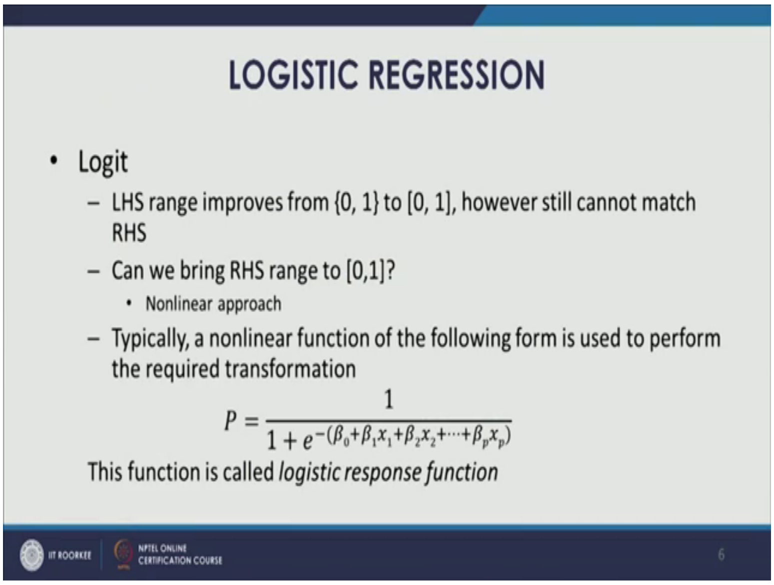
Business Analytics & Data Mining Modeling Using R
Dr. Gaurav Dixit
Department of Management Studies
Indian Institute of Technology, Roorkee

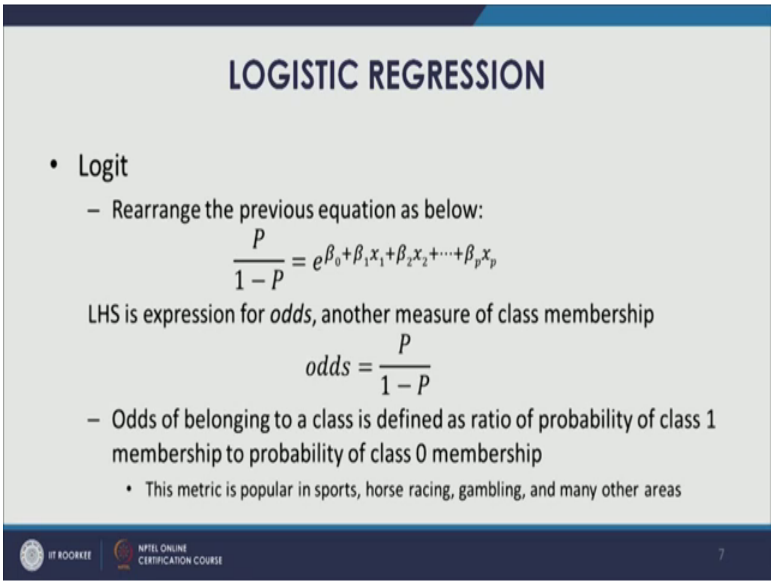
Lecture - 47
Logistic Regression - Part II

Welcome to the course business analytics and data mining modeling using R. So, in a previous lecture, we started our discussion on logistic regression. So, we talked about logistic regression model, the standard formulation, we talked about different issues; for example, categorical outcome variable and inability to you know model it as a function of linear function of predictors, then we discussed further steps how the standard logistic regression formulation is actually done. So, we talked about this probability you know this particular probability model, right.

(Refer Slide Time: 00:56)



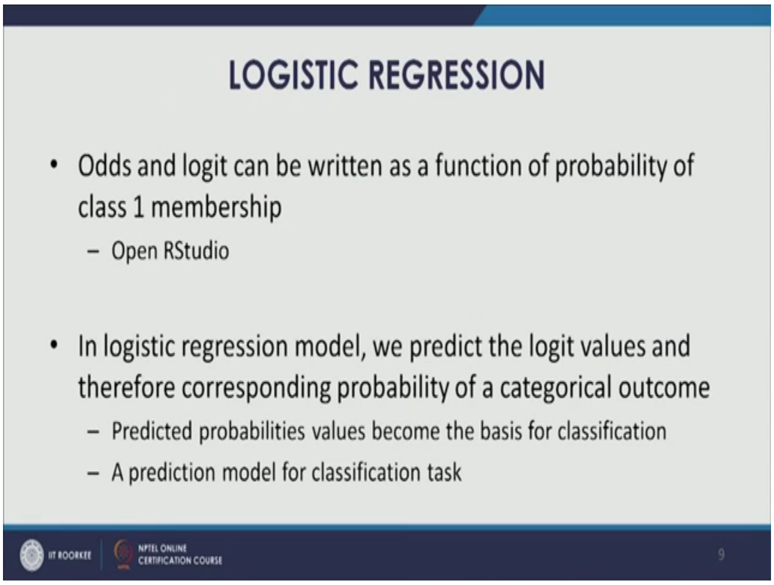
(Refer Slide Time: 01:01)



That is a using logistic response function, then we looked at this Odds model this one; the Odds a Odds Odds model as a function of these a you know exponential and then the power and in the power term the a linear a function of predictors, right. So, this was the Odds model, then Logit model that is the last one that is the standard formulation log Odds also called Logit.

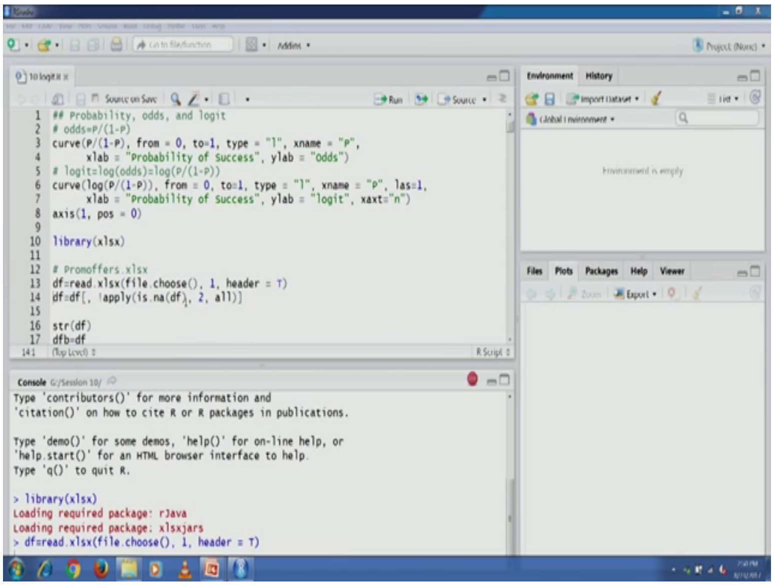
So, Logit equal to a beta 0 plus beta 1 x 1 up to beta p x p. So, we talked about these 3 models and the log Odds are Logit being the standard formulation for logistic model. So, this Logit model the log Odds model that is used to estimate a parameters coefficients beta zeroes and then from this we can always derive Odds model and also probability model. So, we also did a small exercise in R where we looked at the where we generated few plots and looked at the looked at the relationship between Odds and Logit and probability and we also did a one simple logistic regression model where we where we build a model a between to promotional offer being the outcome variable and income being the single predictor right and we also plotted the we also plot at the a fitted model in previous lecture.

