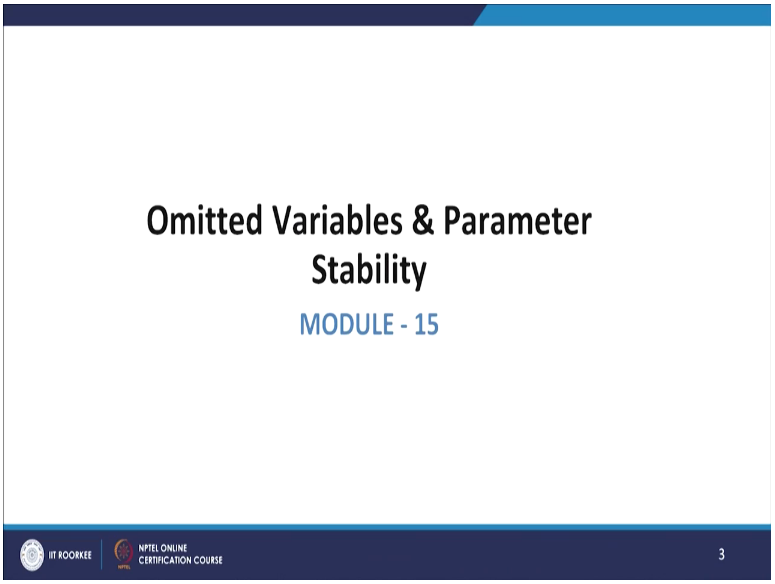
Econometric Modelling
Professor Sujata Kar
Department of Management Studies
Indian Institute of Technology, Roorkee
Lecture 15
Omitted Variables and Parameters Stability - I

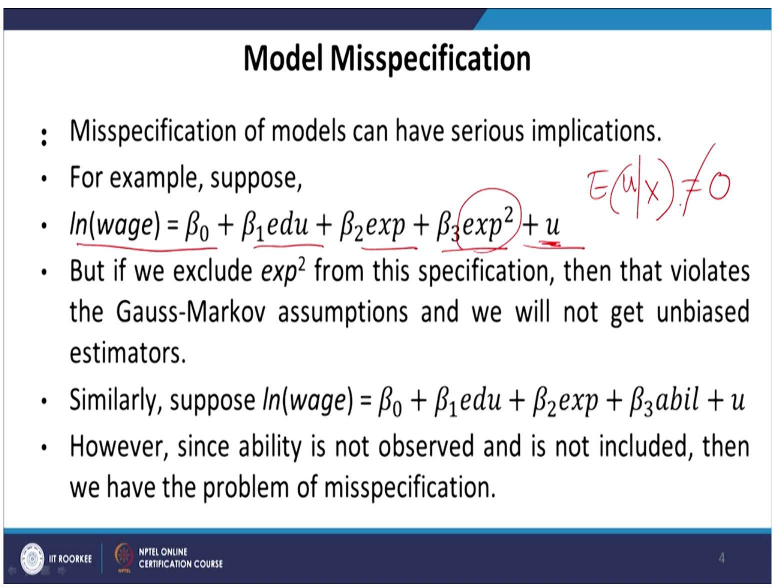
Hello and welcome back to the course on Econometric Modelling. This is Module 15 and it deals with Omitted Variables and Parameters Stability.

(Refer Slide Time: 00:36)

Module 15 and 16 are both devoted to the problem of basically model misspecification. So, in model 15, I present two types of model misspecification, omitted variable problems and measurement errors. While in module 16 we will discuss how to deal with such problems that if variables have been omitted or there is a measurement error, then what are the steps required to be taken in order to correct for those kinds of model misspecifications.

(Refer Slide Time: 01:14)



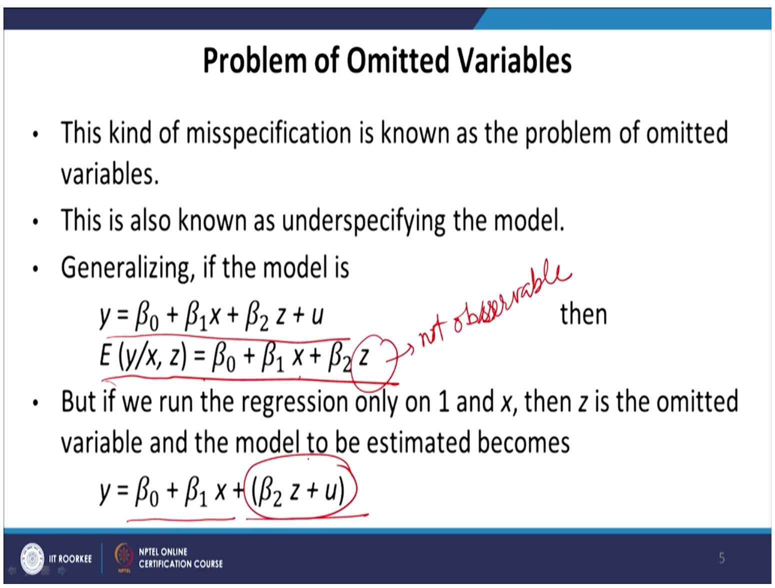


Misspecification of models can have serious implications. For example, suppose this is our model of the world *(refer slide time:01:14)* that is the logarithm of wages regressed on education experience and experience square plus there is an error term, u in the population.

