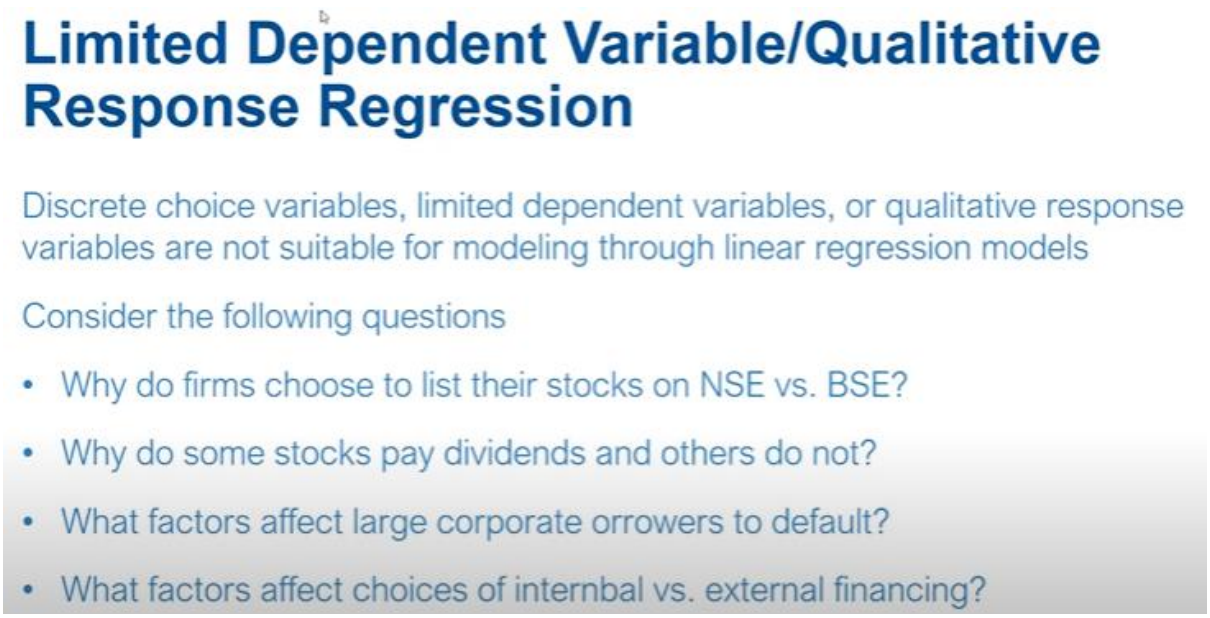
Artificial Intelligence (AI) for Investments
Prof. Abhinava Tripathi
Department of Industrial and Management Engineering
Indian Institute of Technology, Kanpur

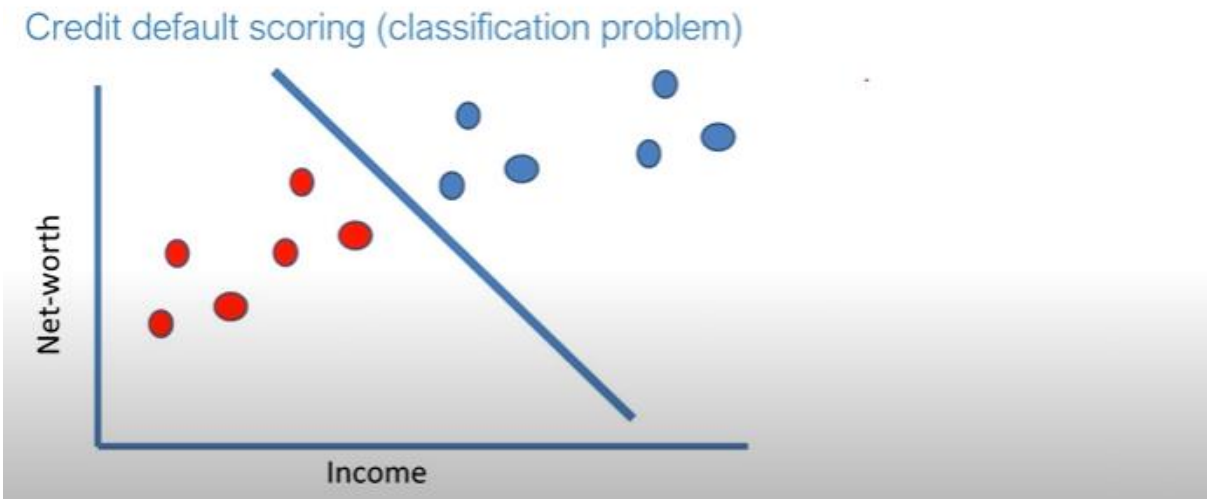
Lecture- 43

Linear regression modeling is less effective in the case where dependent variable is binary and of the form like yes, no or 0, 1 kind of variable. Such dependent variables are called limited dependent variables or binary choice variables. A simple OLS, ordinary least square regression model approach is referred to as linear probability modeling approach since the predicted variable is in the form of probabilities. However, these linear probability models are inadequate to satisfactorily model limited dependent variables such as 0, 1 or yes, no. These limited dependent variables are often modeled using logit probit class of models. We will discuss the logit probit class of probability distribution functions that provide theoretical underpinning to these classification algorithms.



Moreover, the predicted probabilities from logit probit models require thresholding to convert them into ones and zeros that are comparable to the observed data. We will also discuss a very important and comprehensive parameter of model performance that is classification confusion matrix. Also we will visualize the model performance under receiver operator characteristic curve and compute the area under the curve to examine the model performance. Lastly, we will interpret the parameters from the logit probit model estimation.

Limited Dependent Variable/Qualitative Response Regression



In this video, we will introduce the classification problem with a background to discrete choice variables or qualitative response variables. Discrete choice variables or limited dependent variables or qualitative response variables as they are called are like 1-0 or yes-no kind of variables. It is very difficult to model these variables through linear regression models as we have seen. Let us examine the following questions. Why do firms choose to list their stocks on NSC or BSC? Or why do some stocks pay dividends while others do not? Or what factors affect large corporate borrowers to default or not default? What factors affect choices of internal versus external financing? Notice that in all these questions, the dependent variable, the fact that is predicted or determined is in the form of yes-no or 1-0.

That means there are two choices and these choices can be modeled in a regression model through yes-no or 1-0 kind of format. For example, one can put the choice of listing a stock on NSC as 1 and BSC as 0 or paying dividend as 1 not paying as 0 or default as 1, non-default as 0. So these choices can be modeled in a linear regression model mathematically by coding them as 1 or 0. Let us understand this through a simple example of a credit scoring problem in banking. In banking, there are often loan applications that default or not default.

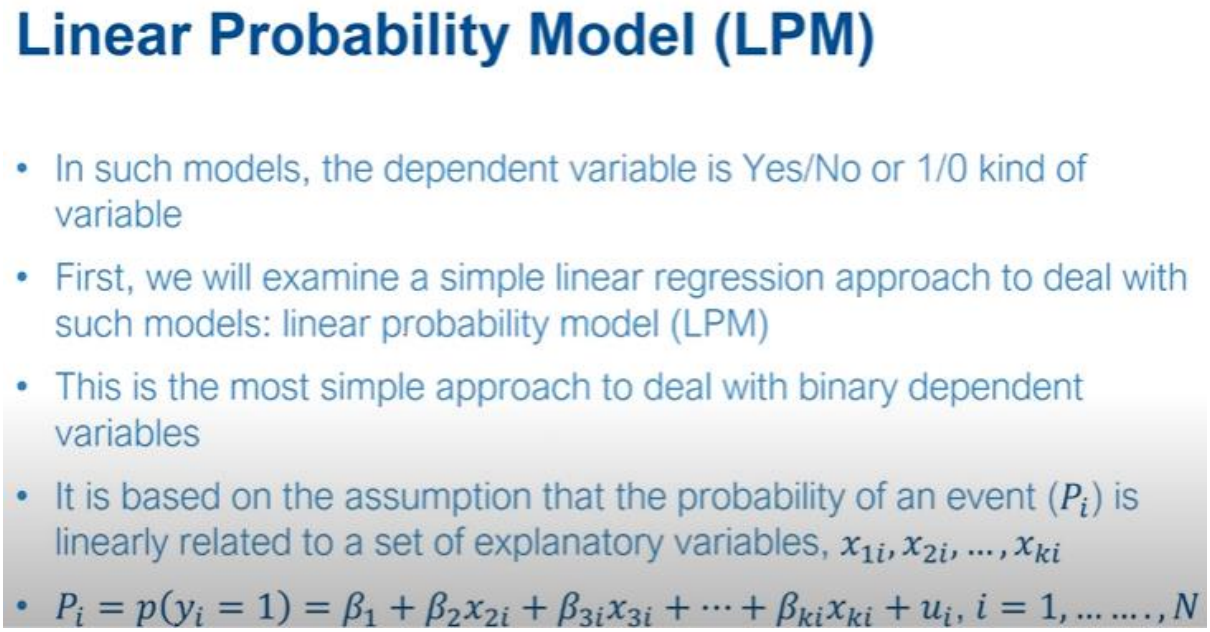
If you want to know or judge the credit worthiness of a particular application, there are certain attributes that you would like to examine. For example, you would like to examine net worth of the borrower or his or her income. Now suppose you want to train a credit default scoring algorithm or a classification algorithm, you would like to have a label data, the data being labeled as defaulted. In this case, the red, these red dots or circles, solid circles are defaulted, let us call them defaulted and the blue ones are non-defaulted. So you have the label data and you also have their features as net worth and income.

Now using these labels and these features, you would like to train your algorithm. There are

different ways to do it. For example, if you are given only net worth, you would have trained your algorithm at different levels based on the level of net worth and tried to see which of these levels classifies or categorizes the data best between red and blue observations. Similarly, if you are given only the income feature, then you would have tried different levels of income to classify data according to its labels and the most efficient way, in the most efficient way possible. You obviously would want to use both of these features and then classify.

For example, in this case, this kind of algorithm appears to be more suitable and therefore in future if some data is available to you, which where the point is here, then you would like to classify it as defaulted. And if it is available here, then you would classify it as non-defaulted. To summarize this question, we noted that the fundamental aspect behind the classification problem is the presence of label data, where data is labeled into two or more categories. For example, here we saw defaulted or non-defaulted. So, data is labeled into two categories and you are given certain features of data.

For example, in our case, net worth and income were some of the features. Features can be demographic aspect of individuals, tastes, preferences, various kind of features can be given to us along with their labels. Now, a classification algorithm would be some kind of algorithm that using these features would try to label the data accurately, higher the efficiency of the algorithm, that means its ability to classify accurately the data into its labels correctly, the better its efficiency. In this video, we will introduce a simple approach to modeling discrete choice variables or limited response variables, which is called linear probability modeling, LPM. Recall that regarding such limited dependent variable modeling, we said that the dependent variable is of the form of yes-no or one-zero kind of variable.



We will start by modeling or introducing the simple linear regression approach to deal with such models and this is the simple linear regression approach in this context is called linear probability model. This is the most simple approach to deal with the binary dependent

variables. But as we will see in the next video, it is fraught with its own issues. But let us introduce this problem. It is based on the assumption that the probability of an event Pi is linearly related to a set of explanatory variables x1i, x2i and so on up till xi, xki.

So the problem formulation would look something like this where Pi, the dependent variable is essentially the probability of an event happening yi equal to 1. For example, if you are talking about bank default, default versus no default case that we discussed in the previous video, we assume that let us say if default is happening, then in that case the event is coded as 1. So the probability that default happens or the variable y takes value of 1, this Pi of yi being 1 is noted as Pi is modeled along the set of independent variables x2i, x3i and so on and their coefficients which reflect the impact of these variables as beta 1, beta 2, beta 3 and so on in an identical manner as we would model the simple linear regression problem. Please note that in such models, the actual probabilities cannot be observed. So the estimates which essentially we model in the form of probabilities have to be converted into 0s and 1s.

$$P_i = p(y_i = 1) = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_{ki} x_{ki} + u_i, \quad i = 1, \ldots, N$$

So this is one problem that you do not observe probabilities, what you observe is the event happening which is 1 or not happening which is 0. For example, consider the relationship between a company i and its ability to pay dividends. So whether a company i pays dividends, let us call it event y equal to 1 if dividend is paid and equal to 0 if it is not paid. Now let us model this event in the form of or its relationship with its market capitalization xi. So xi is here is the market capitalization of the firm which is captured by its this relationship is captured by this coefficient beta 2.

Linear Probability Model (LPM)

In such models, the actual probabilities cannot be observed, so your estimates (or dependent variables) would be 0s and 1s

- Consider the relationship between the size of a company "I" and its ability to pay dividends

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

where X_i= market capitalization of the firm, and Y_i=1 if the dividend is paid and 0 if the dividend is not paid.

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Here essentially what we are modeling when we write yi equal to beta 1 plus beta 2 xi plus mu i essentially we are modeling probabilities, but those probabilities will not be observed. What we observe is whether yi equal to 1 dividend has been paid or 0 which is dividend is not paid. Let us clarify this linear probability model in more detail. So what we just saw

that is yi equal to beta 1 plus beta 2 times xi plus error term is basically here in this model what we will get in the form of beta 1 plus beta 2i is the conditional expectation. Recall our discussion on simple linear regression modeling.

