Handling Large Scale Unit Level Data Using STATA

Professor Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee

Lecture 32

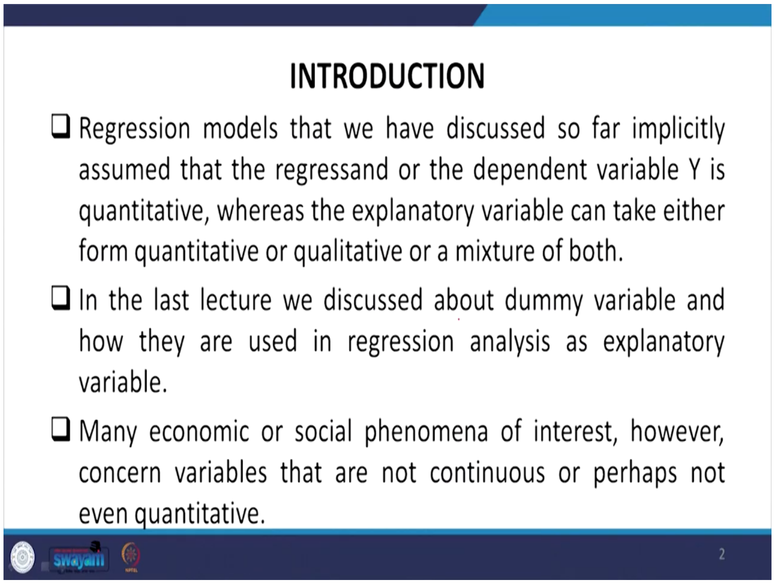
Binary Response Models - I

Welcome friends once again to the NPTEL MOOC module on a very specific module called Handling Large Scale Unit Level Data using STATA. And this is one of the very rarest module available in India. So, I will suggest that irrespective of your discipline whether you are from social science, management or economics it does not make much difference so far as learning is concerned. I just wanted to mention that especially this module is very useful for better and systematic presentation in your paper using databases.

So, let us start the lecture on binary responses of the dependent variable. We are continuing with the analysis of qualitative data we already started in the last class. Last class was on the understanding of dummy variable and dummy variable trap. And we clarified with the help of different diagrams that what do you mean by trap and why trap occurs, and then how to avoid the trap we already suggested.

In this lecture, these successive four lectures are going to be discussing on different types of binary response models. So, binary response model we are only sticking at this moment for better clarity but not necessarily the dependent variable to be binary always. For simplification only we are sticking to binary. It might be multinomial type as well. So, broadly there are four types we are going to discuss based on the nature of the change in the binary structure and the nature or the approach of dealing with the model is very important for the clarity of understanding of the binary response aspects.

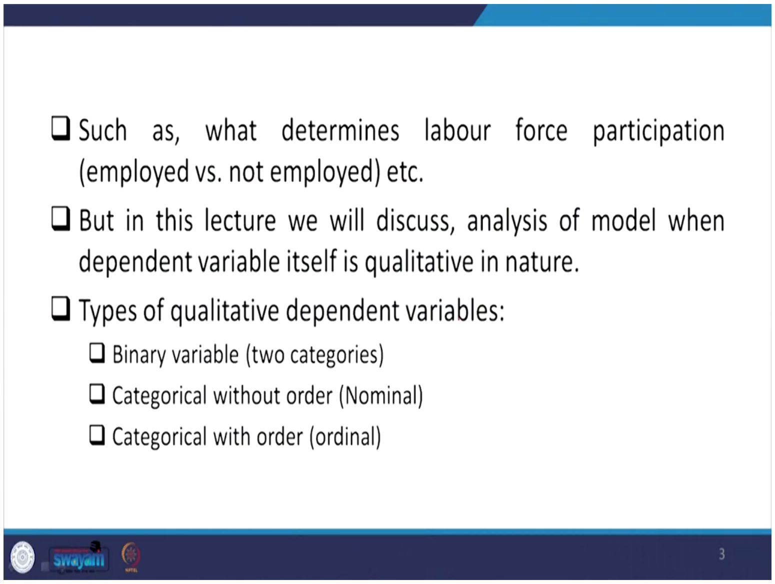
(Refer Slide Time: 02:57)

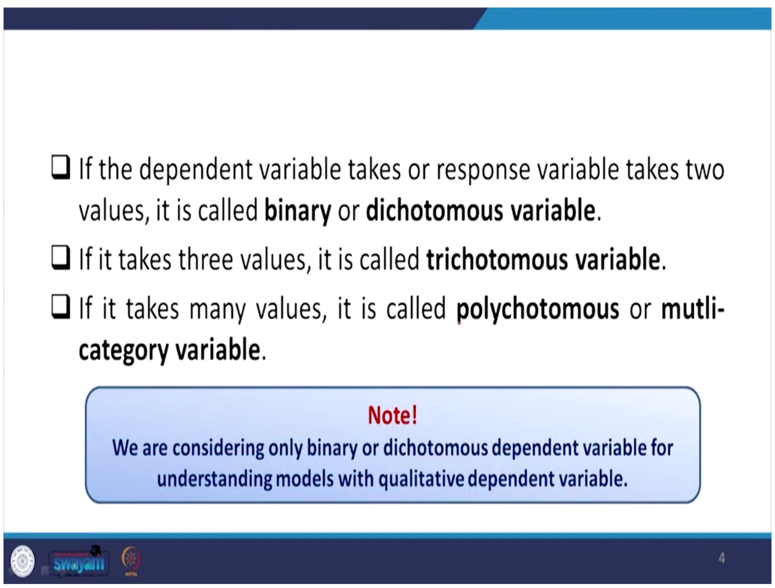


Let us introduce you once again that the regression models we have discussed so far simply discusses the relationship between dependent variable and independent variable. And then it implicitly assumes that the regression or the dependent variable that is why is quantitative, whereas the explanatory variables can either take quantitative or qualitative or mix of both the kind of variables. So, that was the assumption we already mentioned many times. So, I hope you are very clear with differences.

