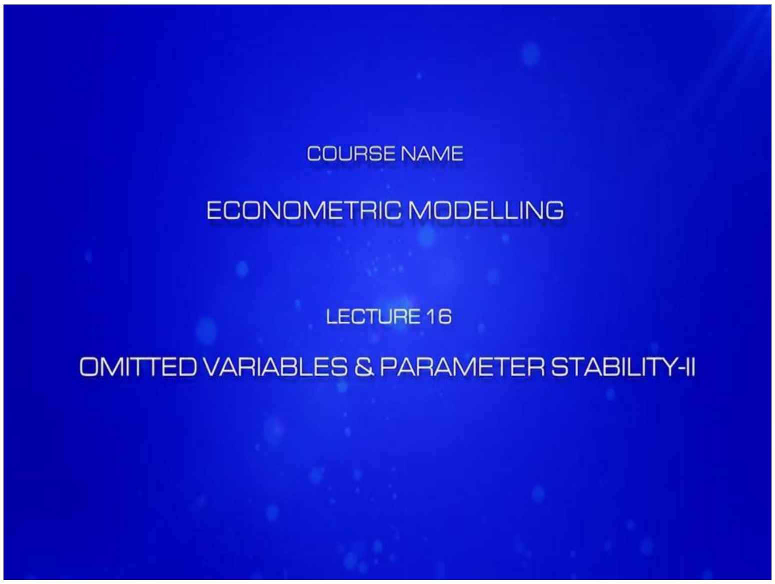
Econometric Modelling
Professor Sujata Kar
Department of Management Studies
Indian Institute of Technology, Roorkee
Lecture 16
Omitted Variables and Parameter Stability - II

(Refer Slide Time: 00:11)

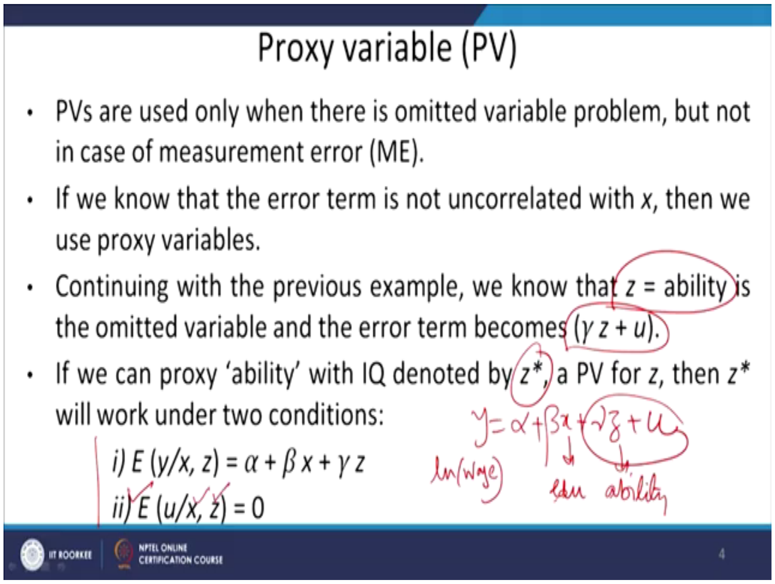


Hello everyone and welcome back to the course on Econometric Modelling. This is Module 16.

(Refer Slide Time: 00:44)

In module 15 we have discussed the problem of Omitted Variable; Module 16 also continues with that but once we have discussed model misspecification in terms of omitted variables and measurement error this module basically checks or suggests the corrections that need to be made in order to control this kind of model misspecification problems. So, first of all, we introduce proxy variables.

(Refer Slide Time: 01:02)

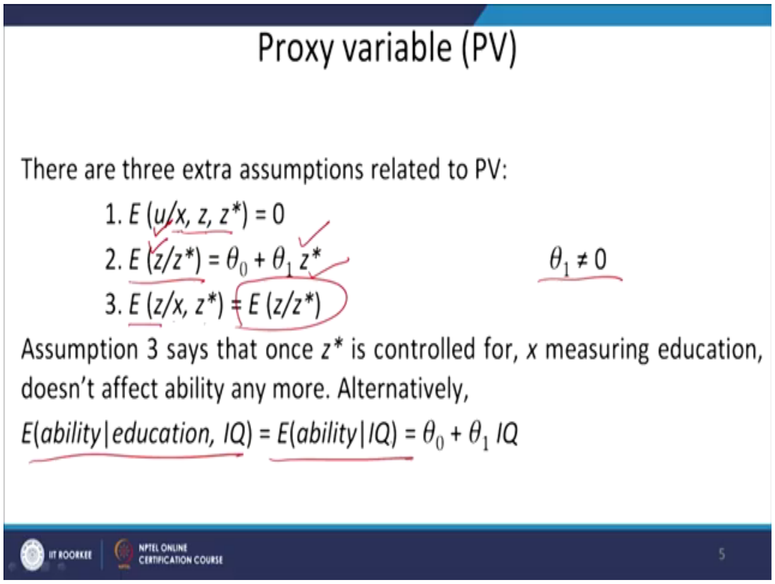


Proxy variables or PVs are used only when there is an omitted variable problem, but not in case of measurement error. If we know that the error term is not uncorrelated with x, then we use proxy variables. Continuing with the previous examples, that is the example that we discussed in the context of omitted variables that if we have an equation or if log which is dependent on education and ability and in case, we do not include the ability in the expression.

