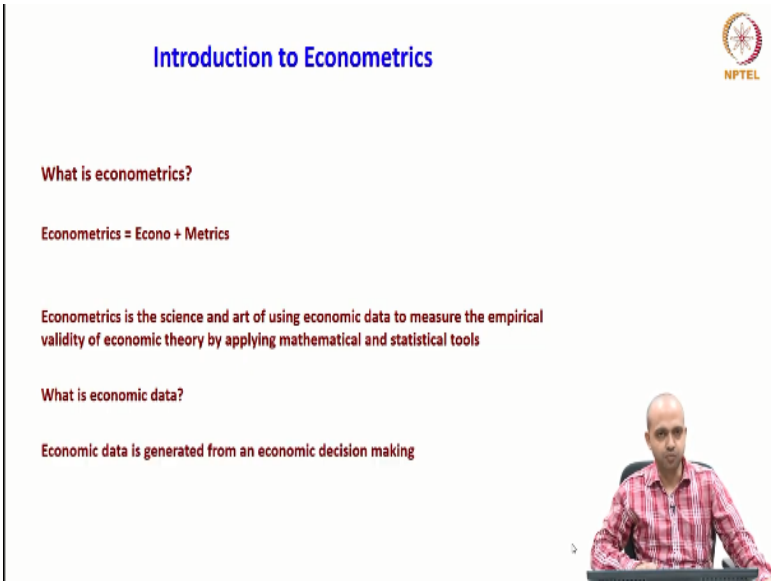
Introduction to Econometrics
Professor. Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras
Lecture No. 05
Desirable Properties of the Estimates of the Population Parameters Part - I

(Refer Slide Time: 0:15)

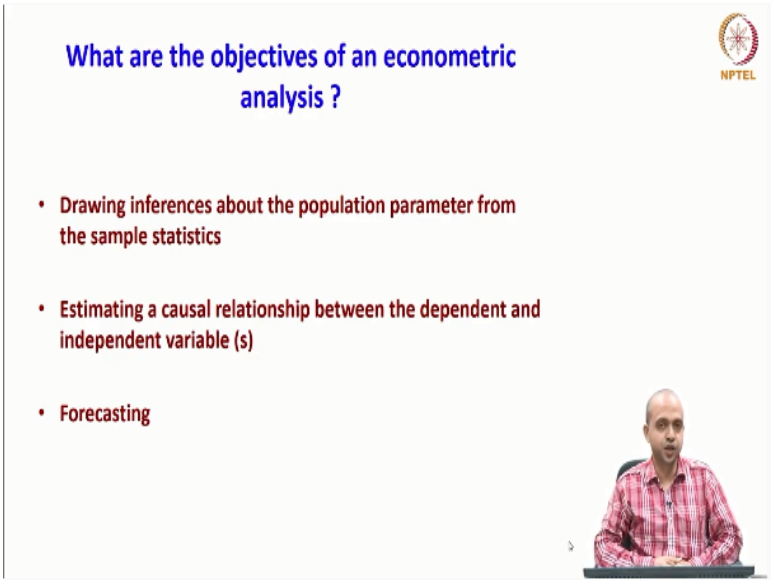


So, welcome to our discussion of Econometrics once again. In our previous class we were basically discussing about what is econometrics and then we said that econometrics is basically an application of mathematical and statistical tools to analyze economic data for the purpose of measuring empirical validity of economic theory. And then we have also discussed about in detail what is exactly economic data and we said that economic data is basically a data that is generated out of economic decision making.

(Refer Slide Time: 1:08)

And we have discussed several examples of economic data and these are the examples that we have discussed.

