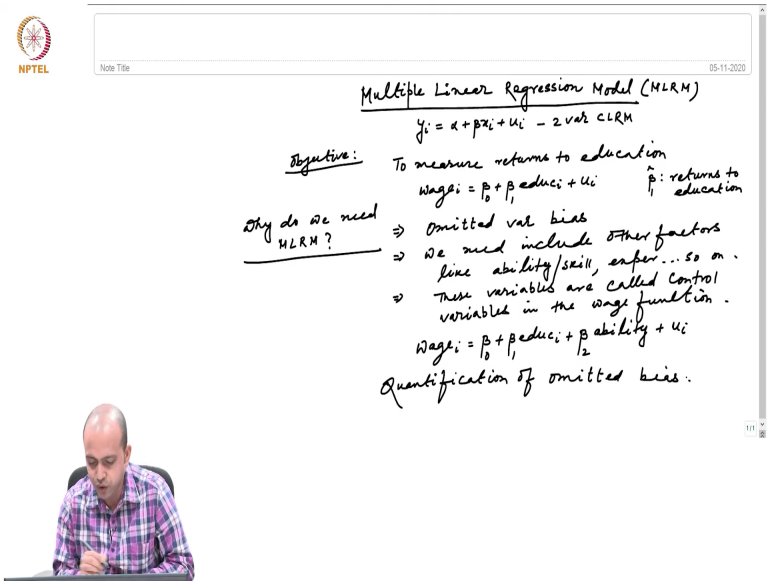
Introduction to Econometrics
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology Madras
Application of STATA for hypothesis testing and introduction to multiple linear regression model Part - 2

(Refer Slide Time: 00:14)



Welcome once again. Today, we will discuss about multiple linear regression model. Now, so far, whatever econometric models we have discussed, they are called 2 variable linear regression model. What does it mean? That means our model was something like this – $Y_i = \alpha + \beta x_i + u_i$. This is called 2 variable linear regression model or you can write 2 variable CLRM (classical linear regression model). Why this is called 2 variable? Because you have one dependent variable and one independent variable.

And as for example, you can think of y is your consumption expenditure, that you are going to explain by your income. Now, let us say that today we will take another example where our objective is to let us say returns to education on wage. What is our objective? Objective is to measure returns to education.

So, basically what is the impact of education on your wage or salary. So, let us say that our model is now, wage is dependent variable for the i^{th} person equals to sum beta 0 plus beta 1 education plus this. This is our objective, to estimate returns to education. So, obviously here

your beta hat would tell you, for a unit change in education, what would be the additional change in wage on an average.

So, beta hat will give you returns to education. Beta hat will tell you returns to education, beta 1 hat, this is called returns to education. This is the model we have learnt so far. This is also a 2 variable linear regression model kind of thing because you have one dependent variable, one independent variable. Now, when we think about multiple linear regression model, first of all we need to understand why do we need multiple regression model.

So, why multiple linear regression model? Why do we need multiple linear regression model? This is in short MLRM multiple linear regression model, so why do we need MLRM? That is the question we need to answer first. Why there is a need to understand multiple linear regression model? Now, when our objective is to estimate education, estimate the impact of education on wage and if we specify this type of model, then what we assume that the individuals are heterogeneous only in terms of education and they are homogeneous in terms of all other aspects.

That is why it is only due to a variation in education, we observe the variation in individuals' wage. But in reality, individuals are not only, individuals are heterogeneous, not only in terms of their education but also, many other factors. For example, their ability or skill, their age, their experience, so on an so forth.

So, when individuals are different in so many other aspects, assuming individuals are different only in terms of education, is quite a restricting assumption. That means even if our objective is to estimate the impact of education on wage, we need to include other factors as well in the model, so as to control the impact of other factors.

Otherwise, your model would be mis specified. And as we have told earlier, model misspecification is a kiss of death. That is how econometricians, they call it. And when individuals are different in terms of, let us say, their ability or skill, education, age, so on an so forth, if we do not include those important variable in the model as control factor, then our model will suffer from omitted variable bias. So, our model will suffer from omitted variable bias.

