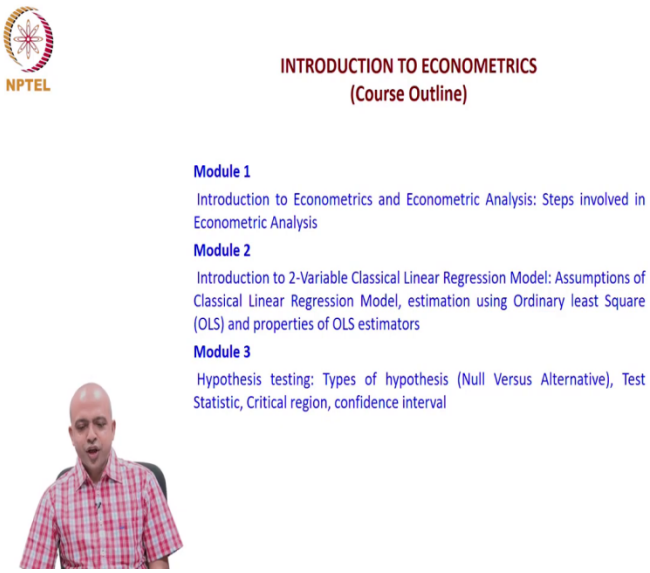
Introduction to Econometrics
Professor. Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras
Lecture No. 01
Introduction to Econometrics and Econometric Analysis Part-1

(Refer Slide Time: 00:14)



Welcome to the Introduction to Econometrics course. Today we are going to discuss about the course outline-what are the things that we are going to discuss in this course. There are around 9 modules that we are going to discuss in this course and I will let you know module by module what is there in each and every module to discuss.

Now, our discussion in module 1 starts with an introduction to econometrics and econometric analysis. Here, I will let you know basically what is econometrics and what does an econometric analysis mean and what are the different steps that is involved in this econometric analysis. Econometrics is basically an application of mathematical and statistical tools with an objective to measure the empirical validity of economic theory. That is the definition of econometrics and there are several econometricians who say that econometric analysis is just like a scientific analysis.

So, that means here also like any other science disciplines we observe certain phenomena-we observe behavior of individuals in our society. The unit of analysis could be individuals or firms

or cities or states or even a country. So, we will first hypothesize some kind of relationship. Then we will try to formulate a mathematical model, try to estimate the relationship and then we will go for higher forecasting or policy making. These are the different steps involved in an econometric analysis and we will discuss in detail about each and every step in module 1.

Now in module 2, what we are going to discuss is introduction to two variable classical linear regression model. Now, at the heart of any econometric analysis there is regression. Regression is the technique-a single most important technique available in the entire econometric literature through which we are actually going to estimate the economic relationship and that is why in module 2 I will let you know what exactly is regression, what is the difference between regression and correlation (another related but different concept) and then what is the original meaning of the word regression, what is the meaning in today's world about the word regression and then I will let you know the statistical procedure that we follow in a regression analysis.

We will be also discussing several assumptions before we actually estimate the model. Those assumptions actually describe an idealistic world based on which if we estimate our model applying regression, then the regression coefficients will give certain desirable properties. The desirable properties will also be discussed in module 2 and we will be applying the Ordinary Least Square (OLS) technique to estimate the relationship in two variable classical linear regression model.

So, OLS is the technique that we are going to learn here. After estimating the model we are also going to learn 3 important properties of our estimates- unbiasedness, efficiency and consistency in module 2. Then in module 3 we will be discussing about hypothesis testing and the need of hypothesis testing in econometric analysis.

So, in econometrics, generally the ultimate objective of any econometric analysis is to draw inference about population with the help of a sample. For example, let us say that we are going to infer something about the consumption behavior of the population for entire Tamil Nadu. For that purpose we will not be collecting data of everyone living in Tamil Nadu because collecting data for each and every individual from Tamil Nadu is a time consuming and huge task.

Rather what we do is that we will collect a random representative sample from the entire population and then we will estimate the regression function from the sample only. At the end of

