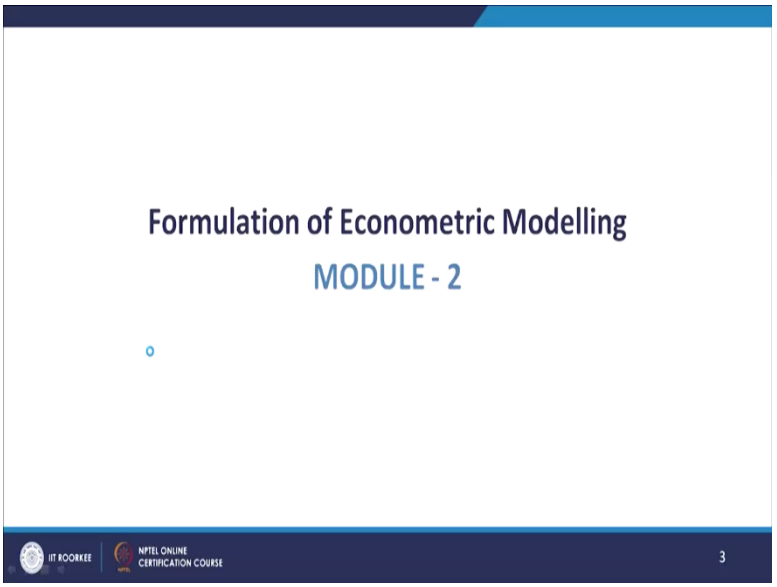
Econometric Modelling
Professor. Sujata Kar
Department of Management Studies
Indian Institute of Technology, Roorkee
Lecture No. 02
Formulation of Econometric Modelling

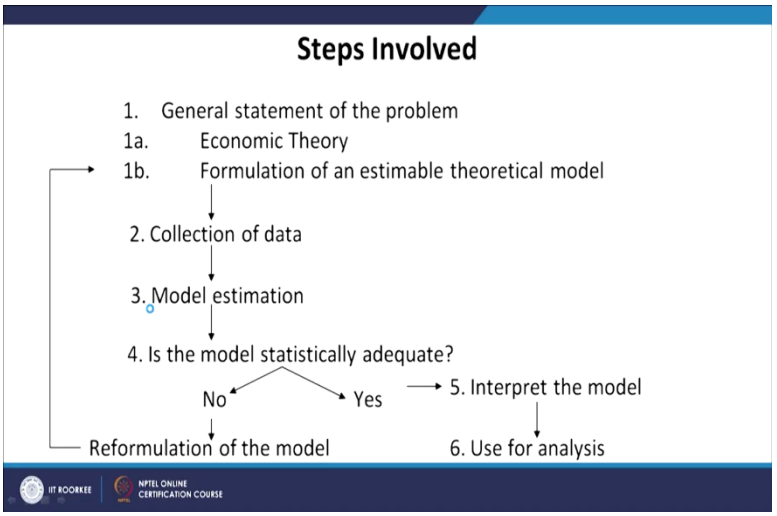
Hello, this is Module 2 of the course, Econometric Modelling.

(Refer Slide Time: 0:39)



Module 2 is on the formulation of econometric modelling. In Module 1, we have discussed that what is basically econometrics, what are its types or categories and what are the areas of applications, and took up some basic examples. Now, I am going to discuss how we formulate econometric modelling, ideally what are the steps involved in econometric modelling.

(Refer Slide Time: 1:13)



So, the first step is, a general statement of the problem. We have to first choose an economic theory or we can also go for the formulation of an estimable theoretical model. Most often the ideas are guided by some economic theory that what do we expect? For example, there is, there are so many things around us. If I consider inflation or price level and say the growth of an economy, so do we expect any relationship between them?

What kind of relationship we expect between them, they are ideally given by our economic theories that whether we can expect any relationship between them or not. So, that is why on the basis of economic theory, we got to first define a general statement of the problem. Once the general statement of the problem is defined, then we need to collect data.

So, ideally, a general statement would actually have a lot of things to tell us. It has a lot of guidance to provide us. The economic theory would tell us, what are the variables involved in this process or in this theory. So, if we are looking for the relationship between inflation and output, so there are two variables involved, but then, if we expect that inflation is going to be impacted by some other variables also.

For example, I expect it to be impacted by the wages of individuals or average wage at the national level, then this is another factor, which is going to impact inflation. So, as a result of which I have a third variable. So, this economic theory is actually going to tell us what are the variables that we need to consider. Once the variables are determined then I need to look into the areas or the sources from where I am going to collect data.

