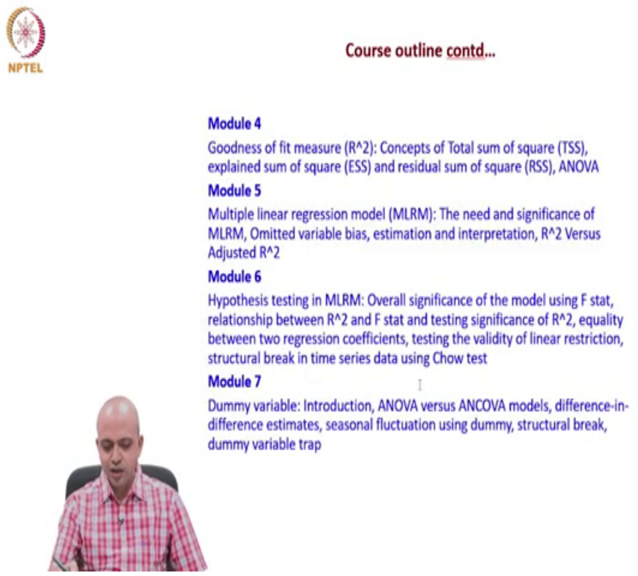
Introduction to Econometrics
Professor. Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology, Madras
Lecture No. 02
Introduction to econometrics and econometric analysis Part - 2

(Refer Slide Time: 00:14)



Course outline contd...

Module 4

Goodness of fit measure (R^2): Concepts of Total sum of square (TSS), explained sum of square (ESS) and residual sum of square (RSS), ANOVA

Module 5

Multiple linear regression model (MLRM): The need and significance of MLRM, Omitted variable bias, estimation and interpretation, R^2 Versus Adjusted R^2

Module 6

Hypothesis testing in MLRM: Overall significance of the model using F stat, relationship between R^2 and F stat and testing significance of R^2, equality between two regression coefficients, testing the validity of linear restriction, structural break in time series data using Chow test

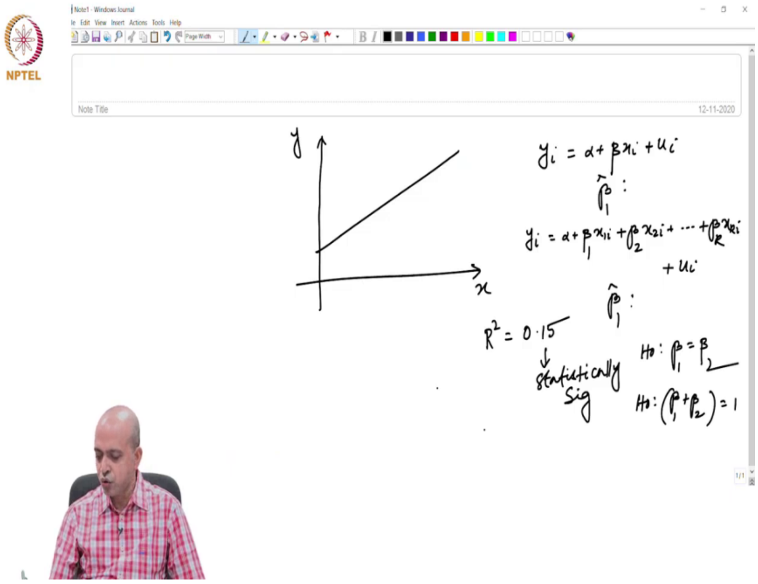
Module 7

Dummy variable: Introduction, ANOVA versus ANCOVA models, difference-in-difference estimates, seasonal fluctuation using dummy, structural break, dummy variable trap

So, all these things you can test could be learned in module 6. Then, there is something called structural break in time series analysis. For example, suppose we are estimating a consumption function where the time span is from 2000 to 2020 for the entire Tamil Nadu population. As you know, in 2008 there was a financial crisis. Now, it may so happen that because of this financial crisis, the income consumption relationship might have changed drastically in the post-financial crisis period.

