Applied Statistics and Econometrics
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology, Kanpur
Dummy Dependent Variable Models (Part 1)

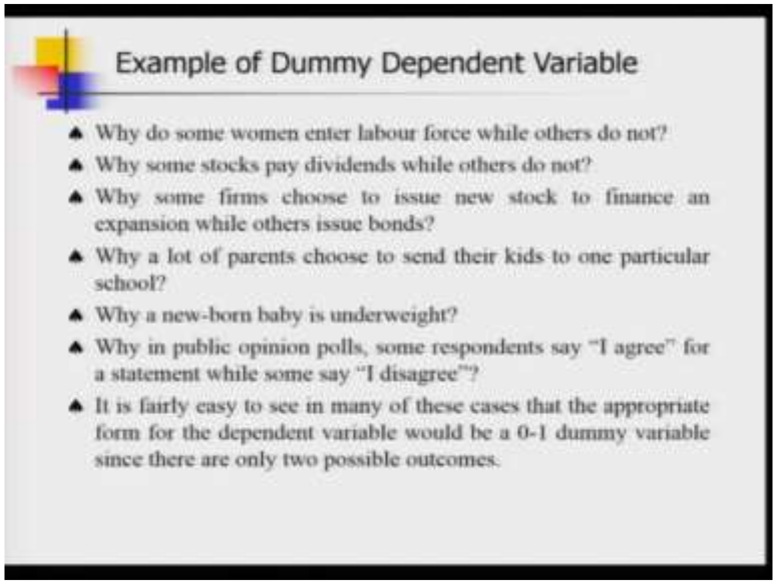
Hello friends. Welcome back to the lecture series on Applied Statistics and Econometrics. So in today's lecture we are going to start discussion on advanced topics in econometrics, and that is called limited dependent variables. Now again this limited dependent variable or it is also called qualitative dependent variable, this type of modeling is a vast area and we do not have enough time to talk about all sort of models which are available in the literature.

But I am planning to take you through some basic models which are like the preliminary fundamental models in this area and then if you are interested you can develop your knowledge base further by consulting different econometrics text books.

So before we go and have a look at the models let us have today's agenda items. So here I am going to first motivate the audience by providing some examples of dummy dependent variable and then I am going to talk about out two types of qualitative dependent variable or dummy dependent variable model, namely linear probability model and probit model.

So far whatever we have discussed in regression analysis we have always assumed that my dependent variable Y is a continuous variable and in the set of explanatory variables there could be a dummy variable which is binary in nature taking 0 and 1 values. Now there could be many, many situations around us where you see that you do not have a continuous variable to model and you are going to model some observations or behavior of human beings around you where the values for that variable is discrete in nature.

(Refer Slide Time: 02:24)



So I am going to show you some examples in the previous slide. Of course the list of examples is not exhaustive. So now I am going to discuss, well not discuss, I am going to briefly expose you to certain situations where a dummy dependent variable is the way to go and continuous dependent variables are just not possible to find or model.

So let us start with this question which is very popular about the labor economists. So why do some women enter labor force while others do not? So clearly it is a choice question. So when woman faces a choice between joining and not to join the labor market the woman chooses one option. And if she chooses to work then she will definitely join the labor force.

So let us move to the second question where we say some stocks are paying dividends and while other company stocks are not paying dividends? So paying dividend is definitely company's discretion. So basically again the company is making a choice whether to pay dividends or not.

Now we move on to third question which is again coming from the field of financial economics. We see that when firms require money to expand their business they collect money from the market either by issuing new stock or by issuing bonds. So if some firm is choosing for stock over the bonds then why they are choosing so?

Now let us move to a different field, say education. So many times we see that a lot of parents choose to send the children to one particular school in a locality. There could be 2-3 more schools but one school is certainly popular as more parents are choosing for that school. Now what are the determining factors that are guiding the parents to choose one particular school over the others?

Now let us talk about different fields, say health economics. So here we are concerned if a newborn baby is underweight or not. Now there could be some factors which are affecting a newborn baby's birth weight. So what are those factors? We want to certainly know that.