(Refer Slide Time: 02:37)



So, let us so, let us go back to our the same the point where we stopped in the previous lecture.

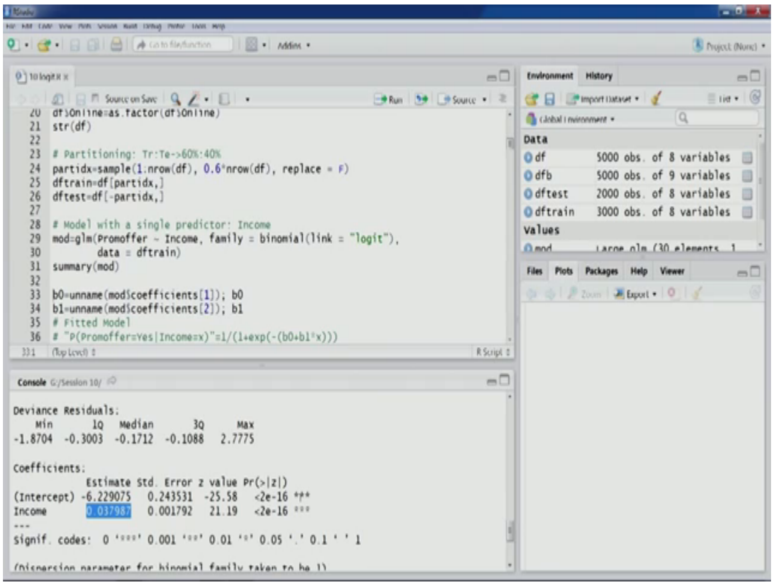
(Refer Slide Time: 02:45)



So, let us load the data set again. So, this is the package, now let us promotional offers is the data set that we had used in previous lecture. So, let us reload it. Now in the previous lecture, we stopped at the plot, right. So, we will start our discussion from the same point will also later on, we will also a develop the model build build the model using all the predictors and then we will discuss some of the interpretation related issues that we generally face in logistic regression model.

So, data is now loaded; let us remove NA columns; this is the structure of data.

(Refer Slide Time: 03:38)

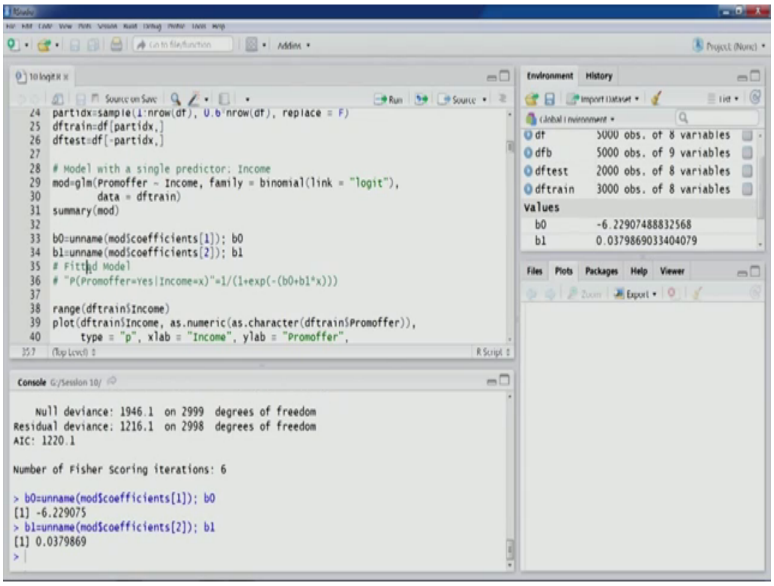


And we would not like to; let us take a backup. Now, we would not like to use pin code variable. So, let us get rid of this then let us convert promotional offer and online to a factor variable and then this structure. So, these are the variables that we use in previous lecture proper partitioning. Now, we talked about the function GLM. So, GLM is the function that can be used to build a logistic regression model in R.

So, let us build this. So, we had done up to this in previous lectures we can see the estimate for the constant term that is intercept and for income as well. So, both being significant right using these two you know a parameters; these two estimates of parameter estimates, we can specify a as we talked about in previous lecture we can specify our probability model in this fashion because the main idea is to estimate the probabilities as we talked about in previous lecture.

The first step is to estimate the probabilities values and then a use that to perform our classification in the second step. So, we can extract these a parameters b 0 for constant term b 1 for the income in this fashion, we can use the unnamed function to remove the that row names that that will have an mod dollar coefficient and then extract the individual values for constant term and income as per this particular code.

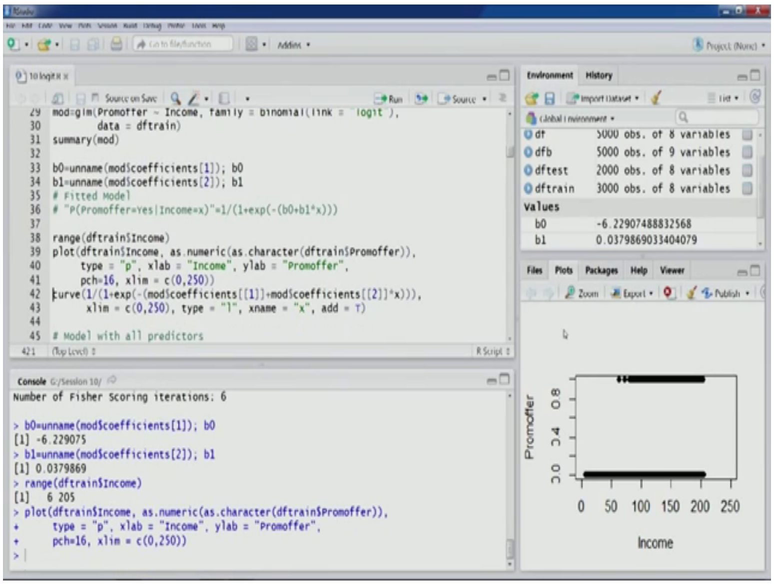
(Refer Slide Time: 05:16)



So, once this is done you can see in the commented out section, here the fitted model can be written in this fashion. So, you can see here the fitted model; the probability model can be can be written in this fashion p promotional offer, yes given the income variable that is x and in this fashion we can express a one divided by one plus exponential of minus b 0 minus of b 0 plus b 1 into x.

