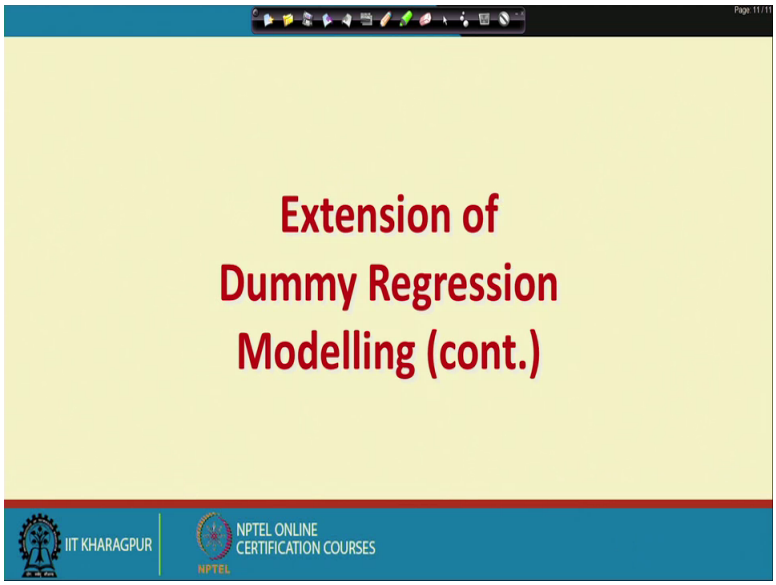
Engineering Econometrics
Prof. Rudra P. Pradhan
Vinod Gupta School of Management
Indian Institute of Technology, Kharagpur

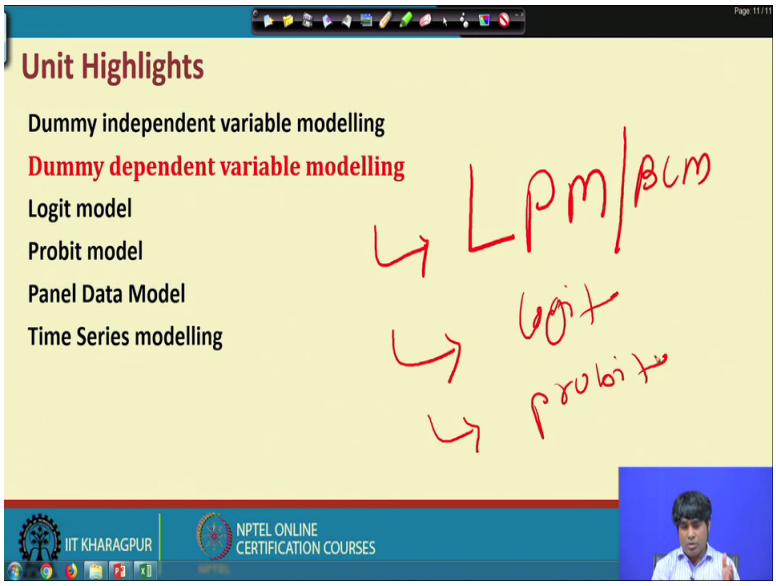
Lecture – 38
Extension of Dummy Regression Modelling- Dummy Independent Variable Modelling

Hello everybody, welcome to Engineering Econometrics will again continue with Non-Linear Regression Modelling and that to Extension of Dummy modelling and the coverage will be specifically dummy independent modelling.

(Refer Slide Time: 00:37)



(Refer Slide Time: 00:40)



So, we have discussed you know, various types of you know models and the applications relating to dummy modeling, where independent variable will be treated as a dummy. And specifically we have discussed couple of you know problems. We have analyzed the problems through softwares and then connected with a particular, you know case. Now we like to see the other side of the game, how is the modelling behavior altogether when the dependent variable will be qualitative type, so; that means, technically we like to address dummy independent modelling.

So, where we specifically have three types of you know models. So, the first one is the Linear Probability Model then otherwise it is called as a Binary Choice Model. The second one is Logit model, the third one is the Probit models, right. So, we have three types of you know models linear probability models, Logit model and Probit models.

So, we like to you know discuss what are the situations we can apply these models and how is this you know models, they are different to each others? So; that means, we like to know the structural form of this models that with respect to BCM, Logit and Probit. One; however, one typical common is the dependent variable which is actually categorical or dummy type that is the, that is the basic approach otherwise you know the models are you know more or less same.

