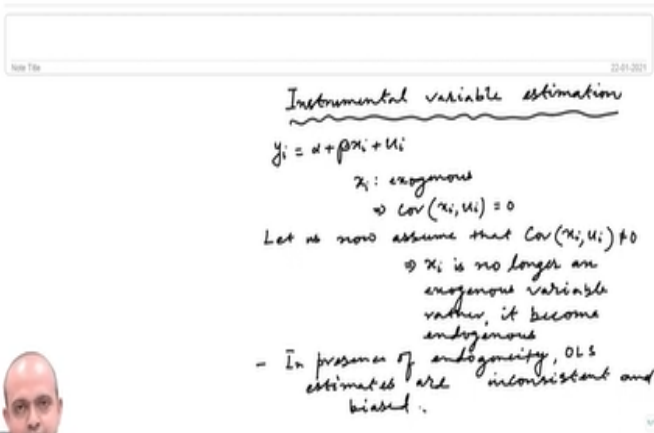



Applied Econometrics
Prof. Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras

Lecture - 01
Instrumental Variable Estimation – Part I

Welcome, to the class of applied econometrics and we will start our discussion with a very important topic called instrumental variable estimation.

(Refer Slide Time: 00:38)



Instrumental variable estimation

$$y_i = \alpha + \beta x_i + u_i$$

x_i : exogenous
 $\Rightarrow \text{Cov}(x_i, u_i) = 0$

Let us now assume that $\text{Cov}(x_i, u_i) \neq 0$
 $\Rightarrow x_i$ is no longer an exogenous variable rather, it becomes endogenous

- In presence of endogeneity, OLS estimates are inconsistent and biased.

Before we start discussing about this topic, let us try to understand why is it required. As I said it is an important topic, we need to understand why instrumental variable estimation technique is so important in econometric analysis. If you recall that, let us say that this is the model we are trying to estimate

$$y_i = \alpha + \beta x_i + \mu_i$$

In ordinary least square (OLS) technique we assume ten assumptions, while discussing about classical linear regression model.

While applying ordinary least square technique to estimate $\hat{\alpha}$ and $\hat{\beta}$ and we said that if those assumptions are satisfied, then only our estimator's $\hat{\alpha}$ and $\hat{\beta}$ they will exhibit the desirable properties. One such assumption was what we assume that x_i is exogenous. What does it mean? This implies covariance between x_i and μ_i is basically 0.

$$\text{Cov}(x_i, \mu_i) = 0$$

There is no covariance or correlation between the explanatory variable and the error term.

$$\text{Cov}(x_i, \mu_i) \neq 0$$

Now suppose, we are in a situation wherein, let us now assume that covariance between x_i and μ_i that is actually not equals to 0.

x_i and μ_i they are correlated. That basically implies, that x_i is no longer an exogenous variable rather it becomes endogenous. When x_i and μ_i they are correlated that means x_i becomes endogenous then in presence of endogeneity, OLS estimates are inconsistent and biased.

It means OLS is no longer applicable in a model, wherein x_i and μ_i they are actually correlated, this is called the endogeneity problem.

(Refer Slide Time: 04:52)

The slide contains handwritten notes under the heading "Reasons for endogeneity problem".

- ① There is simultaneity y_i & x_i
 $y_i = \alpha + \beta x_i + u_i$
 A circular arrow indicates a simultaneous relationship between y_i and x_i .
- ② Omitted variable:
 $y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$
 but we have no data on x_{i2} ,
 as a result of which we estimate
 $y_i = \alpha + \beta_1 x_{i1} + u_i(x_{i2})$
 $\left. \begin{aligned} \text{Cov}(u_i, x_{i2}) &\neq 0 \\ \Rightarrow \text{Cov}(u_i, x_{i1}) &\neq 0 \end{aligned} \right\}$

In the bottom left corner, there is a video inset showing a man in a red and white checkered shirt speaking.

There are three different reasons for endogeneity problem. Number 1, there is simultaneity between y_i and x_i . So, that means when I am estimating a model,

$$y_i = \alpha + \beta x_i + \mu_i$$

Generally what we assume the direction of causation runs from x_i to y_i . But in case where y_i also causes x_i then there is a simultaneous relationship between y_i and x_i .

And as a result of which, x_i becomes endogenous. For example, let us say that we are estimating a demand function where demand is a function of price and price is also a function of the demand. So, price and demand they are simultaneously determined. So, if at all we are interested in estimating a demand function, we need to estimate that model using a

simultaneous equation framework because in that model, price is no longer an exogenous variable. There is simultaneity between y_i and x_i that is one reason for endogeneity.