So, x is a income. So, b 0 and b 1 values; we have already estimated using logistic regression model. So, now, these values can be plugged in into this particular expression and for any value of x that is income in this particular case, we can compute the corresponding probability value using this particular expression right.

(Refer Slide Time: 06:20)

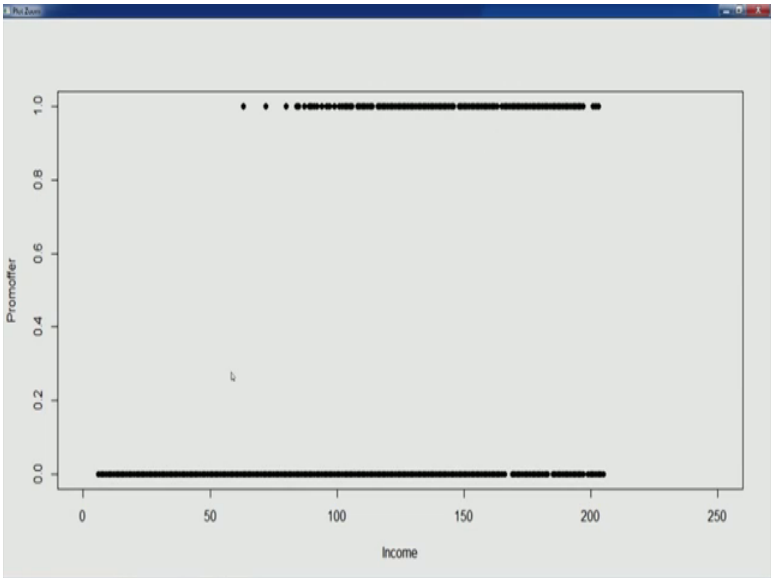


So, when we in the previous lecture then we moved forward to creating a plot for the same. So, let us do the same thing again. So, range is 6 to 205. So, appropriately will be capture this through x limit argument and this particular scatter plot is between income and as you can see here the promotional offer.

Now, the promotional offer is actually the categorical variable as we had converted into a categorical variable now we want to plot this particular variable in a in a scatter plot then we will have to convert it into a numeric variable and when we do. So, we will have just 2 values 0 and 1. So, this is how we are trying to a convert it back to numeric variable. So, you can see promotional offer. Now, first it is being converted, it is being coerced into a character vector using as dot character function and then once that is done this character vector is now being coerced into as dot a numeric this is to get the desired result right for desired result for example, numeric variable we want to get the numeric variable for promotional offers.

So, once this is done we can plot this. So, you can see type of plot is p; that means, we just want to plot the points. So, x and y values a values corresponding corresponding to x axis and y axis, we just want to plot them. So, we are not plotting a you know a line rather we are plotting points. So, let us create this plot.

(Refer Slide Time: 07:57)



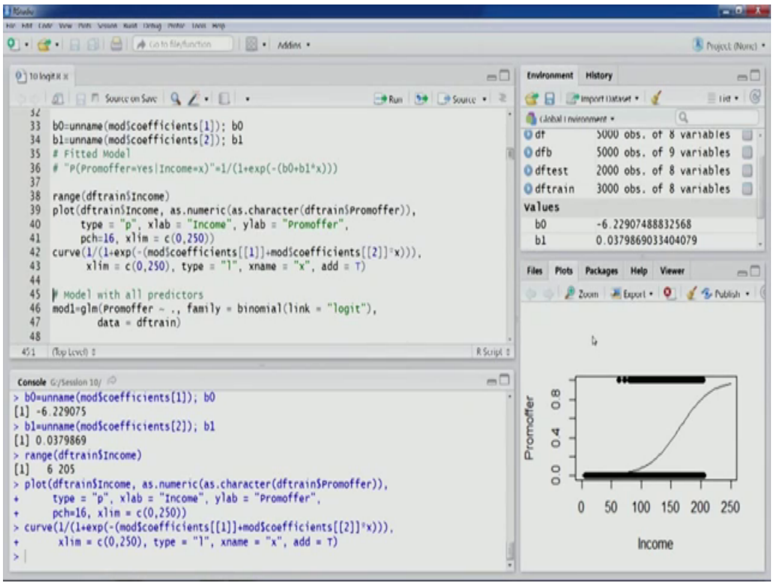
So, we create this part, we see that because the probability promotional offer, it takes just 2 value 0 and 1. So, some of the observations they are at 0, you can see here you know majority of observations are at this level 0 level 0 value and then the remaining observations they are at 1. So, all the points and they seem to be you know on these two extreme values 0 and one because promotional offer offer while you categorical variable we converted into a numeric form 0 and 1. So, just two extreme values for all the points are depicted here.

Now as we talked about as we talked about that and we just saw also that the probabilities value there in essence become because we use the logistic response function. So, the relationship is non-linear because of this that non-linear function is being used to fit these points. So, essentially if we consider promotional offer as numeric and income is already numeric variable and just now we have created this scatter plot now we try to fit this particular plot using our logistic response function right that is the non-linear.

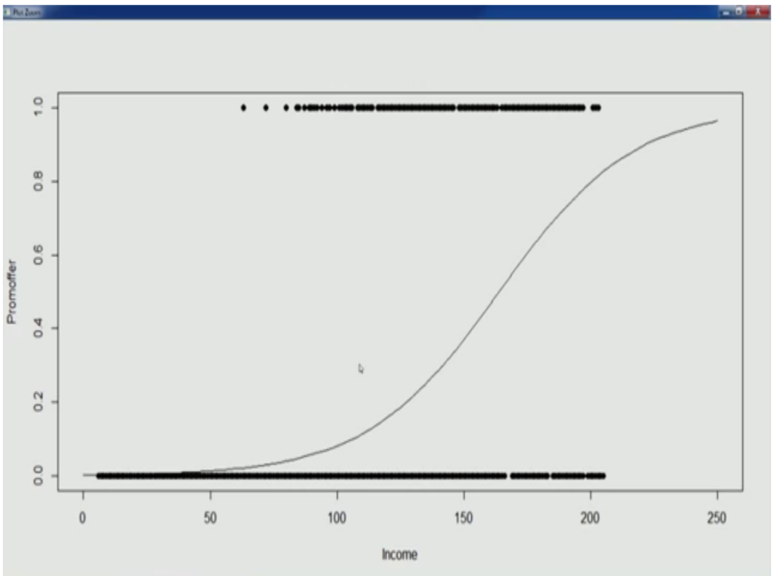
So, this is how we can do we can add this particular curve and the estimates the b, b 0 beta 0 and beta 1, we have already computed we have already estimated these parameters. So, these are directly being used using these expression mod dollar coefficient one and mod dollar coefficient two using double bracket. So, that we you know a do not have to worry about that names that are there.