The next step is the collection of data. Once we have collected the data, we know what is our problem and how we are going to approach the problem, we have the third step of model

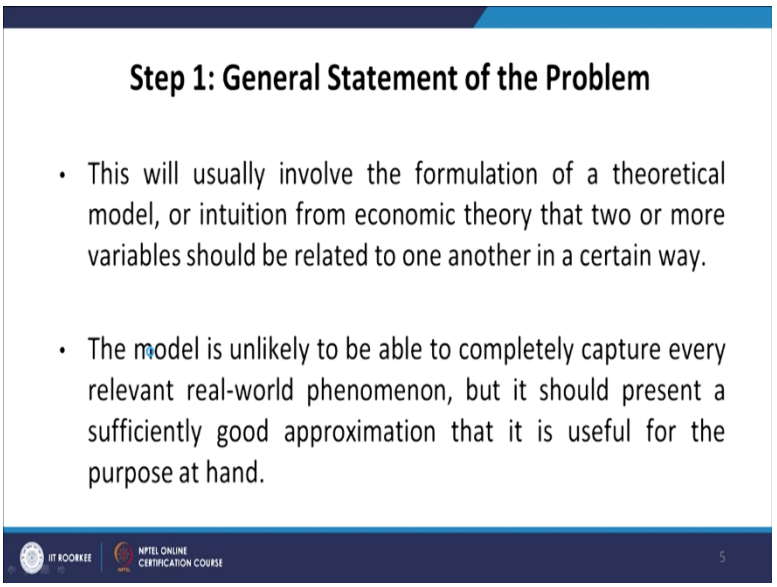
estimation. Here we see if the model is statistically adequate? Is it actually telling me or giving me what I am expecting it to be? Are all my requirements fulfilled? Am I happy with the model performance? If no, then we need to reformulate the model. So, we go back to the first step because there is some problem in economic theorizing and maybe as a result of which our model was invalidated. It is also possible that some theories which are yielding results in the context of some countries, may not yield results in my country or in some other country. For example, the kind of Phillips Curve relationship we can expect in the U.S. we cannot expect or we may not expect the same in the Indian context. So, it is always possible that the theory we began with that did not yield in our contest.

So, we need to reformulate and go back to our first step that is a general statement of the problem, reformulate our theory in order to understand what best suits the Indian context or the context I am probably dealing with. If the model is statistically adequate, so we have yes, in that case. Then, we need to interpret the model and use it for further analysis.

So, interpretation is simply, what are the observations I have got, and what these observations are telling me. On the basis of the interpretation, what actually I am obtaining out of it. So, maybe what are the policy decisions, one can go for policy recommendations, one can go for the insights this model is giving me, so they are basically analysis.

So, I use the interpretation of the model to go for analysis. And then, this interpretation of the model can further feed into the general statement of the problem. We can also observe some variations of that initial model or a general statement of the problem, and then accordingly the thing can be further modified added, or contributed. So, that is how contribution to the literature comes and things like that. Now let us take up these individuals' steps one by one.

(Refer Slide Time: 6:47)



So, first of all, general statement of the problem. This will usually involve the formulation of a theoretical model or intuition from economic theory, that two or more variables should be related to one another in a certain way, the model is unlikely to be able to completely capture every relevant real-world phenomenon, but it should present a sufficiently good approximation that it is useful for the purpose at hand.

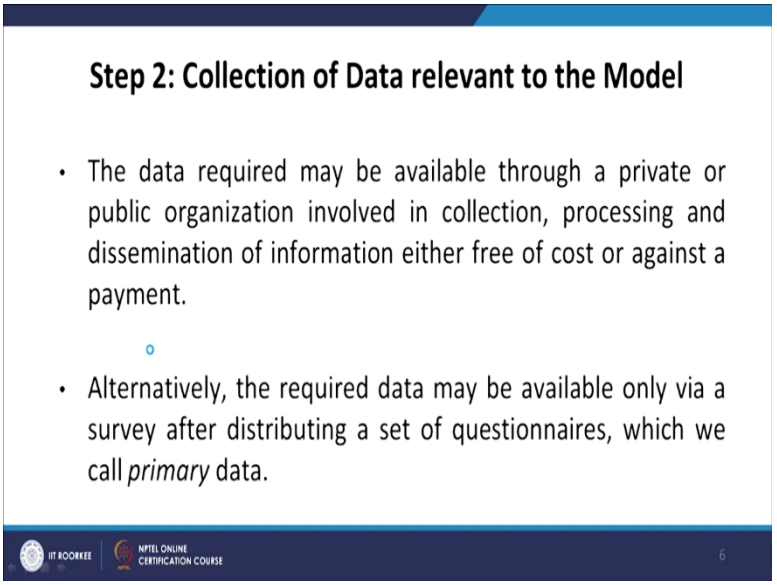
By this what is meant, is that it is actually not possible to capture real-world phenomenon through model building. For instance, we know that in the real world we see that, prices are changing on a daily basis or maybe we perceive that prices are changing. Then our general price index data also reflect certain changes in them on a monthly basis or on a yearly basis or on a quarterly basis. And while trying to explain that, we aim to find out the factors that might be contributing to it, using some macroeconomic data. We may not get sufficient results, which implies that there are certain things that are actually affecting this, but we are not able to get it. So, real-world phenomena are much more complex. When we try to model anything, then that possibly is the closest approximation or they are generally the closest possible approximation to what I am observing.