So, in that case, how will you test the structural stability in a time series data, and there comes the F statistic. We will be using the F statistic suggested by the famous econometrician Chow, which is popularly known as Chow test to test the structural break of time series analysis in module 6. Then, in module 7 we are going to discuss about Dummy variable. The example what we are discussing in this income consumption relationship is where both income as well as consumption are quantitative variables.

(Refer Slide Time: 01:39)



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

$\hat{\beta}_1 :$

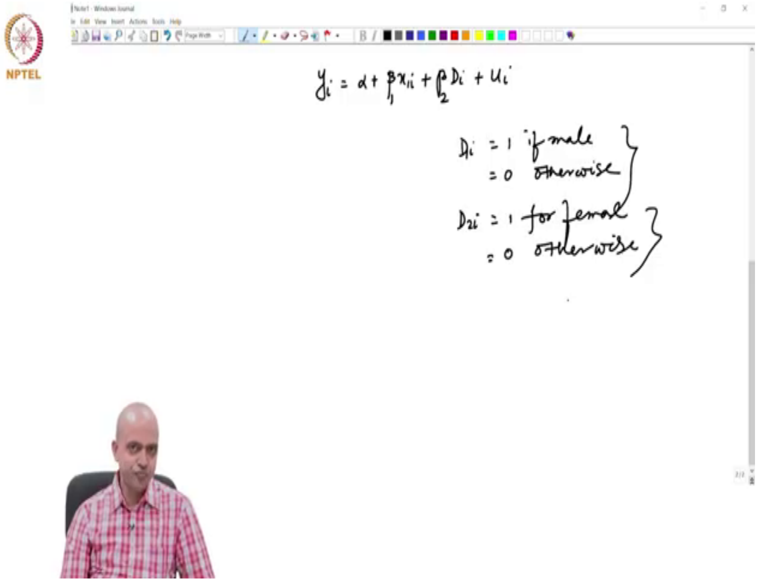
$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + u_i$

$\hat{\beta}_1 :$

$R^2 = 0.15$ ↓ statistically sig

$H_0 : \beta_1 = \beta_2$

$H_0 : (\beta_1 + \beta_2) = 1$



$y_i = \alpha + \beta_1 x_{1i} + \beta_2 D_i + u_i$

$D_i = 1$ if male
$= 0$ otherwise

$D_{2i} = 1$ for female
$= 0$ otherwise

But it may so happen that sometimes our variable of interest is not always quantitative. For example, let us say I am giving a model where $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 D_i + u_i$ .What is D_i?

Suppose, now I am hypothesizing that apart from income, there might be a significant difference in the consumption between male and female. So that means I am hypothesizing that gender might also play an important role in consumption but gender is not a quantitative variable. The technique that we have learned so far is applicable only when your variable is quantitative in nature but many times as I said our variable of interest would be qualitative in nature.

Here is a case. Let us say that you are interested to check whether gender has an impact on consumption patterns or suppose you are interested in estimating wage function and you are interested to check whether there is labor market discrimination across the gender. So, it might so happen that given the same amount of qualification male workers are earning a higher salary than the female workers. That means there is gender discrimination in the labour market. How are we to answer those questions? Because this gender is basically a qualitative variable, you need to somehow make it quantitative one. How will you quantify? And there comes the dummy variable technique. You can easily define a variable D_i equals to 1 if male, and 0 otherwise. You have converted your qualitative variable into a quantitative one now and you can apply the technique what you have learned so far in this context.

So, this is how the dummy variable technique is very important because many times our explanatory variables of interest are qualitative in nature and dummy variable converts that qualitative information into quantitative one and that is what we are going to learn in this module 7 which talks about the dummy variable. So we will give an introduction of the dummy variable and then discuss different types of dummy variable-ANOVA versus ANCOVA model.