It is the conditional expectation of yi given xi that means expected value of yi given xi and in this linear probability model case it can be interpreted that the event will occur given xi that means conditioned upon xi the probability that event will occur is given by this expression. So probability that event will occur what is that event yi equal to 1 that means dividend is paid which is represented by y equal to 1 conditioned upon xi the probability this is what we are estimating and in the simple linear regression or even multiple linear regression if you remember we called it expected value of yi given xi which is this. So here the probability that event will happen is nothing but simply the condition conditioned upon xi the expected value of yi. So we can also write this expression in this form expected value of we have already seen this expression in the context of simple and multiple linear regression modeling that expected value of yi given xi equal to beta 1 plus beta 2 times xi and here the assumption is as we have seen earlier in this linear regression modeling case the expectation of error term is 0. To summarize this video we introduced the linear probability modeling approach to modeling what we called binary response variable or discrete choice variable for example if yi is whether a company will pay dividend that means yi equal to 1 or not pay dividend which is y equal to 0 we model this kind of binary response variable we model this kind of binary response variable or discrete choice variable with the help of simple linear probability modeling by this kind of expression where we said that the probability whether yi equal to 1 given xi xi here is our independent variable can be modeled in the form of following expression beta 1 plus beta 2 times xi here we also drew parallels between this approach and simple linear regression modeling approach where we said that the dependent variable is also written in the form of expectation of yi given xi which is equal to beta 1 plus beta 2 times xi where this expectation of yi is nothing but simply the event happening that means the probability of event happening that means y equal to 1 the probability that this event will happen so we are simply modeling probabilities in the form of linear regression model.

Linear Probability Model (LPM)

In such models, the actual probabilities cannot be observed, so your estimates (or dependent variables) would be 0s and 1s

- This is called linear probability model. The conditional expectation of Y_i given X_i, i.e., $E(Y_i|X_i)$, can be interpreted that the event will occur given X_i: that is, $P(Y_i = 1|X_i)$
- $E(Y_i|X_i) = \beta_1 + \beta_2 X_i$ (assuming $E(u_i)=0$)

$$E(Y_i|X_i) = \beta_1 + \beta_2 X_i$$

In this video we will examine some of the issues with LPM that is linear probability modeling. Recall our discussion on non-normality and heterosidasticity on the video topics related to heterosidasticity and normality in linear regression modeling. There are two possibilities for the event yi that is our dependent variable either 0 or 1. Remember we said that if yi equal to 1 as per our regression model it has a probability of pi basically pi if you recall the discussion in the previous video pi was the probability that yi equal to 1 and the remaining because probabilities have to sum up to 1, 1 minus pi is the probability that yi equal to 0. Let us calculate the expected value of yi given xi this is quite simple if you recall our calculation of expected value it is 0 with 1 minus pi probability and 1 with pi probability and therefore the expected value should be equal to pi 0 into 1 minus pi plus 1 into pi which is equal to pi which is our conditional value of yi given xi.

When your dependent variable is of this nature this kind of model is fraught with several econometric issues. To begin with notice when yi equal to 1 essentially your error term is of this nature why because yi equal to beta 1 plus beta 2 in our simple model xi plus error term. So if yi equal to 1 this is the form of your error and if yi equal to 0 you can rearrange and your error term will be like this. Now this kind of error this the nature of error when formula is of this form and resulting form is this error is not normal anymore this kind of jumping behavior from this to this so there are only two options for error either this or this. So your error term is not exactly normal this should be clear looking at this these values when yi equal to 1 error is of this form and yi equal to 0 error of this form so there are two very distinct behavior of error terms.

$$E(Y_i|X_i) = 0 * (1 - P_i) + 1 * (P_i) = P_i$$

Issues with LPM

Non-normality and heteroscedasticity of error terms

- Y_i has the following distribution

  $E(Y_i|X_i) = 0\times(1-P_i) + 1\times(P_i) = P_i$

- This kind of model has a number of econometric issues
- What is the nature of errors: $u_i = Y_i - \beta_1 - \beta_2 X_i$?

Y_i	Probability
0	$1 - P_i$
1	P_i
Total	1

	u_i	Probability
When $Y_i = 1$	$1 - \beta_1 - \beta_2 X_i$	P_i
When $Y_i = 0$	$-\beta_1 - \beta_2 X_i$	$(1 - P_i)$

However it is still not a problem with large samples so if your sample is large this normal distribution of error term is still not a problem. I request you to look for a term called CLT

central limit theorem which says that in large samples the normality of error term is ensured because if you have sufficient amount of large sample then in repeated samples your sample estimate sample estimates of beta 1 and beta 2 will be normally distributed even though your error term may not be normally distributed but if you have large samples in repeated sampling your estimates of beta 1 and beta 2 will be normally distributed this comes from the central limit theorem and therefore central limit theorem or CLT ensures that mu i even if it is not normally distributed if you have a large sample this is not an issue. However at the same time notice mu i's are also heterosidastic recall our discussion on media topic on heterosidasticity in linear regression. Here mu i's they vary with yi so either yi equal to 1 then mu i have a separate function and if yi equal to 0 then mu i has a separate function. It means that mu i's are varying with yi which is clear cut violation of homosidasticity assumption which means the errors or mu i's are heterosidastic and therefore this can create number of issues with estimation of coefficients like beta 1 and beta 2.

$$u_i = Y_i - \beta_1 - \beta_2 X_i$$

Issues with LPM

Non-normality and heteroscedasticity of error terms

- u_i is not normally distributed; although in large samples, it is not a problem
- u_is are heteroscedastic, i.e., they vary with Y_i

Y_i	Probability
0	$1 - P_i$
1	P_i
Total	1

	u_i	Probability
When $Y_i = 1$	$1 - \beta_1 - \beta_2 X_i$	P_i
When $Y_i = 0$	$-\beta_1 - \beta_2 X_i$	$(1 - P_i)$

Another very critical issue that is getting validated here is the fact that expected value of yi which is nothing but the probability Pi should be between 0 and 1 because this is probability and by nature it should be between 0 and 1. However if the model in the linear probability model there is nothing that stops this probability to be greater than 1 or less than 0. Let's look at an example here. So if we have yi as a function modeled like this let's say our estimated function is minus 0.

$$P_i(Y_i = 1 \,/\, X_i = 1) = \beta_1 + \beta_2 X_i$$

3 plus 0.012 xi and recall in the previous video we noted that here xi is the market cap and yi here is what we are modeling is the probability whether dividend is paid or not. So if dividend is paid then yi equal to 1 if dividend is not paid then yi equal to 0 and essentially we are modeling that probability of dividend being paid. In this function if the resulting

function of this form then every 1 million dollar increase xi is in million dollar so every 1 million dollar increase in the capitalization indicates that the probability that form will pay dividend increases by 0.

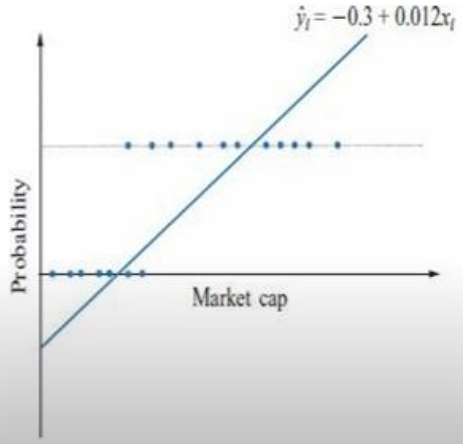
012 or 1.2 percent. In this expression if x is less than 25 dollar then this value of the probability value that is computed from this expression is less than 0 and if x is more than 88 dollar then the probability value Pi will be greater than 1. Now this is a very problematic case where probabilities that are obtained or estimated from the regression are less than 0 or greater than 1. Why so? Because a probability of less than 0 indicates that small size function this form for which this probability is coming out as less than 0 recall that number 25 million dollar that means these forms are rather small forms they will never be dividends and similarly for all the large forms Pi greater than 1 indicates that they will always pay dividends why? Because as a solution you would rather as a simple solution to this problem of less than 0 and greater than 1 you would rather put all the observations less than 0 as equal to 0 and all the observations greater than 1 as equal to 1 which means you are a essentially assuming that small forms with a certain market cap in this case 25 million dollars less than 25 million dollars will never pay dividend will be Pi equal to 0 and greater than 88 million dollars will have will always pay dividend that means their Pi is equal to 1. This is a very problematic assumption and understanding of the model. Another very critical problem with this kind of modeling is the diminishing utility of R square as a goodness of fit measure.

$$Y_i = -0.3 + 0.012\,X_i$$

Issues with LPM

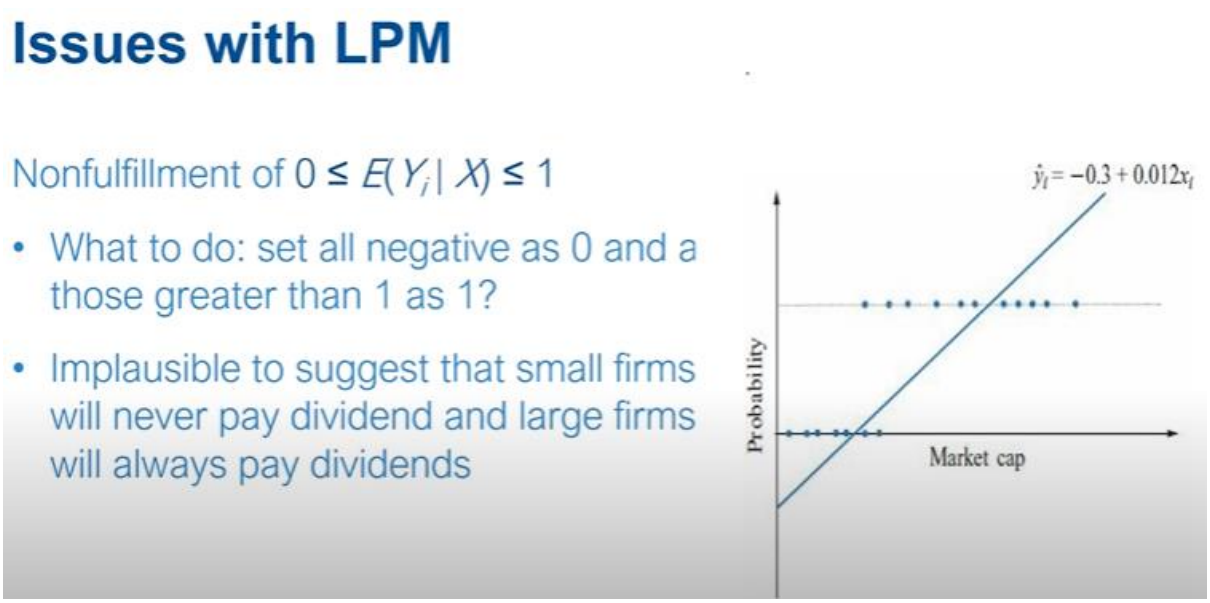
Nonfulfillment of $0 \le E(Y_i \mid X) \le 1$

- $Y_i = -0.3 + 0.012X_i$; where X_i is in million dollars
- For every $1 million increase in size, the probability that the firm will pay dividend increases by 1.2%
- However, for $X <$ $25 million and $X >$ $88 million, the probabilities are less than 0 and more than 1



Recall our discussion about linear regression model. We said that R square measures how well the model explains the variation in the dependent variable. Notice in this kind of setting the values of the dependent variable can be either only 1 or 0 that means they will be at this or this and therefore any linear probability modeling like what we are doing here LPM linear probability model which fits a line like this will have extremely high the error terms will be

very high and therefore because of such large variation in error terms the R square will be very very low therefore R square is a goodness of fit measure will give very poor performance because by nature of the fit of model all the observations will be either Y equal to 1 or 0 and therefore such large errors will result in extremely poor R square and therefore the conventional linear probability model is not expected to fit well with the actual observations. Only for very few cases where observations are very close to this point the fitted point here close to this point B and this point A will probably give lower errors and therefore only few observations will have low magnitude of errors for all the other observations the magnitude of errors is extremely large and very frenetic jumping up and down. To summarize this video we noted that linear probability modeling approach that is translating our simple linear regression modeling or multiple linear regression modeling approach to modeling discrete choice variables like 0 1 or yes no is fraught with issues to begin with the nature of error is non-normal and heterostastic by very nature of design of LPM models.



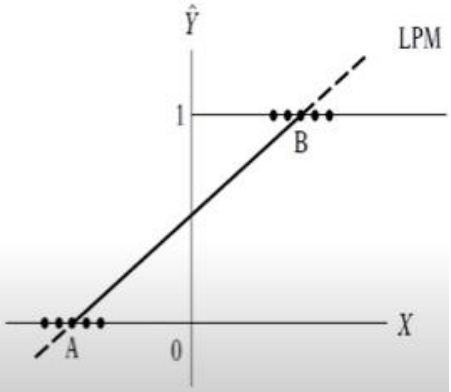
Second the utility of R square becomes very low as a goodness of fit measure R square doesn't work well with LPM linear probability models and third and also very important that the variable being modeled here is of 1 0 kind of form which is modeled in the form of probabilities and probabilities are restricted between 1 and 0 and there is nothing in the linear probability model which is an extension of simple linear regression that stops these probabilities between 0 and 1. So, the probabilities can actually extend beyond 1 and less than 0 as well which is also problematic. In this lesson in next series of videos we will discuss the logit or logistic regression and probit approaches logit and probit models that are over to that are able to overcome the limitation of linear probability modeling where it produces values less than 0 and more than 1 and also account for other issues that we discussed that is non-normality and heterostastic of error term and different goodness of fit measures in the context of logit and probit approaches that overcome all these problems. In this video we will introduce a very important classification algorithm which is often referred to as logistic regression or logit modeling approach to model binary response or qualitative

response variables or what we call discrete choice variables such as 0 1 1 0. Recall one of the major problems faced by LPM approach that it resulted in dependent variable which was beyond boundaries of a probability that is less than 0 or greater than 1.