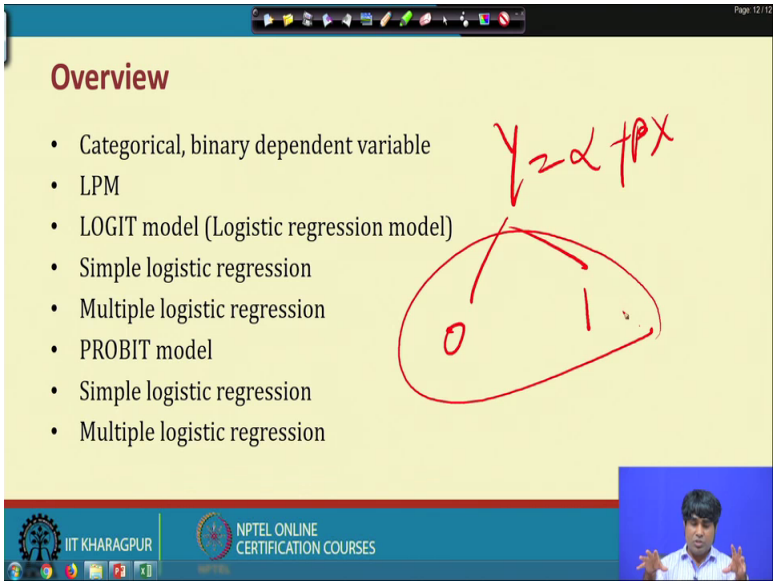
So, we means technically the difference between these 3 models are with respect to various functional form only. So, the first one that is the linear probability model, where

it is a kind of you know linear structure and the other two forms are non-linear structure. So, that is how it is called as you know linear probability model. And, the second one, it follows with the logistic function, the third one, it is for, it follows with you know normal density function.

So, we, let we, let us discuss the first one, of course, it is not in the non-linear form of models, but you know we should know the basic you know background behind this linear probability model. Though, it is actually dependent variable qualitative, the independent variables may be some kind of you know different functional form which we have already discussed in various way, having dependent variable constant, independent variables a particular independent variable or couple of independent variables may be in log format or the kind of you know square root format or expansible format.

So, likewise we have a different kind of you know clustering and that is how we can also discuss the linear probability model. Then we will move to Logit models and Probit models. So, let us see how is this particular you know structure?

(Refer Slide Time: 04:21)



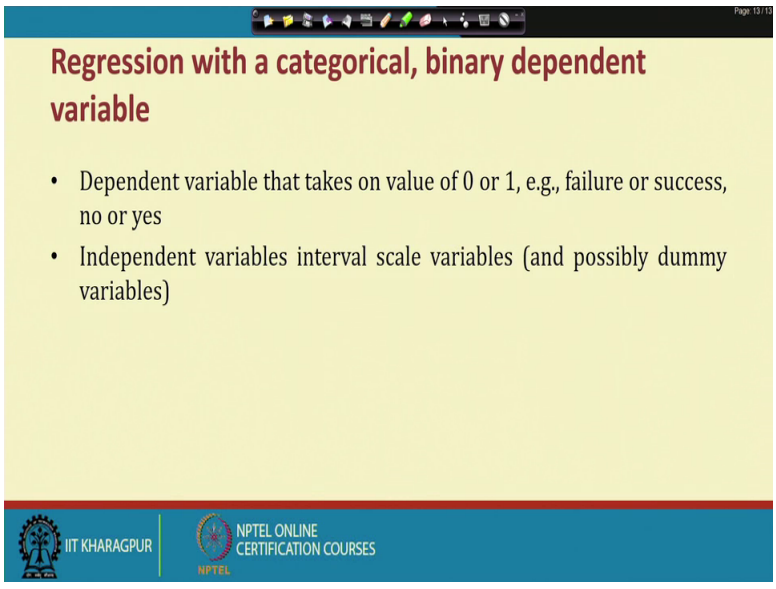
And,, so, technically linear probability is a very simple one, where the dependent variable will be categorical. And all these models are connected with probability, in fact. And in the case of linear probability model, the look will be like this. So, here Y equal to alpha plus beta X. So, the y will be represented as a 0 1.

And most of the instances while using linear probability model, so, the Y content will be either 0 or 1 so; that means, actually like you know we have discussed gender. So, now, the gender will be in the left hand side, that is the dependent variable and then we have a independent variable which may be qualitative, may not be qualitative, not necessarily qualitative which it involves you know both quantitative and you know qualitative. But dependent variable will be qualitatives again varies between 0 to 1 only.