It is not possible to capture 100 percent of what we are observing. If we can capture 100 percent, then that is actually a very good model and but most often real-world observations are very difficult to model because we generally try to go as close as possible to an approximation.

So, the formulation of a theoretical model would come from economic theory, and that would just tell us that two or more variables are interrelated. So, we will try to see, how much only

these variables are contributing to the changes in one variable or how these three variables or two variables are related, and not a complex entire world model, where everything will be taken together. Those kinds of modeling are also done, but even then 100 percent accuracy or results are not possible. So, in this first step, we simply try to come up with an economic theory or general statement based on economic theorizing where certain relationships are laid down. And through that, we try to get as much close to the real-world phenomenon as possible.

(Refer Slide Time: 10:09)

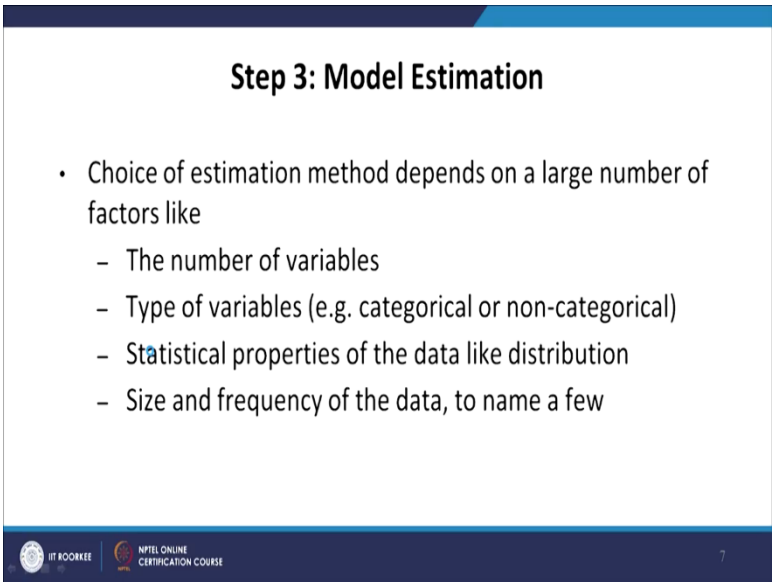


Then the next step is the collection of data. The data required may be available through a private or public organization, involved in the collection processing, and dissemination of information either free of cost or against payment. For example, there are certain organizations, specifically government organizations where data is available free of cost. For example, the Reserve Bank of India provides macro-level data or macroeconomic data, as well as financial data free of cost. Then, CSO, which is handled by a ministry also provides us economic survey, which is basically available free of cost, but then there are certain other private entities, a large number of private entities those who sell similar or more extensive data, for example, CMI Center for Monitoring the Indian Economy. It has many data, and those data need to be subscribed to in order to access them.

So, basically, data collected by some organization and made available to us is known as secondary data, because we are not the first person to collect the data. Someone has already collected it for us and just making it available to us. So, that is why they are secondary data.

While the primary data is basically the data, which is available only via a survey after distributing a set of questionnaires, so that is when we collect data first hand. We are the primary or the first collector of the data, as well as users of the data, then such type of data is known as primary data. So, in order to collect primary data most often what we do is that we prepare questionnaires and ask individuals to fill in those questionnaires. A Questionnaire can be on many matters. They may include certain personal and demographic information. Other than that, they may have information on problems that are specific to the research area, for example, that may be related to consumer behavior, consumers' perception, demand for a particular product, etc. So, that is how we can get questionnaires filled in by individuals. And then, when we consider those data, they are called the primary data.

(Refer Slide Time: 12:37)



After data collection, our third step is model estimation. Choice of estimation methods depends on a large number of factors like the number of variables, type of variables, statistical properties of the data, the distribution of data, size, and frequency of the data, etc. So, as all these characteristics or parameters change, the methodology can also change.

The simplest form of methodology that we think of is, suppose the first case, the number of variables. When we have only one variable and we want to test a very basic relationship, then we use a simple regression method. When the number of variables changes, we use the multiple regression method. There can be many other methods used with a large number of variables.

Type of variables: When the dependent and independent variables are categorical or non-categorical, then, of course, the models change, for example, when we have a dependent variable as a categorical variable (simply means that we have categories). So, instead of having numbers, for example, income levels are given by numbers of individual numbers.