So, this is the population specification of the model. But if we exclude experience square from this specification, then that violates the Gauss Markov theorem assumptions and we will not get unbiased estimators because you can see, in that case, experience square will be a part of the error term and as a result of which the assumption that expected value of u conditional upon all sorts of values of x that is all functional forms of x equals to 0 is actually not valid anymore.

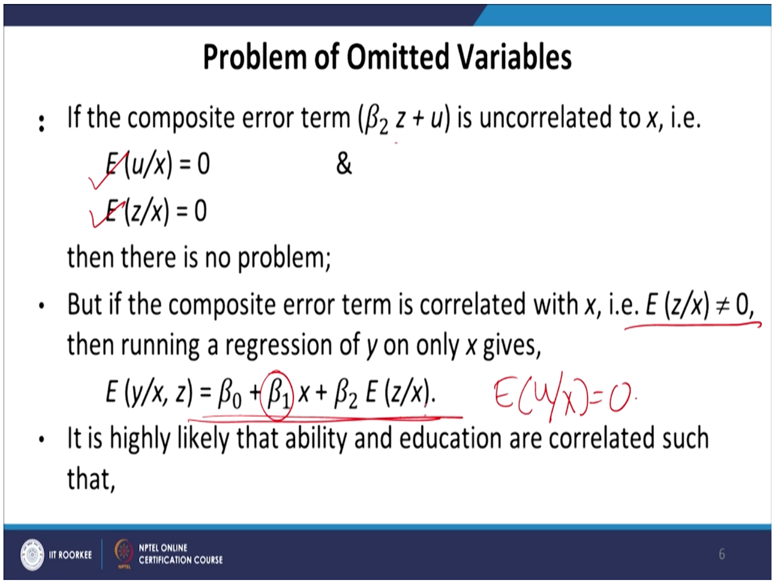
So, similarly, suppose the logarithm of wage is equal to beta naught plus beta 1 education plus beta 2 experience plus beta 3 ability plus an error term. However, since the ability is not observed and is not included, then we have the problem of model misspecification. So, the ability is actually difficult to measure. So, it is quite possible that we can exclude this variable or it will not be included in the expression.

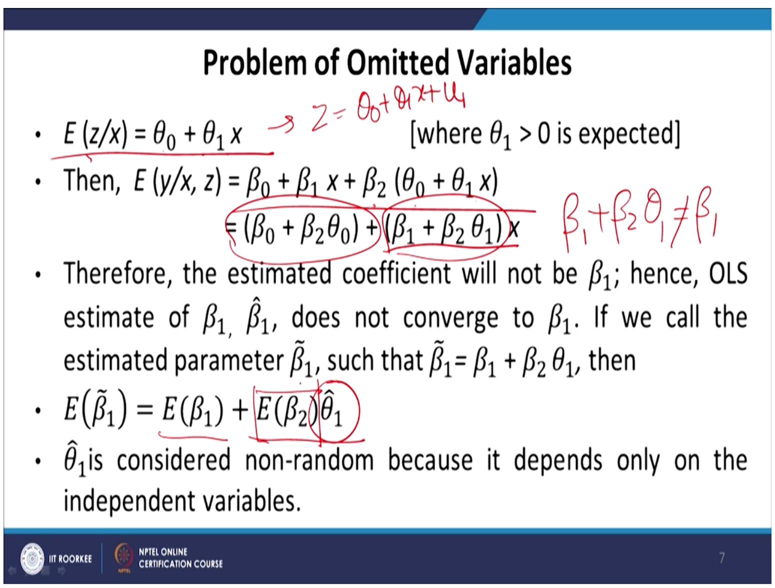
(Refer Slide Time: 02:39)



This kind of misspecification is known as the problem of omitted variables, this is also known as underspecifying the model. So, generalizing it, if the model is y equals beta naught plus beta1 x plus beta 2 z where z is ability then we have expected value of y given x and z is this that is beta naught plus beta 1x plus beta 2 z but if we run the regression only on 1 and x because z is not observable and z is actually the this is not observable. So, z is the omitted variable then the model to be estimated becomes beta naught plus beta 1 x plus beta 2 z plus u so this becomes my total error term.

(Refer Slide Time: 03:32)





If the composite error term beta 2 z plus u is uncorrelated to x, that is expected value of u given x is 0 and the expected value of z given x is 0, then there is absolutely no problem we still get an unbiased estimator of beta 1. But, if the composite error term is correlated with x such that the expected value of z given x is not equal to 0, that is, if ability and education are correlated with each other, then running a regression of y on x only gives this expression.

So, we will be left with beta 2 expected value of z x because the expected value of z x is not equal to 0, we still assume that the expected value of u given x is 0, but since the expected value of z is given x is not zero (*see the slide above*). So, this is actually my original expression, it is highly likely that ability and education are correlated such that the expected value of z given x equals to theta naught plus theta 1x. So, we are assuming a linear relationship between z and x such that the expected value of z given x is equal to theta naught plus theta 1x.

