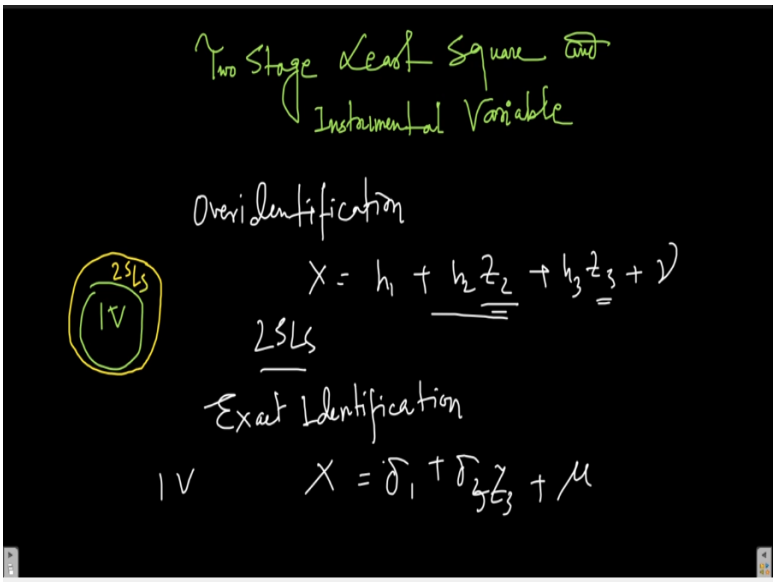
Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology – Kharagpur

Lecture – 106
Two Stage Least Square and Instrumental Variable

Hello and welcome back to the lecture on applied econometry.

(Refer Slide Time: 00:28)



We have been talking about instrumental variable. In the previous lecture we spoke in detail about the identification of the I V regressive and two stage least square. Now, in this lecture we are going to detail this two stage least square and how we use this technique for regression using I V estimator. Now, let us distinguish these two terms. So, when we say two stage least square and I V regression they are basically same.

But two stage least square is something that we say when we have an over identification strategy. So, in case of over identification where we have let us say two our instrumental variable. They are actually determining X variable or endogenous variable. So then in that case let us say X = h 1 + h 2, z 2, h 3, z 3 and let us say I have some error term. Then in that case the strategy that I use the ols studies that I use that is basically or we call it a 2 S L S.

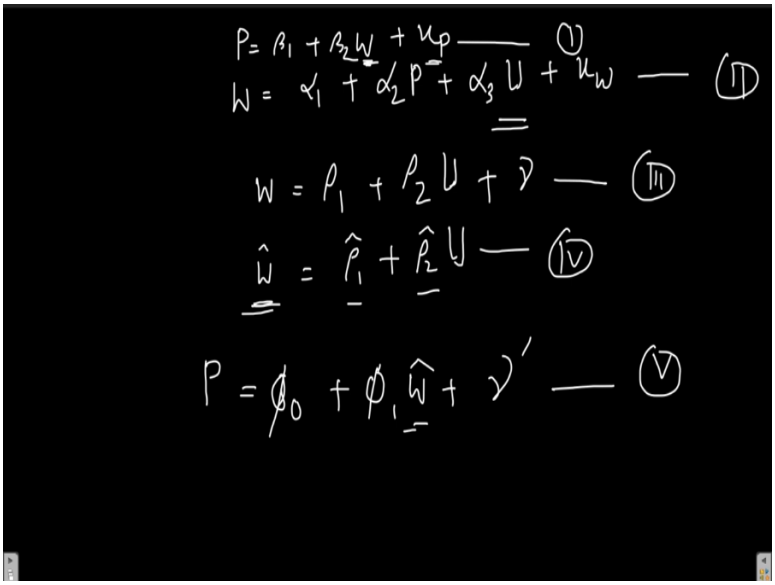
Now, here the z 2, z 3 they all have to have all these properties of I V satisfied whereas for exact identification when I have my number of endogenous variable is equal to my number of instruments. So then we call it a case of I V. But the strategy remains same in the sense that

we have this first stage, we do the ols estimation. And in the second stage the, we use this estimation to do the actual regression equation.

So, here we can write something like maybe delta 1, delta 2. Let us say or delta 3 z 3 and some error time. So, let us say mu. So, in this case this is an I V whereas in over identification we call it 2 S L S. But they are basically same and we can just say that 2 S L S might have a larger ambit whereas I V is a part of this 2 S L S strategy. I V and the outer part is basically 2 S L S. Now, let us write it mathematically how it will look like.