Issues with LPM

Diminishing utility of R^2 as a goodness of fit measure

- All the Y values will be on a line $Y = 0$ or $Y = 1$
- The conventional LPM is not expected to fit well with such observations, except those cases where all the observations are scattered closely around points A and B
- Both logit and probit approaches are able to overcome the limitation of LPM that it produces values less than 0 and more than 1

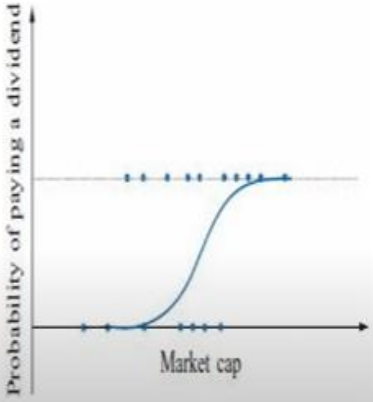


These logit and also the probit approaches they try to overcome this problem or the limitations of the linear regression model by transforming the model to a function that is bounded by 0 and 1. And therefore, the fitted function as per the logit approach and also later on as we will see with the probit approach looks like a S shaped curve here which is bounded on the left side by 0 and on the right side by 1. And the precise function in the case of logit model the precise function is f z i function of z i is e to the power z i upon 1 plus e to the power z i. We can also rewrite this equation as by dividing it with e to the power z i and it will result in 1 upon 1 plus e to the power minus z i. Let us examine this function in slightly more detail.

Introduction to Logit Model

The logit (and probit) approaches overcome the limitations of the regression model by transforming to a function so that fitted values are bounded within (0,1) interval

- The fitted function looks like an S-shape curve
- The logistic function for a random variable z is: $F(z_i) = \frac{(e^{z_i})}{(1+e^{z_i})} = \frac{1}{(1+e^{-z_i})}$



$$F(z_i) = \frac{(e^{z_i})}{(1+e^{z_i})} = \frac{1}{(1+e^{-z_i})}$$

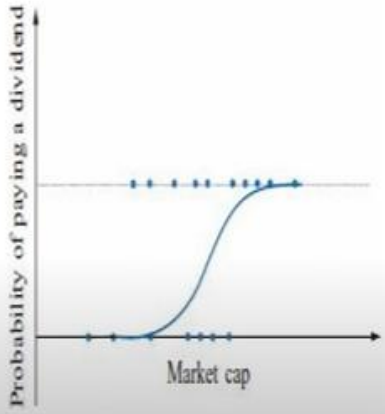
Here, this function f z i is nothing f z i is nothing but a cumulative logistic distribution what we call cumulative logistic distribution where z i is equal to beta 1 plus beta 2 times x 2 i and so on up till beta n times x n i plus error term. This is a familiar expression to us. This is same expression as we saw in the case of linear probability modeling. Only now that we have transformed this into the cumulative logistic distribution or logit function which is probability of y i equal to 1 is equal to 1 upon 1 plus e to the power minus z i where z i is this expression. So, this is sort of non-linear transformation of the original probability linear probability model.

$$P_i\,(y_i = 1) = \frac{1}{(1+e^{-(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_{i)}})}$$

Introduction to Logit Model

The logit (and probit) approaches overcome the limitations of the regression model by transforming to a function so that fitted values are bounded within (0,1) interval

- Here *F* is the cumulative logistic distribution
- The final logit model: $P_i(y_i = 1) = \frac{1}{(1+e^{-(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i)})}$

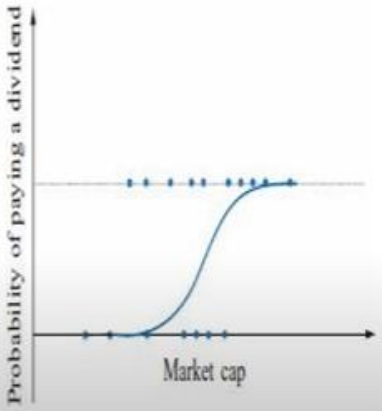


Let us discuss the sum of the properties of this logit function which is p i is y equal to 1 that is probability of y i being 1 in this expression where this part is z i. So, it is 1 upon 1 plus e to the power minus z i. Now, if z i let us look at the asymptotic values when z i tends to plus infinity and tends to minus infinity. If it is tending to plus infinity, this value will become 0 and the overall function the cumulative logistic function distribution function will approach a value of 1. So, it will tend to 1 wherever z is approaching infinity and if it is approaching minus infinity, then this value will approach to infinity and overall expression will approach to 0.

Introduction to Logit Model

$$P_i(y_i = 1) = \frac{1}{(1+e^{-(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i)})}$$

- Model asymptotically touches 0 ($Z \rightarrow -\infty$) and 1 ($Z \rightarrow \infty$)
- Is this model linear? Hence, not amenable to OLS estimation
- The model would predict that the probability, e.g., probability of bank loan default (dependent variable = y)



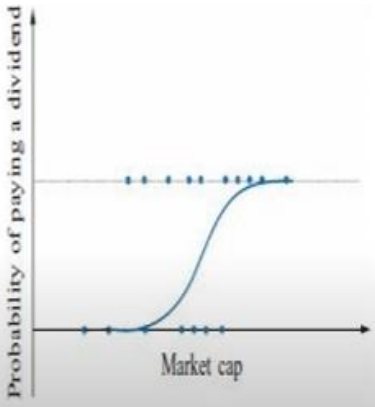
$$P_i\ (y_i = 1) = \frac{1}{(1 + e^{-(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i)})}$$

So, the limiting cases of 0 and 1 are achieved and therefore, whatever the value of this expression z i which is z i is represented by this expression the linear function in variables whatever irrespective of these values, the limiting cases of probability will remain between 0 and 1. However, we can easily see here that now the model is not linear in parameters. So, this is not a linear model and hence it is not amenable to OLS estimation. At the end of this lesson, we will discuss an approach called maximum likelihood approach. At the end of this discussion, we will discuss an approach called maximum likelihood approach or MLE, maximum likelihood approach, which is employed to model and estimate such models, but important here to note that these are not model using OLS estimation because these are not linear models, not linear parameters.

Introduction to Logit Model

$$P_i(y_i = 1) = \frac{1}{(1+e^{-(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i)})}$$

- $P(y = 1)$, then $P(y = 0) = 1 - P(y = 1)$
- Here independent variables are x_{2i}, x_{3i}, x_{4i}, x_{5i}, and so on
- This is essentially a non-linear transformation of the model to produce consistent probability results



$$P_i\ (y_i = 1) = \frac{1}{(1 + e^{-(\beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + u_i)})}$$

$$P(y = 1), then\ P(y = 0) = \ 1 - P(y = 1)$$

Importantly for us, this model will predict the probability that y i equal to 1. For example, we can take an example where the bank loan default let us say default is y equal to 1, where y i is our dependent variable which is the probability of default. So, this model would predict the probability that y i takes a value of 1. So, now that we have said that probability of y equal to 1 is the probability of interest that is let us say probability of default and so on. Probability of non-default can be easily estimated as 1 minus probability of default.

Here, the dependent variables x 2, x 3 and x k and so on, these are our dependent variables. And essentially, this is a non-linear transformation of the model to produce probability consistent result. So, that means we are transforming our model, which was originally the linear probability model into a slight, into a non-linear model, which provides us with the probability consistent results. To summarize this video, we introduce the logit model or logistic regression model, wherein we transformed our original function into cumulative logistic distribution function, which appeared something like this, where z i here, this z i is of that similar to that linear probability model, which was of this form beta 1 plus beta 2, x 2 y and so on up till beta n x n i plus the error term. The key desirable properties of this function included that this expression when z i tends to infinity, this expression tends to go to 1 and when z i tends to minus infinity, this expression tends to go to 0, which means now these values are probability consistent.

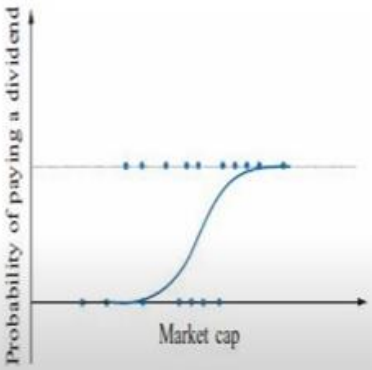
In this video, we will improve our understanding of logit function and relate it with the odds ratio, which is very familiar with the concept of probability. Let us re-examine our logistic function, logit function here, which is simply probability of y i being 1, which is the event of interest as 1 upon 1 plus e to the power minus z i, where this expression represents the z i, the linear part of the model. Now, recall, if the values of z i, which is this expression, that is this one is extremely low and negative. So, if the values of z i extremely low and negative, then that would indicate the probability of no dividend, recall our example where no dividend case was probability of y i equal to 0. So, extremely low values of this expression would suggest probability of no dividend or another example probability of non-default cases with a higher probability.

$$\frac{1}{(1 + e^{-z_i})}$$

Understanding the Logit Function

$$P_i(y_i = 1) = \frac{1}{(1+e^{-(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i)})}$$

- Here extremely low and negative values of the linear function $\beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki}$ would predict No dividend (or non-default cases) with a high probability or $P_i(y_i = 0)$



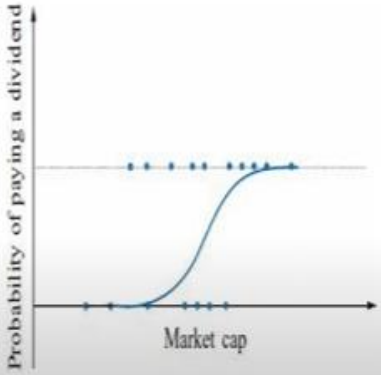
$$P_i\ (y_i = 1) = \frac{1}{(1 + e^{-(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i)})}$$

So, higher chances of no dividend or non-default if this value is very low or negative because remember 1 plus e to the power minus z i, if this value is very low negative minus it tending to minus infinity, this expression will become, tend to very large, tend to infinity and overall value will be 0. Similarly, extremely high and positive values of this expression, extremely high and positive values of this expression which is z i, which means if z i is tending to positive infinity, then this expression, this value will tend to 0 and overall value will tend to 1, which indicate a high probability of dividend payment or default cases which is of high interest. So, remember in our examples, the cases of interest were defaulters with y, which we classified as even y equal to 1 or payment of dividend which we classified as even y equal to 1. So, a high probability of these events if the value of z i or this expression is very large. Now, let us develop this understanding in the terms of our very familiar odds ratio.

Understanding the Logit Function

$$P_i(y_i = 1) = \frac{1}{(1+e^{-(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i)})}$$

- Extremely high and positive values of the linear function $\beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki}$ would predict dividend payment (or default cases) with high probability or $P_i(y_i = 1)$



$$P_i\ (y_i = 1) = \frac{1}{(1 + e^{-(\beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + u_i)})}$$

Odds ratio is nothing but the ratio of let us say there are two possible events, in this case default, non-default or payment of dividend and non-payment of dividend, then the event of interest on the numerator and event of not interest, opposite to that interest in denominator will give us the odds ratio which is probability of y equal to 1 divided by probability of y equal to 0. Now, if this odds ratio is greater than 1, then it is obvious that y equal to 1 event which is either default or payment of dividend is more likely, while if odds ratio is less than 1 that means event which is coded as y equal to 0 that means either non-default or in some other example, non-payment of dividend is more likely. So, this is how we define our odds ratio. So, now let us substitute our logit function in the odds ratio then remember what was our odds function, e to the power minus zi, this was our function of interest. So, we put this in numerator, this was for py equal to 1 and for py equal to 0, what should be the relevant expression that should be 1 minus 1 upon 1 plus e to the power minus zi.

So, if you simplify this expression, eventually it will work out to e to the power zi. So, you can, it is a very simple computation, numerator divided by denominator, recall this function, this function, numerator divided by denominator we get this expression e to the power zi which is and where zi is this expression. So, if you simplify the odds ratio, if you simplify the odds ratio by putting the values for p equal probability of y equal to 1, we have 1 upon 1 plus e to the power minus zi and p y equal to 0 which is nothing but 1 minus p of y equal to 1. So, if we put these expression, we get e to the power zi and therefore odds ratio is nothing but e to the power zi or if you take the natural log on both sides, we get ln odds equal to this expression. Natural log we have taken on both sides, left lh and rh, so we get natural log of odds as this expression.