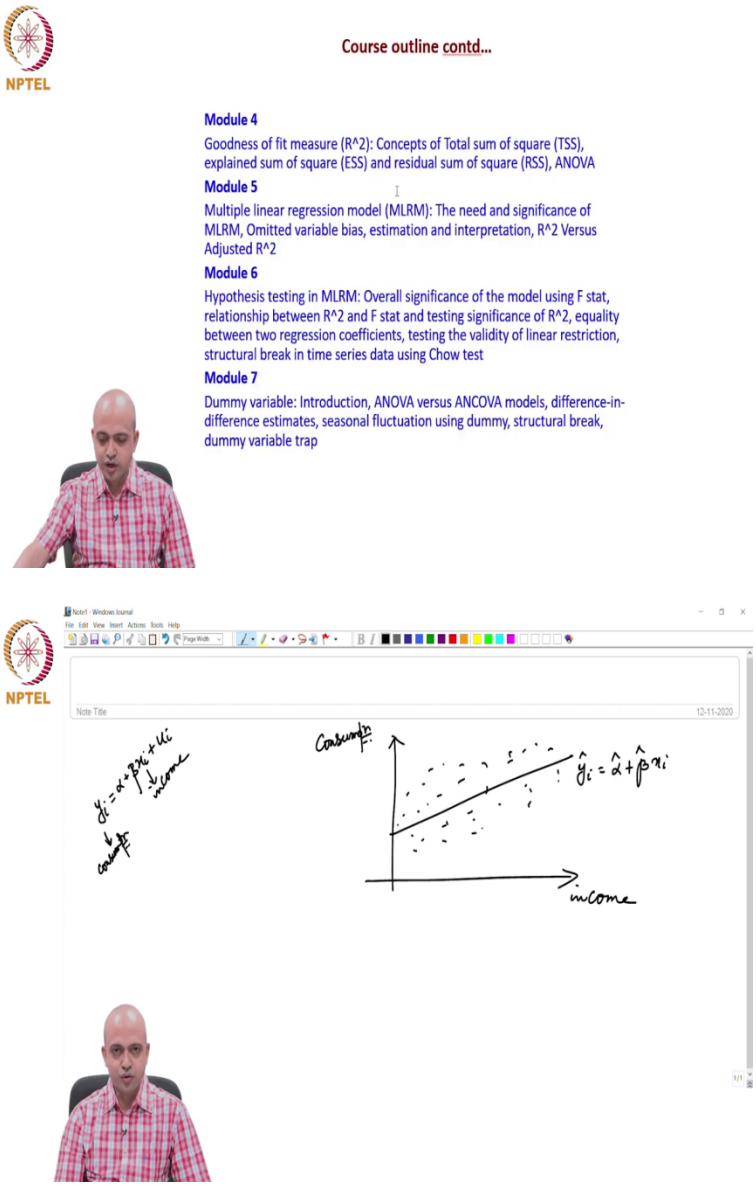
the day we will be getting sample statistic-which is basically the sample counter part of the population parameter and this sample statistic will help to make inference about the population parameter.

Since from the population you can draw a sample in ‘n’ number of ways, what we need to know is whether the result that we get out of one sample is valid for each and every sample that we collect from the population. So, that means we have to rule out the possibility of a chance factor that may lead to a particular result that is observed from a particular sample.

For example, suppose you are interested in consumption behavior thinking income is a significant factor impacting consumption. From that particular sample collected, the equation estimated using regression shows a significant impact of income on consumption but there is no guarantee that you will be getting the similar type of significant impact of income on consumption if we collect another sample or n-1 number of other samples from the same population.

So, that means in the process we have to rule out the chance factor. We have to establish the statistical significance of income and consumption. This is where the hypothesis testing becomes important and enters into the picture. So, we will be discussing different types of hypothesis- Null hypothesis and Alternative hypothesis, different concepts that are very important like test statistic, critical region and confidence interval. These are the 3 important concepts that we must learn to know in the procedure of hypothesis testing which will be discussed in module 3.

(Refer Slide Time: 07:39)



Then in module 4 we discuss the Goodness of Fit measure. Here we collect a sample and we are going to estimate a regression line. For an example let us say we measure income (independent variable) in the X axis and consumption (dependent variable) in the Y axis. The estimated regression line is basically that you have estimated which is $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$

Let us say, the original equation was $Y_i = \alpha + \beta X_i + u_i$, where Y is basically consumption and X is income of individual i. This is the function you have estimated. Now, this regression line when you collect and estimate, the obtained regression line represent the data as the scatter plot.

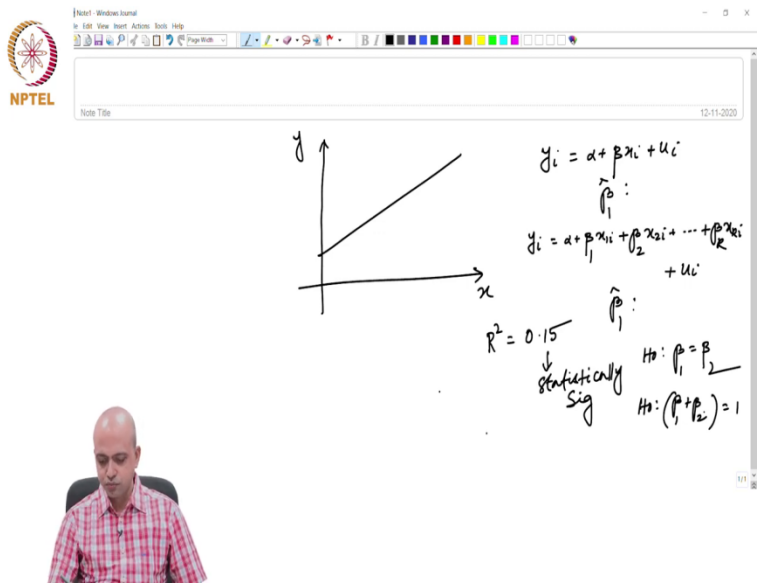
So, that means you are trying to represent the raw data with the help of this regression line and in that sense I can tell you this line what you are fitting is basically you are trying to fit a line to represent the data what you have in your hand. So, then the question that comes to our mind how good is this line to represent the data that you have collected and how good is this line to represent the sample that you have collected?

