Transcriber Name: Saji Paul
Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology, Kharagpur

Lecture - 46
Logistic Regression (Part - Ⅱ)

Hi guys, so in continuation of our discussion on logistic regression last lecture we have discussed logistic regression. So, there we have mentioned different type of logistic regression but we did not discuss in detail so today we will be discussing that.
(refer time: 00:37)

So, one type of logistic regression that we have seen is binary logistic regression. That is binary logistic regression again there were two different types that is one is one explanatory variable two categories or more than one explanative variables and still two categories binary means it will have to get two category pass fail success failure happy sad whatever it is two categories. But explanatory variable may be one or explain a variable may be more. So, today we will be discussing binary logistic regression with one explanatory variable.
(refer time: 01:09)

So, now say this same example I bought here again so a group of 20 students spends between 0 and 6 hours 0 to 6 hours studying for an exam. The table shows the number of hours each student spends studying and whether they passed or fail. So, if it would have been a linear regression instead of pass or fail if a score is given the amount of hours, it is the student has put what is the score.

So, if there is a relationship between the hours and dress code and we would have done gone for a linear regression analysis and simple regression analysis. And if the what is the relationship is linear then it is a simple linear regression analysis this is just one variable but unfortunately no we are not given the test score what is given is the student has passed or failure two categories pass category fail category.

So, now here meaning we want to find a model given the student has put this many hours of study say some x what is the probability he will be passed whether we will succeed. So, we have to find it out. Here also we are predicting there also we are predicting but here prediction is different their prediction was different, their means I mean regression, simple regression.

(refer time: 02:25)

So, how does the number of hours spent studying affect the probability of the student come to the fag end of this course still I am not comfortable using this pin actually. So, how does the number of hours see again I think it is done, now how does the number of hours studying affect the probability of the student passing the exam. That is what we find out we need a module for that.

The reason for using logistic revision for this problem is that values of the dependent variable pass and fail while represented by 1 and 0 and are not cardinal numbers. The values are not cardinal numbers the values are not some scores 120 140 whatever it is. If the problem was changed so that pass field was replaced with grade from 0 to 100 certain in some score that is cardinal numbers then simple regression analysis could be used but unfortunately here we cannot use simple regression analysis.
(refer time: 03:33)

So, we used to fit a logistic function, logistic function means a logistic function means a sigmoid function to the data consisting of the hours studied and outcome of the test y i = 1 for pass and 0 for fail. We need basically probability of that probability that the student will pass that is y i = 1 and probably that the student will fail that is y i = 0. We need to find such a sigmoid function the data points indexed by the subscript i which runs from i 0 1 to 20 because there are 20 students.

The x variable is called the explanatory variable and the y is called the categorical variable. The x variable is called the explanatory variable it is also called a predictor because this variable predicts the outcome. So, x variable is also called the predictor or explanatory variable and y variable is called the categorical variable because we have different categories output is in different categories. Consisting of two categories pass or failed corresponding to category values 1 and 0 respectively.
(refer time: 04:39)

So, here see I have if you find it difficult you may not use this also, I have used a different sigmoid function. It is not a different σ function but just a different form we are used to having this sigmoid function f of $x = 1 / 1 - e^{-z}$, we are used to this sigmoid function. So, if you are comfortable with that go with that but just for binary logistic regression, we can use this form as well.

It is just the same thing just since it is a binary, we have bring out the concept of μ I will just tell you. So, now if this is the thing then what is t that is the logistic is log of odds that is the odd or

this p / 1 – p. We have seen it in the last lecture is it what is odd, odd is p / 1 - p so logic is log of p / 1 - p. So, here for generalized sake I have written b base is b usually base I you can use base e you can base use this 2, 10 it totally depends on the range of the data.

If you have very high at a high range of data then definitely will go for log 10. So, it depending on the range of the data you can use whatever base it is. So, it is not necessarily you will use base e only. Usually when we come for more explanatory variable more than one so it may be that log base e may not suffice then maybe we have to use a higher base maybe 2 or maybe 10. So, that is why just for generalizes I have kept here base b this is obvious like all of you know that.

