Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 09

Lecture – 43

Logistic Regression Model

Hello friend, welcome to the course Multivariate Procedure with R. So, up to now you can see we have considered different aspects of multiple linear regression model. And one of the basic assumption which we took was that Y was measured on a continuous scale that means Y is a continuous random variable. Thus same was true for X1, X2, Xk they can be discrete, they can be continuous, but for Y we had made that Y was following a normal distribution. So, it is a continuous random variable. Now in practice there are many situation where the outcome is only binary that means there will be only two outcomes whether this will happen or this will happen.

But the outcome depends on several factors. For example, if I say the example of there is a bank who wants to take a decision whether a credit card is to be issued to any customer or not. Now if the bank wants to take a decision on the credit card then there are different variables which are going to influence the decision. For example, the salary of the customer.

What was the past? Means in the past if the customer has any loan the customer has repaid it in time or not because in credit card what is happening? You are trying to first spend and then you have to pay it back after a month or so. And in case if the customer does not have a sufficient salary then possibly the customer will spend from the credit card but then there is a risk that person may or may not be able to pay it and the bank will be at loss. So when the bank is trying to take the decision it tries to consider different factors salary, what was the past and what is the age etc. But the final decision is only 1 whether the credit card has to be issued or not. So your response variable will take only 2 possible values and they can be indicated by say 0 or 1.

For example, y = 1 means yes credit card can be given and y = 0 means credit card cannot be given. So this is a binary there is no value between 0 and 1. In such a case if the y value takes value 0.7 it has no meaning. So the next question is how to do the regression modeling when our response variable is binary in nature? And response variable is depending on say more than 2 or say several independent variable or several explanatory variable.

So in simple words whatever multivariate or say multiple linear regression model we have considered y = X beta + epsilon now there is a constraint that the values in the y vector they takes only 2 possible values 0 or 1. So now in such a situation how to do the regression modeling? This is the topic which we are going to discuss in this lecture. So this can be achieved by logistic regression. So let us try to understand about the logistic regression its basic concept and how to implement it in the R software in this lecture. So let us begin our lecture.

Okay, so now we are going to talk about the logistic regression model in this lecture. So in the linear regression model y = X beta + epsilon there are 2 types of variable. One are this k independent variable or k explanatory variable X1, X2, Xk and there is steady variable which is indicated by here y. And these variables can be measured on a continuous scale as well as like as dichotomous variable or an indicator variable where they are trying to indicate a category only. For example, if I say a variable gender, gender takes value 1 if the person is male it takes value 0 or it takes that when the person is female or and so on, right.

So if the person has got say this good marks, average marks, bad marks then they can be indicated by the category 1, 2, 3, right. So when the explanatory variables are qualitative then their values are expressed as indicator variable and then we try to use the dummy variables model. We are not handling here the concept of dummy variable model in this course but these are well established model and available in most of the books on regression analysis and econometrics. On the other hand when the steady variable is qualitative in nature then its value can be expressed using an indicator variable which takes only 2 possible values 0 and 1. And in such a case the logistic regression modelling is used.

For example, this y can denote these values for example, this success or failure. Success can be denoted by 1, failure can be denoted by 0 or vice versa also. The answers yes or no, like or dislike which can be denoted by 2 values 0 and 1. Yeah, there is no necessity

that okay yes will always be taking value 1 and no will be taking value 0, they can interchange also, right. So now you get us try to understand how we can do it in the case of multiple linear regression model and how are we going to handle such a situation when my y is taking dichotomous variable 0 and 1.

So we are trying to consider here the model. This is the same model which we have considered y = x beta + epsilon but now I am trying to write it in the form of yi so that I can explain you in a better way, right. So this can be written here as Xi transpose beta + epsilon i where xi transpose is the i-th row containing observation xi1, xi2, xik on each of the k explanatory variables, right. And beta is a cross 1 vector having the component beta1, beta2, beta k without any problem. Usually if I try to take the first variable xi1 = 1, this will indicate the, which correspond to the intercept it come in the model.