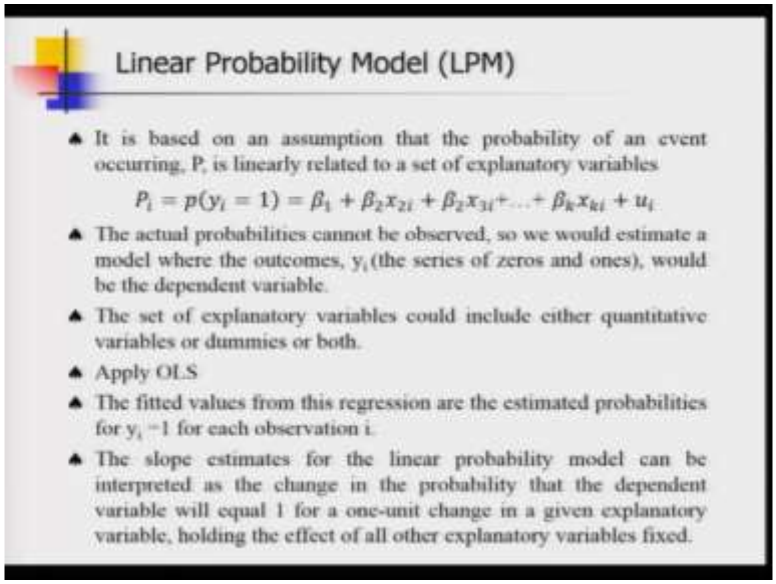
And then last we are going to take an example from public opinion polls. Many times in applied research you see that surveyors are asking statement type questions and then they are asking whether the respondent agrees to these statements or not. And the respondent says yes, I do agree. Or the respondent says no, I do not agree, or I disagree.

So you see here in all the cases that we have talked about so far in this lecture are going to be indicating towards one single thing. And that is know the Y, the variable that you are attempting to model, it cannot take continuous values. So most likely they are going to be qualitative in nature so that you cannot have numbers for those variable.

So there could be labels. And when you are dealing with qualitative variables and there are labels then how do you really take them to regression. So we already know that if you have qualitative information in regression setup you can actually adopt the dummy variable technique and define dummy variables. So here also if you see there is a qualitative variable to explain then you can define dummy variables for its labels.

So as it is an elementary course I am not interested to bring sophisticated discrete choice models where there are many, many labels for one qualitative or attribute variable. So here in this lecture we are going to restricted ourselves to qualitative explanatory variables for which we have two labels or two attribute values.

(Refer Slide Time: 06:56)



Now we are going to start with linear probability model, the simplest possible regression model which can handle this kind of qualitative explanatory variables. So here the idea is that we start with an assumption that we are going to model the probability, we are not going to model the event as such. And then we write one regression equation wherein we are modeling probability of occurrence of an event and there is set of explanatory variables as I am showing here. Now note that here I can link this topic to what we have learnt before.

So remember those cases of, like Bernoulli trials and binomial distributions and all there also we were talking about probability of success and probability of failure. So if you are conducting a random experiment there could be either a success or there could be failure. These two outcomes are possible. So from there we can actually draw a link. So these two areas or concepts are kind-of related.

So here also you see your Y as I am restricting myself to qualitative dependent variables taking only two label values. So there could be two possibilities. So basically either the unit on which you have data has chosen for option number 1 or label value 1, or option value 2 or label value 2. So you can attach success failure story from the part 1 of this course to this dummy variable, dummy explanatory variable case as well.

So basically what we are going to say that these labels that you are seeing for this qualitative explanatory variables some probabilities associated with them such that one individual cross-sectional unit will choose that particular value or label or option whatever you want to call it over the other.

So here we see the role of probability. But note that, this is very interesting here, although we are aiming to model probability but actual probabilities cannot be observed. So we can only see whether person or a cross-sectional unit, say a firm has chosen a particular label or a particular option, out of two options or two label values of qualitative dependent variable.

So we observe the final outcome and this final outcome by y variable, so y i actually is a series of 0s and 1s for observations in the sample. So if you have a sample size of n then you may have n1 number of observations showing the value of the dependent variable y to be 0 and then n2 number of observations which is basically capital N minus n1 observations are there which are showing the value 1 for your dependent variable y.

But the set of explanatory variables can take either continuous variables or the discrete dummy variables or both. There could be interactions between the continuous variable and the dummy variable. So suppose theoretically speaking, no problem. We have this regression equation in mind and we do not observe the values for the probabilities although we want to model probability. So we are desperate. We want to say that, no we want to run a regression.

So what will be my y? So what will be my y? So I am going to use the dummy qualitative variable as the y. So now y we will take only two values, 0 and 1. So now what we are going to do? You forcefully apply OLS. You do not know the consequences of applying OLS but as this is the only tool as of now you know you apply OLS. So after you run the OLS what will you get? You will get the regression coefficient estimates.