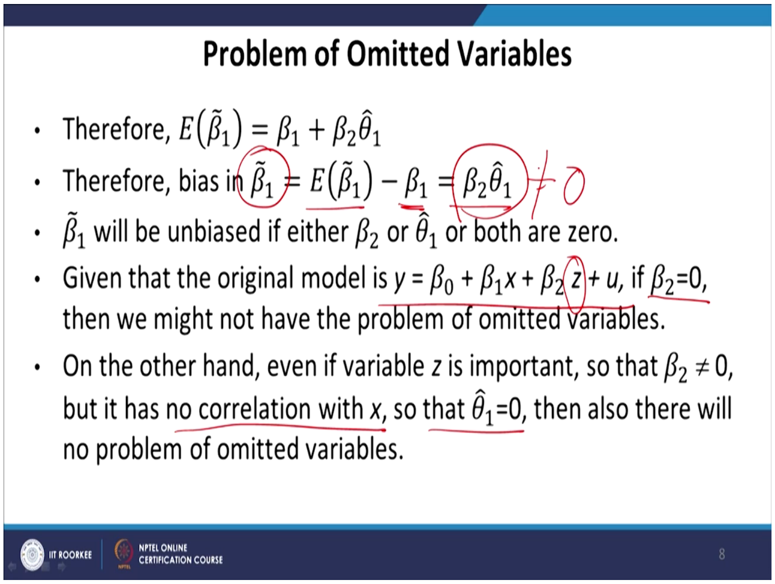
Alternatively, this can be written as Z is equal to theta naught plus theta 1x plus u or here we can denote the error term by u1 say, in order to distinguish it from the original model error term. So, now here theta one is greater than zero that is what is expected, because, we can say that as education increases ability also increases. In that case expected value of y given x and z will be, replacing the inner expression for expected values of z given x with theta naught plus beta 1x.

So, beta 2 into theta naught plus theta 1x, expanding this is the intercept term and this is the slope term. So, in case I have omitted a variable that is correlated with the existing regressor that is x in that case, the coefficient estimate of the regressor that is existing regressor, that is x is actually not equal to beta 1. So, beta 1 plus beta 2 theta 1 and this is not equal to beta 1 *(refer slide time 03:32).*

So, that is the problem that we will get in case we have an omitted variable and the variable is actually correlated with the existing regressor. Therefore, the estimated coefficient will not be beta 1. Hence, the OLS estimate of beta 1 that is beta 1 hat does not converge to beta 1. If we call the estimated parameter beta 1 tilde such that beta 1 tilde is equal to beta 1 plus beta 2 theta 1 then the expected value of beta 1 tilde is equals to the expected value of beta 1 plus the expected value of beta 2 theta 1 hat.

Now, theta 1 hat is taken or kept outside the expected expression, because it is considered to be non-random as it depends only on the independent variables and we replace theta 1 with it estimated counterpart that is theta 1 hat.

(Refer Slide Time: 07:11)

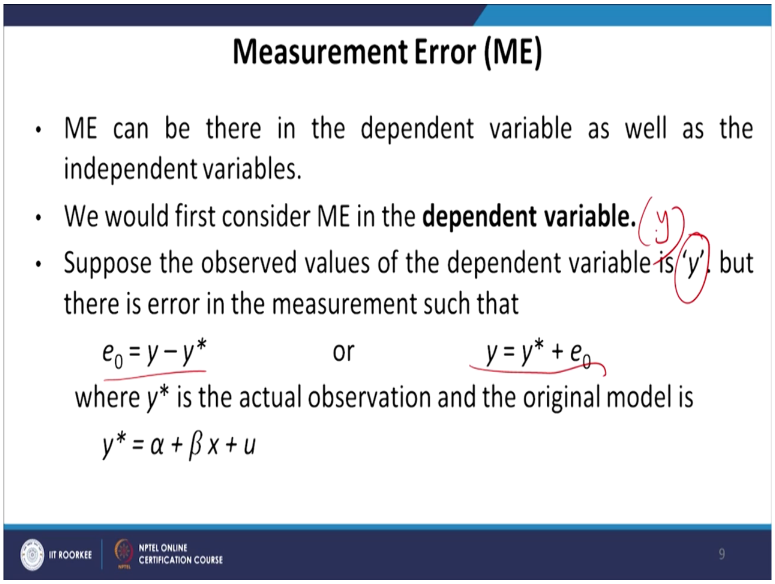


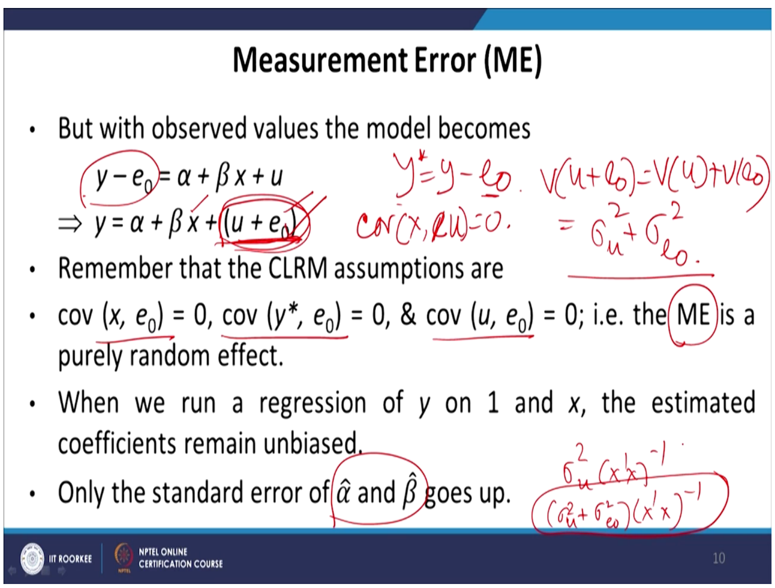
Therefore, the expected value of beta one tilde is equal to beta 1 plus beta 2 theta 1 hat therefore, the bias in beta 1 tilde is actually the expected value of beta 1 tilde minus beta 1 which is beta 2 theta 1 hat *(refer slide time: 07:11)*.

