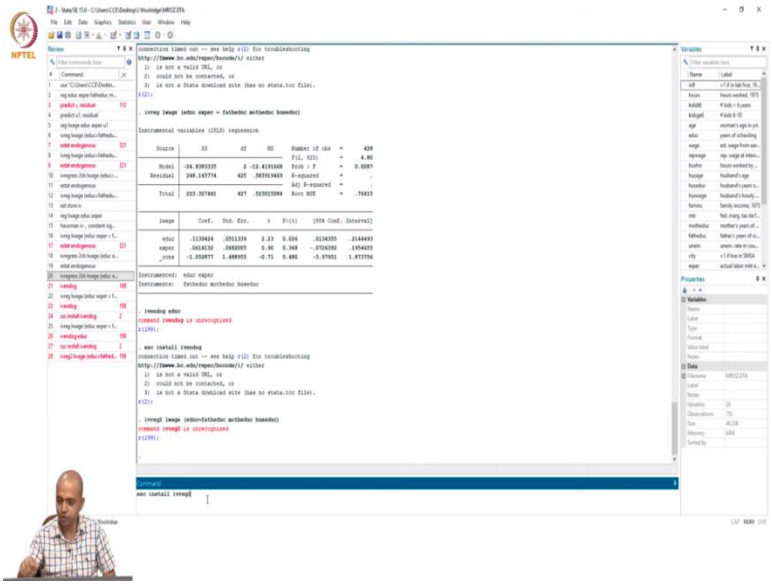
Applied Econometrics
Prof. Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras

Lecture - 11
Instrumental Variable Estimation – Part XI

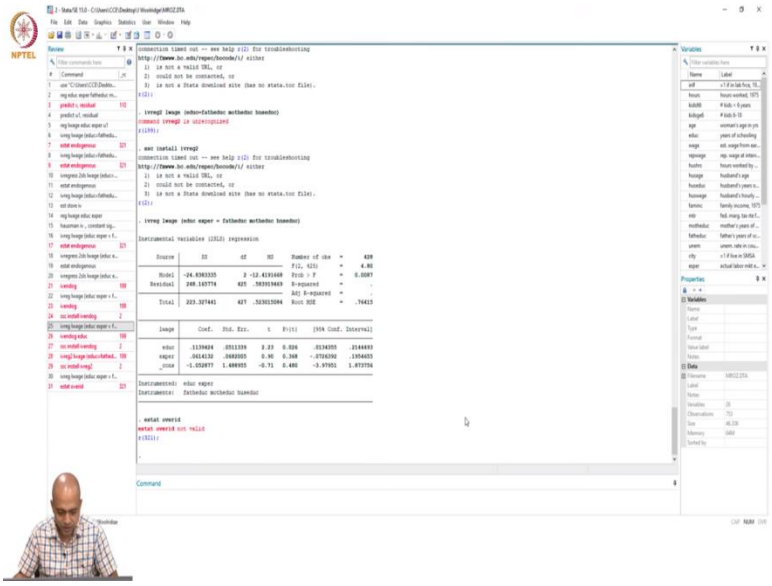
(Refer Slide Time: 00:15)



Now one thing lastly what we need to do is the test for over identification, what we said earlier that if we had more than one instrument then our system is over identified that means we have one endogenous variable. For example, if we do this let us ivreg lwage and then education equals to father's education, mother's education, husband's education then we have one endogenous variable for which we are using three instruments.