So, the graph of the logistic regression curve fitted to the line does this is the curve the curve shows the probability of passing an exam versus hours of studying. So, versus hours of studying business hours of studying and decide it is the x axis is the hours of studying y axis is the probability of passing the exam. So, this is the sigmoid function this graph gives us that for how many hours of studying what is the probability of passing.

So, in this usually 1 - 3^{-z} when z goes to ∞ our probability value goes to 1 when z goes to minus ∞ our probability value goes to 0.
(refer time: 07:00)

So, this is the logistic function which we have just now considered. So, this basically this expression since if you are not comfortable you can just convert it into this form. It is the same thing because it is just one parameter so it will be β 0 + β 1 x so β 0 + β 1 x. If we; just substitute this then what will be my β 0, β 0 is nothing but μ / s and intercept is nothing but 1 / s.

So, what is μ actually why I have used a μ, μ is a location parameter the midpoint at a curve where p of μ is equals to half. In binary there will be just two categories which one to use as a classifier so maybe we can use this (07:43) no issue if you are finding a trouble you can just use that only no issue at all. So, μ is nothing but it is just a location parameter that is the midpoint of the curve, μ is the middle of the median of the curve.

So, p μ is half probability of μ is will be always half so and s is the scale parameter. So, what is β 0 here? If I substitute it in this the standard format this is the standard format so β 0 is - μ / x β 1 = 1 / s. So, now we have already seen what is p of x i what is the probability given that x i values, what is the probability that it will this student will pass or the it will hit the target.

Basically, we are trying to find out the success since in categories it is more than two categories then we will take it differently. So, we may define fit to y i, fit to y i maybe means that a success pass that is one at a given axis p i we can define is this way just for simple calculation instead of carrying out p of x i all the way just simply I can just write p i, p i means p of x i. What is the probability given x i what is the probability that it will be success do not get confused with that.
(refer time: 08:57)

So, now p i had a problem is that the corresponding y i will be unity and 1 - p i are the probabilities that they will be zero, nothing no doubt here, I hope. So, we used to find the values of β 0 and β 1 which gives the best fit to the data. We have seen one example in my last lecture we had a concentration under the odds and the if a person has certain concentration of some particular whatever it is and then what is the thing that he has (09:27).

So, there from the data given I have already shown you that we can find out the different values of β 0 and β 1 but we want to find out that the best values of β 0 and β 1 which will maximize the likelihood of the occurrence of the observed value because the values that we have observed as a real value that we have observed that is. So, we want such parameters which will maximize the likelihood of occurrence of this observed value.

So, there will be different β 1 so now we need the best β 1 β 2. Best β 1 β 2 means we will have to use MLE estimation method the MLE is nothing just that it finds out that parameter value it finds out the parameter value in such a way that is maximize the occurrence of the observation what we have seen. It maximizes the likelihood of the observation that we have observed likelihood of the occurrence that we have observed.

So, in case of linear regression whatever we have seen that is written here we know how we try to find out the minimize the error and then we try to find out the parameter. Similarly, here we will try to find out the, we will try to maximize the likelihood of the observation.
(refer time: 10:43)

So, this is the likelihood function, so this likelihood function what is that is the probability that the given certain test is produced by the logistic function, so this is the likelihood function. So, we want to maximize the likelihood of occurrence of the observed value. So, what is that, so basically it is a joint probability distribution of all the values what we have seen join probability distribution of all the values what we have seen.
So, this is product of all the p i, p i where y i = 1 that means where we got the success probability of all the products all the product of all the probabilities where we got the thing there is a success and 1 - p i where y i = 0 this is nothing but the likelihood function. So, when we maximize this likelihood function that is after maximizing this likelihood function whatever value

we get for the parameters that is the best parameter for β 0 and β 1.