(Refer Slide Time: 1:19)



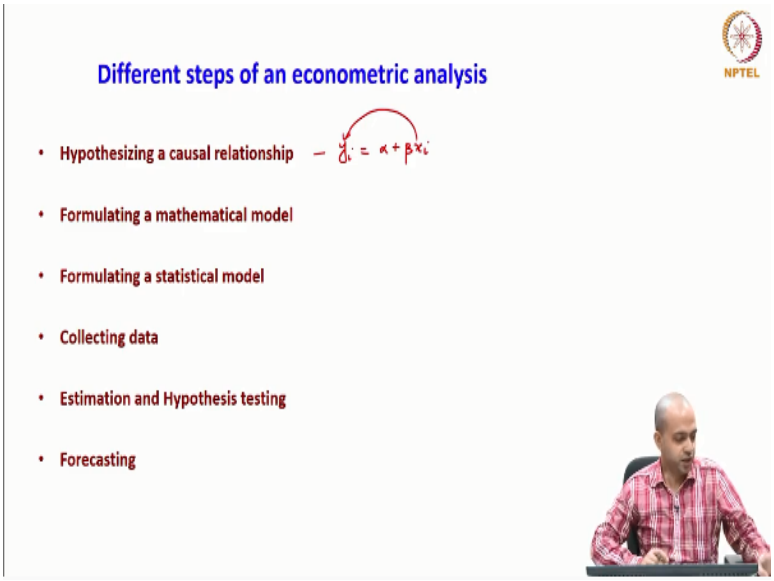
And then we have also discussed about the three major objectives of econometric analysis. Firstly, drawing inference about the population parameter. In the context of consumption function that we are discussing, where we said that $Y = \alpha + \beta X$, α and β are the true population parameters which are unknown and we are trying to draw some kind

of inference about the true population parameter on the basis of our sample statistic $\hat{\alpha}$ and $\hat{\beta}$ that we are going to estimate. And then we are also going to establish some kind of causal relationship between the dependent and independent variable. In this example, we are trying to establish the causal relationship between income and consumption through econometric analysis.

And then lastly, econometric analysis helps in forecasting and regarding this forecasting we have to keep in mind that a related concept is prediction. So while prediction is applicable mostly in the context of cross sectional data meaning when you have information on income and consumption let us say for 100 individuals, so after estimating the model we can predict if someone or some individuals income is 750 rupees per week, then what should be the individuals consumption? So, that is what is called prediction which is applicable in the context of cross sectional data.

But when it comes to forecasting we have to keep in mind that forecasting is basically applicable for time series data, where we are going to use the past data of a time series macroeconomic variable and then we are going to predict the future values of that. For example, if we have India's GDP for let us say 10, 15 or 20 years then what we are going to do is that we are going to forecast India's GDP in next year or next to next year like that. So, that means between forecasting and prediction the only difference is the time dimension. While prediction is applicable for cross sectional data, forecasting is basically applicable for a time series data. So, forecasting is nothing but prediction about the future outcome. That is what we have to keep in mind in this context. Though we are using these two concepts interchangeably in this basic econometrics course- prediction and forecasting, they have a difference in terms of their time dimension. So, forecasting is nothing but prediction about the future. Like we say that weather forecasting and not weather prediction. So we use past data of weather and then we try to predict about the future weather which is called weather forecasting.

(Refer Slide Time: 4:31)

