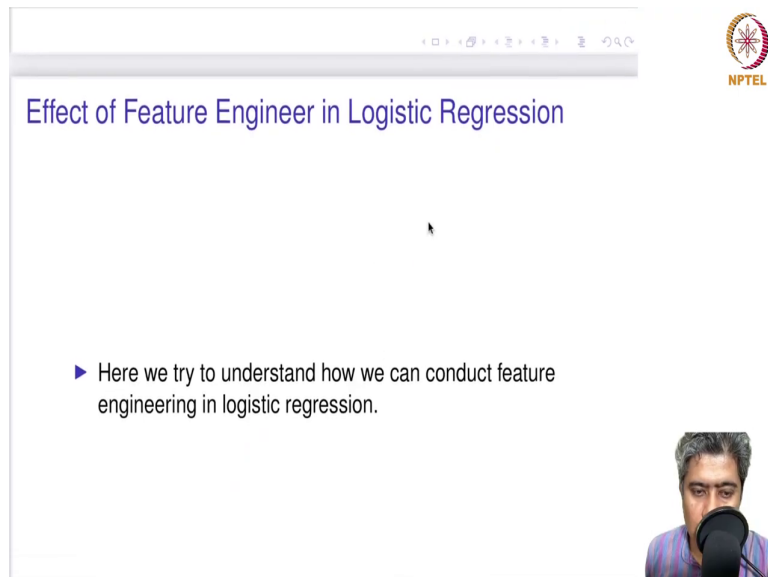


Predictive Analytics - Regression and Classification
Prof. Sourish Das
Department of Mathematics
Chennai Mathematical Institute

Lecture - 44
Effect of Feature Engineer in Logistic Regression

Hello all, welcome back to the Predictive Analytics Regression and Classification course. In lecture 13, in this video, we are going to talk about the Effect of Feature Engineering in Logistic Regression.

(Refer Slide Time: 00:30)




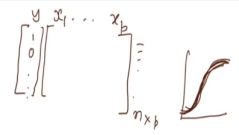
The image shows a presentation slide with a title bar at the top containing navigation icons. The title of the slide is "Effect of Feature Engineer in Logistic Regression". Below the title, there is a bullet point that reads: "► Here we try to understand how we can conduct feature engineering in logistic regression." In the bottom right corner of the slide, there is a small video inset showing a man with grey hair, wearing a blue and white striped shirt, speaking into a black microphone. To the right of the slide, there is a circular logo with a star-like pattern and the text "NPTEL" below it.

Here, we are try, we will try to understand how we can conduct feature engineering in logistic regression.

(Refer Slide Time: 00:42)

Logistic Regression



logit link / Sigmoid fn.



Consider data set $\mathcal{D} = (y_i, \mathbf{x}_i | i = 1, 2, \dots, n)$ (y_i, \mathbf{x}_i^T)

$$z_i = \mathbf{x}_i^T \beta_1$$
$$p_i(1) = \frac{\exp(z_i)}{1 + \exp(z_i)} \text{ and } p_i(0) = \frac{1}{1 + \exp(z_i)}$$
$$y_i = \begin{cases} 1 & \text{with } \mathbb{P}(y_i = 1) = p_i(1) \\ 0 & \text{with } \mathbb{P}(y_i = 0) = p_i(0) = 1 - p_i(1) \end{cases}$$

► β_1 is the $p \times 1$ matrix, *mathematical representation of logistic regression.*



So, typical logistic regression model will have a dataset D. So, typical it will have a, if you think from a sort of excel or data frame kind of thing, you have a y where y is like 1 0s or something like that and then you have bunch of features, x 1 to x p. So, these are your features and you have n such rows, 1, 2, 3 up to n rows and you have, so; that means, you have about n cross p, x columns and bunch of vectors of y 0 1s.

So, that is what I have written. So, the ith row, you have 1 y i and x i transpose. So, the i. So, now, once you have that, then z i equal to x i transpose beta and once you compute z i, then you put e to the power z i divided by 1 plus e to the power z i. This is sometimes in statistics; we call it logit link. In statistics, we call it logit link. In ML, we call it sigmoid function.