To avoid this omitted variable bias, even if our objective is to estimate only the impact of education on wage, we need to include other factors as control variable in our model, that is what

something we have to keep in our mind whenever we estimate any kind of econometric model. So, this is like, in pure or natural science, when scientists, they conduct experiment, what they do?

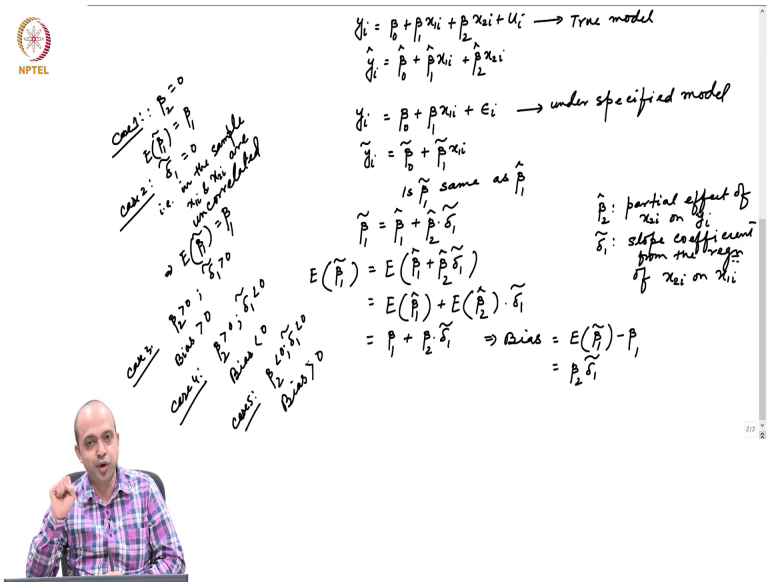
They do the experiment in a laboratory, and in that laboratory setup, impact of other factors are kept constant. For example, if the scientists' objective is to estimate the impact of temperature on gas volume, in the laboratory setup, the scientist will keep the impact of pressure on gas volume, impact of pressure and other factors constant.

Because that is a controlled environment. So, natural or pure, natural experiment are conducted in a controlled environment. You have a control, you know how to control those factors. But when we are doing econometric analysis, basically we are analyzing behavior of the individual. And here, the data what we get, it is not experimental data coming out from a laboratory setup. Rather, what we get is observed data.

So, that means there is a difference between experimental data and observed data. Since we do not have control over what would be the individual's age, what would be the individual's skill, what would be the individual's test and preference so on and so forth. When we are analyzing observed data, we need to include other factors also in the model to avoid omitted variable bias which is not the case in pure science when we conduct natural experiment because there other factors are automatically controlled in the controlled laboratory setup.

So, to avoid omitted variable bias, we need to include other factors as well, other factors like ability or skill, then experience so on and so forth. And these variables are called controlled variables in the wage function. So, that means, let us say that we, our true model would be then wage = $\beta_0 + \beta_1$education + $\beta_2$ability + u_i. This is our true model. We have another variable, ability factor, that should also be included in the model. If we do not include, then our model will suffer from omitted variable bias. Next, the question is how do you quantify this omitted variable bias? So, quantification of omitted variable bias, how do you quantify?

(Refer Slide Time: 11:33)



So, for quantification of omitted variable bias, we will write our true model like this. Let us say $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$ is our true model and if you estimate the true model, you will get $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$. So, that means in this model, our wage is determined by x_{1i} and x_{2i} where x_{1i} is education and x_{2i} is ability. Now, let us suppose, because of ignorance or data non-availability, we specify a model which looks like = $\beta_0 + \beta_1 x_{1i} + \varepsilon_i$. This is some, this model we call under specified model. Why this is under-specified?

Because in the true model, we require x_{2i} also, but because of our ignorance, we do not know that ability also matters. Or even if you know, sometimes it is very difficult to observe ability and get measurable data. How do you measure individual's ability? Which is something, some inert factor, right?

Individual's ability is something which is inbuilt and you cannot observe and collect observable data on ability. So, because of that, you might have omitted this variable and you have estimated this type of model yi equals to beta 0 hat, instead of hat, we will put a different notation, let us say tilde. Since this model is under-specified, estimate from the under-specified model is denoted by tilde, plus beta 1 tilde, x1i.