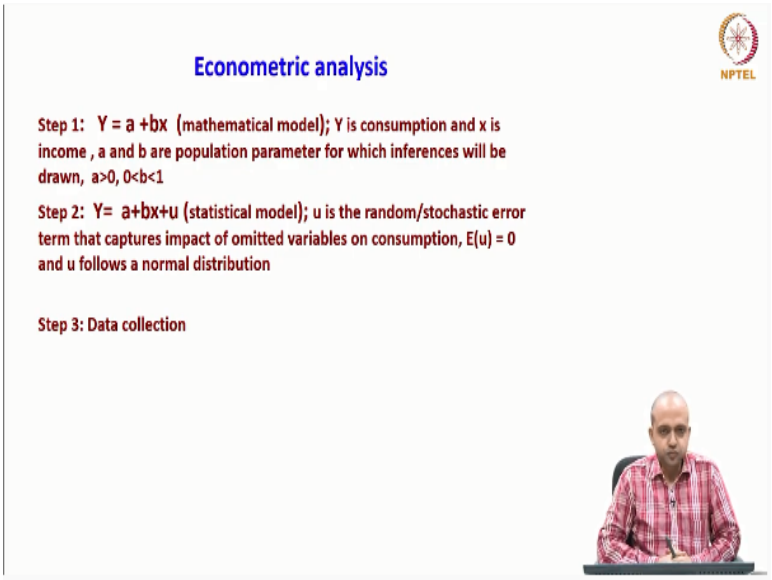


And then today what we are going to discuss is about the different steps that are involved in an econometric analysis. If you recall we said initially that econometric analysis is like any other scientific analysis. So firstly what you have to do is you have to observe certain things, for example, when you are talking about income and consumption, we observe how people consume, how people earn and then with increased amount of income how people spend. And after this type of observation of individuals consumption pattern, then what we have to do in econometric analysis as a first step is to hypothesize some kind of causal relationship between income and consumption. What we are hypothesizing is basically based on observation. We are observing certain things about the social individuals' behavior and then out of that observation we are trying to formulate some kind of hypothesis. That is the first step-hypothesizing a causal relationship.

Now, as we discussed earlier that econometric analysis per se will not tell you the direction of causality meaning when you have income and consumption, we do not know whether income is causing consumption or consumption is causing income. That means you have to set the model, you have to specify which is going to be your independent variable and which is going to be the dependent variable and that kind of hypothesis you have to make based on your own knowledge or based on the existing theory or existing literature.

Let us for example assume that in our context we are hypothesizing a causal relationship between income and consumption and we are also assuming the direction of causation runs from income to consumption. That means income causes consumption. Consumption is the dependent variable and income is the independent variable. When we saying $Y_i = \alpha + \beta X_i$; Y is consumption and X is income. So it is basically income causing consumption. This is the kind of hypothesis that we are making.

(Refer Slide Time: 7:32)



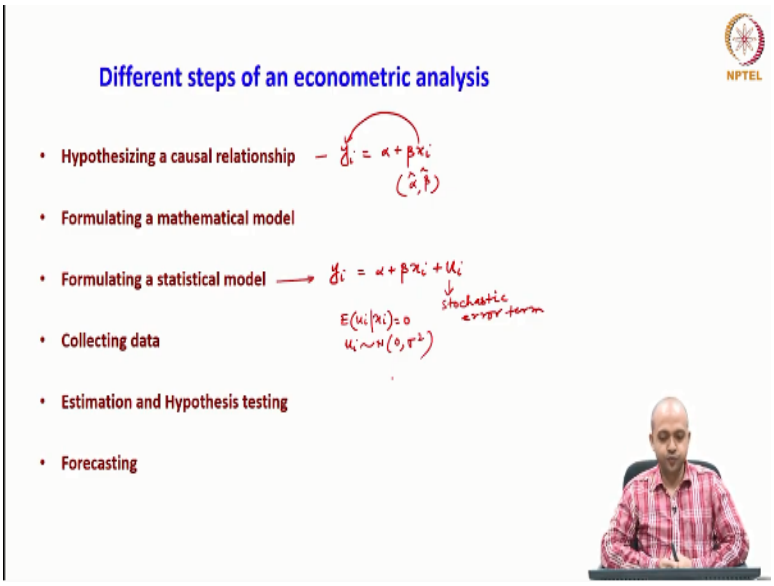
So, this is the step one. Once you specify this type of hypothesis that income is causing consumption, then the next step as we have mentioned here $Y = \alpha + \beta X$ is the next step which is formulating a mathematical model. What is the mathematical model? This is the mathematical model $Y = \alpha + \beta X$.

So, our hypothesis here is that we are assuming income causes consumption. That means when income increases, with increased amount of income individuals spend more on consumption. As income increases, consumption also increases but less than proportionately. If there is an increase in your income by 1 rupee, probably individuals' consumption will increase by 50paise, 60paise, 70, 80 or 90paise. That means by less than 1 rupee and that is what our hypothesis is.

Long back John Maynard Keynes made this kind of hypothesis and that is why it is called Keynesian consumption function actually. He made this hypothesis and then once you made the hypothesis you have to make the mathematical model which is the second step. This is the second step which is formulating a mathematical model where $Y = \alpha + \beta X$. Where Y is consumption, X is income and the restriction we are imposing on the α and β is that a is greater than 0 and b is greater than 0 but less than 1. b is greater than 0 but less than 1 and that restriction is actually implying our hypothesis.

