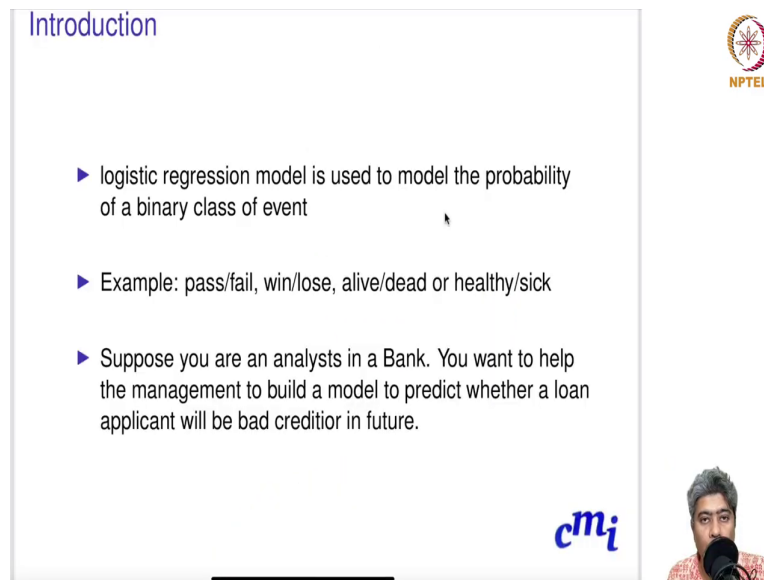


**Predictive Analytics - Regression and Classification**  
**Prof. Sourish Das**  
**Department of Mathematics**  
**Chennai Mathematical Institute**

**Lecture - 36**  
**Introduction to Logistic Regression**

Hello all, welcome to the part A of lecture 11. So, today I am going to start Logistic Regression. So, logistic regression model is used to model the probability of binary class of event.

(Refer Slide Time: 00:35)



The slide is titled "Introduction" in blue text at the top left. It contains three bullet points: "▶ logistic regression model is used to model the probability of a binary class of event", "▶ Example: pass/fail, win/lose, alive/dead or healthy/sick", and "▶ Suppose you are an analysts in a Bank. You want to help the management to build a model to predict whether a loan applicant will be bad creditor in future." The slide features the NPTEL logo in the top right corner, the CMi logo in the bottom right corner, and a small video inset of a man speaking into a microphone in the bottom right corner.

For example, if a student would pass or fail; if a player will win a tennis match or lose a tennis match; if a patient will be still alive by the end of the year or she or he will pass away before the end of the year, if a patient will be healthy or sick. Suppose you are an analyst in a

bank and you want to help the management to build a model to predict whether a loan applicant will be bad creditor in the future.

(Refer Slide Time: 01:20)

Motivating Example




▶ So you look into historical data  $\mathcal{D}$  on  $n$  existing customers in Bank's book

$$\mathcal{D} = \{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \dots, (y_n, \mathbf{x}_n)\},$$

where

$$y_i = \begin{cases} 1 & \text{Bad loan} \\ 0 & \text{Good loan} \end{cases}$$

$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$  covariates or predictor or features of  $i^{\text{th}}$  customer in the bank's book.



So, you look into the historical data  $\mathcal{D}$  of  $n$  existing customer in Banks book. So,  $y_1 \times \mathbf{x}_1$  is information about the first customer,  $y_2 \times \mathbf{x}_2$  is the information about the second customer,  $y_n \times \mathbf{x}_n$  is the information about the  $n$ th customer. Where  $y_i$  equal to 1 means it was bad loan. That means the loan was given and the customer did not return the money with interest on time. 0 means it is a good loan, so that means loan was given and customer return the money with interest on time.

(Refer Slide Time: 02:21)

So, determine whether data on a given customer in Bank's book

$$\mathcal{D} = \{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \dots, (y_n, \mathbf{x}_n)\},$$

where



$$y_i = \begin{cases} 1 & \text{Bad loan} \\ 0 & \text{Good loan} \end{cases}$$

*Edu age*  $\nearrow$   $\nearrow$  *previous loan, etc.*

$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$  covariates or predictor or features of  $i^{\text{th}}$  customer in the bank's book.

*cm*

Objectives



So,  $X_i$  is  $X_{i1}, X_{i2}, X_{ip}$  which are covariates of or predictors of or feature of the  $i^{\text{th}}$  customer in the bank's book. Now, feature could be you know this could be education, this could be age, this could be if a previous loan, if the customer has any previous loan etcetera etcetera. So, all these information bank do ask for.

