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 - 48
Tutorial on Logistic Regression

(refer time: 00:30)

Hi guys, so today is the tutorial lecture. So, first as usual first will be starting with some objective type questions and then we will be doing some problem as well.
(refer time: 00:39)

Now first coming to the objective type equations. Objective type questions always I am repeating it again do not see the do not click the means do not see the what to say answer immediately first try to answer yourself, then you can understand your level of understanding. So, which of the following statements is or are true or false. So, logistic regression may be used for classification, so is it true or is it false?

Say logistical equation basically we have seen it is basically it is not a classifier but of course we can use this for classification. How? If it is a two cluster so we can have a threshold value say 0.5 if my value is probably is 0.5 or above then I can sell it is it in a class one if my value is less than 0.5 probability then I can say it is in class 0, if it is a two class. Similarly, if there is three class accordingly, I can set two threshold for the demarcation between these three classes.

So, it is not a classifier per se but we can definitely use it as a classification as also so logistic may be used for classification this is true. Next is the logit in the logistic regression is a linear function in terms of independent variables. This question I have a objective behind keeping this because whatever I have discussed till now all you have seen the logit is a linear function of the independent variable.

But that is actually it is not always the case that logit is a linear function sometimes there is a non-linear logic maybe a non-linear function in terms of the independent variable as well. So, in the example what I have taken is just linear function. So, that is why logit in a linear function a logic in a logistic regression is a linear function in terms of independent variable that is not true it may be linear it may be non-linear, so this is false.
(refer time: 02:41)

Which of the following is necessary for auto regression analysis? Remember auto regression, so the dependent variables should be normally distributed, applications to both periodic and a predict time series data. Time series data should satisfy non-stationary property, time series data should satisfy both stationary and uniform property, which one is following is necessary for auto regression analysis.

Auto regression analysis we have seen for auto regression analysis our data is not a simple data it has to be time series data. And we have also seen and for auto regression we have also assumed some we have made certain assumption about the time series data that is. What is the assumption? It should satisfy both stationary and uniform property remember what is stationary property stationary property means?

First let us talk of uniform property what is uniform property? Uniform property means suppose we were considering from say 2000 to 2022 we are considering monthly data so that means so when we are considering from 2000 to 2020 monthly data that means we need 12 months 12 data for each year. So, there should not be any missing months data or in some cases there will it is given instead of months it is given days data no we do not need that, it should be any form that is the uniform property.

And stationary means the probability distribution should be stationary it should not vary with (04:13) so that is the stationary property. Probability distribution of means probability for each time each data for each data it should have the same probability distribution. So, probability distribution for all the data of the time series of different years maybe if we consider years or months or whatever it is, it should be it has a uniform probability distribution.

So, that is the stationary property. So, time series data should satisfy both stationary and uniform property. Now if you treat to see the other options the dependent variables should not be normally distributed the dependent variable should not be normally distributed. We did not in time series data we did not talk about the dependent variable distribution of the dependent variable because that would have come.

When we would have specifically talked about the distribution of the dependent variable but that was we did not assume anything of that so, this is definitely this is not necessary, applicable both for periodic and a periodic time says that are definitely logistic auto regression is applicable only for periodic data it is not applicable for a periodic data. And time series data should satisfy non-stationary property that is definitely when it satisfies when our assumption is just non-stationary will be (05:31).
(refer time: 05:31)

The auto correlation is defined as correlation of a given time series with another time series data definitely not given time series it another time series data. How? Why? No, the correlation of a given time series with lag value of another time series data that is also definitely not it may be lagged value the same time series data. The correlation of a time series data with its own lagged value, this is correct. The correlation of one time interval data with another time interval data that is also not correct.
(refer time: 06:06)

Logistic regression transform the output probability to be in a range of 0 to 1 we have seen logistic regressions that is the our model we need such a model that if we get the output probability will be in the range of 0 to 1. Given an input our output would be a probability value it should be value within the range of 0 and 1. Which of the following function is used by logistic regression to convert the probability in the range of 0 and 1?

So, which of the following function we use? We use sigmoid function, hyperbolic sine, logarithmic, hyperbolic tan, which function we use? Definitely we use sigmoid function because sigmoid function f x f x outputs a value between 0 and 1. If we consider hyperbolic sign, hyperbolic sign what is the how is the hyperbolic sign it is something this way this is the hyperbolic sine curve.