So, we know that$z = ability$, or if we denote ability by z this is omitted, and the error term becomes $\gamma z + u$. So, my original model was (refer slide time: 1:50). Now, z is not observed and that is why the entire thing becomes my error term. If we can proxy 'ability' with IQ scores say, and we denote it by z^*, which is a proxy variable for z, then z^* will work under two conditions.

So, first, this is my original model and the original model should fulfill this condition (refer slide time: 2:36). So, both original ability and education were uncorrelated to the error term.

(Refer Slide Time: 02:48)



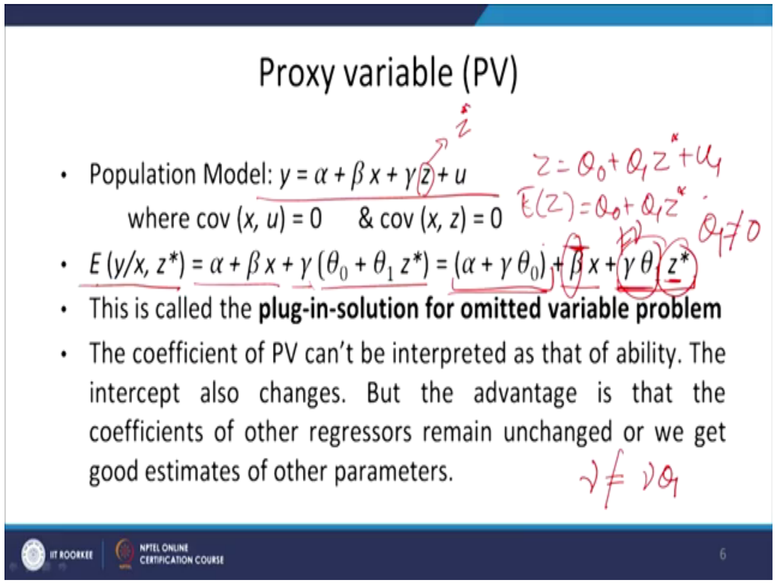
Now, to include the proxy variable, we need to have basically three additional assumptions. The first assumption is that (refer slide time 2:56).

So, the proxy variable additional to the x and z variables is also uncorrelated to the error term. The second assumption is that (refer slide time: 3:12). This implies that there is a linear relationship between the proxy variable and the variable which is omitted, that is here in this case ability and IQ score. So, IQ score explains ability in a linear format where $\theta_1 \neq 0$ because if $\theta_1 = 0$ then this implies that IQ score does not explain the ability.

So, θ_1 has to be non-zero in order for z^* to qualify as a proxy variable or PV. And the third assumption is that (refer slide time: 4:00). This assumption basically says that once z^* that is IQ score is controlled for education or x measuring education does not affect ability anymore.

Alternatively expected value of ability given education and IQ is equal to the expected value of ability given IQ, which is equal to (refer slide: 4:32). So, this actually may not seem very realistic or that may vary from situation to situation, but then this actually can be a close approximation to reality.

(Refer Slide Time: 04:49)



Now, this is our population model again, to begin with (refer slide time: 4:54).

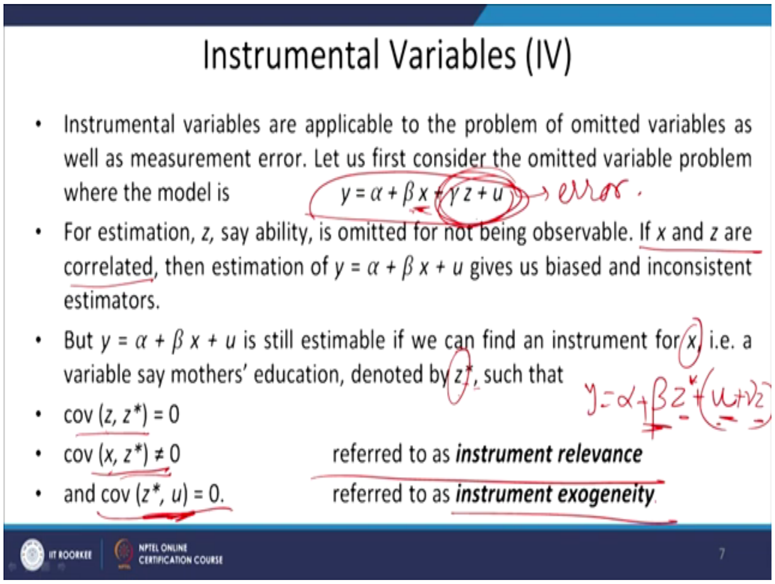
Now, (refer slide time: 5:02). So, I am taking the expected value of y, we are having the expected value of z equals to (refer slide time: 5:33) and that is why it is replaced here.

And then by collecting terms we have (refer slide time: 5:43). So, this is called the plug-in solution for omitted variables. The coefficient of a proxy variable can't be interpreted as that of ability. The intercept also changes, but the advantage is that the coefficients of other regressors remain unchanged, or we get good estimates of the other parameters.

So, we can see that β remains unchanged, though of course the parameter estimate of $z^* = \gamma$. And as a result of which because in order for z^* to qualify as a proxy variable θ_1 has to be non-zero and if θ_1 is non-zero then $\gamma \neq \gamma\theta_1$. As a result of which we do not have an estimate of ability given by the alternative proxy variable, which is IQ score.

And our intercept term also changes, but most often we are more concerned about the parameter estimates of the other regressors. So, the longer the other regressors tend to have unbiased and consistent estimates there is no problem, so proxy variables can be used. Now, we talk about the other tool used to take care of both omitted variables and measurement error problems and that is instrumental variables.