So, I collect data on the income of individuals and that is given by numbers. But if my dependent variable is gender, then gender cannot be a number, it can be only two numbers. So, either 0 or 1. So, 0 suppose is female and 1 is male or vice versa. So in this case my dependent variable is a categorical variable. When a dependent variable is a categorical variable, the standard methods are actually not applicable. We need to use methods, which are very specific to this kind of problem.

When the independent variables are non-categorical then standard methods are most often applicable, but then there are also methods that help us improve the results if we use certain other methodologies. Of course, that depends on what are the problems that we have in at hand.

Then we have statistical properties like distribution. Most often we assume that the data follows a normal distribution. There are ways to test whether data follows normal distribution or not, and a large number of methods can be readily applied. But if the data does not follow normal distribution then either we can go for some modification, so that the data is made to follow a normal distribution or we may need to actually apply methods, which are applied when the data does not follow a normal distribution.

So, statistical properties like distribution of data are actually something very important and that helps us in determining, which method to be applied. The sample size is another thing, which also helps us in understanding which methods to be applied. For example, these days when we deal with microdata, but at the aggregate level then the data sizes are very large. When the data sizes are very large, for example, data running into lakhs of people then we can also use some advanced methodology. So, that is how statistical properties matter.

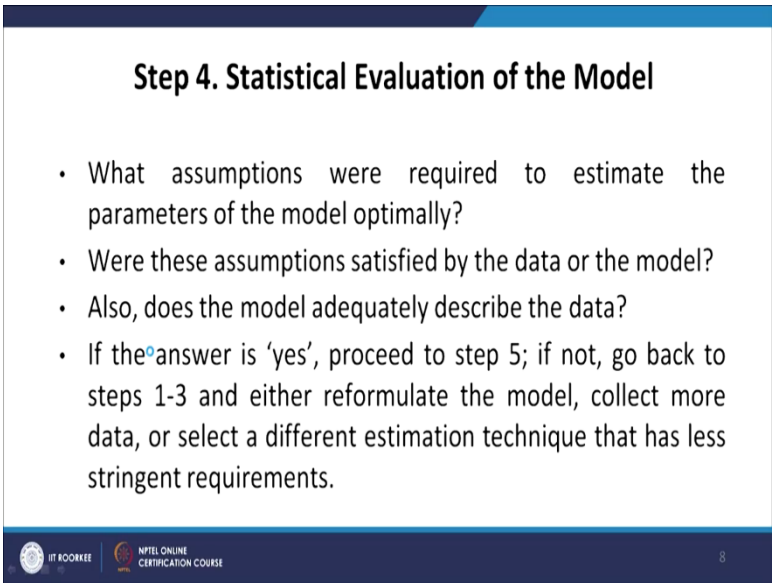
And, finally, we can also talk about the size and frequency of the data. Size I have already mentioned, frequency is another aspect, which is very important. Frequency measures how frequently we are observing the data. So, when we consider stock prices, then that data can be as frequent as hourly data or it can be daily data and then things can be averaged or summed up to weekly data, monthly data, annual data, and things like that. But for some of the variables, we do not have very high-frequency data.

So, stock price data can be called very high-frequency data. Low-frequency data is on the other hand are, for example, gross domestic product. We have data available for the quarter and then annual data. Inflation or price level data are available on a monthly and weekly basis. So, this is basically the frequency of data.

Now, when we analyze the data, its frequency matters, because, depending on how frequently we observe a particular data, we can actually change the methodology, because not all methods are appropriate for high-frequency data. The methods that we commonly learn are appropriate only for low-frequency data, most often annual data. Even for quarterly data or monthly data, we should apply different methods, so that they capture the characteristics of the data better and provide us with better results.

So, model estimation basically takes into consideration these factors in choosing which model to go for. Once we have chosen the model and estimated the model then we need to go for a statistical valuation of the model.

(Refer Slide Time: 18:30)

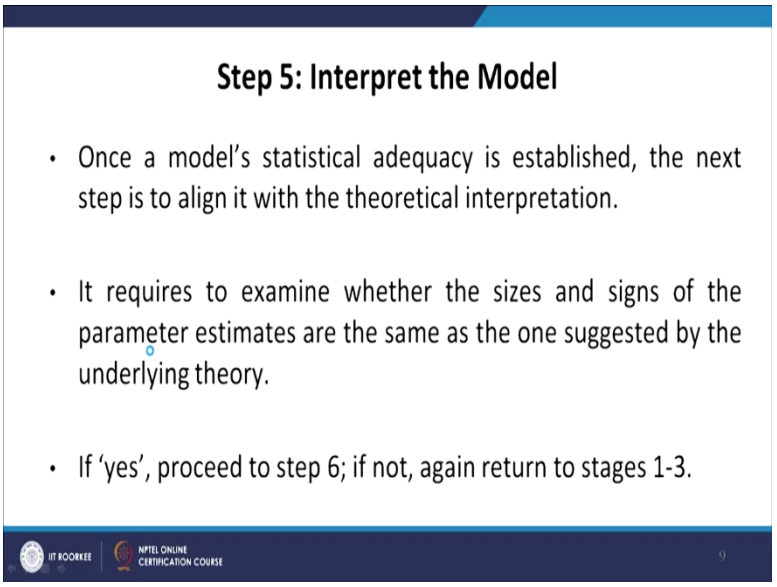


What assumptions are required to estimate the parameters of the model, optimally? So, when we go for evaluation of the model then we basically reassess the assumptions required to estimate the parameters of the model optimally. Where these assumptions satisfied the data or the model? Did my data and the model actually fulfill the assumptions that we made? Also, does the model adequately describe the data?