So, we can see that beta one tilde will be unbiased if either beta 2 or theta 1 hat or both are 0, given that the original model is y equals beta naught plus beta 1 plus beta 2 beta 1x plus beta 2 z plus u if beta 2 is equal to 0, then we actually do not have an omitted variable problem altogether, because in that case, it shows that z is actually not explaining why. So, the problem of omitted variables is not there.

On the other hand, even if variables that are important, beta 2 is not equal to 0, but that z and x are not correlated i.e. z do not have any correlation with x. So, that theta 1 hat is equals to 0 then also there will be no problem of omitted variables. But, if a variable is actually omitted, and it is correlated with the existing regressor x then the estimated parameter beta 1 tilde is actually a biased estimator. Because in that case, our beta 2 theta 1 hat will not be equals to 0. This was the problem of the omitted variable.

(Refer Slide Time: 08:43)





Now, we talk about the problem of measurement error another kind of misspecification i.e. measurement error. Measurement error can be there in the dependent variable as well as in the independent variables, we would first consider measurement error in the dependent variable that is variable y. Suppose, the observed value of the dependent variable is y, but there is an error in the measurement such that E naught equals y minus y star or y equals y star plus e naught this is the same thing written right by rearranging terms *(refer slide time :08: 43; first slide).*

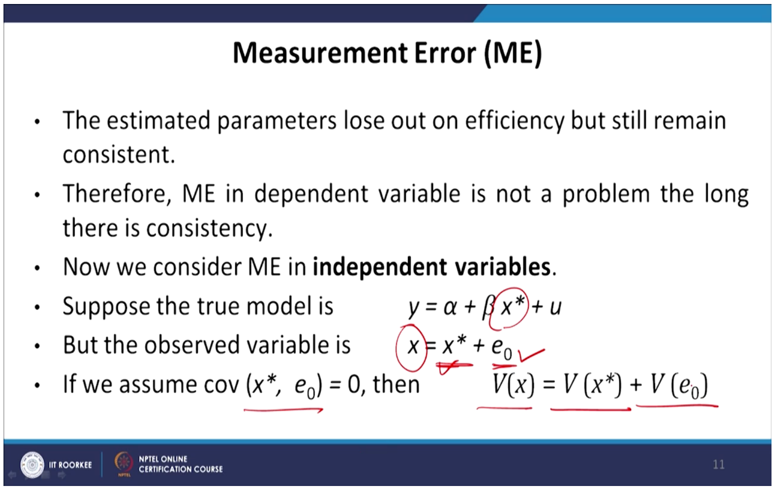
So, the thing is that y star is the actual observation, but we are observing the only y. So, the original model is y star equals alpha plus beta x plus u, but what we are observing? We are observing y. So, since y star is equal to y minus e naught, that is why we have y minus e naught equals alpha plus beta x plus u which implies alpha y equals alpha plus beta x plus u plus e naught. So, now, I have a greater error term or larger error term.

Remember that the classical linear regression model assumptions are covariance between x and e naught should be 0. So, the covariance between x and u is anyway assumed to be 0, the covariance between x and u are assumed to be 0. Now, when we have an extended error term, so, additionally we must have covariance between x and e naught equals to 0.

Covariance between y star and e naught is 0, the covariance between e; and u and e naught is also 0 which implies that the measurement in the measurement error in the dependent variable is purely a random factor or random effect. When we run a regression of y on 1 and x, the estimated coefficients remain unbiased if the measurement error is actually completely random, only the standard error of alpha hat and beta hat goes up (*refer slide time :08: 43; second slide)*.

Now, since you know population error has gone up, population error variance is also supposed to go up even if u and e naught are uncorrelated with each other. The variance of u plus e naught will be a variance of u plus the variance of e naught covariance between u and e naught is 0. We can write as sigma squared u plus sigma squared e naught.

Now, if you remember that, the variance of the estimated parameters is sigma squared u into x prime x inverse. Now, in the case of this situation, I have sigma square u plus sigma square e naught into x minus x prime x inverse. So, now, you can see that the error variance has actually gone up. (Refer Slide Time: 11:53)



The estimated parameters lose out on efficiency but still remain consistent. Therefore, measurement in the dependent variable is not a problem, as long as there is consistency. Now, we consider measurement error in the independent variables, suppose, our true model is y equals alpha plus beta x star plus u, but what we observe is x which is x star plus e naught *(refer slide 11:53)*. So, again there is a measurement error in the independent variable. Instead of observing x star, we are observing x, which is the original value plus some error term added to it. Now, there could be many situations of this kind of measurement error.

