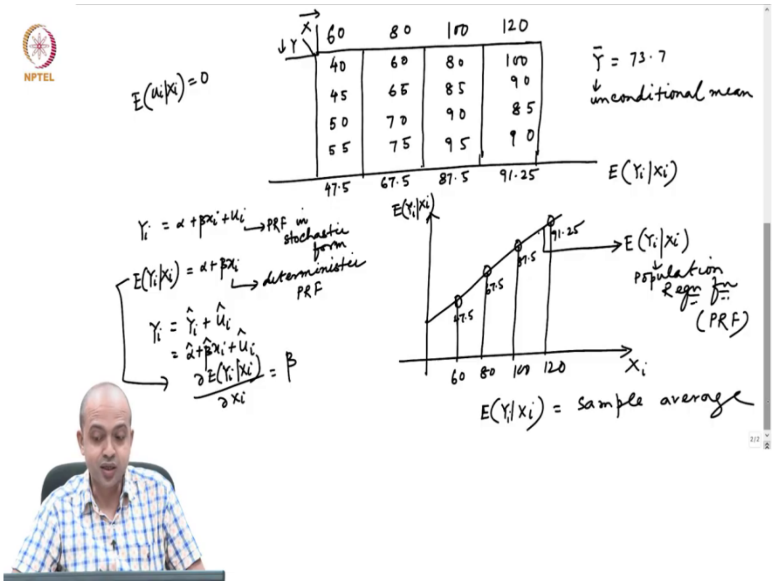
Introduction to Econometrics
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras
Lecture 8
Classical Linear Regression Model Part-2

(Refer Slide Time: 00:14)



$E(u_i|X_i) = 0$

$\downarrow Y \quad \overrightarrow{X}$	60	80	100	120
	40	60	80	100
	45	65	85	90
	50	70	90	85
	55	75	95	90
$E(Y_i|X_i)$	47.5	67.5	87.5	91.25

$\bar{Y} = 73.7$ $\downarrow$ unconditional mean

$Y_i = \alpha + \beta x_i + u_i \longrightarrow$ PRF in stochastic form

$E(Y_i|X_i) = \alpha + \beta x_i \hookrightarrow$ deterministic PRF

$Y_i = \hat{Y}_i + \hat{u}_i$

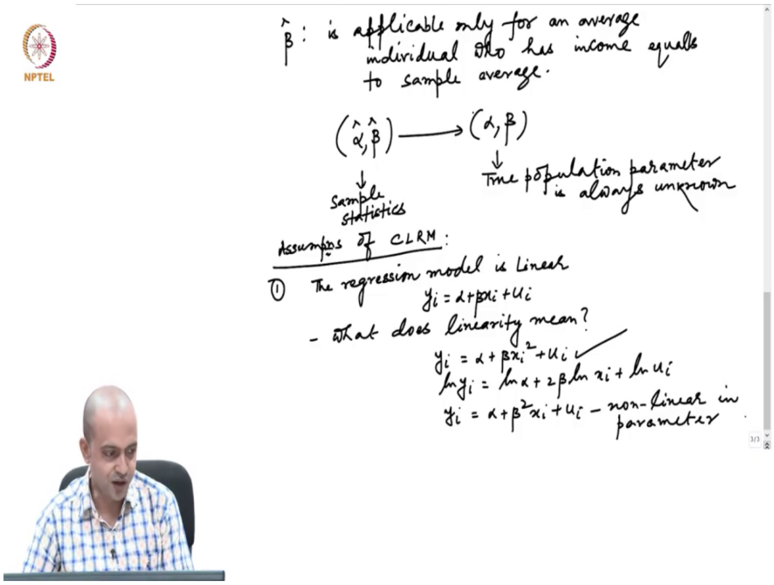
$= \hat{\alpha} + \hat{\beta} x_i + \hat{u}_i$

$\dfrac{\partial E(Y_i|X_i)}{\partial X_i} = \beta$

$\longrightarrow E(Y_i|X_i)$ $\downarrow$ Population Regn fn (PRF)

$E(Y_i|X_i)$ = sample average

(Refer Slide Time: 00:24)



$\hat{\beta}$: is applicable only for an average individual who has income equalls to sample average.