Then the question is, what is the relationship between $\hat{\beta}_1$ and β_1 tilde? Are they same or different? So, that means the question here is, whether, question here is, is β_1 tilde same as $\hat{\beta}_1$?

Now, from the relationship from this $\hat{\beta}_1$ and β_1 tilde, the way we have specified, we can write a simple relationship between these two where β_1 tilde is basically equals to $\hat{\beta}_1 + \hat{\beta}_2 * \delta_1 tilde$

And $\hat{\beta}_2$ is partial effect of x_{2i} on y_i and $\delta_1 tilde$ is basically slope coefficient from the regression of x_{2i} on x_{1i}. That means you can think of $\delta_1 tilde$ is basically the sample correlation between x_{1i} education and x_{2i} which you have actually excluded from your model.

So, if you take expectation of β_1 tilde, what you will get? Expectation of β_1 tilde would be then expectation of $\hat{\beta}_1 + \hat{\beta}_2 * \delta_1 tilde$. This we can write as expectation of beta 1 hat plus expectation of beta 2 hat into delta 1 tilde because why we are not taking expectation? Because delta 1 tilde is the sample correlation between x_{1i} and x_{2i}, that is why it would become expectation of $\hat{\beta}_2$ into this.

And assuming that beta 1 hat is an unbiased estimate of beta 1 and bet 2 hat is an unbiased estimate of beta 2, what we can write? This would become beta 1 plus beta 2 into this. Then what is the bias? This implies the bias, this implies bias equals to expectation of beta 1 tilde minus beta 1, is the amount of bias which is equals to beta 2 into delta 1 tilde.

So, that means the question what we were asking earlier, whether $\beta_1$tilde is same as $\hat{\beta}_1$, they are actually not the same. Because look at here, expectation of $\hat{\beta}_1$ equals to $\beta_1 + \beta_2 * \delta_1 tilde$. Now, when they will become, that means when there would be no bias? So, that means the bias would be 0. We can discuss 2 cases.

Case 1 - let us say that β_2 is actually 0. If β_2 is 0, what does it mean? That means ability variable does not actually qualify to be included in the model itself, that means for the population, I hypothesize ability or skill has no impact on wage. If that is the case, then β_2 equals to 0. And obviously, in that case what will happen? Expectation of β_1 tilde equals to β_1. So, that means β_1 becomes an unbiased estimate of β_1, which is quite natural. If some variable does not qualify to

be included in the model, why should I include that? Obviously, if you exclude that variable also, it does not matter.

Case 2 is important or interesting. Let us say, $\delta_1 tilde$ equals to 0. That means, in the sample what does it mean? That means, in the sample, x_{1i} and x_{2i} are uncorrelated.

So, that means even though in the population we hypothesize that some x_{2i} has some impact on y_i and the sample what we collect, we see that education and ability, they are actually not correlated for the particular sample we collected, then also, we can ensure that expectation of $\beta_1$tilde equals to β.

So, under these two cases, either β_2 equals to 0 or $\delta_1 tilde$ equals to 0, then only we can say that expectation of $\beta_1$tilde equals to β, that means the bias is 0. However, in reality, these two cases, they do not arise. Rather, we get β_1 positive and $\hat{\delta}_1$ also not equals to 0. And if that is the case, we can arrive at alternative cases.

Case 3 both are positive, that means the bias amount, if that is case, in this case bias is also positive. Case 4, β_2 is positive but $\hat{\delta}_1$ negative. Then bias is, bias would be negative. And case 5, β_2 is negative and $\delta_1 tilde$ is also negative. So, bias is greater than 0. So, these are the alternative cases that we can have.

So, this particular example makes us understand the importance of omitted variable bias. So, that means, even if our objective is to estimate the impact of education on wage, we should include other relevant factors also like ability. We have taken only 1 example. If you have many other variables like age, experience, we have to include those variables also otherwise again your model will suffer from model misspecification.

So, this example nicely explains the importance of omitted variable bias and that is how they justify the need for a multiple linear regression model even when our research interest is, we can understand multiple linear regression model.