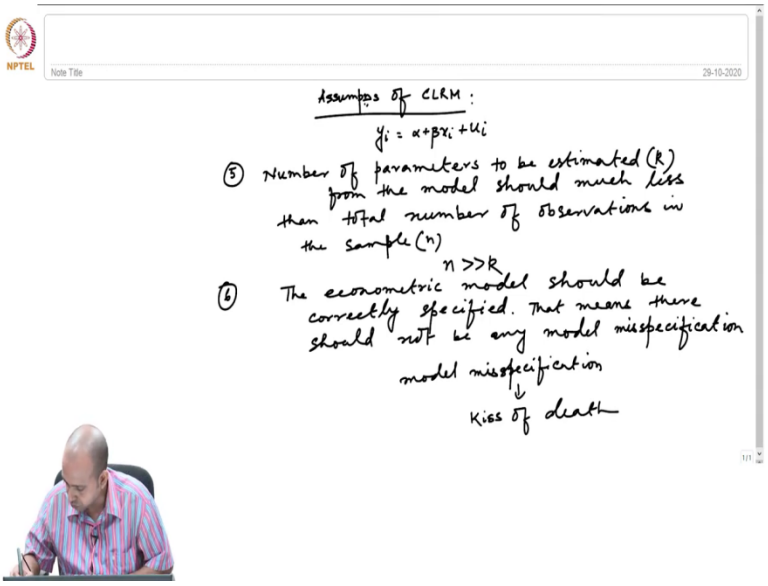
Introduction to Econometrics
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras
Lecture 9
Classical Linear Regression Model Part-3

(Refer Slide Time: 00:14)



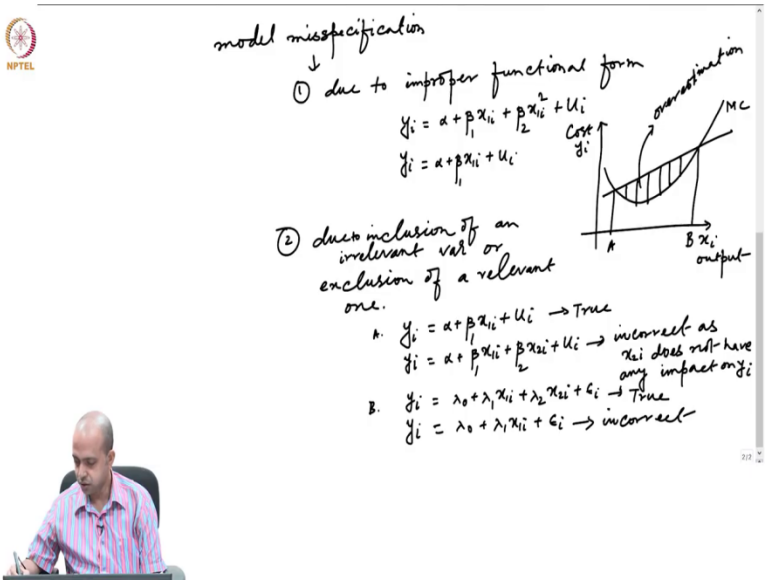
Welcome. So we are discussing about the 10 assumptions of Classical Linear Regression Model. So, this is basically assumptions of CLRM that we were discussing yesterday and once again I would like to repeat that these assumptions are required because they describe an idealistic situation under which the CLRM estimates exhibit the three desirable properties.

So, if these assumptions are maintained and then we estimate the model, our estimates of alpha hat and beta hat will exhibit the desirable properties of efficiency, unbiasedness and consistency. The assumptions discussed yesterday were, firstly the model is linear that is linear in parameters. Then we have also discussed that in our model where $y_i = \alpha + \beta x_i + u_i$, then the expectation of u_i given x_i is equal to 0 and error term does not have any covariance with the explanatory variable. That means covariance between x_i and u_i is equal to 0 and when this assumption is maintained x_i's are said to be exogenous in the sample meaning x_i's are fixed over repeated sampling. We have also discussed that there should be enough variation in the independent variable x_i. So if there is no variation in x_i then x_i does not qualify to be included as an explanatory variable in our model.