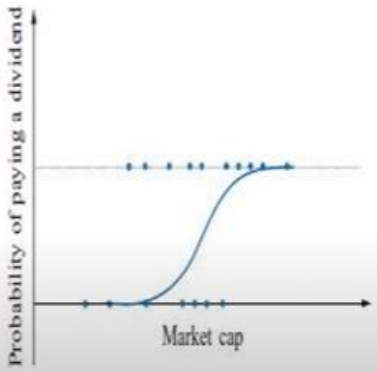
Now, notice the higher this value, the higher this expression or natural log of odds, the higher the probability that the event likely event is y equal to 1. So, the higher the value of this expression, the higher the odds in favor of event y equal to 1 and vice versa that means lower extreme low negative values of this expression, lower the odds in this expression, this probability and more odds are in more favor of this event. And this is what we said earlier as well. To summarize in this video, we discussed the logit function in greater detail, we transformed the logit function and related it to the familiar odds ratio. We already know that odds ratio is the ratio of two events where the numerator is the event of interest and denominator is the opposite event.

$$P_i\ (y_i = 1) = \frac{1}{(1 + e^{-(\beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + u_i)})}$$

Understanding the Logit Function

$$P_i(y_i = 1) = \frac{1}{(1+e^{-(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i)})}$$

- This can also be expressed in the form of Odds
- $\text{Odds} = \frac{P(y=1)}{P(y=0)}$;
- Odds > 1 if $y = 1$ is more likely
- Odds < 1 if $y = 0$ is more likely



$$Odds = \frac{P(y=1)}{P(y=0)}$$

We found that natural log of odds is nothing but the same linear function, which is zi. So, zi, the expression corresponding to zi is nothing but the natural log of odds and therefore we simply and very intuitively understood that higher value of zi, extremely high value of zi, positive high value of zi would increase the odds in favor of the event of interest which is y equal to 1 and extremely low and negative values would favor the odds corresponding to probability y equal to 0. In this video, we will discuss the concept of thresholding in the context of logistic regression. In the case of logistic regression or discrete choice or limited dependent variable models, essentially the outcome of the model is probability. For example, remember our dependent variable where we estimated essentially what we estimated was probability that event of our interest will happen, that is, that was the probability that we wanted to estimate, probability that yi equal to 1.

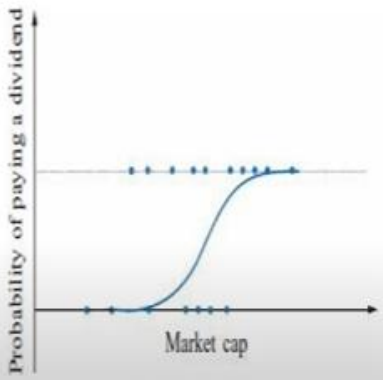
For example, if you remember, recall our bank loan default that we have continued in this lesson we said that if the application defaults, then that case which is yi equal to 1, if yi takes value of 1, we estimate the probability from our regression model which is the logit, which we estimated using logit function of the following form, e to the power minus zi. Now, in real life, you would not only want to make prediction about probabilities, but actually what you observe is 1s and 0s. For example, 1 may be a defaulter and 0 may be a non-defaulter. So, you would like to convert your estimated probabilities using logit or even if it is a probit or any such model and convert them into cases of 1s and 0s.

$$\frac{1}{(1+e^{-z_i})}$$

Understanding the Logit Function

$$P_i(y_i = 1) = \frac{1}{(1+e^{-(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i)})}$$

- If we substitute the logit function in Odds equation, then
- $\text{Odds} = \exp^{(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i)}$ or
- $\ln(\text{Odds}) = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + u_i$
- The higher this logit (or ln(Odds)) form, the higher the probability for $P_i(y_i = 1)$



$$P_i\,(y_i = 1) = \frac{1}{(1 + e^{-(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i)})}$$

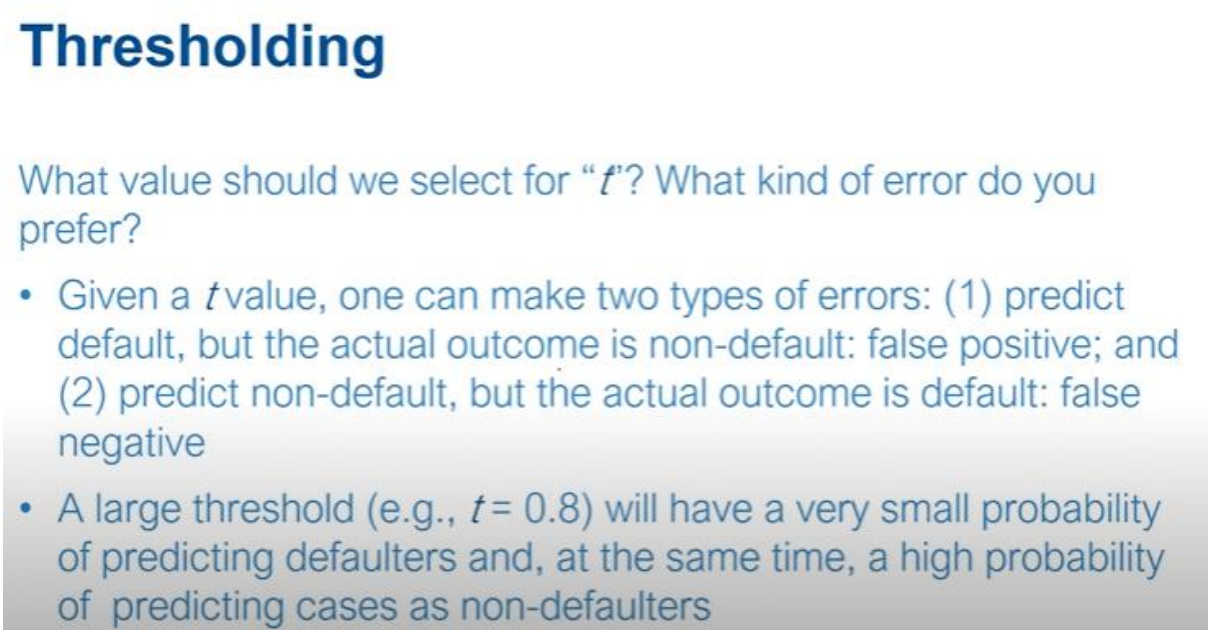
This requires some kind of threshold value t. This t value is needed to convert your probability predictions into binary predictions of 1s and 0s that is for example default or no default. A simple example to build the intuition would be let us say you select a value of threshold tau, if this tau is, if the value that you obtain the probability value that yi equal to 1 is greater than tau, greater than equal to tau, then you say that it is a default or the event that is of interest, you can change it also default or non-default. And in case if it is less than t, if the value, the probability value is less than t, then you predict as a case of yi equal to 0 that means you predict it as a non-default case if the probability is less than t. Now, an interesting question here is what should be the appropriate value of tau?

Thresholding

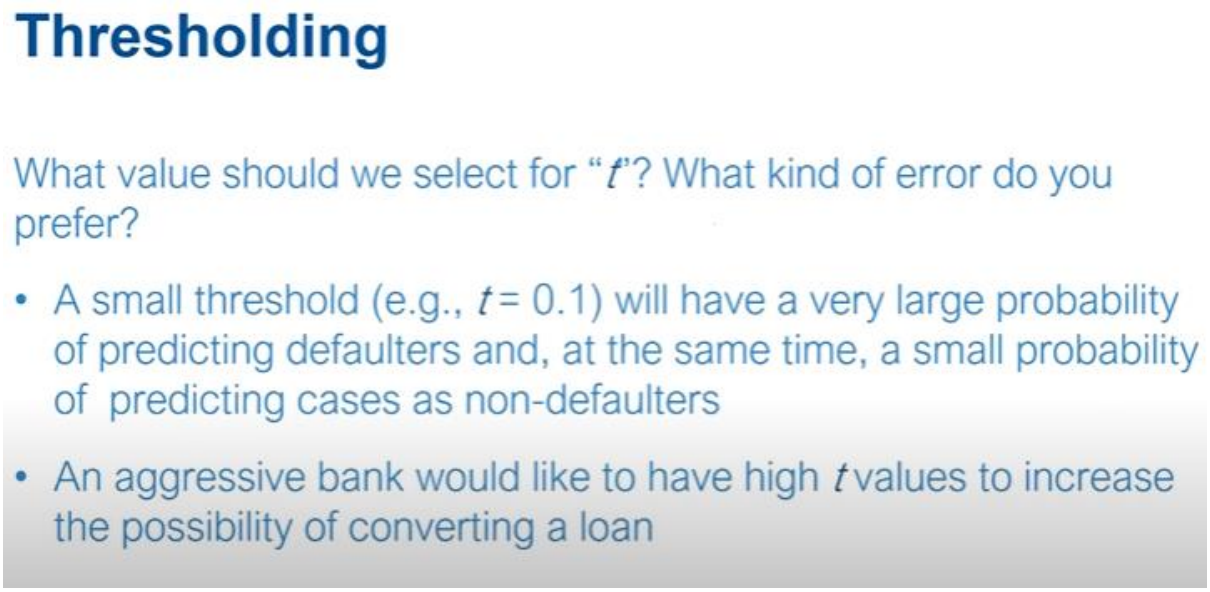
The outcome of the regression model is a probability

- In real life, you would want to make a binary prediction, e.g., default or no default
- For this, we may consider a threshold value "*t*"
- If *P*(Default = 1) >= *t*, then predict a default case
- If *P*(Default = 0)< *t*, then predict a non-default case

What kind of error would you want to prefer? Please notice given a t value, there is a possibility of making two types of error. One, whether the actual outcome is of non-default but you still predict it as default which can be considered as false positive.

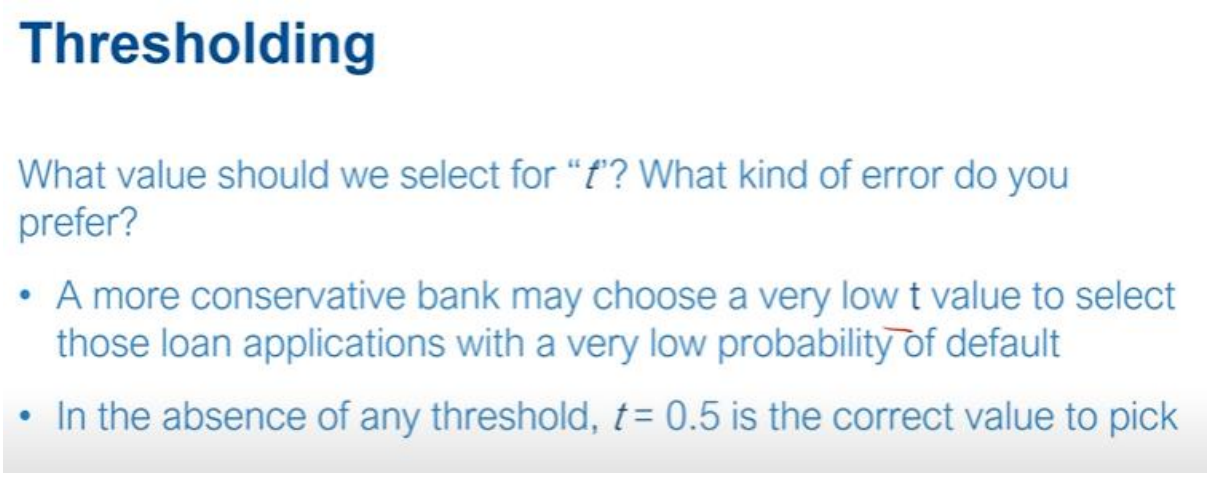


You consider it falsely as a case of yi equal to 1 that is default while it is actually a non-default case. And the second kind of error is you predict non-default which is actually a default. So it is actually a yi equal to 1 case but you predict it as yi equal to 0, you predict non-default.



So, this is false negative. Let us try to visualize this. So for example, consider there are, this is your tau value and you have some, these are all your non-defaults. Let us call these xs as non-defaults. And let us consider on the higher side these are defaulters in 0s, these are defaulters. See I am putting visually them on the higher side because they have a higher tau

value, a higher probability of default I am associating with them on the higher side. Now suppose you select a very large threshold, suppose you select a very large threshold of let us take an extreme threshold of 0.



`

8. So let us say you pick a very high value of t, let tau or threshold let us say a very high value like 0.8 and since it is a very high value you will classify a very small fraction as defaulters. So now in this case, what about these default cases? Because you have taken a very high value of tau while you would select certain number of cases which are defaulters, there are number of default cases 0s here which will be classified as non, incorrectly classified as non-defaulters. So a large number of defaulters will be incorrectly classified as non-defaulters. That means with the large probability you will classify large number of default cases as non-defaulters.

So a number of false negatives will be occurring. So this is a challenge. However, the positive point about taking a high tau value is that a sizable, most of the non-defaulters, these non-defaulters will be classified as non-defaulters only. That means your false positives will be very very less. So your false positives are very less but your false negatives are very high if you are taking a very high value of tau or thresholding.

Let's take another example where you take a very low value of threshold. Let's say tau value of 0.01, this kind of small value or sorry 0.1. You will, your cut-off will be from here. Now notice in this case all the default cases, default cases you will correctly and most of them you will accurately classify them as defaulters only.

So most of the defaulters are accurately classified as defaulters. So the error on this side, the false negative error would be very less because most of the, all the defaulters where actual outcome is default are being classified as default only. But what about these cases which are non-defaulters? It seems that many of these non-default cases now are being considered as default. That means false positives. So now we are high on the side of positive error.

So our positive, false positive is high while false negative is low. So as we can see there is a trade-off. As we increase the value of tau or threshold, we make one kind of error while being good on the other side and vice versa for the positive case. That means if we take a very high value of tau, our false negatives are higher but false positives are lower. But when we take a very low value, our false negatives are lower but false positives are higher.

So that is a trade-off. Now let us try to answer the question which kind of bank would select what kind of value or how this value would be selected. So let us take two examples. One a very conservative bank that want to give loans only to very creditworthy borrowers and a very aggressive bank which wants some kind of expansion, make some new relationships and therefore want to loan out in large volumes. So these two, let us consider two cases. A very conservative bank would like to select only and only very niche, very small number of borrowers and therefore it would like to classify even a small risk borrower as defaulter.