So, let us say what, if we do the first stage least square. So, let us think about our example of wage rate.

(Refer Slide Time: 00:35)



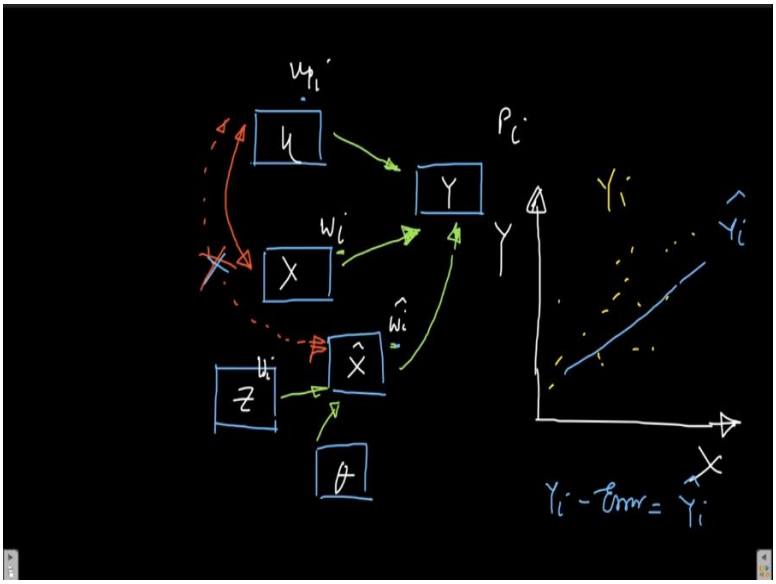
So, wage rate W is equal to we have written alpha 1 + alpha 2 and then we had our P and then we had alpha 3 and then unemployment and then we had some error time. Now, here for I V in the first stage what I am going to do? I am just going to use U as my z. So, I will have my W, I will be writing this equation. So, let us say some row 1 + row 2 U and some other error term. Let us say nu. I will be repeating this all the same terms.

And here what I do is I actually get a W estimated. So, when I estimate W. So, I get a row 1 hat, row 2 hat and U. So, I actually estimate this coefficients row 2, row 1 hat and row 2. Now, when I get these values what I do is I actually use this in my original equation. And there is a price equation I am basically estimating price. So, let us write down the structural equation for price here.

Price is equal to I had beta 1 + beta 2 W + u p. Now, here I am supposed to replace W. So, let us number the equations also 2, 3 and this is a 4. Now, I will use equation 4, W hat and replace the W here. So, what I am going to do is now, my P is going to be let us say some phi 0, 1, W hat and then I have some other error. Let us say nu prime. Now, I am simply using this value from this equation to equation 1.

So, this is going to be my equation 5. And this is where I am claiming that I have actually solved the endogeneity problem. So, this W hat is no longer having this relationship with u p. So, it is no longer covering with the error. Now, how exactly I can claim that? To illustrate that let me actually take help of a diagram.

(Refer Slide Time: 05:34)



So, let us actually try to see all our, you know different components of our regression. So, let us say we have our Y, let us say Y is here, X is here and my error term is here. Here my Y is P price rate here is wage rate and this is my u p i. Now, I want to see the impact of wage rate on the price rate. So, this is something I am really interested in and I know this error term is actually influencing the Y and that is fine.

But which I do not want is that a channel here. This is something I do not want so, I mark it as red. Now, basically, if I have a channel here what will happen? The part of X influence will go through u and that will again influence Y. So, the coefficient I will get for X is going to be distorted. So, I do not want that. So, in order to avoid this what I am going to do? I am

basically claiming that let us say, if I can have a z here and if that z can help me to estimate X hat.

So, what I can do is, I can actually basically create or I can have a variable which will not be related with you. So, let me actually write it down. So, let us say this is my z, this is my X hat. So, X hat is nothing but my W i hat and j is nothing but U i and here let us say some other error term we have. So, maintain the consistency of color. So, let us say this is some theta is the error here.