So, and similarly if we consider hyperbolic 10 that will also that is also something this way so definitely it will not be in the range of 0 and 1. Logarithm to definitely not base log what is log of zero all of you know that so logarithm function is also definitely not so it has to be when we since we in logistic regression we want our output to be in the range of 0 and 1. So, sigmoid is the function for that.
(refer time: 07:39)

Which are the following options are true? For linear regression error values have to be normally distributed but in case of logistic regression it is not the case. This for logistic regression error values have to be normally distribution but in case of linear regression it is not the case. For linear and logistic relation error values have to be normally distributed. Both linear and logistic relation error value have not to be normally distributed. Which is the correct answer?

See here for linear regression error values have to be normally distributed that is very true for linear regression we have assumed that a error value is normally distribution it has a constant variance with σ^2 is not it but its logistic regressions we are not considering the error values here we are just giving the probability we are not considering the error values and we are not considering that it is normally distributed or whatever for linear regressions we are considering that so, answer is the correct answer is a.
(refer time: 08:39)

Then which of the following is a sigmoid function? this very easy which of the following is a sigmoid function? This is the sigmoid function, is not it? This also we can write it in this form e^x / $1 + e^x$ same.
(refer time: 08:58)

So, what is the value of the sigmoid function already we have seen so value of the sigmoid function as x equals to ∞ as x goes to ∞ sigmoid function it gives 1 as x equals to basically minus ∞ so it gives 0, so 1 and 0 is respectively.
(refer time: 09:19)

In which of the following case of logistic regression the target variable can have three or more possible values without any order, the target variable can have three or more possible values but without any order there is no ordering but it can be more than two basically if more than two definitely it is not binary if model 2 definitely it has to be multinomial but if it maintains a order then we would have called it ordinal logistic regression if it maintains order. But here it is selling without any order so it is the multinomial logistic regression.
(refer time: 09:57)

The ratio of the probability of event occurring to the probability of the event not occurring event occurring to the problem not occurring not that is odds and log of odds its logit.
(refer time: 10:12)

Which of the following is a time series problem? Estimated number of guest room booking in next 6 months estimating total, next six months what will be the number of guest rooms booking we want to estimate that is it a time series problem? Definitely yes because to estimate that we need the value for previous years data or previous months data based on that we will find out it is in all this periodic data will need based on that we will be able to estimate that this is basically very much a time series problem.

Estimating total sales in next two years of an insurance company this is also very much a time series problem. Estimating the number of calls for next one week that also we will estimate it based on the regress data based on the lag data is not it. So, this is also a time series problem, so which is the answer all three other time series problem so 1, 2 and 3 this is the answer.
(refer time: 11:13)

So, now we will see certain problems. See here the question looks a bit big but nothing to worry very simple. In a study of urban planning in a country a survey was taken of 50 cities 24 use tax base funding and 26 did not out of 50 cities have surveyed 50 cities out of 50 cities 24 cities have used tax-based funding and so there is in software writing is TF and 26 did not use tax base funding.

One part of the study was to investigate a relationship between the presence or absence of TF and the median family income of the city. So, is this tax-based funding as it is depend on the medium and median income of the family. So, there are total 50 cities out of 50 24 use tax-based funding for different work maybe different what to say different type of work for setting up of some public organizations, different utility anything it used tax-based funding and for 26 other cities it did not use.

So, now the study is to investigate under which sort of city are using tax based funding, where the median income is greater there we are using tax base funding is. Basically, is there any relationship between the tax base funding and the family income of the particular city. So, data is given in table one with median income in order of 1000 dollar so here it is in order of 1000 dollar so here you see the income here it is given access 9 point it is in the order of 1000 dollar, so 9.2, 9.2, 9.3.

So, when this sort of data's are given so you see it is very difficult to find out for the 9.2 there are two there are 9.3 9.4 9.5 very small intervals it is very difficult to find data for each and every data likewise. So, in if when the data in such a way in statistical method we usually group the data. So, first thing is that we will group the data we will make certain group maybe from 9 to 10, 10 to 11, 11 to 12 like this whatever group if you find out whichever group is best.

So, accordingly we will we have to find a grouping in such a way so that our group also group should not be too many at the same time groups should not be too less also. If the group is two less then too many members will the frequency of the group will be will increase. So, that way we will have to make and make the grouping, so first for this we will make a grouping.
(refer time: 13:45)

So, here we have income category from 9 to 10, 10 to 11, 11 to 12, 12 to 13, 13 to 14, this is how we have made the income category and then midpoint of this income category 9 to 10 midpoint is 9.5, 10 to 11 is 10.5. So, now in this category how many is are number tax based funding zero we are indicating for one of the classes zero here 0 is for tax base 0 is where city which does not have tax base funding and one for tax base funding.