Today we will discuss about another assumption which says that the number of parameters to be estimated from the model should be much less than total number of observations in the sample. Alternatively we can say that number of observations in your sample should be much greater than the number of parameters to be estimated in the model. So, you can say that n (number of observations) should be much greater than k (number of parameters). For example, in this model we are going to estimate only two parameters-α and β. So k is equal to 2. There is no clear cut rule set by the econometrician on the exact number of observations but as a rule of thumb we can say that it should be minimum 20 times larger than the number of parameters. So, if you are estimating two parameters then probably you should get minimum 40 observations and more is better. If you try to estimate two parameters and your total number of observation is let us say 5, that means the sample estimate of α and β would be unreliable-we cannot rely much on the α and β estimated from a very very small sample where number of observations is just greater than the number of parameters. So, n should be much-much greater than k.

Assumption 6 says that the econometric model should be correctly specified. That means there should not be any model misspecification. Now, model misspecification is also a serious problem in econometric analysis. Sometimes econometricians say that model misspecification is basically dangerous that they term this model misspecification as case of death. So, model misspecification is known as case of death indicating that model misspecification problem is so subtle in nature that you do not even realize that you have made the mistake and misspecified your model, but the consequence is so severe. Since it is very difficult to identify that you have committed the mistake of model specification but the consequence is highly dangerous, it is basically case of death in econometric analysis. So, we should first ensure that our model is correctly specified before we actually estimate the model using data and all.

(Refer Slide Time: 08:47)



Now there could be two types of model misspecification. The first type of model misspecification is due to improper functional form. For example, let us say that our true model is $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + u_i$. Now, we are specifying here a model say marginal cost function where in the x axis we are measuring output which is denoted by x_i. I am interested in estimating a marginal cost function which is a u shaped relationship between cost output (y_i, measured along the y axis).

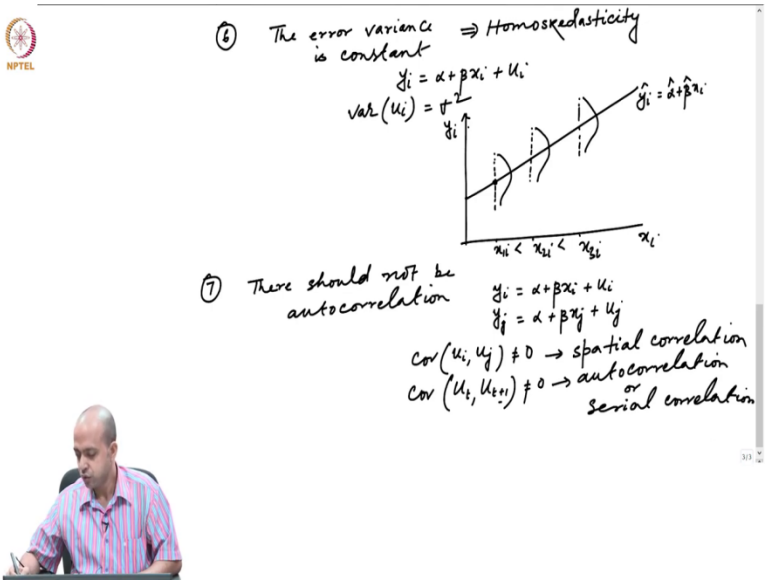
So, that means we know that the relationship between y_i and x_i is non-linear and it is u_i. As production increases marginal cost comes down initially and then it goes up in a u shape. If you want to estimate a u shaped relationship, then you need to specify this type of model where $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + u_i$. It is coming from our standard quadratic equation form which is ax^2+bx+c. This is our true cost function. But by mistake let us say that you are specifying this type of model where $y_i = \alpha + \beta_1 x_{1i} + u_i$. That means when the true relationship is non-linear and you are specifying a linear relationship like this you are either overestimating or underestimating the cost of production.

For example, let us say this is point A and this is point B. At the left side of A your true cost is actually higher than what you have predicted through this linear function, and beyond B your true cost is actually lower. In between A and B your true cost is actually lower than the predicted

one. That means your model is basically over estimating. There is a problem of overestimation in between A and B. Because, this is the line you have fitted where the true relationship is this and beyond B and left side of A also it is underestimating the cost. So in between A and B there is overestimation. So, this is type of model misspecification arises due to improper form where my functional form says my variable should appear in the equation in its level form as well as in its quadratic form. That means when the functional form requires non-linearity in variable and you are specifying a linear relationship between y_i and x_i, that is why you are either overestimating or underestimating the cost. This is model misspecification due to improper functional form.