But we know probability and probability varies between 0 to 1 so; that means, these are two extremes. If your problem involves you know 0 1 only so; that means, like you know yes, no, if yes equal to 1 and no equal to 0 or something like that, then you know by default the information behind Y will be only 0 and 1, but it will not in between 0 and 1. If in case the information will be in between 0 and 1, then we can use the Logit models and the Probit model.

So, that is how the two different you know options with, with yours while using linear probability model and non-linear probability models. So, in the case of linear probability models, so, information will be between 0 and 1 only. So, in between contents are not there, but if you know no 0, no 1 and in between content, then in that case you can use either Logit models or you can use Probit model. So, now, let us see how is this particular you know structure? So, this is the basic background behind this modelling.

(Refer Slide Time: 06:51)



And then, we will discuss various forms of you know linear probability models. So, typically like you know failure success, male female, then you know having house not having house, having you know bank account having not bank account, so, these are the cases through which actually you know, you apply linear probability models.

For instance, let us say financial inclusion. So, one way to measure the financial inclusion is a you may have a bank account, right. So, now, will, will have with this survey and then we will ask whether the guy has a bank account or not, a bank account then this one, if not then it is a 0. Now, if the dependent variable will be the financial inclusion then; obviously, the problem will be what are the factors responsible? For maybe income is a factor, employment is a factor.

So, these are all you know in the independent basket, but here dependent variable will be the financial inclusions where the representation will be the you know the presence of you know bank account. If they have bank account then this will be yes and the numeric transformation will be 1 and if they have not then the numeric transformation will be 0. So, as a result it will be only lying between 0 and 1 and that that specifically in lines with you know probability model, where we are only two extremes 0 and 1.

So, now, in between you know 0 and 1. So, if it is in between 0 and 1, then we can use Logit model and Probit model. These same problems can be you know, you know can be applied to Logit model and can be applied to Probit model provided the sampling structure will be different. So, now, you know let us say, we have actually 20 different you know sample points so; that means, we are, you know surveying a 50 individuals or 10 individuals. So, with the intension that you know or to know what are the factors responsible for the financial inclusion? So, you can start with you know 1 2 3 and 50. And every time you ask whether they have a account or not account? So, like you know if he has 1 then if not zero. So, that is how the transformation is all about.

So; that means, in the first hand you will get actually, the information which involves only 0 1. So, by default the first choice is the linear probability model, but you know if you use something you know means, if you are very much interested to use a non-linear regression modelling and that to with the dummy involvement and that to against the dummy involvement with you know or dependent variables. So, in that case the same problem can be.

So, what we can do? You have to change the sampling structure. So, now, instead of you know in let us say 10 individual or 30 individuals. So, now, you have to, you have to select 30 different location altogether. Then within the 30 different location, every locations you may have a 30 different you know sample again. Then out of 30 different samples you check how many having account and how many having not account. So, having account if it is actually say 30 people having all these 30 people having accounts then by default the ratio will be 1.

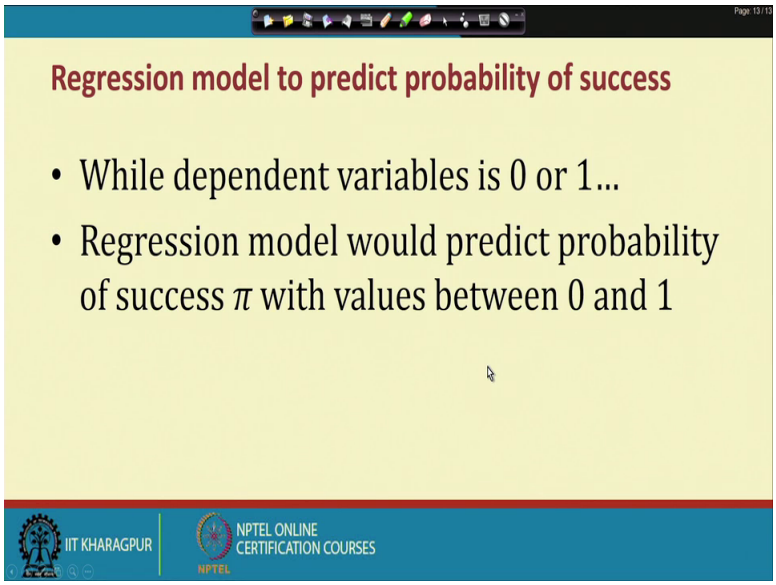
So, in that case this can be discarded simply because we need to use the non-linear probability model and that to non-linear regression modelling with you know qualitative dependent mode, dependent modeling or qualitative response regression modelling.