This is another way to un name unnamed the those variable names and the coefficient result and then for x is the variable that is going to be varied right. So, in the x name argument you can see x is mentioned and this is going to be added to the existing plot right add argument is there add is true. So, to in the existing plot this particular curve is going to be added. So, let us create this like we did in previous lecture as well.

(Refer Slide Time: 10:10)



(Refer Slide Time: 10:15)



So, now from this plot, what we are trying to a do here is we first create a scatter plot between outcome variable of interest that is promotional offer versus the income the single predictor and then we have added our added the fitted curve. So, the fitted curve here is this is logistic in a small function and expression and estimates, we have already seen for this.

So, now you can see some part of these 0 values; 0s are you know are or on this plot and then this plot is as we as the values of income increase this probability value keeps on increasing. So, these probabilities value are the one that we are going to use to classify the observation into class 0 or class 1, right. So, this is the log plot that we are using to fit the data that we have right. So, this one eventually as the value income value increases further to the right side this particular plot is approaching the one.

So, as you can see as you can see that after you know first few observations belong into class 0. So, they are you know they are fitting to the, to this particular non-linear curve and in and there might have been if there have been a few observation on this particular top right corner. So, they might have also been fit into the curve as it approaches one, but in this case more or more of them are on this side, right.

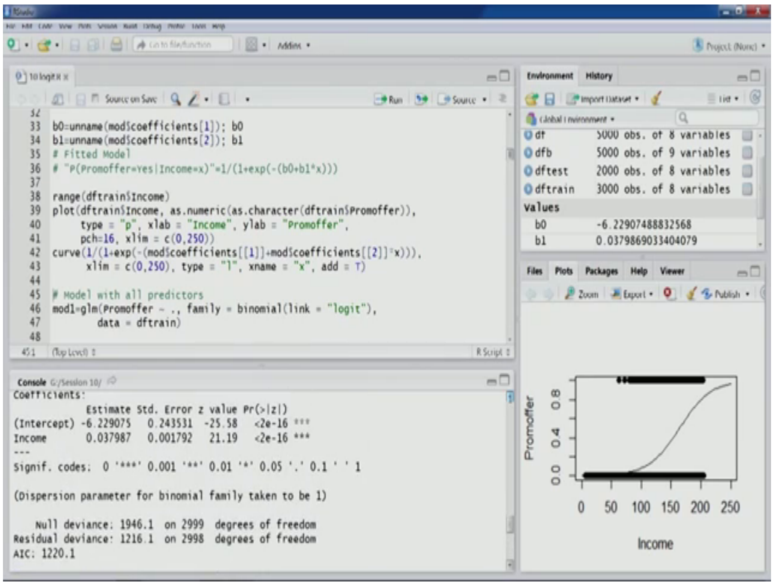
So, in the another problem and other data set you know some of the points could have been here as well. So, some observations as the this particular non-linear curve is approaching one right fit to the curve some observations as it this particular starts from 0 right you know we all might be fitted by this curve. So, the remaining observation they will a lie in between you know.

So, now, for new observations when we estimate probabilities value using this logistic response function; so, we will have to use cutoff as you can see from here to classify these observations in into either class 0 and class 1. So, I from this particular plot, we can also understand other thing for example, income is just one predictor that we have just that we have modeled in our single you know predictor logistic regression model.

So, in case it had been you know in case we had more. So, the x axis in a way is also representing the Logit values a that a and therefore, this particular and this logistic response function is actually the probabilities value. So, this is also this particular curve also giving us the indication about the relationship between when Logit is on y axis and the probability values on in Logit is on x axis and the probabilities value on y axis, what is going to be the relationship.

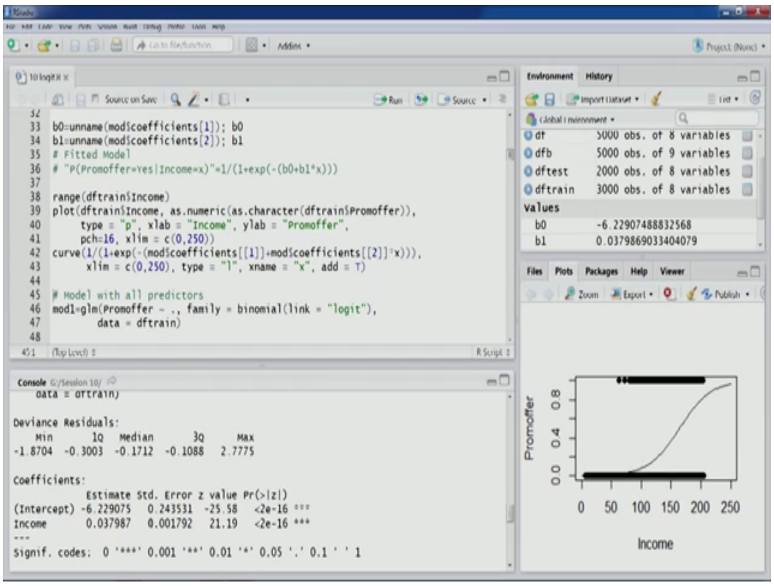
So, as the Logit value values increase along the x axis the probabilities values also increase. So, let us come back to our discussion here in the slides. So, in logistic regression model we predict the Logit values as we saw in the output.

(Refer Slide Time: 13:52)



So, let us scroll the output I once again; so, this was the single predictor model that we had.

(Refer Slide Time: 13:56)



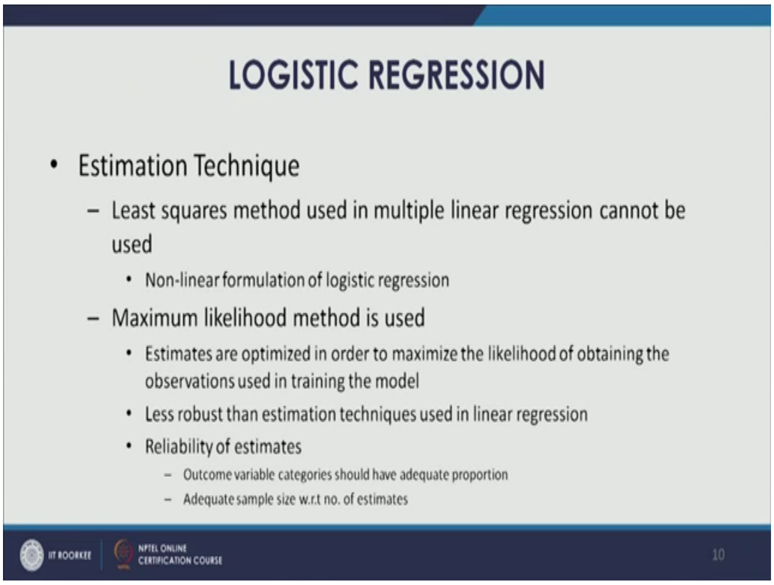
So, we predict the Logit values and therefore, corresponding probability of a categorical outcome right. So, using those a i output values and using the logistic response function a we can use the same parameters and compute the probabilities value. Now these predicted probabilities value become the basis for classification right first step was to estimate the probability now once we have these probabilities value they can be used for classification. So, in essence, if we look at the logistic regression model essentially we are fitting a model just like a multiple linear regression model right we have Logit values which can we can is considered as a numeric numeric outcome variable and the on the and it is being fitted as a linear function of predictors just like linear regression.