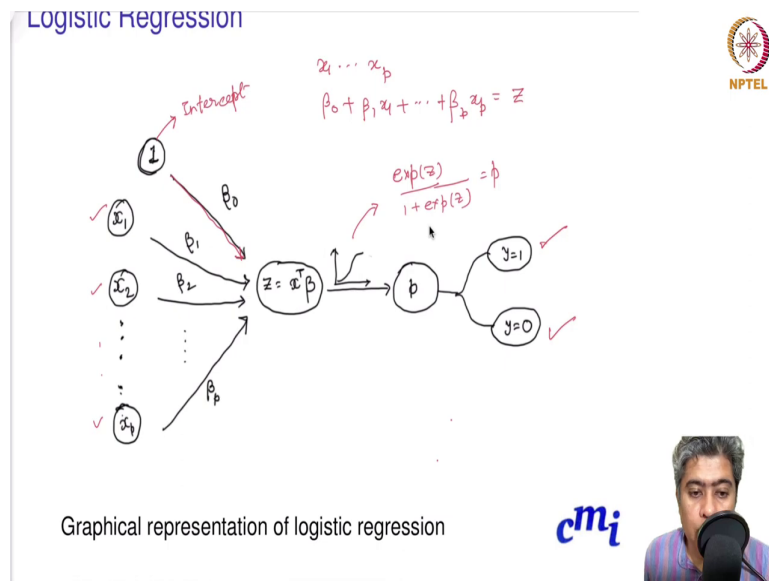
And why we call it sigmoid function? We will know, but we will know that y, ah, but I personally feel logit function is much more appropriate name because sigmoid function, there

are other functional form which also follow like a sigmoid function. So, sigmoid function because this behaves like this. Its kind of elongated s, this particular function, but there are other than this function, there are functions like probit also behaves like elongated function.

So, probit link can be called sigmoid functions, but in ML, particularly logit link is being called as sigmoid function which is not necessarily as an unique thing. So, I prefer to call it logit link. And then once you define p_i and p_i is effectively $1 - p_i$. And then for p_i , you observe y_i , value as 1 with probability p_i equal to 1 and 0 with probability $1 - p_i$.

So, here beta, this should be beta, beta is p cross 1. So, ok now this is a mathematical representation. This is mathematical representation of logistic regression.

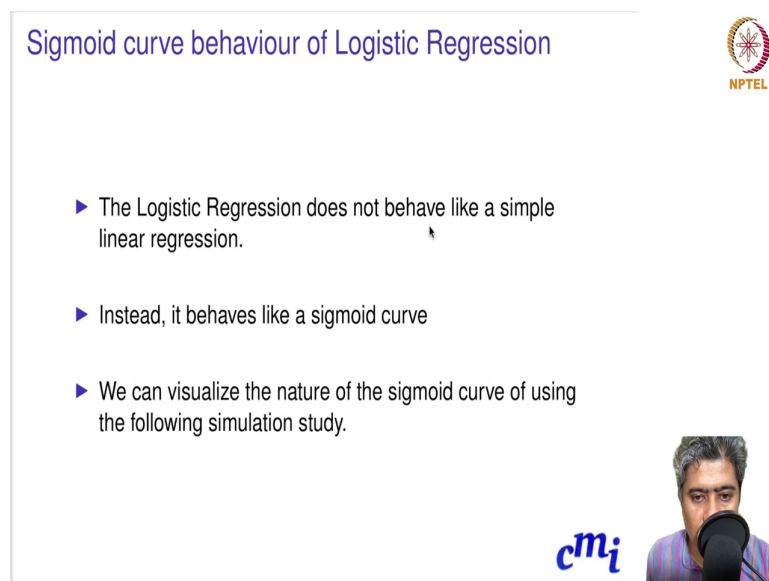
(Refer Slide Time: 04:13)



Now, the same logistic regression I am going to represent as a in a graphical point of view. Now, we have x_1 , let me use a different color. So, this is x_1 , this is x_2 dot dot dot x_p . So, we have x_1, x_2, x_p . And then each of their weight, these arrows represents their weight. And if I multiply them with their weight. So, β_0 , this is 1, this is intercept, this is intercept.

So, $\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$. Let us call it z . And then we are putting it into a sigmoid function, z is e^z by $1 + e^z$. That is my p , I am getting p . And then with some probability, I am getting observing 1 I am observing 0. So, that is my graphical representation of the model ok.