In the last particular lecture we address dummy variables and the issues in estimating the beta coefficient attached with the dummy. And many economic or social phenomenon of interest have a concern, variables that are not continuous or perhaps not even quantitative. We are highly concerned now to address this issue because in many situations the variables are not continuous or difficult to be quantified. So, that is the reason why we are trying to stick to. Alright so, what is the example behind it? We wanted to suppose understand labor force participation. Here I am referring to employed versus not employed.

(Refer Slide Time: 04:38)

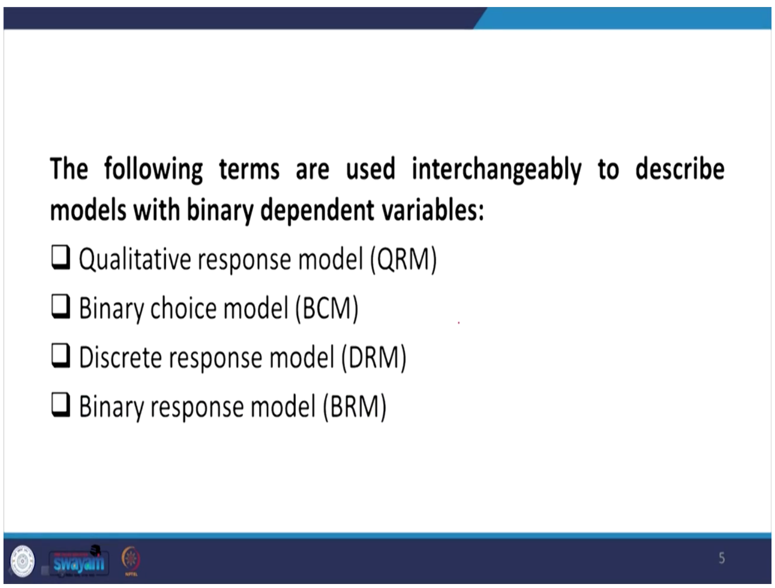




But in this lecture we will discuss analysis of model when dependent variable itself is qualitative. Like if independent is qualitative or mixed is not so problematic and we already addressed. But now onwards when we introduce the dependent variable to be qualitative we are committing many errors if we simply go by the standard ordinary regression techniques. And there are four types broadly, models we are going to discuss. But types of qualitative dependent variable to be explained. What are the different three types majorly we are going to discuss.

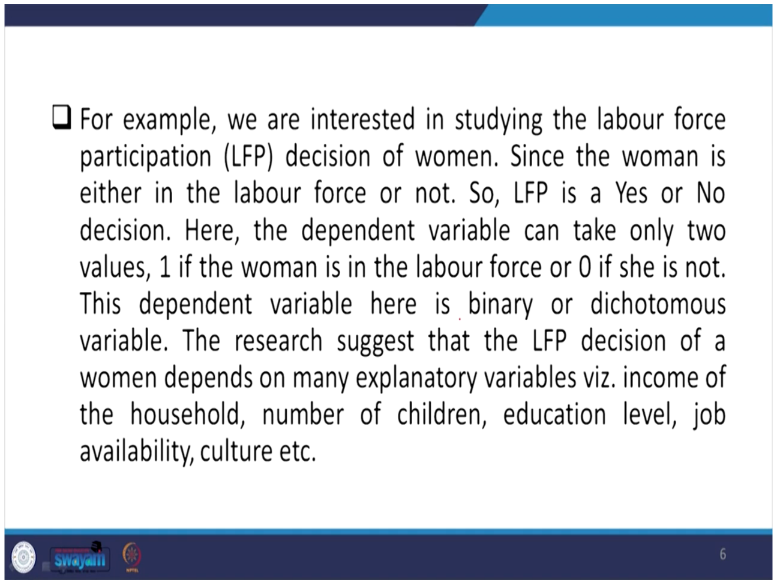
One is called binary variable having 2 categories; categorical without order if it is nominal type, then categorical with order that is called ordinal type if any are there. Then accordingly we have different regression specifications. If the dependent variable takes or the response variable takes two values then it is called binary or dichotomous variable. If it takes 3 variables then it is called trichotomous. Or if it is more than that then it is called polychotomous or multi category variable. As I already mentioned that we are dealing with the binary choice model here or the dichotomous dependent variable for understanding the models with qualitative dependent variable.

(Refer Slide Time: 06:06)



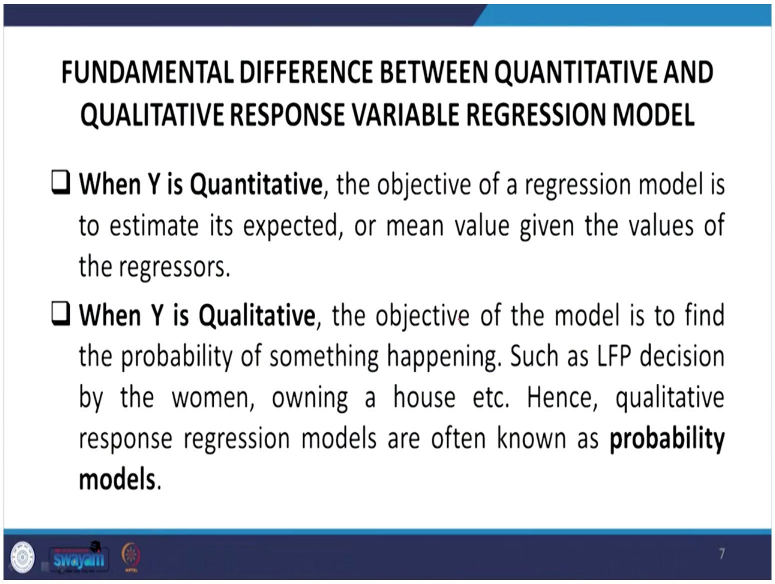
The following terms are used interchangeably to discuss or describe the models with binary dependent variables. These are used interchangeably. So, like first we are mentioning qualitative response models from where you might be confused what is this qualitative response model as compared to binary choice model. So, like in our title we mentioned that binary response models, so these are all similar; discrete response model, binary choice model, so all four terms are interchangeably used.