So, here what will you do? So, in the second samples. So, you can actually, you know survey for 30 individuals again and then check how many having you know bank account. So, let us say 20 out of 30, 20 is having bank account so; that means,. So, the financial inclusion quantification will be 20 by 30. So, that will not 0, that will not 1. So, then it will be in a, in, in between 0 and 1. So, that will be in a ratio format.

So, if all the 30 samples are having no bank accounts, so, then by default it will come down to 0 and if all are having bank account, it will come down to 1. And if few are having back account by default, it will be lying in between 0 to 1. So, the, the locations where the information will be in between 0 to 1 so, that will be the counted samples for the Logit model and the Probit model.

So, that is how it means, same problems we can analyze, but the samplings, I mean sample restructuring is requires. So, means the way actually highlighted the issue same problems, but your survey structure will be little bit you know restructure or you know redesign. Then once you have the samples, then as usual the, you know processing will be similar, ok.

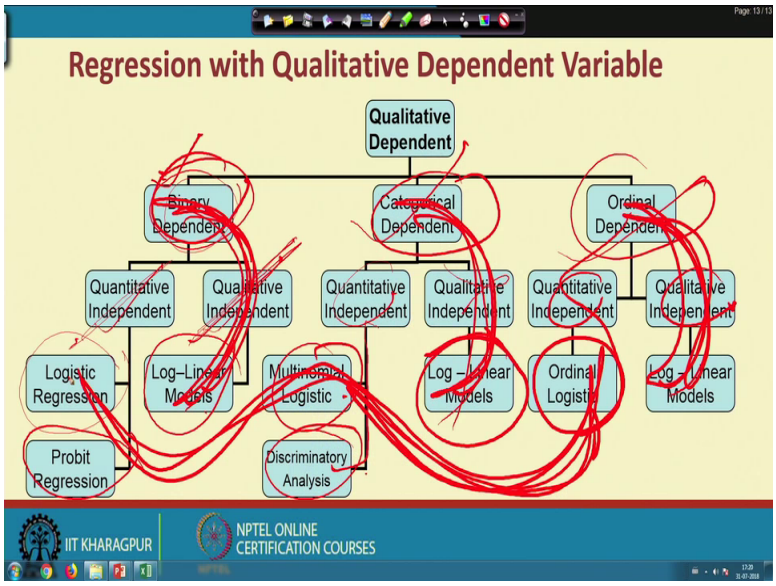
(Refer Slide Time: 12:07)



So, this is the basic idea behind the linear probability model then Logit model and Probit models. Of course, we, we, we are interested to know how is this logist, Logit models and how is the Probit models? Of course, the functional form will be different. In fact, the right hand side of the structure will be more or less same in that case of you know linear probability models, Logit model and Probit model, but what is the difference is the left hand side of the game.

The first one is the simple linear ones, so, where the data will be in between 0 and 1, the second one followed the logistic function. So, here this sample will be again 0 to in between 0 and 1, not 0 not 1. And the third one is the you know, it follows with you know normal density functions, again the sample or the data will be in between 0 and 1, no 0 no 1. So, that is the a typical difference, you know among these 3 models.

(Refer Slide Time: 13:17)



So, where you know the first one is the linear one and the second and third one are the non-linear one. Here is the you know typical classifications so; that means, this is the simple way of you know understanding in between the complexity will be coming into the pictures. If you use you know multinomial logistic functions or something kind of you know dummy qualitative response regression modelling with you know more number or dummies, in the right hand side that is the independent clusters.

So, likewise you will be find, you know plenty of different kind of you know structures through which you can actually analyze the situation. In fact, the you know the structure which you have discussed earlier. So, in the first and there are 3 classifications. So, binary dependent, that is what we are actually discussing right now then categorically dependent and then ordinary dependent.

So, that is, that is what actually the level of you know grade a grading actually. So, binary dependent means against it is a you know 0 1 only, that is the binary numbers, categorical means it is here the, the kind of you know classification will be more than more than 2. So, in the binary case, it is the two way classification yes, no, 0 1, success failure like that male female.

So, so these are the kind of you know examples, your binary dependent will be work. And categorical dependent means like you know different category like Hindu, Muslim,, you know other groups Christians like that you know or you know say 3 types of you

know educational levels like you know PHD, M. Tech, B. Tech or you know 3 different forms of you know age groups of the peoples like you know, let us say people having you know age of you know more than 40, then more than 30 more than 20. So, these are all various ways you can actually a you know, you can, you know categorize.

