Exploring Survey Data on Health Care

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

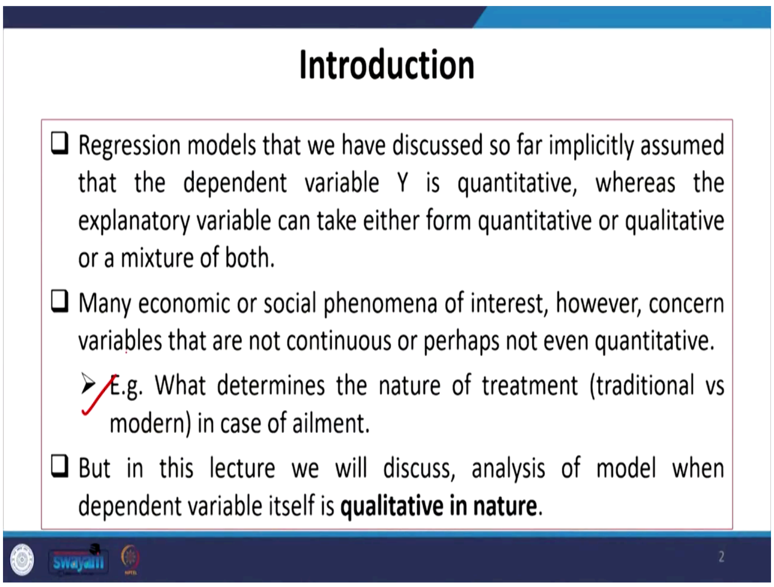
Lecture - 27
Regression Models of Qualitative Dependent Variables

Welcome participants to my NPTEL MOOC module on Exploring Healthcare Survey Data. We are in on the week explaining influential statistics in healthcare. In the previous lecture, we discussed regression models using the quantitative dependent variables. Here we are tending towards explaining the dependent variable which is qualitative.

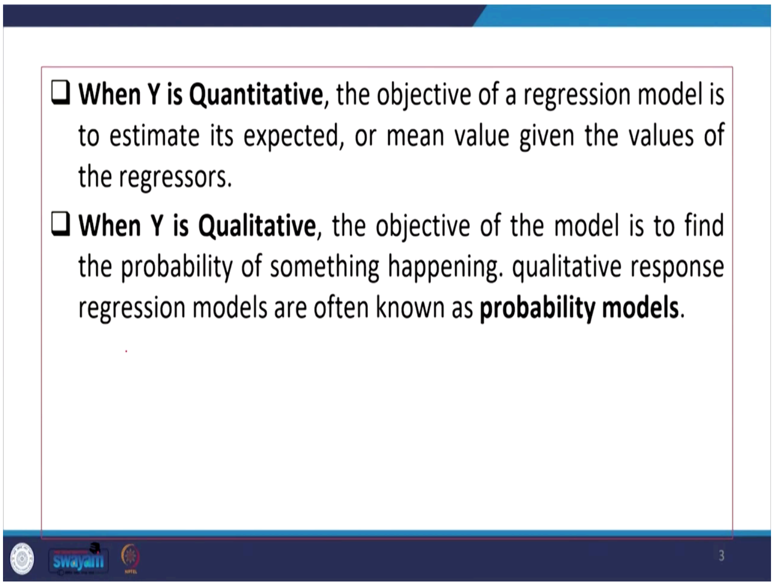
The regression models are not usually these models are also referred to as some generalized form of the least square model. We will discuss all those things now without consuming more time. Regression models that we have discussed so far implicitly assumed that the dependent variable Y is quantitative whereas, the exponential variable can take either form of variables may be a mixture, or maybe all quantitative maybe all qualitative.

Many economic or social phenomena of interest; however, concern variables that are not continuous or perhaps not even quantitative. But in this lecture we will discuss an analysis of the model when the dependent variable itself is of qualitative nature, you can just refer to this example for your own understanding.

(Refer Slide Time: 01:41)

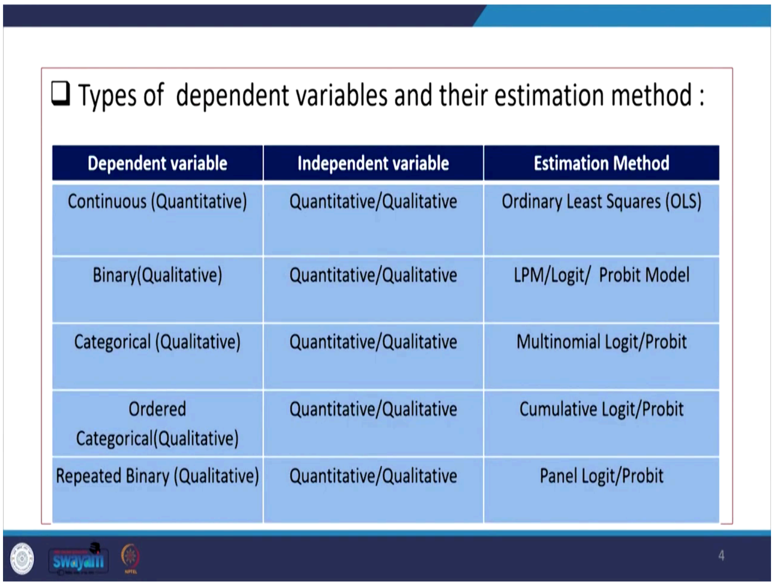


(Refer Slide Time: 01:42)



When the dependent variable is quantitative the objective of a regression in a model is to estimate it is expected or mean value given the values of the regressors. When the dependent variable that is Y is qualitative the objective of the model is to find the probability of something happening, qualitative response regression models are often known as probability models.

(Refer Slide Time: 02:11)



The types of dependent variables and their estimation methods we have made a comparison and you can understand very clearly what methods and what context is usually applied. If your dependent variable is continuous that is quantitative in nature and the independent variable may be mixed or in of the qualitative or quantitative, you are supposed to apply the OLS regression model.

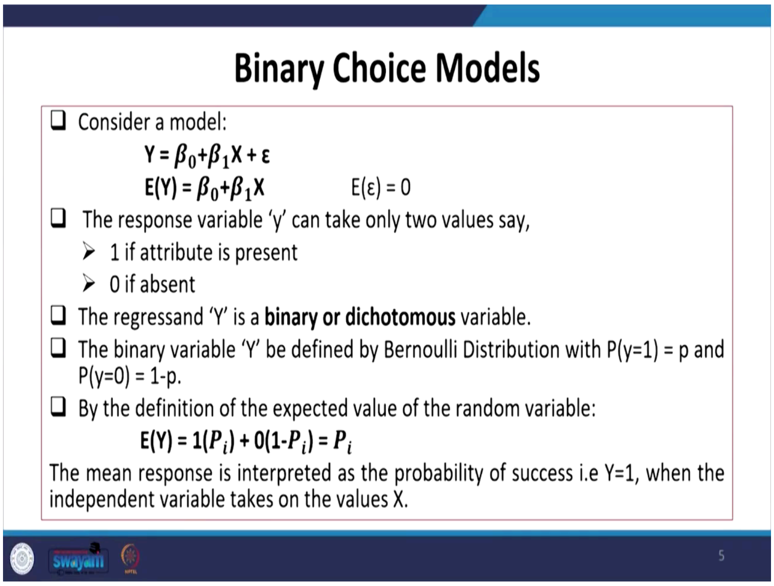
Where your dependent variable is binary, but qualitative binary in an independent variable may be of a mixed kind. You are supposed to apply the linear probability model logit model or probit model. There are, there might be, not necessarily the dependent variable to be binary it might be categorical if it is categorically usually discussed as qualitative data, irrespective of the independent variables, multinomial logit regression is applied.

If your dependent variable is ordered; ordered categorical, but order like education is the variable educational argument. In that case, you are supposed to apply the cumulative logit model or cumulative probit model these are called cumulative models if your data that is dependent variable is binary, but repeated binary your responses are in fact, repeated. In that case, as we already discussed from the beginning of the data panel logit or probit is applied.

Since data is a kind of longitudinal or a kind of repeated one irrespective of the kind of independent variable you carry. First of all, we will discuss the starting of binary choice models. I am not discussing all the things in detail I will just discuss with you the basic overview of the theory from the beginning, then I will suggest you to go through my previous lecture.

Last year I floated this year also that is running that is on exploring handling large scale data set that is all also of aid credit. You will get detailed guidance and their practical applications as well.

(Refer Slide Time: 04:38)



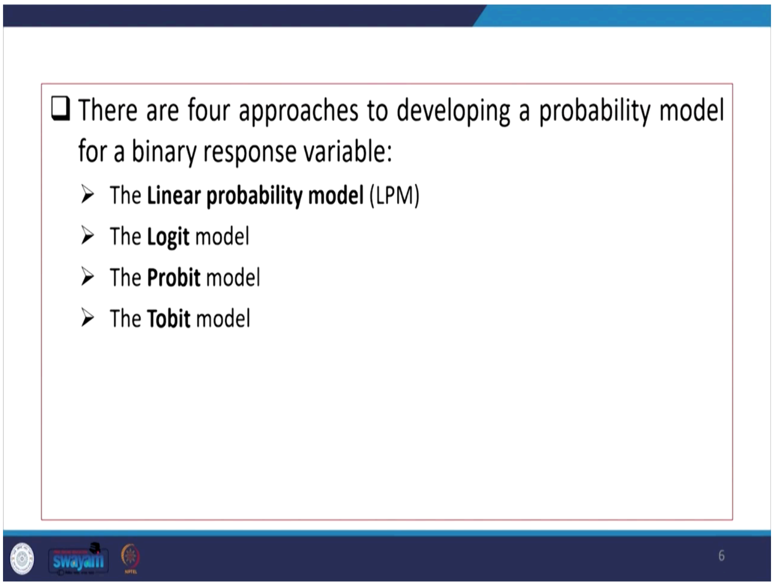
At this moment, I am just clarifying if it is a binary choice model the choice function is binary that is why if it is linearized with beta naught, beta 1 X plus epsilon the expected value of Y is in fact, defined as beta naught plus beta 1 X with the assumption that the estimated or expected value of the error term is equal to 0.

The response variable Y can take only two values for say binary as 1 if its attribute is present, else it is 0, that is, why the regression Y is called a binary or dichotomous variable which we just said. it has a probability with the distribution called the Bernoulli distribution.

When we have a binary response the distribution follows called Bernoulli distribution with p is equal to 1 that is small p and then p if it is equal if the dependent variable is with Y equal to 0 then other that is 1 minus p is the answer.

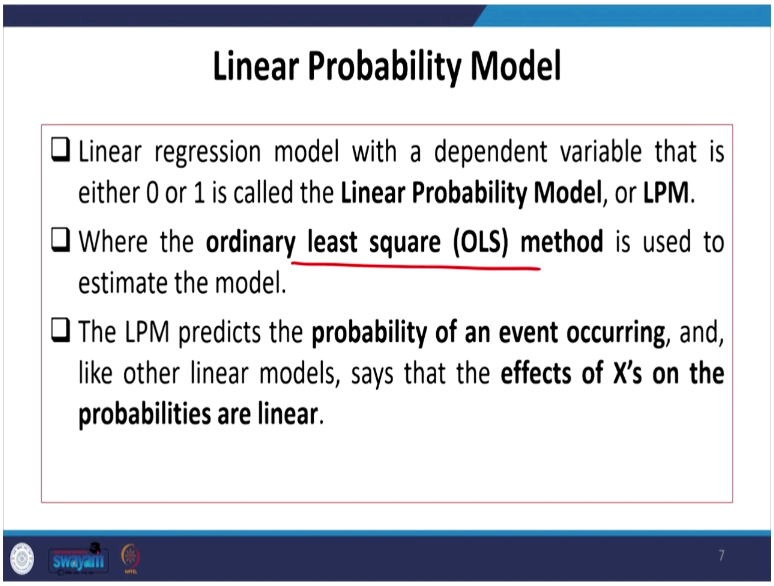
By the definition of the expected value of the random variable expected Y is, therefore, would be equal to 1 time Pi plus 0 that is 1; 0 0 times 1 minus Pi. So, 1 minus Pi is the probability of 0 as the occurrence Pi is the probability of one is the occurrence. Finally, the expected value of the dependent variable is nothing, but the probabilities. The mean response is interpreted as the probability of success that is Y equal to 1 when the independent variable takes on the values of x.

