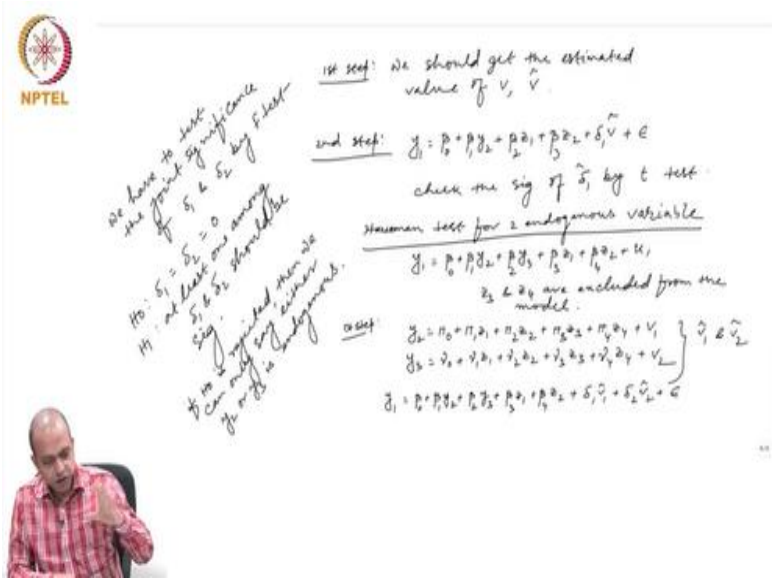
Applied Econometrics
Prof. Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras

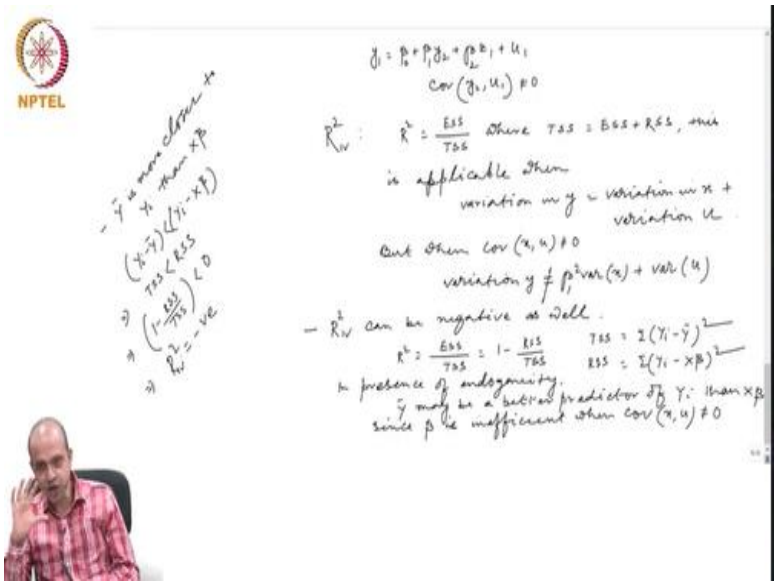
Lecture - 06
Instrumental Variable Estimation – Part VI

(Refer Slide Time: 00:15)



Now this is how we can actually test endogeneity. Now once all this discussion about endogeneity instrumental variable estimation technique are over then we need to understand couple of more things about this model.

(Refer Slide Time: 00:39)



That means, once we estimate this model then we are saying that $Y_1 = \beta_0 + \beta_1 Y_2 + \beta_2 Z_1 + U_1$ and we say that covariance between Y_2 and U_1 is not equals to 0. We have so far not discussed anything about the R square that we actually get by estimating this model by IV estimation technique. So, that means the goodness of feet or R square IV that we need to discuss R square IV. Now what is R square generally?

R square is basically ESS by TSS where TSS = ESS + RSS. So, that means this standard definition of R square is applicable, when the total variation in the dependent variable total variation in y is actually possible to decompose into two components. So, that means this is possible, this is applicable when variation in y = variation in x + variation in u. But when covariance between x and u is actually not equals to 0.

Then what happens variation in $Y = \beta_1^2$ into variation in x plus variation in u this is actually not true. Because there is some covariance term also between x and y, so unless that variation in $Y = \beta_1^2$ variation in x plus variation in u is possible to write, we cannot say that the standard interpretation of R square is actually applicable. Because the standard interpretation or meaning of R square is coming from ESS by TSS, where TSS is actually = ESS + RSS.

So, that means variance and in y = variation in x + variation in u. This type of clear decomposition of total variance in y is not possible when x and u are actually correlated. That is why the standard

definition of R square is not possible in this context. And sometimes what happens R square IV can be negative as well. Why the R square IV can be negative? Because R^2= ESS by TSS = 1 - ESS is TSS – RSS so, 1 - RSS by TSS now $TSS = \sum(y_i - \bar{y})^2$.

And RSS is $\sum(y_i - x\beta)^2$. Now in presence of endogeneity, $\bar{y}$ may be a better predictor of the actual y_i then x beta since beta is inefficient when covariance between x and u is actually not equals to 0. So, in presence of endogeneity beta is actually inefficient that is why $\bar{y}$ in a regression what we are trying to do, we are actually trying to predict our y_i.

So, when x and u they are correlated in presence of endogeneity the average value sample average value of y_i which is $\bar{y}$, that is a better predictor than x beta. So, that means $\bar{y}$ is more closer to y_i then x beta. If that is the case what happens that $y_i - \bar{y}$ is lower than $y_i - x\beta$ because $\bar{y}$ is more closer to y_i, so this implies that TSS is actually lower than RSS. When TSS is lower than RSS that means 1 - RSS by TSS then may is less than 0.