So, understood this is the likelihood function nothing it is just the joint probability distribution of all the probabilities, basically we are trying to multiply all the probabilities. Now here when you try to multiply all the probabilities it might the equation might become unstable because when some probabilities if you try to multiply some very small probabilities maybe the value becomes maybe and may become zero also.

So, when we try to multiply too many probabilities it senses that the equation become unstable. So, in such case what we do is that so this is my likelihood function and I need to I want to maximize this likelihood function. So, because there it involves a lot of multiplication of probability, joint probability distribution is nothing but a multiplication of the conditional different conditional probabilities.

So, what I do is that I take the log of this that is called the log likelihood. So, if I take the log of this if I take the log both side then this product becomes sum is not a log when the product if we did not take the log this side now this product will become sum that is all.
(refer time: 12:50)

So, this is the thing log likelihood. So, this also I can write it in this form. So, this is only instead remember what a p i was I am using p of x i I am writing this p i. Now just for your understanding I have kept this step you see what does this step means this step means just this and so up here I bought again p x i what was because if you get confused what is p i, p i is basically probability of success given x i probability of success means y i = 1 given x i probability of y i = 1 given x i this.

So, if we simply and whatever p of x i p of x i has this b 0 and b 1 term what is p of x i p of x i is $e^{t/1+et}$ and what is t, t = β 0 + β 1 x. So, simple simplification here nothing I will not explain these simple simplifications but actually you do not have to do it manually all these ready-made programs are there, in the net nowadays all these ready-made programs are there you will have to you do not have to do this many MLE you do not do it manually just feed the data and you get the results.

But even if you should know also how to do just, I have simplified it, now when I am talking of maximize this likelihood, I made log likelihood so that my equation remains stable since I am multiplying the probability is so I have made it log likelihood now to maximize this. So, basically to get the optimum value. So, what I will have to do? I will have to differentiate these equations partial differentiation with respect to each and every parameter here there are just two

parameter I will do the partial derivation with respect to β 0.
I will do the partial derivation with respect to β 1 and then I will find out the value of β 0 and β 1 by equalizing to 0.
(refer time: 14:54)

So, that is what, I am just differentiating it with respect to the different parameters and once I have what to say differentiated it with respect to the different parameters then I will equalize it to 0 to get the values of the β 0 and β 1 that is the technique standard technique which you know but here there is a problem. What problem is there? You see the expression here this p i when I talk of p i there is a term of exponential terms.

So, this is this whole this expression it is not an algebraic expression. So, when this is not an algebraic expression so when we differentiate it, we cannot because since it is not an algebraic expression it is not it will not give us a finite sequence of some algebraic operation is not it exponential terms are there and the exponential terms there then sin term cos and when these terms are there.

When it is we will not get a finite series of algebraic operations it will give an infinite series. So, when this is an infinite series definitely though I have shown it here because this is the standard way of equalizing is to zero and then finding out the value but actually not actually you cannot equalize it to zero because this is not a finite series. It is not an algebraic expression that is why this is not an finite series.

So, what you will have to do? You will have to after differentiation you will have to use one of the numerical techniques methods the best is Newton Raphson method you will iteratively using some approximations iteratively you have to solve this expression and find out the values of β 0 and β 1 by using Newton Raphson method. Newton Raphson method will best suit here. So, now please do not expect me to explain the Newton Raphson method.

That is a totally a concept of numerical techniques numerical methods that all of you should know as engineering student all of you should know that, not only in this subject in many subjects you will be needing this numerical technique. So, please go to it if you have not gone through it please go through it.
(refer time: 17:03)

So, now so by that whatever Newton Raphson method we have used and we found out the value of β 0 and β 1, we found out the value of β 0 and β 1. So, now what was for this since it was binary, we have used the concept of μ and s and all those stuff (17:28) so I can find out

also what was μ you can if you can remember here see here β 0 is - μ / s β 1 = 1 / s from there if I find out μ what will be μ.