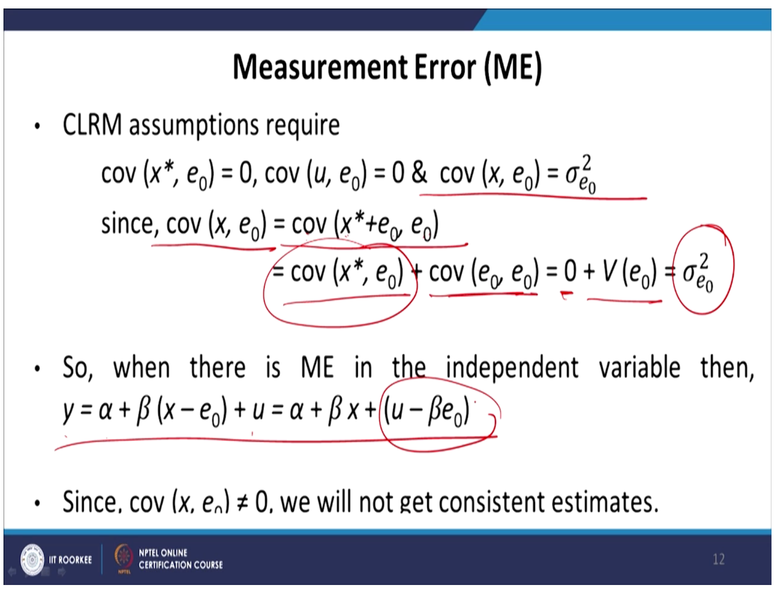
And measurement errors are many times completely unintentional. We can take the example of household savings reported by individuals. So, individuals can be different household members. The household saving reported by husbands can be different from those reported by wives and can be also reported by other family members. Similarly, we can talk about things which are reported by individuals, for example, the number of days ill in a particular year.

So, the number of days actually ill could be different from the number of days reported by an individual and that difference could be a completely random term. It is possible that I remained ill for say 10 days in a year, but instead of mentioning it at 10 days, I reported it to be 20 days. If it is intentional, then it is intentional, not only for me but for a large number of people. If we are biased towards exaggerating it, or if we have a tendency to report a lower number of days, in both cases, we would see a systematic error.

And in that case, that error is actually not completely random. But the long errors are completely random, which is also possible that I forgot actually how many days I remained ill. So, in that case, that is probably not much of a problem at least in the case, when we have a measurement error in the dependent variable.

Now, considering the case of measurement error in the independent variable, so, this is my observed model or observed variable which is x star plus e naught *(refer slide 11:53)*. If we assume that covariance between x and e star, x star and e naught which is 0, then the variance of x is equal to the variance of x star plus the variance of e naught. Again we do not have a covariance term coming out.

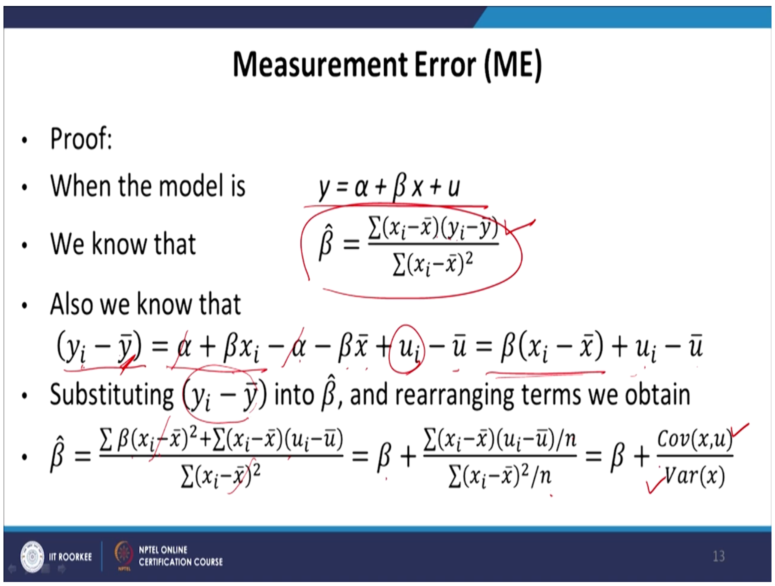
(Refer Slide Time: 14:40)

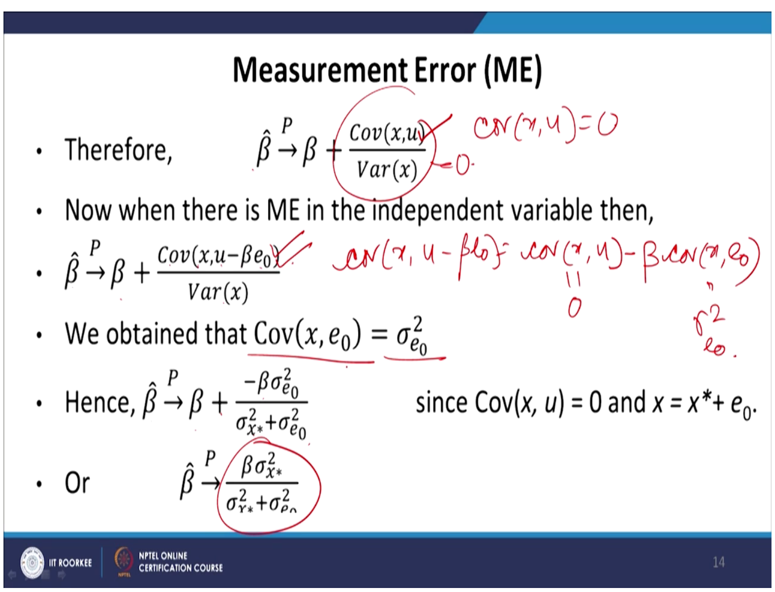


So, the classical linear regression model assumptions required covariance of x star and e naught to be 0; covariance between u and e naught to be 0 and covariance between x and e naught is equal to sigma e naught squared, why this is so? Because you can see that covariance of x and e naught is basically covariance x star e naught and e naught. Now, I expand it, we have covariance x star and e naught, covariance e naught, e naught and covariance x star *(refer slide time 14:40*).