(Refer Slide Time: 06:40)



One such example, let me cite for the understanding is on labor force participation decision of women. Since the women is either in the labor force or not so LFP is a yes or no decision here the dependent variable can take only two values that is 1 and 0. 1 stands for yes in the labor force or 0 if she is not there in the labor force. And so it is binary and dichotomous. So, the research suggests that the LFP decision of women depends on main explanatory variables. Explanatory variables could be income of the household, number of children, educational level, job availability, culture etc.

(Refer Slide Time: 07:42)

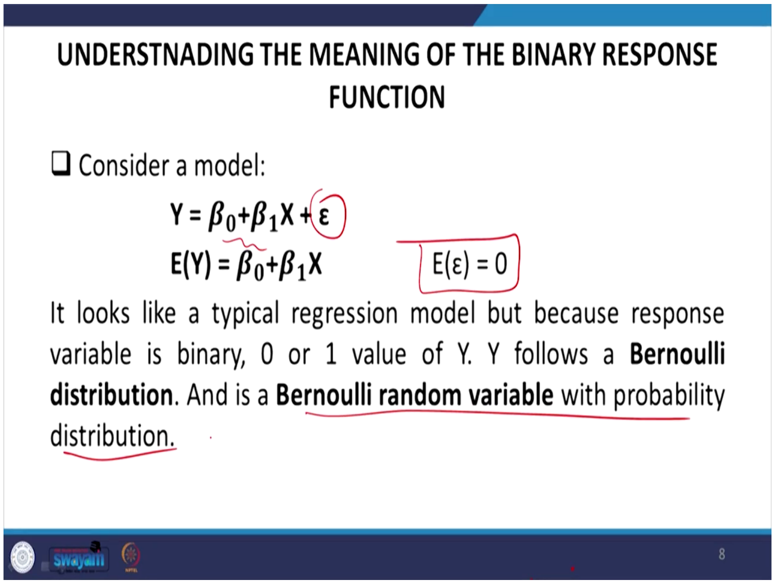


So, what are the fundamental difference between quantitative and qualitative response variable regression models? so first one is when Y is quantitative, we are now differentiating quantitative and qualitative response variable model to establish a very good understanding of the binary choice model as compared to the earlier existing models. When Y is quantitative the objective of a regression model is to estimate its expected value or mean value given the values of the regressors.

But in case of Y is qualitative having some limited choices or limited entries or specific discrete entries you cannot just go by Y to be averaged. There are some approximation through, may be through probability approximation, through yes or no if its binary or there is some likelihood estimation somewhere aligning with the probabilistic structure. So, likelihood is used where the marginal changes are little different, than that of the changes in the probabilistic structure.

So, in our case labor force participation decision by the women owning a house etc is more important. Hence qualitative response variable models or response regression models are often known as probability models since probabilistic structures are attached with the dependent variable.

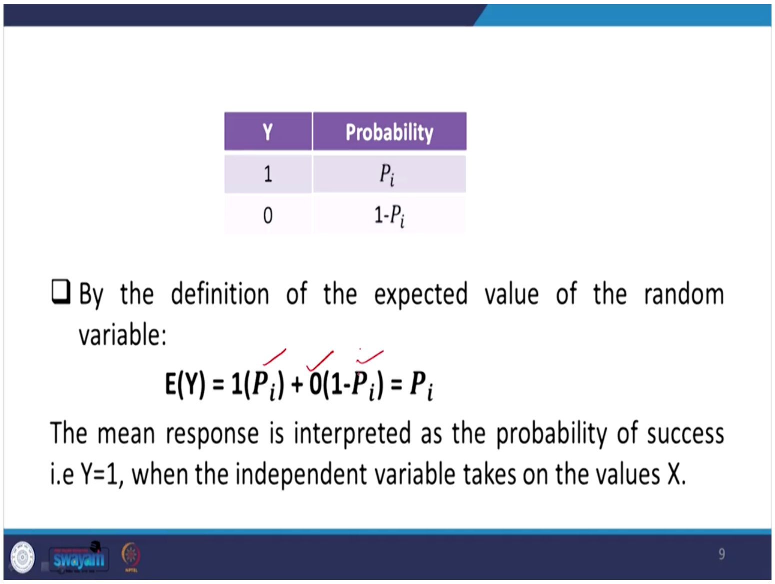
(Refer Slide Time: 09:13)



Let us clarify it, understanding the meaning of binary response function. Consider a standard linear regression model Y is equal to beta not plus beta 1 X1 plus error term epsilon. Expected value of Y is nothing but this portion only because this boils down to 0, because of the assumptions.

So, it looks like a typical regression model but because response variable is binary with 0 and 1 values so Y follows, had it been a standard Y continuous variable case then there is no problem. But since in this case only two responses are there that is 0 and 1 so Y follows a Bernoulli distribution. So, Bernoulli random variable with probability distribution. So, what do you mean by that? Bernoulli distribution usually takes values 0 and 1, or the binary entries.

(Refer Slide Time: 10:26)

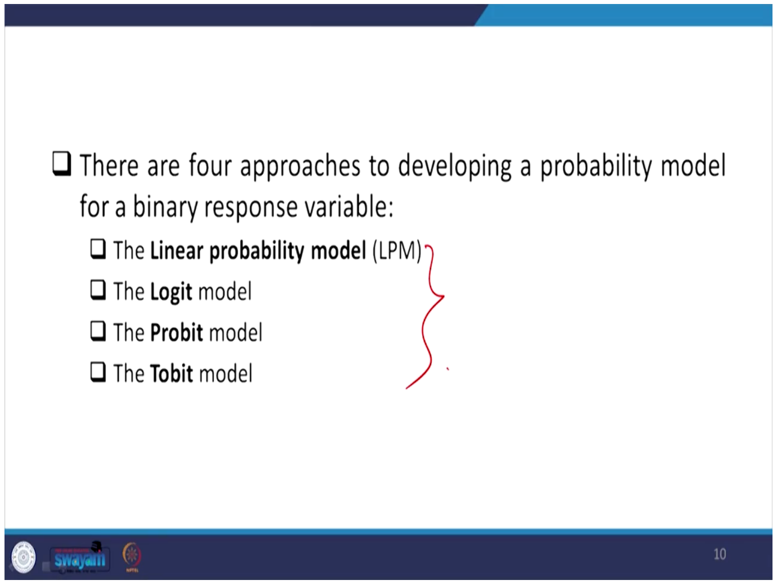


So, here we are discussing the Bernoulli distribution. How it looks like? We already stated with 1 as having or not having or success or failure. So, accordingly the probability is attached. If it is 1 probability is attached with Pi otherwise it is (1-Pi). So, by the definition of the expected value of the random variable, that is expected value of Y, either probability to be yes that is 1 times its probability and 0, not occurrence, 0 times it is another probability.