So, in this category there are total 13 is non-tax based funding and one is tax-based funding. So, we are trying to find out log of 1, log of 1 what it will be 1 / 13. So here similarly here for this case 10 to 11 category number of zeros at three that means total three cities which are in this income range in this income range three cities which has which does not have tax base funding four cities have tax base funding.

So, what is log of text based when we find out odds of tax based funding so it will be 4 / 3.

Similarly, for this 11 to 16 in this range odds of tax base funding will be 6 / 6 this is also 6 this is 6. And log of tax base funding for this range it will be 10 / 4, similarly this is, here you see in this range from 13 to 14. So, log of tax based funding will be 3 / 0, when there is a zero term in any of the frequencies see here it is given.

One of the cases has zero occurrence creating an undefined ratio odd ratio that is 3 / 0 it becomes undefined. Since a log of odd is undefined, we followed a common practice of adding one to both numerator and denominator counts in a calculation of all the odds. In such case when there is zero occurrence what we do is that we add one both to numerator and denominator. So, that and it will not make a difference if we add 1 to both the numerator and denominator.

So, that way we can avoid the undefined results. So, this is similarly we found the logouts of this so we have for now we have the logouts for different values then now the question is that we have the logouts. Now what we need we need to find out the parameters value? How will you find out the parameters value? Parameters value will be using maximum MLE method. So, one other way also we can use parameters but that is we will we may not get the best parameter it is just if we plot the value.
(refer time: 16:30)

See here we have plot the values midpoint of income because what is that independent variable here what is the explanatory variable here explanative variable is the income category. Is not it? So, here we have taken the midpoint so this is my explanatory variable. So, I will plot it based on this is my explanatory variable and this is my log odd so I will plot the data. So, by plotting the data if we could fit a straight line then it is very good.

Then from there or if the maximum points if I can find out the line which crossed or the maximum point then from the line itself, I can find out the intercept I can find out the slope and then that will give me the values of the different parameter that is one way. But of course, with that it is we cannot be sure that we got the best value of the parameters. Another wave is that we have the log odds value we have the since independent variable.

So, we can have different expression, we can have different equations, two unknowns two equation we can find out a value. But then from this many number of logouts we can have if I take two to each so I will be having different combination of equation. From different combination of equation, I will see I will get different β 0 β 1 values now which β 0 vitamin I will consider which one is best I need to consider which one is this.

So, best method is go for MLE and then you find out the β 0 and β 1. Now suppose here we

found out the β 0 and β 1 value we found this is my β 0 value and this is my β 1 value. So, how we have found out by using your software to use you definitely use MLE what are your software and you found out this is your β 0 value decision β 1 value. Now see β 1 value is 0.8529, what does this indicate? It indicates a strong correlation.
(refer time: 18:23)

So, does the logistic model chosen is in this form where x is the city's median income odds and odds is the probability that a city will have tax base funding divided there probably that it will not have tax base funding.
(refer time: 18:34)

So, we got this is my β 1 value, so we find that there is strong evidence of a relationship between the tax based funding and the median income because this is β on value is quite high value it is not close to zero it is close to one. So, when the β 1 value is the high value then we can see that a strong evidence of relationship between the tax base funding and the median income.

So, more value cities have a higher probability when there is a evidence that means when the median income is more that means cities have a higher probability of adopting tax base funding. So, now if the city has a median income of 12,500 the probability of adopting TF is just put it in the value this is how we got the T value once we know the T value, I can find out the $e^t / 1 + e^t$.

So, this is the probability so there is 0.704% probability that the city with 12,500 median income will adopt tax based funding. So, once we have basically logistic regression once we find out any other regression model basically once you find out the parameters then there is nothing. Just find out the parameter, once you find out the parameter then given the predictive variable find out a response variable that is all.

Any of the regression analysis whether you call simple linear regression, auto regression, logistic regression whatever it is just that logistic regression gives a probability and other regressions gives us a value, that is all.
(refer time: 20:05)

Now next question Time Magazine used data from USA to compare whites and blacks opinions of the death penalty. The data consists of response from 32937 participants collected from 1972 to 1996. The outcome variable is whether the respondent did or did not support a death penalty. The survey provided a table of the percentage of whites and black each year that supported the death penalty.