(Refer Slide Time: 07:17)



Instrumental variables are applicable to the problem of omitted variables as well as measurement error. Let us first consider the omitted variable problem, where this is our original model for estimation z say ability is omitted for not being observable. If x and z are correlated then the estimation of (refer slide time: 7:37) gives us biased and inconsistent estimators, because you can see that this will be our error term and the error term will be correlated with x because x and z are correlated.

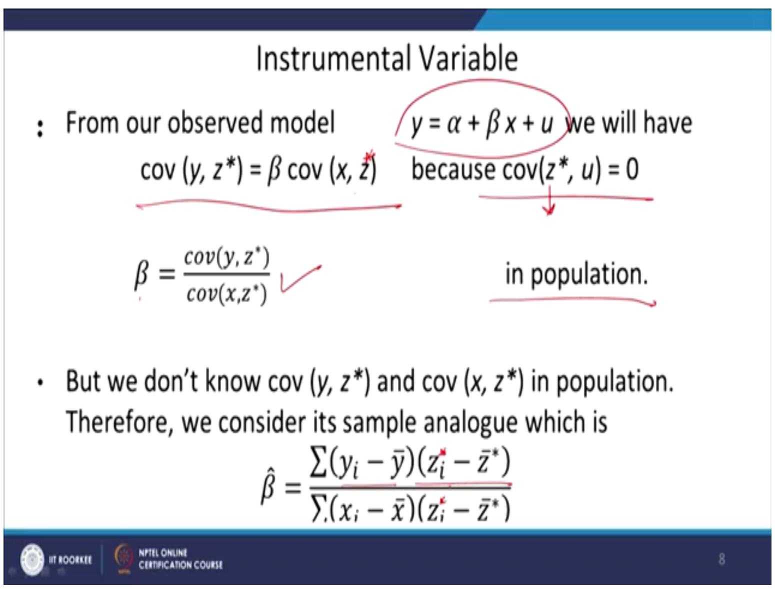
So, we cannot have unbiased and consistent estimators that have already been discussed in the previous module. But (refer slide time: 8:01) is still estimable if we can find an instrument for x that is a variable, say mother's education, so remember x was the education of the individual. Now, we are trying to find an instrument for the variable x. Suppose it is the mother's education and we denote it by z^* such that (refer slide time: 8:25).

What I want to mean is that this is my error term. So, when I am replacing x, instrumental variables replace the existing regressor. So, unlike proxy variables it does not have the original regressor, it completely replaces the original variable. So, now instead of x will be having (refer slide time: 8:53).

Now, of course, this z^* has to be uncorrelated to u that is our classical linear regression assumption then we must have z and z^* also uncorrelated because you remember that our error term contains the ability also. So, this is $u + \gamma z$, so it has to be uncorrelated to u it has to be uncorrelated to z in order to give unbiased and consistent estimators of β.

And we also must have x and z^* correlated because we are actually finding an instrument for x so unless and until x and z^* are correlated, thus purpose is not served. So, this is called instrument relevance and this assumption is called instrument exogeneity. So, the instrument must be exogenous, but the instrument also must be relevant; relevant in the sense of an instrument that can replace x. So, it is, if it is completely uncorrelated to x then it cannot be used as an instrument for x.

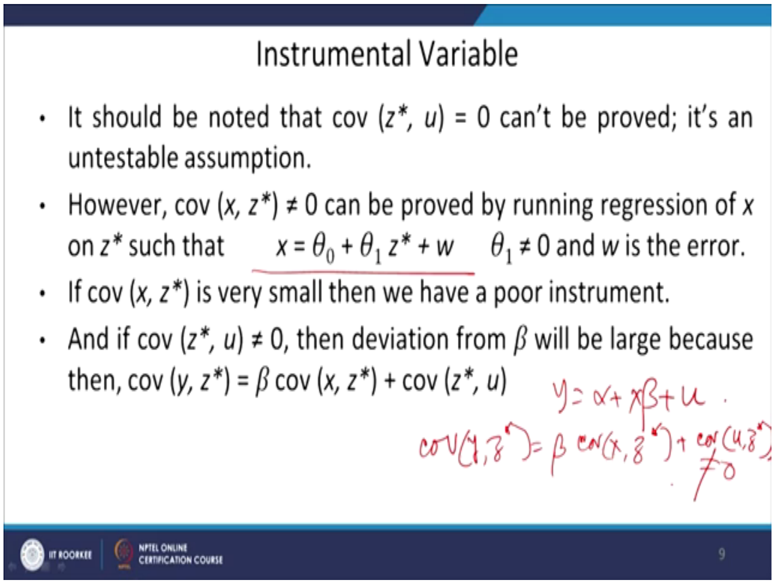
(Refer Slide Time: 10:09)



So, for the observed model (refer slide time: 10:13), we will have covariance between (refer slide time: 10:18), which is one of the classical linear regression assumptions that must the instrumental variable, z^*, fulfill.