(Refer Slide Time: 06:07)



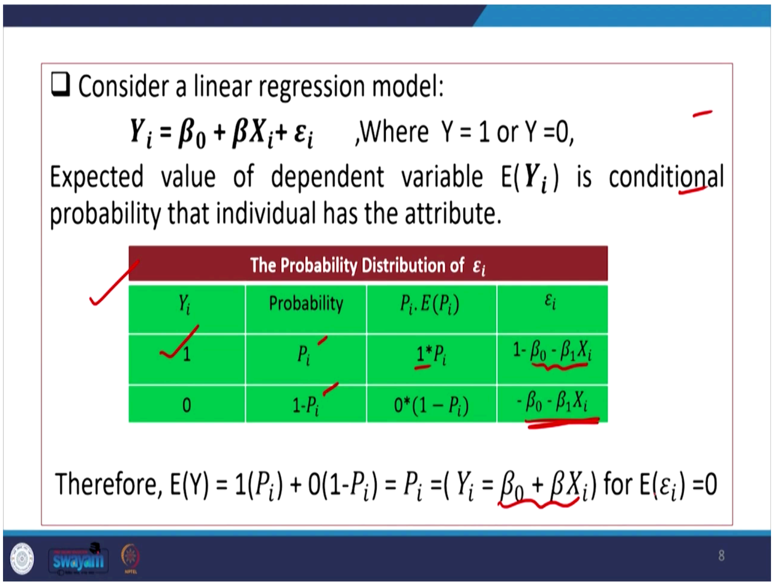
There are four approaches to developing a probability model for the binary response kind of variables. Broadly they are discussed as LPM Linear Probability Model the logit model, the probit model, and Tobit model.

(Refer Slide Time: 06:25)



Linear probability model LPM that you have to refer the book also to my previous module as I have already suggested it has some linearity in its responses where the ordinary least square method is actually applied. OLS is applied, the LPM predicts the probability of an event occurring and like other linear models say that the effects of Xi on the probabilities are in fact linear.

(Refer Slide Time: 06:54)

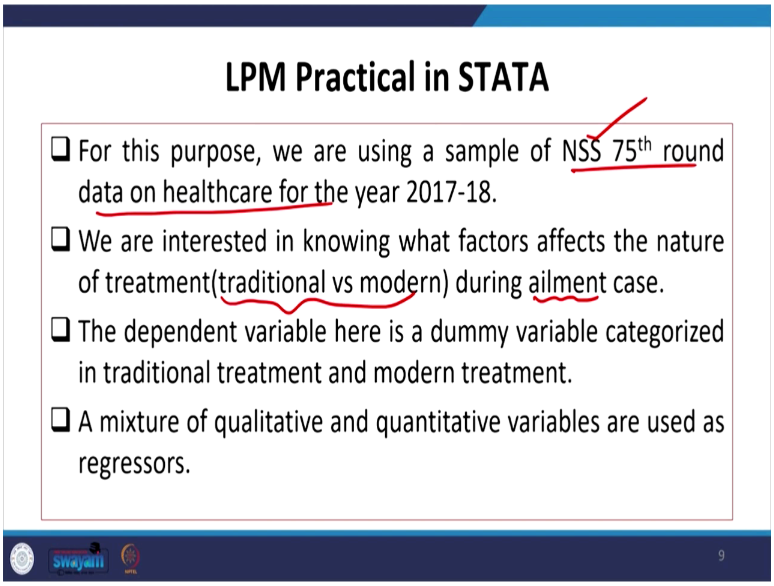


We are just explaining here to the model as we already explained I am not going to repeat I am just explaining the information given in the table. The probability distribution of the error term is equal to if Y is equal to 1 that is dependent variable is equal to 1 the probabilities are in fact, called P i. Otherwise if it is 0 the probability is 1 minus P i.

So what are the expected probabilities? Then one times P i if 1 is the occurrence in Y and 0 times of this. So, the error is basically 1 minus 1 minus the expected values. So, 1 minus that is the expected value we have already mentioned that this one 1 minus the expected value 1 Y 1 since Y i is 1.

So, 1 minus this 1 is the error term in the case of the one value of Y and otherwise 0 minus this. So, this is the expected value of the error term therefore, the expected Y is equal to 1 time Pi plus 0 times 1 minus Pi equal to P I, which is equal to Y i the expected value of this 1 for each of the expected value of the error term which is assumed to be 0.

(Refer Slide Time: 08:21)

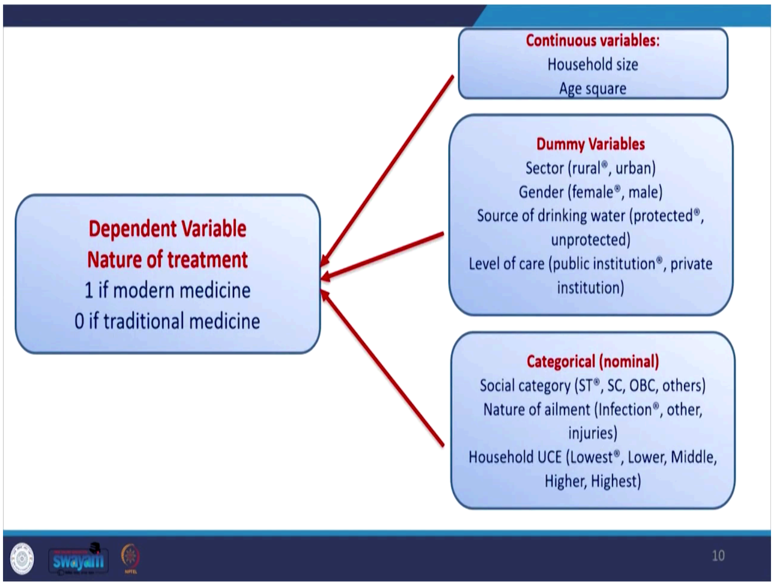


LPM and the practical in Stata were also explained in my previous model. for this purpose we are using a sample of NSS 75th round data on health care, this is the difference compared to my last year module on the previous module on handling large scale data using Stata.

This year we have included the NSS 75th round on health care data. We are interested in knowing what factors affect the nature of treatment that is traditional versus modern during the case of an ailment. If the ailment person is reporting access to some forms of medicines nonmedicinal treatment or nature of treatment those are traditional or modern.

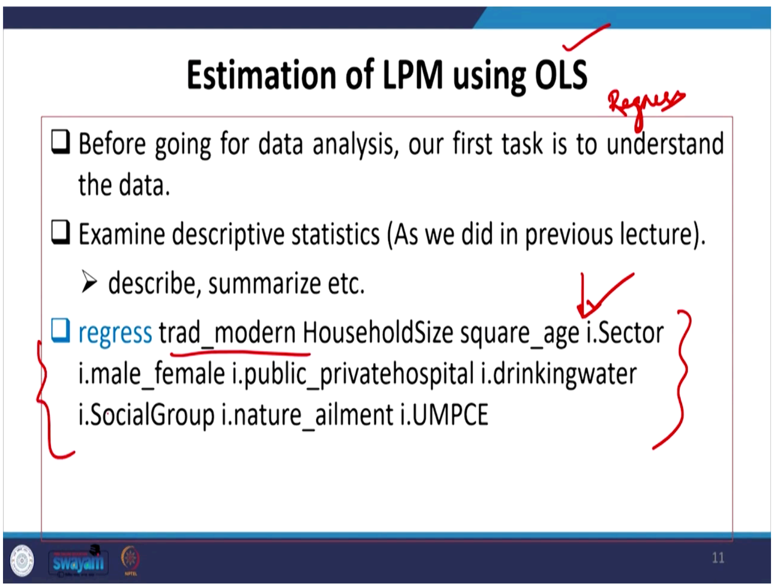
we have categorized into two the dependent variable here is a dummy one that is yes or no. A mixture of qualitative and quantitative variables is used as the regressors.

(Refer Slide Time: 09:31)



Here are the dependent and independent variables. dependent variables we have just defined as 1 is for modern medicine, 0 is for traditional medicine. And the independent variable we have taken for explanation. We have already discussed independent variables in our regression model in the previous lecture. Here we have to consider some continuous variables, some dummy variables, and some categorical variables. You can read between these lines for your better understanding.

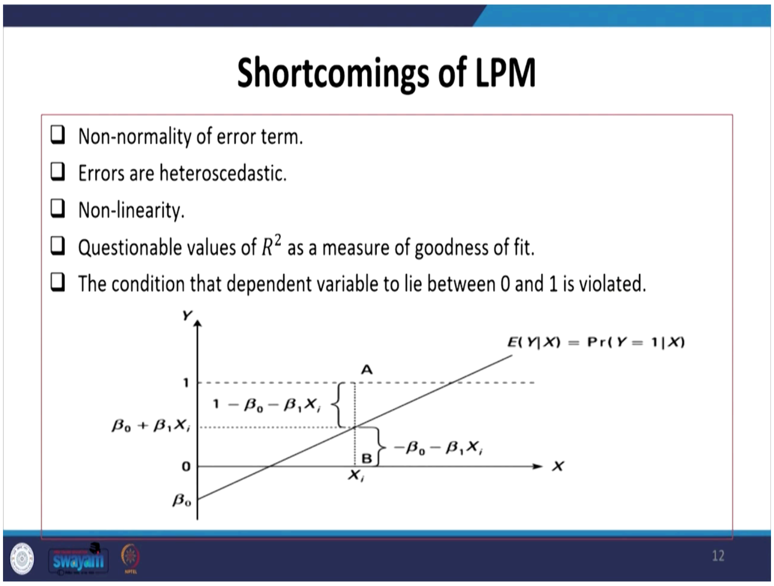
(Refer Slide Time: 09:58)



the estimation of LPM using a simple OLS technique is taken. So, as we know when OLS is applied, when the OLS result is derived, simply apply the command called regress. So, simple regress with the first variable which is the dependent variable followed by the independent variable some i dot commands are given.

We have already explained that i dot commands are assigned or added just to identify whether that variable is in fact, categorical or not. Whether this has considered any base category for comparison and it impacts on the medicine on the choice of treatment. So, this is what is made we will also do our practical at the end. And I am just clarifying each of the concepts, and you can also run like these using the practice data set that is going to be uploaded.

(Refer Slide Time: 11:06)



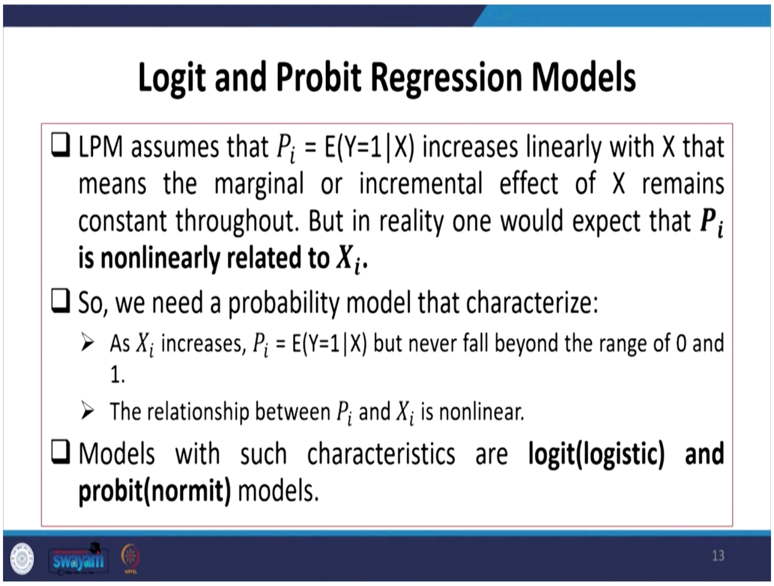
The shortcoming of this LPM are that the distribution that has been considered is not normal. Because it is following a Bernoulli distribution with 1 and 0, it is not normal. So, it is not avoiding not the errors are not estimated to be 0. So, errors are basically heteroscedastic that is important you may be asked the question if the distribution is not normal. Then what would be the error distribution? Does it have homoscedasticity or heteroscedasticity whether it violates or not?