So, this can be one form of you know qualitative response modelling then or ordinal dependent. So, here actually the same levels we have a different kind of you know rating. For instance good, better, best like that you know very goods, goods not so good. So, like that you know different level of (Refer Time: 15:33) typically a different scaling, ok.

We use like it, scale to do these kind of you know categorization and sometimes some of the problems, engineering problems will be such a way that the dependent variables having actually categorical in character or ordinary in character or binary in character. If it is binary in character then we will come to this group, if it is categorical you can come to this group and if it is ordinal dependent we can come to these groups. So, then against each basket we have a different forms of you know models.

So, in the binary dependent it may be quantitative independent or qualitative independent, but the dependent variable will be binary, it is qualitative it is only in between 0 1. However, the game will be connected with the either you know quantitative independent or qualitative independent or it can be both quantitative independent and qualitative independent. So, it can go together also.

So, again if it is a quantitative independent so, it may follows actually logistic regression again and it can follows with also probability regression. Similarly, in the case of qualitative independent, so, the form of the model will be log linear models which we have already discussed. So; that means, this actually brings different kind of you know non-linear forms of you know regression modelling and the entire you know structure of non non-linear regression modelling can be summarized here.

So, now, in the case of you know categorically dependence again we have 2 different first hand classifications. So, quantitative independent and qualitative independent that is also equally true in the case of you know ordinarily dependent. So, quantitative independent and qualitative independent.

So that means, all these 3 categories, we have first hand classification quantitative independent qualitative independent and then it may have that you know the mixture of you know 2. So; that means, binary dependent with the quantitative independent or binary dependent with the qualitative independent or binary dependent with the both quantitative independent and qualitative independent. So, this is the same structuring can be applied to categorical dependent and again it will be applied to ordinary dependent.

And in between you will find again various forms, so, against the way we have discussed in the binary dependent in the case of categorical dependent. So, we have a quantitative independent suppose there is no qualitative independent. So, in the case of quantitative independent, it can have 2 different forms. So, one form is called as a multinomial logistics and the other form is called as a discriminator discriminatory analysis, ok.

So, discriminatory analysis means again so, it is kind of you know, yes no type of you know situation then in the qualitative sides it is like you know log linear models. So; that means, technically if the dependent variable will be qualitative and the independent variable will be qualitative then the, the structure of the model will be more or less same that is the log linear models, every case this is also true here again this is also true here, this is also true here, so, ok. So, this is one form, this is another form, this is one another form. So, that is the kind of you know uniqueness, you know among these 3 models

So, binary dependent categorical dependent ordinal dependent however, in the case of you know quantitative independent, in the case of you know categorical dependent you may have a multinomial logistic or discriminatory analysis and in the case of binary dependent we have a logistic regressions and Probit regression. And again in the case of actually ordinal dependent with the quantitative independence we have ordinal logistic function.

So; that means, the logistic function, one logistic functions will be is common against for the kind of you know this is how the commonality, we find in these 3 categories, right. So, this is how the kind of you know classification we have in the case of you know a qualitative response regression modeling, where the common structure is a dependent variable will be qualitative and it differs with you know 2 different ways a first is the various functional forms that to the simple linear format and 2 different non-linear formats that to Logit and you know Probit. And a against the dependent variables may

have a 3 different forms that is the binary dependent, categorical dependent, ordinal dependent depending upon the kind of you know classification, different levels of classification, it means the starting with you know at least 2 then more than 2 and against little bit you know more flexible more than 2, right.

So, it is kind of you know the games then in between the involvement of you know independent variables with the quantitative structure and qualitative structure and both quantitative structure and qualitative structure.

So, we have actually a pool of you know non-linear regression modelling and against the way I have cited already. So, it is actually very specific a special kind of you know models and it cannot be applied each and every engineering problem. Depending upon the particular problems where the particular specification will meet, then you can just step like after you know peaks you know, you know doing the structuring of all these you know in a model requirements.

Then the process of estimation will be more or less seems like you know you know original you know or you know simple regression modelling. So, what, what is more important is that you know, the kind of you know structuring. So, like we have actually instead of you know X we can use log X, instead of X you can go for you know square root of X, instead of X you can go for you know, you can go for you know exponential of X.