As a result of which, we have (refer slide time 10:39- 11:10)

(Refer Slide Time: 11:10)



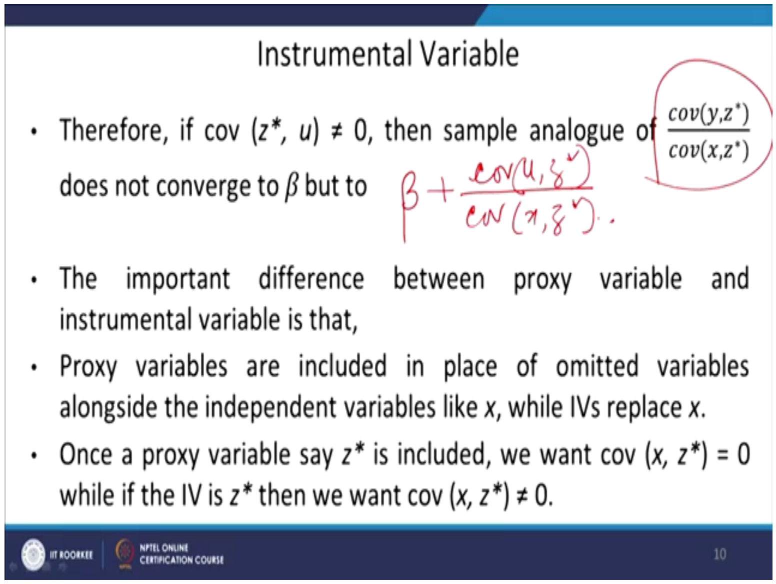
It should be noted that cov $(z^*, u) = 0$ cannot be proved, it is an untestable assumption, because u is population error and population error is not observed. So, whether cov $(z^*, u) = 0$ or not that cannot be proved. However, the covariance between cov $(x, z^*) \neq 0$ can be proved by running a regression of x on z^* such that x is a linear function of z^*.

And we have this you know the format, (refer slide time 11:45). Now, $\theta_1 \neq 0$ is important because otherwise, x will be completely unrelated to z^*. Now, if cov (x, z^*) is very small then we have a poor instrument.

If cov $(z^*, u) \neq 0$, the deviation from β will be large because then (refer slide time: 12:11).

Now, you can see that if this is not equal to 0 then what we are saying is that deviation from β the parameters will be large.

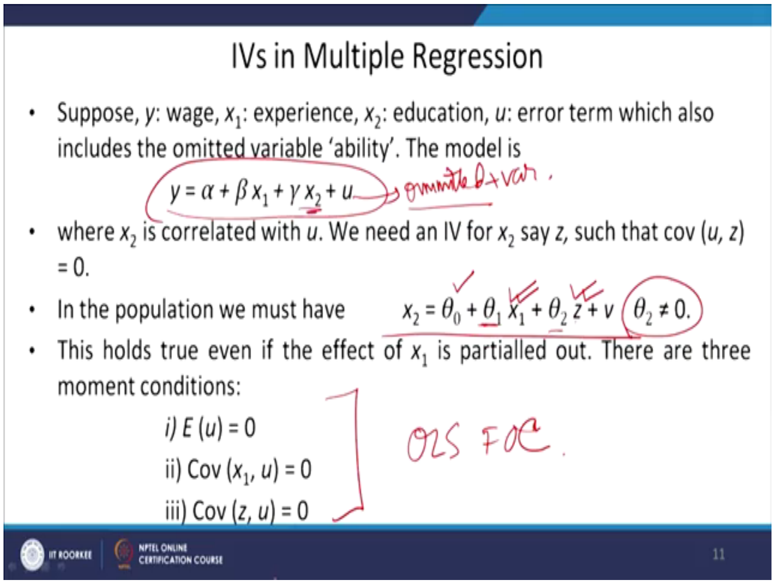
(Refer Slide Time: 12:55)



We are now going to prove that, therefore if cov $(z^*, u) \neq 0$t then the sample analog of (refer slide time: 13:06-13:21)

The important difference between proxy variables and instrumental variables is that proxy variable are included in place of omitted variables alongside the independent variables like x, while instrumental variables replace x. And second thing is that once a proxy variable say, z^*, is included, we want cov $(x, z^*) = 0$, while if the instrumental variable is , z^* then we want cov ($x, z^*) \neq 0$.

(Refer Slide Time: 13:53)

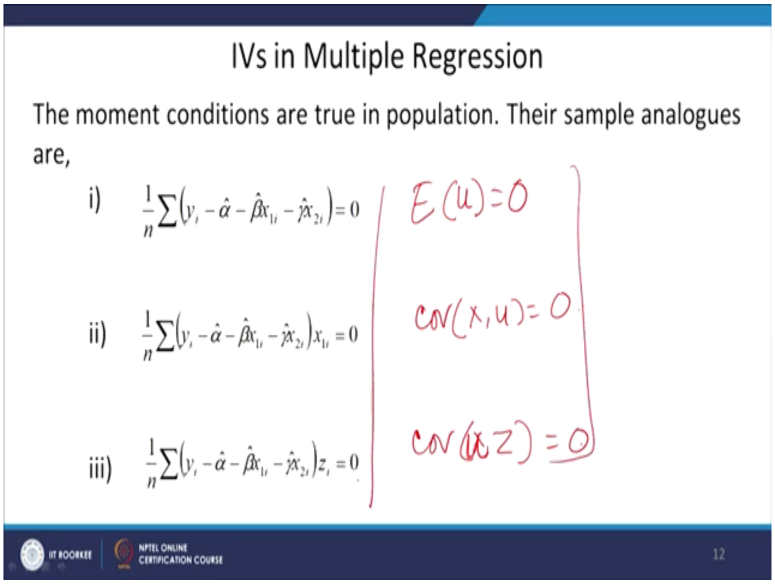


Now, we consider instrumental variables in multiple regression. Suppose y is wage or log of wage, x_1 is experience, x_2 is education and u is the error term which also includes the omitted variable ability. Therefore, this is our model where this model also includes the omitted variable. The error term also includes the omitted variable where $x_2$is correlated with you and we assume that $x_2$that is education is correlated with the omitted variable, which is the ability, and as a result of which x_2 is correlated to the error term.