(Refer Slide Time: 03:01)

Objectives

$$\mathcal{D} = \{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \dots, (y_n, \mathbf{x}_n)\},$$




where

$$y_i = \begin{cases} 1 & \text{Bad loan} \\ 0 & \text{Good loan} \end{cases}$$

$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$  covariates or predictor or features of  $i^{\text{th}}$  customer in the bank's book.

1 Which covariates has impact on  $y_i$ ? **Statistical Inference**

2 For a new loan applicant  $\mathbf{x}^0 = \{x_1^0, x_2^0, \dots, x_p^0\}$  - what is the  $\mathbb{P}(y^0 = 1)$ ? **Prediction**



Now, given this setup there are 2 kind of questions you will be looking for. First question, which covariate has impact on the  $y_i$ ? So, which covariate has impact or has an effect on deciding whether the it is a good loan is going to be a good loan or a bad loan. This kind of questions typically fall in the category of statistical inference.


And the second kind of question you will be interested in for new loan applicant given a  $\mathbf{x}$  naught, what is the probability of  $y$  naught equal to 1?  $y$  naught equal to 1 means it is going to be a bad loan, what is the probability? So, it is a prediction problem.

You want to predict what is the probability that a new customer will be turn out to be a bad customer, based on all these features that is being provided based on the financial status of the customer maybe last few years of IT return statement, maybe what is the education level

etcetera etcetera. Based on that you want to predict whether the customer will default or not. So, this will be belong to prediction problem.

(Refer Slide Time: 04:39)



Latent Variable


$$y_i = \begin{cases} 1 & \text{Bad loan} \\ 0 & \text{Good loan} \end{cases} \quad z_i : \text{potential score}$$

Equivalently, we can write

$$y_i = \begin{cases} 1 & z_i \geq 0 \\ 0 & z_i < 0 \end{cases}$$

$z_i$  is the unobserved latent score.



Now, in order to model this, we will bring the concept of latent variable. So, we are defining  $y_i$  equal to 1 if it is a bad loan and 0 if it is a good loan. We can also define it as  $y_i$  equal to 1 if  $z_i$  is greater than equal to 0 and 0 if  $z_i$  is negative or less than 0,  $z_i$  is some unobserved latent score ok. Sometimes  $Z_i$  is also known as potential score.

(Refer Slide Time: 05:32)

**Probit Model**

We can model  $z_i$  as  $z_i = x_i^T \beta + \epsilon_i$  *unobserved*

What we want to model: *linear regression we can not really fit a linear regression*

$$\begin{aligned} \mathbb{P}(y_i = 1) &= \mathbb{P}(z_i \geq 0) = \mathbb{P}(x_i^T \beta + \epsilon_i \geq 0) \\ &= \mathbb{P}(\epsilon_i \geq -x_i^T \beta) \\ &= \mathbb{P}(\epsilon_i < x_i^T \beta) \text{ by symmetry of the distribution} \end{aligned}$$

*$\epsilon_i \sim f \quad E(\epsilon_i) = 0$*

1. If assume  $\epsilon \sim N(0, 1)$  then it is known as **probit model** or **logistic regression with probit link**

$$\mathbb{P}(y_i = 1) = \mathbb{P}(\epsilon_i < x_i^T \beta) = \int_{-\infty}^{x_i^T \beta} \phi(\epsilon_i) d\epsilon_i = \Phi(x_i^T \beta)$$

*cmj*



Now, I can model this  $z_i$  as  $x_i^T \beta + \epsilon_i$  is unobserved remember that, but still, I can model it as  $x_i^T \beta + \epsilon_i$ . It looks like a linear regression kind of model, linear regression kind of model. However, my  $z_i$  is completely unobserved. So, I cannot really fit we cannot really fit linear regression. So, we cannot really fit a linear regression. So, but what we want to model is probability  $y_i = 1$ , that will be same as probability  $z_i > 0$ .

Why? Because if you go up, so what is  $y_i = 1$ ? Your  $y_i = 1$  is  $z_i > 0$  same as  $z_i > 0$ . So, I can say probability of  $y_i = 1$  is same as probability  $z_i > 0$ . Now, what is  $z_i$ ?  $z_i$  is  $x_i^T \beta + \epsilon_i$  ok. So that means, I can replace  $z_i$  by  $x_i^T \beta + \epsilon_i > 0$ .