$(\hat{\alpha}, \hat{\beta}) \longrightarrow (\alpha, \beta)$

$\downarrow$ Sample statistics

$\downarrow$ True population parameter is always unknown

Assumptions of CLRM:

① The regression model is Linear

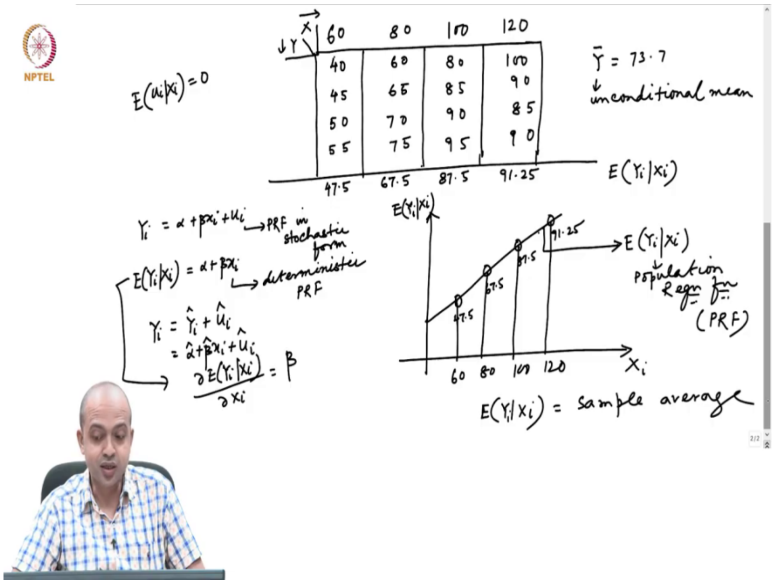
$y_i = \alpha + \beta x_i + u_i$

- What does linearity mean?

$y_i = \alpha + \beta x_i^2 + u_i$ ✓

$\ln y_i = \ln \alpha + 2\beta \ln x_i + \ln u_i$

$y_i = \alpha + \beta^2 x_i + u_i$ – non-linear in parameter.



$E(u_i|X_i) = 0$

$\downarrow Y \backslash \overrightarrow{X}$	60	80	100	120
	40	60	80	100
	45	65	85	90
	50	70	90	85
	55	75	95	90
$E(Y_i|X_i)$	47.5	67.5	87.5	91.25

$\bar{Y} = 73.7$ → unconditional mean

$Y_i = \alpha + \beta X_i + u_i$ → PRF in stochastic form

$E(Y_i|X_i) = \alpha + \beta X_i$ → deterministic PRF

$Y_i = \hat{Y}_i + \hat{u}_i$

$= \hat{\alpha} + \hat{\beta} X_i + \hat{u}_i$

$\dfrac{\partial E(Y_i|X_i)}{\partial X_i} = \beta$

$E(Y_i|X_i)$ = sample average

So, the interpretation of β or $\hat{\beta}$ is applicable only for an average individual. So, please keep in mind that the interpretation of $\hat{\beta}$ is applicable only for an average individual in the sample. The average individual is one who has income equal to sample average. So, that means my entire interpretation of the regression coefficient is valid only for the representative individual or average individual and not for each and every individual.

The regression line itself is an average line which is expectation of Y_i given X_i. So that means I am predicting the income using that average line when income changes from 60 to 70 and the change is measured with reference to this line which is basically an average line. That means the average line itself indicates the features of an average individual who is completely different from actual individual. That is why you can see from the diagram that when my income is 60 my model says that consumption should be 47.5 and that consumption is for that average individual.

