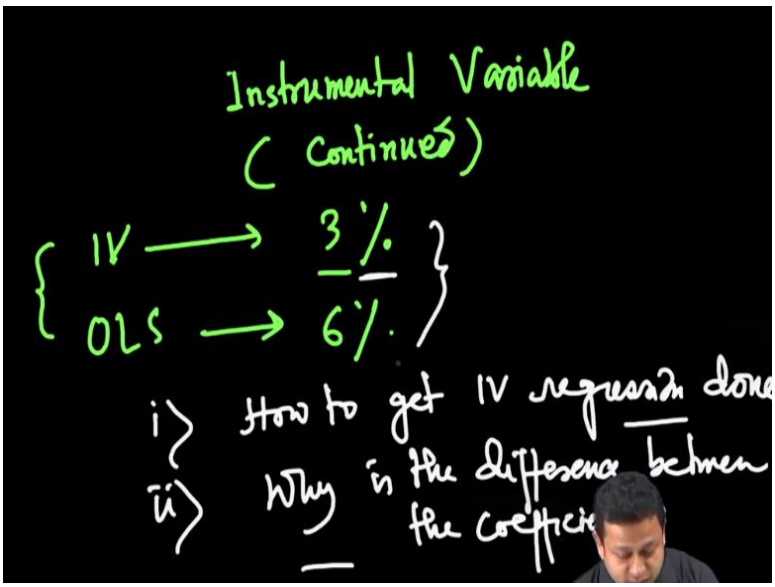
Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology – Kharagpur

Lecture – 99
Instrumental Variable Continued

(Refer Slide Time: 01:06)



Hello and welcome back to the lecture on applied econometrics. We have been talking on instrumental variable and we said that instrumental variable is a concept which is very dear to the economists. And we have seen with an example that how a national experiment, an intervention that is not designed by the researcher can actually be so impactful in terms of understanding question that you have been asking all throughout what is the true impact of education on wage.

And that we have talked about in the context of United Kingdom in 1947, they brought in legislation where they actually increased the minimum age of schooling that mandatory age of schooling that one has to stay in the school and we sort of estimated the effect of schooling on wage and we found that after the use of IV the estimate to be 3% that is what we have calculated previously.

Whereas when they actually did a normal OLS, ordinary OLS, so what they found is that ordinary OLS says that the impact of schooling on the wage is 6%. Now, the fundamental question is how this difference has come to play? So, in fact, there will be two questions; one

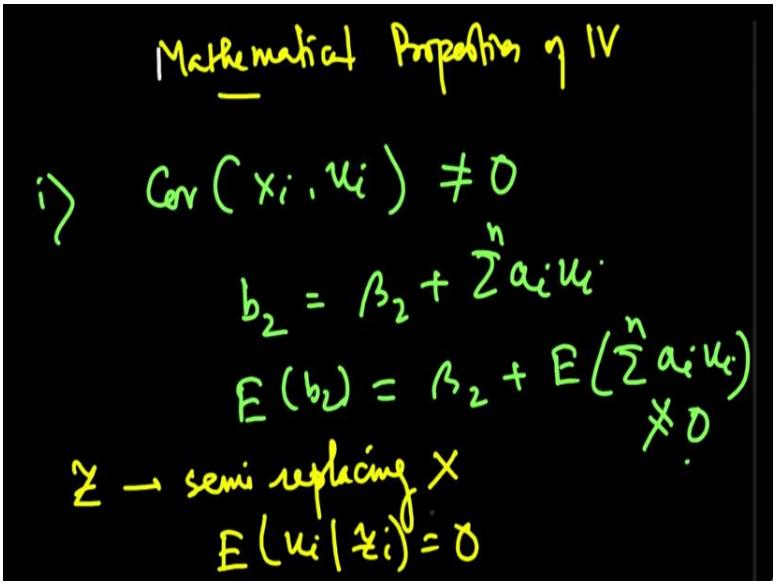
is this why this difference happened and second how really we have estimated this? We kind of showed you the values and in terms of the graphical representation, what happened in terms of the average years of schooling, the school leaving age and so forth.

But how really if we have to conduct this; if we have to use instrumental variable how really we do that? So that is the question we want to address. How to get IV regression done? How we will do that? And second what I asked you is why is the difference between the coefficients? Now the how question, actually how question is relatively easy because we as an econometrician will learn how to run a software and how to provide the inputs to get some output.

And that how part is rather easy, but as an econometrician where our role will lie is in terms of understanding the concept of instrumental variable and in terms of our ability to interpret the result that we get after we feed the data in our software. So, software we will definitely learn but will also learn to explain the result, so that is what we are going to do in the next lectures. To answer the second part, essentially you need to have knowledge in economics domain.