(Refer Slide Time: 05:45)



The slide is titled "Sigmoid curve behaviour of Logistic Regression" in blue text. In the top right corner, there is a circular logo with a star and the text "NPTEL" below it. The main content consists of three bullet points, each starting with a blue right-pointing triangle:

- ▶ The Logistic Regression does not behave like a simple linear regression.
- ▶ Instead, it behaves like a sigmoid curve
- ▶ We can visualize the nature of the sigmoid curve of using the following simulation study.

In the bottom right corner, there is a small video inset showing a man with grey hair speaking into a black microphone. To the left of the microphone is a blue logo that looks like "cmi".

So, now, we will try to understand the sigmoid curve behaviour of the logistic regression law. So, logistic regression does not have behave like simple linear regression. Instead, it behaves

like a sigmoid curve. So, what we will try to do, we will try to visualize the nature of the sigmoid curve using some followings simulation study ok.

(Refer Slide Time: 06:16)

$x \in (-\pi, \pi)$

- ▶ We consider the predictor variable x between $(-\pi, \pi)$.
 $x \sim \text{unif}(-\pi, \pi)$
- ▶ We simulate the latent variable z using the relation as $z = 0.01 + 0.45x + e$, where $e \sim N(\text{mean} = 0, \text{sd} = 0.3)$.
- ▶ Now we define the response variable $y = 1$, if $z > 0$, and $y = 0$, if $z \leq 0$.
- ▶ **Ask yourself** Is it a logit or probit model?

logit / probit
 (y, z)


So, we consider predictor variable x between minus pi and pi. So, x is a variable which takes value x ranges value between minus pi and pi. So, now what I am going to do, I am going to define z , where z is 0.01 plus 0.45 times x plus e , where e is some random number generated from normal distribution with mean 0 and standard deviation 0.3. Now, we define a response variable y equal to 1 if z is strictly greater than 0 and y equal to 0 if z is strictly less than 0 ok.

Now, ask yourself, pause the video for 5 minutes, ask yourself is this model, a logit model or probit model? I hope you tried it and you got the answer. So, it is actually a probit model because you see e here it follows normal distribution with mean 0 standard deviation 0.3 ok.



So, that means, it follows logistic regression with probit link ok. So, it is logist probit model ok. So, not logit it is a probit model.

(Refer Slide Time: 08:20)

Sigmoid curve behaviour of Logistic Regression



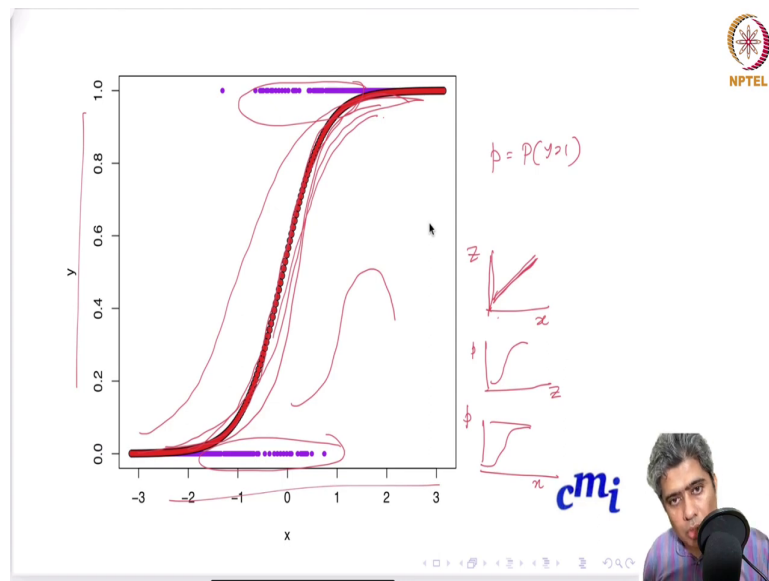
- ▶ Now we pretend actual response variable z is unobserved.
- ▶ The only data that we observe are x and y in D as `data.frame`.
D = data.frame(x, y)
- ▶ We model the relationship between the x and y using the logistic regression and we use the glm function in stats package of the R.



Now, we pretend that actual response variable z is unobserved, this is very important. In your real life, we will never see observe z. So, this is my data generation process. So, I am just simulated some random numbers between from uniform minus pi to pi, then using this model, I generated z and giving z, I generated y. Now, I will just pretend that I have only y and x and I will delete all the z values ok.