But, if you look at in this group- when the income level is 60, not all the people are having consumption level of 47.5. Rather there are consumption levels of 40, 45, 50, 55 etc. That means we are committing some kind of mistake in predicting somebody's consumption with the help of income. Similarly, when income changes from 60 to 80, the model says that the predicted consumption should be 67.5 but actual consumption is 60, 65, 70 and 75. When it goes from 80 to 100 then it becomes 87.5 but actual consumption in this group is 80, 85, 90 and 95. So, that means we are committing mistakes in predicting individual's consumption at each and every level of income. Actual consumption is Y_i and predicted is $\hat{Y}_i$ and the difference basically indicates your $\hat{u}_i$ which is the predicted error term or the mistakes committed.

So, whenever we do the interpretation of $\hat{\beta}$, we have to say that for a unit change in income, consumption changes by β amount or $\hat{\beta}$ amount on an average. 'On an average' component should not be missed because that is valid only for the average individual. This is the meaning of regression analysis.

Now, the next question is the need to do the regression. Our entire objective of this regression analysis is drawing inference about the true population parameter with the help of the sample counterpart. So, this is the true population parameter and this is the sample counterpart or sample statistic- $\hat{\alpha}$ and $\hat{\beta}$. Now, the true population parameter is always unknown. Then the question is when the true population parameter α and β are always unknown, why we are spending so much of time in estimating $\hat{\alpha}$ and $\hat{\beta}$? Even after estimation also how do you know that we are actually predicting the true population parameter because they are always unknown?

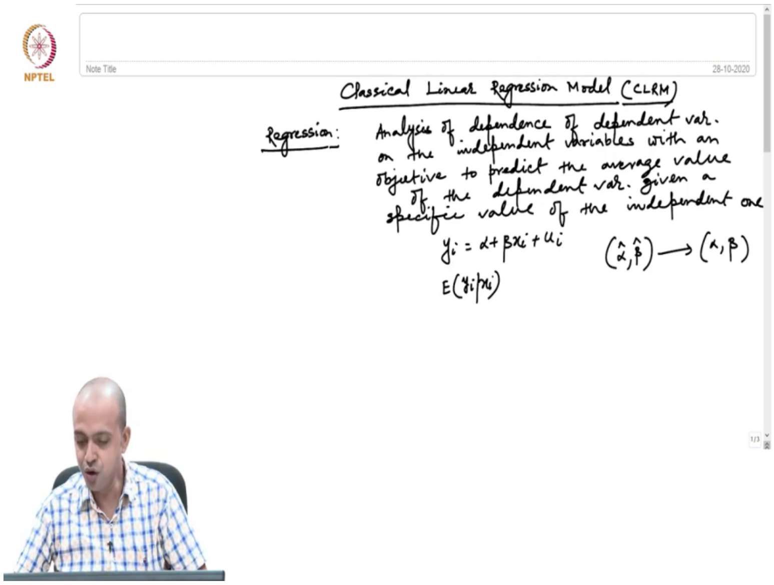
Now, I would like to introduce a simple analogy here. True population parameter is like our god. All of us we believe that there is god but none of us have ever seen god. That is why god is unknown. We do not know how does he or she look like but we believe that there is a god and we make certain assumptions and those assumptions actually generated different religions Hindu, Muslim, Christian so on and so forth. In every religion there is a set of assumptions like how you should lead your life so that you can reach to god at the end. Similarly, econometricians' also say that there is a god which is population parameter, which is always unknown. We can always estimate $\hat{\alpha}$ and $\hat{\beta}$ like each and every religion.

Hindus have a shape of god, Muslims have a different shape, and Christians have a different shape. These gods what we see is the estimate or estimator of the true population parameter α and β, and each religion says that if you perform religiously and if you behave religiously throughout your life then at the end you will reach god.

Similarly, econometrician say that if you follow certain assumptions while estimating your $\hat{\alpha}$ and $\hat{\beta}$, then the statistical procedure will ensure that your $\hat{\alpha}$ and $\hat{\beta}$ would be as close as possible towards your true population parameter α and β. So, that means we need to make certain assumptions before we estimate the regression line. Only then we can say that this line is actually reliable and this line is based on $\hat{\alpha}$ and $\hat{\beta}$.