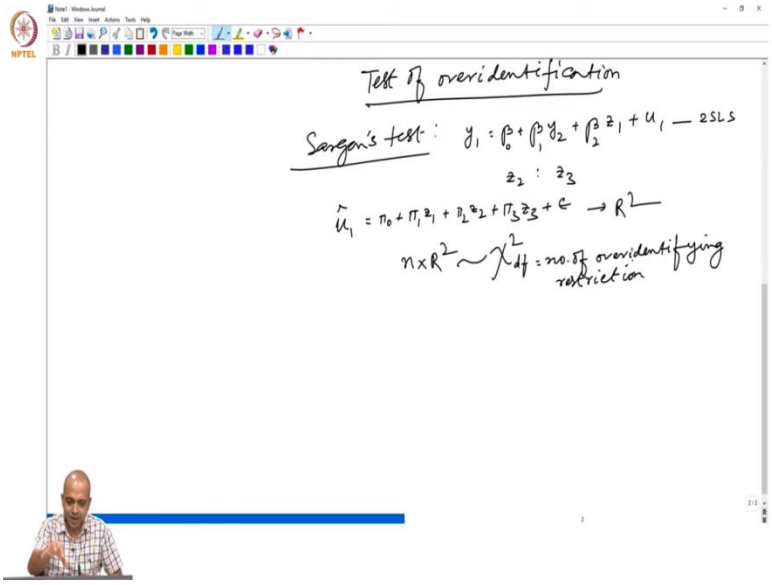
So, these are called this system is then called over identified system. And this over identification test will tell whether all these instruments are actually exogenous or not. So, that means whether all these instruments they are correlated with the error term or not. So, if we do this iv reg 2 is also I am not able to implement here because again the same problem iv reg 2 I am not able to install because it is not connected with the internet. I will try another command I will take this then let me see.

(Refer Slide Time: 02:21)



Now after this you can use estat over id, over id is also not a valid because these commands we have to download actually since I am not able to connect it with the internet. So, what you do basically you download this command over identification, and what is the test, this test was developed by Sargan that is why it is called Sargan's test.

(Refer Slide Time: 02:59)

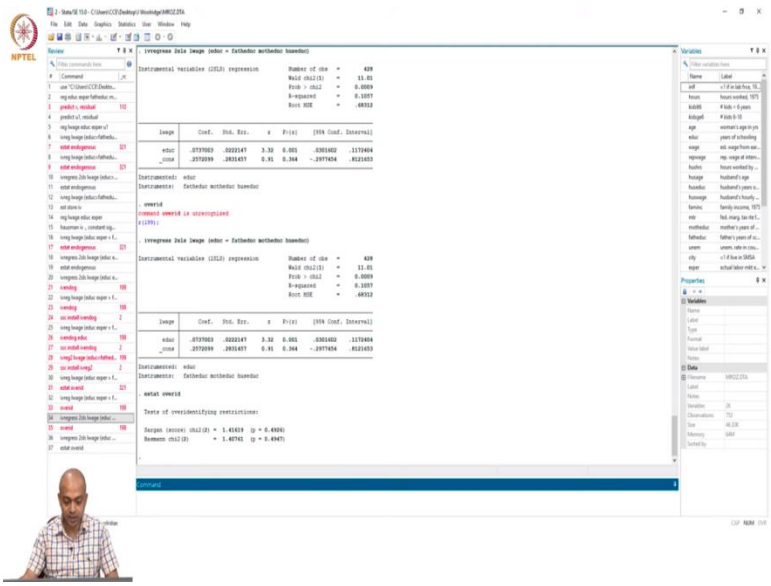


So, over test of over identification this is called Sargan's test, so what we have to do let us say that this is our model $Y_1 = \beta_0 + \beta_1 Y_2 + \beta_2 Z_1 + U_1$ and we have 2 instrument Z_2 and Z_3, these are the two instruments. So, we have to estimate this model using 2 SLS and then we have to get U_1 hat. So, this we have to estimate using 2 SLS and then we have to get the predicted value of the error term.

And that we have to regress on all the exogenous variable $\pi_0 + \pi_1 Z_1 + \pi_2 Z_2 + \pi_3 Z_3$ plus let us say epsilon. And then from here we have to get the R square then n into R square we have to get which you will follow a chi square distribution with degrees of freedom equals to number of over identifying restriction. What is the number of over identifying restriction here? It is 1, because we have one endogenous variable but we are using two instruments.

So, that is why there is one over identifying restriction. And that Sargan test we can implement in stata using the stat over id or iv reg 2.

(Refer Slide Time: 05:44)



Also, if we use then we will automatically get the Sargan's test statistic. Unfortunately, I am not able to implement the test because I am not able to install neither iv reg 2 or a stat over id. So, I will try once iv reg this over id let me see, over id is also not working, so iv regress 2 if I put iv regress 2 SLS let me see iv regress 2 SLS then over id over ideas cannot take, estat over id. Here it is working.

