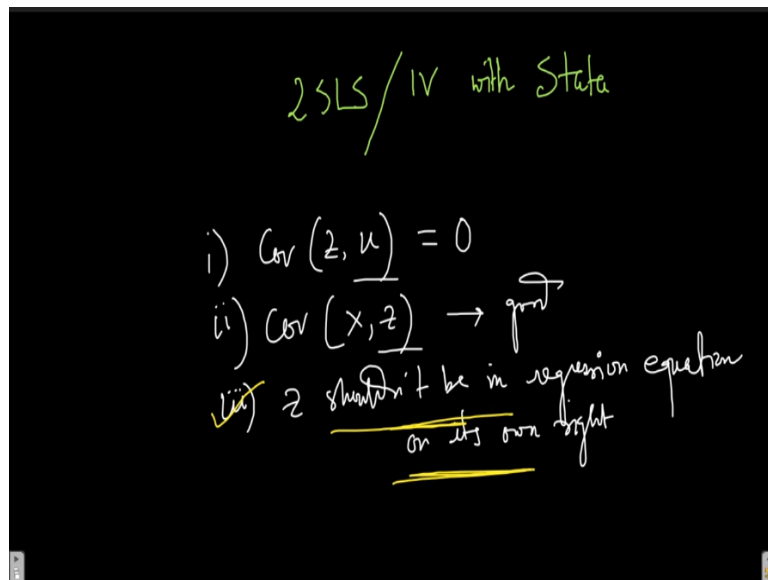


Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology – Kharagpur

Lecture – 107
2 SLS and IV with Stata

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 talked in detail about the 2 stage least squares strategy. Now, in this lecture we are actually going to implement this 2 stage least square strategy using stata. So, this is as I said earlier is the easiest part of the whole series of this lecture on instrumental variable. It is easier because we have already learned the concept.

We have already learned how to interpret the I V regressor and how to actually see understand the results. So, we will try to actually use the data that we have been using all throughout and the data is basically national sample survey data. So, I already have this data here. **(Video Starts: 01:07)** I have already loaded this data and this national sample server data for West Bengal.

And I have my do file ready. I also have my stata output window ready. So, we are all set to actually do this work **(Video Ends: 01:23)**. Now, we know few conditions that we have to keep in mind, just to sort of for our own convenience, I am writing it down. So, first thing is

that covariance of the error term and the u in the original second stage least square regression, it has to be equal to 0.

And the second condition is that the covariance of z and X has to be high or is to be, you know significantly good, should be good. And the third is that the z should not be in regression equation on its own right. That these three conditions we already know but we are just going to actually check these three conditions when we are going to implement the code. Let us just do that.

Let us first open our do file (**Video Starts: 02:40**). So, basically I again will run the wedge regression equation. And what I am going to do? I am going to have my log of wage total as the outcome variable. And then I have my explanatory variable general education, experience square sex. And what I am going to think is that like Kruger and Angrist in their work and on later on also we found education is not really an exogenous variable.

But rather, it is an endogenous variable. So, in this exercise we are going to actually assume education to be the endogenous variable. Education to be the X variable. So, we have to find a z here instead of you know as an instrument for education. Now, let us first before we begin let me actually go back to the code here. So, what I have done? I have create, I have basically run a simple ols with log of wage total as the outcome variable.

And education is the endogenous regressor. And these are let us say exogenous regressor that we do not have any problem with all these regressors. And we do this regression for the people who are having casual employment or they are earning wages or salaries, they are not self-employed. And the sector that is basically the rural sector. So, I will talk about why I have chosen rural sector? So, let me first run the regression here.