So, this is my μ value I have the μ value I can find out since I know β 0 and β 1, I can find the as value similarly I can find out the logic expression that is the t. So, now the next question you have to find out I found out the coefficient. Now again I need to do here also I need to do the significance test; significance test means here and here what is my null hypothesis. My null hypothesis is that there is no correlation there is no the what to say the predictor does not the coefficient of the predictor to give the dependent variable the coefficient is almost equals to zero.

That is there is no dependency between this predictor and the dependent variable and the response variable. So, that is my null hypothesis. Alternate hypothesis the predictor contributes for the dependent variable. So, now how do I find a better at this whether there is a dependency between the predictor and a dependent variable significant dependency or it is not. So, for that I need to find out the p value.

If my p value is very small that means if it falls in a rejection region given a significance level usually significance level is 0.05 is considered so even if my significance p value is much less than this it is less than the 0.05 that means if it falls in a critical region then I reject the null hypothesis. Null hypothesis means that there is no dependency between the predictor variable and the dependent variable.

So, now how do I find a p value here? So, I found out the coefficient so after finding out the coefficient I will have to find out the standard error. What is the standard error? Standard error means the how the coefficient the variance of the coefficients. So, remember the first example in my first class on logistic regression when I have given so concentration cancer concentration and having tumour.

So, as I told you we will be able to find out different β 0 β 1 values is not it. So, now let us take just one parameter suppose the β 0 or let us not consider the intercept because what is important is the slopes the different slopes that we have. The parameters that are that basically that is involved with the independent variable that is the predictors. So now this β 1, so now we will have different values of β 1 is not it so there will be variance among these values.

So, this variance among these values is the standard error. So, now how do we find out the standard error? So, it is really not possible to find out the standard error manually there are soft wares for that, so you know you should know the technique how we find out is that supposed to

find out the coefficient for we have used maximum likelihood estimation and whatever it is and we found a cost coefficient for that suppose we took a sample a bigger size sample.

Suppose we took a sample of size 100 for example of size 100 and we have found out the coefficient β 0 and β 1. So, now what from this sample we will take a sub sample, say maybe of size say we have taken a sub size say 60 or size 50 whatever it is. We will take another from this sample we will take is another sample that is lesser than our original sample. We will by maybe will take a sample in such a way so that by replacing we will take one again will replace that value again will take will replace that value.

So, we will take a sample and then from that sample we will calculate another β 0 and β 1. Similarly, again we will resample again we will take another sample from that so that way we have taken many samples. For each sample we have used this MLE estimates and we got the β 0 a β 1 value. One largest on the largest sample we have calculated β 0 β 1 values, now from this smaller samples from the same sample but we have what to say done sampling of this sample.

We have done the sampling of the sample means we have taken a sub sample of the sample this sub sample we have taken maybe say size 50 so we will take differences sample of size 50. We will take sample it different times (21:43) so we will take a different sample of size 50 also be size equal so then we will find out different values of β 0 and β 1. Now this you know different values let us consider β 1.

So, we found out different values of β so we will find out the variance among this β 1 there will be this β 1 the values are different this difference what is this difference this variance is nothing but this is the standard error that is the standard error. So, here we have this is the coefficient value what we got it by (22:13) then we calculate the standard error. Once we calculate the standard error then we will be able to find out the z value.

What is the z value? There is one technique by proposed by well this is called well technique to find out the z value is nothing but the coefficient divided by standard error gives us the z value. So, once we get the z value definitely, we will be able to find out the P value from the z value given a z value we will be able to find out the p value that is no issue at all. So, now here from the z value itself we can come to the conclusion.

If the value of z is further away from zero that is the value of z is away from zero then only, we can say this it is large enough to be significant that means there is a relation between the predictor variable and the dependent variable. So, this z value is quite large this z value is quite

large so this itself it shows that that null hypothesis is rejected. So, when we got the p value also which is very less that is how we do the significance test if we have to do the significant test.