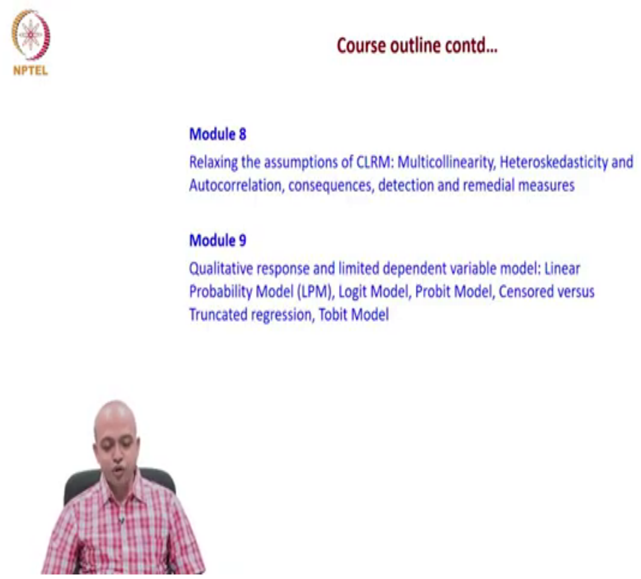
It may so happen that in your model you have only a qualitative variable. That type is called ANOVA model. Then we have ANCOVA model wherein you accommodate not only the qualitative variable but also quantitative one. That means a mix of qualitative and quantitative variables in the model makes it ANCOVA. Then we will also learn about how to apply this dummy variable technique for impact evaluation and that is called Difference in Difference estimates. For example, let us say that your research objective is to check what is the impact of a job training on labour productivity. Suppose in your organization some of your employees have gone for a job training and you want to see whether the performance of those who have participated in the job training has improved significantly compared to those who have not participated. That is called some kind of impact evaluation. And in the context of impact evaluation, we will see how the dummy variable technique is quite helpful to estimate the impact of that particular policy. Then, we will also discuss about the seasonal fluctuation in this particular context.

Sometimes it may so happen that the variable of your interest is showing some kind of seasonal fluctuation. For example, if you are estimating a demand for let us say, household utilities or home appliances, you will see that there are seasonal fluctuation like in summer season there might be a drastic increase in the sale of refrigerators and then air conditioning and so on and so forth, then also at the time of Diwali, then in Christmas time etc. So in different seasons you will get some kind of seasonal impact on your dependent variable. How will you identify those seasonal fluctuations? There also we will be using the dummy variable technique to estimate that type of seasonal fluctuation.

And then, we will also give an alternative to Chow test to test the structural break what I was talking about in the module 6. So, that means there we were estimating the structural break using the F statistic that is we are estimating a restricted versus unrestricted model we and then applying the Chow test which is nothing but a F Test. That is a time-consuming process if you apply Chow test for structural break analysis. Chow test has several limitations also in the process of estimating the structural break. We can overcome these limitations by applying our dummy variable technique. We will see how dummy variable technique can be applied to identify the structural break analysis. Then, we will discuss about dummy variable trap a rule which says that while assigning the dummy variable you should include dummy variable which is 1 less than the number of categories.

For example, here when you have 2 categories-male and female, I have assigned only 1 dummy. So, if you include dummy like this, let us say this is D_{1i} , and D_{2i} equals to 1 for female and 0 otherwise. That means I have 2 categories and I have introduced 2 dummies also and you will end up with a problem which is called dummy variable trap. So, here we will discuss what is dummy variable trap and what is the exact problem of dummy variable trap. All these things would be discussed in module 7.

(Refer Slide Time: 9:08)



Course outline contd...

Module 8

Relaxing the assumptions of CLRM: Multicollinearity, Heteroskedasticity and Autocorrelation, consequences, detection and remedial measures

Module 9

Qualitative response and limited dependent variable model: Linear Probability Model (LPM), Logit Model, Probit Model, Censored versus Truncated regression, Tobit Model