You need to know what exactly influences the wage and what are the theories saying? What previous researches are saying? So from there, you will be able to explain the differences between two coefficients. So, we will see that. Now, in this lecture as I mentioned previously what I am going to do is to actually the as an econometrician as we need to understand the properties of IV, so in this lecture we are basically going to dedicate to understand the mathematical properties of IV. So let me write it down, let me use a different colour.

(Refer Slide Time: 03:50)

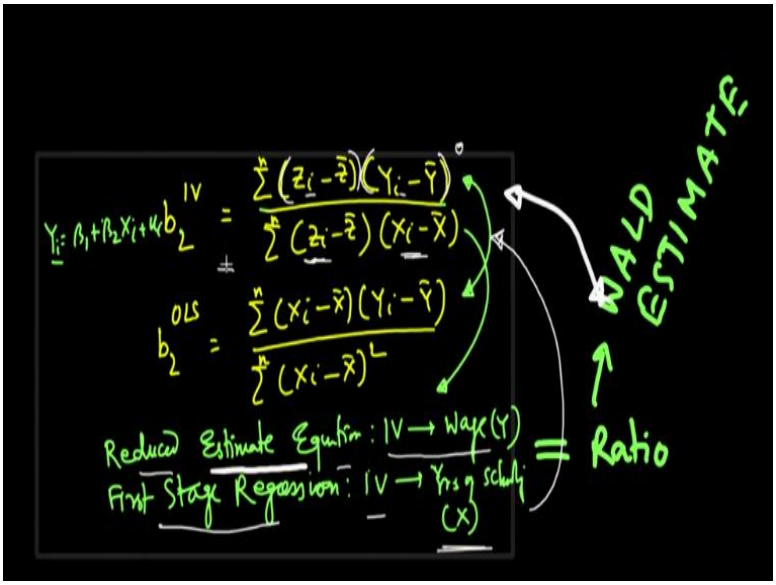


Mathematical properties of IV: Now, remember in one of our previous lectures, we spoke in detail about the stochastic regressor. And one problem that we mentioned when you use stochastic regressor is that the X i's could be actually related with the u i's and that is essentially a big problem because if my X i's and u i's are related or essentially I can write if in the covariance of X i and u i are not equal to 0.

So what I will have is that my model, in my regression model the coefficients that I will obtain they are going to be biased, and I proved it using this formula you already by now remember this formula summation ai u i over all n. So if my X i and u i are not independent, so what is going to happen is this. If I take an expectation of this, so I can have beta 2, but then, this term is not going to be 0.

This term is not going to be equal to 0 and when this term is not going to be equal to 0 that is essentially creating the bias term. So similarly, we know that the IV, let us say my Z is actually semi replacing X. And I claim that if my Z is not independent of the u, so then I will face a similar problem, I will have the problem where my regression coefficient is going to be biased. So, I essentially need to have this condition satisfied expectation of all the u i given Z i = 0 and I will actually show you why.

(Refer Slide Time: 05:45)



But before that, let me actually show you really how we can actually express mathematically the IV coefficient, instrumental variable coefficients. So the way we write it is this summation Z i – Z bar into Y i – Y bar by summation Z i – Z bar and X i – X bar. And let us say my regression equation is Y = beta 1 + beta 2 X i + ui. Now in this regression, so you can actually see the way we can express beta 2 IV is somewhat similar to the way we express beta 2 OLS.

So, we know beta 2 OLS = summation X i – X bar into Y i – Y bar divided by summation X i – X bar square over all n. Now, there are kind of we can see the similarities in the numerator and there are some similarities in the denominator as well. Now where are the difference is coming from? Now, remember here the IV is actually same replacing the X variable, so we are not really running the regression with Y on Z.

We are running the regression with Y on X, but then there are some way we are also incorporating Z into the equation. Now, what are you seeing here if you remember in the previous lecture when we talked about the reduced estimate equation, we essentially saw the impact of IV on wage on Y, alright. Whereas in first stage regression, we see the impact of IV on let us say years of schooling or X.