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i = x_i'\beta + \varepsilon_i, \; i = 1,2,...,n$$

where $x_i' = [x_{i1}, x_{i2}, ..., x_{ik}]$, $\beta' = [\beta_1, \beta_2, ..., \beta_k]$.

The steady variable now takes 2 values 0 or 1, that is yi = 0 or yi = 1 for all i goes from 1 to n. Now we know from statistics knowledge that when yi is taking such binary variables then it will not follow normal distribution but it will follow a Bernoulli distribution with parameter pi. So its probability mass function can be indicated or can be written as yi takes value 1 with probability say pi, it takes value 0 with probability 1 - pi, which is sometime you will see it is written as qi also in many books. So that is what you have to now keep in mind, probability that yi = 1 is pi and probability yi = 0 is 1 - pi, right. Now we will assume here that expected value of epsilon i is 0.

So now under this case if you try to find out the expected value of yi, this is xi transpose beta which is equal to here probability of yi = 1, which is here pi, right. So you can see here now because this is the probability and expected value of yi is the mean value or the average value of yi. So now there is a constraint that the average value is lying between 0 and 1. And if you try to recall your general setup of multiple linear regression model, there was no such constraint. The average value can lie between - infinity and + infinity.

And in this case the variance of yi, it is the variance of a Bernoulli distribution which is given as say p * (1 - p). So this will become here expected value of yi * (1 - expected value of yi), which I am trying to indicate by the quantity sigma square yi. Now if you try to recall earlier in the case of multiple linear regression model, your variance sigma square was not dependent on the observation. We have not used the symbol about like sigma i square. But in this case the ith observation has got a variance sigma square yi and which is changing from one observation to another observation.

So now the question is how to model it. So when y is a dichotomous variable, that means it takes only two possible values, then the empirical evidences suggest that this expected value of y on the whole real line can be mapped to 0 on 1 and through a non-linear shape which is called as a shape like as this one. So here a different color plane like as here this is between 0 and here 1 and this is here x and this curve is going to turn like here this or it is here like this. So this type of a shape curve can indicate such a process where y is dichotomous variable. So now in case if you try to say this compare the two cases of multiple linear regression model and logistic regression, then suppose my y is dichotomous.

Then in that case all the data will be concentrated at two places. Here where the probability of y = 1 is 0 and here which is here on the y axis and here where the probability of y = 1 is 1 because it takes only two values. There is no data between 0 and 1. Now there are several constraints that you are simply fitting here one line like this. Do you think that is it a good decision? Certainly not.

Number 1, you are trying to take here these values but the range of y can exceed the 0 and 1 range also. So this is not good enough for us. So now we try to look at the logistic regression where y lies within the range of 0 to 1 and the two sets of data are lying here on the two extremes and we are trying to fit here s curve like. So this is the logic behind this logistic regression. So now if you try to understand all this process from the statistics point of view then we are trying to find out here probability of y = 1 for given x.

So this probability can be found here as say like this. Probability of y = 1 given x is exponential of x transpose beta divided by 1 + exponential of x transpose beta. I am not going to give you here the details that how I have found it out but you can assume that it is correct and you can try to indicate this value by here p because probability that y = 1 that we have indicated here you can see in this case here by here pi. You can see here. So that is why I am indicating it by here p.

So now if you try to see here if you simply try to solve it you will get here p upon 1 - p = exponential of x transpose beta. Remember here x is here a vector. And so I can see here X transpose beta = log of p upon 1 - p and log is the natural log. So now you can identify what is this X transpose beta. If you try to recall your expected value of yin the case of multiple linear regression model was expected value of y = X beta.