So, now, we will pretend as if we never observe z, we never observe z values. The only data that we have is x and y in D as some data frame right. So, we will just define some data dot frame x comma y as D ok. So, we model the relationship between x and y now using logistic regression. So, we will use glm function in the stats package of R.



(Refer Slide Time: 09:54)





So, if you do that, turns out that these purple colors point, these are actual observed y for different values of x , these are actual observed values. And this red color that you are seeing, this red color curve, this is actually your estimated p of probability y equal to 1 ok. So, you can see that this behave this p behaves like a elongated S, that is why its called sigmoid curve.

(Refer Slide Time: 10:36)

Non-monotonic Relation with Logistic Regression

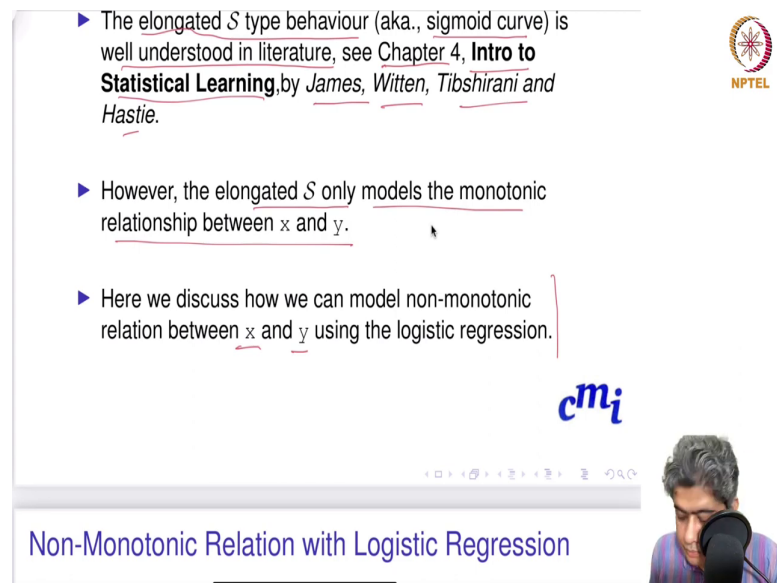


- ▶ The elongated S type behaviour (aka., sigmoid curve) is well understood in literature, see Chapter 4, **Intro to Statistical Learning** by James, Witten, Tibshirani and Hastie.
- ▶ However, the elongated S only models the monotonic relationship between x and y .
- ▶ Here we discuss how we can model non-monotonic relation between x and y using the logistic regression.



So, the non-monotonic relationship with logistic regression, the elongated S type behaviour also known as sigmoid curve is well understood in the literature. If you want to know more about the detail about it, then you can see the chapter 4 of Introduction to Statistical Learning by James, Witten, Tibshirani and Hastie ok. I will recommend this everybody to read this chapter, this is a beautifully written chapter.

(Refer Slide Time: 11:08)



The slide contains three bullet points:

- ▶ The elongated S type behaviour (aka., sigmoid curve) is well understood in literature, see Chapter 4, **Intro to Statistical Learning**, by James, Witten, Tibshirani and Hastie.
- ▶ However, the elongated S only models the monotonic relationship between x and y .
- ▶ Here we discuss how we can model non-monotonic relation between x and y using the logistic regression.

The slide also features the NPTEL logo in the top right, the CMU logo in the bottom right, and a small video inset of a man speaking into a microphone in the bottom right corner. The title 'Non-Monotonic Relation with Logistic Regression' is at the bottom.

However, elongated S only models the monotonic relationship between x and y . Here we discuss how we can model the non-monotonic relation between x and y using logistic regression.

(Refer Slide Time: 11:29)




Non-Monotonic Relation with Logistic Regression

$$p = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

- ▶ We can model the non-monotonic relation between x and y by using the higher-order polynomial, i.e.,

$$p = \frac{\exp\{\beta_0 + \beta_1 x + \beta_2 x^2\}}{1 + \exp\{\beta_0 + \beta_1 x + \beta_2 x^2\}} \quad (1)$$

- ▶ In Equation (1), the quadratic equation is being presented to model the relation between x and z .
- ▶ We can use higher order polynomial model to capture the underlying relationship between x and z .