If we follow those assumptions then our $\hat{\alpha}$ and $\hat{\beta}$ are reliable and they would be as close as possible to the unknown true population parameters and then only the $\hat{\alpha}$ and $\hat{\beta}$ would be useful for any inference making or policy making. So, this line is estimated based on certain assumptions which we need to know.

(Refer Slide Time: 9:52)



Classical Linear Regression Model (CLRM)

Regression: Analysis of dependence of dependent var. on the independent variables with an objective to predict the average value of the dependent var. given a specific value of the independent one

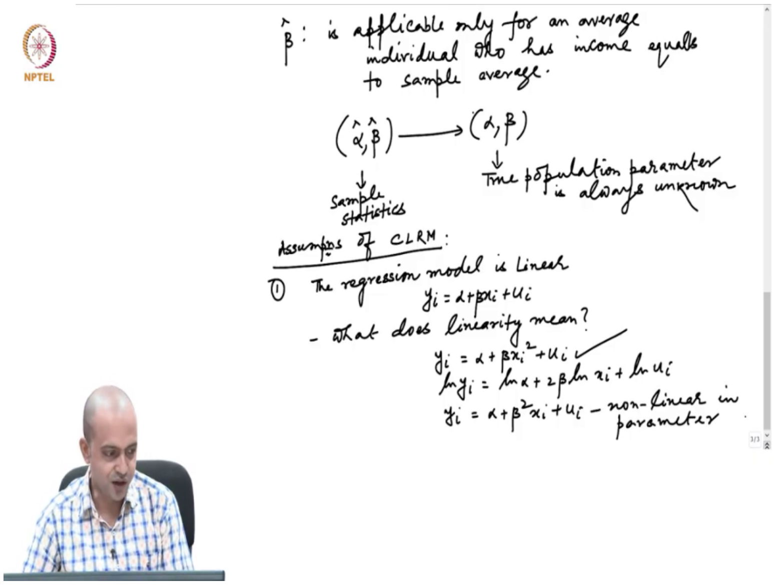
$y_i = \alpha + \beta x_i + u_i$ $(\hat{\alpha}, \hat{\beta}) \longrightarrow (\alpha, \beta)$

$E(y_i|x_i)$

That means estimation of this classical linear regression model is based on certain assumptions and only if we maintain those assumptions our alpha hat and beta hat will exhibit the desirable properties of unbiasedness, efficiency and consistency what we have discussed yesterday.

So, those assumptions are actually an idealistic situation like in reality every religion says that you should be honest, you should have tolerance, you should do this and that etc. Those things are like assumptions which describe an idealistic situation that will lead you to the god. Similarly, these assumptions of classical regression linear regression model will estimate $\hat{\alpha}$ and $\hat{\beta}$ and will lead to α and β as close as possible and then only we say that $\hat{\alpha}$ and $\hat{\beta}$ is reliable and exhibit the three desirable properties of unbiasedness, efficiency and consistency.

(Refer Slide Time: 11:09)



$\hat{\beta}$: is applicable only for an average individual who has income equals to sample average.

$$\begin{tikzcd} (\hat{\alpha},\hat{\beta}) \arrow[r] \arrow[d] & (\alpha,\beta) \arrow[d] \\ \text{Sample statistics} & \text{True population parameter is always unknown} \end{tikzcd}$$

Assumptions of CLRM:

① The regression model is Linear

$y_i = \alpha + \beta x_i + u_i$

- What does linearity mean?