And therefore it would set a very small threshold. So in that matrix, in the diagram it will set a very small threshold. So a large number of cases would be classified as defaulters. And therefore those that are selected, a very small population of loan applications will be selected, small proportion will be selected and definitely these would be very creditworthy borrowers. That means this bank, this bank is taking a chance in order to select only most creditworthy borrowers, it is taking a chance and classifying even the relatively safer ones into the defaulter category. The advantage is aligned to its objective that means it will get only and only those applications as non-defaulter that are very safe.

Contrast this to an aggressive bank strategy. That bank would like to give loan to most loan applications and therefore it would like to reject only and only high risky borrowers. So it will select a higher threshold, probably on the higher side here maybe, so that those that are rejected, it is very sure that these are definitely not the good ones. However in this process, it is taking a chance that while some of the good ones, most of the good ones are getting selected but there are some of the bad borrowers, not so good creditworthy borrowers are also falling in that list because of this high threshold value it has put which is rejecting only some of the extremely bad borrowers. So some of the good borrowers are also, or not so good borrowers are also falling inside and get their loans accepted or approved. However please remember this, the value of tau changes its interpretation if your case of Y equal to 1 and Y equal to 0 changes.

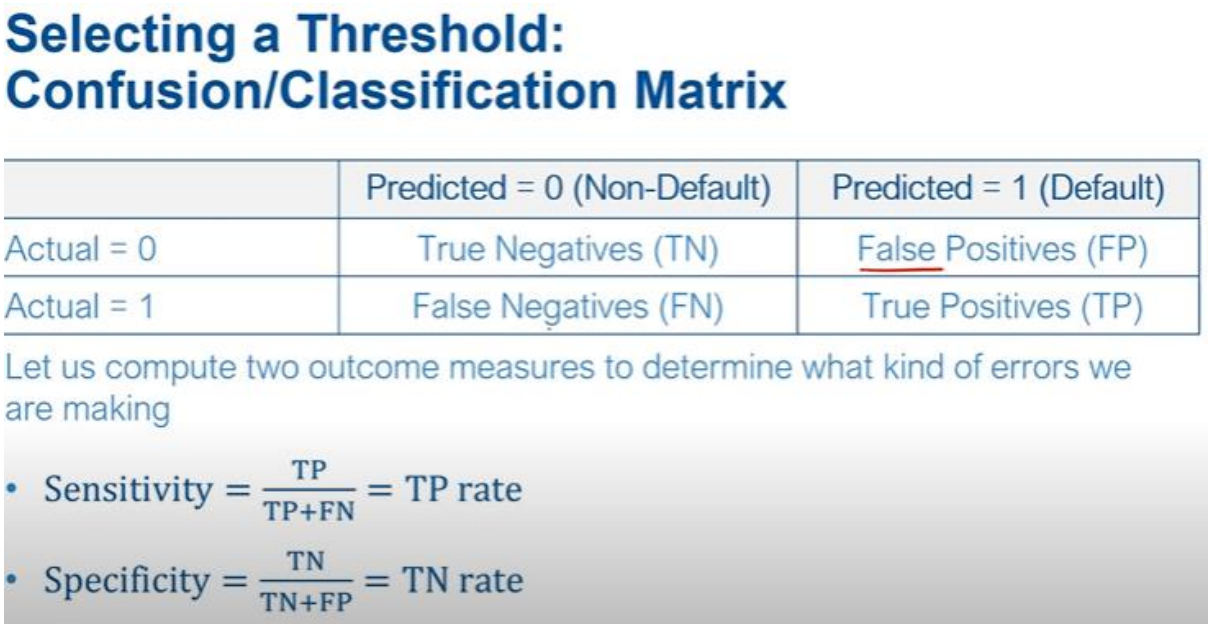
For example, all our discussions we are considering Y equal to 1 as the case which is default and 0 not default. As soon as you change this interpretation, you take Y equal to 1 as not default and 0 as default, then the entire interpretation changes to the opposite side. So we said that if the bank was more aggressive, it wants to accept most of the applications, expand business, then it will take a higher threshold. A conservative bank may choose a very low threshold value.

So this is the in a nutshell what we understood. To summarize, in this video we discussed the

concept of thresholding. We noted that while the model, the logit and similar class of models on discrete choice variables or limited dependent variables model and give dependent variable in the form of probabilities, we need to convert those probabilities into binary outcomes or something like 1, 0, just, no's and so on. We took an example of bank loan default and non-default application to explain the case. However, in order to convert these probabilities into 1's and 0's, we need a thresholding or number, a cut-off range above which the values can be classified if probability lies above in from that we can classify it to 1 and if it is low then 0.

So we need a threshold number which can be used to classify. We also noted with the example of bank that there is a trade-off. Depending upon the value of tau, you have a certain probability of making error called false positives that incorrectly classifying as 1's and false negatives that means incorrectly classifying as 0's and there is a trade-off between them which depends upon your selection of tau value. As a general rule, in the absence of any guidance, t equal to 0.5 is a good threshold, usually considered a reasonable threshold.

In this video, we will discuss the concept of classification matrix. Classification matrix is a very important concept employed to examine the efficiency of a classification algorithm such as logistic regression. Recap, in the previous video, we introduced the concepts of or the errors known as false negatives and false positives. Let us elaborate in more detail now. We noted that there are cases and we will carry with our example of default and non-default.



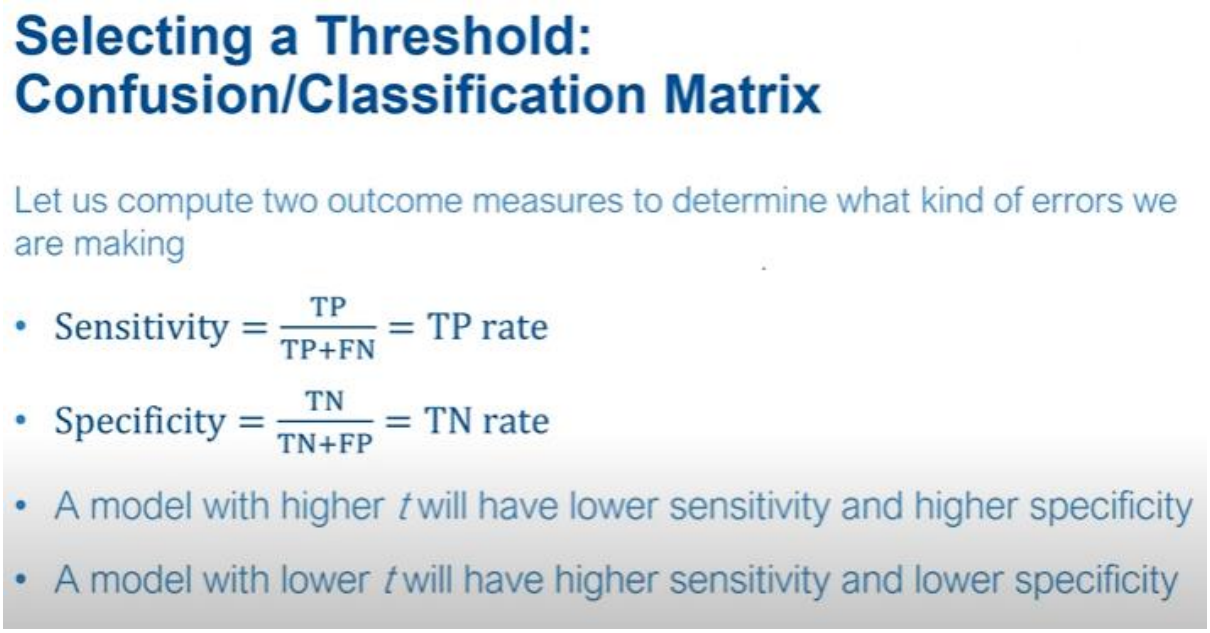
Remember, we classified y equal to 0 as cases of non-default. Now cases where your regression algorithm like a logistic regression or classification algorithm like logistic regression, those cases that are actually non-defaulters and correctly predicted as non-defaulters are considered as true negatives. Similarly, those cases which are actually non-defaulters that means 0's but incorrectly classified as defaulters or 1's, they are considered as

false positives. Similarly, those cases that are actually defaulters or 1's but incorrectly classified as non-defaulters of 0, they will be called false negatives. And conversely, those that are actually 1's but predicted and also predicted as 1's that means defaulters, correctly predicted as defaulters are considered as true positives. Now, this matrix is often referred to as classification matrix or confusion matrix and various efficiency measures, various efficiency or accuracy measures to understand the accuracy of a classification algorithm like logit model is derived from this matrix.

Let us look at two very important parameters that are derived out of this matrix. First is sensitivity. Sensitivity is nothing but the ability of our model to correctly classify true positives out of total positive cases. So for example, total positive cases are true positives plus false negatives. Actual ones are true positive, false negatives plus true positives which are here in denominator and in numerator we have true positives.

$$Sensitivity = \frac{TP}{TP + FN} = TP\ rate$$
$$Specificity = \frac{TN}{TN + FP} = TN\ rate$$

So this is called true positive T-rate. Similarly, we have another measure defined as specificity. Specificity is nothing but overall true negativity rate which is true negatives on the numerator divided by true negative plus false positive. Now, as you would have understood now by our discussion of thresholding, there is a tradeoff between sensitivity and specificity depending upon our value of threshold. Recall that diagram of tau thresholding discussion where we said that let us say these 0's are our defaulters, these 0's are our defaulters and these crosses are non-default cases. And remember, remember we said that a high value of t, if you pick a very high value of t, let us say 0.



8 like this, there is a possibility, a high possibility that some of the default cases or y equal to 1's will be incorrectly classified as y equal to 0's. That means a high false negative rate, false negative. This means because your false negatives are larger, your sensitivity will be low. What about specificity? Once you have taken a very high value of t or tau, possibility that you will incorrectly classify non-defaulters as defaulters is very low.

That means your false positives are very low and you score high on specificity. So you score high on specificity but low on sensitivity if you choose a high t. Vice versa discussion applies for a lower t. That means if you pick a very low t, low value of tau or t thresholding, we are interchangeably referring to this t or tau. And in that case, remember most of the default, most of the default, in fact all of the default chances are you will classify correctly as defaulters.

So your false negative will be very less and your sensitivity will be high. However at the same time, if we are talking about specificity, there is a chance that some of the non-default cases may be classified as defaulters. That means a high false positive number and therefore your specificity will be low. So if your tau or thresholding t value is low, then you may score high on sensitivity but low on specificity. So depending upon your selection of tau, your sensitivity and specificity measures will vary. There are some other important measures of efficiency or accuracy of model prediction that are also computed.

You have not looked at one measure but you look at number of measures. For example, there is a measure of overall accuracy which sees how many true negatives and true positives. That means how many zeros you collect, true negatives are the zeros that you actually classified as zeros and true positives are the ones that you correctly classified as one, divided by total number of observations. So this is your overall accuracy. Similarly, overall error rate. So false positives, those are zeros that are incorrectly classified as ones and false negatives, those are ones that are actually classified as zeros.

Selecting a Threshold: Confusion/Classification Matrix

- Overall accuracy $= \frac{(TN+TP)}{N}$, where N = number of observations
- Overall error rate $= \frac{(FP+FN)}{N}$
- False negative error rate $= \frac{FN}{(TP+FN)}$
- False positive error rate $= \frac{FP}{(TN+FP)}$

$$Overall\ accuracy = \frac{(TN + TP)}{N}, where\ N = number\ of\ observations$$

$$Overall\ error\ rate = \frac{(FP + FN)}{N}$$

$$False\ negative\ error\ rate = \frac{FN}{(TP + FN)}$$

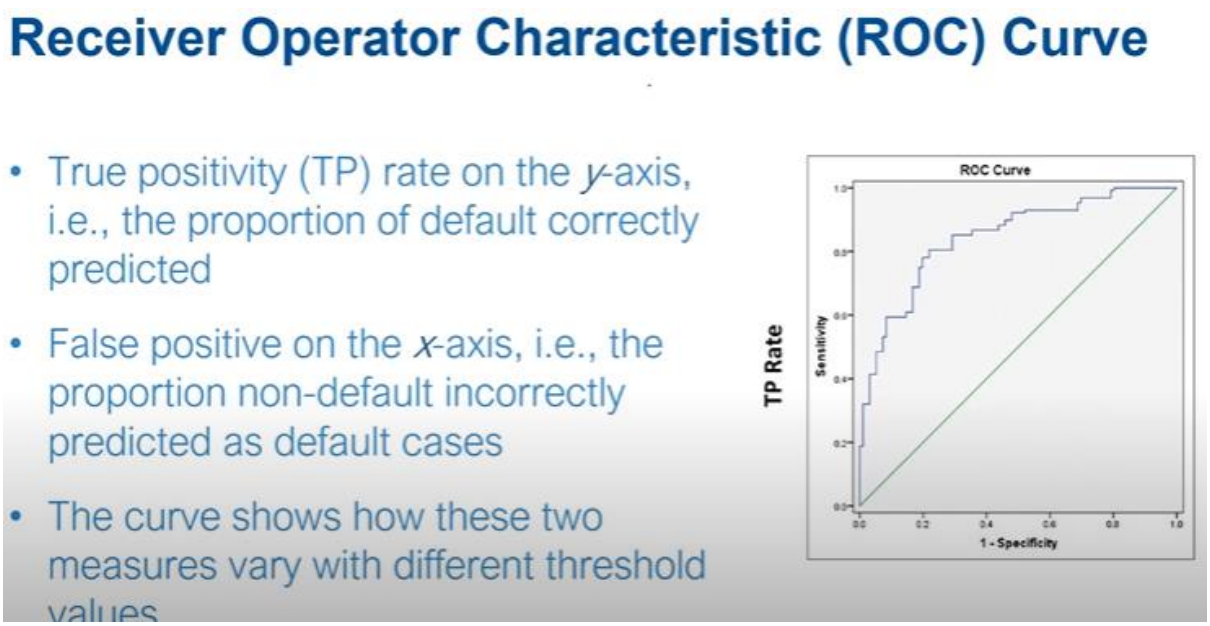
$$False\ positive\ error\ rate = \frac{FP}{(TN + FP)}$$

So both are inaccurate classification divided by total number of observations. So this is our overall error rate. Another measure is false negative rate. That means overall, what is the incorrect classification for the actual y equal to ones in this case, the falters that is incorrectly classified ones so that those that are ones incorrectly classified as zeros divided by total number of ones. How do we arrive at total number of ones? That is true positives, ones that are classified as ones and false negatives, ones that are incorrectly classified as zeros.