Now, I can write it as  $\epsilon_i$  is greater than equal to minus  $x_i$  transpose beta. This is same as probability of  $\epsilon_i$  less than strictly less than  $x_i$  transpose beta. I am assuming that  $\epsilon_i$  is some residual, so it will have some distribution  $f$  with expectation of  $\epsilon_i$  equal to 0.

So, it will be some bell shaped distribution with mean at 0. Now, if you assume if you assume that  $\epsilon_i$  is normal 0 1, then the model is known as probit model or logistic regression with probit link. So, because probability of  $y$  equal to 1 you can write it as probability  $\epsilon_i$  less than  $x_i$  transpose beta, which is same as integration minus infinity to  $x_i$  transpose beta  $\phi(\epsilon_i) d\epsilon_i$ .

(Refer Slide Time: 08:31)

We can model  $z_i$  as  $z_i = x_i^T \beta + \epsilon_i$  *linear regression*

What we want to model: *we can not really fit a linear regression*



$$\begin{aligned} \mathbb{P}(y_i = 1) &= \mathbb{P}(z_i \geq 0) = \mathbb{P}(x_i^T \beta + \epsilon_i \geq 0) \\ &= \mathbb{P}(\epsilon_i \geq -x_i^T \beta) \\ &= \mathbb{P}(\epsilon_i < x_i^T \beta) \text{ by symmetry of the distribution} \end{aligned}$$

*$\epsilon_i \sim f$   $E(\epsilon_i) = 0$*

1. If assume  $\epsilon_i \sim N(0, 1)$  then it is known as **probit model** or **logistic regression with probit link**

$$\mathbb{P}(y_i = 1) = \mathbb{P}(\epsilon_i < x_i^T \beta) = \int_{-\infty}^{x_i^T \beta} \phi(\epsilon_i) d\epsilon_i = \Phi(x_i^T \beta)$$

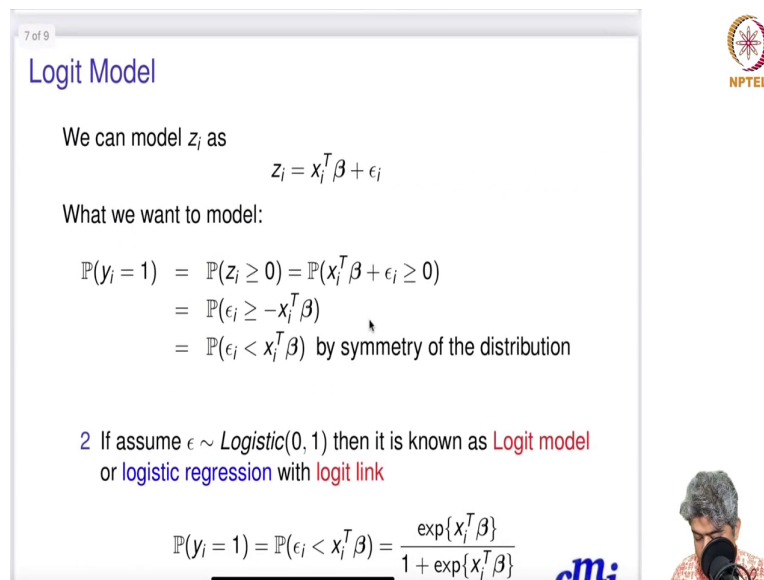
*pdf of standard normal* *Cdf of standard normal dis* **cm<sub>i</sub>**

Now, what is phi? Phi is the simple pdf of standard normal distribution pdf of standard normal distribution. And this is capital phi of  $x_i$  transpose beta. So, this is CDF of standard

normal distribution Cumulative Distribution Function CDF of standard normal distribution ok.