Secondly, model specification may also arise due to inclusion of an irrelevant variable or exclusion of a relevant one. From this example itself we can understand that your true model requires x1i square also to be included but you are excluding it when you are estimating. So you are specifying this equation as $y_i = \alpha + \beta_1 x_{1i} + u_i$ thereby omitting x_{1i}^2. You can think of a completely different example say, your true model is $y_i = \alpha + \beta_1 x_{1i} + u_i$ but unnecessarily you are estimating a relationship where $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$ which is incorrect as x_{2i} does not have any impact on y_i. This is basically inclusion of an irrelevant variable. Similarly, it may also happen that your $y_i = \lambda_0 + \lambda_1 x_{1i} + \lambda_2 x_{2i} + \varepsilon_i$ which is my true model but you are specifying your $y_i = \lambda_0 + \lambda_1 x_{1i} + \varepsilon_i$ which is incorrect because you are omitting a relevant variable x_{2i} from the model. So, these two examples are about the two cases by which model misspecification may arise- either due to improper functional form or due to inclusion of an irrelevant or exclusion of a relevant variable. This is assumption number 6.

(Refer Slide Time: 17:50)



Assumption number 7 says that the error variance is constant. In your model when you are specifying $y_i = \alpha + \beta x_i + u_i$, that means you are saying variance $(u_i) = \sigma^2$. i is basically indexed for individual. This is the equation that I have specified for the i^{th} individual. Similarly, you can write another equation for the j^{th} individual where $y_j = \alpha + \beta x_i + u_j$ and variance of ui, uj, uk everywhere is σ^2. That is why there is no subscript i over here. The error variance is constant and if that is the case this indicates a situation of homoscedasticity. In a simple diagram I can make you understand this situation.

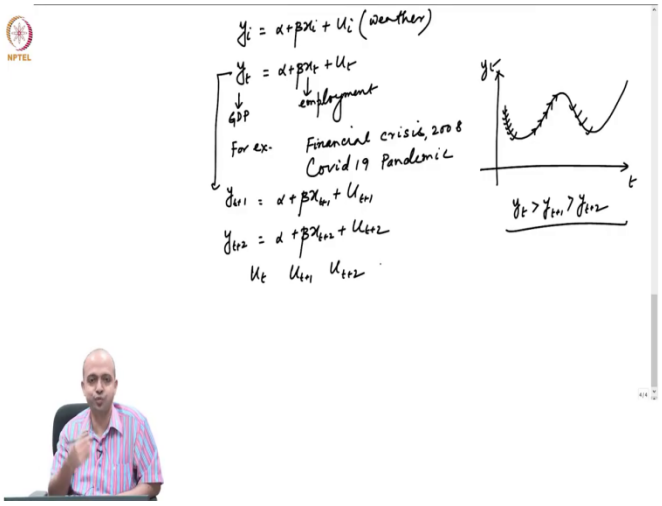
Let us say, this is x_i (income) and this is y_i (consumption) and let us say this is your line that you have specified where your $\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$. For any level of x_i let us say this is x_{1i}. You will have line point like this. That means while your line predicts that the predicted income should be there, your actual consumption should be here but if you look at your data set you can see that your actual consumption is either above the predicted line or below the predicted line.

So, that means either you have made positive error term or negative error term. These you can think as the spread of the error term. From this diagram you can understand the spread of the error term. The spread is called the spread or variance of the error term and it is actually same variance. When your income is increasing from x_1 to x_2, x_2 to x_3 at higher level of income also the error variance is constant. That is what we assume and is called constant error variance or homoscedasticity. But, in reality we may not have this type of situation. We may get a situation where your error term becomes heteroskedastic in nature. Actually that is the realistic situation. The reason for heteroscedasticity, remedial measures, detection of heteroscedasticity problem

will be discussed in detail later on when we come back to this particular chapter called heteroscedasticity. That means later on we will try to relax these assumptions one by one and we will discuss their detection, remedial measures and consequences. So, for the timing you just understand the simple meaning of heteroscedasticity which means that at different levels of income the variance of the error term is actually constant. There is no change in the spread of this error term.