$y_i = \alpha + \beta x_i^2 + u_i$ ✓

$\ln y_i = \ln \alpha + 2\beta \ln x_i + \ln u_i$

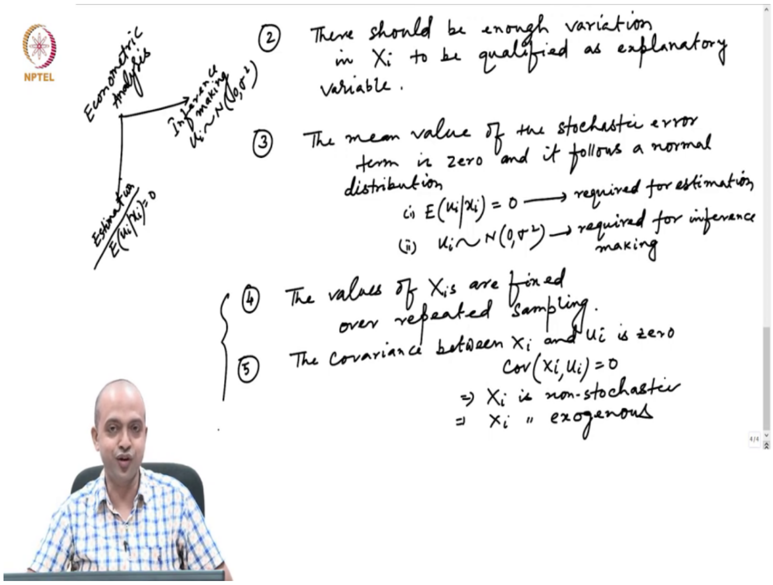
$y_i = \alpha + \beta^2 x_i + u_i$ – non-linear in parameter

Now then it becomes important to know those assumptions while we estimate $\hat{\alpha}$ and $\hat{\beta}$. So these are called the assumptions of CLRM. First assumption says that the regression model is linear. I am writing this type of model where $Y = \alpha + \beta X_i + u_i$. When I am saying that the model is linear, I am saying the model is linear in parameters. So, here I have two parameters- α and β and I am saying the model is linear in parameters. I am not saying about the linearity of the variables. That means it may or may not be linear in variables and we can still estimate the model using the procedure of classical linear regression model even if the variable is non-linear in nature.

For example the model $\alpha + \beta X_i^2 + u_i$ is non-linear in variable but still we can estimate the model using the procedure suggested by CLRM which will discuss later because this model can always be linearized by proper transformation. So linearity is not required for the variables but linearity is required for the parameters. For example, if you take log of both side this would become log(Yi) = α + 2β log(Xi) + log(u_i). Thus you can always linearize when the model is non-linear in variable but if the model is where Y = α + $\beta^2 X_i + u_i$, then this model is called non-linear in parameter and these types of models cannot be estimated using the technique what we will learn in CLRM. That requires a specific technique called Non-linear estimation method.

So, that means whenever we are thinking about classical linear regression model as the name itself says it is the linear model which takes the form Yi = α + β $X_i + u_i$ or Yi = α + $\beta X_i^2 + u_i$ because linearity or non-linearity in variable is not a matter of concern. We can always linearize that model using a proper transformation. But it requires the model to be linear in parameters like That is the first assumption we must make.

(Refer Slide Time: 15:38)



Econometric Analysis

Inference making $u_i \sim N(0,\sigma^2)$

Estimation $E(u_i/X_i)=0$

② There should be enough variation in X_i to be qualified as explanatory variable.

③ The mean value of the stochastic error term is zero and it follows a normal distribution