Similarly, the non-linearity aspect is also identified since the data is not following a linear pattern, questionable values of R square, of course, because your distribution is not normal.

So, the goodness of fit in question the condition that the dependent variable to lie between 0 and 1 is in fact, violated.

So, the 0 and 1 that we have said may not be actually derived correctly. So the rest I have already explained in my previous model you can just follow.

(Refer Slide Time: 12:22)

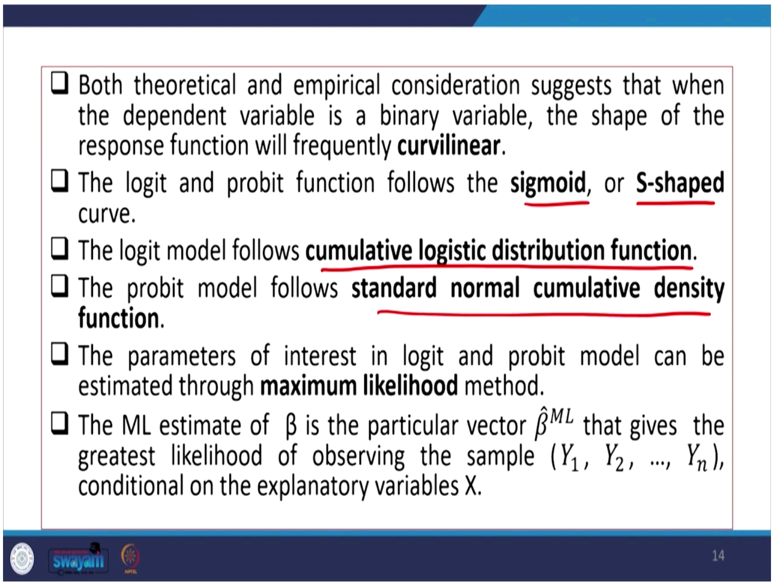


Now, coming to other two important and widely applied regression models in this context are called logit and probit regression models. LPM assumes with 1 and 0 P i is equal to the expected value of Y if it is 1 given the conditional values increase. When that P i increases linearly with X; that means, the marginal or incremental effect of X remains constant throughout, but in reality, one would expect that P i is non-linearly related to the regressors.

So, we need a probability model that characterizes the Xi with the Pi values but never falls beyond the range of 0 and 1. The relations between Pi and Xi is non-linear models with such characteristics called logit and probit. If normal distribution in case of the normal distribution is followed.

A probit model is also called normit model, normit distribution model and here it is called logistic distribution models. Both theoretical and empirical considerations suggest that when the dependent variable is a binary one the shape of the response function will frequently be curvilinear. The logit and probit function follows the sigmoid curve.

(Refer Slide Time: 13:53)



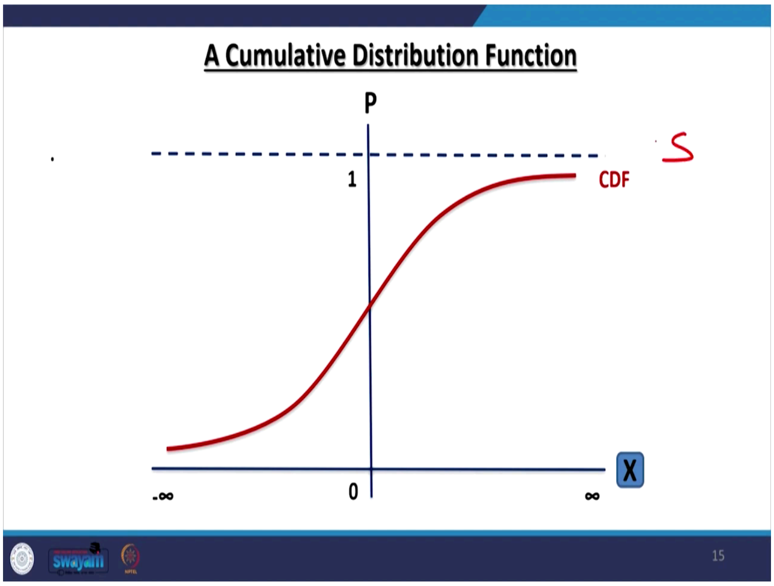
Sigmoid you might be asked a question that which distribution it follows. In LPM we have already said that that is following a Bernoulli distribution, here it is s-shaped as a sigmoid curve the logit model follows a cumulative logistic distribution function. Again the curve that is explained is called the cumulative logistic distribution function.

The probit model follows the standard normal cumulative density function which is the difference between logit and probit. Standard normal, cumulative density function where in case of logit it is called cumulative logistic distribution function. The parameters of interest in logit and probit can be estimated through the maximum likelihood estimation method.

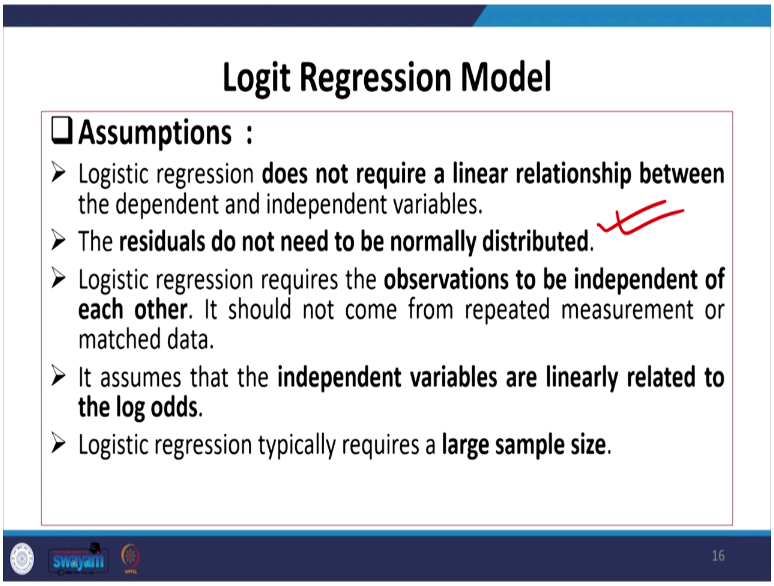
I think you might have heard about different estimators one is called the ordinarily least square method. Another method is called the maximum likelihood method.

What is the likelihood of being included in that particular model? The maximum likelihood estimator of beta is the particular vector that is also written as estimated value beta hat ML that gives the greatest likelihood of observing the sample conditional on the explanatory variables.

(Refer Slide Time: 15:17)



(Refer Slide Time: 15:27)



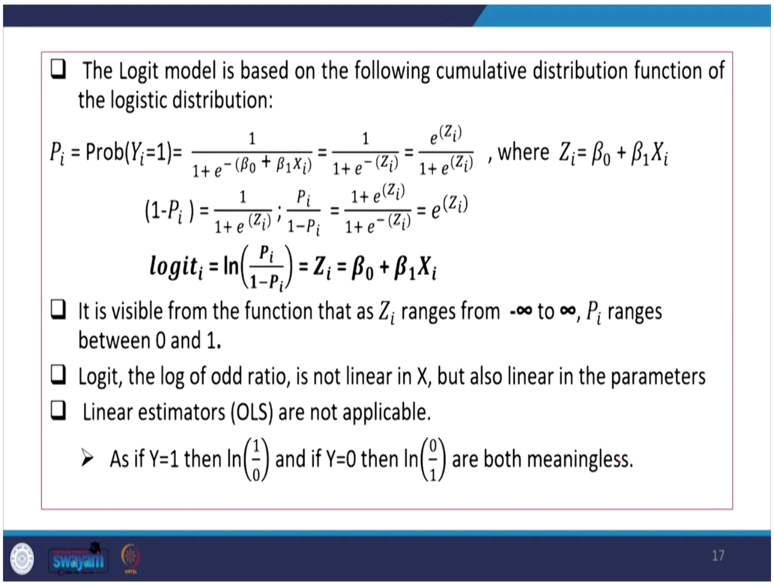
This is what it looks like this is a S-shaped curve called a sigmoid curve. This is a kind of logistic map the logistic model have certain assumptions that it does not require a linear relationship between the dependent and independent variable. That is why we said if OLS is violated then we may go for some form of non in the non-linear model.

But again these are called transform linear models. These are also called generalized forms of model they are also called GLM. In some cases of GLM we will explain these to some extent in one of the modules, In one of the lectures the residuals do not need to be normally distributed.

It is a distribution is s-shaped this is sigmoid kind. So, not necessarily following a normally distributed one this is quite important to note. Logistic regression requires the observations to be independent of each other. It should not come from repeated measurement or mass match data this assumes that the independent variables are linear related to the log of odds.

If the odds are taken as a log; that means, the distribution comes out to be linear. So, we will show that also in our derivation. Logistic regression typically requires large sample data that has to be understood.

(Refer Slide Time: 16:42)



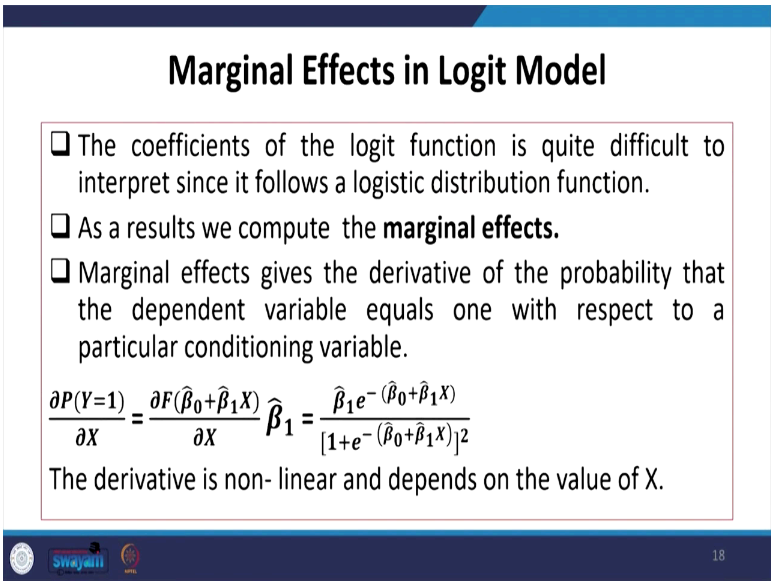
Here is the model when we take the log of this one basically you just go through this I have explained very clearly in my previous module on handling large-scale data, this is the logistic function. It has the expected value the exponential values, X exponential opens the standard normal distribution, but it is exponentiated.

So, then exponentiated then another one actually the function is, in fact, presented like this. And it is probable and it is odd of the probabilities or 1 minus P i is represented here. Finally, the logit is actually called the transformation of the function of P i to that of 1 minus P i this is also called the odd ratio.

This is when we take a log of all those things we are finally, getting a linear function. So, the log of the odd is nothing, but the OLS nothing, but in fact, the linear regression model. So, it is visible from the function that the Zi is ranging from minus infinite to plus infinite that you can just read between the line.

So, logit that is basically a log of odd ratio is not linear in X, but also linear in the parameter the way we discuss in the assumption of OLS regression. So, linear with the parameter and non-linear with the variables. So, the ordinary least square method is not applied when you have this kind of function.