And this b, the coefficient which is attached with income, it has a specific name in the literature of macroeconomics. It is basically called marginal propensity to consume. That means the increase in consumption for a unit change in income is called marginal propensity to consume or MPC. If you differentiate this function with respect to X you will see that $\Delta Y/\Delta X$ equals to b which is nothing but the marginal propensity to consume. Then, what is the next step?

(Refer Slide Time: 10:34)



Next step after formulating the mathematical model is we need to convert that mathematical model into a statistical one. Why this is so? Because in the mathematical model where $Y_i = \alpha + \beta X_i$, this is kind of exact relationship between income and consumption. That means if you estimate this model you will get $\hat{\alpha}$ and $\hat{\beta}$ as the estimates of alpha and beta. So, once you estimate your model, if I tell you some values of X which is, let us say 600 rupees, and then you would be able to exactly predict what is going to be the individuals' consumption. But in reality even if you know somebody's income, do you think that it is possible to predict exactly somebody's consumption? Probably no because in reality individuals consumption depends not only on income but also on many other factors like taste and preference, the price of related products, individuals education, gender, sex so and so forth.

So, that means there are many variables (factors) which will also have some kind of impact on the individuals' consumption which we have not included in this model. Even if you include 100 such variables there will always be some variables left out of this model because it is impossible to think and imagine all the factors that will have some impact on individuals' consumption. So, you may not have data for all the factors or you may not imagine all the factors and even if you can imagine all the factors then some

factors might be unobservable. For example, I can imagine test and preference but how to measure somebody's taste and preference? So, because of all these things, there will always be some omitted variables in your model and for that you cannot exactly predict somebody's income. That means, we need to convert the mathematical model into a statistical one by adding some kind of error term and how do you do that? This is the mathematical model. Then the statistical model would be $Y_i = \alpha + \beta X_i + u_i$, this u_i is called a stochastic error term. u_i basically captures the influence of all other omitted factors that might also have some impact on your consumption.

The moment you add the error term, your mathematical model becomes a statistical one and journey of econometric analysis starts from here actually. So the entire econometric analysis evolves around this stochastic error term. So even though u_i is called the error term it has the most significant influence. It is the most important factor in an econometric analysis. We have to clearly keep in our mind that the moment you add an error term in the mathematical model, your mathematical model becomes a statistical one and the journey of econometric analysis starts from here that is from this statistical model.

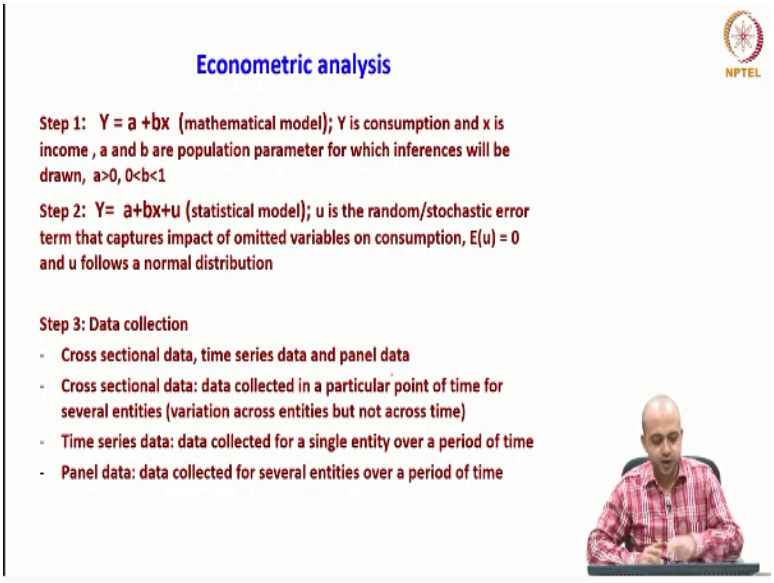
And the error term captures the influence of all other omitted variables that might also have some impact on consumption and what do we know about this error term? Error term is not a complete black box. Rather we know certain things. Two important things that we assume about the error term is firstly, we assume that expectation of u_i given X_i is basically 0. Conditional mean of the error term is 0 and then u_i follows a normal distribution with 0 mean and some sigma square variance. That is what we know about the error term.