(i) $E(u_i/x_i) = 0 \longrightarrow$ required for estimation

(ii) $u_i \sim N(0,\sigma^2) \rightarrow$ required for inference making

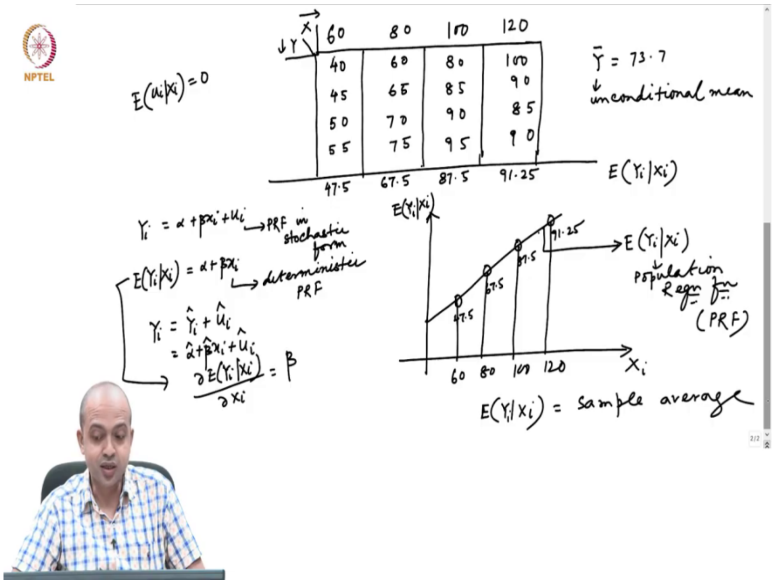
④ The values of X_is are fixed over repeated sampling.

⑤ The covariance between X_i and u_i is zero

$Cov(X_i, u_i) = 0$

$\Rightarrow X_i$ is non-stochastic

$\Rightarrow X_i$ " exogenous



$E(u_i|X_i) = 0$

$\downarrow Y \backslash \overrightarrow{X}$	60	80	100	120
	40	60	80	100
	45	65	85	90
	50	70	90	85
	55	75	95	90
$E(Y_i|X_i)$	47.5	67.5	87.5	91.25

$\bar{Y} = 73.7 \downarrow$ unconditional mean

$Y_i = \alpha + \beta x_i + u_i \longrightarrow$ PRF in stochastic form

$E(Y_i|X_i) = \alpha + \beta x_i \hookrightarrow$ deterministic PRF

$Y_i = \hat{Y}_i + \hat{u}_i$

$= \hat{\alpha} + \hat{\beta} x_i + \hat{u}_i$

$\frac{\partial E(Y_i|X_i)}{\partial x_i} = \beta$

$E(Y_i|X_i) = $ sample average

Then the second assumption says that there should be enough variation in X_i to be qualified as the explanatory variable. For example, when we are talking about take this sample where the income level of several group of families is 60, 80, 100 and 120. Now, instead of that if the income level for all the groups were 60, 60, 60 and 60 then the observed difference in their consumption cannot be explained by income. If all the people have same level of income how can you claim their difference in their consumption is due to income? Everybody is having the same level of income.

So, let us say if it is not 80, let us say 60, 61, 62 and 61.5. That difference is also not significant. That is why it says there should be enough variation in your X_i to be qualified as an explanatory variable and to be included in the model. So, when you specify your model you must ensure that all your explanatory variables have enough variation in them. Without having enough variation I cannot include a variable in my model as an explanatory variable.

Same is true for Y. If everybody have the same level of consumption the necessity of regression itself does not arise. Why should I bother about regression analysis when everybody is same in terms of their consumption? The question itself does not arise in our mind. So, that is why I have I assumed that there is enough variation in the consumption level and that is why it gives some kind of motivation to us to think and understand why different people have different consumption patterns and that we are trying to understand with their level of income. There should be enough variation in the level of income because without a variation in level of income we cannot include that variable as explanatory variable in my model. So, this is assumption number two.

The third assumption says that the mean value of the stochastic error term which actually captures the influence of omitted variables in the model is 0 and it follows a normal distribution. So, that means the expectation of u_i given X_i is equal to 0 and u_i follows a normal distribution with let us say 0 mean and σ^2 variance. This is what we know about the error term. Now, we have to keep in mind again that there are two assumptions that we make about the error term. This is let us say 1 and this is 2. Now, assumption number 1 is required for estimation. That means without making this assumption we cannot even estimate $\hat{\alpha}$ and $\hat{\beta}$ because our model says $Y_i = \alpha + \beta X_i + u_i$. Unless I know that expectation of u_i equals to 0, I cannot say that expectation of Y_i given X_i equals to α and β and unless I remove that u_i it is difficult to estimate that model.