Suppose it is 50 percent then you can calculate it accordingly. So, the average here in Bernoulli distribution, even the average is nothing but the mean that is probability, that is P. in Bernoulli distribution P is for success and Q is for failure. So, P, the average of the distribution is nothing but P, in Bernoulli distribution.

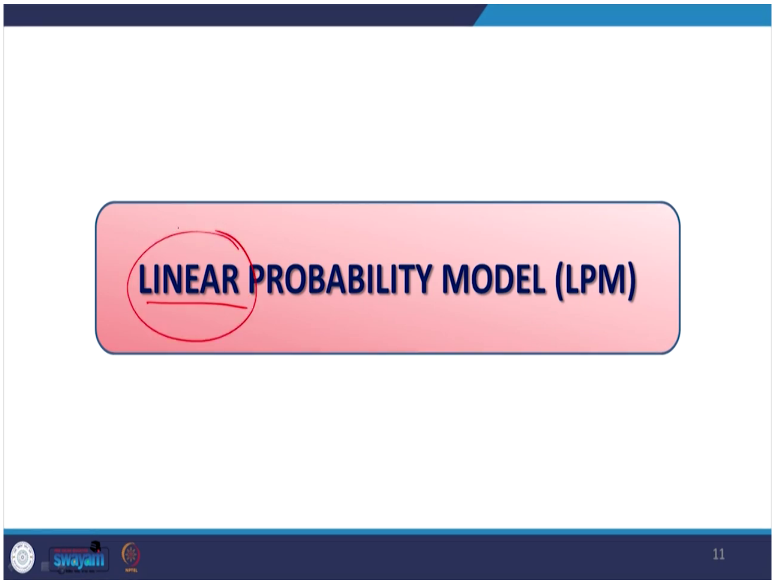
Here also we simply define that the average is nothing but the Pi, alright. But the variance if you remember, it is nothing but (P*Q). That is simply (P*Q) in the Bernoulli distribution. The mean response is interpreted as the probability of success when it is 1 and when the independent variable takes the values of X.

(Refer Slide Time: 11:58)



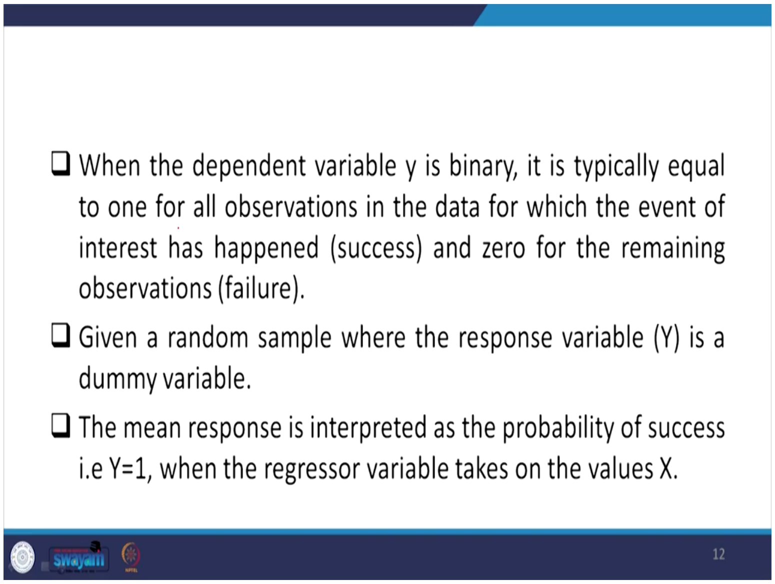
There are four approaches to developing probability model for a binary response variable. When a binary response variable is there we have broadly four models. So, these four we are going in discuss in our successive lectures. First one we will explore and scan its all the details of probabilistic model but established with ordinal regression format. We will go by ordinal structure and why ordinary regression model is not 100 percent guaranteeing the validity of the qualitative variables through the LPM. The more sophisticated models for this approach is Logit and Probit and somewhere specific to Tobit given the limited boundaries.

(Refer Slide Time: 12:46)



So, let us start with LPM model. LPM is single clarification given as probabilistic with linearity. Linearity is attached. We are going to define through the linear structure of the model.

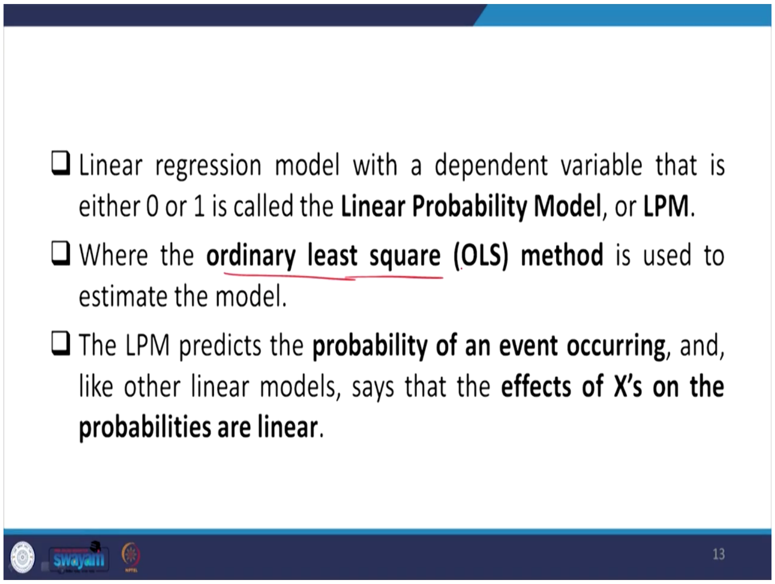
(Refer Slide Time: 13:06)



When the dependent variable is binary and with equal 1, if it is binary it is typically equal to 1 for all observations in the data for which the event of interest has happened that is success. Otherwise 0 if there is a failure. Then given a random variable when the response variable is a dummy variable. The response variable is a dummy variable and we are discussing LPM only.

