International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6S2, August 2019

Logistic Regression for Health Profiling

Ambika P., E. Laxmi Lydia, K. Shankar, Phong Thanh Nguyen, Satria Abadi

Abstract*: in an event when there is lots of risk factor then the logistic regression is used for predicting the probability. For binary and ordinal data the medical researcher increase the use of logistic analysis. Several classification problems like spam detection used logistic regression. If a customer purchases a specific product in Diabetes prediction or they will inspire with any other competitor, whether customer click on given advertisement link or not are some example. For two class classification the Logistic Regression is one of the most simple and common machine Learning algorithms. For any binary classification problem it is very easy to use as a basic approach. Deep learning is also its fundamental concept. The relationship measurement and description between dependent binary variable and independent variables can be done by logistic regression.

***Keywords* : logictic; regression; medical; binary variable.**

I. INTRODUCTION

A regression class where variable that is independent is utilized to foresee the dependent variable is known as Logistic regression [1]. It is called a logistic regression of binary type when the variable that is dependent has two classifications. And it is called logistic regression of multinomial type when the variable that is dependent has more than two classes. When the dependent variable category is to be ranked than it is called OLS (ordinal logistic regression) [6]. By transforming the variable that is dependent in the logit function, it can get maximum likelihood measurement. The logit function is generally defined as natural log of the dependent variable and it show the event occur or not. In between dependent and independent variable there is no linear relationship. This function doesn't expect homoscedasticity.

The logit model is a kind of statistical analysis and it is often used reaches out to applications in machine learning and for prescient investigation and displaying. in this approach the dependent variable is consider as categorical or finite.

A. they can be either A or B, this show the binary regression

B. or there is several finite options like A, B, C or D, they show multinomial regression

The relationship between one or more Variables that is independent and the variables that is dependent it is used statistical software, by using logistic regression equation it measures the probabilities [1].

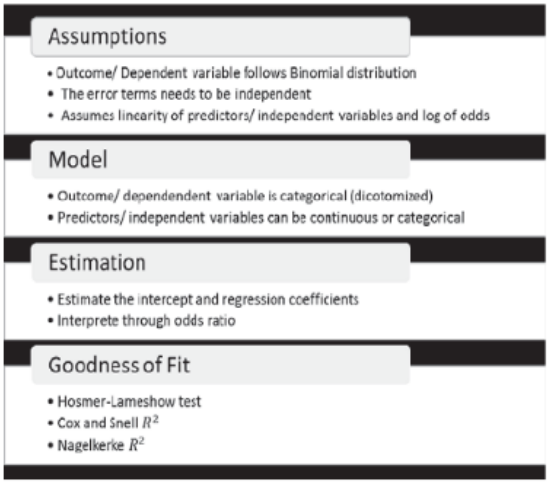

Figure 1: Binomial Logistic Regression Essential

II. LINEAR REGRESSION VS. LOGISTIC REGRESSION

The logistic regression gives a static or constant output but the linear regression provide a continuous output. The price hose and stock price are the example of continuous output. Predicting the patient has a disease or not is the example of discrete output. Logistic regression is measured through Maximum Likelihood Estimation (MLE) method and linear regression is calculated using Ordinary Least Squares (OLS).

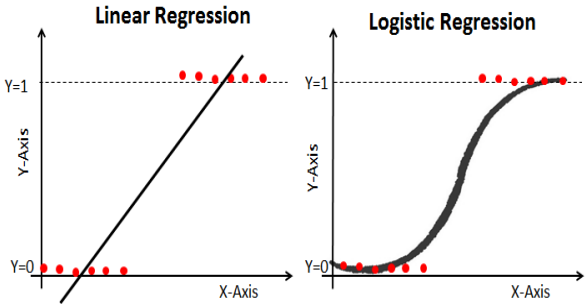

Figure 2: Linear Vs logistc regression

III. TYPES OF LOGISTIC REGRESSION

There are several types of logistic regression. Some of them are as follows:

Manuscript published on 30 August 2019.
***** Correspondence Author (s)
Ambika P., Department of Computer Science, Kristu Jayanti College, Bangalore-560043, India.
E. Laxmi Lydia, Professor, Vignan's Institute of Information Technology(A),Department of Computer Science and Engineering, Visakhapatnam, Andhra Pradesh, India.
K. Shankar*, Department of Computer Applications, Alagappa University, Karaikudi, India. E-mail: shankarcrypto@gmail.com
Phong Thanh Nguyen*, Department of Project Management, Ho Chi Minh City Open University, Vietnam. E-mail: phong.nt@ou.edu.vn
Satria Abadi, Department of Information Systems, STMIK Pringsewu, Lampung, Indonesia.