So, in a way we can define logistic regression as a prediction model for classification task. So, we first fit a prediction model. So, the outcome variable is instead of you know the category using the categorical outcome variable we are using a form a form of you know a particular function based on that categorical outcome variable that is Logit and that Logit variable is actually being you know modeled as a function of linear function of predictor. So, essentially we build a prediction model and, then that is based on using that those values we compute the probabilities and then use it for classification tasks.

Now, let us talk about the estimation technique. So, the estimation technique that we use in multiple linear regression is a typically ordinary least square ols. So, a that method cannot be used in this particular case because essentially the logistic regression a the formulation is non-linear non-linear it is a non-linear formulation because we are looking to estimate probabilities values through a logistic response function which is which is a non-linear function therefore, least squares a estimation method has not been a not being considered as an appropriate technique to estimate parameters in logistic regression model..

So, the another method which is called maximum likelihood method that is used to estimate parameters. So, this idea about a main idea behind this particular maximum likelihood method or m l m is actually to estimates to find out the estimates in order to maximize the likelihood of obtaining the observations using the training the model, right.

(Refer Slide Time: 16:40)



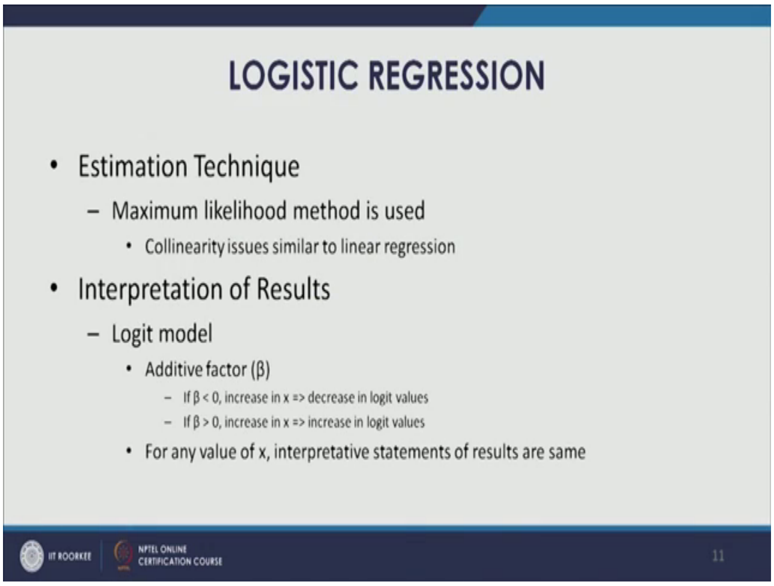
So, the estimates are we will we look to compute estimates which are going to increase the like likelihood of obtaining the observations if you using a those particular estimate itself.

So, of course, this particular method will require a number of iteration to reach to those estimates. So, this particular method is used a as you can see a from the main idea itself this particular method looks a less robust than the estimation technique that we use the ordinary least square least square method that we use in linear regression.

So, m l m maximum likelihood method, it is less robust than that and reliability of estimates for MLM also depends on a number of things for example, outcome variable a categories; this would have a adequate proportion. So, you know probably logistic regression model is not suitable for situations where we have a very few records a in the training partition belonging to class 1 or class 0. So, we would like to have adequate number of records in all the classes and then the you know reliability of estimates would be improved adequate sample size with respect to number of estimates.

So, quite this particular technique is going to be sensitive because of the estimation method we are using maximum likelihood method MLM. So, because of that we would require an adequate sample size with respect to number of parameters that we need to estimate. So, number of depending on the number of coefficient we require adequate sample size. So, that the estimates that we get that those are reliable now some of the some of the issues that we encounter that we face in logistic regression are similar to a linear regression for example, co linearity.

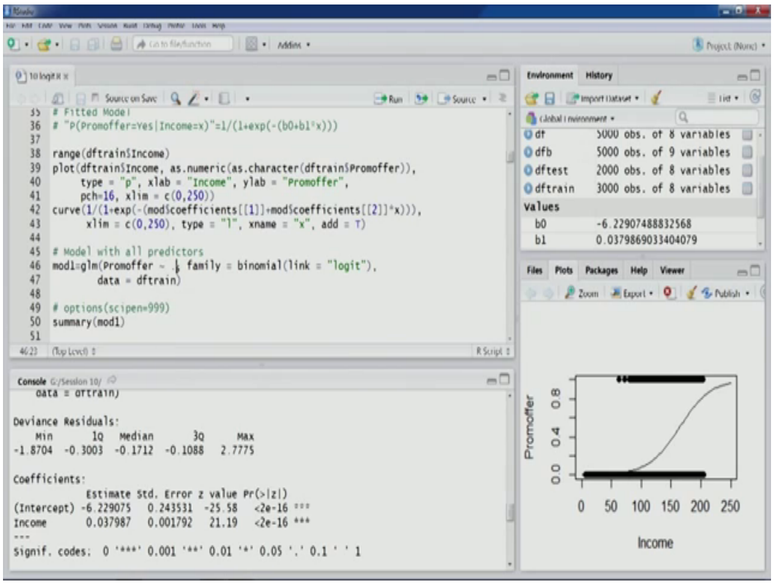
(Refer Slide Time: 18:46)



So, co linearity related issues are very much similar to linear regression. So, if you know predictors, if some of the independent variables in our model if they are collinear then that could be problematic for our logistic regression results as well just like in linear regression so much of the discussion that we did in a multiple linear regression technique so that also applies here as well.