We need an instrumental variable for x_2 say z such that covariance between u and z is equal to 0. In the population, we must have x_2 as a linear function of the existing regressor or the other regressor that is x_1 plus the instrumental variable z. So, it is not important whether we have this one equal to 0 or not, but this has to be not equal to 0, otherwise z cannot be an instrument for x_2.

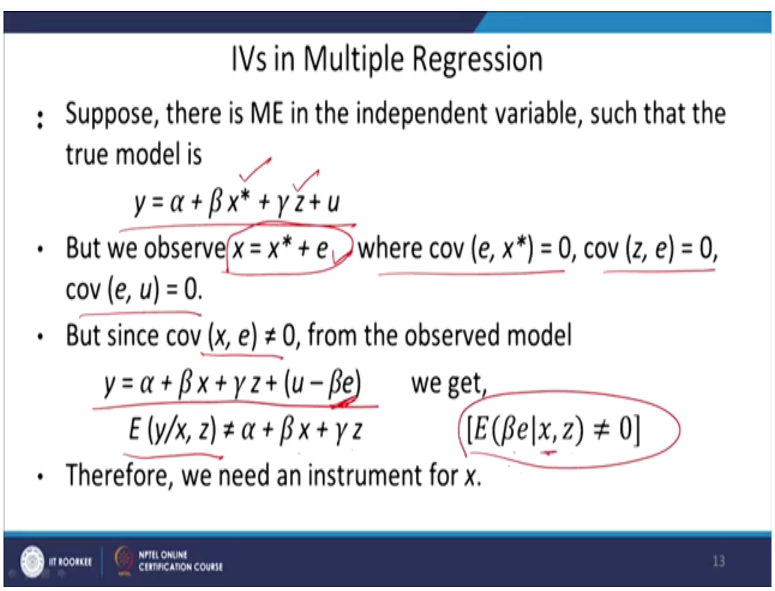
This holds true even if the effect of x_1 is partialled out. There are 3-moment conditions. Moment conditions are equivalent to OLS first-order conditions. So, the 3-moment conditions we have because we have one constant term, one independent variable x_1, and one independent variable z or instrumental variable z. So, as a result of which we have 3-moment conditions (refer slide time: 15:32).

(Refer Slide Time: 15:40)



So, the moment conditions are true in the population their sample analogs are given by this. So, this is equivalent (refer slide time:15:50- 16:07). So, these are basically the three-moment conditions and their sample analogs.

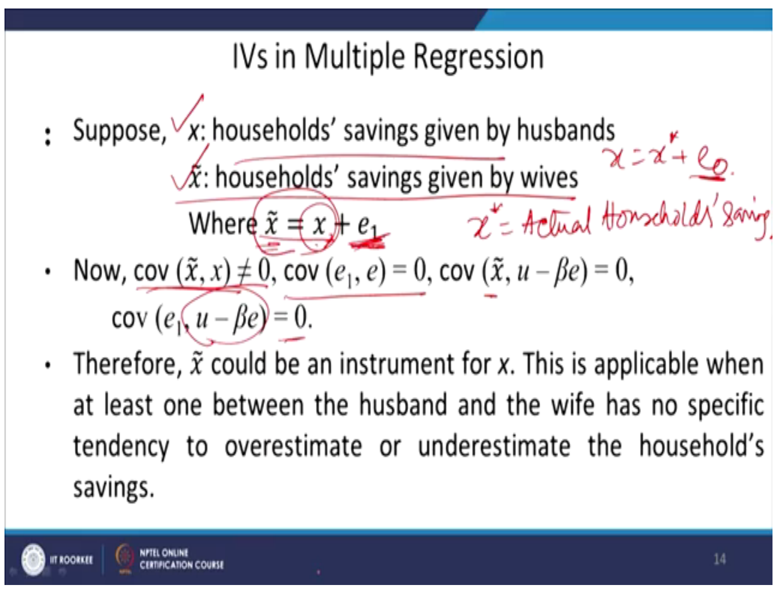
(Refer Slide Time: 16:15)



Suppose there is a measurement error in the independent variable such that the true model (refer slide time: 16:19). So, so far, we were talking about omitted variable problems, now we will talk about the use of instrumental variables in the context of measurement error. So, if measurement error is there in the dependent variable, but it fulfills the assumptions of the classical linear regression model then we have noticed that there is actually no problem because the estimates are unbiased and consistent.

Now, the problem is there if we have a measurement error in the independent variable. So, we observe (refer slide 16:57), is my original value or variable. (Refer slide time 17:07-17:54). Therefore, we need an instrument for x.