So, that means after estimating the Y_i, that is after estimating the regression line or regression function or sample regression function we must know what is the goodness of fit that is how good this line represent your data. R^2 is the goodness of fit measure and that is what we are going to learn in this module.

So, along with the goodness of fit measure there are 3 related concepts-Total sum of square, Explained sum of square and Residual sum of square. In econometric analysis basically the objective of the regression analysis is to explain the total variation in the dependent variable with the help of the explanatory variable that you have collected. So, in this example consumption is the dependent variable and income is the explanatory variable.

So, this R^2 is going to tell you what is the percentage of total variation in consumption that you are going to explain with income. That is what we are going to learn in this particular module.

(Refer Slide Time: 11:34)



So, your Y_i is the dependent variable and X_i is the independent variable. With the regression line our objective here is to explain the variation in Y with the help of X. In this context I would be explaining out of that total variation in y how much my model is able to explain given by explained sum of square and the residual if any that is given by residual sum of square.

So, that means the total variation in Y_i would be then explained by and can be decomposed as sum of explained sum of square and the residual sum of square. So TSS is actually equals to ESS plus RSS. We will also learn about the ANOVA-Analysis Of Variance-that means what are the different components and what is their contribution, and we will be deriving several other concepts from this ANOVA table. So, that is what we are going to learn in module-4.

Then in module 5 we will extend our two variable classical linear regression model into multiple linear regression model. The example of income and consumption is called a two variable linear regression model because you have one independent variable and one dependent variable. That is why it is called two variable classical linear regression model. Then we extend this into a multiple linear regression model and in that multiple linear regression model first of all what we learn is, what is the need and significance of a multiple linear regression model.

Generally, we think that if we have data on multiple variables then we can easily run a multiple linear regression model. That is all right but first of all we need to learn what is the objective and

significance of a multiple linear regression model. We would be learning the importance of multiple linear regression model in the context of omitted variable bias so, that means there might be several other factors apart from income that might also have some impact on consumption.

If, we do not include these other factors then your model may suffer from an omitted variable bias. What is the omitted variable bias, how to measure the extent and amount of omitted variable bias, those things also we will be learning in the context of multiple linear regression model. Then we will learn the estimation and interpretation of the coefficient.

For example, when you talk about the model $Y_i = \alpha + \beta X_i + u_i$, this is called a two variable classical linear regression model and the interpretation of $\hat{\beta}$ in this two variable model, and the interpretation of $\hat{\beta}_1$ in a multiple linear regression model like $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_k X_{ki} + u_i$ is actually different.

So, we will learn what is the difference in the interpretation of $\hat{\beta}$ in the context of multiple linear regression model and two variable linear regression model. Then we will be learning another concept called Adjusted R^2. In this context it means that if you increase the number of explanatory variables in your model, it may so happen that the explanatory power of your model or R^2 will increase but that does not mean that we will be entering into a game of maximizing R^2. So, that means we cannot really compare the R^2 of two different models when your number of explanatory variables are different.

So, in a two variable model and three variable model R^2 is not comparable because higher the number of explanatory variables obviously higher would be the R^2. So we need to derive a different but related concept that will help us to compare the goodness of fit measure across different models with different number of explanatory variables. Here the adjusted R^2 enters into the picture. We will learn the definition and meaning of adjusted R^2, and the relationship between R^2 and adjusted R^2 in the context of multiple linear regression model.

Then in module 6 once again we will bring hypothesis testing. So, the hypothesis testing that we applied in a two variable linear regression model is basically to test the individual significance of the variable. Here, apart from testing the individual significance of the variable what we also learn is the overall significance of the model. That means when you have several variables in the model how to test the overall significance of the model using F statistic is learned here in module 6.

Then we will also learn a fantastic relationship between the F statistic and R^2 and that particular relationship will help us understand even the significance of R^2 also. For example, let us say that you have estimated a consumption function and your R^2 turns out to be 0.15. So, that means out of the total variation in Y your model is able to explain only 15 percent.

Now, if somebody ask you is it a good measure or a bad measure, always you have to keep in mind that this value is giving only a mathematical value. That means by one value we do not really know whether 15 percent that we are going to explain is a sufficient amount or not. So, that means we need to know whether this is a statistically significant value. So that particular relationship between F and R^2 will not only help us to understand or estimate the overall significance of the model but also will tell whether a particular value of R^2 is statistically significant or not.

So, that is what we are going to learn here after deriving the important relationship between F and R^2. Then we will also learn about the equality of the two regression coefficients and testing validity of linear restrictions, structural break in time series analysis using chow test. By testing linear restriction it may so happen that you are interested in whether both X_1 and X_2 have same impact on Y_i or not.

So, that means your null hypothesis may test whether β_1 equals to β_2 or not. This is called testing the equality of two regression coefficients. This is something which you can test. You can also test this type of hypothesis, let us say, I am testing this hypothesis $\beta_1+\beta_2=1$. So, this is called testing linear restriction.