So you will get all this, intercept and slope parameter estimates and you also know the values of xs, the explanatory variables for each individual observation in your sample. So now you make use of whatever information you have. So the values of xs and the obtained regression coefficients and get the fitted value from the estimated regression equation.

Now what are these fitted values? So these fitted values are actually the probabilities. So this is basically modeling probability. So the fitted value will give you the predicted probability for an individual observation in the sample for specific set of values for the explanatory variables. So here we can look at the coefficient interpretation and try to link that with whatever we have learnt from the previous lessons when we discussed the case of OLS.

So what is regression coefficient measures? So it measures marginal effect. So if there is an one unit change in one of the explanatory variables how my y is going to change? So specifically how many units of change in y is expected to happen? So here what is the interpretation of the slope coefficient in this LPM context? So that is we are going to discuss now.

So as I said that although we are making use of y values 0 and 1 but my conceptual regression model is probability model, so basically the regression slope coefficient will tell us the change in the probability that the dependent variable y will take value 1 for a one unit change in a given explanatory variable.

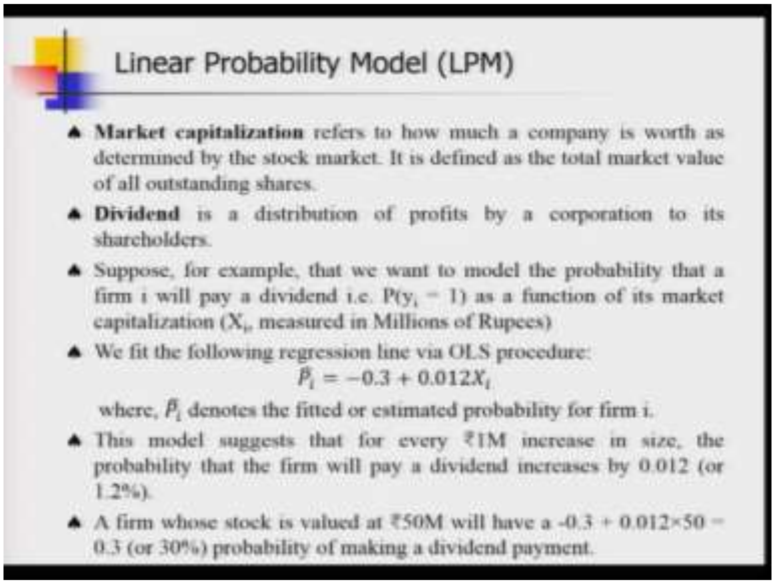
But do not forget that you have to keep all other explanatory variables fixed as I did for the OLS standard OLS case as well. So it is very important point that you must remember. So when you are talking about regression coefficients that is marginal effect. So the concept marginal is linked with calculus. So basically what do we mean by marginal? So basically we are talking about partial derivatives of the dependent variable in the multiple regression with respect to an explanatory variable of that same regression equation.

So basically there is a small incremental change in the explanatory variable then how much of y is going to change. But that is partial derivative. So to calculate the partial derivative or to measure this marginal effect we have to keep all other explanatory variables as fixed. So they are not going to change. Only one explanatory variable at a time is allowed to change when you are talking about the marginal effect.

So now we are going to continue our discussion with LPM again. So here I am going to give you an example from financial economics. So suppose we have got a data set and we can observe that there is a set of firms and some of them have paid dividend and some of them have not paid dividend in a particular year. And as an analyst or as a researcher you want to know why some firm is paying dividend and while others have opted for not paying a dividend.

So of course paying dividend is firm's choice but what could explain their behavior? Why some of them are paying and some of them are not paying? So you want to frame this problem in a regression setup. So you are looking for explanatory variables. What sort of explanatory variables you can bring in?

(Refer Slide Time: 16:19)



So there is a concept called market capitalization. So that basically talks about the value of the company or the firm in question. And we can assume that if we know the firm's market capitalization value then that may have any impact on firm's decision or firm's behavior. So suppose we have a very simple model where we have only one explanatory variable and that is market capitalization.

So we have talked about market capitalization as an explanatory variable predicting or determining the values of y is which is basically 0 and 1 in this case. 1 is basically that the firm paid dividend and the value 0 for dependent variable y means that the firm did not pay dividend. So before we move forward it is not a bad idea to wait for a minute and then have a clear understanding what we mean by market capitalization and dividend, because some of you may not come from financial economics background.