So the denominator is total number of ones and numerator is only those ones that are classified as zeros. Similarly, we have false positive rate. False positive rate is theoretically those that are zeros incorrectly classified as ones. So numerator we have zeros incorrectly classified as ones and they are divided by, denominator by total number of zeros. That means true negatives, zeros that are classified as zeros plus false positives that means zeros that are classified as ones.

So this is false positive error rate. To summarize this video, we discussed a very important measure of model accuracy, which is confusion or classification matrix. Based on confusion classification matrix, we derived two important measures of sensitivity and specificity. We noted that selection of thresholding value, a very important concept we discussed in the previous video is a trade-off between these sensitivity and specificity. We also derived a

number of very important measures of overall accuracy of the model prediction, overall error rate, false negative rate and false positive rate. Essentially these, all these measures are derived out of confusion and or what we call classification matrix as we discussed in this video in detail.

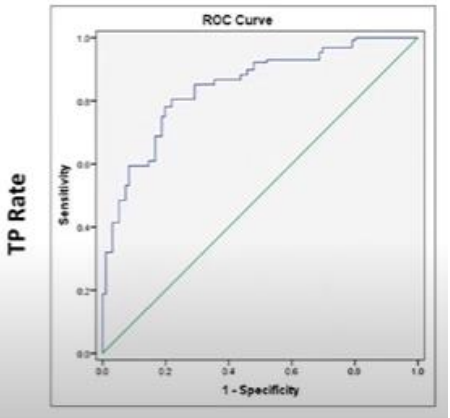


In this video, we will discuss receiver operator characteristic curve, which is a very important visual measure of the accuracy of classification algorithm. Recall our discussion on true positives and false positives. Remember, true positives are same as sensitivity measure and false positives are 1 minus specificity measure. We discussed these formulas in the previous video.

Now let us plot true positive and false positive rates on the y and x axis respectively. True positives are essentially the proportion of default cases, that means ones correctly classified as ones or sensitivity. False positives are non-default cases, incorrect or that means zeros in our case, incorrectly predicted as default cases or ones, which is 1 minus specificity also. On x axis, we have false positives and y axis we have true positives. So x also sensitivity on y axis or 1 minus specificity on x axis. Now a curve like this indicates the model performance, how these rates of sensitivity and 1 minus specificity vary depending upon different threshold values.

Receiver Operator Characteristic (ROC) Curve

- For $t = 1$, TP = 0, and FP = 0 → will not be able to predict any default cases but correctly predict all the non-default cases
- For $t = 0$, TP = 1, and FP = 1 → will be able to correctly predict all the default cases but incorrectly predict all the non-default cases
- As we move from $t = 1$ to $t = 0$, different combinations of TP and FP are obtained

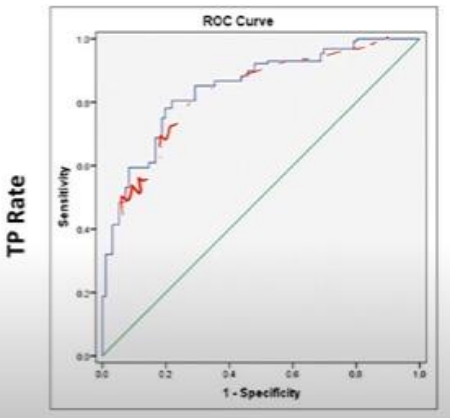


Let us discuss this in more detail, but this behavior, this movement is essentially what we call as ROC or receiver operator characteristic curve. Let us start with the case of thresholding value of tau or t equal to 1. In that case, our true positives will be 0. t equal to 1 is an extreme case where all the cases will be considered as non-default.

So all the loan applications or what we call as whether they are zeros or ones, they will be considered as zeros. That means none of the cases will be classified as ones, even the ones, all the ones will be classified as zeros and therefore true positive rate is zero. So on the true positive rate or sensitivity of the model is here zero. But at the same times, because we are not classifying any case as one, even those that are zeros, all the zeros are collectively classified as zeros. So none of the zeros are also classified as ones. So this is one, one can say a sort of positive from this model that none of the zeros will be classified as ones and therefore false positives will be zero.

Receiver Operator Characteristic (ROC) Curve

- ROC curve captures all the complete threshold behavior
- High threshold: high specificity and low sensitivity
- Low threshold: low specificity and high sensitivity
- Thus, it is a tradeoff between cost in failing to detect default cases vs. incorrectly considering non-default cases as defaulters



So this point will be here as well. So for t equal to 1, irrespective of the logit regression or a classification regarding the model, its coordinate will be 0, 0 for t equal to 1. This is irrespective of the model, it is all it will always be 0, 0 here. Now think of another extreme case where t equal to 0. In this case, if t equal to 0, all the whether it is zeros, defaulters or one zeros or non-defaulters or one or defaulters, all of them will be classified as defaulters that is ones.

That means all the defaulter cases ones are of course classified as ones. So our true positivity rate is one, our sensitivity is one, true positivity and sensitivity are same. So we are standing here. But at the same times, our zeros are also getting classified as ones that means false positive rate is also one.

All the zeros are classified as ones or specificity or false positives is also one. So we have this coordinates. Again, this is irrespective of the model, whatever model you choose the for t equal to 0, this point will always be 1, 1. However, for all the other cases, different models will show different behaviour except these two points, all the models will show different points, different behaviour. So as we move from t equal to 1 to 0, we will move on some curve like this. And one can judge the accuracy of model as per area under this curve.

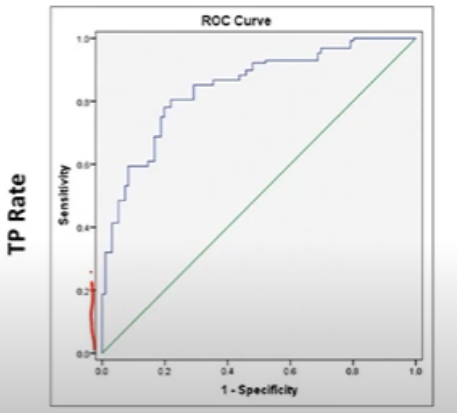
For example, whatever area under this curve as a percentage of overall area will reflect the accuracy of this model. Let us discuss this in more detail. Here this ROC curve, the behaviour of this ROC curve captures the complete threshold behaviour of our classification algorithm. For example, at different threshold values, this model will have high or low specificity and sensitivity. For example, if you have high threshold value like t equal to 1, this model will have very high specificity that means low 1 minus specificity, it will here somewhere here and low on sensitivity. At the same time, if you have very low threshold

value the model the point is somewhere here, that means low on specificity but high on sensitivity.

So there is a trade-off between the cost in failing to detect cases versus incorrectly considering non-default cases as defaulter. So there is always a depending upon threshold value there is a trade-off and this trade-off is very nicely captured by this ROC curve. For example, if I move from initially if I increase the threshold value in this region notice how sharply because the curve is very steep in a slope, how sharply my true positive rate or sensitivity of the model increases that means my ability to classify correctly once into once increases very sharply. At the same time, the fall in specificity here because we are plotting 1 minus specificity so the rather the fall from 1 to these lower values is very small. So notice in this region my sensitivity gain is very high as I increase my thresholding value but my loss of specificity is very less.

Receiver Operator Characteristic (ROC) Curve

- A 100% score area under the curve will indicate complete accuracy, i.e., all the observations are correctly identified
  TP = 1 and FP = 0
- A 50% score will indicate random guessing, that is, half TP = 0.5 and TN = 0.5 (FP = 0.05)



At the same time as I move in this region please notice my gain in sensitivity is very small because now the curve has become flatter but my loss is in specificity very high. So I would rather be somewhere here this may be one of the optimum points but again it is subjective depends on individual objectives of the entity which is doing the modeling but still it seems to be a good place to be here because here we have lot of gain in sensitivity and the loss in specificity is very small. Finally an ideal but a more of a theoretical case would be 100 percent score that means all the observations are correctly classified and in that case your movement that kind of model if that were to exist although it will not but if there is a model that completely and accurately classifies with 100 percent accuracy all the observations it will move like this and therefore the area under the curve for this kind of curvature which is moving like this is 100 percent it captures everything. That means all the observations are correctly classified so true positivity rate is 1 all the time and false positivity rate is 0 for all the thresholding values but this is more of a theoretical discussion. Let us take a random model where there is a 50 percent probability which indicates almost like a random guessing

a coin tossing a fair coin toss where 50 percent probability you can get a head or 50 percent probability you can get a tail.

In this kind of model you have for all the times your true positives is 0.5 and false positive is also 0.5. So on this behavior which is like a 50-50 model your movement will be reflected by this kind of curve so it is always 50-50 and therefore area under the curve will be 50 percent of the overall area. This is like a coin tossing game and this model is no good than a coin toss random guessing. For more practical models they will lie somewhere in between that is from 50 percent to 100 percent they will be somewhere in between and therefore the accuracy of the model can be very nicely measured as the percentage of area under this curve this ROC curve. To summarize this video we discussed the ROC receiver operator characteristic curve.

In this curve essentially provides a visual examination of the efficiency of the classification algorithm. Here on the y-axis we have true positive or sensitivity rate on the x-axis we have 1 minus specificity rate or what we call false positives. We noted that there are two extremes to this model one those models are purely theoretical models which classify every observation correctly and therefore the area under the curve of such model is 100 percent while then there are cases where 50-50 accuracy is there for them like this kind of 45 degree straight line area under the curve would be 50 percent. This ROC curve essentially captures the behavior of any classification algorithm depending upon different threshold values. So it captures the thresholding behavior at different thresholding levels for a given classification algorithm and the accuracy of any classification algorithm can be measured as area under its curve and we noted this area in normal practical cases should lie between 50 percent to 100 percent where these two 50 percent and 100 percent are two extreme cases and it lies somewhere in between. We also noted that for all the models two points two extreme points that is thresholding value of t equal to 1 and t equal to 0 are the same.

In this video we will talk about parameter interpretation of logic class of models. Recall our discussion about simple linear regression model of this form where we said yi equal to alpha naught plus alpha 1 into xi plus the error term mu i. We said that here the interpretation of parameter alpha 1 goes like this if xi and yi are absolute form then one unit change in xi will result in alpha 1 unit change in yi. If xi and yi are in log-log form then percentage change in xi, 1 percentage change in xi will reflect in alpha i percentage change in yi. And if we were to plot this it would appear like this x, y on xy axis this would appear like a straight line where alpha naught and the slope of this slope of this line will reflect alpha 1.

$$Y_i = \alpha_0 + \alpha_1 x_i + u_i$$

So, this is our OLS fitted line. However the interpretation is not as simple in the logic class of models because essentially these are non-linear in parameters. Why this is important is because remember our logic function which essentially the probability that yi equal to 1 is reflected as 1 plus 1 upon e to the power minus zi where zi was a linear function of xi's this is a vector of various x's. So, it will be something like beta naught plus beta naught plus beta

1 x1 plus beta 2 x2 and so on up till beta n xn plus the error term. So, recall this is the form of function which was non-linear in nature and its behavior on xy axis is captured like this.

Like a s-curve where the limiting cases are 1 and 0. So, on the left side the limiting case is 0 and on the right side the limiting case is 1. Now such behavior which is non-linear in nature it requires high school mathematics to understand that the impact of this zi impact of this zi on yi varies. Here this x essentially it captures the dynamics of zi which is this function. So, here the impact of zi on yi varies depending upon where you are on this zi. For example, if you are here the if you want to compute the impact here recall our high school mathematics that the slope of yi that means the slope of this function dp let's call this function as pi dpi upon dzi this slope which is affected by the slope here on this curve measures the impact on yi at this particular point.

$$P(Y_i = 1) = \frac{1}{(1 + e^{-z_i})}$$
$$Z_i = f(x_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \mu_i$$

Similarly, if you want to measure impact on at any specific point you need to compute the slope. So let's say you want to compute the impact of a particular variable x2i on zi the most appropriate way to find that impact is the differential with respect to that particular variable and this is what we call as marginal effect. So, to put things in perspective in a in this kind of model it is not correct to say that one unit increase in x2i will cause beta 2 percentage increase in the probability of yi which we said in the LPM model. Remember in linear probability models LPM models we simply said that beta 2 the coefficient of that model yi equal to beta 0 plus beta 1 x. In case of linear probability modeling we said that this beta 1 is nothing but one unit increase in xi will result in beta 1 into 100 percentage increase rather I should write this as beta 2 and this as beta 1.