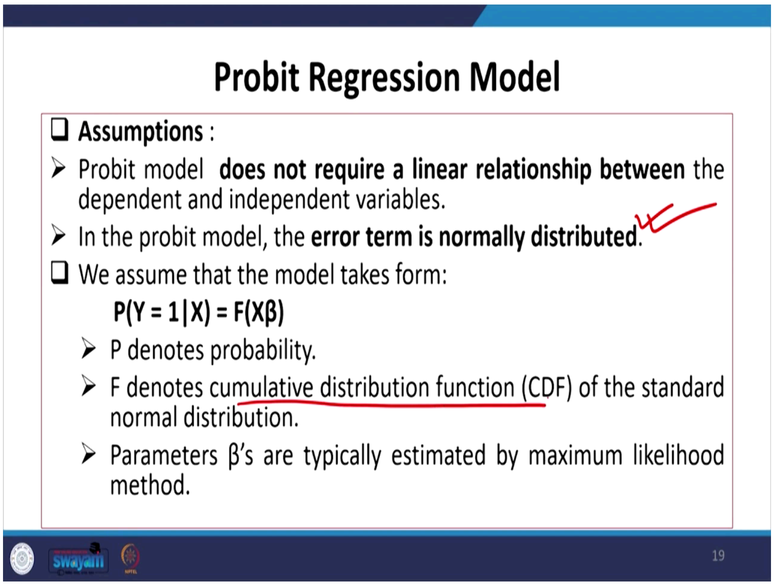
(Refer Slide Time: 18:32)



whenever you derive a logit model it is always suggested to derive marginal effects because logit is not have with continuous variables. by ordinal least square regression we get all marginal changes marginal effect of each conditional variable or control variables on it is the dependent variable, but since in logit, it is a sporadic change from 1 to 0.

It follows a path different than that of a continuous since the dependent variable is not continuous. We are supposed to further take the marginal effect. So, the marginal effect basically the derivative again is taken like this. So, d P by d X and you can just follow between the lines and how it is interpreted.

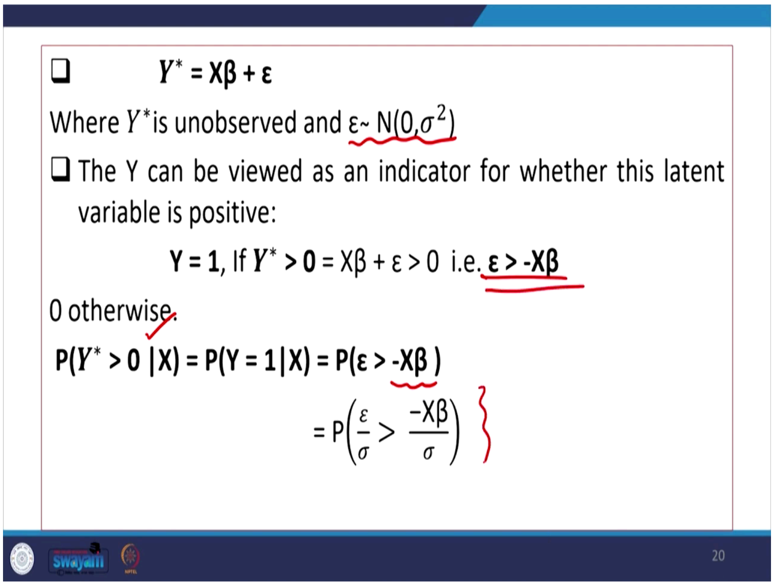
(Refer Slide Time: 19:27)



The next one is called the probit regression model, probit regression model. This does not require a linear relationship between the dependent and independent variables probit model the error term is normally distributed that is most important. But in the case of logit, it is not normally distributed.

the probability function is Y is equal to 1 given X. So, in sort, it is this cumulative frequency are written as X beta rest of the detail the is given here. This is the distribution follows as CDF Cumulative Distribution Function.

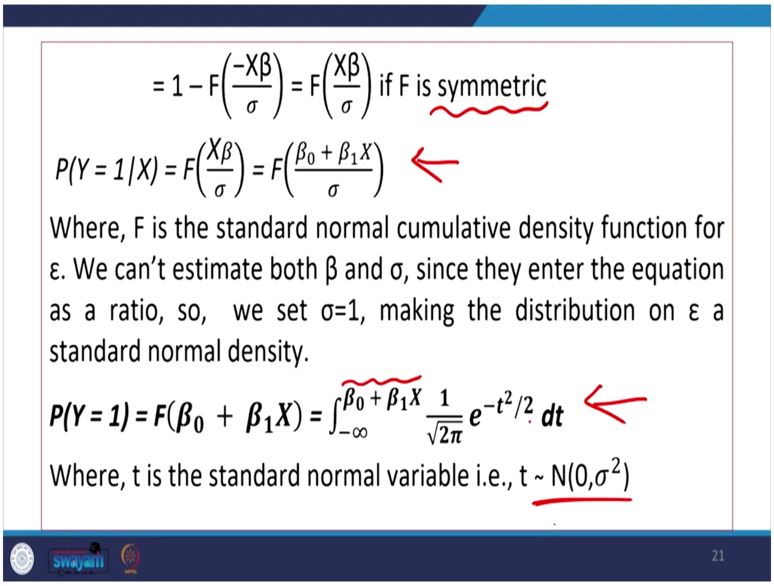
(Refer Slide Time: 20:08)



And the probabilities, in this case, are defined we have already said that it should follow it is error distribution to be normally distributed. And the probability is having greater than 0 values, given different Y is greater than 0 values given different control variables.

It suggested that p is equal to Y is equal to 1 given X should equal to the probability of the error distribution and the error distribution should be at least greater than that of it is estimated values. Therefore, this is how it is defined you can just follow it therefore, the probabilities will be presented with this range.

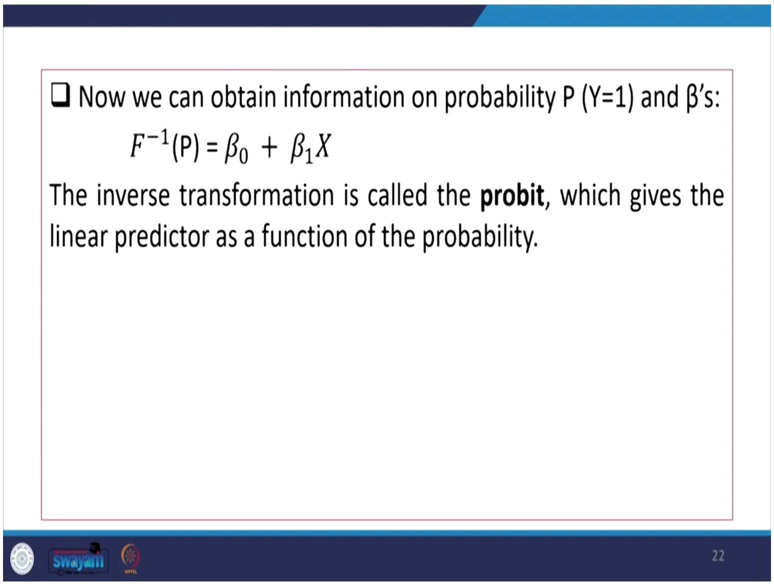
(Refer Slide Time: 20:58)



Now, accordingly, we arrive at the density function you can just see once how density function is defined it is basically defined as symmetry. The, therefore, the error distribution is defined to be normal. So, it follows the normal distribution with these functions.

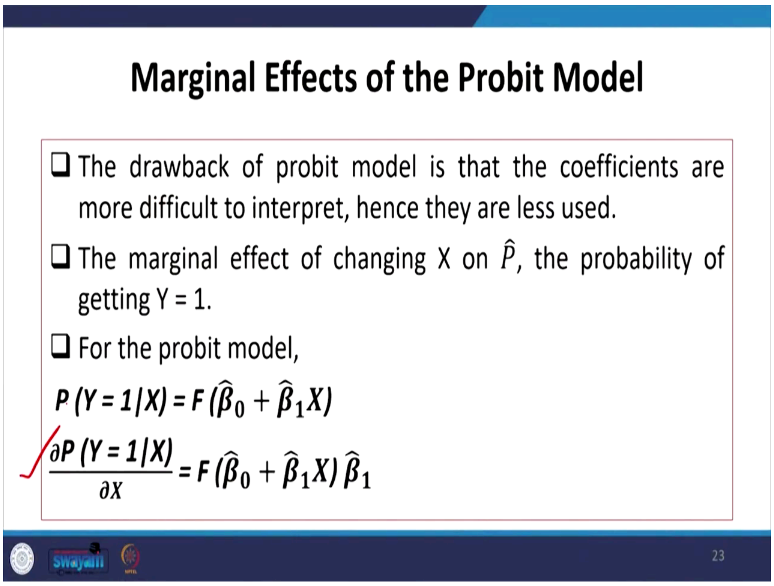
We know that normal distribution that varies from minus infinity to the estimated value of the beta one of 1 upon square root of 2 pi times it is its exponential value of the Z value or t values here with it is marginal changes. So, t is the standard normal variable with the function is varies from 0 with 0 mean and sigma square standard deviation or sigma square as the variance.

(Refer Slide Time: 21:47)

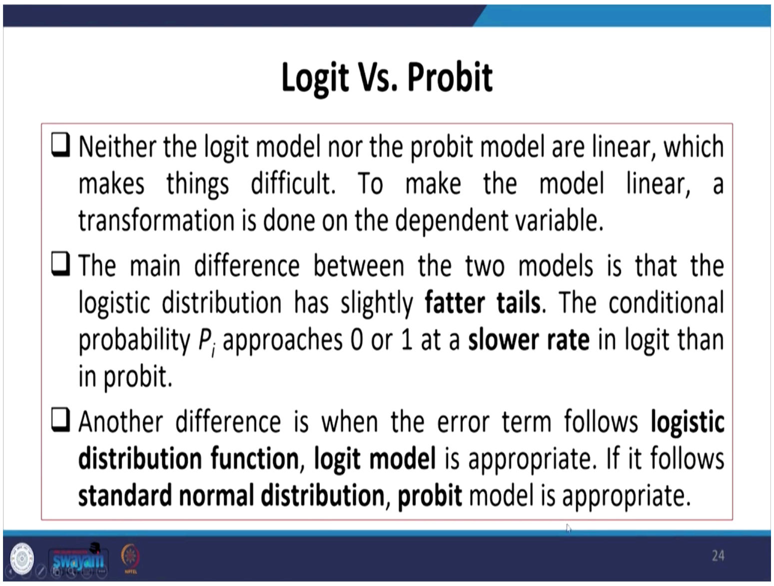


Now, we can obtain the information and probabilities rest you just can follow I will come to the practical aspect of it. Similarly, since it has probabilities and it is extreme values with 1 and 0 it is always suggested to go for marginal effect.

(Refer Slide Time: 22:09)



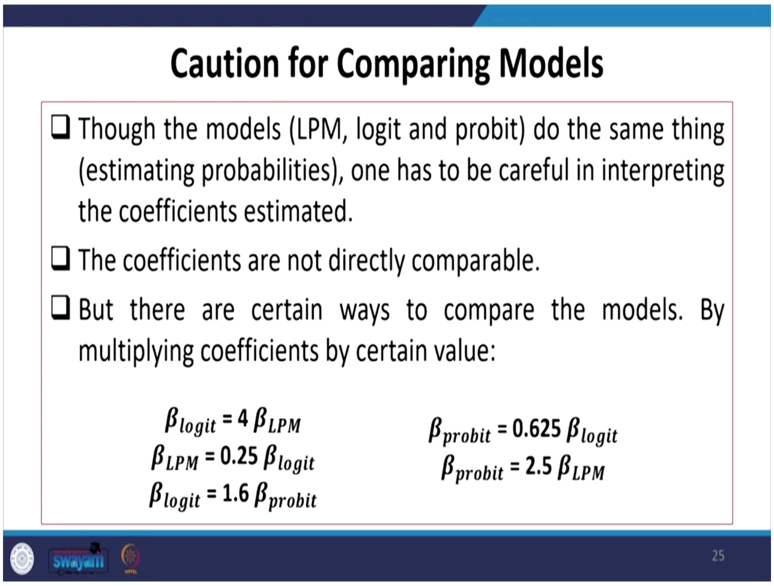
(Refer Slide Time: 22:13)



Again we take the derivative of the probabilities function with respect to it is X and accordingly, it is interpreted. Now it is important to know the difference between logit and probit. Neither the logit model nor the probit model is linear which makes things difficult to make the model linear transformations are taken usually log transformation is taken.