So those of you who have come from social science background then these two terms may not be very clear to you. So I am going to first provide a definition for these two concepts and then I am going to continue with econometric analysis. So formally speaking market capitalization can be defined as how much a company is worth as determined by the stock market. So it is basically the total market value of all outstanding shares of the company. Now we move on to dividend. Dividend is a distribution of profits by a company to its shareholders.

So now let us continue with our econometric analysis. So suppose we want to model the probability that firm i will pay a dividend. So the probability of y i taking value 1 that is what we are going to model. And it is a function of only one explanatory variable, and that is the market capitalization value of the firm, and that is measured in millions of rupees.

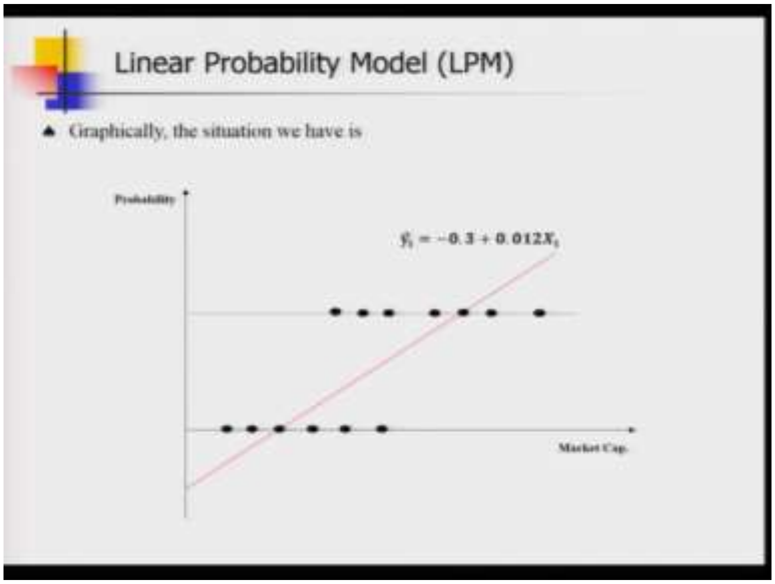
But note that this is a very hypothetical case that we are going to discuss. So there could be 5 more explanatory variables which can affect a firm's decision to pay dividend or not to pay a dividend. But let us make the story simple. So although we are risking ourselves with this omitted variable bias and all that I am sure that you are probably thinking in mind that this is such a simple model that probably it will be affected be plagued by omitted variable bias.

But just to show how LPM works and what are the advantages or disadvantages whatever you want to call are there, so for that purpose two-variable simple linear equation model is good enough. So hence I am restricting myself to one explanatory variable. So suppose I assume that the classical linear regression model assumptions are holding and I apply OLS to fit the regression line which actually is given as follows.

So P hat i is the fitted or estimated probability for firm i paying dividend. And that is equal to minus 0.3 plus 0.012 times x i. So now what does this model tell us? So this model suggests that for every 1 million rupee increase in the size of the company which is the market capitalization value, the probability that the firm will pay a dividend increases by 0.01 to, or you can express it in percentage terms and you can say that this is 1.2 percentage.

Now we take another value of x and this time we take a very high value of x, say 50 million. So here also we can compute the predicted or fitted or estimated probability and it shows that in that case the probability is 0.3 so there is a 30 percent chance that the company worth rupees 50 million is going to pay a dividend.

(Refer Slide Time: 20:41)



Now at this point I want to express this problem or illustrate this problem through a graph because then the lacunas, the loopholes of the simple linear equation model will be much more clear to you. So in this slide I am trying to explain the same story with the help of a graph. So here you see I am plotting the explanatory variable market capitalization along the horizontal axis and the probability along the vertical axis.

And you note that there is a broken line in this two-dimensional plane and that broken line is drawn parallel to the horizontal axis at the value 1. So now note one interesting thing. So we observe only two values of y which are 0 and 1. We are not going to observe any value on probability. Probability is not available as a data to us. We can maximum get fitted values.

So what happens if you plot this dataset that you currently have, you have basically one y that is showing numbers like 0 1, 1 0, 0 0 1, 1 1 and something like that. So it is series of 0s and 1. And then you have some continuous numbers for your explanatory variable. So if you plot this data in this two dimensional plane you are going to see that your ys are going to be clustered.