Second one, is called omitted variable. Let us say that we are estimating this model, let us say this is the model,

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \mu_i$$

This is the complete model. But we have no data on x_{2i} as a result of which, we estimate

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

Why we have no data on x_{2i} ? May be x_{2i} cannot be observable. it means x_{2i} is an unobserved variable. There is no proxy for x_{2i} so on and so forth. As a result of which, we are not able to include x_{2i} , even though theoretically x_{2i} also is an important variable determining y_i . So, if we cannot include that variable in the model what where will that variable go? That variable will be here in μ_i only. So, μ_i , the error term, as it captures the impact of important variable which are omitted from the model.

The moment x_{2i} is omitted, it would be there in the error term and if we assume that this x_{1i} and x_{2i} , they are correlated. If we assume,

$$Cov(x_{2i}, x_{1i}) \neq 0$$

that in turn will indicate that covariance between

$$Cov(x_{1i}, \mu_i) \neq 0$$

that means x_{1i} will become an endogenous variable because of this omitted variable problem. So, we have to keep in mind what is omitted variable and how this omitted variable may lead to endogeneity problem.

(Refer Slide Time: 10:34)



③ Measurement error in y_i & x_i

Omitted Variable:

$$\ln(\text{wage})_i = \alpha + \beta_1 \text{educ}_i + \beta_2 \text{ability}_i + u_i$$

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

$$\ln(\text{wage})_i = \alpha + \beta_1 \text{educ}_i + u_i \text{ (ability)}$$

$\Rightarrow \text{Cov}(\text{educ}, u_i) \neq 0$

\Rightarrow educ is actually an endogenous variable

Third reason for endogeneity is called measurement error in y_i and x_i . It means there is no guarantee that all the time we will be able to measure our y_i and x_i correctly. And if there is some measurement error, that means we are partially captured the impact of a particular variable. Then that variable will also become endogenous because the other part will be there in the error term.

So, this measurement or error in variable y_i and x_i may also lead to endogeneity. Now out of these three reasons, in this course we will discuss only these two while discussing about endogeneity problem and we will start with the omitted variable problems. Simultaneity bias or when endogeneity arises from simultaneous relationship that we will discuss in our next topic when we will be discussing about simultaneous equation model.

So, let us now talk about the omitted variable problem in econometric model, omitted variable. So, we will start with an example. Let us say that we are estimating an wage function, where wage is a function of

$$\ln(\text{wage})_i = \alpha + \beta_1 \text{education}_{1i} + \beta_2 \text{ability}_{2i} + \mu_i$$

So that means we can think of

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \mu_i$$

economist they say that wage is determined not only by the education of the individual but also the individual's inert ability or skill.

Now ability or skill is a factor which is not directly observable. Just by looking at an individual we cannot determine what is the individual's ability or skill factor. So, that means ability is a variable, which is highly important in determining wage. But it is not directly observable and there is no direct proxy also to measure the ability variable. As a result of which what the model in turn we will estimate is,

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

Some people they say that, ability can be measured by IQ. But IQ is also something which is not directly observable. How will you measure IQ of an individual? Some people they say that we can measure IQ of an individual just by asking some questions and then we will say that this individual's IQ is twice higher than Einstein, or thrice lower than the Einstein, there is some way.

But that is also not a proper way to measure somebody's IQ just by asking some question. So, that means this ability factor is not directly observable, and we cannot measure this variable. As a result of which, this ability will always go into the error term and if we believe that education and ability they are correlated, positive or negative that I am not commenting here. Sometimes we think that highly educated people would be with high amount of ability.

Sometimes education makes people enable also. So, I am not commenting on positive or negative, but at least education and ability they have some kind of correlation which we cannot deny. So, in education and ability they are correlated we cannot include the ability factor in the model that is why μ_i will capture the factor ability. And as a result of which, this correlation between education and ability will lead to covariance between education and μ_i that is not equals to 0 and that implies education is actually an endogenous variable.

Education is an endogenous variable. So, when education is endogenous variable, then the OLS is no longer applicable because it will lead to inconsistency and biased estimates of β_1 and α hat.

(Refer Slide Time: 17:00)



$y_i = \alpha + \beta_1 x_{1i} + u_i$ $Cov(x_{1i}, u_i) \neq 0$
 $E(u_i | x_{1i}) = 0$
 \rightarrow OLS is not applicable as it will lead to inconsistent and biased estimates of $\hat{\alpha}, \hat{\beta}_1$.
Solution: Instrumental var (IV) estimation
 Let us assume z_i is var such that
 $\left. \begin{array}{l} \textcircled{1} Cov(x_{1i}, z_i) \neq 0 \\ \textcircled{2} Cov(z_i, u_i) = 0 \end{array} \right\} z_i \text{ is an instrument for } x_{1i}, \text{ educ.}$
 Ex. father's educ (z_i)



Now what is the solution? So, that means this is our model to start with a simple model

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

where, $Cov(x_{1i}, \mu_i) \neq 0$

$$E(u_i | x_i) = 0$$

Zero mean assumption is still valid. If this is the case, then OLS is not applicable, as it will lead to inconsistency or inconsistent and biased estimates of $\hat{\alpha}$ and $\hat{\beta}$.