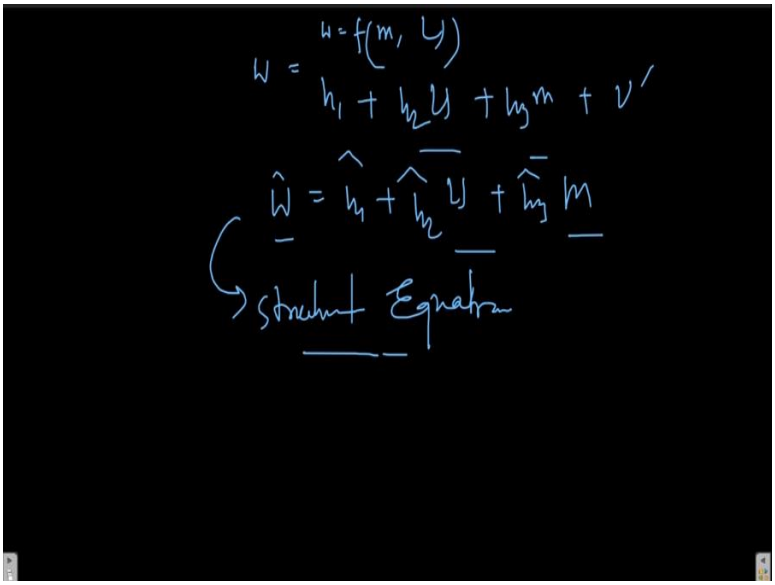
Now, what I am claiming is that the z and this error is actually helping me to estimate my X hat. And X hat is nothing but my W i hat. So, this instead of using W i am using W i hat to predict my price. So that is what I exactly did previously. So, I am basically creating this and I am saying that now, since I have X hat so, this relationship is no longer there, this relationship is no longer existing.

Now, how I can claim that this relationship is no longer existing? So, if I have X hat so, does that mean I have freed my equation from the error? Well so, we can think it in terms of the very simple regression that we have done at the beginning. So, where we had simply plotted in two dimensional graph X and Y. So, we had X and we had Y and then we plotted all these points here and there.

And these are my Y actuals. So, Y i's and then I had my regression line. This regression line is basically representing all these Y i hats. Now, this Y i hat, I got this Y i hat only after removing the error part. So, Y i minus the error is basically the Y i hat. So, when I remove the error, I basically when I get my W i hat. So, it is an estimate of W i. So, I can claim that no longer it is related with u p i. So, basically that is how we can claim that it is free of the error term.

And then when I take the regression using X hat so, we basically solve the endogeneity problem. So that is basically the intuitive understanding behind how we can use the estimated endogenous variable.

(Refer Slide Time: 09:47)

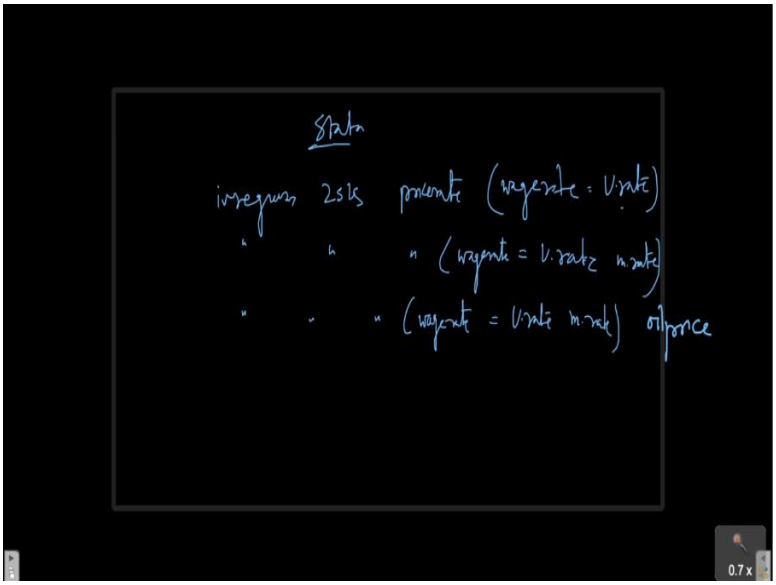


Now, if this is the case of over identification we can use the same strategy since for example in our previous case W was related to let us say both money supply rate as well as unemployment rate. So, what we can do is, we can you know instead of one instrumental variable I can have let us say h 1 + h 2 u and h 3 m and some u prime. Now, when I basically do it, I basically replace it.

So, I by regressing I get W hat which is h 1 hat, h 2 hat, u + h 3 hat m ok. So, this is something is a case of over identifications, we call it 2 S L S. But the strategy remains same. This one instrumental variable I get my W hat and I substitute my W hat into my structural equation so that exactly what you have done previously. So, this is how we use 2 S L S to solve the I V regression case.