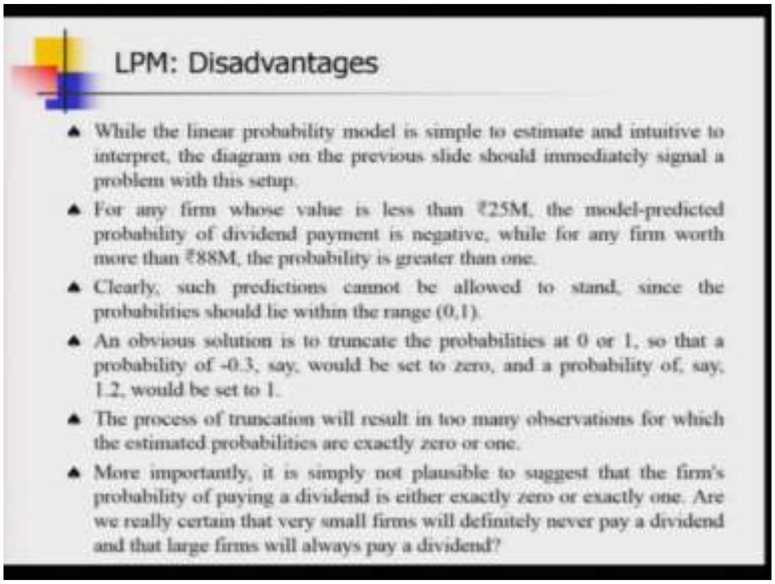
And there will be one cluster that you will found at the horizontal axis and there will be another cluster of points that you will find along this broken line which is drawn at probability value equals to 1. So basically you are observing this scatter plot. Now think about fitting this regression line through this kind of scatter plot because the scatter plots are such that they are concentrated at two different values of y or probability of y and y equal to 1 and that is basically 0 and 1. So this is a very peculiar case.

Now if you want to fit the regression equation through the scatter plot. So of course the regression equation will pass through the scatter plot and it will present you some values of intercept and slope parameter estimates. But there are many problems with these estimates. So let us have look at what sort of problems we can encounter. So here you see we have plotted this fitted line which is y hat equals to minus 0.3 Plus 0.012 x.

So note that this straight line is a positively sloped straight line with a negative intercept. Everything is fine here up to this point. But now the trouble arises if you look at the two ends of the straight line. So if market capitalization value is above a particular threshold value then this fitted straight line will not lie within these two horizontal lines which are bounded by 1 and 0. So it is going to in fact move up towards the north east direction and the fitted values are going to be higher than 1. Now you concentrate on the other end of the fitted straight line.

So if we are talking about very small market capitalization values, of course there will be a threshold, so if your market capitalization value is less than that particular threshold value then corresponding fitted probability is now going to be negative because you see the fitted straight line actually is passing through a quadrant where you have negative values. So basically here also there is a problem.

(Refer Slide Time: 24:47)



So now we are going to summarize what we have seen in that diagram in the previous slide in terms of words. So although the LCM model is very intuitive to interpret but interpretation is not everything. There are severe problems with this linear probability model. So here we can actually calculate the threshold values that I was talking about in the previous slide.

So now look at the threshold values and let us see what sort of problems are associated with these two threshold values. So for any firm whose market capitalization value is less than rupees 25 million the model predicted probability of dividend payment is actually negative. And if the firm is worth more than rupees 88 million then the predicted probability or the fitted probability that this firm will pay a dividend is actually greater than 1.

So you see we will only get a proper probability measure which will be bounded by 0 and 1 if my firm in the dataset has market capitalization value lying between 25 million rupees and 88 million rupees. But what is the guarantee that there will be all firms which will have the market capitalization value within this range? Of course, no guarantee. And also see that there is another problem. And the problem is also very funny.

So suppose you have one firm whose market capitalization value is rupees 100 million. So there is an obvious solution to this problem that comes to our mind that we can truncate the probabilities at 0 or 1. So if say the probability is negative then we can cap that at 0 so we will not allow the predicted probability to fall below 0. And if the fitted probability is say value 1.2 or some other number which is greater than 1 we can also cap that number at 1. So we are not going to allow the fitted probability to take a value greater than 1.

So do you think that this is a very good solution? No, it is not a very good solution because of two reasons. First of all if you cap the probability values artificially like truncation then you will have lots of points which will be clustered together. So there will be many firms which will have the same predicted or fitted probability value either 1 or equal to 0. That is not good.