If the answer is yes, then we proceed to step 5. If no, go back to steps 1 and 3 or formulate the model, collect more data or select a different estimation technique that has less stringent

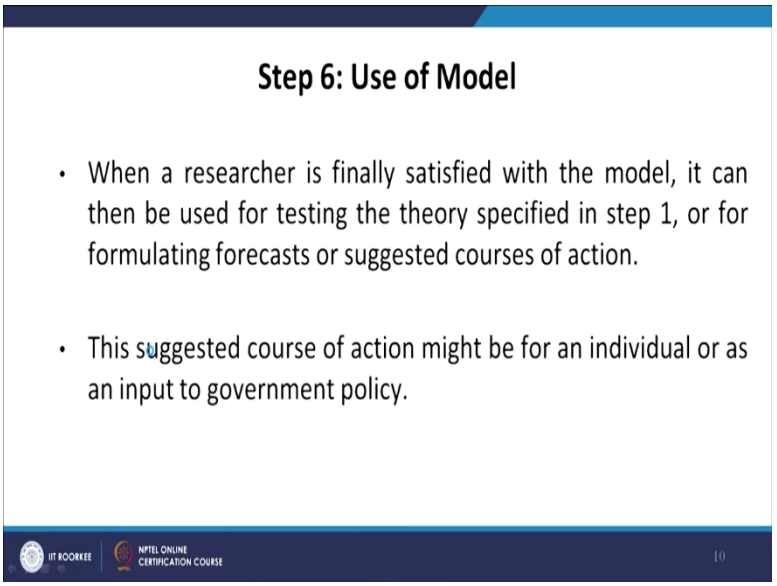
requirements. So, these are the steps that have been already explained very briefly in the first slide, that if the model is found inadequate or model evaluation is not satisfactory, then what we need to do is that we basically need to go back to the first step. But if we find it sufficient and adequate and we are happy with the model performance, then we proceed to step 5.

(Refer Slide Time: 19:43)



Step 5, talks about the interpretation of the model. So, once a model's statistical adequacy is established the next step is to align it with the theoretical interpretation. We have observed that we have got desired results, but now we see to what extent these desired get aligned with the theoretical interpretation. It requires to examine whether the sizes and signs of the parameter estimates are the same, like the ones suggested by the underlying theory? If yes, then proceed to 6, if not again return to stages 1, 2, 3.

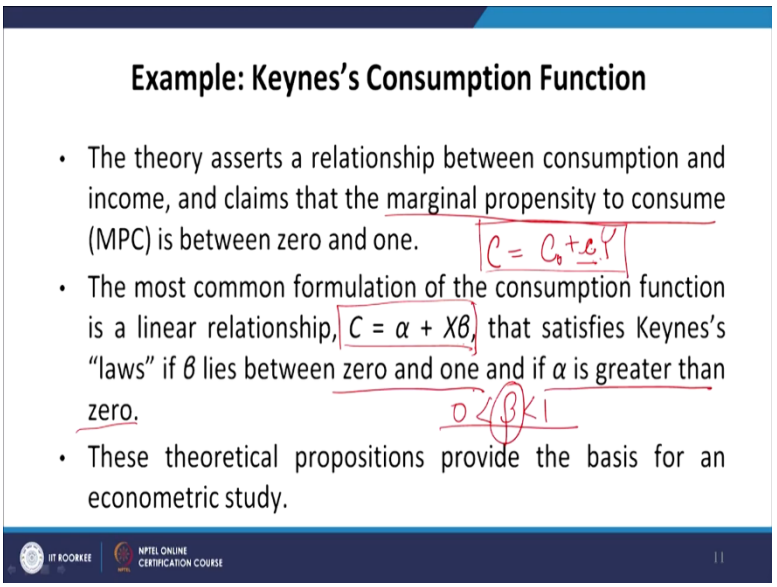
(Refer Slide Time: 20:25)



Now, step 6, talks about the use of models. When a researcher is finally satisfied with the model, it can then be used for testing the theory specified in step 1 or for formulating forecasts or suggested courses of action, or to come up with policy recommendations. So,

these are broadly part of the analysis. Based on the observation. We come up with the recommendations and implications it has for the targeted audience. The suggested course of action might be for an individual or an input to government policy.