Now covariance x star e naught is 0 and covariance e naught, e naught is actually variants of e naught, if we denote it by sigma square e naught so, that is how we have covariance x e naught is equal to sigma square e naught. So, when there is a measurement error in the independent variable, then this is our model i.e. y equals alpha plus beta x minus x naught plus u, and my error term which is u minus beta e naught *(refer slide 11:53)*. Since covariance x e naught is not equal to 0 we will not get consistent estimates.

(Refer Slide Time: 15:53)





So, this is proof now, that when the model is y equals alpha plus beta x plus u, that is the standard model, then we know that this is the expression of our beta hat, which is basically if I write it in matrix format, it would be x prime x inverse x prime y. Otherwise, this is summation x i minus x bar into y i minus y bar divided by summation x i minus x bar whole square (*refer to slide time 15:53; first slide).* This has been already derived and mentioned several times so far.

Also, we know that y i minus y bar is equal to alpha plus beta x i, plus u i minus alpha minus beta x bar minus u bar corresponds to this y. Now, by connecting terms, we can see that alpha-alpha cancels out, I collect beta. So, beta has beta x i minus x bar plus u i minus u bar *(refer slide 15:53; first slide)*. Now, if I substitute this expression into the expression for the beta hat and rearrange terms, then you can see that, first of all, I will be having beta into x i minus x bar whole square that is the expression is put here plus x i into x bar into u y minus u bar *(refer slide 15:53; first slide)*.

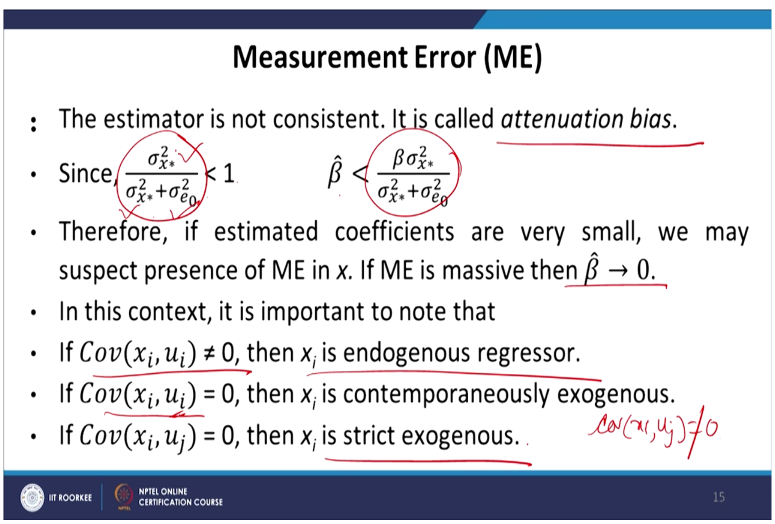
The denominator remains as it is, these two things cancel out *(refer slide 15:53; first slide).* I have beta here plus the entire thing divided by the denominator. Now, I am dividing both the numerator and the denominator by n. And what I am getting? In the numerator, I have covariance x u and in the denominator, I have a variance of x. Therefore, in the probability therefore, beta hat in probability limit tends to beta plus covariance x u divided by variance of x.

Now, when there is a measurement error in the independent variable, then we have beta hat with probability limit tending to beta plus covariance between x and u minus beta e naught because now, my error is not u, it is u minus beta e naught. You can see that the covariance between x and u equals to 0 this term is equal to 0 and the beta hat in probability limit tends to beta. So, we say that the estimates are consistent*(refer slide 15:53; second slide)*.

Now, we need to see whether this term tends to 0 or not or this term is equal to 0 or not. So, we know that covariance between x and e naught is sigma e square, e naught square. So, hence beta hat in probability limit tends to beta plus we have covariance between x and u equal to 0. We can expand this, that is the covariance between x u minus beta e naught as the covariance between x and u minus beta covariance between x and e naught right this is 0 and this is equal to sigma square e naught *(refer slide 15:53; second slide)*.

Therefore, we have in the numerator minus beta sigma u naught square divided by variance of x was observed to be sigma x squared plus sigma e naught squared. Therefore, probability beta hat with probability limit tends to rearranged term and it is not actually tending to beta.

(Refer Slide Time: 19:34)