So, the we can model the non-monotonic relationship. Now, why before going into the how non-monotonic? So, I hope you understood why I am calling it monotonic relationship because as x increases. So, what is the relationship between x and z ? The relationship between x and z , we are considering is straight line right. And what is the relationship between z and p ? z and p has a elongated x and that what we are seeing in x and p .

So, if you put that, that is this put this putting it into some sort of a you know elongated curve, but it is continuously monotonic function, it is not a non monotonic. Means, it been like going up and then coming down. It is always increasing constantly, it is increasing. It is bounded between 0 and 1 because it is probability, but it is continuously increasing, it is a monotonic function.


Now, the relationship between x and p not necessarily has to be monotonic. It can be in real data; it can be going up and then make them down. So, the elongated curve in the monotonic relationship. So, if you just use a simple linear logistic regression, it will only model the monotonic relationship between x and y .

So, now we will discuss how can we model the non-monotonic relationship between x and y . We can do it by using higher order polynomial. How? So, previously what we were doing? We were using beta equal to e to the power, say $\beta_0 + \beta_1 x$ divided by $1 + e$ to the power $\beta_0 + \beta_1 x$.


Now, what I am suggesting, why you stop here, you add the higher order polynomial. Here, I have added quadratic, you can add cubic or polynomial for the 4, 5, whatever you want, you can put it there in the equation 1. Quadratic equation is being presented to model relationship between x and z . We can use higher order polynomial model to capture the underlying relationship between x and z .

(Refer Slide Time: 14:19)

Non-Monotonic Relation with Logistic Regression

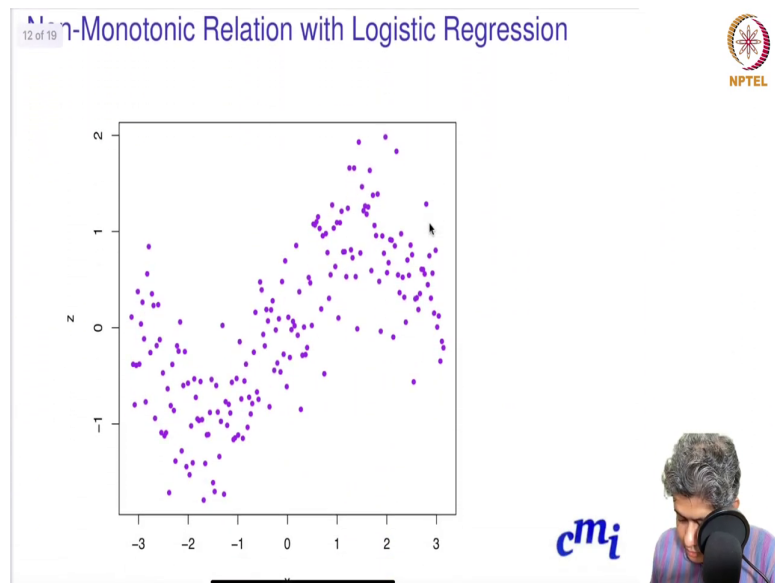


- ▶ First we assume the relationship between x and z are sinusoidal.
- ▶ We consider the predictor variable x between $(-\pi, \pi)$.
- ▶ We simulate the latent variable z using the relation as $z = \sin(x) + e$, where $e \sim N(\text{mean} = 0, \text{sd} = 0.5)$.