(Refer Slide Time: 21:05)



Now, I explain this with an example. We take a very common and popular example that is Keynesian Consumption Function $C = C_0 + cY$. Keynesian Consumption Function basically tells us that, consumption C is a function of, first of all, an autonomous consumption denoted by C_0, and then the rest of the consumption basically depends on the income of the country.

It is a fraction of the income that is generated during a particular period that is expected to be dependent on or spent on consumption. So, this is something, which is actually of our interest. So, the most common formulation of the consumption function is a linear relationship. So, what exactly I have written here is $(C = \alpha + X\beta)$ that satisfies can Keynesian law, if beta (β) lies between 0 and 1 and if alpha (α) is greater than 0.

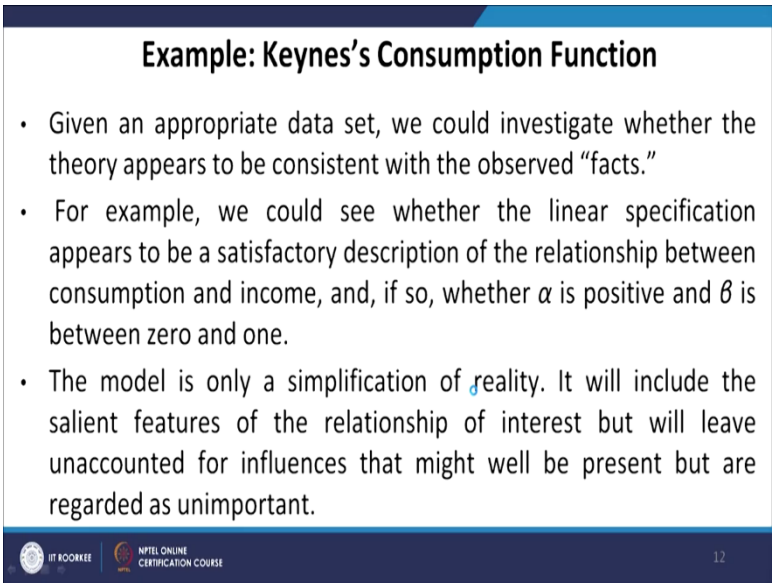
So, I simply give you the mathematical interpretation that even if income is 0 then what would be the basic level of consumption subsistence level of consumption which is given by C_0 and what is C? C is the small c, basically the marginal propensity to consume, which is written here. What does it measure? It measures that if income increases by one rupee then what percentage of it is going to be spent on consumption or by what amount consumption is going to go up or if income increases by 1 percent then by what percentage consumption goes up at the aggregate level. We are not talking about individuals.

So, this is a linear relationship that is actually that specifies the Keynesian consumption function, which says that β here or small c here, which is the marginal propensity to consume that lies between 0 and 1 and alpha is greater than 0, which is obvious that everybody needs to consume something in order to survive. So, alpha is greater than 0 and beta is between 0 and 1.

Now, again, there is empirical evidence where beta is actually not between 0 and 1, but again, they are exceptional cases. Most often we would expect that beta lies between 0 and 1 which implies that only a fraction of the increase in income is spent on increasing the consumption or spent on consumption.

These theoretical prepositions provide the basis for an economic study, sorry, econometric study. So, what we can do is that we can try to estimate a functional form like this and see that whether beta fulfills this or not. We can do it in the context of a large number of countries.

(Refer Slide Time: 24:43)



Now, given an appropriate data set, we could investigate whether the theory appears to be consistent with the observed facts. For example, we could see whether the linear specification appears to be a satisfactory description of the relationship between consumption and income. If so, whether alpha is positive and beta is between 0 and 1.

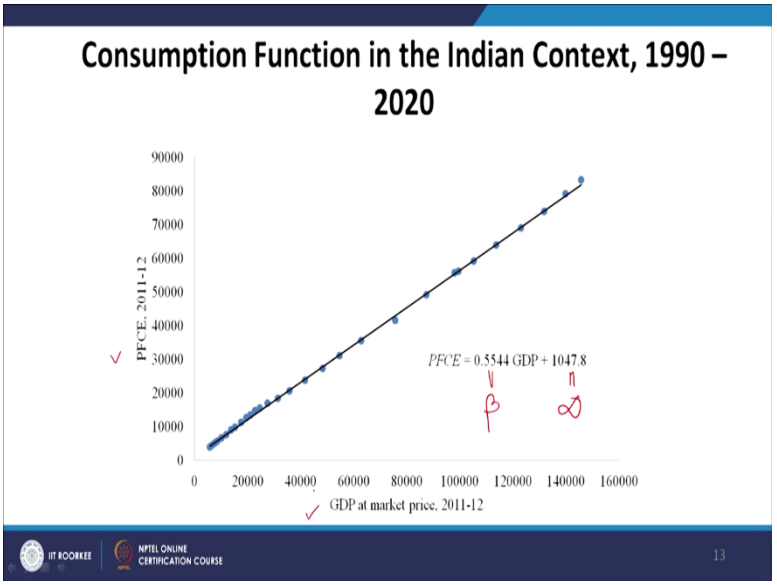
The model is only a simplification of reality. It will include the salient features of the relationship of interest but will leave unaccounted for influences that might be present but are

regarded as unimportant. So, here in this consumption function, we are focusing on the impact of income on consumption.