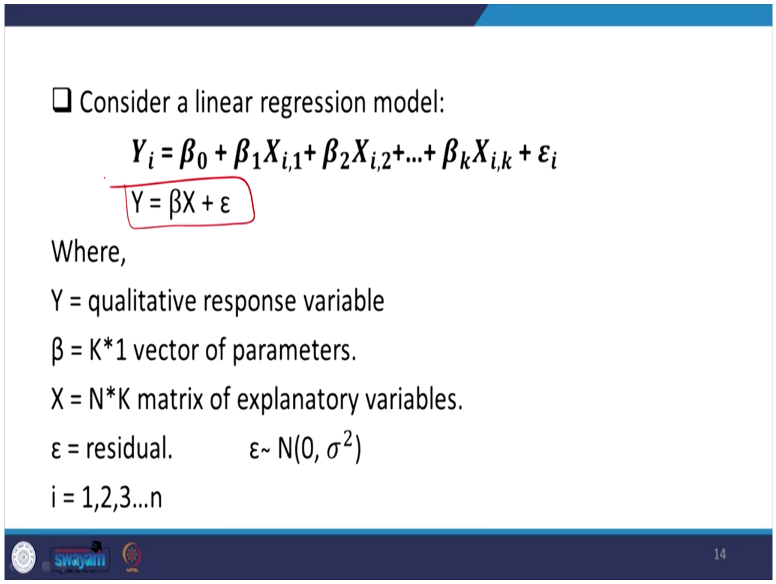
The mean response is simply interpreted as the probability of success that we have already proved in a couple of minutes back and where Y is equal to 1. So, the regressor variable takes any values of X.

(Refer Slide Time: 14:02)



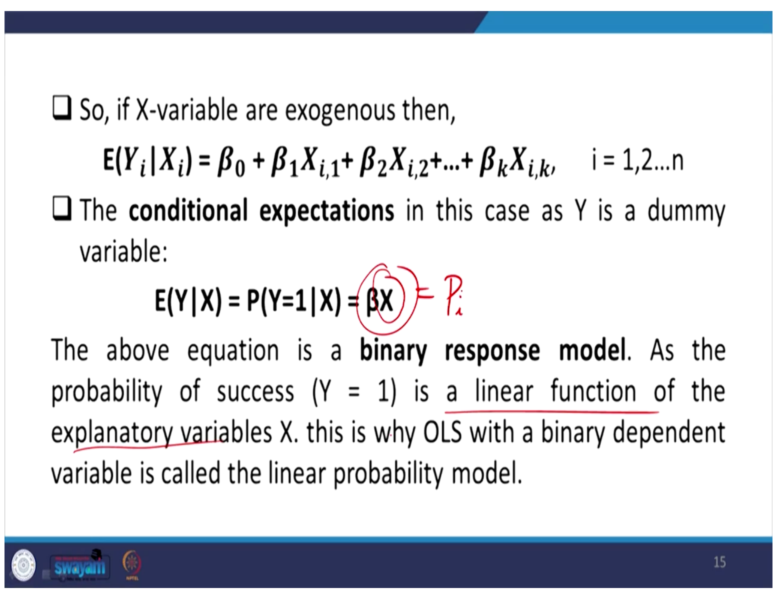
So, linear regression model with a dependent variable that is 0 or 1 is called linear probability model or LPM. So, where the ordinary least square method is used to estimate the model. So, as I already mentioned, since linear model we are mentioning and so OLS is applied. The LPM predicts the probability of occurrence of an event and like other linear models we say that effect of X on the probabilities are basically linearly related.

(Refer Slide Time: 14:45)



So, let us understand through an equation. Equation here we have different explanatory variables, we explained as Xi's. Then with epsilons for the error term. The expected value of Y is defined to be Y is equal to beta X. But in short the entire equation can be mentioned like this in matrix format as well. Beta times, beta of X, so here Y's are qualitative response variable. Beta has K times 1 vector of parameters. Whereas the X variable has N times K matrix, so N times its row and column is N*K. And epsilon represents the residual, follows a normal distribution 0 and, standard as 0 and with sigma square as the variance. Where i varies from 1 to n.

(Refer Slide Time: 16:01)

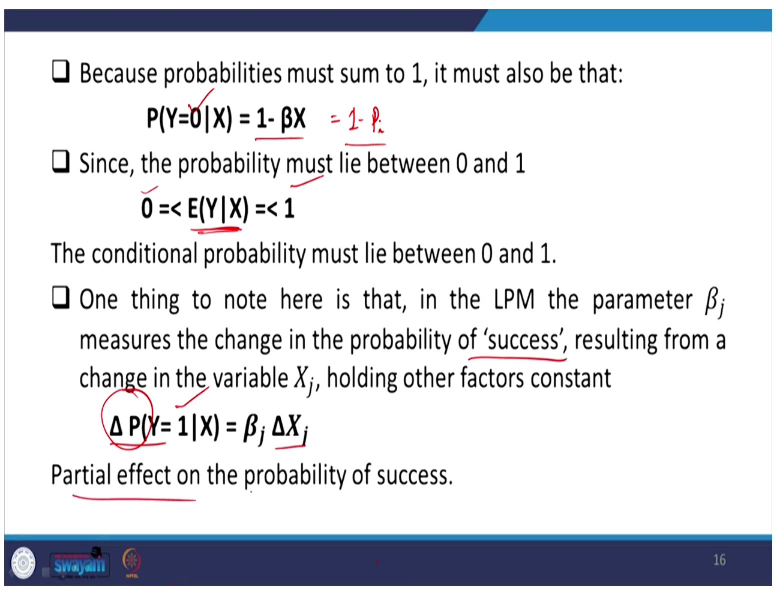


If X variable are exogenous then what is going to happen? If X variables are exogenous then we need to find out the conditional impact of X on Y. In this case, expected value of Y_i given X_i, the way we calculated. So, Y_i only represent 0 and 1 here instead of continuous variable. So, in this case the conditional expectation is Y, Y is a dummy variable we mentioned several times.