So, in different year what is the percentage of people who have supported the death penalty is

given white who have supported the death penalty and number of blacks who has supported the death penalty percentage basically percentage is given, so what we need to find out?
(refer time: 20:54)

It is given comment on any if there is a trend in time does it appear linear or quadratic, so convert the percent is given into table so log odds within each race and year and plots log on versus year. Comment on a pattern you see is there a trend in time does it appear linear or quantity. So, we have to see whether how this what to say how does opinions with respect to year it is here is the independent variable their opinion is the dependent variable is there any relationship between that.

So, and is if there is any relationship is it linear or is it quadratic? So, what we will do that? We will first find out the log odds of white and log odds of black we will do it separately.
(refer time: 21:51)

First, we will find out this is given white so we will find out the odds which supporting the death penalty odds of white. How do we find out odds of white? Percentage of white person supporting percent of white person opposing, a 57.4 is supposing the 100 - 57.4 is the opposite. If you would have found out the probability what we had have done 57.4 /100. So, now it is odds 57.4 / 100 - 57 that is how we find I found out the odd of 1972.

Similarly, I found the log odds similarly for all the years we have found out so all the years we have found out the log odds.
(refer time: 22:27)

So, once we found out the log odds then basically once we found out the log odds then what happened we found a logit function then we will have to find out the MLE estimation MLE and then estimate that parameter this is we have estimated the parameter. This actually I will not go in details of this question because I have not explained here non-linear when the logit is non-linear we have only seen the logit is linear.

This example we have I have solved it some using the software. So, I could just show you the result. So, here once first you assumed as a linear logit as a linear on the function of all the parameters and explanatory variables linear function of all the explanative variable basically that is why we have used the parameters. So, then using MLE methods we found the parameters and um we found out the parameter that is one model we have seen.

Using the that model we got that model then we wanted. Here the question is that does it appear linear or quadratic. As I told you so sometimes just from the plot, we can we can say whether it is a linear the dependency is linear or non-linear. But sometimes it is very difficult to

tell just from the plot. So, what we have to do first we will definitely try linear and then we found out the R^2 value.

Remember we have talked of R^2 value; R^2 value is the quality of fit. What is R^2 value? 1 minus of if variable in error by the variability in the data. Remember we have done that 1 minus of sum of that is variability in the error divided by the variable in the data that we have variability and error and variables in data we have taken in terms of sums squares we have discussed that.

So, we found out the R^2 value more R^2 value it is a better is not it, so we found out for linear we found out R^2 value. Then we will try for non-linear, so non-linear is the say $\beta 0 + \beta 1 x 1 + \beta 2 x$ if I have just one parameter lamb $\beta 0 + \beta 1 x + \beta 2 x^2$ so it is non-linear of order two. So, again using MLE we found out the parameters of $\beta 0$ $\beta 1$ $\beta 2$ then we found out again within found out the R^2 value we got a particular R^2 value.
So, if we found out that the quantity R^2 value is better than the linear R^2 value then definitely it is not linear it is non-linear. Now non-linear means it is of degree 2 or degree 3 first we have tried with 2, then we will see with three non-linear with degree 3 we will some using degree 3 we will find out the parameters then we will find out the R^2 value. Suppose using degree 3 with our value of R^2 reduces then we understood that no.

It is that means it is non-linear and it is of degree 2. So, similarly for this example we have found out the R^2 value for quadratic is 0.842 weight and for linear is 0.6532. So, it is from this you can see that it is a logistic linear model whose logit function is quadratic because this R^2 value is greater than the linear value.
(refer time: 25:41)

So, similarly we will do for black for black also we will find out the odds of the black odds means basically black with supporting the death penalty. So, how it will be here if we see it is 28.8 odds will find out 28.8 / 100 - 28.8. So, then we will find out the log odds so once we find out the log odds. Then we will get the logit function now we have to find out the value of the parameter same way we will find out the value of the parameters.

And once we found below the parameters, we got the regression models then we try to find out the R^2 value.
(refer time: 26:21)

So, similarly for black also we will find out the R^2 value, here we have seen for our black also when we try to find out the opinions on the regarding the death penalty of the black. Here also we found out that R^2 quadratic as quadratic R^2 value is greater than linear R^2 value. So, this is also it is quadratic line fits the data more even from the plot also we can see that so, with that we come to the end of this lecture.

(refer time: 26:47)

So, then thank you guys.