The assumption number 8 says that there should not be autocorrelation. The model $y_i = \alpha + \beta x_i + u_i$ is written for the i^{th} individual. For the j^{th} individual the equation is then $y_j = \alpha + \beta x_j + u_j$. The assumption of autocorrelation say that covariance between u_i and u_j is actually 0. Now, though this term is actually is covariance between u_i and u_j, it reflects the correlation between u_i and u_j. This is actually not autocorrelation. This is known as spatial correlation and if you are dealing with a time series data then you will have this type of situation where u_t and u_{t+1} or u_{t-1} should also be 0. So, if this is not 0 then only it is called special correlation and this is called autocorrelation or serial correlation. But our assumption says that there should not be this type of relationship.

(Refer Slide Time: 25:57)



So, presence of either autocorrelation or spatial correlation is ruled out by the assumption so that our estimates of $\hat{\alpha}$ and $\hat{\beta}$ shows the desirable properties. Why does the problem of spatial correlation or autocorrelation arise? So when you are dealing with cross sectional data that means let us say this is your y_i (consumption) which is a function of income plus u_i and this u_i basically captures the omitted variable. As we mentioned, the omitted variable might also have some impact on y_i apart from your income. Now, when you are collecting data it may so happen that let us say the i^{th} individual and the j^{th} individual are residing in a same place. So that means i^{th} individual and j^{th} individuals' consumption might be influenced by some other factor which is not there and that is why it is there in the error term.

Since, i^{th} and j^{th} individual reside in the same place it might so happen that u_i and uj are correlated. So, this is the problem in cross sectional data and this is the reason for spatial correlation since the two individuals are residing in the same place their consumption pattern might be influenced by a same variable which is not included in our model. For example, then i^{th} and j^{th} individuals consumption are influenced by same type of weather it might also happen that i^{th} and j^{th} individuals' taste and preference is also same since they are residing in a neighborhood. In that case also there is a possibility that u_i and u_j will be correlated and then when you are dealing with time series data your equation would be $y_t = \alpha + \beta x_t + u_t$.

When you deal with time series data sometimes our variable, let us say, this is your GDP and this is your employment data. So many a times these types of macroeconomic variables get some kind of external shock. For example, financial crisis in 2008 or the recent covid-19 pandemic. These external shocks are given in the economic system and the impact of this shock does not die down immediately. That means the impact persist in the system for some time.

Since this shock cannot be actually observed, the error term captures actually the impact of these external shocks which is given in the economy. So, starting from 2008 onwards to 2009, 10, 11 and all other years your variable will get impacted by the same type of shock which was given earlier. Since ui which is your error term is influenced by same type of shock, it is quite likely that they get correlated in a time series data.

Let us say, this is your x_t and this is your y_t. When external shock is given, your time series macroeconomic variable is time and it will show some kind of pattern like this let us say, this is your y_t so, this pattern is actually known as business cycle in macroeconomic literature.

So, due to external shock when the economy is going down that means when the economy goes for recession that means in each successive period GDP of the country becomes lower than the previous period. So, that means y_t is greater than y_{t+1} and y_{t+1} is greater than y_{t+2} like this or when it is going up then the reverse pattern will observe then again the pattern like this.

So, in each successive period here y_t shows a pattern where y_t is greater than y_{t+1} and y_{t+1} is greater than y_{t+2} which is greater than y_{t+3}. Obviously, from these we can write that y_{t+1} would be $\alpha + \beta x_{t+1} + u_{t+1}$ and then $y_{t+2} = \alpha + \beta x_{t+2} + u_{t+2}$. Since, y_{t+1} and y_{t+2} show some kind of pattern, we will observe that u_{t+1} and u_{t+2} will also show a similar type of pattern because of which we will end up with this type of autocorrelation problem. This is the reason. Of course we will discuss again in detail about the autocorrelation problem, how to detect autocorrelation problem, the other reasons of autocorrelation problem and how to solve autocorrelation problem in a later part of our discussion when we will come back in detail with autocorrelation.