So now you have somehow found a function which is similar to x beta in the case of multiple linear regression model where y is simply a continuous random variable. So this I can define here by introducing a new term logit. So now I can say here y = here logit of p and logit of p is defined here like this one which is here log of natural log of p upon 1 - p. So now if you want to fit such a model what you have to do that you have to obtain the value of beta as beta hat and then you have to substitute it in place of beta to obtain the fitted values. So now looking at this expression y = this thing right which is here means here this thing also I can write down here the fitted value as say yi hat = pi hat = exponential of xi transpose beta hat divided by 1 + exponential of Xi transpose beta hat right.

And this can be written as say 1 upon 1 + exponential of - xi transpose beta hat. So you can see here now the difference earlier in the case of multiple linear regression model your y hat was x beta hat but now it is changed right. It is now here like this. And the next question comes here how to understand the interpretation of this thing. The interpretation of beta j in the case of multiple linear regression model was that is the rate of change in the average value of y when there is a unit change in the value of x or say xj.

To understand this the interpretation of this beta in the case of logistic regression model we try to attempt in a different way. First try to consider a simple case which has only one variable right. And suppose if I say here eta x = beta0 + beta1 X right. There is only one variable something like a simple linear regression model beta0 + beta1 x + epsilon right. Now we obtain the value of beta0 and beta1 as say this here beta0 hat and here beta1 hat.

Now the fitted value of this eta at x = xi is obtained just by substituting x = xi in the eta x function like eta hat xi = beta0 hat + beta1 hat xi. And this is the log odds at x = xi. This is how we call it. Because if you try to see here this is here the log of p upon 1 - p, and p upon 1 - p is called actually here is say odds. Those who are familiar with the probability theory they will know it otherwise it is a very simple definition.

So this is log of p upon 1 - p is called as log of odds. Now you try to do the same exercise at x = xi+1. So obtain the fitted value of eta at this xi+1 as beta hat xi+1 will be = beta0 hat + beta1 hat * xi+1 which is the log odds at x = xi+1 right. Log odds mean log base e of p upon 1 - p.

This is log odds. So now if you try to take the difference of eta hat computed at xi+1 and xi this will come out to be here like this. Odds at xi+1 in the natural log of odds at xi+1 - natural log of odds at xi. This will come out to be here natural log of the odds of xi+1 divided by odds of xi. And if you try to see here, if you try to transform it then this here odds of xi+1 divided by odds of xi is simply coming out to be here exponential of beta1 hat. And on the other hand if you try to look here, if you try to take this difference here say eta hat xi+1 - beta hat xi it = here beta0 hat + beta1 hat xi+1 - here beta0 hat - beta1 hat xi.

$$\hat{\beta}_1 = \hat{\eta}(x_i + 1) - \hat{\eta}(x_i)$$

$$= \ln\left[\text{odds}(x_i + 1)\right] - \ln\left[\text{odds}(x_i)\right] = \ln\left[\frac{\text{odds}(x_i + 1)}{\text{odds}(x_i)}\right]$$

$$\Rightarrow \frac{\text{odds}(x_i + 1)}{\text{odds}(x_i)} = \exp(\hat{\beta}_1).$$

So you can see here that this beta0 will get cancel out, beta1 xi will get cancel out this xi. So we are getting here as a beta1 hat. So this is what I have written here. So now you have obtained this quantity like as here. So this is terms as the odds ratio which is the estimated increase in the probability of success when the value of explanatory variable changes by 1 unit.

So this is how we try to interpret it. Now the question is the way you have conducted the test of hypothesis in the case of this multiple linear regression model using the t-test etcetera etcetera, but here you cannot use it because the model is now different. So the test of hypothesis for the parameters in the logistic regression model is based on the asymptotic theory. I am not going into details of this asymptotic theory. Asymptotic theory means when the sample size becomes very very large. So and for that test of hypothesis we have a very general procedure which is the likelihood ratio test.

