will put $\hat{y}_2$. So, what will happen then in the original model if I put so our original model was if we put $\hat{y}_2$in the structural equation, we will get we will get

$$y_1 = \alpha + \beta_1 \widehat{y_2} + \beta_2 z_1 + \beta_3 v + u$$

that means the structure in the new structural form equation the error term is this which was earlier u, now it is becoming $\beta_3 v$ this is the composite error term.

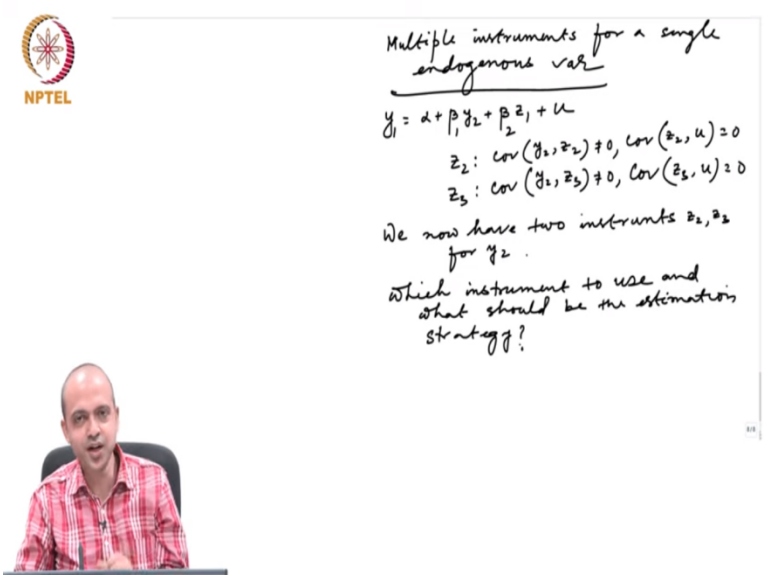
So,

$$E(\beta_3 v + u) = 0$$

$$Cov(\hat{y}_2, \beta_3 v)=0$$

So, in this equation if we plug in then this earlier the problem what y_2 hat was correlated with the error term it is no longer correlated with this because we have solved this problem by getting the estimated value. This process, this is the actual estimation process of the instrumental variable estimation.

(Refer Slide Time: 27:44)



Now suppose we will consider another model where we have multiple instrument for a single endogenous variable. So, let us say that this is our model

$$y_1 = \alpha + \beta_1 y_2 + \beta_2 z_1 + u$$

but we have two excluded variable which are z_2 and z_3 such that

$Cov(y_2, z_2) \neq 0$; $Cov(z_2, u) = 0$

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

So, we will say that we have now two instruments z_2 and z_3 for the single endogenous variable y_2, if that is the case when we have multiple instruments which instrument to use and what should be the estimation strategy, that we will discuss in our next class. So, far we assumed that we have one endogenous variable for which we could identify only one exogenous instrument.

But when you have multiple instruments, it is possible that sometimes there are two instruments two exogenous variable which are highly correlated with the endogenous variable in a model. If that is the case, should we use one instrument or both the instrument and what should be the estimation strategy that we will discuss in our next class, thank you.