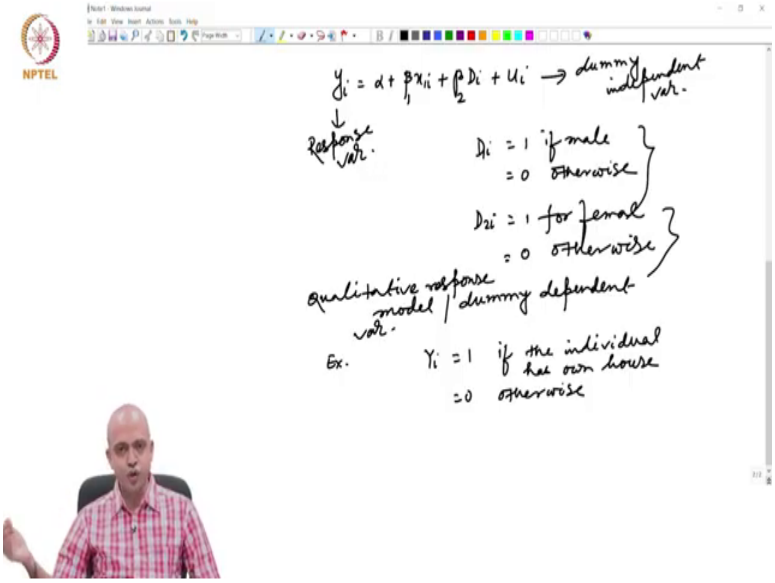
As we said earlier before estimating the classical linear regression model, we assume certain things. Why do we assume certain things? Because those assumptions make the world realistic and idealistic, and in that ideal world only if you estimate your function, the estimated coefficient will exhibit certain desirable properties which are unbiasedness, efficiency, and consistency. But there is no guarantee that in real-world when you collect a sample and data, your sample data will always follow the assumption what we make. Rather most of the time at least 1 or more than 1 assumptions would be violated. And if the assumptions of classical linear regression model is violated then you no longer expect that your estimates will exhibit the desirable properties. They will not exhibit the desirable properties and what will happen if the assumptions are violated? In that context, we will be relaxing 3 important assumptions multicollinearity, heteroskedasticity, and autocorrelation which will be discussed in module 8.

Multicollinearity basically says that when you include the variables there should not be perfect linear relationship among the two variables but in reality it may so happen that two of your explanatory variables are highly correlated. What will happen? What is the consequence of multicollinearity problem? How will you detect that the data is actually suffering from multicollinearity problem and what are the remedial measures? If you are sure that your data is suffering from multicollinearity problem, then how would you rectify that problem so that your estimates exhibit the desirable properties like unbiasedness, efficiency and consistency.

Then another important assumption is called heteroskedasticity which assumes that error variance is constant. That means even if your X values increases or decrease across individuals the variance of the error term for the i^{th} individual and j^{th} individual are actually same. So, this is called constant error variance assumption but many times when you are working with empirical data you see the constant error variance assumption is actually violated. And in that situation, we call it heteroscedasticity. We will all again learn the consequences of heteroscedasticity, how to detect heteroscedasticity and the remedial measures. Lastly, we will discuss about autocorrelation problem. Particularly when you are working with time series data we assume that no 2 error terms that is error term for the t^{th} period and error term for the t+1 period should not be correlated. But in reality when you are working with time-series data, many times we observe that the inner terms actually show some kind of pattern. That means they exhibit some kind of interrelationship between these 2 time periods. So error terms of 2 different time periods get correlated and that is what is called autocorrelation. Once again like multicollinearity and heteroscedasticity, in the context of autocorrelation we will learn the consequences of autocorrelation, how to detect autocorrelation and the remedial measures.

The last module is 9 where we are going to discuss about qualitative response or limited dependent variable model. For an example, when we talk about the dummy variable.

(Refer Slide Time: 13:42)



$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 D_i + u_i \rightarrow \text{dummy independent var.}$$

$\downarrow$ Response var.

$D_i = 1$ if male
$= 0$ otherwise

$D_{2i} = 1$ for femal
$= 0$ otherwise

Qualitative response model / dummy dependent var.