So, the distributional assumption about the error term is known in this entire econometric analysis. Unless we know the distributional assumption or properties of the error term, it is really not possible to go for econometric analysis. And why do you assume the error term to follow a normal distribution? Because of some convenience and some statistical properties as you know that any distribution tends to follow normality as sample size increase because of the central limit theorem which we are all aware of and secondly,

normal distribution has some standard properties for which the remaining steps of our econometric analysis becomes easier.

So, we will discuss the importance of normality in a later part of our discussion. For the timing we just have to keep in our mind that error term captures the influence of omitted variable and secondly, we know the distribution properties of the error term that it follows a normal distribution with 0 mean and sigma square variance.

(Refer Slide Time: 17:47)



Once you formulate the statistical model, then the next step is basically to collect the data. When you are interested in estimating a consumption function, then you need to collect the data on income and consumption and how do you collect data? There are 3 ways by which you can collect data. The first type of data what you can collect is cross sectional data. There are three ways-Cross sectional data, Time series data and Panel data.

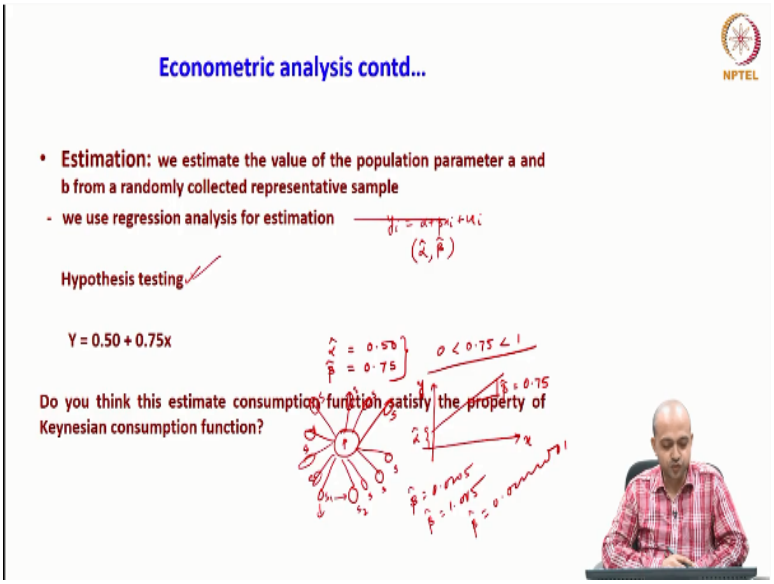
So, let us see each and every type of data because depending on your data structure we will have a separate econometric analysis. So, what is cross-sectional data? Cross sectional data is basically data collected across several individuals but for a particular period of time. For example, if we are interested in consumption function, let us say that

we are collecting income and consumption data from 500 individuals residing in Chennai in a particular point of time, let us say in November 2020.

So that means there is variation in your observation, in your data across individuals but there is no variation across time because this data is collected only in a particular point of time. But there is variation across individuals. Individual i might be having 500 as consumption and 1500 as income, and when you go to j^{th} individual, j^{th} individual's income is 2000 and consumption is 750. So, that means there is enough variation across individuals but there is no variation across time. If that is the structure of your data, then that is called cross sectional data. Then time series data. It is just the opposite of cross section. That means here you have variation across time period but there is no variation across entity. Why? Because you are collecting data only for a particular entity and that entity might be an individual, might be a state or might be a country, and you are collecting the data on income and consumption for the same entity over a period of time, let us say for the last 10, 15 or 20 years, then the data will become a time series data.

If you combine this cross section and time series, then you are going to get a cross sectional time series data or panel data. That means in panel data you have characteristics of both time series as well as cross section. There is variation across entities and there is variation across time periods. If that is the case, then it is called panel data but for our course of basic econometrics what we are going to discuss here is called cross sectional data. That is why our entire focus would be on cross sectional econometrics and the other two types of data we are not going to consider here because they come under the scope of advanced econometrics. We are going to discuss only cross sectional data in this particular course.

(Refer Slide Time: 21:47)



Once you collect the data, the next step is basically estimation. So, what is estimation? That means you remember we have specified a statistical model $Y_i = \alpha + \beta X_i + u_i$. This is the model we are going to discuss and we need to estimate this $\hat{\alpha}$ and $\hat{\beta}$ and that is called estimation.