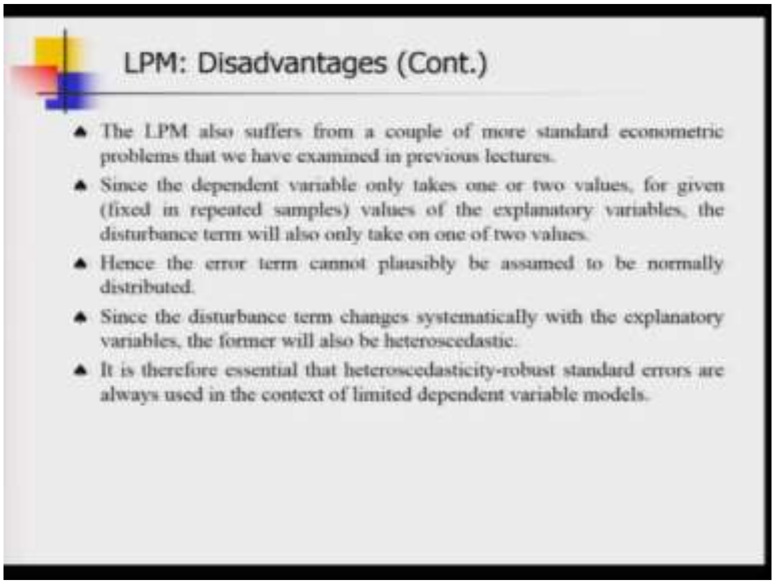
And the second is even more funny. So suppose you observe a firm with market capitalization value of rupees 90 million. So as per the regression equation we know that 88 million rupees is the threshold value. So any firm which is reporting higher capitalization value, for that firm the probability, fitted probability is going to be greater than 1.

So now you are kept it 1 as a solution. But then interpretation is absurd because then you are telling the audience or your manager or your supervisor that there is a certain, it is a certain case that this firm is going to pay a dividend because the fitted probability is 1. But who guarantees on earth that this firm is actually going to pay a dividend.

Similarly you take a case of a small firm, say 10 million rupees worth. And then your model says that as the market capitalization value for this particular firm is less than 25 million rupees, the threshold value the predicted probability value is negative. So I am capping it at 0. So then you are saying the audience that this firm is not going to pay a dividend.

But again how can you say that this firm is not going to pay a dividend? And you just cannot make this kind of certain statements because your model is completely mis-specifed. So basically we have to go for some developed or sophisticated models.

(Refer Slide Time: 29:47)



But before we land up doing that kind of sophisticated model let us also look at other problematic areas of linear probability model. So here in this slide I am going to show that LPM suffers from couple of more standard econometric problems that we have already discussed in previous lecture. So let us talk about them very briefly. So our dependent variable takes values 0 or 1 only.

So basically for a given fixed repeated sample values for explanatory variables the disturbance term is now going to take only two values. So suppose the actual observation is 0 and your fitted value is saying that you can say that this particular firm is certainly going to pay a dividend. So you say my fitted value of y is 1. So there is one difference.

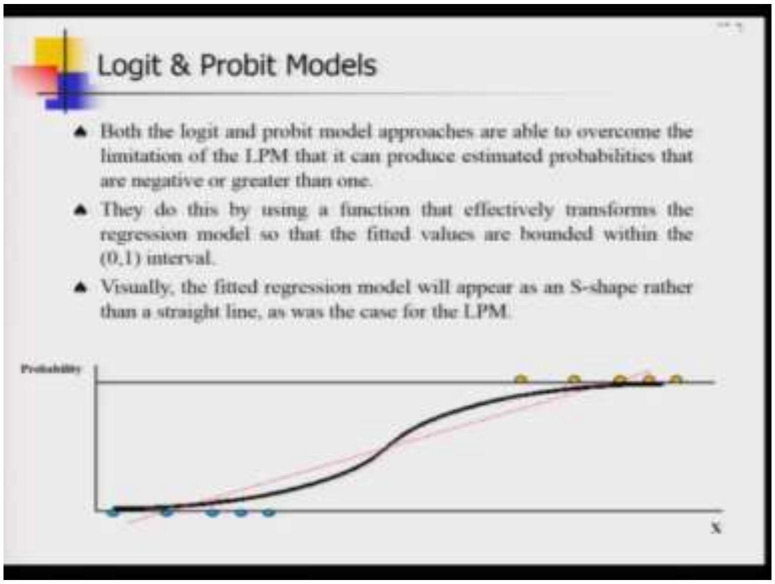
Similarly you can calculate the other type of residual values that are possible and you see the residual will not now take a range of values. It is going to take only two values as you are dealing with a discrete dependent variable. So it is not at all rational to assume that error will follow a normal distribution, because if a variable, if a random variable takes only 2 or 3 values then you canot say that it is a continuous variable and the values are going to be in a interval, any number can come.