So, that means I we regress 2 SLS we have to use the command and then instead over id, look at this test of over identifying restriction Sargan score 1.41 p value is 0.49 so that means we cannot reject the null but we have to be very careful what is our null. So, our null is over identifying

restrictions are valid that is the null. So, whenever you are implementing Sargan test we have to be very careful about our null hypothesis.

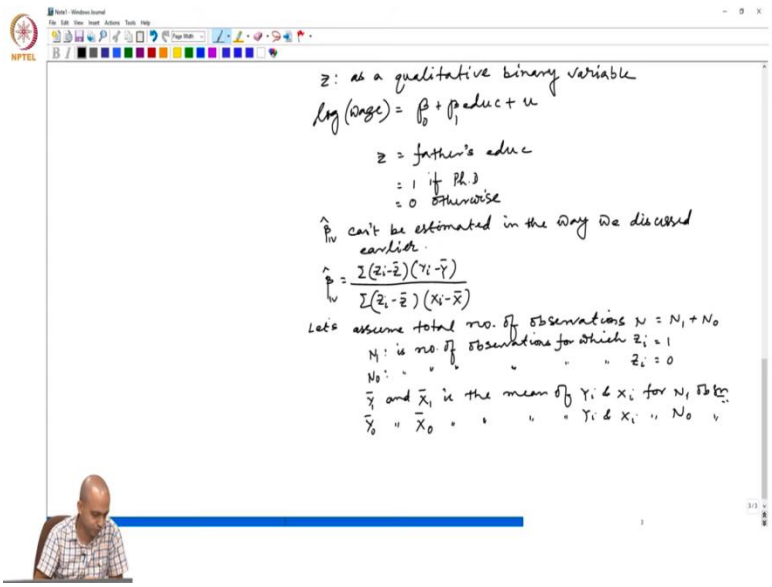
So, test of over identifying restriction that is what we are testing that means we are saying what identifying restrictions are valid so that means if we cannot reject this null that means we are saying over identifying restrictions are valid that means father's education, mother's education, husband's education all of them are uncorrelated with the error term. So, they all are valid instruments.

So, in this way we can actually test the validity of the instruments when we have more than one instrument to work with. So, 1.41 is the chi square value and as I said it follows a it follows a chi squared distribution with degrees of freedom equals to over identifying restriction, we have one endogenous variable but we have used three instruments that is why 3 minus 1 2 would be the degrees of freedom in the Sargan test of over identifying restriction, this is how it works.

So, that means we have discussed everything, we have discussed test of endogeneity, we have discussed how to test endogeneity, we have discussed manual estimation. We have also discussed how to test work with the test commands stata command then lastly, we have discussed about over identification that is also that is also there. Then lastly what we will do? We will discuss one more case that is so far whatever instruments we have discussed they all are quantitative variable that means father's education.

How do you measure father's education? Father's education may be measured as number of schooling for the father. But now let us say that our instrument is actually a qualitative variable.

(Refer Slide Time: 09:47)



So, we are using z as a qualitative binary variable. So, this is our model log which equals to $\beta_0 + \beta_1 \, education + u_1$ and z is actually father's education equals to 1 let us say if PhD so this is our model. So, z is now a qualitative binary variable whether the father is a PhD or not, in this case what would be our beta hat iv, so beta hat iv we cannot be estimated in the way we discussed earlier. So, that means we need to do some modification.