(Refer Slide Time: 17:56)

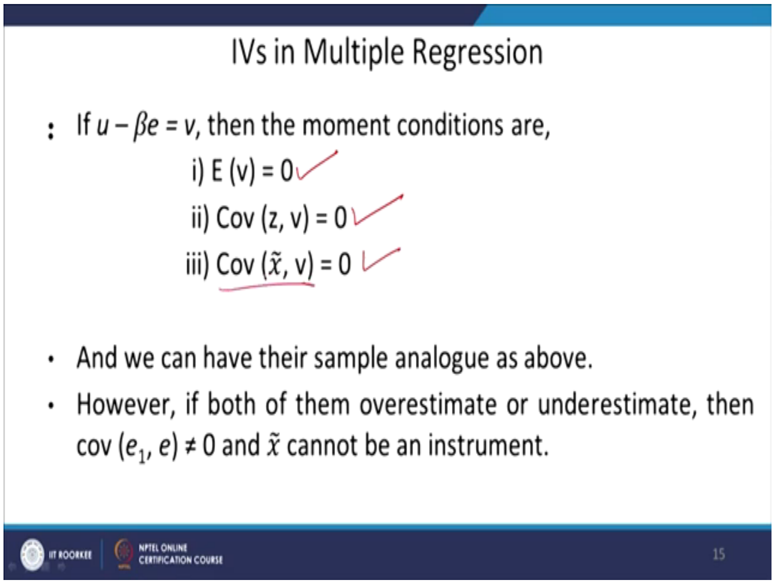


Suppose x is household savings given by husbands and x delta is household savings given by wives, (refer slide time: 18:05). So, now you remember that x^* is actual household savings. So, one such measure we were taking from the husbands, which were denoted by x and (refer slide time: 18:29). So, some errors are reported by the husbands. Now, if we consider household savings given by the wives then we can see that with respect to husbands' estimates there is some error that is wives' and husbands' estimates of household savings actually differ.

Now, (refer slide time: 18:54-19:28), which we observed with respect to our household savings given by the husbands.

Therefore, (refer slide time 19:34-19:45). So, it should not be equal to 0, and given this relationship it is not equal to 0. This is applicable when at least one between the husband and the wife has no specific tendency to overestimate or underestimate the household savings.

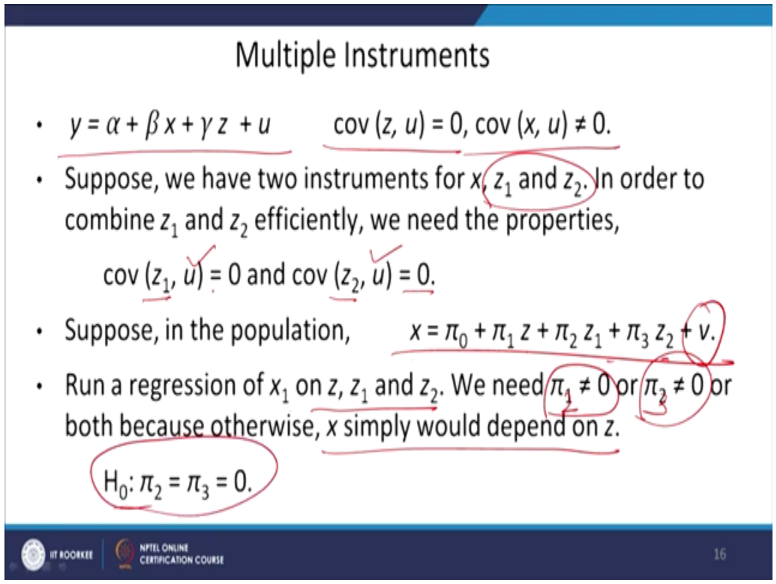
(Refer Slide Time: 20:00)



If (refer slide time 20:01- 20:19). So, the moment conditions are fulfilled in the population this implies that the assumptions of the classical linear regression model are fulfilled. And we can have their sample analogue as previously shown.

However, if both of them overestimate or underestimate then (refer slide time: 20:38) and this violates the assumption of a classical linear regression model.

(Refer Slide Time: 20:52)

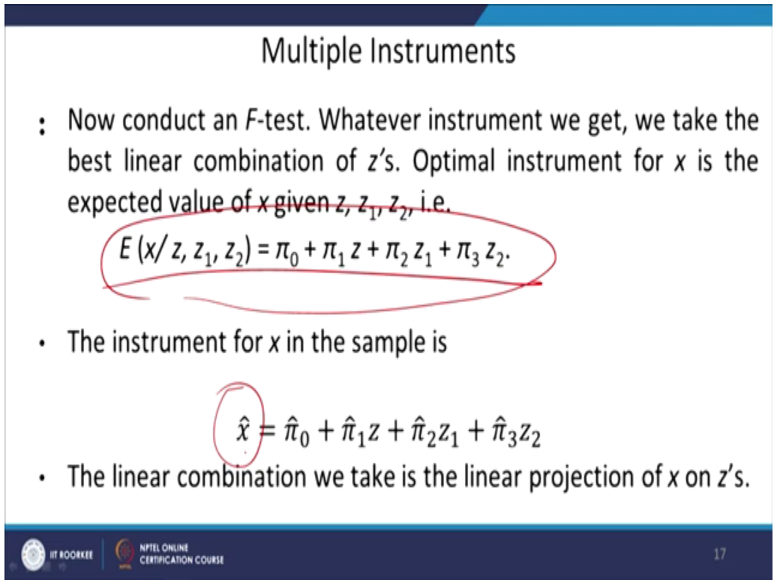


Now, we talk about multiple instruments, so instead of having one instrument, we can also have multiple instruments. So, if this is our original model while (refer slide time: 21:02) that could be because of measurement error or omitted variable problem.

Now, suppose we have two instruments for x and they are denoted by z_1 and z_2. To combine z_1 and z_2 efficiently, we need the properties that both the instruments z_1 and z_2 are uncorrelated to u so that their covariance with respect to u is 0.