So, that is why this first assumption is required for estimation. Normality of the error term is not required for the estimation purpose. So, in an econometric analysis there are two stages or two steps where first one is estimation and second stage is basically inference making.

So, this is required here in this stage- expectation of u_i given X_i is equal to 0 and this requires inference making or hypothesis testing which we will discuss in detail later on. This requires u_i to follow a normal distribution with 0 mean or σ^2 variance or some other normal distribution- I am not very particular about what should be the variance of this but, it should follow a normal distribution with either sigma square variance or unit variance whatever. So, that is why I am saying this is required for inference making. This is the assumption that we make about the error term.

The fourth assumption is again very important and says that the values of X_i's are fixed over repeated sampling. Now, this assumption is little difficult to understand and I will explain that with the example once again. We will use the same example here. The X values are 60, 80, 100 and 120 and they are fixed over repeated sampling. Now this entire 16 observations of 16 families consist of my entire population and from that population I can always draw a sample of 4 families and how can I do that?

Let us say that this is my Y_i and this is my X_i. For the first case let us say that I am taking only the first element of each group that means 40, 60, 80 and 100 randomly but x values are 60, 80, 100, 120 which is one sample. Now, I can draw another sample taking let us say second element from each group randomly. So, what will that sampling look like? If you plot Y_i and X_i then that will become 40. This is now becoming 45, 65, 85 and 90.

I am taking second element from each group but my x values are still 60, 80, 100 and 120. What does it mean? It means that in repeated sampling while you can draw a particular family randomly from each group, your X values are fixed-60, 80, 100 and 120. Whatever sample you draw, first of all you must fix your X values otherwise I cannot fix the group from which I am going to take a particular household randomly. That is why it says that X_i's are fixed over repeated sampling from the given population because yesterday we discussed that from a given population we can randomly draw n number of samples.

But, of course in reality we are not going to collect n number of samples. Rather we will draw only one sample but we should draw the sample randomly so that I can get a distribution of my estimated parameter beta hat. Unless the sample is drawn randomly, I can get only one such sample non randomly drawn from the population. That is why I say that X_i's are fixed over repeated sampling.

I am saying that since the sample is drawn randomly I can generate many samples from the given population but my X values are fixed at 60, 80, 100 and 120. That is what it means when I say that my X values are fixed over repeated sampling. Assumption number 5 (which is actually derived from assumption number 4) say that when the values of X_i's are fixed over repeated sampling, the covariance between X_i and u_i is 0. That means covariance between X_i and u_i is equals to 0 and that indicates X_i is actually non-stochastic or X_i is exogenously given.

So, income level of the individuals is given and that determines somebody's consumption which is actually endogenous in this model. So, this is assumption 5 and these two assumptions are very important but they represent mostly the same thing. When I am saying that X_i's are fixed over repeated sampling you can fix values of X only when that is exogenously given. If the variable is determined from within the system then that is called an endogenous variable for which you have no control and you cannot fix. That is why I say that X_i's are fixed over repeated sampling and for that reason X_i's are called non-stochastic and exogenous variable. That means covariance between X_i and u_i is actually 0.

But in reality, many a times this X_i may become endogenous in nature. That means in reality we may get situations wherein the covariance between X_i and ui is actually not equal to 0 and if they are not equal to 0, X_i becomes endogenous and we need to apply a different estimation technique which is different from what we are going to learn in CLRM to ensure that our estimates alpha hat and beta hat exhibit the three desirable properties of unbiasedness, efficiency and consistency. That is why I said these are the assumptions that we are making to describe the idealistic world but reality might be completely different and if the reality is different, then obviously we cannot reach towards god.

So, we must ensure that if there is any problem we must rectify in the data. If these assumptions are not satisfied then we need to rectify those. That is why first of all we should know the idealistic situation and the assumptions that we make so that our estimates exhibit the desirable properties. There are another 4 to 5 important assumptions describing this idealistic world before we estimate the model using CLRM which we will discuss in our next class.

Thank you very much.