The main difference between the two models is that logit distribution has slightly fatter tails the tails are fatter. Whereas, in the case of the conditional probability P the conditional probabilities Pi approaches if 0 or 1 at a slower rate in logit. So, in the probit model, we will have fatter tails, but in the case of logit, it has a slower rate.

(Refer Slide Time: 23:06)

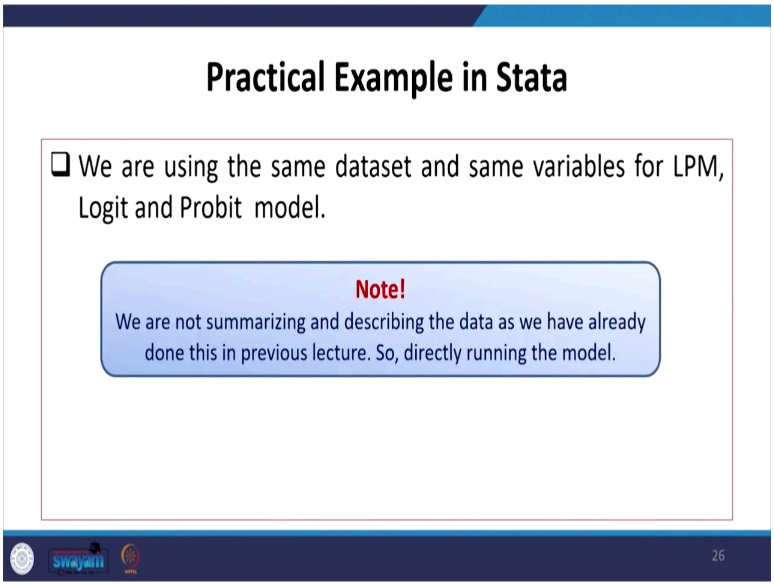


No, I think I have said it a little differently I will show it here maybe I think in my previous model I have said let me show you once again yes. The logistic distribution would have the tail should be fatter. Whereas, in case of probit, it is more is less fat.

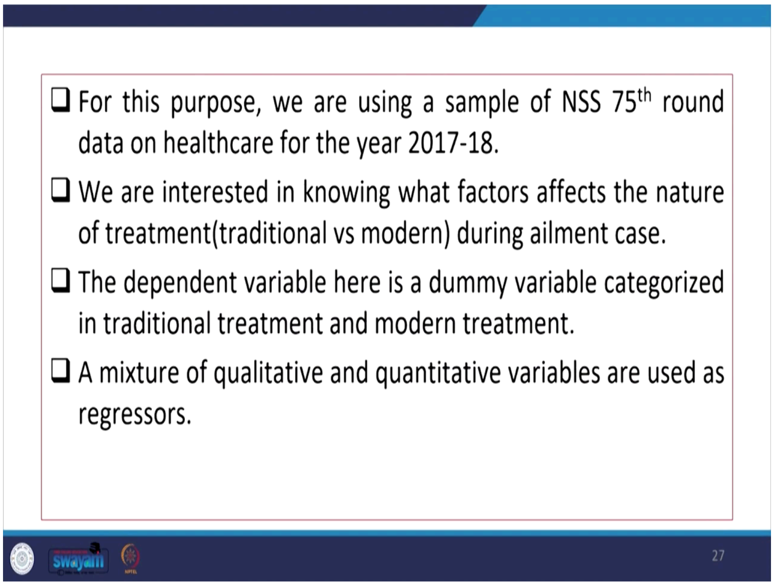
Another difference is the is when the error term follows the logistic distribution function logit is more appropriate. Whereas, when there are distribution is normal or normally distributed cumulative density function is normal then probit is most appropriate. These two points are quite important.

Now, some questions for comparing models so I have said in earlier models how these are estimated, what are the comparison points beta value in case of logit is 4 times the beta value of the LPM model linear probability model? Beta value in probit is 0.625 times of it is logit similarly other values you can just compare and for your record, these are already estimated and made the comparison.

(Refer Slide Time: 24:17)

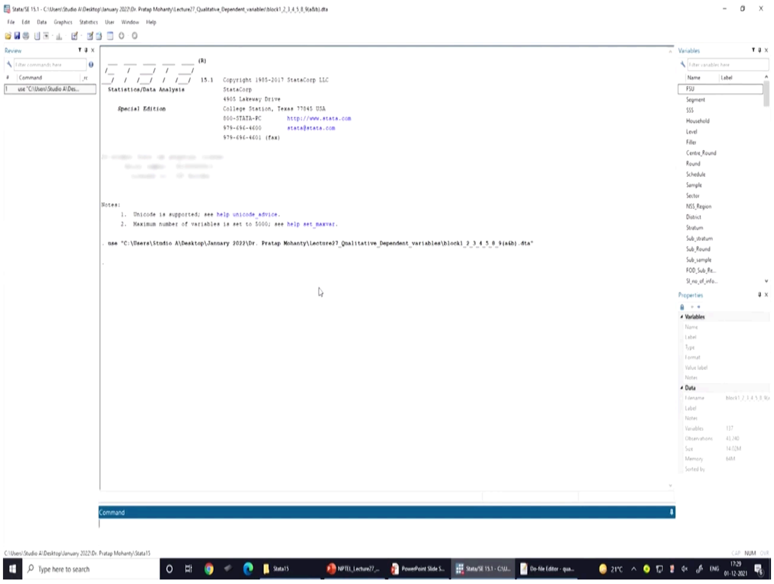


(Refer Slide Time: 24:24)

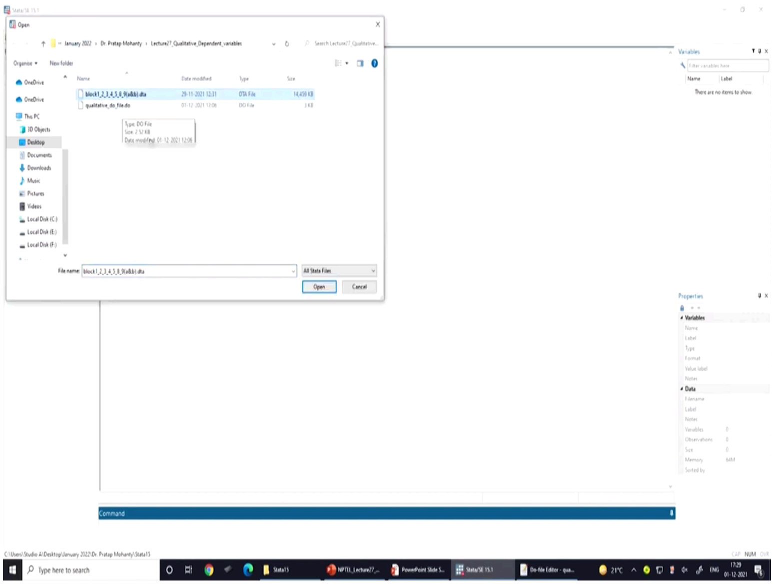


Now, we are going to operate with our practical example that we have taken the NSS 75th round data and we will also explain you.

(Refer Slide Time: 24:29)

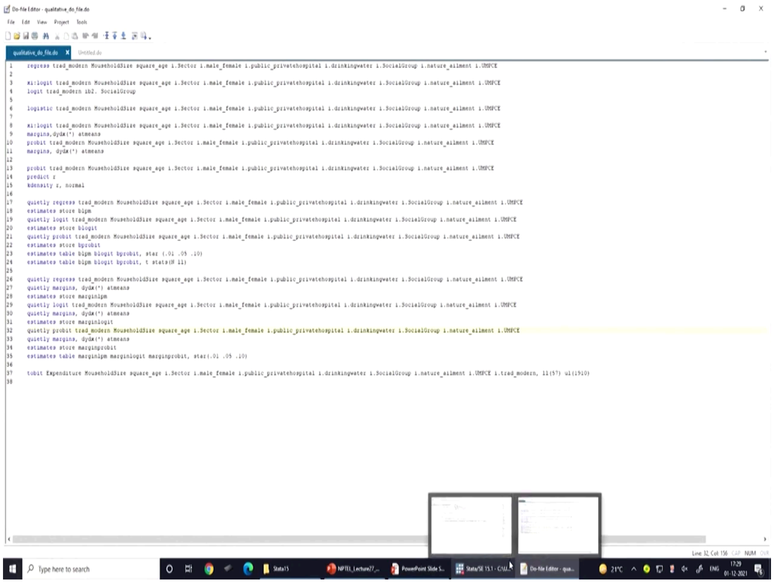


(Refer Slide Time: 24:36)



Here is your stata we are going to open the data on your screen. The sample data that we are also going to share with you these are the MOS data 1, 2, 3, 4, 5, 8, 9 of NSS 75th already the fine tuned data after merging.

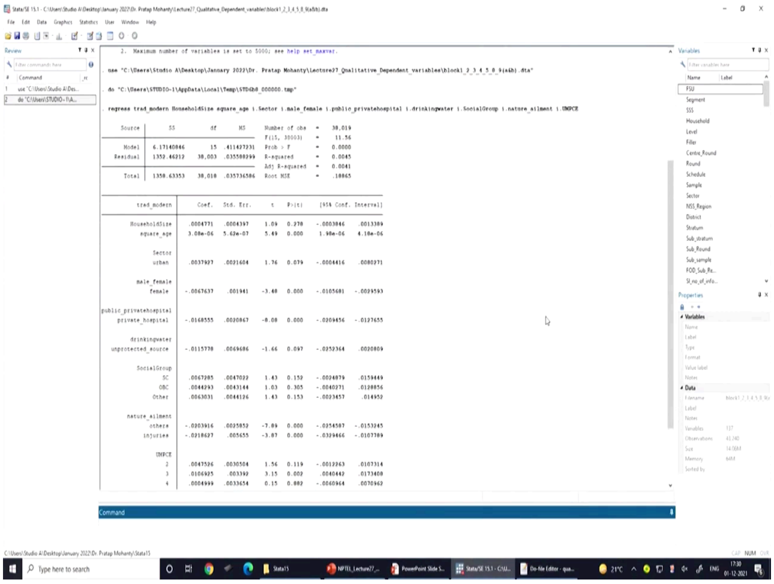
(Refer Slide Time: 24:48)



Now, we will also open the do file. We have made just for your reference you can we have already discussed several times about how to develop a do file and how it helps a lot in running the model. We will first regress. Why regress is required?.

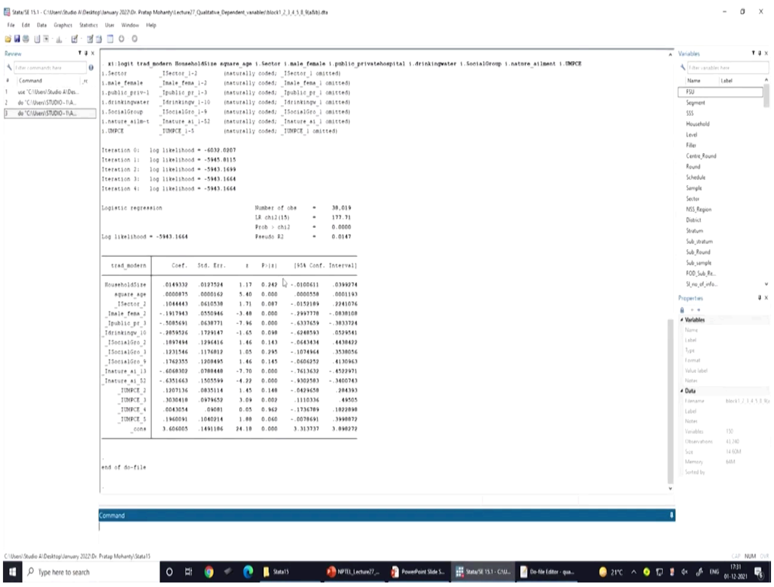
Because if you regress with the limited dependent variable that is dependent variable is categorical the result we are deriving is basically called LPM first one we will regress the LPM result on your screen. Therefore, that is simply taking the command as we take in wireless regression.