(Refer Slide Time: 09:08)



7 of 9

### Logit Model



We can model  $z_i$  as

$$z_i = x_i^T \beta + \epsilon_i$$

What we want to model:

$$\begin{aligned} \mathbb{P}(y_i = 1) &= \mathbb{P}(z_i \geq 0) = \mathbb{P}(x_i^T \beta + \epsilon_i \geq 0) \\ &= \mathbb{P}(\epsilon_i \geq -x_i^T \beta) \\ &= \mathbb{P}(\epsilon_i < x_i^T \beta) \text{ by symmetry of the distribution} \end{aligned}$$

2 If assume  $\epsilon \sim \text{Logistic}(0, 1)$  then it is known as **Logit model** or **logistic regression with logit link**

$$\mathbb{P}(y_i = 1) = \mathbb{P}(\epsilon_i < x_i^T \beta) = \frac{\exp\{x_i^T \beta\}}{1 + \exp\{x_i^T \beta\}}$$


Now, on the other hand instead of assuming epsilon i follow normal, if you assume epsilon i follow normal 0 1, if you assume epsilon i follow normal 0 1 then it is known as Logit model or logistic regression with logit link. Now, the interesting thing is for logistic distribution logistic 0 1 distribution, we know the CDF exactly. The cumulative distribution function is analytical form is known. So, for normal distribution the analytical form is not known, this analytical form is not known.

But so, we have to resort to some kind of the you know computational ability through R or Python. But for logistic distribution, the analytical form is exactly known and we can model



probability  $y$  equal to 1 or equal to probability  $\epsilon_i$  less than  $x_i$  transpose  $\beta$  is exactly  $e$  to the power  $x_i$  transpose  $\beta$  by 1 plus  $e$  to the power  $x_i$  transpose  $\beta$ .

(Refer Slide Time: 10:33)

$$\begin{aligned}\mathbb{P}(y_i = 1) &= \mathbb{P}(z_i \geq 0) = \mathbb{P}(x_i^T \beta + \epsilon_i \geq 0) \\ &= \mathbb{P}(\epsilon_i \geq -x_i^T \beta) \\ &= \mathbb{P}(\epsilon_i < x_i^T \beta) \text{ by symmetry of the distribution}\end{aligned}$$

2 If assume  $\epsilon \sim \text{Logistic}(0, 1)$  then it is known as **Logit model** or **logistic regression with logit link**

$$\mathbb{P}(y_i = 1) = \mathbb{P}(\epsilon_i < x_i^T \beta) = \frac{\exp\{x_i^T \beta\}}{1 + \exp\{x_i^T \beta\}}$$

*Handwritten notes:* "cdf of logistic dist.", "cmj"

Logistic Regression

This is the exact CDF of logistic distribution. So, this is optimum time called logit model or logistic regression with logit link.

(Refer Slide Time: 10:52)

Logistic Regression

$p = P(Y = 1)$

▶ Logistic Regression with **logit-link** : Sigmoid fn. (Binary)

$$\log\left(\frac{p}{1-p}\right) = x^T \beta = \beta_0 + \beta_1 x_1 + \dots + x_p \beta_p$$

▶ Logistic Regression with **probit-link**

$$\Phi^{-1}(p) = x^T \beta = \beta_0 + \beta_1 x_1 + \dots + x_p \beta_p$$

▶ How to estimate  $\beta$ ?

NPTEL

So, now, we have logistic regression with 2 variation. One is logistic regression with logit link and logistic regression with probit link. Now, natural question is how we estimate the beta? So, if you see these two model if you see these 2 model, you will see that effectively all you need to know that how can I estimate. If we know the value of betas, then all I have to do, I have to just plug in the x i's.

If I can plug in the x i's, it will give me the phi transpose X, phi transpose beta or log of p over 1 minus p and from there we can estimate the p. That is what we want to estimate. What is p? P is probability that Y i equal to 1, Y equal to 1, that is what we want to estimate. But we have to estimate beta.

So, that I am going to discuss that how to estimate beta that I am going to discuss in the next video. But before I end this video you may ask me another question that what about using any

other distribution? So, here we are assuming 2 distribution, epsilon can follow either logistic distribution or it can follow normal distribution. Of course, you can assume other distribution like Cauchy distribution and that will give you Cauchy link and other links are also available.

However, like extreme value link with through extreme value distribution. However, these 2 link logit link and probit link, these 2 links are most popular in any typical textbook these are the 2 mod link functions you will see. So, I am providing I am just for in the theoretical class, I am just providing you these 2 link, but you can have you can use any other link function that you want.

Sometimes logit link is also known as the sigmoid function. So, in some of the machine learning sigmoid function, sometimes it is called soft max function also soft max function also. In machine learning typically, they are either called sigmoid or soft max. Soft max is when you have multinomial kind of situation, we will come to it later multinomial, where as sigmoid function is known to for the binary situation like the one we are considering.

So, you can see all these variation in the terminology. So, I will stop here in the next video we will discuss how to estimate the coefficient of logistic regression.

Thank you very much. See you in the next video.