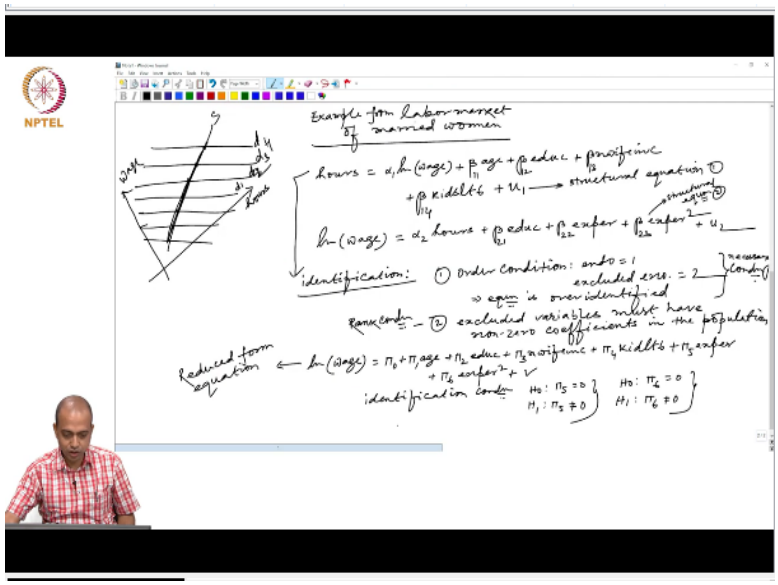
Applied Econometrics
Prof. Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology-Madras

Lecture-17
Simultaneous Equation Model-Part VI

(Refer Slide Time: 00:16)



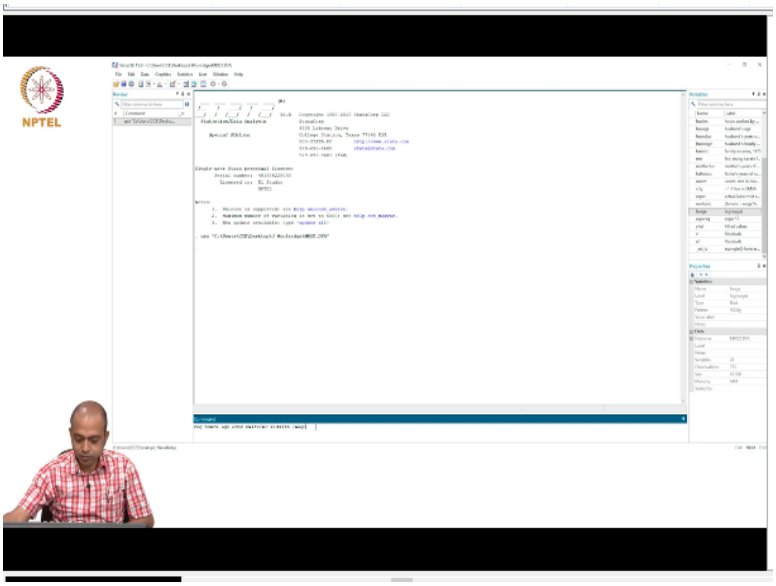
This is order condition and this is called rank condition. So, the reduced form equation if we formulate it would be very easy for us to understand the rank condition. So, let us now formulate the reduced form equation for the endogenous variable which is

$$\log(wage) = \pi_{io} + \pi_{i1age} + \pi_{i2}education + \pi_{i3}nonwifeincome + \pi_{i4}kidslt6 + \pi_{i5}experience + \pi_{i6}experiencesquare + v.$$

Reduced form equation is an equation where an endogenous variable is a function of all excluded and included exogenous variables. So, the best idea is you first include all the exogenous variables which are appearing in the equation, age, education, non-wife income , kidslt6, then you include the excluded one which is experience and experiences square. They are not appearing in this equation and some V. So, this is the reduced form equation. What is the identification condition then? Identification condition can be checked by testing this hypothesis of π_{i5}=0 against the alternative π_{i5} not equal to 0; π_{i6} equal to 0 against the alternative π_{i6} not equal to 0. These are the identification conditions. So, that means either π_{i5}or π_{i6}should have non-zero coefficient. So, this is what is called identification condition with an example.

Now we will first take one data set and we will try to estimate these two models. So, first we will estimate the labour supply function without bothering much about the simultaneity bias.