So, that means the expected value of Y given X if Y is equal to 1, given X is nothing but beta times X. That is basically the expected value and nothing but the P, the probability of success. Then the above equation is a binary response model as the probability of success is equal to 1 is a linear function of the explanatory variables, alright. So, basically this is linear function of the explanatory variables.

Now, this is why OLS with a binary dependent variable is called a linear probability model. That is only reason why we are saying it is a linear probability model, because this is also the Pi, isn't it, the probabilities of it.

(Refer Slide Time: 17:37)



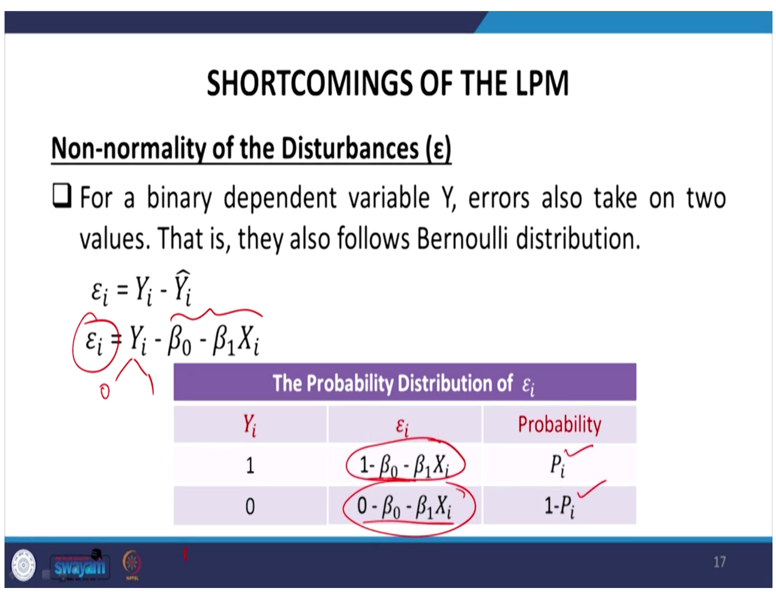
Let us clarify it further. Because probability must sum up to 1 it must also be that the probability of Y, if Y is equal to 0 given X, since the probability was beta X so non-occurrence of the event, that means probability of failure, it should be 1 minus beta X or 1 minus Pi, isn't it, probability of success.

So, here 0 since we are saying that if it takes 0 as the value then it will be 1 minus Pi or 1 minus beta X. Since the probability must lie between 0 and 1 so the expected value should lie between these two brackets 0 and 1. But if we are violating this bracket that means this module is not correct. We are going to test with the data. We have the real life data from the server data where we will test, these brackets are not followed by LPM correctly. We are going to tell it. The conditional probability must lie within this bracket.

One thing to be noted here is that in the LPM, the parameter B_j that is beta measures the change in the probability of success resulting from the change of the explanatory variable that is X_j holding other factors constant. So, it is not like probability of the dependent variable given the independent variable and their changes. Rather it is specific to probability of a success only, when it is equal to 1 like this.

So, that means the coefficient attached here is beta j times, if I take the first order derivative or the change of it so delta P of the equation, delta P should be equal to delta X here. So, beta j is equal to simply delta P upon delta X, alright. So, here it is just different than that of the Y. It is not delta Y; it is delta P. So, these are also called partial effect of the probability of success.

(Refer Slide Time: 19:47)