Now, in stata how we do it? We usually the command is like this. We will do the hands on in the next class. But the command is something like this. I wanted to introduce the command just to have a sense of how we actually do that. And we will see that in the first stage in this most of the softwares, we actually do not do the first is separately.

(Refer Slide Time: 11:16)

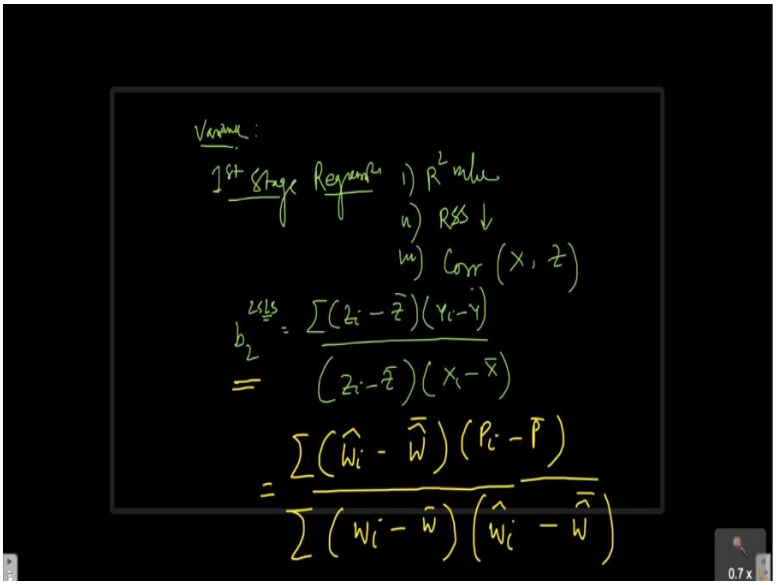


So, let me write down the command. So, we basically write I V regress 2 S L S then I write the Y variable. So, here let us say price rate and then I write the X variable my endogenous variable which is wedge rate and then I write the I V that I want to sort of assign for this wage rate. So, let us say unemployment wage rate writing and then I did not have any other X variables so, it is fine. But, if I let us say, if I want to have a two I V.

So, I can write I V regress 2 S L S the same price rate but in the bracket I write wage rate is equal to let us say U rate and m rate and. So, 2 I V I can just assign like this. Now, let us say I in my price equation there was another explanatory variable let us say that was let us say oil price. So, price rate can actually vary with oil price. So, if the oil price is increasing by prices of other goods usually increase in the economy.

So, if that is the case. So then I can write my I V the regression I can do is like this. Price rate, wage rate is equal to U rate m rate and then I can close the bracket. Let me reduce the size a little bit. I can close the bracket and then I can write let us say oil price. So, this is how we do the regression I V regression in stata. Now, some mathematical properties.

(Refer Slide Time: 13:06)



So, one thing is that usually you may see the variance or standard error for the I V regressor might be actually more than the ols and that is because it depends the standard error since the I V regressor you are not actually using the original X variable, original regressor but you are using some estimate of that. So that is actually going to be less precise and the standard error is going to be high. So, there is nothing to worry about on this.

The second part is when we do the first stage regression. So, I said the in software you usually do not it usually do not give you the result for the first stage regression. So, what you do? You basically run the first stage regression separately and you see, if the properties are satisfied. So, you will what properties you see? A very good R square value then secondly you see, if the residual sum square is low.

And third you want the correlation coefficient or the covariance between X and z to be a good one It has to be good correlation coefficient. Now, how do we represent the beta 2? Let us say i V in this case. It is going to be or beta 2 S L S we can also write 2 S L S. So, 2 S L S is the general case for both it basically encompasses I Vs also. So, how I am going to write? So, usually the usual equation is z i – z bar into Y i – Y bar divided by z i – z bar into X i – X bar.