Now, when I try to understand the impact of actually the true estimate, true impact of let us say schooling on wage, so what I do is I actually take a ratio of these two and this ratio is also called WALD estimate. Now from this, if you try to compare this WALD estimate this ratio

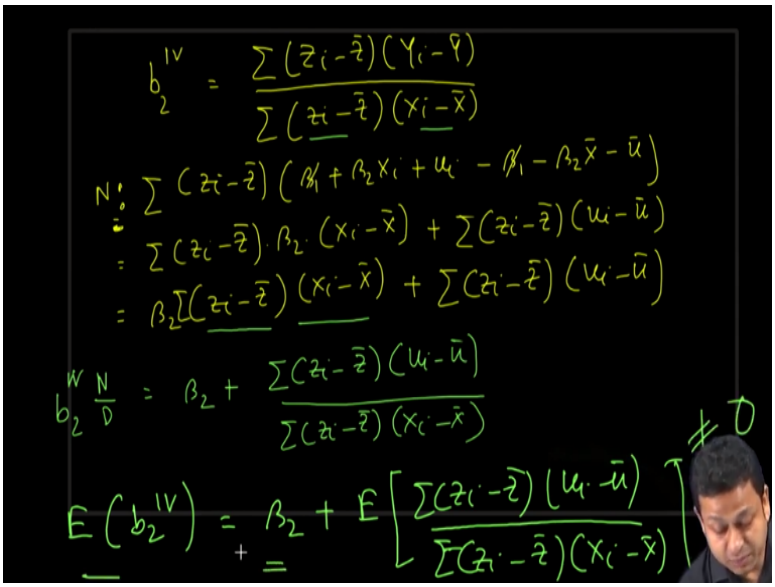
with this, so what we see here? So you see here in the reduced estimate equation what you are doing, you are basically seeing the impact of IV on the wage that is on Y.

So here also you see the numerator what you are doing is you are actually trying to see the strength of relationship between X and Y, so your IV and the Y. So intuitively, it makes sense, it looks like the reduced estimate equation is somewhere reflected in the numerator term. Whereas in the first stage regression, what you are doing is you are trying to see the impact of your IV on the true treatment that is basically your schooling.

Now, here also in the denominator if you just compare what you will see here that the strength of relationship between Z and X, alright. So, then it actually makes sense when I compare beta 2 OLS and beta 2 IV you can actually see since my Z; I need to see the strength of relationship between Z and X as well as strength of relationship between Z and Y, I actually have this mathematical expression for beta 2 IV.

Now, it makes sense from what you seen previously and what you are writing now, but also I need to ensure that my beta 2 IV is actually following the blue properties. It has to be consistent, it has to be unbiased, so do whether it follows it, so that that is something that we have to show now. So let me actually do a little bit of maths around this.

(Refer Slide Time: 10:25)



So let me write down again, let me use a different colour. Beta 2 IV is equal to; I will just rewrite the equation Z i – Z bar Y i – Y bar divided by summation of Z i – Z bar into X i – X bar. Now, if we expand the numerator, so let me explain the numerator, so what I will have

is, let me write down numerator. So, I will have Z i – Z bar and in Y i – Y bar I can write actually from my regression equation beta 1 + beta 2 X i + u i – beta 1 – beta 2 X bar –u bar, alright.

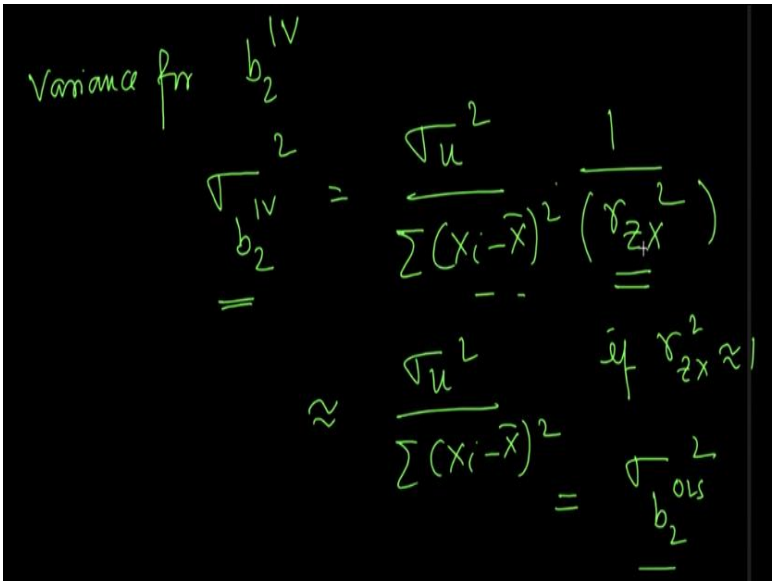
Now, so beta 1, beta 1 they cancel out and I can see that I can take Z i – Z bar into beta 2 into X i – X bar here and then for the other term I can also take Z i – Z bar into u i – u bar. Now, if I take beta 2 out, I will have beta 2 summation Z i – Z bar into X i – X bar + I have as it is Z i – Z bar into u i – u bar. Now, if I now do numerator by denominator, if I bring in the denominator component, let me use a different colour N by D is going to give me, so look at it.