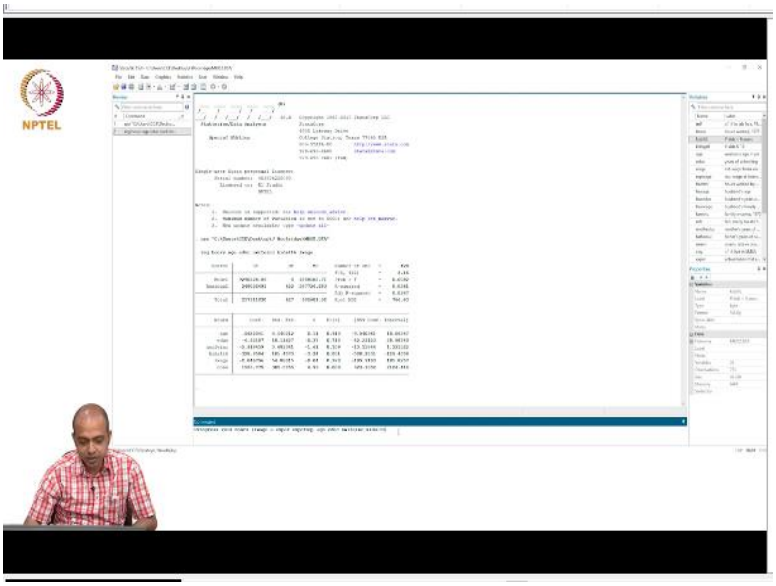
(Refer Slide Time: 04:13)



So, we will open this data set and our function is: hours is a function of wage, age education, non-wife income, kidslt6. So, what is the command: reg hours. Hours is actually a function of your age, your education, then non-wife income and then number of kids less than 6 given by kidslt6.

So, this is the labour supply function. Labour supply is a function of your age, education, non-wife income and kidslt6 and also log wage. So, labour supply is a function of wage, age, education, non-wife income and kidslt6.

(Refer Slide Time: 06:45)



So, what is happening here? Look at the coefficient. The coefficient of log wage is actually negative. That means wage and labour hours are negatively correlated which is difficult to explain. When wage increases, the labour hours decreases and if you look at the variable is not significant-the p value it is 0.97. So, that means log wage is actually not significant here which is difficult to explain in the labour supply function. How much labour I will supply will very much depend on what is the market wage rate. But that is not significant here. What could be the reason? Probably this model is suffering from simultaneity bias, because wage and hours are simultaneously determined, jointly determined and we are trying to estimate the model using a single equation method which might be the reason for which this wage is having a negative coefficient which is insignificant. Now what we will do?

To solve this you will try to estimate this equation using the instrumental variable method. Now instrumental variable method means, if you go back again, we have to use instruments for this wage (log wage) and the variables which are the instruments, that are not appearing in this equation is experience and experiences square. We will estimate the same equation using the command: **ivregress 2sls hours (log wage = exper sq experiences square) age education nonwifeincome kidslt6.** These are the variables. So, we have used two instruments for the endogenous wage variable and all other exogenous variables are appearing just like that. **(Video Starts: 10:50)**

Then put enter, now we will see what is happening. Log wage which was insignificant earlier is very well significant now and it is also giving a positive coefficient. So, that means there is some kind of positive correlation between wage and labour supply which we actually expect. So, that means the moment we solve the simultaneity bias our model gives a better prediction. In the same equation if you put comma first then what is stata is doing, when I put the first command stata is showing the first stage and the second stage. In the first stage Stata is actually running the reduced form equation where log wage is actually a function of age, education, non-wife income, kidlt6, experience and experience square.

This equation is exactly the same equation what we are talking about-the reduced form equation. The variables are: Age, education, non-wife income, kidslt6, experience, experience square. All these variables are appearing in the reduced from equation of wage or rather log wage.