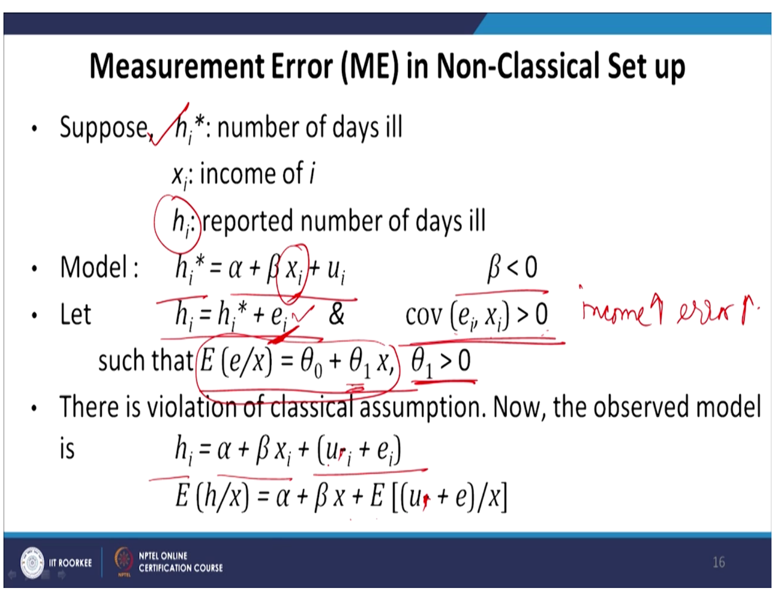
The estimator is not consistent and it is called attenuation bias. Since this expression is less than 1 because you can see all of them are positive numbers and the numerator is smaller than the denominator. So, we have this expression less than 1 which implies that beta hat will be less than this *(refer slide time: 19:34)*. Therefore, if estimated coefficients are very small, the beta hat is actually less than beta multiplied by a fraction. So, if estimated coefficients are very small, we may suspect the presence of measurement error in x.

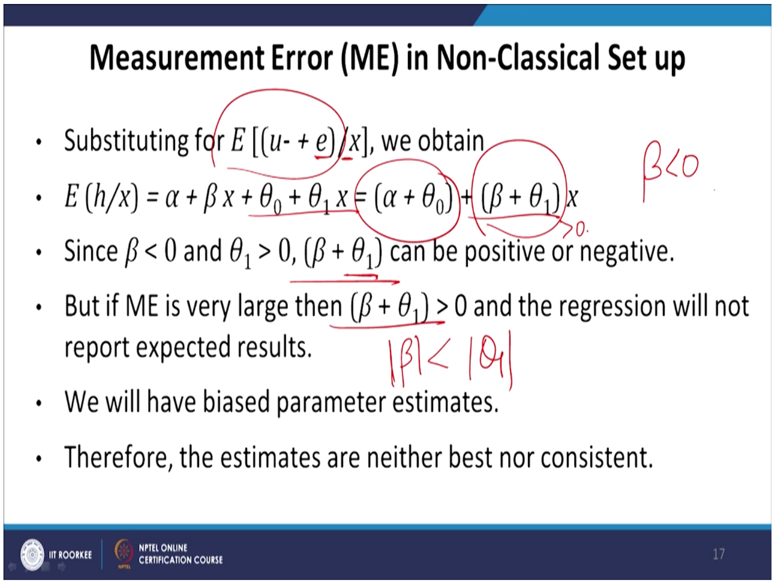
If the measurement error is massive, then the beta hat pretends to 0 because this number actually becomes very small. In this context, it is important to note that if covariance between x i and u i is not equal to 0, then x i is endogenous regressor or we call x i to be an endogenous regressor. Endogenous is something that is explained by the system. So, generally we assume independent variables to be non-random, or they are independent in the sense that the model is not trying to explain them.

But if they are correlated with the error term, then this implies that the model contains additional information that may explain the variations in x as well. That is why we call them endogenous regressors. If xi and ui are equal to 0 or covariance between xi and ui equals to 0, then xi is said to be contemporaneously exogenous that is this is a situation where covariance between x i and u j not equal to 0, only this is equals to 0.

So, in that case we say that x i is contemporaneously exogenous and if covariance xi and uj are 0. That is all the xi's and all the ui's for any i and j is 0, then x is strict exogenous that is the model does not explain or contain any information that may explain the independent variable x.

(Refer Slide Time: 21:48)





Now, we talked about measurement errors in the non-classical setup. Suppose hi star is the number of days ill, I will explain why we call it a non-classical setup. Now, xi is the income of i. And hi is the reported number of days an individual remains ill *(refer slide 21:48)*.

So, the model is actually the number of days ill is explained as a function of the income of individual i. When we expect the beta to be less than 0, that is, as income increases, the number of days a person remain ill actually reduces, but this is actually the reported number of days ill which is different from or which could be different from the actual number of days the person remains ill.

And suppose this is how they are related. So, h i is equals to hi star plus ei. And now, we are assuming that the covariance between this error term and income, are correlated, which implies, if this is positive, then this implies that as income increases the tendency or the error actually increases, and as income increases, the error also increases such that expected value of E given x is theta naught plus theta 1x where theta 1 is greater than 0 *(refer slide 21:48; first slide)*.

So, we are expecting a positive correlation and that is why a positive parameter estimate. If we have a linear relationship between e and x, the error here and the independent variable there is a violation of classical assumption. So, that is why this is called non classical setup, because now, the observed model is hi alpha plus beta xi plus ui plus ei, where this xi and ui are correlated.