(Refer Slide Time: 25:27)



Here on the screen this is the one. This gives the result of the LPM, since our dependent variable is limited. Now we are again explaining another one logit. A logit, but another aspect is we are given conditional logit conditional logit Xi command is given on the next do file command. We have mentioned is Xi, Xi command is basically giving you some conditional information about it is about it is reference categories.

(Refer Slide Time: 26:10)

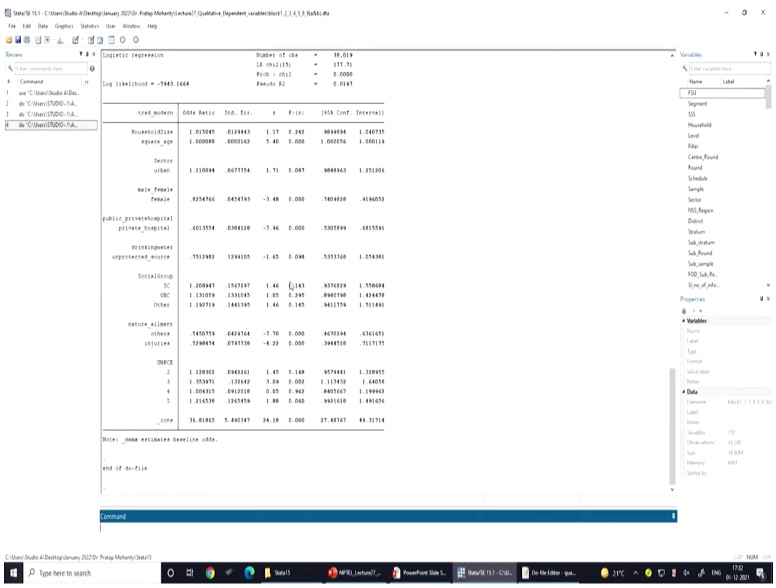


Now we are just running it over here the result is on your screen now this is derived it is here. We have derived the logit with it is conditional variables like for example, the conditional you can just read here sector on 2. One is rural and two is urban what is the interpretation of it the coefficient is given not the odd ratio you should note it very clearly. These are not odd ratio these are simply called coefficient and it is significance value is given as P values, P values column are there.

It says that the probability of the likelihood of the patients or the persons with element located in urban areas are less are more likely to access modern medicines. The 0 and 1 we have taken, the 0 and 1 in our dependent variable traditional versus modern.

Similarly, you can also follow others. Second one is female as compared to male, then public versus private etcetera rest of the details you can find from our paper their paper link we already shared to you earlier. Now, we are also trying to give logit estimation with it is simple this is simple logistic regression equation is it is you can also derive and find out here this is result is displayed.

(Refer Slide Time: 27:44)



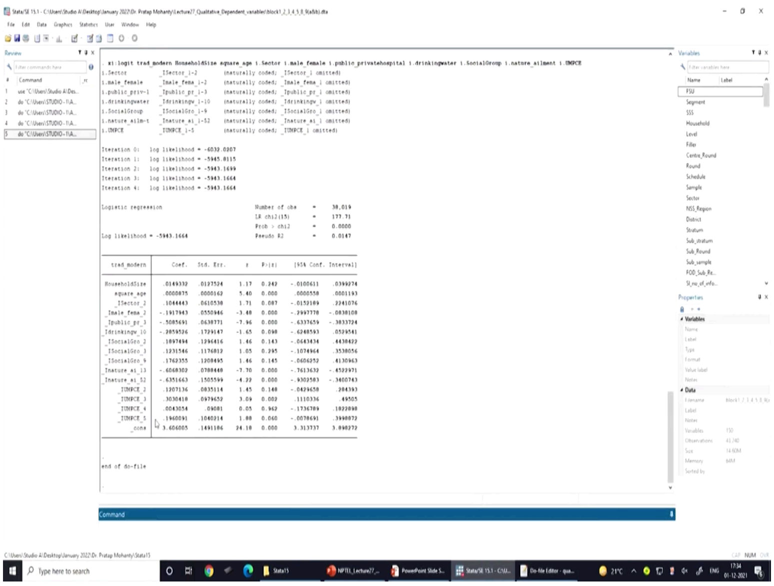
So, basically simple one it takes the average without having like look at the here. When you have let me tell you the command here logistic regression, when you give logit it gives you the coefficient, when you take logistic command it gives you the odd ratio.

So, here is the odd ratio on the first column of the result is your odd ratio. So, odd ratio if it is greater than 1; that means, your result is in favor if it is less than 1 it is against. So, like P and 1 minus P and whether they are significant or not you can interpret accordingly. So, then we will have margin command as well.

So, margin at it means after running the logit model when you run logit command you are getting coefficient not the odd ratio. But when you have applied the command with logistic you have got the odd ratio in case of odd ratio no margin effect is required.

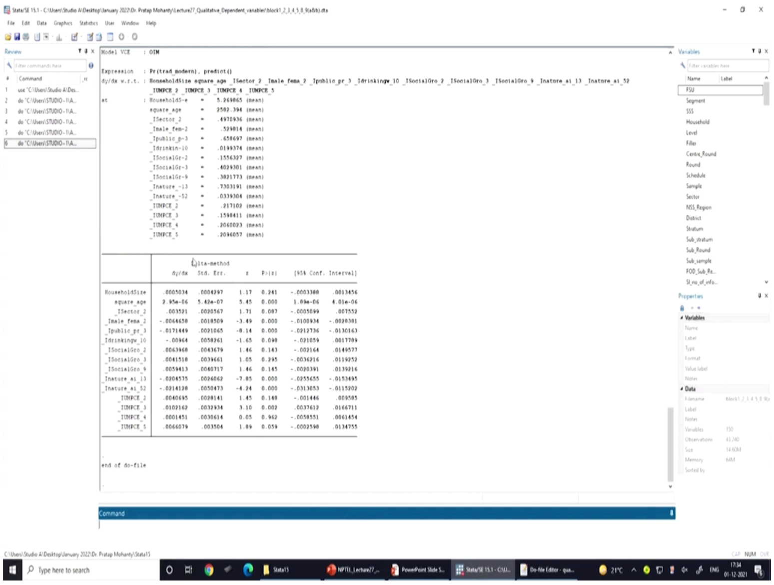
If you have got the value as coefficient you have to run the margin command that is has to be interpreted better. So, once again we will run the logit and show you followed by it is margin effect X i command. Once we have taken Xi logit. So, we are just operating it over here that I can show it to you how it works.

(Refer Slide Time: 29:22)



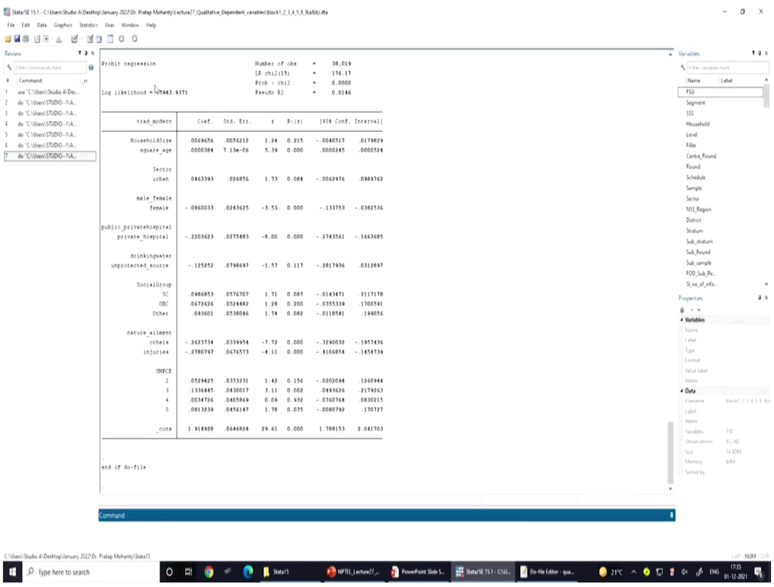
So, this is in fact, our coefficient values this is what is the logit command now followed by that we will run the margin command margin with dY by dX and at it means. So, it means we are deriving the result has already been derived.

(Refer Slide Time: 29:42)



So, this d Y by d X the first order derivative and it is value and its significance level is derived. but the interpretation are similar to that of the coefficient, but these gives most appropriate values than that of the logit coefficients. So, margin effects has to be derived alright. Then similarly you can run the probit model probit the probit coefficient is here probit and it is dependent and independent variable. Here are the results.

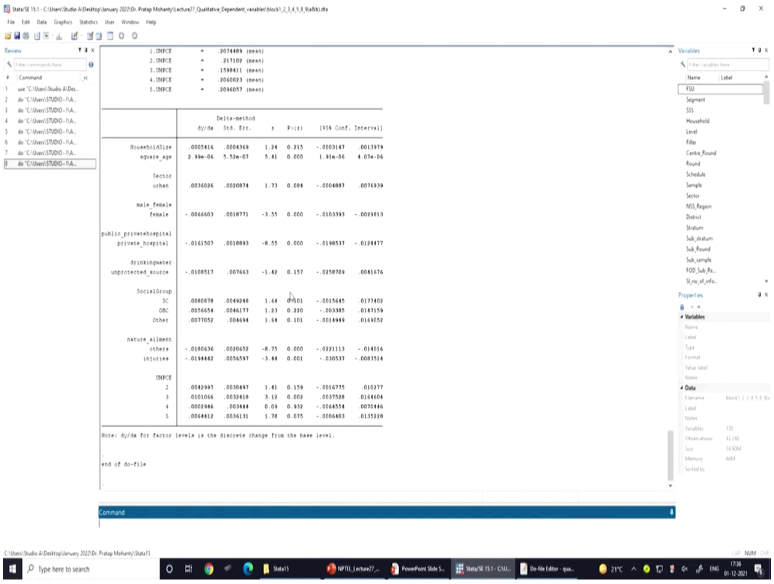
(Refer Slide Time: 30:14)



Now probit regression on the top is written the in the probit regression when your distribution is symmetric your distribution cumulative distribution function is following normal distribution or error distribution normal then probit is most applied.

And accordingly, we have derived the coefficient when you have got the coefficient I already suggested to you that you are supposed to run marginal margin probit as well. So, margins we are also deriving here alright when you have got the margin values then you can check over here. I think we have not, we will run it once again and the result is on your screen.

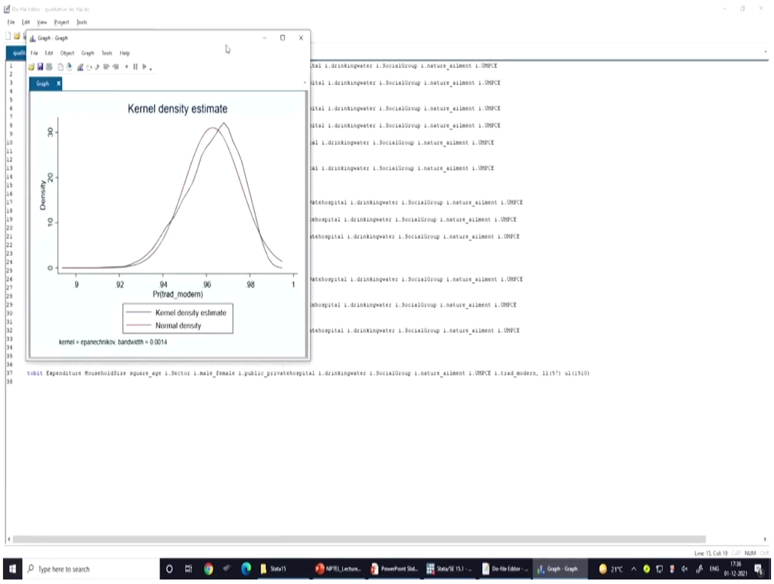
(Refer Slide Time: 31:01)



So, margin at means is already derived and another interesting part in probit is that you can predict since you have probabilities. So, prediction of the variables are possible you can predict for a different time. So, first we will draw the logit probit regression. Then we will predict then that prediction will give us whether the distribution is actually normal or not normal.