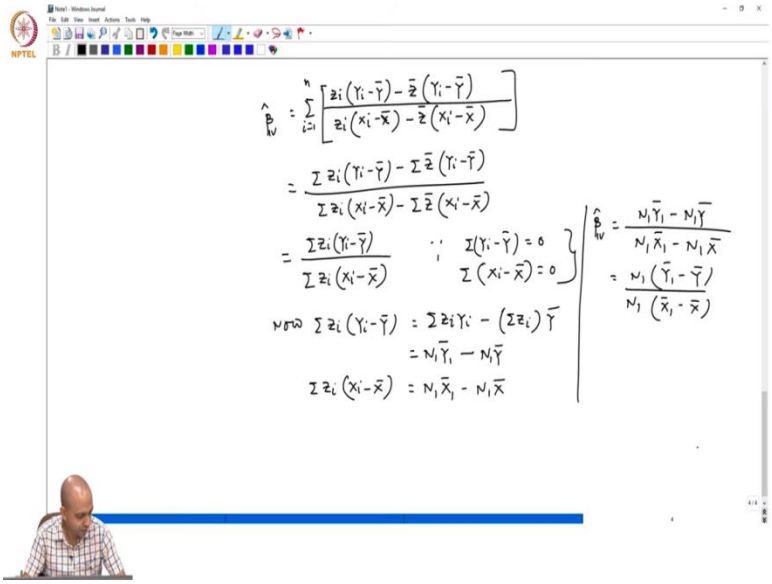
So, earlier what is the beta hat iv we discussed? Beta hat iv we said that this is nothing but summation $Z_i - \bar{Z}$ into $Y_i - \bar{Y}$divided by summation $Z_i - \bar{Z}$ into $X_i - \bar{X}$ that is the iv method we discussed earlier. So, we need to do some modification in this case. So, let us assume that total number of observations in is actually equals to $N_1 + N_0$ $N_1 + N_0$. What is N_1? N_1 is number of observations for which for which Z_i is Z_i = 1.

That means for fast N_1 number of observations their fathers are having PhD degree and N_0 is number of observations for which Z_i = 0. Then we also assume that Y_1 bar is basically or Y_1 bar and X_1 bar is the mean of Y i and X i for N_1 observation and Y bar 0 Y_0 bar and X_0 bar is the mean of Y i and X i for in 0 observation. That means we have two group for the first group mean is Y_1 bar X_1 bar for the second group mean is Y_0 bar X_0 bar this is very simple.

So, we have divided the entire sample into two first group and second group for the first group we have N_1 observation, second group N_0 observation and N_1 is basically the observation for all these

for the first group all of them have PhD for the second group there is no PhD that is what we mean. For the PhD group the mean of Y is Y_1 bar mean of X is X_1 bar for the non-PhD group mean of Y is Y_0 bar and mean of X is X_0 bar.

(Refer Slide Time: 16:16)



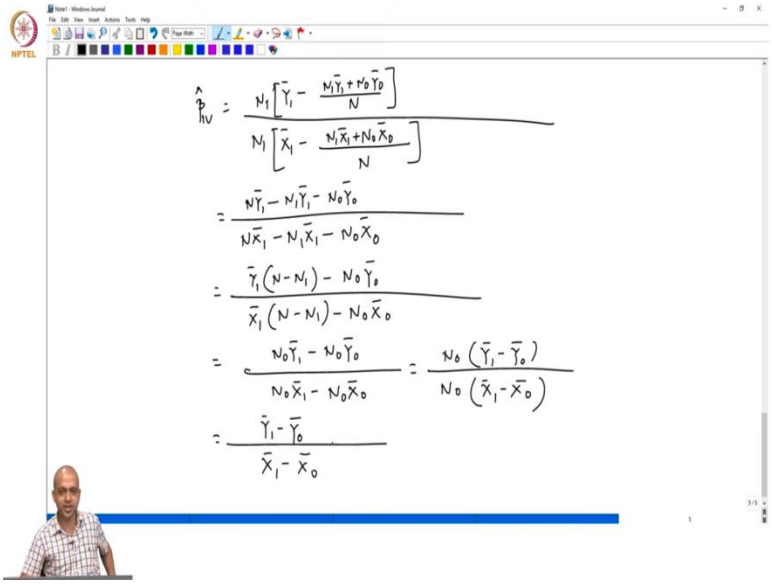
Now beta hat iv as we said it is basically a summation, i running from 1 to n it is summation, so Z_i minus so if you previously I am simply Z_i Z_i into $Y_i - \bar{Y}$ - z bar into $Y_i - \bar{Y}$. What I am doing simply decomposing the term because the numerator where $Z_i - \bar{Z}$ into $Y_i - \bar{Y}$, so I am just saying that means it is equals to Z_i into $Y_i - \bar{Y}$into z bar into $Y_i - \bar{Y}$and divided by Z_i into $Y_i - \bar{Y}$ minus.