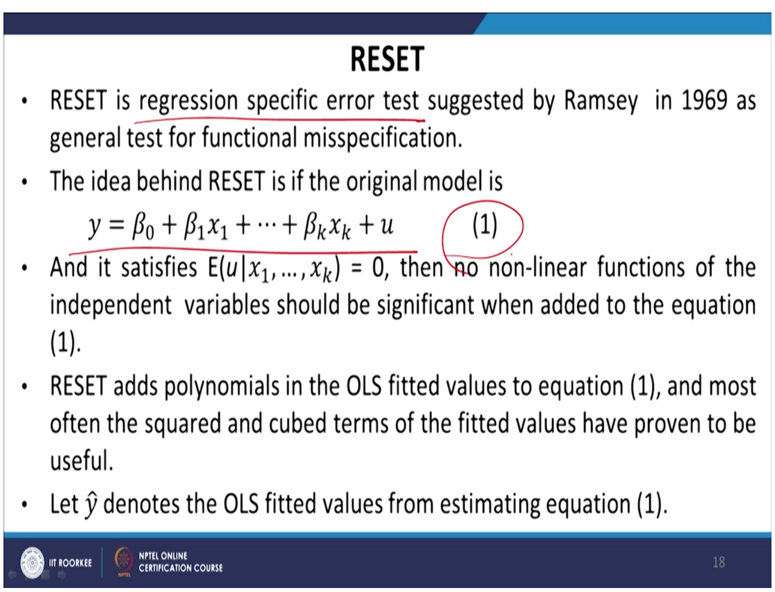
So, the expected value of h given x equals alpha plus beta x plus the expected value of u plus e given x.

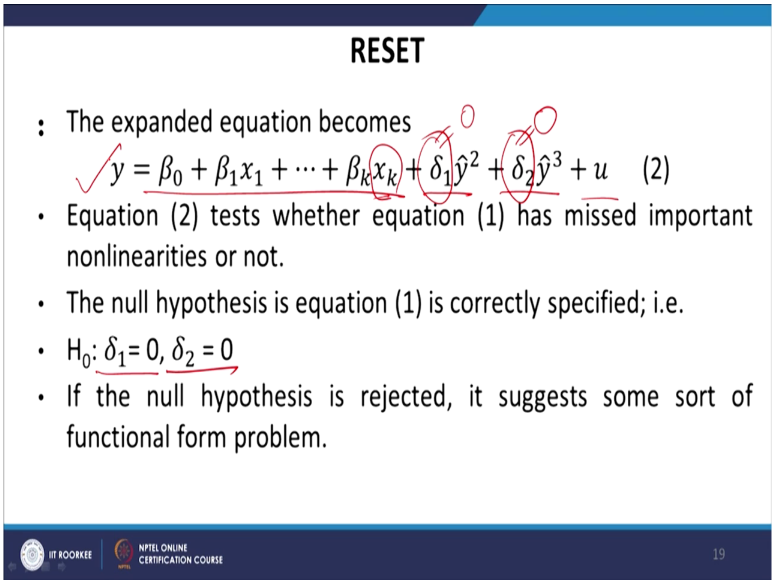
Substituting expected value of u plus e given x we obtain the expected value of h given x equals alpha plus beta x plus theta naught plus theta 1x. So, this expression, this is the linear relation we obtain between e and x and that is what is being replaced here as a result of which this is my constant term and this is my parameter of x *(refer slide 21:48; second slide)*.

Now, this parameter of x is actually not equal to beta. Since beta is expected to be less than 0 and theta one is greater than 0. Beta plus theta 1 can be either positive or negative. But if the measurement error is very large, then we expect theta 1 to be large. And as a result of which mod value of beta could be less than the mod value of theta 1 and as a result of which beta 1, beta plus theta 1 will be greater than 0.

And the regression will not report expected results because originally we expected beta to be less than 0 if we observe a coefficient of x which is greater than 0, then this is actually not fulfilling our expectations following the theoretical argument or for what we have initially postulated. So, we will have biased parameter estimates. Therefore, the estimates are neither best nor consistent. So they have neither minimum variance nor consistency.

(Refer Slide Time: 25:31)





Now, how do we test whether there is misspecification or not? RESET is regression specific error test. RESET was suggested by Ramsey in 1969 as a general test for functional misspecification. The idea behind reset is that, if the original model is $y = \beta_0 + \beta_1 x_1 + \ldots\ldots \beta_k x_k +$ u, I consider a general model with k plus 1 variable and it satisfies the expected value of u given x 1 to x k equals to 0 that is there is no correlation between the error term and the independent variables, then no nonlinear functions of the independent variables should be significant when added to the equation 1 *(refer slide time 25:31; fisrt slide)*.

So, now, we are basically talking about or trying to find out a model misspecification where the model could have some nonlinear functions of the independent variables. That is an example, we gave in the beginning that there could be experienced quite in the specification also. So, whether that experience square has been missed out or not could be given by reset. Reset adds polynomials in the OLS fitted values to equation 1, this is our equation 1 *(refer slide time 25:31; fisrt slide)*.

And most often the squared and cubed terms of the fitted values have proven to be useful. So, that is why instead of considering a large number of polynomials, it at the max considers the squared and cubed terms of the independent variables or the fitted values. Let y hat denotes OLS fitted values from estimates obtained from equation 1, the expanded equation then becomes the original equation plus squared term of the fitted values plus a cubic term of the fitted values plus the error term.

So, equation 2 tests whether equation 1 has missed important nonlinearities or not, the null hypothesis in equation 1 is correctly specified as delta 1 equals 0 and delta 2 equals 0 *(refer slide time 25:31; second slide).* So, we are actually focusing on these two parameter estimates. If these two parameter estimates are equal to 0, then there is actually no model misspecification in terms of exclusion of nonlinear functional forms of the independent variables, existing independent variables. If the null hypothesis is rejected, it suggests some sort of functional form problem i.e. some sort of functional or nonlinearities in one or more independent variables would have been present in the model specification. But if the null hypothesis is not rejected, then probably we do not have any problem.

(Refer Slide Time: 28:21)

So, these are the references. Thank you.