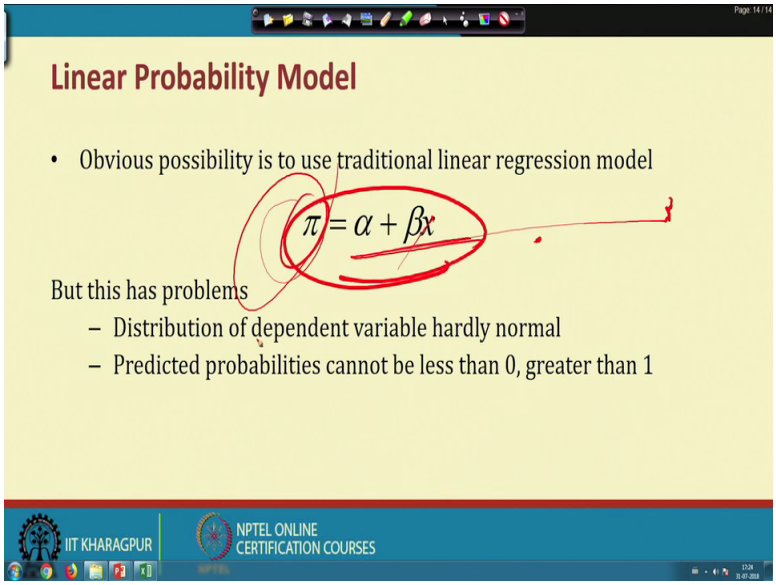
So, likewise various forms of you know and again instead of X, we can follow the logistic pattern. So; that means, technically we have actually a different kind of you know structuring to analyze this problem. And what is the requirement? The requirement is you know as per the a, you know problem structure and the kind of you know data structure because ultimately the function for functional form of the types of model which like to use or to apply, to estimate and then analyze the engineering problem.

Exclusively depends on the problem structure and the data structure until unless you know the problem structure and data structure you cannot just blindly use this model for the estimation where you know the model outcome will be different and it will not pass through blow theorem and as a result diagnostic will be lost and the you know the kind of you know reliability will be lost.

So, as a result, so, what is the best is that? So, you have to check the structure and check everything, then you know means everything means the data structure and the kind of you know types of models, the requirement. Then we finally, use and again go for you know data transformation, if it is necessary to process this model because ultimately if you choose let us say Probit regression model or logistic regression model.

So, it is an actually non-linear format. So, you do the proper transformation that with respect to variables. And as a result the ultimate the, the transformation will go to the data and then you will trans and then you will have a new form of the data which can be simply estimate as usual the you know ordinary regression analysis. So, that is the kind of you know structure which you have to follow.

(Refer Slide Time: 24:15)



So, to estimate the process and to analyze the problem and as per the particular engineering requirement and on the basis of you know the requirement of you know management decision. So, the kind of you know linear probability model is like this. So, simple model is like this. So, over here the structure will be like this is what the y dependent variables and this is the independent variable structures, but necessarily the independent variable will be 1, it can be many and all are may be quantitative in natures and few can be qualitative.

And there is a high chance that you know all can be also categorical and there is a high chance again both categorical and you know numeric. So, that is not the big deal what is

the big deal heres, the y classification that is the you know; that means, very - very special kind of you know problem this one.

So, here you know sample structure must be 0 and 1 that is the only way you have to use the linear probability model. For examples, you know lots of people you know staying in outside from the home and sometimes you know they have the own house and they may not have their own house, they used to stay in a rented house and obviously, you are interested to know what are the factor responsible for you know having own house and what are the factor responsible against people staying in a rented house.

So; that means, people having house was staying in own house or staying in a rented house that is the two way classification. If they are staying in a let us say own house then you can say that you know 1, if not then you will be put a 0, whether it is a rented house or something else.

So, the simple interpretation is if whether people having own house or not that is all that is the question which you approach to the respondents, And then you like, like to find out or you, you, you may have sufficient knowledge to know what are the factors through which you can actually explain whether the you know people having you know house or not having house or people staying in own house or not staying in one house.

So; that means, first questions you have to post that you know whether to, whether you are staying in own house or not. Maybe people you know I mean say the there are, so, many factors maybe one of such factor is income.

So, if people have a more you know sufficient income or you know heavy income and definitely they have their own house because usually as a human being they are very much interested to stay in one house rather they are staying in a rented house because rented house means they will pay the money, right. So, what they you know what they used to do they, they try to have their own house and for that you must have money.