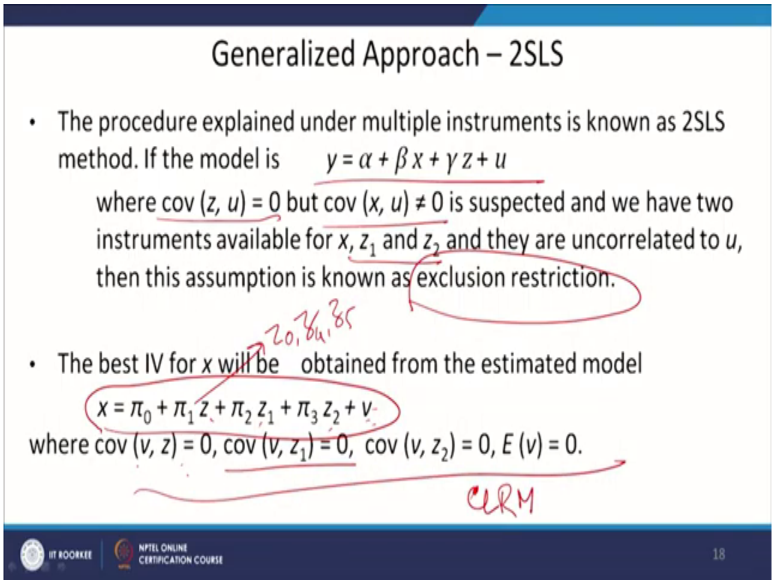
Suppose in the population, we have this linear relationship between x and the instruments and the other regressor. So, (refer slide time: 21:45- 22:24).

(Refer Slide Time: 22:22)



Now, conduct an F-test. Whatever instrument we get we take the best linear combinations of z's. The optimal instrument for x is the expected value of x given z, z_1 , and z_2 that is the expected value of x given z, z_1 , and z_2 we have (refer slide time: 22:42). The instrument for x in the sample is the estimated value from this equation or rather from this equation, the estimated value of x is denoted by x tilde. So, this is our instrument for x in the sample, the linear combination we take is the linear projection of x on the z's.

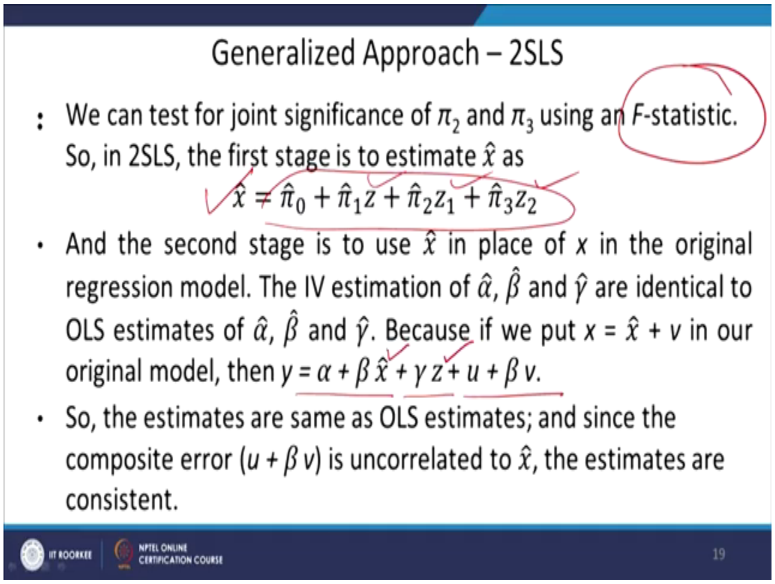
(Refer Slide Time: 23:08)



Now, we talk about a generalized approach, which is also known as 2SLS or 2-stage least squares. The procedure explained under the multiple instruments is actually known as the 2-stage least squares or 2SLS method. So, if the model is (refer slide time: 23:24) then this assumption is known as exclusion restriction.

The best instrumental variables for x will be obtained from the estimated model of this that is where x is expressed as a function of the existing regressor z or the other regressor z, z_1 and z_2 are the two instrumental variables. If there are more regressors like z, if we have more regressors like z_0, z_1, $z_2$then all of them should have been included here. Now, where (refer slide time: 24:17). So, this implies that we must have the CLRM assumptions fulfilled.

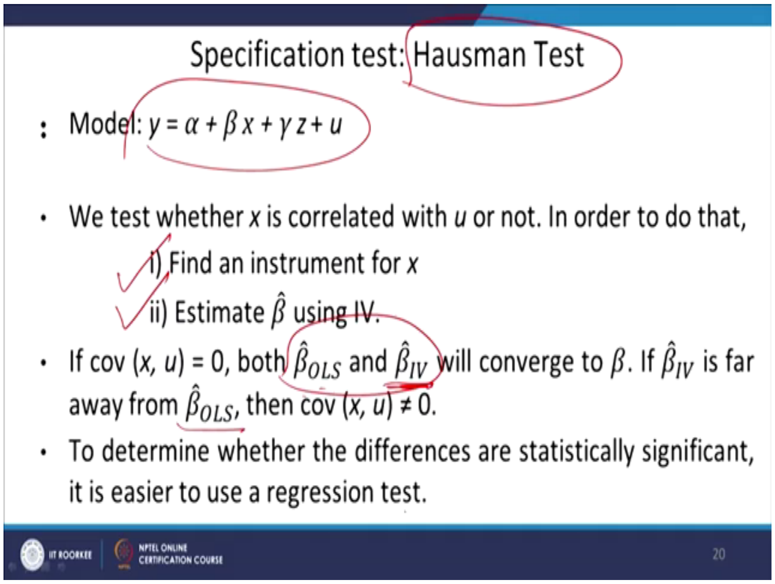
(Refer Slide Time: 24:34)



We can test for joint significance of π_2 and π_3 using an F-statistic. We have not talked about F-statistic at length till now. So, this is just a testing procedure that will be dealt with at length in the next part. In 2SLS, the first stage is to estimate $\hat{x}$ that is to estimate x as a function of z, z_1, and z_2, and the second stage is to use $\hat{x}$ in place of x in the original regression model.