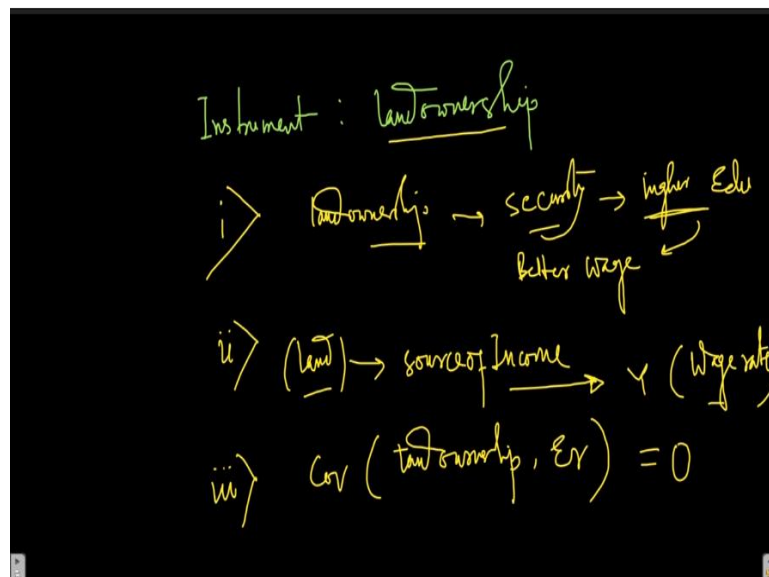
I have run the regression and let us actually look at the regression result. So, I will go to the output file and I see that general education. So, basically the R^2 is 0.4, descent general education is 0.15 coefficient which is a significant and all other variables are significant. If females are running less than men basically that all makes sense. Now, what I said is that we are basically going to assume that general education as the endogenous regressor.

Let us say from theory we already know that. And then what I am going to do? I am going to actually find out the error term from this regression equation basically I will subtract the fitted Y from the actual Y. So, if we do that. So, let us say predict wage total and this is basically, I first predict wage total \hat{Y} and then I create a variable error and error is wage total minus the predicted wage total.

So, I do that sorry, actually I have run the regression so that is why, it is actually giving this error. So, I just will run this part. So, already I have these results with me. So that is the reason why I am basically I will show the results. So, basically this telling 0 changes are made. So, we already have the results stored. So, let us actually use this result and see, if my error term is going to be correlated with the instrument.

Now, what is the instrument here? Now, I will bring a little bit of economics here. The instrument I am assuming here (**Video Ends: 05:46**) is the land ownership and why I am assuming land ownership could be an instrument. So, let me give you two arguments here. So, land ownership as instrument. So, land ownership.

(Refer Slide Time: 05:58)



So, this instrument the z variable is land ownership for education. Now, two possible reasons, two different arguments rather. So, argument one. So, if I have land ownership. So, think about a rural setting you do not have a lot of sources of income but you have land. So now, when you have land you get a sense of security. So, you have something to fall back on in your bad time. So, when you have a big amount of land, you can afford to send your kid to higher schooling.

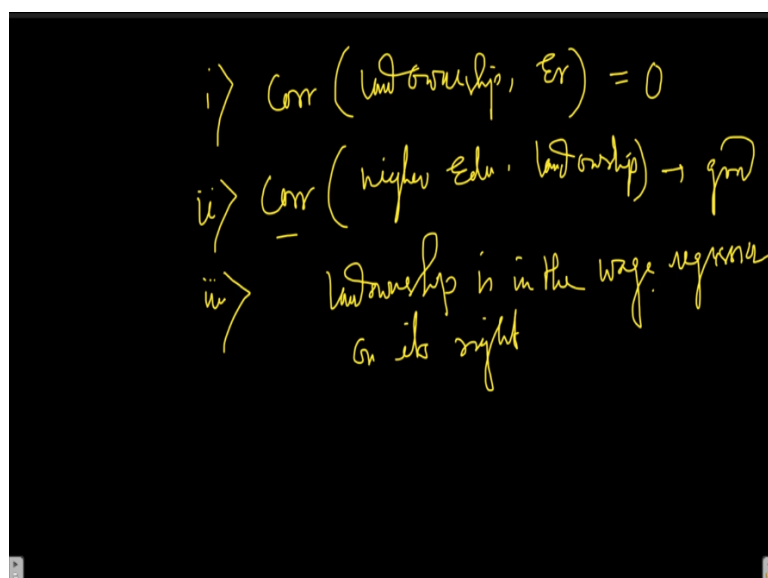
So that gives a sense of security and that helps children from the family to go for higher education. So, basically land ownership is a security, some sort of security and that actually helps the children to go for higher education let us say. And when children are getting higher education that is actually leading to wage, a better wage. Now, the other option is that the land itself could be a source of income.