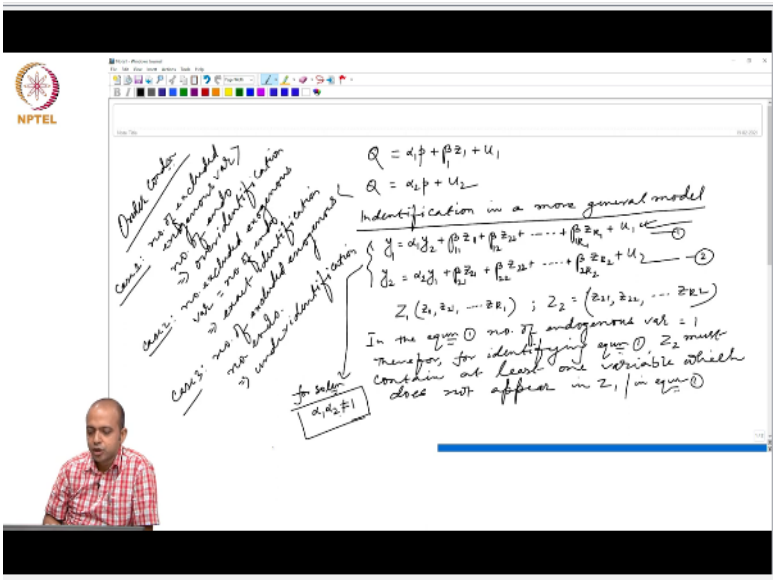
And then in the second stage we are putting the predicted value of log wage and estimating the equation and STATA is giving the second stage final structural equation. So, the idea is, if you recall, we will estimate the predicted value of the endogenous variable from the reduced form and then we will put that variable in the structural equation as an additional variable here to solve the endogeneity problem, because log wage was correlated with your error term u_1. And that is why instead of using log wage we are using the predicted one and this is the reduced from equation. So, this is exactly what STATA is doing. Now one thing we have to keep in mind is that as I told you earlier, STATA is helping you to understand what is instrumented. So, instrumented means the endogenous variable which is log wage and the instruments are experience and experiences square. But the other variables are also appearing. If you look at the command, I have mentioned only experience and experience square as my instruments but STATA is saying age, education, non-wife income, kidslt6 are also instruments. You have to keep in mind very clearly that exogenous variables by themselves are instruments for them. That is why STATA is reporting those variables also as instruments.

Experience and experience square are excluded exogenous variables but other exogenous variables are also exogenous since they are exogenous explanatory variables and are used as instruments for themselves. So, age does not require any instrument because age itself is its own instrument, education is its own instrument, non-wife income is its own instrument, kidslt6 its own instrument, experience and experiences square are however but instruments for the wage actually. So, that means, from this equation since there are two instruments, I am trying to get a linear combination of those. So, this equation basically gives a linear combination, so I will get a fitted value. So, the linear combination of all those explanatory variables will keep the fitted value of log wage and that will be used in the equation.

But as I said we will not do manually we have to use STATA only because if you do manually You will put only the log wage in this equation and error term will remain e_1. But actually if you look at this equation, you have log wage plus this error term also. So, ideally the error terms would also be included there which will not be included manually. That is why manual estimation of 2sls is not suggested, we have to go for STATA estimation only, even though they are logically the same. So, there is a huge change in the sign and coefficient of this log wage and since this is expressed in logarithmic form you can take this coefficient as elasticity measure, interpretation would be for 1% in wage, there would be this much unit change in labour supply. That is the idea.

So, far then we have discussed about identification in general and why identification is required prior to estimation. Because without identification you do not know which equation you are identifying, and then we have discussed about rank condition and order condition wherein we said that while order condition is necessary it is not sufficient, so sufficient condition is given by the rank condition which says that the excluded variable must be significant. **(Video Ends: 19:56)** Then we have generalized the equation.

(Refer Slide Time: 20:02)



This is the generalized equation and in the generalized equation also for solution of **y_1 and y_2**, the condition what is required is same. That is α_1, α_2 not equals to 1. Since one is demand function and another one is the supply function, so one is positively sloped and another one is negatively sloped, it is not quite unrealistic to assume that α_1 and α_2 , product of this is not equals to 1.

Rather they are negative and then we have also discussed about three types of order condition- over identification, under identification and exact identification depending on the number of excluded variables and number of endogenous variables. When the number of excluded variables is just equal to number of endogenous variables, we said that the equation is exactly identified. When the number of excluded variables is more than the endogenous variables we said that the system is over identified and when the number of excluded variable is less, we said that the equation is under identified. Then we took an example from the labour market and we establish and we understood the consequence of not considering simultaneity. So, when we have not considered the simultaneity bias that means simultaneous nature of wage and labour

supply, then the labour supply equation will produce biased estimates-the sign and magnitude of log wage will become actually wrong which is not supported by the theory. The moment we estimated the equation by IV regression, 2SLS IV regression, then we got significant value and sign is also is accordance with the theory.

So, with this we are closing our discussion today and in our next class we will discuss the rank and ordered condition in a system of equation. When there are more than two equations- when the number of equations are 3, 4, 5 or 6, then in a system of equation it is very difficult to understand the rank and order condition. So, that is why we need to have a more generalized rank and order condition for a system of equations.