Ex. $Y_i = 1$ if the individual has own house
$= 0$ otherwise

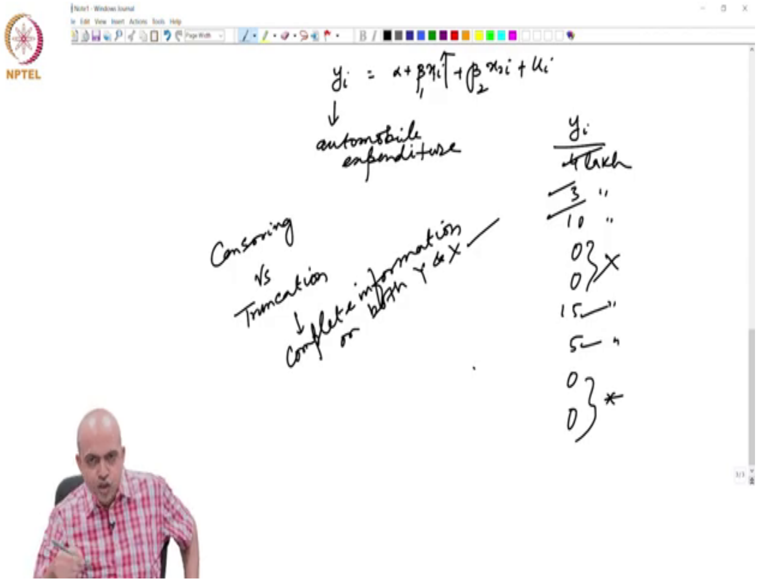
Here we say that, if you look, your dummy variable is an explanatory variable. So this is called dummy independent variable meaning some of your independent or explanatory variables are qualitative in nature but it may so happen that your dependent variable Y is also dummy for an example. That is why this Y variable is actually called response variable. That is why when your response variable becomes qualitative it is called qualitative response model or this is also called dummy dependent variable.

For example let us say that we are interested to see what are the factors that determine whether the individual will have his or her own house or the individual will stay in a rented apartment or not. What I am saying is that you are interested to estimate the factors that will determine whether an individual will posses his or her own house or not. That means here your dependent variable is qualitative in nature. That means you go to an individual and you ask whether he/she has their own house or not. The individual will say either yes or no. That yes or no response is basically a dummy kind of situation. That means Y_i equals to 1 if the individual has their own house and 0 otherwise. In this context our dependent variable itself is a dummy variable and that is why this is called qualitative response model or dummy dependent variable.

So, the objective of these models is to estimate basically the probability of owning a house given the individual socio-economic and demographic factors. If I know the individual's income, education, age, gender, so on and so forth, what is the probability that the individual will own a house? So, 3 types of probabilistic models we are going to discuss here. They are actually the Linear Probability Model, Logit model, and Probit model. Linear probability model is a model where probability of owning a house or probability of owning a car is linearly characterized that is the probability is a linear function of the explanatory variable. For example, probability of owning a house is let us say a linear function of income. That is the starting point of a probabilistic model but this probabilistic model LPM has several limitations.

First of all, when you characterize probability as a linear function of X, that means, we assume that as income increase probability of owning a house keep on increasing but that is an unrealistic situation. When your income increases from 10,000 to 20,000, 20,000 to 25,000, 25,000 to 50000, 500000 to 1 lakh, 1 lakh to 1 lakh 50, and so on, after some point of time you will see that probability is not changing because it does not work in that way. When your income reaches to 1 lakh 50,000 per month, probably at that income range everybody will have his or her own house. So there is not much of change in probability once you achieve a certain level of income. That is why linear characterization of probability with the explanatory variable does not give a realistic picture. It has several other limitations also because of which econometricians have suggested two alternative models where the probability is nonlinear function of the independent variables. In that context logit and probit models would be discussed. Then we will be discussing another important type of model which is called Tobit model or censored model.

(Refer Slide Time: 19:17)


NPTEL
$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$
automobile expenditure
Censoring vs Truncation
complete information on both Y & X
y_i
4 lakh
3 "
10 "
0
0
15 "
5 "
0
0

Sometimes it may show up that we are not interested in probability. Rather let us say this is your $Y_i = \alpha + \beta X_i + \beta_1 X_{1i} + \beta_2 X_{21} + u_i$. This Y_i is, let us say, is your automobile expenditure. So, here I want to know as income increases what is the change in an individual's automobile expenditure. And if you recall your understanding of basic microeconomics it is nothing but elasticity- elasticity of automobile expenditure with respect to income. When your income changes by 1 percent, what is the change in your automobile expenditure.