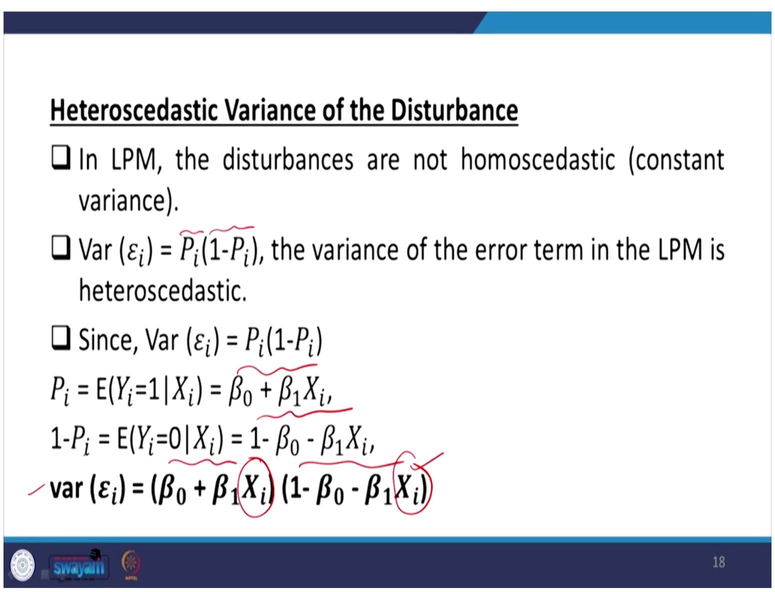


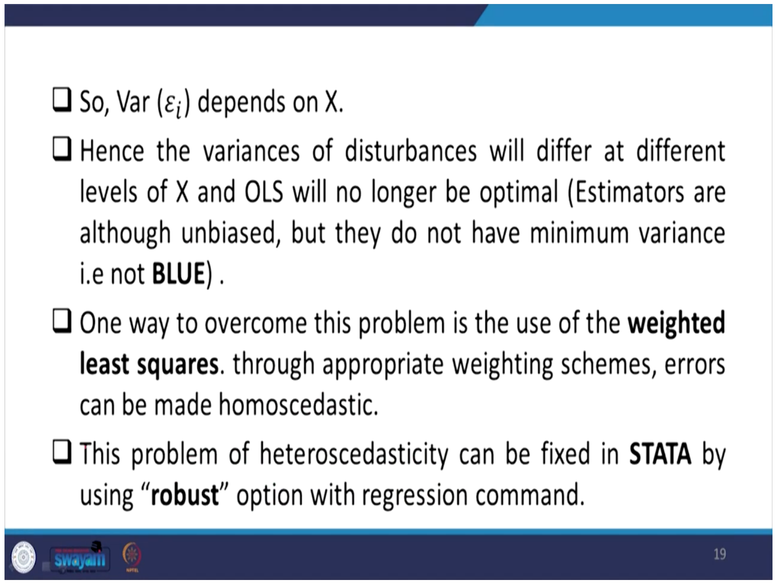
What are the shortcomings of the LPM then? After saying the theoretical, mathematical theory of the LPM model we are trying to mention some shortcomings of this model. Where, why LPM is not a good model so far as binary choice variables are concerned, so far as qualitative dependent variable is concerned? So, if the short comings are very clearly understood then you can go for any other advanced or sophisticated models.

First shortcoming of LPM is that there occurs non-normality of the disturbance term. The disturbance term has to be normally distributed. We usually say 0 mean with sigma standard derivation, sigma square as the variance but does not follow. It is not normal. All the entries are not there. Rather the entries of this is only with two points that is probability of success and probability of failure. So, let us calculate the error term. Okay. So, error will be what then? Y_i minus the estimated value of Y_i or Y_i hat. So, this is Y_i hat. We already said that it is beta not, Y1 hat is basically beta not plus beta 1 X1. If you subtract it we will get it for sure.

When we know that Yi only considers two values. When Yi considers only two values that is 0 and 1, so similarly our error term will also be consisted with two boundary points. That is 1 minus this or 0 minus this. So, with their probabilities, with Pi is the probability of success and probability of failure. So, error is not randomly distributed. So, it has only, since it depends upon Y_i's and Y_i values have a specific point that is 1 and 0 so the variability is not there. So, for a binary dependent variable Y errors also taken only two values, that is two values we have defined this and this, alright. So, they also follow, the error distribution also follows a Bernoulli format or distribution function.

(Refer Slide Time: 22:34)





Coming to another shortcomings of LPM model is related to heteroscedasticity. So, heteroscedasticity is one of the standard assumption in OLS. That is the error distribution, the disturbance term must be homoscedastic. But the standard deviation or the variance, standard deviation should be near about constant. But in LPM the disturbances are not homoscedastic or constant variance.

What do you mean by this? Let us examine the variance. Let us calculate the variance. So, how to calculate the variance then? Variance is equal to, as I told you in Bernoulli distribution it is P times Q simply. So, probability of success into probability of failure. So, P times 1 minus P if you multiply it, we get the variance. What is this then? What is this Pi?

Pi already we mentioned that is probability of success, probability of success of the dependent variable. So, given the Xi which is nothing but this we already derived before. So, 1 minus this then is the probability of failure. If I multiply these two terms together, this is the first part, this is the second part.

Look at this very clearly. We have two terms okay. In this case the variance we wanted to find out. What we exactly expect from the variance? That it has to be a constant term. After calculation, the variance should be boiled down to a constant coefficient. But in this case we have attached with a variable. Variable may be a scale, basically there are variabilities.

In the variance result we have a variable attached called Xi. So, Xi no longer defines the variance to be constant. Since Xi varies so variance is varying. So, it is not homoscedastic. Hence the variance of the disturbances will differ at different levels of X and OLS is no longer optimal.

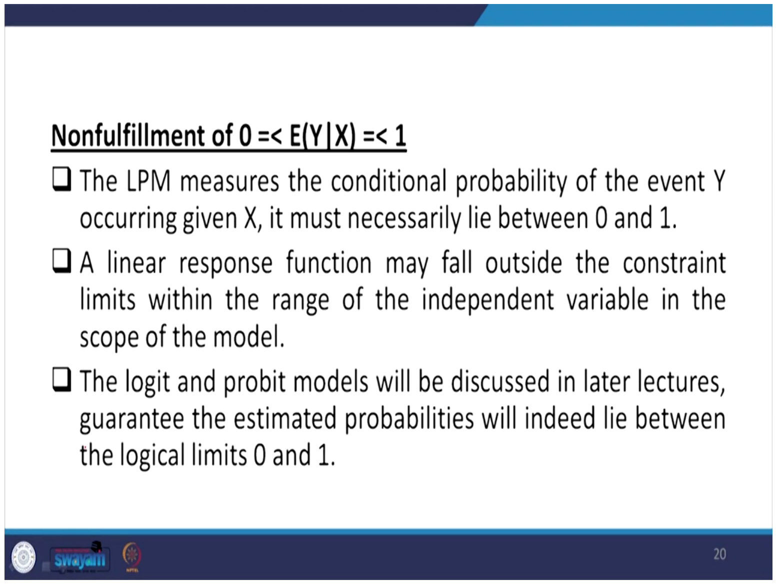
When variance is not followed or not constant then estimators are not optimal. May be unbiased, unbiased estimator as per the BLUE properties, expected value of the sample distribution is matching with the population mean that basically defines the biased or unbiased estimator. But coming to optimal, optimality is defined when our disturbance term is constant or near about 0. When disturbance are very less or constant then that defines optimality, optimal variance. So, that is problematic. So, BLUE properties does not follow.

One way to overcome this problem is the use of weighted least square method, I think WLS it is famously known. So, weighted least square method can erase some of the problems. Then this is through weighting scheme errors can be made homoscedastic. With some weight errors can be, the errors which we said can be nullified to a constant term.

Then coming to the problems of heteroscedasticity again, that can be fixed in STATA by using a robust option with the regression command. So, if you find out standard errors like every coefficient attached with a standard error. So, just standard errors since these are not optimal in our case, so we are not going to define standard error every time.