There might be other factors, which impact consumption level that, for example, interest rates are a factor or could be a factor that impacts the consumption level because if interest rates are lower people may borrow for in order to consume more. So, that kind of possibilities we are not considering, we are simply focusing on the relationship between income and consumption.

So, that is what is being stated here, that this model is only a simplification of the reality, it does not anywhere emphasize that income is the only factor that impacts the level of consumption or other factors are unimportant.

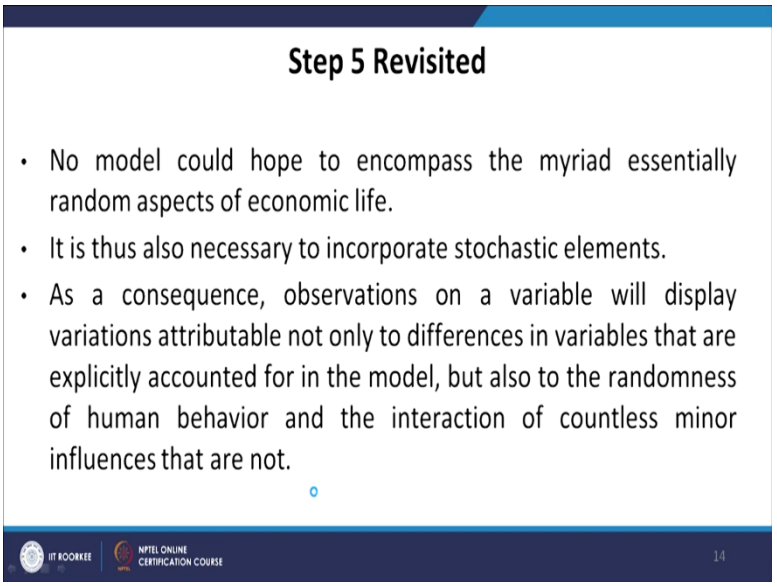
(Refer Slide Time: 26:26)



Now, if we consider the consumption function in the context in Indian context that is we collect or consider data from 1990 to 2020, and I plot them. So, this is what we observe. Here PFCE is Private Final Consumption Expenditure, GDP of course, refers to Gross Domestic Product and when we fit a line through these scatterplots then we observe that we have an alpha value which is equal to 1,047.8 and we have a beta value which is basically 0.5544.

So, roughly this is 0.55. This implies that if income increases by 1 percent, then consumption increases by 0.55 percent. If income increases by 1 rupee, then 55 paise would be spent on private final consumption expenditure given this data in the Indian context for the period 1990 to 2020. And also note that we are measuring PFCE on the vertical axis and GDP at market price. 2011-12 base period in both cases on the horizontal axis.

(Refer Slide Time: 27:43)

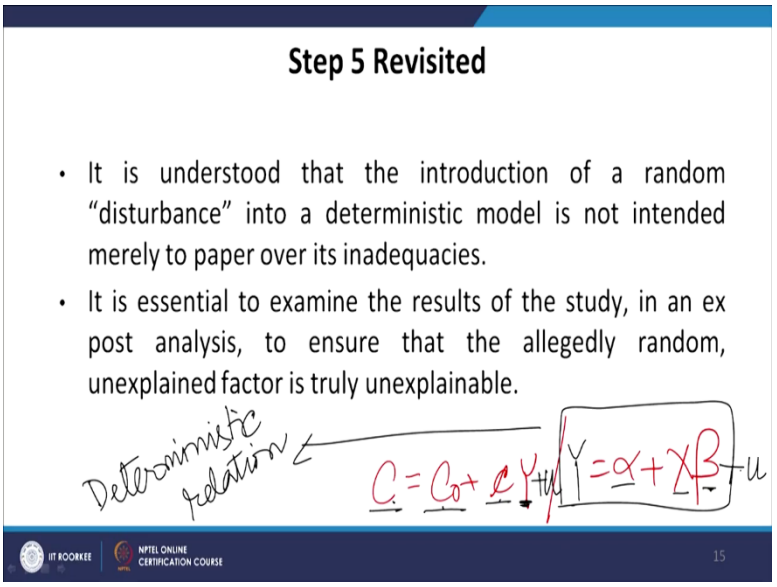


But we need to revisit step 5 because no model could hope to encompass the myriad essential random aspects of economic life. So, as I was trying to tell you that this is a simplified model, this is a simplification of reality. It is thus also necessary to incorporate stochastic elements. What are stochastic elements? Stochastic elements are random elements.