So, what we will do? We will go to each and every household and we will ask what is their expenditure for automobile? But in that, let us say you are asking the question for 400 individuals and out of those, let us say, 100 individuals do not have any vehicle at all. So let us say the automobile expenditure is 4 lakh for first individual and after that 3 lakhs, then 10 lakh, then some individual will say 0 meaning he/she does not have any automobile expenditure, then again 15 lakh, then 5 lakh, then again 0, 0, 0. So we will observe automobile expenditure only for those individuals who are actually having an automobile and others, since they do not have any vehicle, their automobile expenditure is 0. If that is the case how are you going to estimate the model? Are you going to delete those observations with 0 automobile expenditure? And if you delete the observation with 0 automobile expenditure, then your estimate is going to be biased. That means this is a typical case of censoring. We have to censor our observations and our sample would be a censored sample.

All the observations are that means we have limited information on Y_i. This household we have information, this household we have information but these two households I do not have information because they have 0. Again, these two households I have 0 information that means limited information. So, when you have limited information on your dependent variable but you have information for all the households regarding their income, gender, education, so on and so forth, we have limited information or dependent variable, but complete information on the independent variable. This is a case which is called censored regression model.

So, we need to apply censored regression model characterized by Tobit model which is suggested by James Tobit. So, that is why it is called Tobit model. So, we cannot simply throw out those observations with 0 automobile expenditure and estimate the model using the Standard Euler's technique that we learned earlier. If you do so, then your estimate will suffer from a bias. To overcome that we need to apply Censored Regression model or Tobit model. We will be discussing censored regression model here. We will also learn about the difference between censoring and truncation. The truncated regression model is another important model and will not be discussed here but at least we should know the difference between censoring and truncation. So censoring means you have complete observation on X but you have limited information on Y.

Now suppose after collecting the sample you do not want to include those observations with 0 automobile expenditure. Then what you do is you delete this observation and after deleting this observation, your sample would become a truncated sample. Since you are not including non-ownership of a car in your sample, then your sample becomes a truncated sample of the population. And you need to apply the truncated regression model in estimating this to get unbiased estimate. That is basically the difference between censoring and truncation.

Mostly truncation happens in the pre sampling period itself. So, for example, my objective is to estimate the elasticity of automobile expenditure only for those households who are having vehicle. So, I will go to the individual and ask if you have a vehicle and then I will interview. If they say no, I will not interview the household at all. So, at the end of the day I will have complete information on both Y and X but in case of censoring, I was having only complete information on X but not on Y. That is the difference between censoring and truncation. We need to understand in detail and then we need to estimate and interpret the coefficients.

And these probabilistic models are quite different from the standard econometric models in terms of interpretation because here you are going to estimate the probability of owning a house. Then we will also discuss how the estimated coefficients are not a direct measure of marginal effect like the standard econometric model. Goodness of fit method is also quite different in the context of probabilistic model as compared to the standard econometric model. Those things we are going to discuss in module 9 which is qualitative response models.

So, Linear Probability Model, Logit and Probit are called qualitative response model while the Tobit Model is called Limited Dependent Variable Model. That is why I said qualitative response as well as Limited Dependent or Tobit Model-both the things would be discussed in module 9. So, this is going to be the entire syllabus of Introduction to Econometrics course. So, here my objective would be to give you a brief overview of the subject to make a foundation in econometrics so that you can apply your knowledge of econometrics in the real world, whenever your interest is to estimate some kind of relationship between dependent and independent variable and that relationship might not come from not only from economics but also from any discipline, any other social science, engineering and pure science discipline. Also, they are applying econometric tools to estimate some kind of relationship. So, whenever you have some query about the causal relationship, cause-effect relationship, and you want to estimate how much is the impact of the cause on effect, this is going to be useful.

So, throughout this course, I will take you to the theory as well as applied econometrics. So, that means you will learn theoretical portion as well as the application of econometrics using statistical software. This is going to be definitely an interesting and rewarding course in a sense that there is a huge demand for this course in academia as well as industry. With this I am just concluding our introductory section. From next day onwards we will be actually discussing all the subjects outlined in the course, module by module. I wish you all the best in learning and please put your effort because this is the course which is quite different from others-little involved, requires to divide reading and understanding but if you do so, if you can invest time, the reward what you are going to get out of this course is huge.

So, with this we are closing our discussion of introduction outline of the course. And from next class onwards we will be actually discussing the subject matter module by module. Thank you. Happy learning.