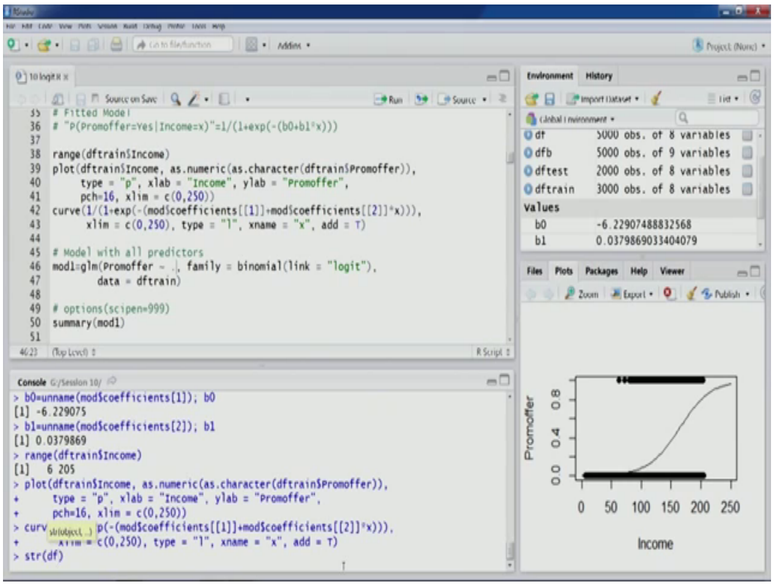
So, let us a let us go back to the R studio environment and what we are going to do.

(Refer Slide Time: 19:30)



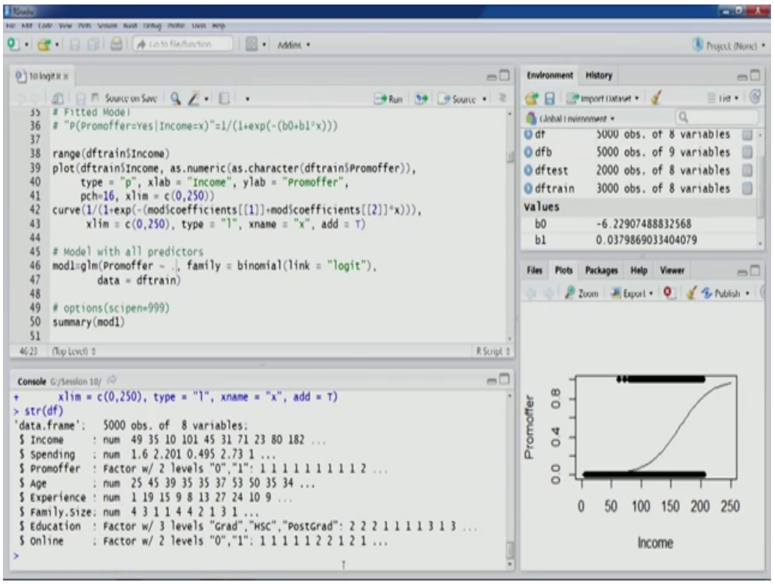
Now, is we will model the same data set we are going to use promotional offered data, data set and we are going to use the same data set now to build a model with all the predictors right; so, the predictors that we have.

(Refer Slide Time: 19:52)



So, just you know we have till now we have built a model with single predictor that is income; so, the predictors that we have; so, these are the predictors as we already know.

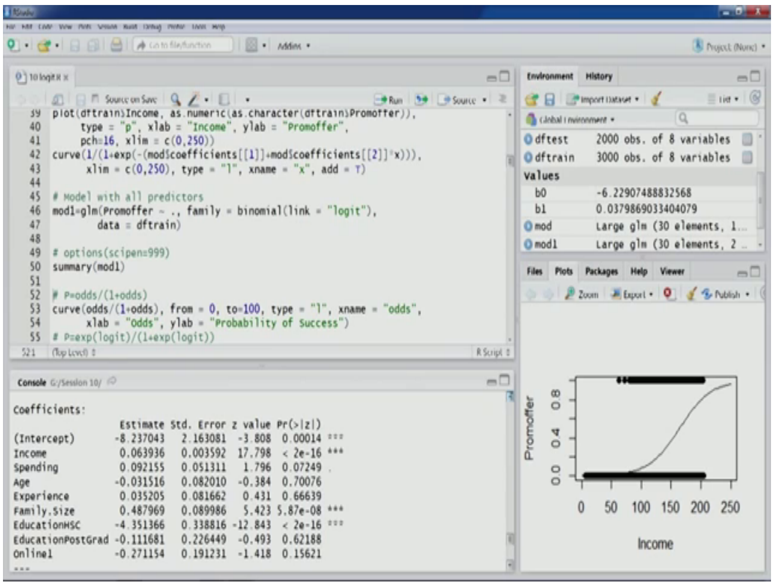
(Refer Slide Time: 19:55)



So, now, we would like to build a logistic regression model promotional offer versus all these other predictors income spending age experience family size education and online. So, the function that we used for this is same glm; so, promotional offer now against all the predictors now other arguments are same.

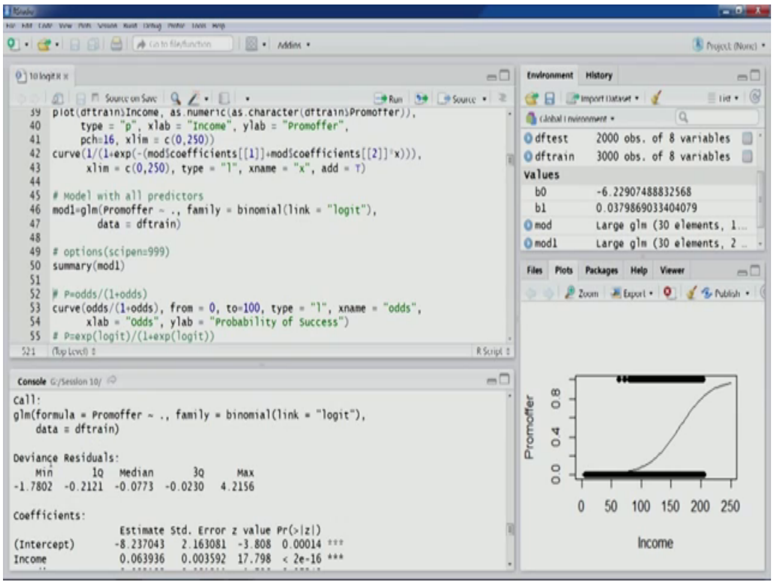
So, let us execute this, fine, let us look at the summary results.