So, we run the probit once again then we will run the predictor. Predict of the residuals are stands for the residuals or the error term. Now, we will see whether they are actually distributed normally or not; k density function is going to give us result over here.

(Refer Slide Time: 31:47)

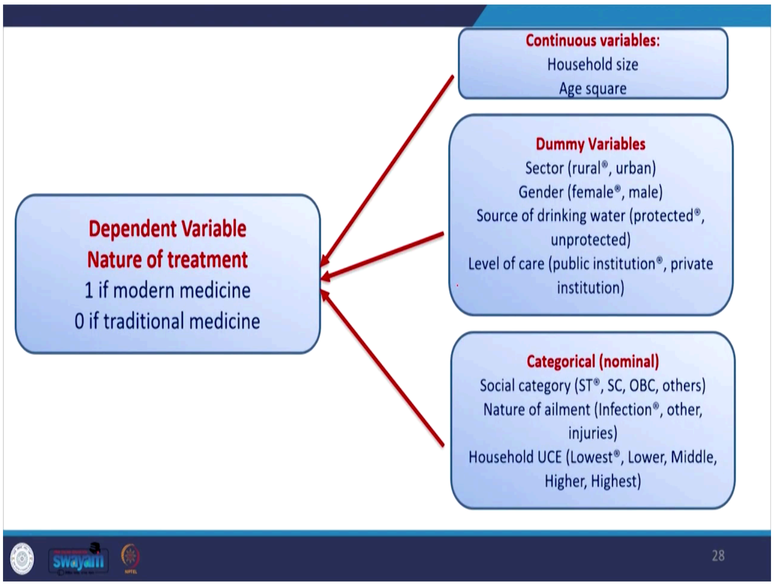


So, here this same session it is almost assimilating with the normal trend. So, in that case probit model could be applied. If it is much deviated and more asymmetric then in that case you need not apply a probit rather logit is more suitable.

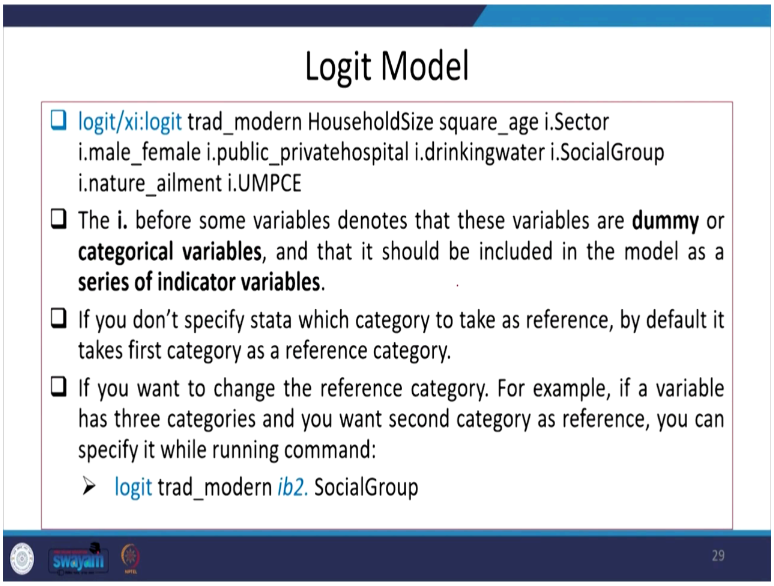
So, other aspects are like how to compare all those things like comparing logit, then probit, then your LPM together. How these values which we are already shown what proportion of logit coefficient is comparable to the coefficient of LPM. These all rest of the commands are actually going to help you because we can store all the commands together and compare their values.

So, at this moment I am not running I am just keeping it with you I am sure that you will be enjoying and running those things. And my if you are having some difficulties I will suggest that you refer my previous NPTEL module that is still this time it is also being floated that is on handling large scale large handling large scale data with stata. The videos are available in YouTube as well all the videos with that title you can search and find out the details. So, I am not going to explain all those things those are care for your interest.

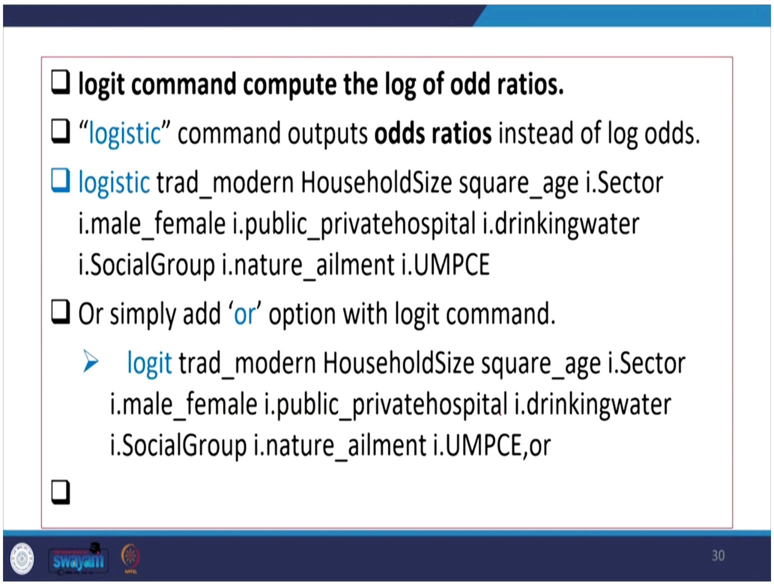
(Refer Slide Time: 33:15)



(Refer Slide Time: 33:17)

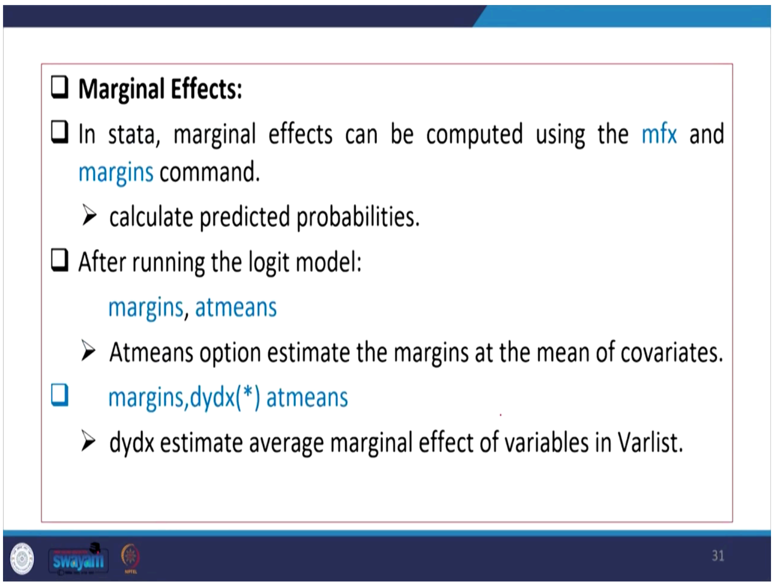


(Refer Slide Time: 33:24)

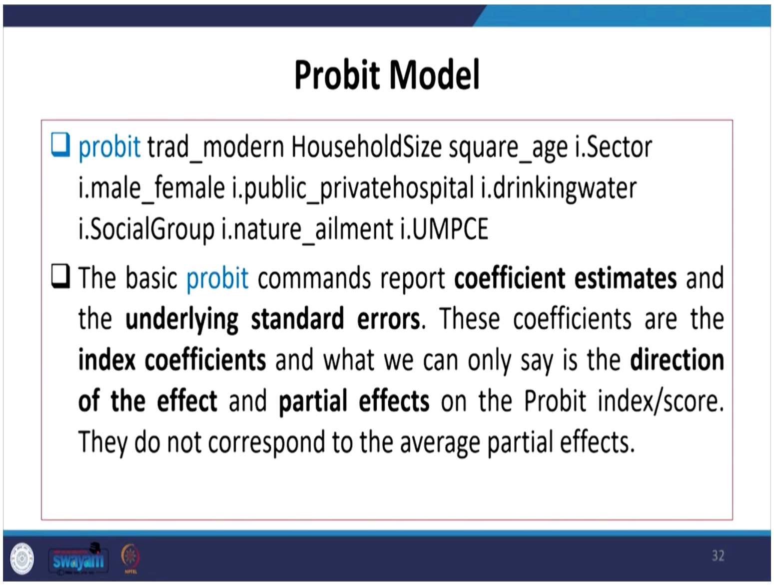


So, we have all explained all those things including mixture of qualitative and quantitative variables. Logit model with these commands we have already discussed. I am not discussing you need to go through on your own.

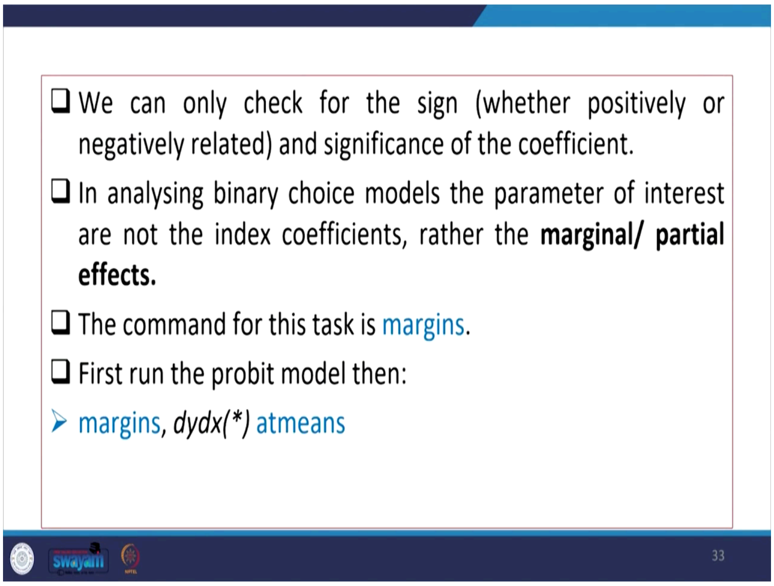
(Refer Slide Time: 33:28)



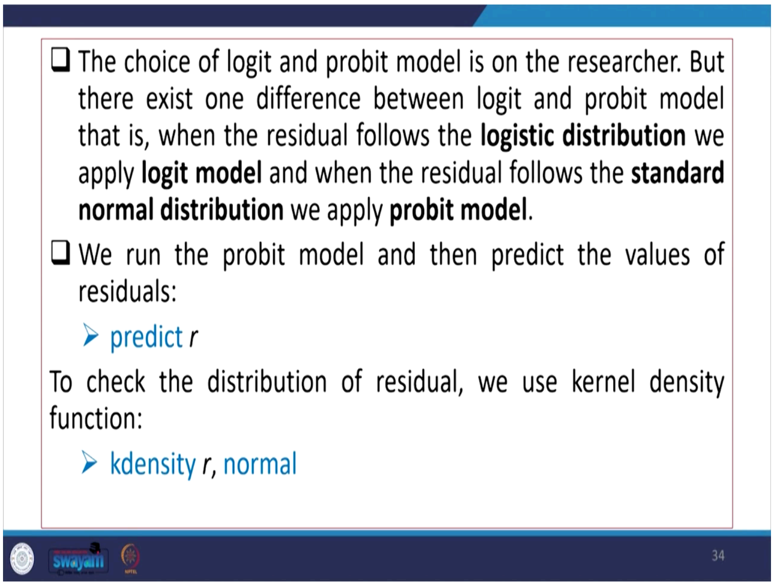
(Refer Slide Time: 33:31)



(Refer Slide Time: 33:33)