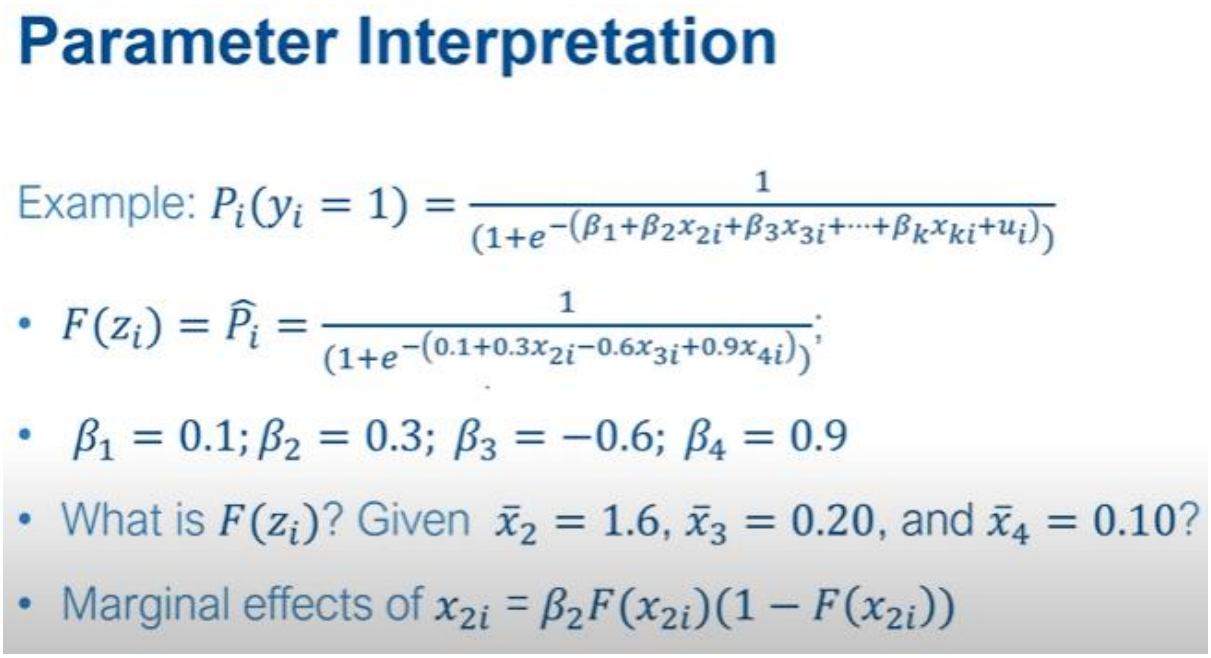
$$\frac{dP_i}{dx_{2i}} = \beta_2 F(x_{2i})(1 - F(x_{2i}))$$

Parameter Interpretation

Unlike LPM, it is incorrect to state that 1 unit increase in x_{2i} will cause $100*\beta_2$% increase in the probability of y_i = 1

- For logit model, we calculate $\frac{dP_i}{dx_{2i}}$; this works out to $\beta_2 F(x_{2i})(1 - F(x_{2i}))$ for the logit model
- So, a 1-unit increase in x_{2i} will increase the probability of y_i= 1 by $\beta_2 F(x_{2i})(1 - F(x_{2i}))$
- Usually, these marginal/incremental impacts are evaluated at mean values

So 100 into beta 2 times percentage increase in probability of yi equal to 1. This kind of statement we cannot make in the logic class of models and like we discussed in the just a few seconds ago in the previous slide this has to be calculated in this form dpi upon d of x2i which is nothing but the slope remember we computed at any point we computed this slope. This slope represents for a variable x2 this would represent dpi upon x2 dx2 where on y axis we have pi on x axis we have x2i variable. Now for logic class of models if one computes using that function 1 plus e to the power minus zi where zi was the linear function of variables x1i x2i and so on this works out to beta 2 times f of x2i where this is f of x2i and 1 minus f of x2i for the logic model. So, a one unit increase in x2i will result in the probability of yi equal to 1 by beta 2 times f of x2 into 1 minus f x2i where f of xi is the logic cumulative logic distribution function which we have discussed in previous series of videos in the lesson and these are called marginal or incremental impacts evaluated and these are often evaluated at mean values of these variables like x1i and x2i as we will see through a simple numerical example. So, have a look at this function of probability function logic function logic class of function where the value of zi is reflected with this expression this is essentially 1 plus e to the power minus zi where zi is captured by this expression.



$$F(Z_i) = \hat{P}_i = \frac{1}{(1 + e^{-(0.1+0.3x_{2i}-0.6x_{3i}+0.9x_{4i})})}$$

$$Marginal\ effect\ of\ x_{2i} = \beta_2 F(x_{2i})(1 - F(x_{2i}))$$

Now here the variable the parameters of beta 1 beta 2 and so on are already estimated here we can see them beta 1 is 0.1 beta 2 is 0.3 and so on. Now here if you are asked what is the value of f zi it is very easy to calculate we will calculate them at the mean values of variable let us say the mean values of x2 bar x3 bar x4 bar it is customary to compute the marginal effects at mean variable although you can compute at any given values. So at these mean values we will compute the value of f zi first and if you are asked to compute the marginal

effect let us say marginal effect of x2i then you will compute using this formula beta 2 into f of x2i into 1 minus f of x2i. So first we will compute the value of f zi at the mean values it works out to mean values that is x2i bar x3i bar and x4i bar at the mean values we will fill the mean values as we saw in the previous slide these mean values are available to us using these mean values we will compute the value of f zi which works out to 0.63.

Parameter Interpretation

Example: $F(z_i) = \widehat{P_i} = \frac{1}{(1+e^{-(0.1+0.3x_{2i}-0.6x_{3i}+0.9x_{4i})})} = \frac{1}{1+e^{-0.55}} = 0.63$

- Thus, a 1-unit increase in x_{2i} will increase the probability of y_i by 0.3*0.63*(1 − 0.63) = 0.07
- Similarly, for x_{3i}, −0.6*0.63*(1 − 0.63), and x_{4i}, 0.9*063*(1 − 0.63)
- Sometimes, these are also called marginal effects

$$F(Z_i) = \hat{P}_i = \frac{1}{(1+e^{-(0.1+0.3x_{2i}-0.6x_{3i}+0.9x_{4i})})} = \frac{1}{(1+e^{-0.55})} = 0.63$$

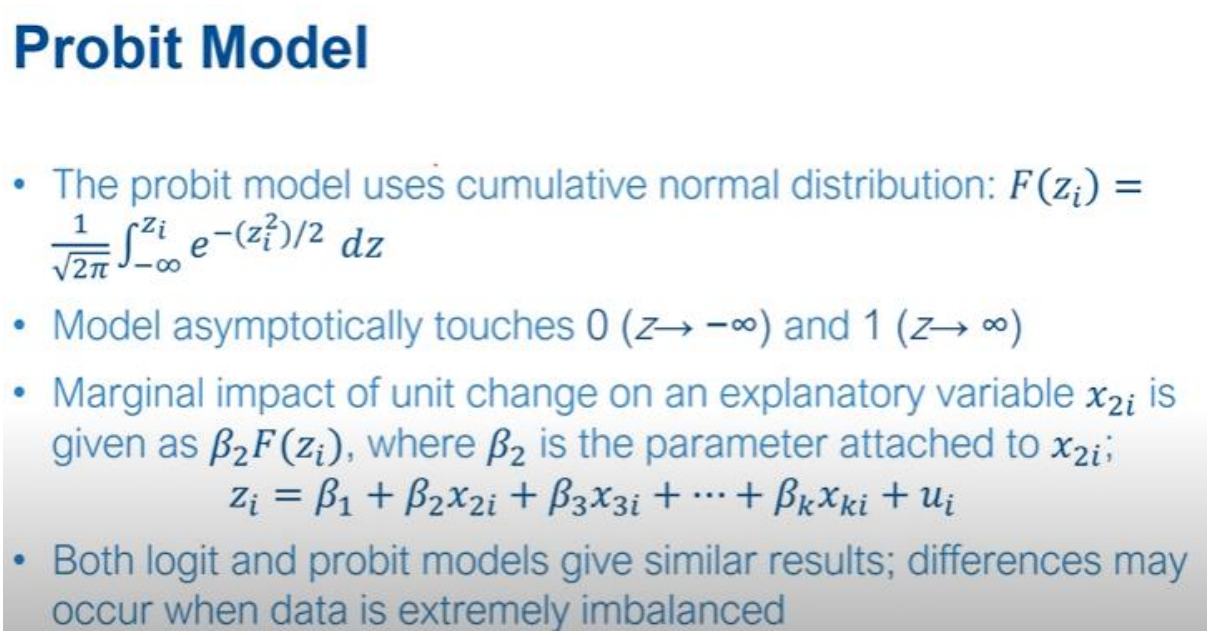
$$1\ unit\ increase\ in\ x_{2i}\ will\ increase\ the\ probabity\ of\ y_i\ by\ 0.3*0.63*(1-0.63) = 0.07, \quad Similarly\ for\ x_3 = 0.3*0.63*91-0.63) = 0.07, \quad etc\ \ldots$$

Now that we have mean values we can simply compute this expression beta 2 times f of x2i into 1 minus f of x2i which works out to beta 2 which is 0.3 into f of x2i this essentially works out to same as f of zi which is 0.

63 into 1 minus 0.63 so this number. Similarly for x3i we will multiply it with beta 3 other values will remain same 0.63 into 1 minus 0.63 for x4i it will be 0.9 into 0.63 into 1 minus 0.63. These are these what we are calling as dpi upon dx2i or dpi upon dx3i these are often referred to as marginal effects or incremental effects or the impact of coefficients one unit change in the coefficient the impact on percentage change in probability this can also be interpreted like that.

To summarize in this video we noted that unlike linear probability models or simple linear regression models the impact of individual variables like x2i, x3i and so on cannot be simply measured in the form of their respective betas like beta 2, beta 3 and so on. In the case of these logic class of models one needs to calculate the differential impact or what we call marginal effects or incremental effect in the form of this expression dpi upon dxi and the value for logic class of model works out to simply for a coefficient xi as beta i into 1 minus f of xi into f of xi and these marginal effects are nothing but simply the slope remember the behavior of this s curve the logic model s curve this is nothing but the slope of the curve at

any given xi at any given xi the slope where y axis is pi the slope is the marginal effect which is computed with this formula in the specific case of logarithm function. In this video we will briefly summarize three important topics which is probit model, maximum likelihood estimation MLE and goodness of fit measures.

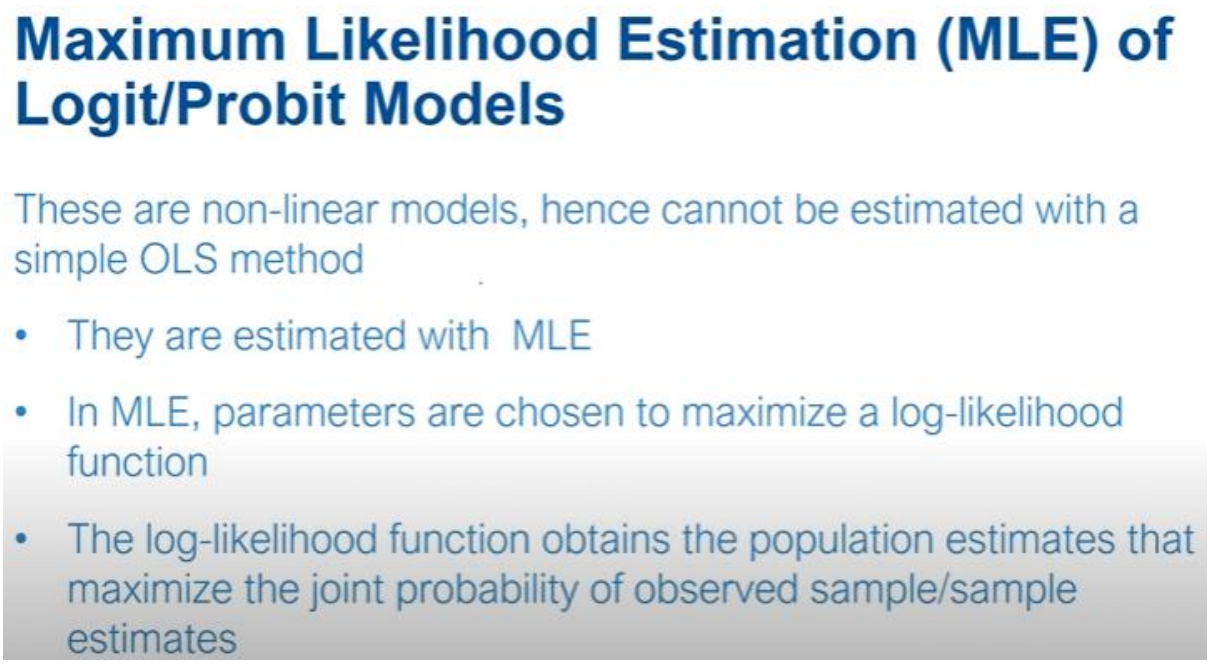


$$z_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_{ki} x_{ki} + u_i, \quad i = 1, \ldots, N$$

The mathematics can be ugly and out of the scope of discussion only a brief summary is warranted here. First let us quickly summarize the probit model. The probit modeling approach is very similar to logit model in fact it is almost identical with the only difference that the cumulative the function the distribution function is changed to normal distribution. So now our f zi which was the logit cumulative logistic distribution function earlier it is now cumulative normal distribution function. Remember the normal distribution function appeared somewhere like this starting it appeared like this and its cumulative this is the normal probability density function its cumulative function would appear something like s curve slightly different than the logit curve but still like s curve starting from 0 lower cut off is 0 and upper cut off is 1 because it is a probability density function. So the cumulative function would of course start from 0 and cut off at 1.