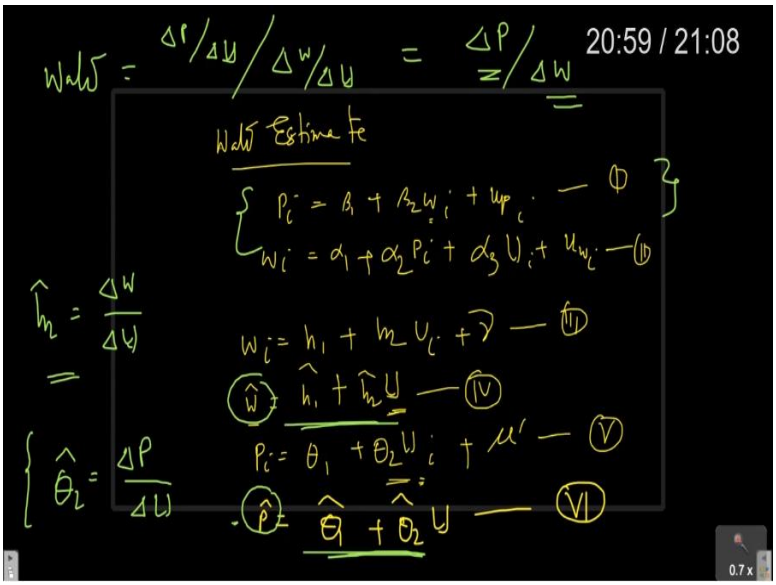
So, in the previous case, if I think about our wage rate and price rate equation. So, how it is going to look like? Let me use let me change the color a little bit. So, it is going to look like. So, instead of z i I am going to have my W i estimated. And then that estimated bar into I

have my P i Y i is here my price rate P bar. That is the numerator and in the denominator what I am going to have?

So, my X i is W i here. So, I will write W i – W this is for X and for z I am going to write W i estimated minus W. So, this is what the beta 2 2 S L S is going to look like. And here again, if suppose we are actually you know having a over estimation case where I have more than one I Vs. So, usually the standard error is going to be less, if I have an exact identification and the reason is when I have more than one I V, I am going to have my first stage least square.

First stage regression going to be more accurate. So that is why my error is going to be a little less. So, one last thing.

(Refer Slide Time: 16:20)



So, there are some other ways actually we also do this I V regression and you remember that we estimated Wald estimate the first example we have given. And this is I am going to show you how we actually calculate that. So, here what we do is, let me write down the equations again, let me write down P i is equal to beta 1 + beta 2 W i + u p i whereas my W i is equal to alpha 1 + alpha 2 into my P i + alpha 3 into U.

And then finally there is an error component u W i let us say. Now, in this case what we do is let us say we have our in the first stage regression I will have W i = h 1 + h 2 U i and let us assume error term. And when I estimate it, I get W i hat is equal to W hat is equal to h 1 hat +

h 2 hat into U. Now, so that is how we used to do previously that is how we have learned we do and then you substitute W hat in the first equation.

Again I number the equations 1, 2, 3, 4. Now, in this case we will do things a little differently. So, what I am going to do instead of actually substituting W into the first equation what I am going to do? I am going to use the I V the U directly into the first equation. So, what I mean by that? So, let us see, P i is going to be let us say I use theta. Now, theta 1 + theta 2 U instead of U i. So, P i U i so, instead of W i am going to use U i.

And let us say some I do not know let us say some error term is some mu prime. So, this is what I am going to do here. Instead of using the W is estimated I am going to use the instrument directly into the second stage regression. So, this is my equation 5. Now, what I get? So, my h 2 hat and my theta 2 what do they really represent? So, here what I am doing is, I am actually estimating my P i. So, I get let us say P hat, if I estimate P hat.

So, it is going to be theta 1 hat + theta 2 hat into U. So, my error term is gone. So, let us say this is equation 6. This is a fitted equation. So now, look at the equation 6 and my equation 4. So, in equation 4, I have estimated my endogenous variable W and in equation 6, I have estimated the dependent variable that I wanted to understand the relationship with the independent variable. So, this is the price rate.

So, essentially what this h 2 hat and theta 2 hat say? So, h 2 hat says basically change in unemployment rate and its impact on the wage rate. So, basically I can write delta wage by delta unemployment. Whereas my theta 2 hat is saying change in the price rate with the change in the unemployment rate. Now, what Wald estimate said? Wald estimate we took the reduced the coefficient of the reduced form equation which is basically this one the main equation here.

So, basically delta P by delta u and then we divided it by the first stage regression. So, basically delta h hat. So, h 2 hat which is delta W by delta u. So, if I simplify it. It is going to be delta P by delta W. So, essentially it means, I am looking at the change in the price rate with change in the wage rate. So that is basically what I wanted to get in my equation 1. So, with Wald estimate exactly that is what it is giving us.

So, it is basically telling the delta change in Y with respect to delta change in the explanatory variable. But unfortunately since the explanatory variable had endogenous problems we had to basically follow through all these different routes. So, this is how we do it and with this we will end this lecture here. And in the next lecture we are going to do some hands-on links data. Thank you.