(Refer Slide Time: 20:38)



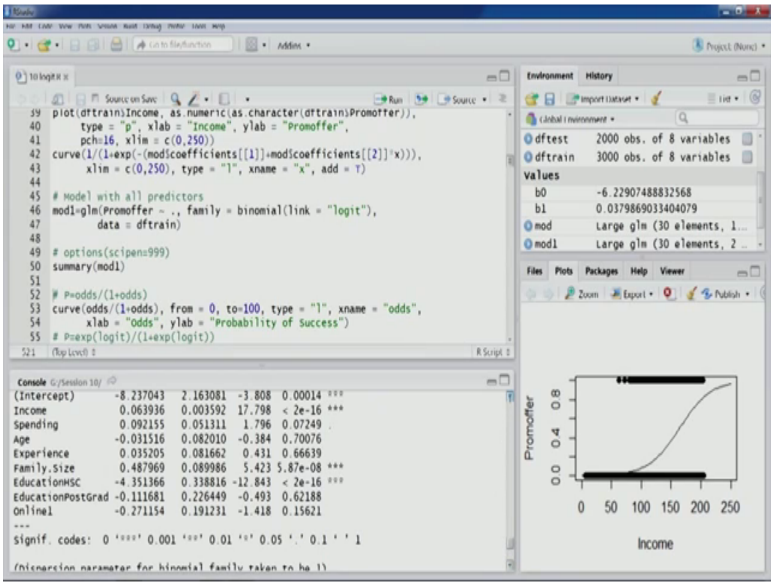
Now if we look here that; then first what we get is the call that actual code that we had used to perform this logistic regression modeling.

(Refer Slide Time: 20:49)



And you would see the deviance residuals here the basic statistics about the residuals and then we see coefficients and for all the Predictors.

(Refer Slide Time: 21:00)

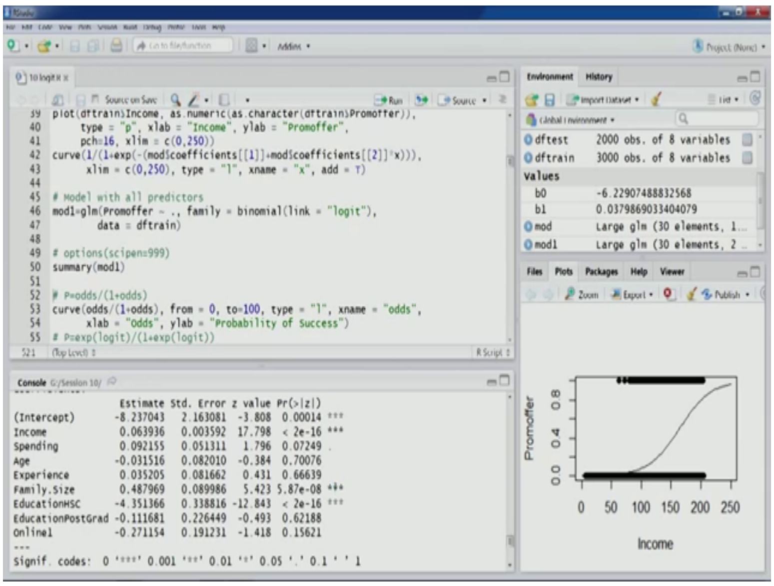


So, intercept significant than we see income which was significant when it was a used as a single predictor and in presence of other predictors as well; this particular variable is still remain significant now the second one is spending you can see is dot here and you can see a dot here. So, this is also this particular predictor is also significant at this particular level you can see the 4 dots; it has to be value has to be less than point 1. So, p value has to be less than 0.1, which is the case for spending you can see the p value is 0.07 to 49. So, at 90 percent significant at 90 percent confidence interval this particular spending variable is also significant.

While the income variable; it is significant at 99.9 percent confidence interval if you look at other variables age, this is not found to be significant experience this is also not found to be significant in this particular problem and data set family size we can see that this particular variable is significant at 99.9 percent confidence interval, this is you know a as we can understand from three star significance now if we look at the next variable that is education HSE, we can see that this is also significant, right. So, this is also significant at 99.9 percent confidence interval and after this we have education post grad this is not found to be significant as you can see the p value is quite high and for this particular variable, then we have a online.

So, online that the variable we had two levels; so, one whether a particular customer is active online and 0 if the customer is not active online; so, you can see online one this particular a dummy variable is being used. So, this also comes out to be insignificant as you can see from the p value.

(Refer Slide Time: 23:17)



So, from a this model as we can see a we can see 3 predictors that is income and the family size and the dummy variable education HSC. So, these three predictors seem to be significant and we had significant at 99.9 percent interval and spending also seems to be significant at 95 90 percent confidence interval.

Now if we look at the interpretation interpretation of these results we can see the income the coefficient is positive the value is 0.063936, right. So, what we can say is that for a unit increase in income right, we would see a some increase the corresponding increase in Logit values and because of that there is going to be you know some increase in the probabilities values and therefore, the probability of accepting the promotional offer right. So, similarly if we look at the spending; so, the this particular coefficient comes out to be 0.0921550 also positive. So, therefore, a unit increase in this particular variable is spending, we will also have the corresponding influence in the Logit values and therefore, but this is at 95; 90 percent confidence interval right.

So, some analysts might not consider this particular variable. So, corresponding influence in Logit values is from a spending as well and therefore, a corresponding change in the probabilities value and therefore, in the acceptance of the promotional offer, if we look at the other significant variable family size. So, this is significant the coefficient is also on the higher side, you can see 0.48, 0.487969.

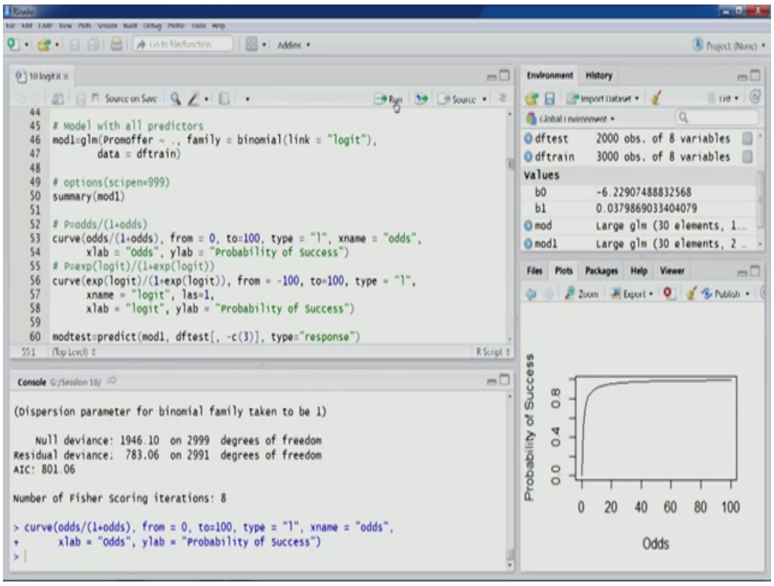
So, this is significant at 99.9 percent confidence interval and the so, therefore, a unit you know the unit to increase in family size will have a corresponding a increase in Logit values and therefore, the probability values will also change and that might a you know that might indicate the increased probability increased level of acceptance of the promotional offer.

If we look at the another in the first dummy variable which is significant the only dummy variable that is significant is education HSC. So, a as we can understand in this particular case the education variable we had three categories right HSC graduate and postgraduate grad and post grad and because of the alphabetical you know ordering of these categories the grad has been taken as the reference category and I just see in post grad we can see in the model.