So the way we try to do it that the test of hypothesis is actually a large sample test which is based on the likelihood ratio test statistics which is called as deviance. And there are certain definitions for example, a model with exactly p parameter that effectively fits to the sample data is terms as saturated model. And the statistics that compare the log likelihoods of fitted and saturated model is called as model deviance and we use this model deviance to conduct the test of hypothesis. It is defined as like this, the model deviance is indicated by this lambda beta which is actually twice of natural log of the likelihood of the saturated model - twice of natural log of the likelihood of the function at beta = beta hat where beta hat is the maximum likelihood estimate of beta and this L is here the likelihood function. So, in case if you assume that the logistic function is correct,

the last sample distribution of likelihood ratio test is approximately distributed as chi square with n - k degrees of freedom when n is large and that is a very standard result in statistics.

So that alpha percent level of significance we can conclude that if your model deviance is smaller than chi square value with n - k degrees of freedom at alpha level of significance then it indicates that fitted model is adequate. And if reverse happens that lambda b is greater than chi square with n - k degrees of freedom then it indicates that fitted model is not adequate. So this is how we try to conduct the test of hypothesis in the this logistic regression model. So, basically I will try to take a simple example and I will try to show you how you can implement it in the R software. So I am trying to use here the data set for from the book statistical analysis and data displays by Heiberger and Holland which was published by Springer in 2015.

This data is available through a package so that is why I am trying to use it here. So if I try to use it here for this thing I need to install here a package which is HH this is the capital H capital H means capital H two times. So you need to install this package using the command install.packages. You need to upload it using the command library and after that I will use here the data whose name is S-P-A-C-S-H-U.

It has got 138 observations on two variables. This data I can show you it here looks here like this. Well I will try to show you in the R console also which has got a variable here temperature in foreign height and then damage which is here only two values 0 and 1 you can see here. Some values are 0, some values here are 1 and that is all. So this is a good example to fatal logistic regression and you have this data is about some space shuttle challenger disaster which was related to you know massage space shuttle which has two booster rocket and each of which has three joints will be O rings. A warm O ring quickly recovers its shape after a compression is removed, but a cold one will not.

So, an ability of an O ring to recover its shape can allow a gas leak which may lead to disaster. So on January 28 1986 the space shuttle challenger exploded during the launch and the coldest previous launch temperature was 53 degree Fahrenheit. So, the temperature forecast for time of launch of the challenger on the morning of January 28 1986 was 31 degrees Fahrenheit and on the evening of January 27 day earlier teleconferencing conference was held among people at Morton, Tickell, Marshall Space Flight Centre and Kennedy Space Centre. There was a substantial discussion about ingenious over whether the flight should be cancelled, but there was no statistician was present for any of this discussion.

Well, I have taken it from the R software. So that is why I just spoken it, right. I am not an specialist in the space shuttle. But anyway I am more interested in using this data set. So, this input data set is there in the name of spacshu. So, if you want to use it you have to use the command data(spacshu) that is space shuttle that is the short form of this S P space shuttle is the full name and spacshu is the short name and it has two variable one is here this temperature which is temp and capital F that mean temperature in degree Fahrenheit at the time of launch and damage was either 0 or 1.

So, it takes value 1 if an O ring was damaged and 0 otherwise. So, each of the launch has 6 cases one for each O ring. So, there are a total of 23 * 638 cases and the O ring for one flight were lost at sea. So, I am simply going to use the data which will look here like this which has values here 0 and 1, right, ok. So, if you try to plot this data this will look like here you can see here the observations are centered here this, this, this, this, this and here.

So, the so I can think of our logistic regression model and the logit model under consideration is logit of p = beta0 + beta temp * f + epsilon where p is the probability of damage, right. So, you can see here this data is concentrated only at two point there is no data here in this graph. So, now, in order to fit this logistic regression model we have a command here glm. This is actually the generalized linear model, right. So, this glm command is used to fit the generalized linear model, but here with by choosing the option here family equal to binomial we can fit here the logistic regression model.