And there may be many you know a ways you can have your own house, but one of the way is to you know, you must have you know sufficient income. If your income is high then you may be in a position to purchase the house and as a result you can stay in your you know one house, that is the kind of you know example which we can have. So, now,

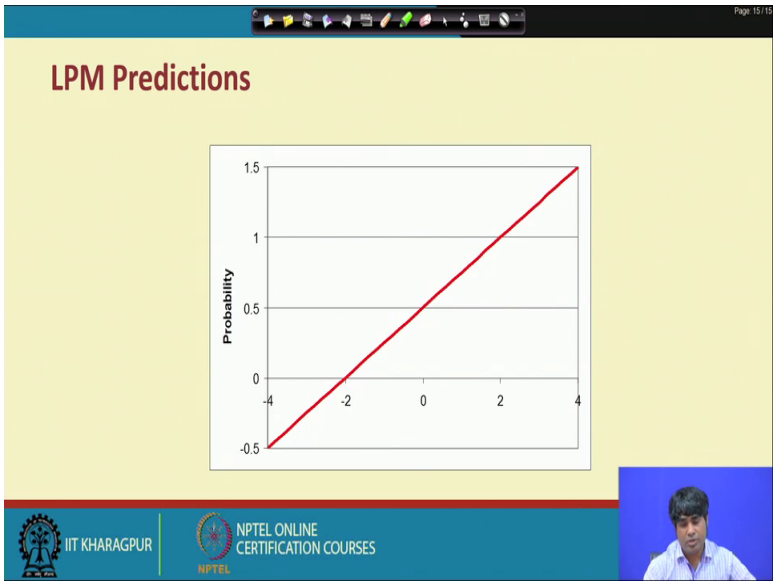
likewise there are couple of factors which are equally responsible for you know whether you have one house or you know staying in a rented house.

So, now until unless you know the particular structure you cannot you know give a kind of you know a statement. So, first you have this survey then you go ahead with the estimation process. Now, analyzing this kind of you know problem by default whatever you know, you know enquiry will do and whatever you know respondence will be capture. So obviously, then have the situation like you know some situation they have their own house and some situation they have not there you know own house ok.

Now, if that is the case then you can you know go ahead with the estimation process otherwise if, if you are having survey and where all peoples are having own house then in this particular sample cannot be used for this particular model. So, that is kind of you know, biased kind of you know process. And against a you are respond you are you know capturing a particular samples and all are having actually you know is all are staying in a rented house; that means, they have you know, they have not their own house. So, this, this kind of you know sample also cannot be used for this kind of you know modelling.

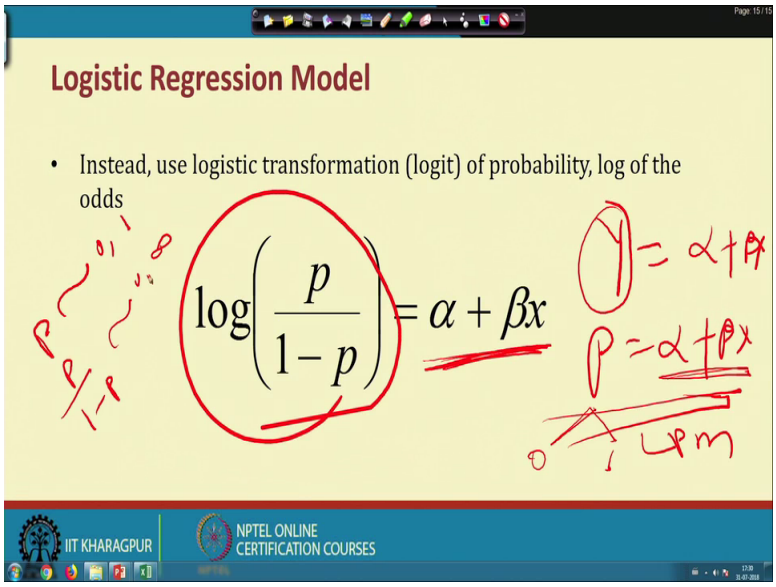
So, this model can be used only when few samples will be having, they you know in such a way that they have their own house and by default other samples they have not there in own house. Then it can go ahead the estimation process and check whatever factors you have identify whether they are you know significantly influencing the kind of you know case or not right, that is the kind of you know structuring you will be had and in that case of you know linear probability models, right.

(Refer Slide Time: 29:13)



So, this is how the a you know graphical look of the linear probability models. So, the predicted line will be like this and by default you will find 2 different data the groups one group with you know you know say, you know yes category another group say no category, this has mean 0 y you know I, y pool with you know 0 coverage and y pool with 1 coverage, that is the kind of you know structuring.