So, at present this is a source of income. Now, if land itself is a source of income that may actually directly contribute to the Y, the wage rate. So, if I already have my land and if I let us say work in my own land and if I get some earning from that land, so that is actually going to increase my wages let us say. Let us say I am working as a wage worker in my own land. So, this is another possibility.

So, in that case what is going to happen? This land ownership is actually going to be present in the regression equation on its own right. So, this is going to happen. So, this is something we have to check. We have to check this assumption that we have made. So, if the land ownership is actually leading to higher education. So, what you have to do? We have to simply run a regression.

So, the regression for obtaining the relationship between education and land ownership. And the third part is that I have to see, if the covariance between this land ownership and the error that I have obtained is equal to 0. If these three conditions are satisfying.

(Refer Slide Time: 08:36)



i) $\text{Corr}(\text{Land ownership}, \text{Er}) = 0$
ii) $\text{Corr}(\text{higher Edu.}, \text{Land ownership}) \rightarrow \text{good}$
iii) Land ownership is in the wage regression on its right

So, basically again let me strategize. So, first I am going to see the correlation between land ownership and error term is equal to 0. If that is satisfying the second condition I am trying to check, if the correlation between higher education or I can actually do a regression here, higher education and land ownership is good. And third I am going to check, if the land ownership is in the wage regression on its own right.

There three conditions I am going to check, on its own right. So, let us do that. So, let us go back to the do file. **(Video Starts: 09:35)** So, what I am going to do? So, land ownership I have chosen based on some of my you know intuitions and first let me see the correlation between a return and the land ownership. So, what is the result, what result we are going to get? So, if we do that correlation between land ownership and the error term.

So, what I get is this and also let me run this one also. So, basically I will be done with two. So, I can also run a regress basically instead of regress Y on X and we can also specify that actually, if let us say this whole condition is there actually that is a good thing to do. So, let me run again. So, this one, I run the first condition z and u, correlation covariance z and u. And the second condition I run the first stage least square regression.

So, what we see here? In the first stage I see that R square is not that great and the land ownership, the coefficient is also very, very small but, it is significant. Whereas, if I see the correlation between land ownership and error term this is going to be a pretty low 0.02. So, I can consider to be a very poor correlation. So, the first condition is fine. The second condition is somewhat fine because it is not really greatly explaining the R square.

So, it might be a weak instrument instead of a strong instrument. Now, the third condition, what I am going to see? The third condition I am actually going to run this ols just to check, if and I am going to include the land ownership into the regression equation just to see, if the land ownership is actually explaining the wage total on its own right. So, let me actually run that and let us actually see the results.

The result is here. So, we see that actually land ownership is present in the regression equation on its own right. So, it is actually explaining the wage earnings that we are, you know people are making in the rural area. Now, with this we can at best say that perhaps this IV that I have chosen is not the best IV probably, it is let us say, it is some sort of weak IV.

So, with this weak I V let us still run the regression just to see how the results are looking like.

So, we know the code by now. So, what we do is, we write I V regress 2 S L S and then the outcome variable log of wage total. And then I have my instrument general education is equal to land ownership and then I have other all list of exogenous variables. So, we can run it, let us run these, actually we can run both of these so, just to compare. But we already have this result. But let us just run the first one.

This is the first one. So, the result that we get here is this. So, we see that general education, the coefficient is significantly increased and the reason is the earlier you see the general division coefficient is 0.15, here is 0.23. That is in, when I am basically using the instrumental variable, it is perhaps the instrumental variable is more relevant in the original regression equation. So, it is basically the instrumental variable itself is contributing in its own right.

So that is the reason why general education coefficient is suddenly increased so much because it is now, instrumented with land holding. So, perhaps the instrument that I have chosen is not the best instrument. But still we get an idea how to run this instrumental variable regression. Now, going back to the do file we can also use a command like just I V reg and we do not have to write 2 S L S and we can do the regression.

But 2 S L S itself cannot be the command. So, this is basically the idea of how we can use I V regression using stata. And you need to, as you could understand is very important and is very critical to actually find out the right instrument to run this I V regression. With this we end this lecture here. Thank you. **(Video Ends: 14:14)**