So basically we are now talking about discrete distributions. So here is a problem. You cannot assume normality for your error distribution and if you cannot assume normality for your error distribution then what is the problem? I mean it will not stop you from executing OLS but definitely it will hurt you because you cannot go for hypothesis testing because we have shown earlier that if you want to conduct hypothesis testing after estimating the regression coefficients then you have to assume normality.

Also one can show that the disturbance term changes systematically with the explanatory variable. So the error term will also be heteroscedastic. So of course one has to take care of this heteroscedasticity factor. So what we have discussed earlier one can go for this White's approach which offers you the heteroscedasticity robust standard errors.

So by now we have understood that although LPM is very simple intuitive it is not going to be of any use to us because you are working with the completely mis-specified model. So all the assumptions that you are making to develop this classical linear regression model, everything will disappear and most of the assumptions are not going to be made. So then what is the point of running such weak regression which is of no use?

(Refer Slide Time: 33:52)



So what is the solution? I mean you can criticize particular model but then you have to come up with some solution which is a better model. So fortunately we also have some better models which are capable of handling discrete dependent variables or dummy dependent variables. And two most important models are called Logit and Probit.

So in this lecture I am going to briefly e discuss about Probit and I save the discussion for Logit for the next lecture. So let us have a look at the Logit and Probit models together philosophically. I mean what actually both of them are trying to do.

So in this slide we are going to talk about the philosophy behind Logit and Probit models. So note that we are very peculiar problem. So concentrate on the diagram first. So look at the diagram presented at the bottom of the slide and it is basically the same diagram that we have seen earlier.

So here, of course, if you remember we are measuring the probability along the vertical axis and the value of the explanatory variable along the horizontal axis and here first of all you look at the scatter. So basically the blue circles that you see on x axis, they are basically the firms which have not paid the dividend. So the value of y for these firms, they are same, they are all equal to 0. So note that x is changing but y is not changing.

Now you note that there is another cluster of data points on the line horizontal of this horizontal axis and this line is drawn at the value 1 on the probability axis. So here you see there are some golden dots and these golden dots are basically representing those firms which have decided to pay the dividend.

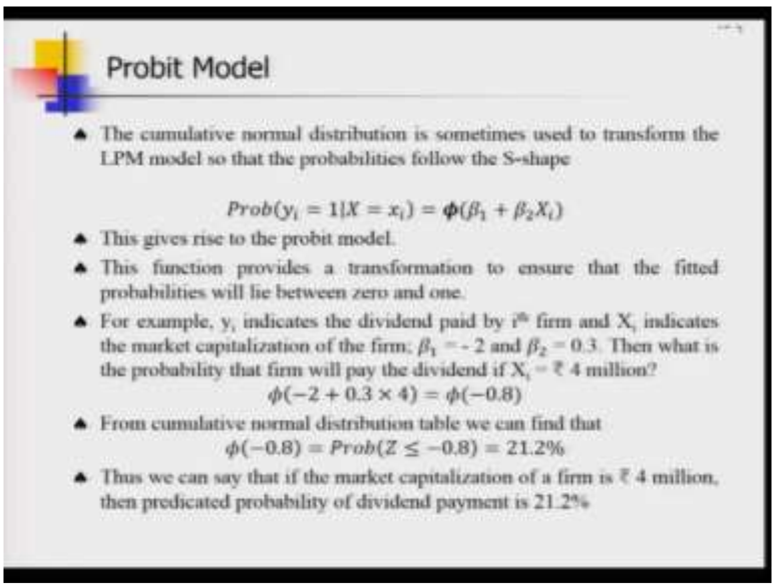
So now you have these data points, ten of them and you want to work with LPM, you have decided. So of course you are going to estimate a line. Nobody can stop you because you have 10 data points. You are going to estimate only 2 regression parameters alpha and beta intercept and slope parameters. So you are perfectly fine. You fit a line and suppose this hypothetical fitted line is shown as the red line here.

Now you note that this line is actually yielding probability numbers which are negative for some values of x and greater than 1 for some values of x. So basically you have to now come up with some curve which is going to be bounded by these two horizontal lines at 0 and 1. So you are looking for a curve which is not going to cross these two lines in either directions such that the fitted values that you are going to get out of your regression model they are going to be all fractional numbers.