So, therefore, interpretation of these dummy variables, are going to be with respect to the reference category that is grad. So, because this particular education is just see significant three start significance and the coefficient is quite high, but negative sign. So, direction is negative. So, the coefficient is minus 4.351366. So, from this we what we can say is that for customers who are having education up to HSC level right there is twelfth customers who are twelfth pass.

So, they have they have a significant lower levels of a acceptance of promotional offer in comparison to the reference category that is graduated; however, among graduates the second dummy variable education post grad. So, this was not found to be a significant. So, we cannot make any distinction between the customers having education grad a being graduate or being postgraduate; however, for customers will twelfth pass education, they will have they have lower level of acceptance in comparison to the reference category that is graduate. So, in this fashion a we can go about we can go about interpret doing the interpretation for logistic regression.

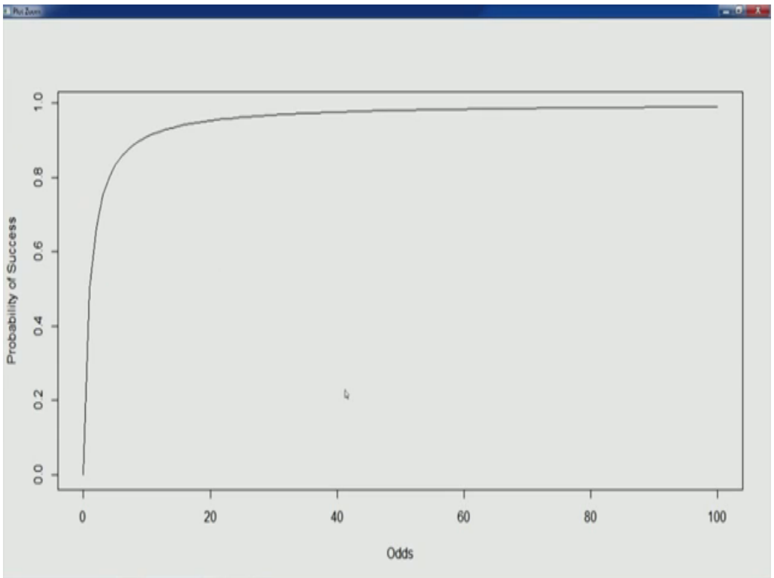
(Refer Slide Time: 28:11)



Now what we can do we can understand few we can learn a few more plots to understand it better, for example, for example, a Odds probability and odd. So, we had in the previous lecture, we had generated two plots where which was Odds versus probability value and Logit versus probability value. So, we will also look at how you know when the Odds are on x axis and probabilities on y axis. So, probability or probability versus Odds and probability versus Logit value; how the plot looks like.

So, as you can see in the curve function will have to write this equation. So, probably now Odds divided by 1 plus Odds that would be equal to the probability value and we are changing is from 0 to 200. So, because the Odds values, they can range from 0 to infinite. So, therefore, we are taking a good enough value 100 a large enough value and the a type of plot is going to be a you know a line and then this change x name is odds. So, with this let us create this plot. So, as you can see in this plot as you can see in this plot as the Odds increase as we move along to along the x axis from left to right as the Odds increase the probability of success also increases right.

(Refer Slide Time: 29:21)



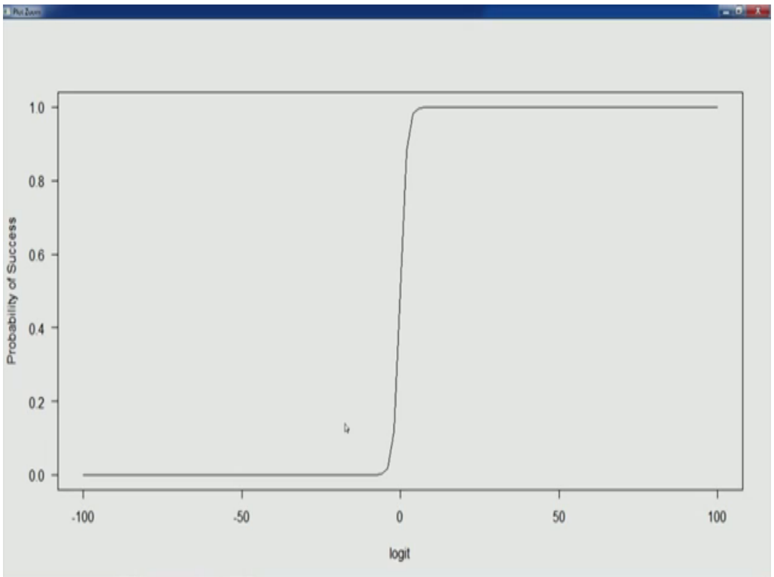
And the change in Odds value change in probability value is you know much higher a change as we increase Odds you know after you know the value after more than one it seems that after once Odds is greater than one the probability of success increases much more, right. So, because that is the for example, the cutoff the fall cutoff value that we have been talking about and the typical two class scenario case where the default value of on most problems.

Last method the corresponding default value could be 0.5 the same cutoff value we used the Odds instead of probability as the cutoff criteria then the corresponding value is going to be one right because the a probability is a Odds divided by one plus odd. So, if we have probability value as point five the corresponding value for Odds is going to be one.

So, because a once the Odds value is one you can see increase in probabilities values this is the relationship and similarly we can have a look at the curve between probability versus Logit. So, you can see now the equation we have the Logit values you know because we are building logistic regression model. So, therefore, from the estimates and a predictor values we can compute the Logit values and therefore, from though that those Logit values we can compute the probabilities values. So, this is the expression. So, exponential of Logit value divided by one plus exponential of Logit values and because this the Logit value can range from minus infinite to infinite. So, therefore, I have given a you know good enough range minus 100 to 100 and the type of plot is again line and Logit x name and other things are quite similar.

So, let us create this plot. So, this is the plot as you can see here.

(Refer Slide Time: 31:41)



So, this particular plot as you would see that this particular plot looks quite similar to what we saw in logistic response function. So, this is. So, quite similar to what we saw in logistic response function and as you can see as value inc increase along the x axis from negative values from minus 100 to minus 50 up to 0 the probabilities; there is very slight increase in the probabilities, whether it remains close to 0 and as we move further from 0 to more higher values 15 and you know 100 a closer to a Logit value of 0, we see sudden increase in probabilities right probabilities values and after that this a kind of remains close to 1.

So, a as the Logit values increase; so, for the negative where Logit values the you know probabilities values are going to be close to 0 for you know a positive Logit values the probability value is going to be close to one and near about 0 there is going to be the switch from you know close about 0 values to close about 1. So, that that significant change happens in here that value. So, a with this we will stop here and we will continue our discussion on logistic regression in the next lecture.

Thank you.