So, in the previous graph, you can see that or in the Keynesian function also, we had only $C = \alpha + X\beta$. If I know the value of alpha (α) beta (β) and X, I can get a perfect prediction of consumption with 100 percent accuracy, but that does not happen in reality, because as we are trying to emphasize that this is a mere approximation to the reality, simplification of the reality. In reality, consumption would be impacted by many other factors.

As a consequence, observations on a variable will display various variations attributable not only to differences in variables that are explicitly accounted for. For example, income here in this case, but also to the randomness of human behavior and the interaction of countless minor influences that are not included in this specification. For instance, we have included only income, and other factors that might be impacting consumption are not included in the model.

(Refer Slide Time: 29:24)



It is understood that the introduction of a random disturbance into a deterministic model is not intended merely to paper over its inadequacies, that is understanding that my model is actually inadequate to capture the real-world phenomenon. I am not incorporating the disturbance term or a random term.

It is essential to examine the results of this study in an exposed analysis to ensure that the allegedly random unexplained factors are truly unexplainable. So, first of all, I have probably included possibly as many factors into the model as possible, but still, there will be some which are left out. So, in order to understand that kind of a possibility, I will include that randomness, so I accept that it is not possible to include all possible factors that might be impacting a variable.

Now, here, in this case, the other alternative is that I focus on a particular variable, and because of that, I do not want to include other variables. So, since here my concern is to understand the relationship between only income and consumption, so I am including income only. I do not want to take into consideration other variables. So, these are possible reasons for which we need to include random variables.

The model is inadequate without the inclusion of randomness. So, we include a random factor, which would actually take into consideration all factors which are not included in the model, either for theoretical purposes or for the understanding that it is not possible to come up with a real-world estimation or analysis in its true sense with 100 percent accuracy.

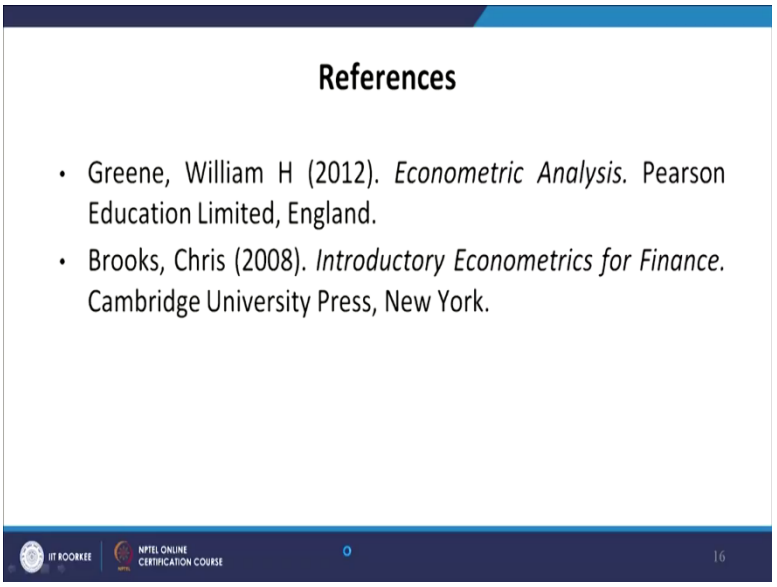
So, how we basically include this randomness into the model that will actually take us to the econometric modelling. Now, here, I would just like to note that the model that we have estimated just now, which is ($C = \alpha + X\beta$) was y equals to $\alpha + X\beta$. (Refer to slide on Step 5 revisited above). So, what happened here is that this entire thing is called a deterministic relation.

This implies that there is no randomness. And I have already mentioned, that this shows as if we know the value of alpha, know the value of X, and also know the value of beta, then I can have a value of y with 100 percent accuracy. Alternatively, this implies that, if I know, the autonomous consumption level, if I know the marginal propensity to consume and if I know income then I know consumption level.

Not only for this period, since I have considered data from 1990 to 2020, I can say that I can predict consumption for 2021-2022 and so on with 100 percent accuracy, provided that in these years also C_0 and C is not going to change much. But this rules out the possibility that consumption could be impacted by other factors as well, as income is the only factor explaining consumption which actually is not true.

Since other factors can also impact consumption, so I need to actually incorporate a random term, and this random term is denoted by u. We call it an error or unobserved term, but all these details will be discussed in the upcoming modules. For the time being, I finish off the formulation of econometric modeling are the steps involved in that.

(Refer Slide Time: 33:53)



These are the references, you can follow. Thank you.