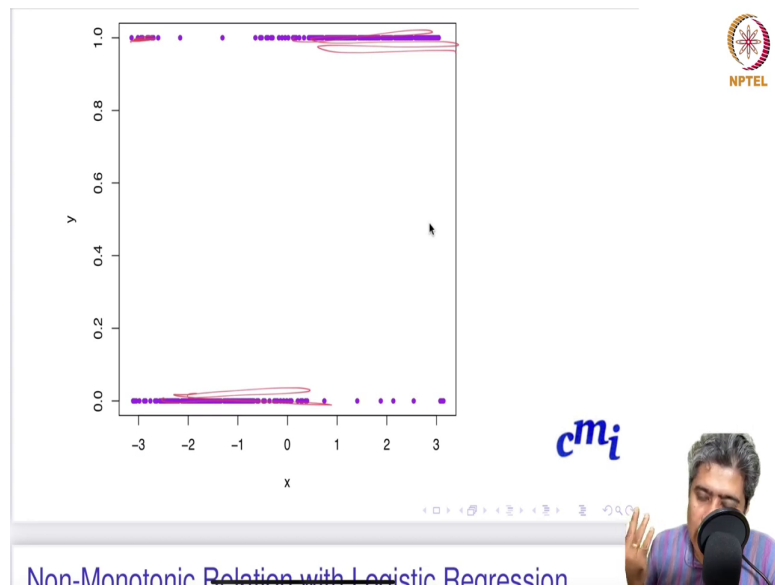
So, the non-monotonic relationship with logistic regression, let us try to capture this behaviour. Let us assume that relationship between x and z are sinusoidal. So, we consider the predictor variable between x of x between minus pi and pi. So, we simulate the latent variable z as z equal to $\sin x$ plus error. So, error is still following normal. So, 0 with standard deviation 0.5, but the relationship between z and x is for sure not linear, its a sinusoidal.

(Refer Slide Time: 15:03)



So, how it is? That is how the relationship between z and x . The latent space has a very non-linear relationship.


(Refer Slide Time: 15:16)





Now, if you when from z , we simulate the y 's and we plot that is how it looks like, that is how it looks like. So, we had some values here, then we got bunch of values here and then we got lots of values here. So, because now what is either all values of z is being converted into y as either 0 or 1. that is all and in real life, we do not know what is the 0, now z values. All we have is the values of y which write as 0 or 1.

(Refer Slide Time: 15:56)

Non-Monotonic Relation with Logistic Regression

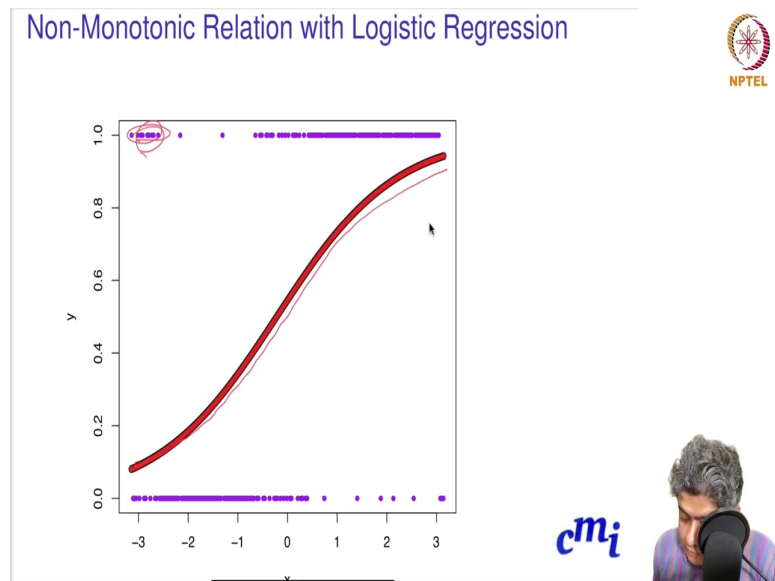


We pretend that we do not know the true relationship between x and z and we fit simple logistic regression:

$$\text{logit}(p) = \beta_0 + \beta_1 x.$$
$$p = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$



So, we pretend that we do not know the true relationship between x and z , ok. And we simply fit a logistic regression of simple $\text{logit } p = \beta_0 + \beta_1 x$. So, that is $p = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$ ok.

(Refer Slide Time: 16:30)





Now, if you fit that, this is fitting a simple non-monotonic, elongated x curve. Fine monotonic relationship, its giving you monotonic relationship, but though we know we have quite a few points here and the underlying relationship is very non-linear.