So, what was our term? We will just go back and see so Z_i - z bar into $X_i - \bar{X}$ which is equals to Z_i into X i - X bar - z bar into $X_i - \bar{X}$. Then it is not Y i Z_i into $X_i - \bar{X}$ - z bar into $X_i - \bar{X}$this is what we get. So, this equals to then what we can write that summation Z_i into $Y_i - \bar{Y}$ - summation z bar into $Y_i - \bar{Y}$divided by summation v into $X_i - \bar{X}$- summation set bar into $X_i - \bar{X}$= summation Z_i into $Y_i - \bar{Y}$divided by summation Z_i into $X_i - \bar{X}$.

Because summation this z bar will come out of the summation then summation $Y_i - \bar{Y}$= 0 summation $X_i - \bar{X}$ is also equals to 0 because summation $Y_i - \bar{Y}$= 0 summation $X_i - \bar{X}$ that is also equals to 0. So, this is equals to this now will decompose summation Z_i into $Y_i - \bar{Y}$ = summation Z_i Y i - summation Z_i into Y bar. And summation Z_i u i is nothing but N_1 into Y_1 bar divided - N_1 into Y bar.

Similarly, summation Z_i into $X_i - \bar{X}$= N_1 into X_1 bar - N_1 into X bar. So, this what we will do? This we will substitute in this will substitute there. So, beta had iv then here if we substitute this so beta hat iv then would be equals to N_1 into Y_1 bar – N_1 into Y bar divided by N_1 into X_1 bar – N_1 into X bar N_1 into Y_1 bar - Y bar. So, that means we can say that this is equals to N_1 into Y_1 bar - Y bar divided by N_1 into X_1 bar - X bar. Now this again, so what we will do? We will use Y bar as the grand mean.

(Refer Slide Time: 23:09)



So, in next page what we will do that means beta hat iv equals to will simply take N into N_1 into what we will do we will take Y_1 bar minus in place of Y bar, what we will write N_1 into Y_1 bar + N_0 into Y_0 sbar divided by N that is nothing but Y bar similarly in the denominator N_1 into X_1 bar - N_1 X_1 bar + N_0 X_0 bar divided by N so equals to what we will get equals to what we will get N_1 will get cancelled.

So, N into Y_1 bar - N_1 Y_1 bar - N_0 Y_0 bar so this should become N into Y_1 bar - N_1 into and in the denominator N into X_1 bar - N_1 into X_1 bar - N_0 X_0 bar. So, now from here what we can do? We can take y one common Y_1 bar we can take common so this would become N - N_1 in the numerator it would become if we take Y_1 common Y_1 bar if we take common in 1 - N_0 Y_0 bar here if I take X_1 bar common N – N_1 - N_0 X_0 bar.

So, this would become N - N_0 is nothing but N - N_1 is N_0 Y_1 bar - N_0 Y_0 bar divided by N_0 X_1 bar - N_0 X_0 bar equals to if I take N_0 , Y_1 bar - Y_0 bar X_1 bar - X_0 bar. So, that means ultimately this would become Y_1 bar - Y_0 bar divided by X_1 bar - X_0. So, this is our beta hat iv which has simply come out as the difference of the Y mean as the numerator difference as the X mean as the denominator that is what is the beta hat iv.

So, that means when we have a qualitative variable as instrument then we cannot use the same technique what we used for quantitative instruments to be used same method. We cannot use this is the method we have to use that means simply we have to segregate the entire sample into two and then for the first group we have N_1 observation second group N_0 observation N = N_1 + N_0 for the first group all the others are having PhD.

Second group they do not have PhD or the first group the way mean of wages let us say Y_1 bar for the second group it is Y_0 bar for the first group let us say educations is X_1 bar and that is X_0 bar. So, you will get nicely Y_1 bar - Y_0 bar divided by X_1 bar - X_0 bar as our beta hat iv. So, with this we are closing our discussion today; will again remaining things we will discuss in our next class, thank you.