And once we find a z value now β 0, we got this value β 1 we got this value these are some high level topics so I am just mentioning it I am not going to explain in details. So, we found a β 0 value we found a β 1 value we found z value. Now remember confidence interval while teaching hypothesis testing while discussing hypothesis then I have discussed confidence interval confidence interval means what value b 0 will take within in the we give the interval of β 0 value.

We cannot because this is we are just giving one estimate point estimate that is not good probability of a point estimate is zero. So, it is better if it is always better if you can give an interval. So, once we have the z value once you have the coefficient value, we can calculate the interval also. Same technique what we; have used to calculate confidence interval while discussing hypothesis testing.
(refer time: 24:20)

So, that is what here this value is 0.0167 it is very less value by Wald test it is the output indicates that hour studying is significantly associated with a probability of passing the exam. Since the p value is very less that means it is significantly hours and passing the exam this association is significant association it is not an insignificant association. If the p value is greater than 0.05 then null hypothesis is satisfied that means hour studying it does not it is not a predictor for the passing the exam.

So, β 0 β 1 coefficient may be entered into the logistic regression equation to estimate the probability of passing the exam. So, for example for a student who studied 2 hours entering the value at x = 2 into the equation gives us the estimate probability for passing the exam. So, here this is t this is my logic equation so I found out I have found the b I have I know the β 0 value β 1 value I know what is x axis 2 so I found out what is t then how what is how do I find out probability of passing given 2 hours.

This is the formula is not it the sigmoid function I know the value of t putting the value of t I got the probability of passing the exam is not it. It is not difficult once you find out the β 0 and β 1 finding out the probability is not difficult at all.
(refer time: 25:43)

So, similarly suppose a student studies 4 hours so what is this estimated probability of passing the exam. If you studies 4 hours that means x = 4 I have β 0 and β 1 value so I found out the t value this is my t value. So, probability of passing the exam $p = 1 + e^{-t}$ it is the same thing as e^{t}

/ 1 + e t same thing this form or this form whichever form you use.

So, 0.87 and probability of this is the 0.87 probability of passing the exam. The table shows the problem you are passing the exam for several hours values of our study. You see here for different hours here I have specifically kept the value 2.71. Why for binary we use this depth form which I have given the sigmoid representation what we have seen no need to use that do not get confused.

But if you use this for your understanding see for μ is equals to 2.71 what I got probability I got 0.05 it is the mean of the curve basically 2.01 I found t value is 0 then log of e 3 is 1 probability is 0.5. So, for different values hours of study I could find out the probability of occurrence and probably have passed in the exam.
(refer time: 26:54)

So, output from the logistic analysis gives a p value which is based on the Wald z score, Wald z score is nothing but the coefficient by standard error. Coefficient for each we find out for each individual coefficient standard error we find out for individual coefficient we find out the individual coefficient for β 0 β 1 β 2 how many other coefficients are there. This simple model is known as binary logistic regressions with one variable and two categories and has one explanatory variable.

The above example of binary logistic regression on one explanative variable can be generalized to binary logistic regression on any number of explanatory variables. Here we had just one explanatory variable the example we have seen is just the hours of study that is one explanatory variable. Now we can generalize it to any number of relative variables maybe hours of study and maybe the age maybe both these ages also may be the predictor of passing the exam.

So, that way many number of expanded variables and any number of categorical values category values may not be only just zero and one here we have seen just 0 and 1 fail or pass there may be more than two categories as well.
(refer time: 28:02)

So, in this lecture we discussed binary logistic regressions, we learned the concept of logic very important what is logic. Now you understood from logit we got the logistic regression logit is nothing but the log of odds, what is odd probability of success by probability of failure ratio of success by failure basically. So, that is I should not use the word probability odd is the ratio of success by failure that is on.

Then we understood a concept of logit we understood a concept of odd then what is and maximum likelihood estimation. And also, the use of sigmoid function in logistic regression, these are very important things which you have learned in this lecture. In the next lecture we learned binary logistic regression with more than one explanatory variable and also, we will learn multinomial logistic regression.
(refer time: 28:50)

And this is the reference and thank you guys.