The appropriate channel of estimation should be attached with a robust standard error. So, in the command if you attach a robust followed by a comma that will give you a robust standard error. Robust standard error considers a weighting technique. So, that is important. We are going to show it, some of these differences later.

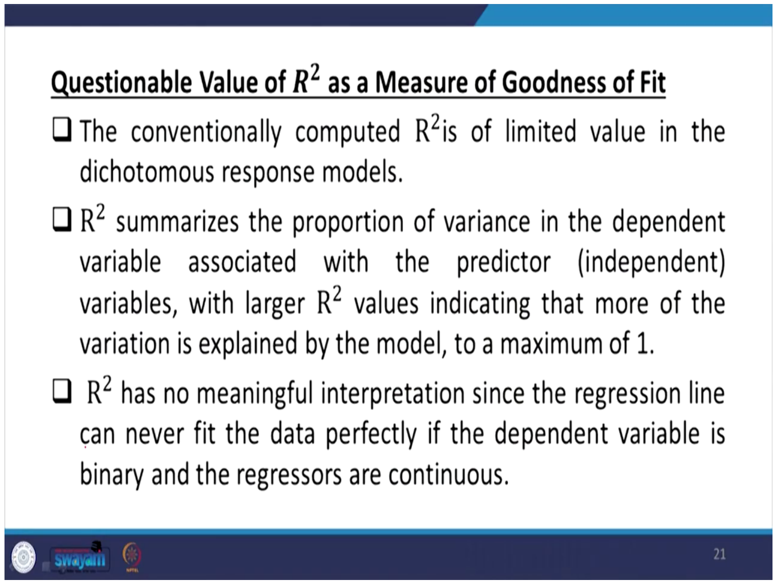
(Refer Slide Time: 27:18)

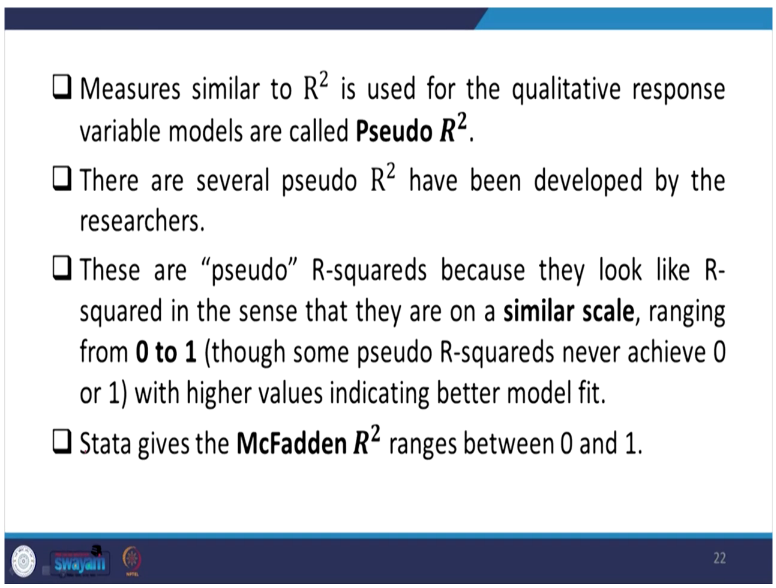


Let us understand other problems with LPM. It non-fulfils the bracket, the estimated value within 0. it has to be 0 and 1, the limit of the expected value, predicted value should be within 0 and 1 but through LPM we may not able to stick to 0 and 1. It might be problematic. The LPM measures the conditional probability of Y occurring X. It must be necessarily be within 0 and 1. But it may fall outside the boundary and we are going to show it with the help of data in next class for sure.

But Logit and Probit model will be discussed in the later classes as well. That guarantees the boundary limit of the probabilities lie within the limit of 0 and 1. So, couple of other things to be followed. We will then, probability is better to carry forward, you may cover up what is, another problem is there, okay. So, let me just explain this. With this we will stop this class and will continue with its practicle applicability and how why non fulfillment is there. Let me just stick to this slide these 2 slides and will move to the next class.

(Refer Slide Time: 29:09)





The last questionable aspect is related to measures of goodness of fit. Usually we go by R square, R square explain sum of the square divided by total sum of squares we mentioned, ESS divided by TSS, if you remember from the trend line, ESS only within the average of the population mean and above to the trend line or distance from the trend line to the population mean is basically all about the variance of the explained sum of squares. And total sum of squares basically the exact frequency, the difference than that of the population average, alright.

Conventionally we compute R square of the limited value in the dichotomous response models. We also calculate R square summarizes the proportion of variance in the dependent variable that is explained with the help of explained variables or the independent variables. With higher the R square indicating the better fit of the model and at a maximum level of 1, R-square has no meaningful interpretation since the regression line can never fit the data perfectly because the dependent variable is completely binary.

So, that is the reason why in binary choice model R square has very less value. Okay, but still there are some technique like Pseudo R square measures have been taken to understand grossly on how closely we are fitting to the model. So, McFadden R square we are going to refer in our next slide that is going to talk about.

Then this measure similar to R square is used, in this case square in the binary choice model R square, the name of that R square is called Pseudo R square and STATA usually gives McFadden R square to define the Pseudo. Pseudo means some representative R square, some other approaches to define a representative R square, McFadden technique is largely used in STATA.

There are several Pseudo R squares have been developed by different researchers. If you simply type in the help command as R square, you will find number of R square techniques. These are Pseudo R square because they look like square in the sense that they are on a similar scale ranging from 0 to 1, though some Pseudo R square never achieve 0 and 1, usually the optimal level reaches. If you remember we already discussed R square, adjusted R square. Adjusted R square can be negative as well but R square and even Pseudo R square should be within the limit of 0 and 1. Alright McFadden I already mentioned.

This is all about the background theory of the lecture on linear probability model and binary choice functions. From the next class onwards we are going to discuss with the data that which model is best fitted and which model to be picked up. So, let me stop here and we will continue with the data from the next class. Thank you so much.