So this term here Z i – Z bar X i – X bar is exactly the same term here. So essentially, the cancel out and what I have is beta 2 and on the other hand for the other term what I have is Z i – Z bar into u i – u bar divided by Z i – Z bar into X i – X bar. Now, look at the star. So this is essentially my b 2 IV. Now, if I take an expected value of beta 2 IV, if I take an expectation on both the sides; so expectation of beta 2 IV and this expectation of beta 2 is going to be beta 2.

And this is going to be expectation of the whole term that is Z i – Z bar into u i – u bar divided by summation of Z i – Z bar into X i – X bar. Now, remember the previous derivation wherever we used this notation of a i and u i, so here also we can actually express Z i and u i in terms of a i and u i. And we can very clearly see if the Z i and u i they are not independent, the star is not going to be is equal to 0.

So, just like in the previous case, just like in case of OLS regression where you had to ensure that X i and u i has to be independent, similarly for IV regression we have to ensure that the Z i and u i are independent. Only then this term will be 0 and only then the expected value of b 2 IV is going to be beta 2 and that is what you want to ensure the unbiasness, so that is the first thing.

(Refer Slide Time: 14:28)



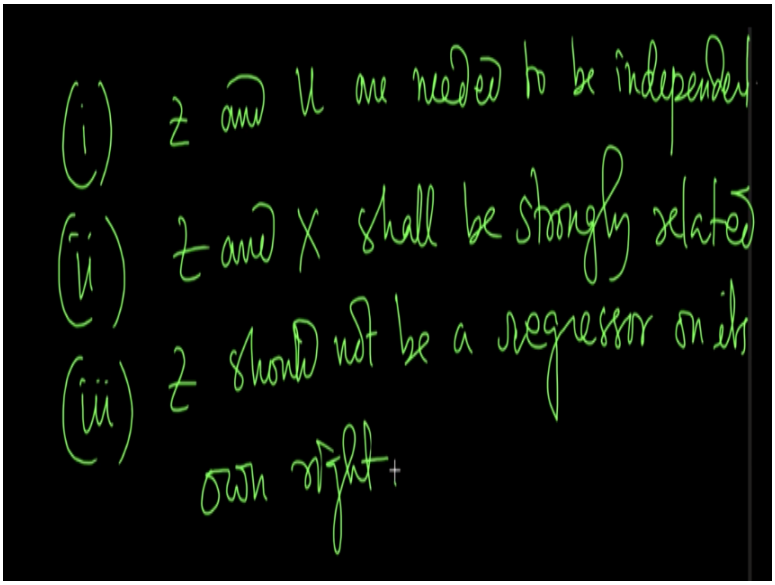
Now, if we also look at the variance part, the variance part for beta 2 IV is actually you can write it sigma b 2 IV square = sigma u square summation of X i – X bar square into 1 by r ZX square. So, r ZX is kind of showing the strength of relationship between Z and X. Now, it makes sense because here the variance is in regression equation, we actually have X whereas we semi replacing X with Z.

So we need to consider the strength of relationship between Z and X. So that is the expression for the variance for the IV regressor. Now, here if the relationship is strong, so then r is going to be close to 1. And if r is going to be close to 1, so it essentially would be let us say if r or r square ZX is close to 1, so then this term is essentially going to be sigma u square by summation X i – X bar square, so which is nothing but essentially sigma beta 2 OLS square.

So that is basically the same variance that we see for my b 2, where my b 2 is actually the correct coefficient that we obtained. So that is why we have to sort of ensure that the strength between X and Z pretty obvious that they are pretty quite strongly related. So, that condition we have to satisfy, alright. So, this we can actually also look at the conditions that we talked about; the conditions of relevance and exclusivity.

So, here you see the IV that we have chosen it has to have some impact on the Y and it has to channel influence through the X variable. The way we mathematically written it the influence should come through the X variable, then only it can be a good IV.

(Refer Slide Time: 17:04)



So, let me write down the result that we get. One is that Z and u are needed to be independent. Second, Z and X shall be strongly related. Now, one more point we have to think and that is if Z should not be regressor, Z should not be an explanatory variable on its own right. So, if Z itself is an explanatory variable, then I will have a simple OLS with Z as a regressor, but this is an IV regression and I want to see the impact of Z through Z.

So that is why we write Z should not be a regressor on its own, right. So, essentially, this sums up the mathematical properties of the IV. And in the next lecture we will actually do a hands-on problem where we will use our software and run two stage least square and we will try to interpret the results that we get and we will try to recall the mathematical properties we just derived here. So, with this we end the lecture on instrumental variable. Thank you.