So, we are basically going to estimate the sample counter part of the true population parameter alpha and beta and what is the technique that we are going to use? We are going to use Regression Analysis. So, throughout our discussion on econometrics there is a single technique for estimation which is regression. The detailed procedure of regression we will discuss in a later part of our discussion but for the timing you just keep in mind that there is only one technique that is available to estimate this $\hat{\alpha}$ and $\hat{\beta}$ which is called regression analysis.

And we also have to keep in mind that we are going to estimate this alpha hat and beta hat from a randomly collected representative sample. That means from a given population, let us say from Chennai, if Chennai is your population, then from that population we can derive or we can draw 'n' number of samples of let us say 200 individuals. So we are going to collect the sample randomly. Why this is so? Without

random sampling we cannot proceed for remaining steps of our economic analysis that we will discuss later.

So, for the timing we have to keep in mind that we will collect a random sample from the population for our estimation and the estimation will be done by the regression technique or regression analysis. And once you estimate the model, the next step is the hypothesis testing. Now, for example, let us say that you have collected a sample of 200 individuals from the Chennai city about the income and consumption and then you have estimated this type of model. So that means here your $\hat{\alpha}$ equals to 0.50 and $\hat{\beta}$ equals to 0.75.

Now, from this $\hat{\alpha}$ and $\hat{\beta}$ do you think that the properties of Keynesian consumption function hypothesis that you made earlier is established or not? This is the result. Now, apparently it may look like our hypothesis is established because value of $\hat{\beta}$ is 0.75. That means when your income increases by 1 rupee, your consumption increases by 75paise which is a significant amount. That means income is having a significant impact on consumption. And secondly 0.75 is lying between 1 and 0. So, that means from this value of $\hat{\beta}$ we can say that the properties of Keynesian consumption function is satisfied because this is your $\hat{\alpha}$ and this is your $\hat{\beta}$ which is equal to 0.75. Let us say this is your consumption and this is your income. So apparently you may think immediately after estimation that the properties of Keynesian consumption function is actually satisfied.

But statistician and econometricians give some kind of caution to draw that type of inferences immediately after estimating the model. Why this is so? Because here you see that you have drawn only one sample from your population and in that population your $\hat{\beta}$ value is 0.75. Now it may so happen that if you draw another sample of 200 individuals from the same population, so if this is your population you can draw 'n' number of samples of 200 individuals, these are all samples from the population.

You have drawn only one sample say s1. You have drawn this sample and in this sample it is 0.75. Now it may so happen that if you collect another sample from this say s2 the $\hat{\beta}$ value is 0.005. So, you can easily understand that for a unit change in income, consumption changes by 0.0005 which is very, very insignificant amount.

That means this particular value of $\hat{\beta}$ which is 0.75, you may get purely by a chance factor. You have drawn one sample and in that sample it is 75 while there is no guarantee that in the next sample what you may collect of 200 individuals, $\hat{\beta}$ will again come out as 0.75 or 0.70 or 0.80 or it may so happen that $\hat{\beta}$ is even 1.005 in the next sample. That means I cannot rely upon the estimate that I got only from one sample.

What I need to guarantee about is that, even if I draw 'n' number of samples from the population, in all the samples my $\hat{\beta}$ value will come out to be significant. When I say significant that means it is significantly different from 0 because if the $\hat{\beta}$ value is 0, let us say 0.00000001, this is almost 0, and in that case what will happen, you have to say that my income does not have significant impact on consumption. So, that means we need to ensure that this 0.75 value what you got is not because of a chance factor but for a statistical reason. That means we need to establish statistical significance of $\hat{\beta}$.

So far what we have seen here is only mathematical significance. That means the magnitude of $\hat{\beta}$ is different from 0.75 is mathematically different from 0 but we need to ensure it is statistically different from 0. That means even if you collect other samples, in all other samples it will come out to be significantly different from 0 and how will you ensure that?

After estimation you need to go for this step which is called hypothesis testing, and that is why when you put this question whether you think this estimated consumption satisfied

the property of Keynesian consumption function, then we cannot say that this is actually satisfying the property of Keynesian function unless and until we go for our next step which is hypothesis testing.