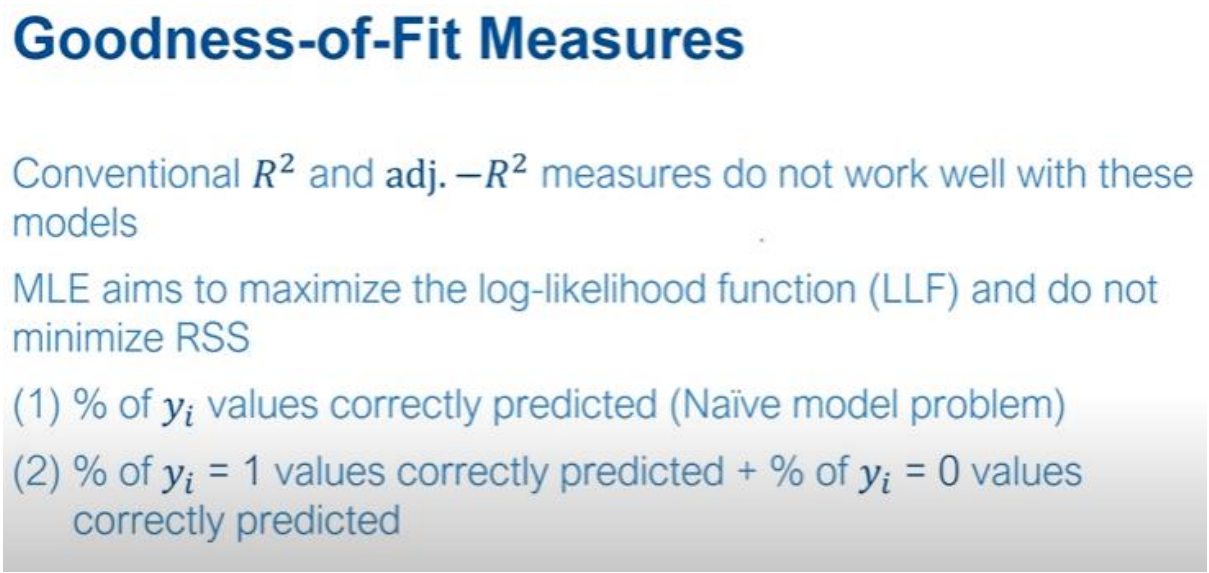
In functional form this appears like this f zi is equal to this complex looking function. However we need not remember that formula there are computer programs and as we will work on or it will be automatically clear. Now the properties are exactly same as logit function as z tends remember that z if z tends to minus infinity this function approaches to 0 and if z tends to plus infinity it approaches 1. So this is a very desirable property for this kind of modeling. Second the marginal impact the interpretation of marginal impact of xi on the probability function remains identical only when we derive this formula of marginal impact that is dpi upon xi the resulting formula is this one which is slightly different it is beta 2 times f zi where beta 2 is the parameter attached to x2i and remember what was here zi? zi was beta 1 plus beta 2 plus x into x2i plus beta 3 into x3 and so on.

Next we will summarize the maximum likelihood estimation of logit probit models. Recall that these are non-linear models in parameters and hence cannot be estimated with a simple OLS method. They are estimated with the help of MLE and in MLE parameters are chosen to maximize the log likelihood function as we will see and this log likelihood function obtains the population parameter estimates that maximize the joint probability of observed sample or sample parameter estimates. Let us discuss this in more detail what exactly this is and recall our discussion that essentially what we have is a set of population parameters that means we have a set of population regression function which maps the observed yi from a given set of variable x2, x3, x4 these are our variables which map to population regression function where parameters may be beta 1, beta 2, beta 3 and so on. However, we do not have the luxury of working with population but we work population which is a very large sample of data which maps from xi's to yi. We work with a rather small sample and we estimate what is called sample regression function SRF which tries to map these variables x1, x2, x3 and so on with our variable of interest yi as dependent variable.



But here what we estimate is beta 1 hat, beta 2 hat and beta 3 hat and so on. Now, please remember for each observation for each yi let us call it each yi observation which is part of our sample a set of parameter estimates may be there. So, each observation yi can be associated with the parameter estimates for all times remember we do not have the population parameters, we have sample estimates and for each observation a set of estimates may be there. So, for example, y1i, y2i these are different observations of the yi which map from these xi observations x1i, x2i and so on where a certain set of parameter estimates may be more likely and so on. Now for each observation yi we have a set of parameters let us call them the entire parameter let us call them beta vector which is a vector of parameters, let us call it beta k vector and for each observation y2 we have a different vector let us call it beta 1 for y2, beta 2 vector and y3 let us put it in capital V.

So, b1, b2 these are the these are vector of parameters for each observation. Now what we are doing here we are trying to find the probability we are trying to find the probability of observing this vector of parameter estimates and let us call it p1, p2, p3 for a given set of original parameter which is b there is a certain probability of observing this parameter estimate p1, p2 for this set of vector, p3 for this set of vector and then ultimately if you want to jointly maximize this probability of observing all this that is p1 what will be the joint probability of observing all these parameter estimates p1 into p2 into p3 and so on up till pn if you have an observations then this is the joint probability of having these observations and each observation maps to a particular set of parameters. Then essentially as a part of MLE what we are doing we are trying to find those population parameters, those population parameter vector b which maximizes the joint probability of observing the sample parameters. I repeat just to summarize essentially as a part of MLE we are trying to find those population parameters that maximize the joint probability of observing these sample observations where each observation corresponds to a set of likely parameter estimates probably p1 with beta 1, p2 with beta 2 and so on and this population parameter which we estimate from MLE method maximizes the probability joint probability of observing this sample data. In the last part of this video we will summarize the goodness of fit measure that are more useful for logit probit class of model. Remember that conventional R square and adjusted R square measures do not work well with the logit and probit class of model for the simple reason that unlike OLS, ordinary least square regression model, ordinary least squares where we minimize the residual sum of squares RSS of the error terms residuals unlike that MLE aims to maximize the log likelihood function that we just discussed which is the joint sort of joint probability maximization.

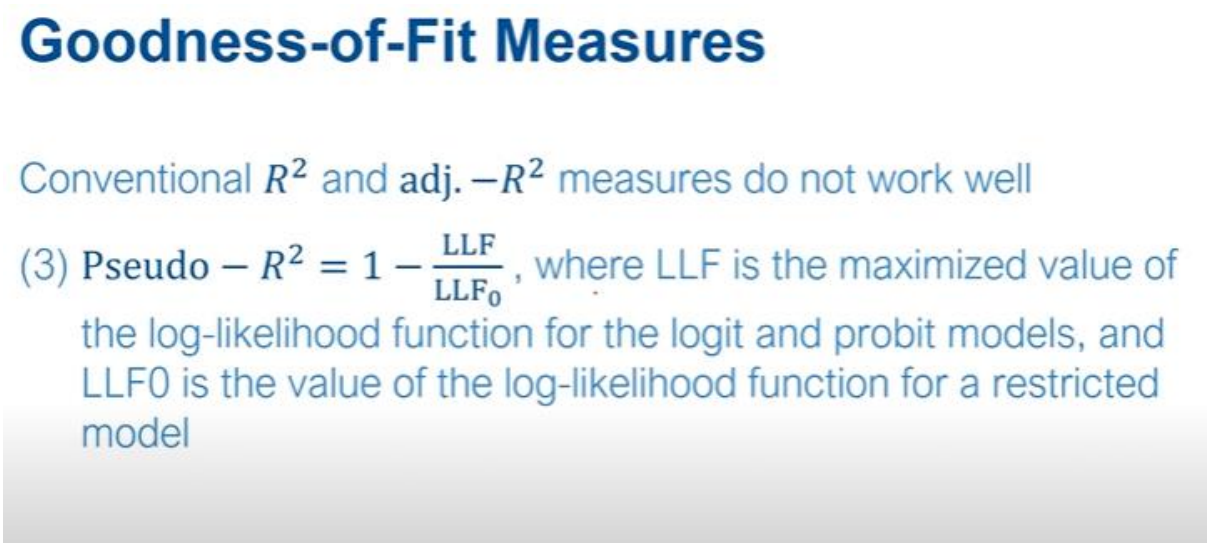


And therefore, the conventional R square and adjusted R square measures do not work well. So, there are two very simple and very intuitive measures that are employed by logit probit functions. One is the percentage of values that are correctly predicted. This is often referred to as naive model problem where you try to find how many what percentage of yi value you

are correctly predicting. So, there are total number of values that are correctly predict divide by aggregate number of total number of observations. The other way would be percentage of y equal to 1 that means positive cases that are correctly predicted and percentage of y equal to 0 cases that are correctly predicted and that is your accuracy of the model.

You can take some sort of average of percentage of y equal to 1 correctly predicted and percentage of y equal to 0 correctly predicted to reflect the accuracy of model. Obviously, you can change the weights also for example, if you feel that y equal to 1 correctly prediction is more value, then you give a higher weight rather than 0.

5, you can give it 0.7 and so on and then 0.3 to this. Similarly, if you feel this y equal to i equal to 0 is more desirable, then you give a higher weight to this one maybe 0.



$$Pseudo - R^2 = 1 - \frac{LLF}{LLF_0}$$

7 or 0.8 and the remaining 1 minus w to this one. So, if you assign 0.8 here then 0.2 here and so on so that w1 plus w2 the weights are 1. Also there is a very interesting measure which is pseudo R square which is employed. Pseudo R square formula is very simple 1 minus Llf on Lf0 where Lf is the maximized value of log likelihood function that we discussed in work p1 into p2 into p3 and so on up till pn is the joint probability that we maximize with respect to parameter beta i we try to maximize this function. So, this is the joint probability maximization for logit and probit class of models and Llf0 is considered to be the restricted model where it is assumed that all parameters are relevant that means beta to beta and jointly all these parameters are 0. So, the restricted function assumes or calculates these this joint probability for a model where all the parameters are assumed to be 0. So, that is your restricted model and then you compute for the maximize value for those parameters for those parameters beta 1, beta 2 and so on where this function this joint probability function of observing the parameters estimates is maximized.

So, Llf is the maximize value of log likelihood function and Llf0 is the likelihood function for a restricted model where all the parameters are assumed to be 0. Now this expression has a very interesting property and that goes like this. If your model is very poor, if your model is very useless that means all the beta 1 and beta 2 and beta 3 are not not contributing to the explanatory power in that case your Llf value will almost be as close as Llf0 that means your maximize value with L will be same as Llf0 because your parameters are not so useful. So, that means even whether you have these parameters in the model or do not the explanatory power of your model is not so good and therefore your Llf maximize as well as Llf0 for restricted model will be close to each other and therefore this value will turn out to be 0. So, this is one limit.

Another extreme case here is that your model is very well specified and this therefore this Llf value is extremely extremely large as compared to Llf0. In that case, please remember that this log likelihood essentially log this Llf is log likelihood function that means it is log of some probability number where probability is moved from 0 to 1. So, if your model is very well specified, let us say it is a very very good model where probability is almost inching towards 1 which is an ideal case where it is almost inching towards 1 then this log of this value Llf will be closer to 0 because the likelihood functions are probabilities and when probability is approaching 1 the log of likelihood this is log likelihood function this will approach to 0. However, because your Llf0 is very poor this value will be very very close to 0 assuming that relatively it is very poor it will be much closer to 0 and therefore essentially if you look at this number the log of this number will approach minus infinity. If this value as Llf0 approaches 0 this value will approach minus infinity and therefore the ratio will approach 0. While these are limiting cases, but you can imagine this Llf0 as long as this ratio this number inches this number the probability inside log of a number of probability which is relatively much higher as compared to the denominator this number will approach 0.

The idea is that in the case where the model fits well the log of probability will approach a reasonably large number while the denominator approaches to very large number and the magnitude of this number inches to 0 and therefore the upper case the upper limit case of this function becomes 1. In this pseudo R square very similar manner to R conventional R square or registered R square will range from 0 to 1. So this is how we use goodness of fit measures for the logit probit class of model which are slightly different from the conventional class of goodness of fit measures such as R square and registered R square. To summarize this lesson among supervised learning models classification algorithm is a very important tool employed in the finance domain for applications such as grade scoring of loan applications.

Classification algorithms are often implemented through logit probit class of models these are very simple yet powerful models. These models account for a number of shortcomings of linear probability models such as first non-normality and heteros-dicity of error terms. Second the values of dependent variable or the probabilities cannot exceed the 0-1 range and third the diminishing utility of conventional measures of goodness of fit such as R square measure. The limited dependent variable models such as logit model employ cumulative

probability functions such as logistic function. These models although non-linear are very useful for modeling limited dependent variables that are probabilistic in nature.

In the case of the logit model, the logit function is essentially the odds ratio. Since the estimated variable is in the form of probabilities the thresholding process is required to convert these probabilities into limited outcomes such as yes, no or 0-1. The conventional measures of goodness of fit such as R square are not very useful for such models. These measures are evaluated on their ability to accurately classify observations correctly. The receiver operator characteristic curve provides another useful tool to examine the efficiency of these models and also facilitates the selection of thresholding values.

Unlike simple linear models, the parameter estimates are interpreted in different manner as we discussed in the form of marginal impact. These marginal impact or marginal effects are computed to interpret the coefficients and their relationship with the dependent variable. Other models such as probit model remain very identical in all other aspects to this logit model except that a different cumulative probability function is considered. In case of probit model, normal distribution is employed. Since the model is non-linear in nature, ordinary least square or OLS cannot be employed for estimation and therefore maximum likelihood method is often employed to estimate these models. Thank you.