So basically what are the solutions at hand? So we know from our statistics knowledge that there are some distributions and there are some transformations. If we can make use of those distributions and transformations then perhaps we can generate this kind of S-shaped curve which will be completely within these two horizontal lines. And this S-shaped curve is not going to cross these horizontal lines for any values of x.

So statisticians say that we should go for a functional transformation. We cannot just take that simple linear regression model. So basically after the transformation the fitted value should be bounded between 0 and 1 values. So here I am showing you that S-shaped function that you can see plotted in black color. This is basically the philosophy behind Logit and Probit models.

(Refer Slide Time: 38:23)



So now we are going to talk about the Probit model first. Why Probit model first, because here you are going to see we are going to make use of one distribution which will lead to a transformation which will give that S-shaped curve. And we all are aware of normal distribution. We have worked with normal distribution. It is a familiar distribution to you. So that is why I have chosen the Probit model to start the discussion. And in Logit we will come next lecture because for that we have to learn new distribution.

So now we are going to start talking about the Probit model and here we are going to see the the use of concept of cumulative normal distribution which is denoted by capital phi symbol or notation in statistics textbooks. And this is used to transform the linear probability model so that the probabilities follow the S-shape.

So now you concentrate on the equation that you see at the end of the first bullet point on the slide. So here you see there is no change in the left hand side. So we are going to model the same y with which we have started. So we are still modeling the probability that y is taking value 1 for given x value for certain individuals. So we are still modeling the probability of paying dividend by one firm if you stick to the example.

And then you see in the right hand side there is a huge change. So here you see that I have that linear equation beta 1 plus beta 2 times x i there. But I have a small phi symbol in front of it and there is no stochastic random disturbance term associated with that. So basically what we do?

So we are going to make use of a link function. So this is the concept of generalized linear model. So when we did OLS in the first of this course then we are actually executing linear statistical models. But in statistics literature when we have non-linear models then there is a term and that is called generalized linear model or GLM. And in GLM, the y, the dependent variable is non-linearly related with the explanatory variables and linear parameterization may not be of great help. So your OLS is not good solution.

And that is what as one of the examples we have already studied this in this particular lecture only. And so statisticians propose that you have to go for some transformation so you convert this nonlinear relationship between your xs and y again in linear equation form so that you can still apply a linear regression model, because nonlinear regression is very complicated to handle.

Now there is this concept link function which is basically bridging this gap between dependent variable and the independent explanatory variables. So here you see the link function is this phi function which is basically the normal PDF. So as we are talking about the form of the population regression function or equation of course we do not expect a stochastic disturbance term to show up here.

So this equation gives me the probit model. So this function provides the transformation to ensure that the fitted probabilities will lie between 0 and 1 as the probabilities are now coming from the standard normal distribution. So in this course we are not very interested to discuss

Probit model at length because natural interpretation of the coefficients are not possible and hence I decided to skip the estimation part of Probit models.

And I save this estimation of this kind of discrete choice models for the next lecture. I will cover that but it in the context of Logit model. So for the time being you assume that there is some technique through which we got the coefficient values beta 1 and beta 2. I told you couple of seconds before that no natural interpretation of slope coefficients are possible. But does it mean that they are of no use? No, it is not actually.

You can make use of the estimated coefficient values to predict probabilities of an event occurring. So let us go back to that same old story with which we have narrated the case of LPM and now you know we are at Probit stage. So we are talking about the financial economics case where we are actually trying to explain that some firms are paying dividends and some are not paying dividend.

So here you have that model where the coefficient estimates are basically beta 1 hat or B 1 taking value minus 2 and beta hat 2 or B 2 taking value 0.3. So these are all hypothetical numbers. I am just trying to show you how these estimated coefficients are going help you to compute probability. Now you are interested to find the probability that the firm will pay the dividend if x i equals to rupees 4 million.

So you see you plug the values in the Probit equation. So basically you know the coefficient values. You plug these values there inside the phi expression. So now you are going to find the value of phi of minus 0.8. I have shown you how to find these phi values. I hope that you still remember how to consult a normal table to find probability values. Now as this is negative number what you have to do?

You have to calculate the value that the standard normal variable Z value will be less than or equal to minus 0.08, the number that you got in the above expression and you see that you get a probability which will tell you in percentage terms. There is 21.2 percent chance that this particular firm will pay a dividend. So we stop our discussion on qualitative dependent or dummy dependent variables right here. And we will be back with the discussion on Logit models. So please join me for that lecture. Thank you so much. Bye.