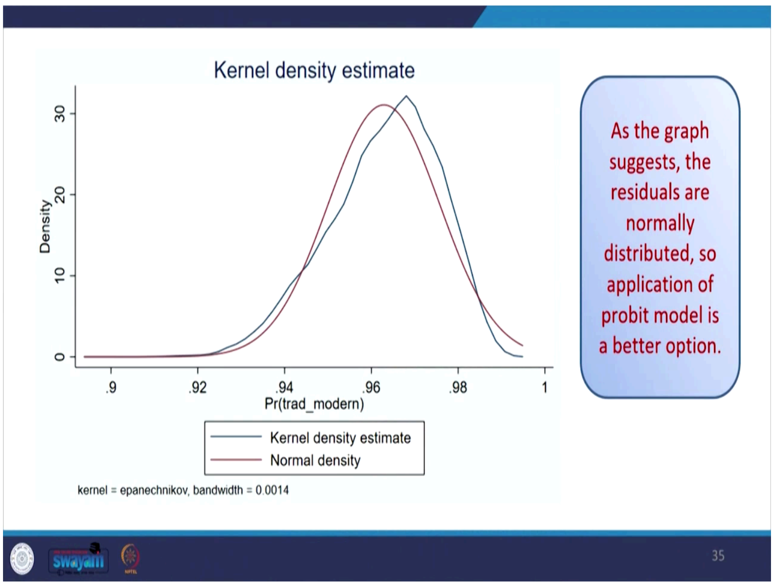


(Refer Slide Time: 33:35)



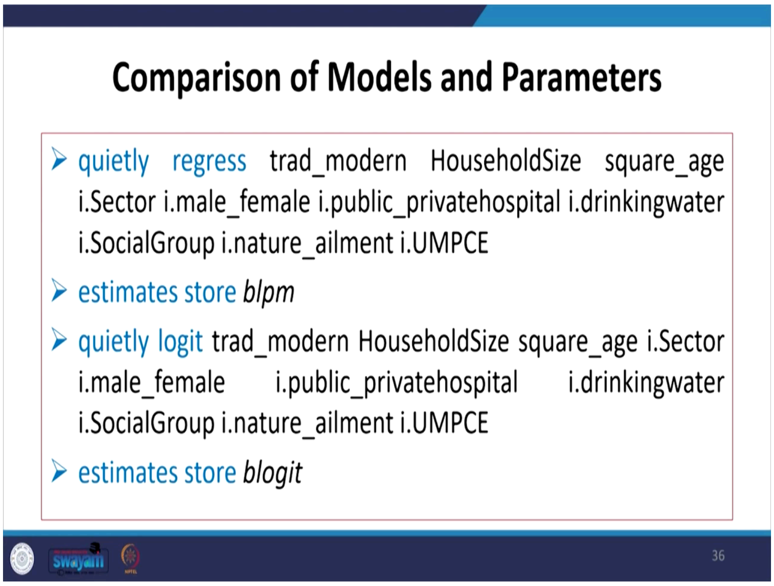
I am deliberately skipping marginal effects and all their details probit and their marginal effects I have already explained. So, no need to explain it further we have already run the command called predict r. Then we explained about k density function and what is the necessity of k density function I already discussed.

(Refer Slide Time: 33:47)



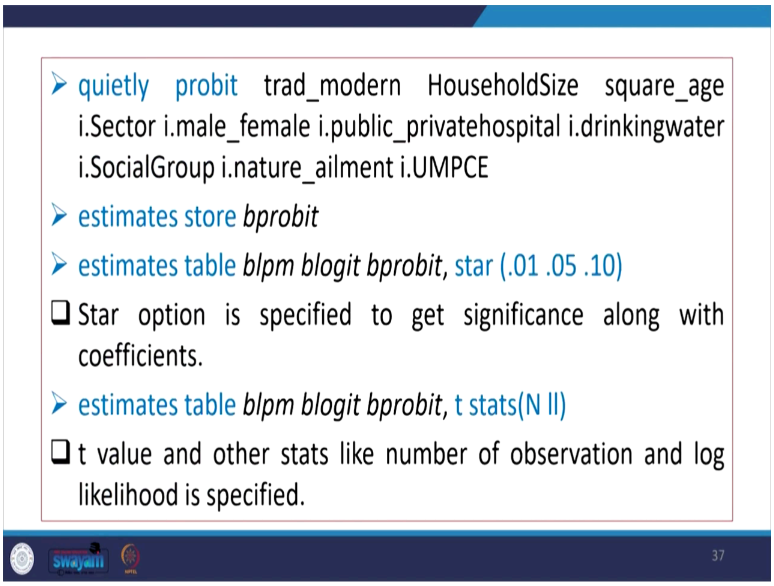
This is what the result looked like and I think you must have understood.

(Refer Slide Time: 33:52)

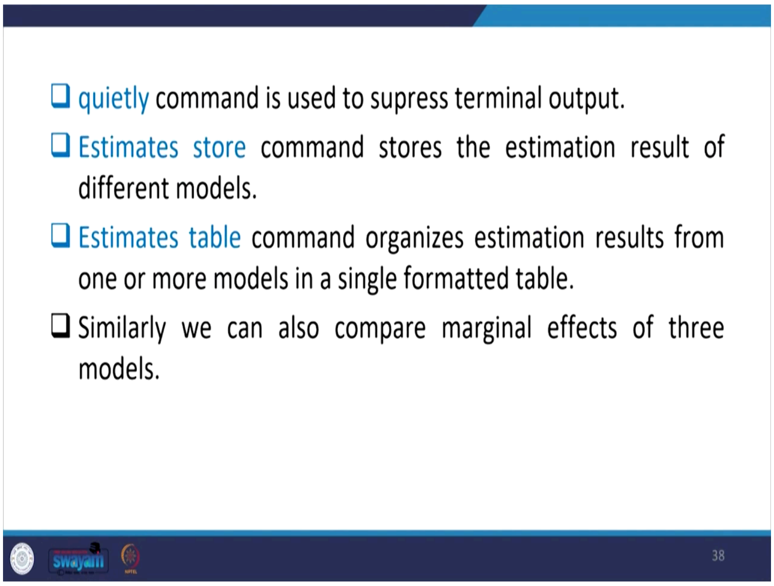


Now, how to make comparison of all those models and their parameters. these are the command we have also given step by step and I am quite sure with these command you can able to estimate and compare.

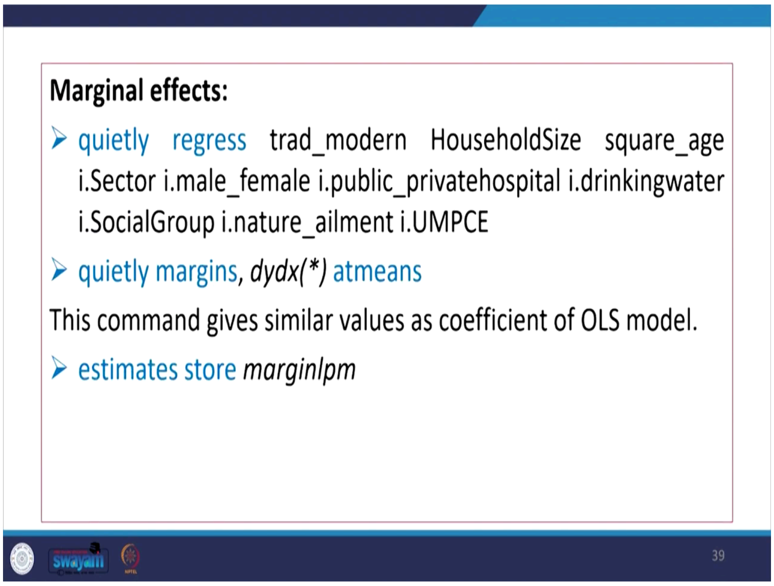
(Refer Slide Time: 34:01)



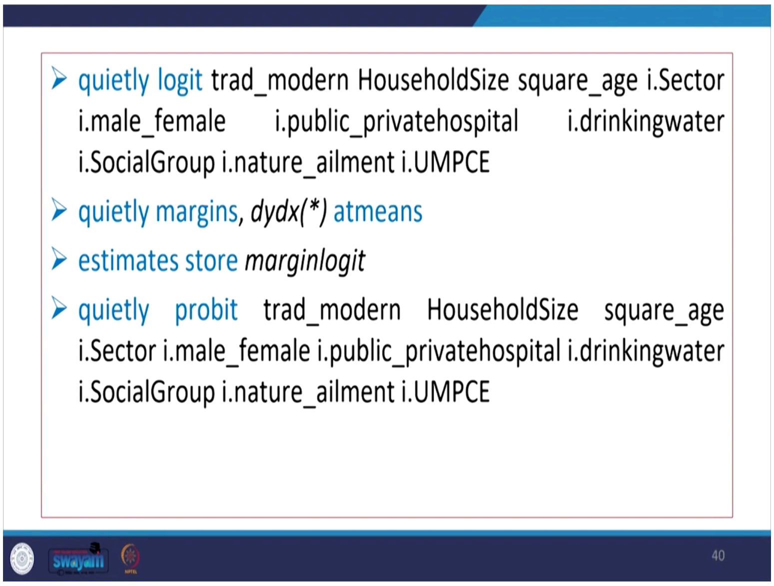
(Refer Slide Time: 34:04)



(Refer Slide Time: 34:09)

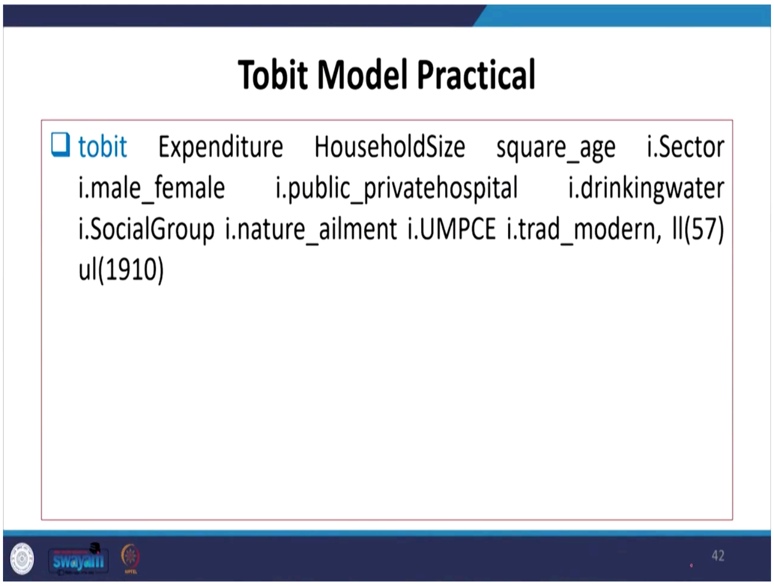


(Refer Slide Time: 34:11)



If you have still difficulties my videos are already floated in YouTube. I am sure you can able to understand very clearly at this moment I am not repeating. Because this is simply consuming time and with this less span of time you guys can able to get all those details.

(Refer Slide Time: 34:27)



Then comes tobit, tobit model is applied when we have the data that is having continuous data the dependent variable is continuous. But those who are not following normal distribution or not properly normal in the sense it is not symmetric.

When it is not symmetric either it will be positively skewed or negatively skewed. That means, there are some what is called data at the extreme points, maybe at the lower extreme, at the upper extreme. So, in that case tobit is taken instead of OLS if you run the OLS the since distribution is not normal OLS it is not applied.

Your expenditure is the variable others we are not going to care for it. So, it simply you take tobit and run the regression I am sure you will get the result. So, like some other details like tobit sensor how it is different, how it is truncated look from the lower side or from the upper side.

You have to read it from my previous lecture, previous module and I am not emphasizing because we have already exceeded the time of this particular lecture. So, with this I so, stop here I am sure you will understand and apply in your day to day research work and develop research paper. If you have difficulties do not hesitate and raise your queries we will be most happy to address it.

Thank you.