The IV estimation of (refer slide time: 25:05- 25:22).So, the estimates are the same as OLS estimates. Since the composite error (refer slide time: 25:22) the estimates are also consistent.

(Refer Slide Time: 25:34)

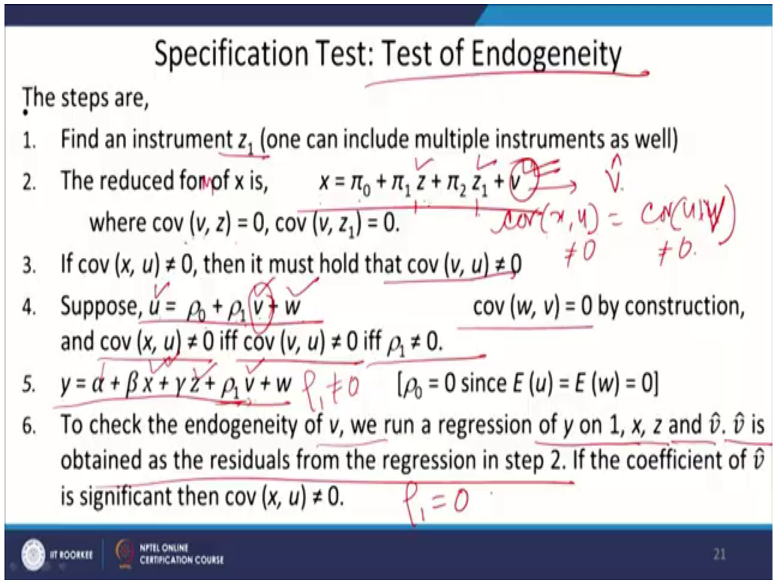


Now, we talk about the specification test. So, this talks about finding out whether we need an instrument or not. So, this is our original model. We test whether x is correlated with u or not. So, if x is not correlated with u, we do not have any problem we do not need an instrument. In order to do that, first of all, find an instrument for x and estimate $\hat{\beta}$ using instrumental variables.

If (refer slide time: 26:04- 26:25). So, if the estimate from 2SLS is the same as the estimate from an OLS, then there is no correlation between x and u, if that is not the case then there is a correlation between x and u.

So, this kind of test was suggested by Hausman and that is what is called the Hausman Specification Test. To determine whether the differences are statistically significant or not it is easier to use a regression test and what we next discuss is a test of endogeneity.

(Refer Slide Time: 26:52)



So, the steps involved are to find an instrument z_1 one can include multiple instruments as well, so there can be z_1, z_2 , and z_3. The reduced form of x is (refer slide time: 27:11- 28:09).

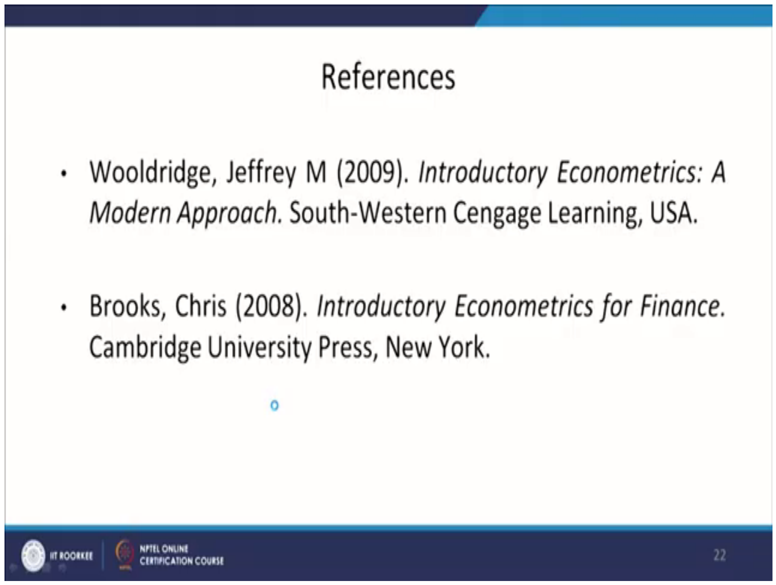
So, suppose we assume a linear relationship between the error terms, so (refer slide time: 28:16- 28:50). So, what we are going to test here is that whether (refer slide time 28:52-
29:19).

What we do is that to check the endogeneity of v we run a regression of y on (refer slide time:29:26). So, instead of v, we are having $\hat{v}$ which is the estimated counterparts of v, $\hat{v}$ is obtained as a residual from the regression in step 2. So, we first run a regression of (refer slide 29:44) collect the residuals that is $\hat{v}$, and then this $\hat{v}$ is included in our original model so that I regress y on 1, x, z and $\hat{v}$.

If the (refer slide time: 29:56- 30:16) So, these are actually alternative ways of finding out whether there is endogeneity problem or not whether we need an instrument for x or not.

So, one was given by Hausman which basically states that we can check for the significance in the difference between $\hat{\beta}$ OLS and $\hat{\beta}$ IV. Alternatively, I can also test for the significance of the coefficient ρ_1 in this kind of test of endogeneity.

(Refer Slide Time: 30:47)



So, these are the references. Thank you.