(Refer Slide Time: 29:41)



And the other form of the you know dummy independent modelling is the non-linear structures and that to the first non-linear structure is the logistic regression you know

model where the regression that too dependent variable structuring will be like this, ok. So, here the regression structuring will be like this, ok. Corresponding to linear probability models where we have Y equal to alpha plus beta x or simply we can say P equal to alpha plus beta x because Y represents you know linear for probability model means it will be only probability values and that to only in extreme conditions that to only in between 0 and 1.

So, that is the linear probability structure, but here, the format is log P by 1 minus P so, but the right hand side is a constant alpha plus beta x, here also alpha plus beta x that is the linear probability model, I am just comparing both the models and check the difference how is that kind of you know scenario. So, here what is happening, this side is same this side same, but the left hand side game is a different. So, it is a simple P where the range will be only 0 and 1, but here again P, but if we will be positioned differently.

So, we instead of directly we are using a ratio log of the ratio where p divided by 1 minus P usually in probability, P represents probability of success and by default q is the probability failure which is the difference between 1 minus P because total probability is exactly equal to 1.

So, if P represents the success and by default 1 minus P will be the failures there is a high chance the success will be the 100 percent, in that case P equal to 1, if P equal to 1 then by default 1 minus P equal to 0. So, in that case, so, 1 by 0 which is actually brings you infinite. But in other case if the probability of success equal to 0; that means, P equal to 0.

So, the ratio between P by 1 minus P, 0 by 1 minus 0 by default it will be equal to 0. So; that means, ultimately depending upon the P, so, you will find the ratio will be bearing. So, one extreme will be 0, another extreme will be you know infinite. So; that means, when P equal to 0 so; that means, in the linear probability models when we put you know P equal to 0 and against when you are putting P equal to 1. So; that means, putting P equal to 0 and 1, so you will find actually the ratio will be say 0 to 1.

So, like that you know we have to see the transformation. Now, what is happening? So, now, corresponding to P value 0 and 1. So, your ratio here P by 1 minus P will also vary from 0 to infinity; that means, well P varies from 0 to 1, then P by 1 minus P varying from you know 0 to infinite. But if any you know items, or you know any data is having

actually infinite quantification and by default you cannot run these models. Software will not actually help you to calculate, if you manually cannot calculate because you do not know the exact quantification, if you do not know the exact, exact quantification prediction cannot be possible.

So, as a result, so, the whole transformation will have actually by sampling. As a result, you have to restructuring the particularly you know means the entire process. So, now, what is actually you know happening here or what is the clear message here that, if the sample points are you know 0 1, then there is a problem here. In fact, if it is 0 there is no issue, but if it is 1 then there is a issue, but ultimately the total sampling will be exactly equal to 1.

So, that is why, so, if you like to use this ratio. So, or, so, this the suggestive idea is that you know your P should not be 0, P should not be 1. So, it should be in between right for that, you need actually group sampling rather than individual sampling. So, we need actually P represent probability of success. So, then check what is the percentage of success? That percentage of success will be the P and by default that will be in between 0 and 1, not the 0 not the 1. If it is 0 and 1 then; that means, you can reject that sample for this particular you know logistic model, ok.

So, so this is one way of you know understanding the structure about the logistic models and the kind of you know linear probability model like why the Probit model is a also similar kind of you know structure. We are the right hand side of the game will be very much similar that is the alpha equal to alpha plus beta, but here it follows the normal you know normal density functions. So, as a result, so, you just you know check how you can bring the transformation all together, ok? So, how you can bring the transformation altogether? So, ultimately, so, ultimately the P value will be again in between 0 and 1 and not exactly 1 not exactly 1.

So, otherwise this can actually affect the entire process again. So, so what is actually required here again? So, again if you will go for you know group sampling; that means, technically for low using logistic models or Probit models, you should have a group sampling rather than individual sampling and in the case of linear probability model you simply use the individual sampling where the dependent variable will be in between 0 and 1, but it will be the probability value 1. It because all these 3 models are you know

probabilistic models and that too in the form of you know linear probability models and non-linear types of you know models like you know Logit model and Probit models

So, now, the beauty is that you know what are the situation you have to use? And again technically it is a special kind of you know models, it will be used in a special engineering kind of you know situations where we need to predict the requirement as per the, as per the you know decision making process is concerned. So, in the next class we will discuss this models against within a, suitable examples. In the meantime, we will stop here.

Thank you very much. Have a nice day.