That means R square equals to negative, that is the logic. So, intuition is very interesting so R square is basically 1 - RSS by TSS and this is TSS, this is RSS $y_i - \bar{y}$ $y_i - x\beta$. In presence of endogeneity when you estimate this model by IV, even in IV this x beta, beta is inefficient that means, standard error of the IV is always higher than the OLS that is the disadvantage of using IV.

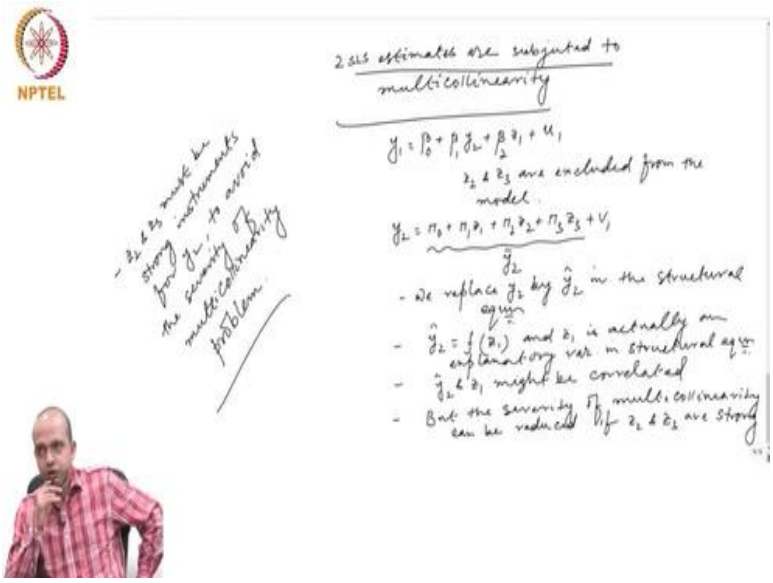
So, obviously the difference between $y_i - \bar{y}$ would be lower than $y_i - x\beta$ since, it is a better predictor means y bar is actually more closer to IV than x beta. And when this is more closer this would be less than this, so TSS is less than RSS 1 - RSS by TSS is less than 0 so R square equals to negative it is possible. So, R square IV can be negative as well, but that does not mean that we will not apply R square.

Because we are not trying to maximize our R square rather, we are trying to get the unbiased and efficient estimates, unbiased and consistent estimates of beta one hat. But one thing we have to keep in mind since consistency is a large sample property instrumental variable estimation technique should be applied when we have a large sample. Otherwise in small sample the R square IV cannot give you the consistency and unbiasedness result.

And if the sample is very small the IV estimate should actually have more bias, small sample bias. These two things we have to keep in mind, that R square IV is consistent and unbiased when we have large sample. But in the small sample since consistency is a large sample property cannot be ensued in small sample. So, in small sample we must avoid this instrumental variable estimation technique.

Because it will produce the biased estimates and that bias is known as small sample bias. This is what you have to keep in mind while applying the instrumental variable estimation technique. There is one more thing that we must keep in mind.

(Refer Slide Time: 12:17)



That is 2 SLS estimates are subjected to multi-collinearity. This is also something we have to keep in mind, how to understand. Why 2 SLS estimates are subjected toward prone to multicollinearity? This is the reason. Let us say $Y_1 = \beta_0 + \beta_1 Y_2 + \beta_2 Z_1 + U_1$ and we could identify Z_2 and Z_3 let us say are excluded from the model. Then 2 SLS estimates means we need to first write the reduced form what is the reduced form $y_2 = \pi_0 + \pi_1 Z_1 + \pi_2 Z_2 + \pi_3 Z_3 + v_1$.

And from this what we will get we will get $\widehat{y_2}$ and this same y_2 that means and then we replace y_2 by $\widehat{y_2}$ in the structural equation. Now what is happening here? In the estimation of y_2 that means we can what is happening here $\widehat{y_2}$ is actually a function of Z_1. So, that means we estimate $\widehat{y_2}$ by

using Z_1 and the same Z_1 is used as an explanatory variable in the structural equation, y_2 is a function of Z_1 and Z_1 is actually an explanatory variable in the structural equation.

So, $\widehat{y_2}$ and Z_1 might be correlated, that is the problem. So, that means we may have multi-collinearity problem in the process of 2 SLS. But the possibility or severity of multi collinearity can be reduced, if Z_2 and Z_3 are strong. What I am saying? In the reduced form equation apart from Z_1 there are two additional explanatory variable which are Z_2 and Z_3. So, if Z_2 and Z_3 are very strong.

That means, if Z_2 and Z_3 are powerful instruments for y_2 then the severity of multicollinearity goes down because it is not only Z_1, but Z_2 and Z_3 are also there as a determinant of y_2. So, even though $\widehat{y_2}$ is a function of Z_1 since in the process of determination Z_2 and Z_3 also played a significant role we can say that yes, some degree of correlation is inevitable between Z_2 and $\widehat{y_2}$ and Z_1. But the severity of multicollinearity goes down if they are strong instrument.

So, Z_2 and Z_3 must be strong instruments for y_2 to avoid the severity of multi-collinearity problem. So, that means 2 SLS estimates are prone to multi-collinearity we cannot help because the process itself produces that multicollinearity. y_2 is estimated using the explanatory variable which are exogenous in the model here it is Z_1 there might be many other factors many other exogenous variables.

But as long as the excluded exogenous variables are strong then we can say that, there is collinearity between y_2 and Z_2. But Z_2 and Z_3 also played an important role, so severity of the collinearity goes down in this way that much we can say. So, with this we are closing our discussion on instrumental variable estimation technique today and rest of the things we will discuss tomorrow, thank you.