So, the expression is here like this glm all in lower case then you have to give the formula then data then family has to be specified as binomial because you have seen that binomial is more suitable because y follows a binary random variable. Y is a binary random variable which takes value only 0 and 1, two values and the model = true method = glm.fit and there will be many more commands, but I am going to restrict only to these things. So, formula is an object that how you want to specify your model. Data the way you the way you want to give you data on which the glm has to be fitted and family is a description of the error distribution and link function to be used in the models, right. The rest you can see in the help, but anyway I want to show you here the application.

So, I try to use here the command here glm(damage ~ tempF). It is something like y tilde x formula as we use in the case of multiple linear regression model like this. Data here is spacshu and family here is binomial and whatever is the outcome this I am trying

to store in this outcome spacshu.glm. You can give any name, right. So, that I can analyze it because I will be using it couple of times.

So, if you try to see here this is here the outcome. So, it is here the G L M function what we have given command, then I have here coefficients, the value of interceptor beta0 and the value of beta1. And then you have here degrees of freedom here 137 and then you have here the value of null deviance, then the residual deviance and the value of AIC. AIC is the Akaike information criteria, right. Because here you cannot use the use here the r square or the coefficient of determination. So, now based on that you can also find out different types of thing.

For example, if you use here the command here coef over this summary of this outcome, then you can get here these details. For example, estimates that means beta0 hat and beta1 hat for interceptor and temperature, then their standard error are here, their Z value are here, their p-value are here, right. So, now based on this I can see here that the fitted model or the fitted logit model under consideration is logit of p hat = 5.085 this is the value of here. beta0 hat - 0.1156 F which is the value of here beta1 from here. So, let me try to show you these things on the R console and you can see here this is here the outcome on the R software, right. So, let me try to first do it. Yeah, I already have installed this package on my computer. So, I am not installing it, but definitely i need to upload it and I need to have this data set.

So, you can see here. So, yeah, if you try to see here spacshu, yeah, you can see here this is the data set here like this. There are 138 observations you can see here, right. So, but anyway this data you can also observe. My objective is to first to make here the plot, right. So, if I try to make here the plot, so you can see here this will come out here like this.

All the points are concentrated on the bottom at 0 and at 1 on the top. You can see here. After that I try to use here by this glm command and I would like to save it here so that I can find out the summary command also. So, you can see here let me try to clear the screen. So, you can see here spacshu.glm.

So, you can see here this is here the outcome. Right. So, this is what you are getting here. This is the same outcome which I explained you. And if you want to get here the coefficients of this thing where from where you can obtain this model, this is here like this, right. So, you can see here this is how you can obtain a logistic regression model

also. And yeah, we have not covered here AIC, but anyway, you can have a look into the books and you can find out about a Akaike information criteria as well as Bayesian information criteria that is BIC.

Yes, I have given you a brief idea about this concept of model deviance. So, null deviance is the model deviance under H0 and its value here is 66.54. Ok. So, now we come to an end to this lecture and you can see here in this lecture, we have talked about the generalized linear model and we have taken a very specific case which is the logistic regression. This glm command is a very generalized command in regression analysis and in linear models and it is used under different types of concepts and we have used it for the sake of fitting logistic regression.

Yes, I agree there were some concept which we have not understood it, but definitely they are very popular concepts and which are available in almost all the books on regression analysis in the chapter of logistic regression model. My objective was to give you a fair idea that under what type of situation you can use the logistic regression model and how to obtain them in the R software. So, I will stop here, but definitely you do not have to stop, but you need to take some data set, you can collect some artificial data set and try to execute these commands and try to see how you can interpret it. And believe me this logistic regression is very useful in real data set because many times you want to take a decision, the decision can be yes or no, you are not interested in the values. In those cases this is a very popular modeling technique whenever the response variable takes the binary variable, binary outcome.

So, you try to practice it and I will see you in the next lecture, till then goodbye. Thank you!