So, what is the solution? The solution to avoid this inconsistency and biased estimates is instrumental variable estimates, or IV estimation in sort. What is this technique? Let us assume, there is a variable called z_i , such that

where, $Cov(x_{1i}, z_i) \neq 0$

z_i is highly correlated with the endogenous variable x_{1i} . But,

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

z_i is not correlated with the error term.

So, these are the two important conditions or properties that z_i is having. If we find out such a variable z_i , which is highly correlated with the endogenous variable in this context let us say education, but not at all correlated with the error term, then we will say that z_i is basically instrument for x_{1i} which is education actually. So, z_i is an instrument for the endogenous variable x_{1i} if these two conditions are satisfied.

If these two conditions are not satisfied, then we cannot call z_i as an instrument. Now, the question is when we are estimating wage function, where wage is a function of education can

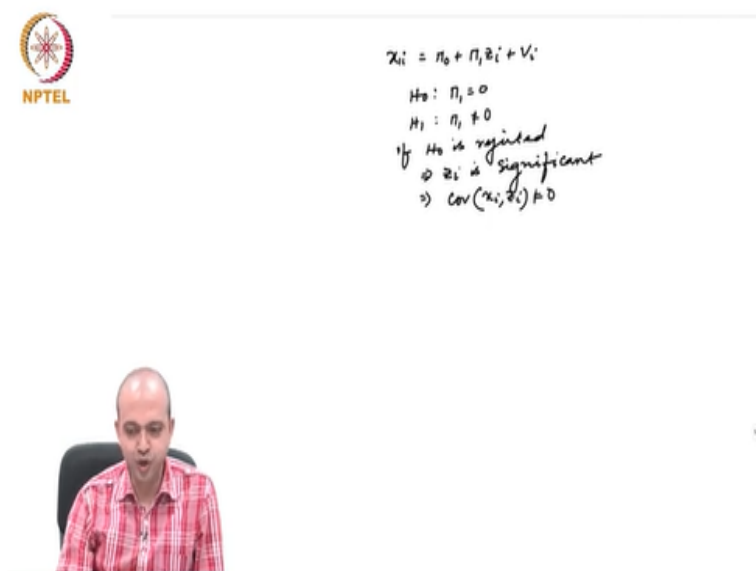
you think of a variable which will be highly correlated with somebody's education but not at all correlated with that individual's wage. What could be the example? Just take a minute and think.

Wage is a function of education and education is endogenous, because the ability factor is excluded from the model, omitted from the model. And we need to identify an instrument for education. And that instrument is highly correlated with education, but not at all correlated with the wage factor. What could be that factor, that variable? Now an example could be, father's education.

It is quite obvious that somebody's education will be highly correlated with that individual's father's education. If father is educated that father will try to make more attention to his kids education as well. But when the kid entered into the job market, then obviously the employer will bother less about that individual's father's education while determining his wage which quite obvious.

It means, we can think of father's education as an instrument for that individual's education. This is an example for z_i , father's education. And then how will you estimate the model? How will you check whether this condition these two conditions are satisfied or not?

(Refer Slide Time: 23:00)



The slide contains the following text:

$x_{1i} = \pi_0 + \pi_1 z_i + v_i$
 $H_0: \pi_1 = 0$
 $H_1: \pi_1 \neq 0$
If H_0 is rejected
⇒ z_i is significant
⇒ $\text{cov}(x_i, z_i) \neq 0$

The NPTEL logo is visible in the top left corner of the slide.

When we assume z_i is an instrument for that x_{1i} , so that means, we can easily run a regression

$$x_{1i} = \pi_0 + \pi_1 z_i + v_i$$

and what is our null hypothesis here? Null is $\pi_1 = 0$ and alternative is actually $\pi_1 \neq 0$. So, if the null is rejected which implies z_i is actually significant. This is all we know from what basic econometrics.

So, that means z_i is actually significant. And what does it mean? z_i and x_{1i} they are actually correlated, not equals to 0. So, this way by running a simple regression of x_{1i} on z_i then we can identify whether the z_i is actually significant. That means whether z_i can be considered as an instrument for x_{1i} . The first condition is satisfied.