(Refer Slide Time: 16:53)



- ▶ We see that simple logistic regression models non-linear but monotonic relationship between p and x
- ▶ Now we fit a cubic relationship between x and z , with
$$z = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3,$$

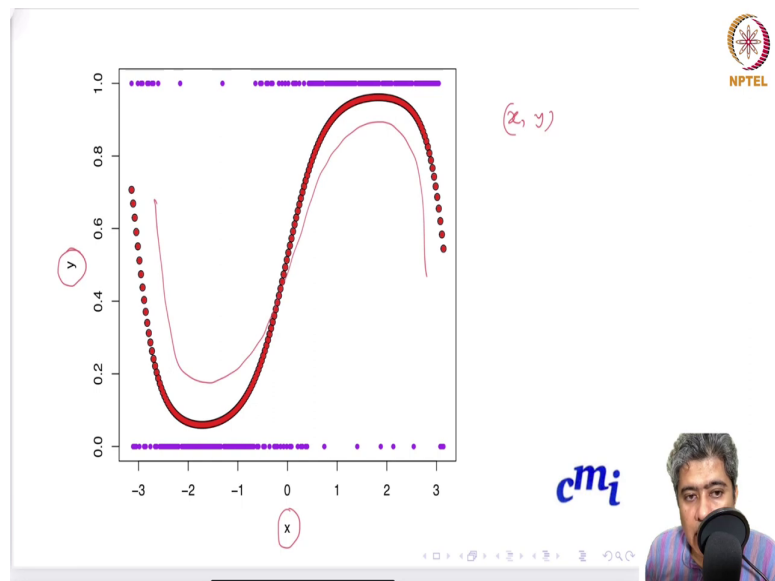
i.e.,

$$\text{logit}(p) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$


So, we thought like, we see simple logistic regression models non-linear, but non-monotonic relationship between p and x . Even if between x and z has a non-monotonic behaviour like you know quadratic cubic behaviour sort of sinusoidal behaviour. So, we fit a cubic relationship between x and z , x and p , x and p ok. And we decided to go for this beta naught plus beta 1 x plus beta 2 x squared plus beta 3 x cube.

And hence the final model is this logit p equal to beta naught plus beta 1 x plus beta 2 x square plus beta. So, final model is logit p equal to beta naught plus beta 1 x plus beta 2 x square plus beta 3 x cube.


(Refer Slide Time: 17:57)



And when we fit that model, the p is now nicely non-monotonic and capturing the sinusoidal behaviour between x and y . So, we never used z here, remember that all we have is only x and y . Just modeling the higher order feature, we put couple of higher order feature x square and x cube and the model logistic regression able to model the sinusoidal behaviour of relationship between x and y .



(Refer Slide Time: 18:43)

Non-Monotonic Relation with Logistic Regression



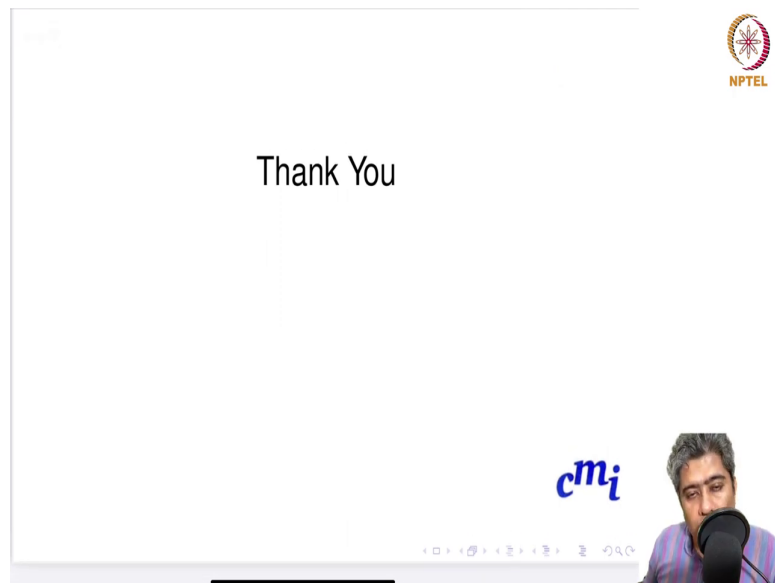
Summary:

- ▶ Above demo shows that we can use "**feature engineering**" technique to capture the non-linear and non-monotonic relationship between x and p (y)
- ▶ The "**feature engineering**" typically helps increasing the out of the sample model accuracy.
- ▶ However, we should be careful about the over fitting.



So, this is, so, ever do not show that we can use feature engineering technique to capture the non-linear and non-monotonic relationship between x and p or y . The feature engineering typically helps increasing out of the sample model accuracy. However, we should be careful about the over fitting because when you put more and more features in your model. So, naturally you have a chance that you end up over fitting the model ok.

(Refer Slide Time: 19:24)



So, thank you very much, see you in the next video.