1. Binary Logistic Regression
In this type of regression the variable has only to output like the document is spam or not and the patient as cancer or not

2. Multinomial Logistic Regression
The Multinomial type of Logistic Regression uses 3 or more nominal output. Like predicting the disease type.

3. Ordinal Logistic Regression
The Ordinal type of Logistic Regression variable uses three or more ordinal output. Like rate any hotel from 1 to

Model building in Scikit-learn
This research work taking an example of building a diabetes prediction model.
Here we are using a Logistic Regression Classifier for predicting the diabetes.
In the initial stage the data is loading in the CSV function. It is shown in figure 3.

```
#import pandas
import pandas as pd
col_names = ['pregnant', 'glucose', 'bp', 'skin', 'insulin', 'bmi', 'pedigree', 'age', 'label']
# load dataset
pima = pd.read_csv("pima-indians-diabetes.csv", header=None, names=col_names)
```

```
pima.head()
```

Figure 3: loading data

	pregnant	glucose	bp	skin	insulin	bmi	pedigree	age	label
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 4: example of data set

IV. LOGISTIC REGRESSION MODEL

In medical and biomedical research area the logistic regression has wide range of applications. To determine whether the output is happen or not for different factors this model is used. In logistic regression model the output is binary or dichotomous variable. Generally to maintain the attributions in predicting the providing output, the data of patient is used [7]. To measure for any new patient by probability of given outcome that is (Y) it place in logistic regression model. The equation of binary logistic regression is as follows:

$$P(Y)=\frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\ldots+\beta_n x_n+\varepsilon)}}$$

Here P(Y) is consider as probability of given output to be predict

The probability of negative output is considered as P(Y).

ß0 is consider as Constance

x1 to xn show the independent variables

ß1...ßn show the models who are always constants for each attributer

As mentioned above in logistic regression the dependent variable is always binary or dichotomous. So the output is only in the form of true or false. The data coded 1 is the result is true or successful and it coded 2 if the outcome is false or failure.

The logistic regression is used to get the best result. Ad it described the relationship between dichotomous characteristic of interest and a set of independent variable [10]. Here

Response or outcome variable is consider as Dependent variable and Independent variable is predictor or explanatory

The Logistic regression generates the coefficients to address the probability of presence characteristic of interest of a logit transformation:

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

The characteristic of interest's probability of presence is defined by p. The definition of logit transformation as the odds logged is:

$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characteristic}$$

and

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

The variable in logistic regression chooses in the way that is maximizing the value of used data instead of data that use to reduce the errors count.

V. HOW TO ENTER DATA

In the given model AGE and SMOKING are two predictor variables. The variable OUTCOME is consider as response or dependent variable. The dependent variable coded 1 for positive and coded 0 for negative [3].

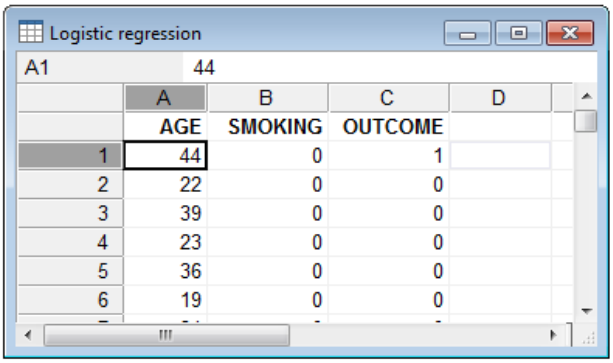

Figure 5: enter the data in model

Required input

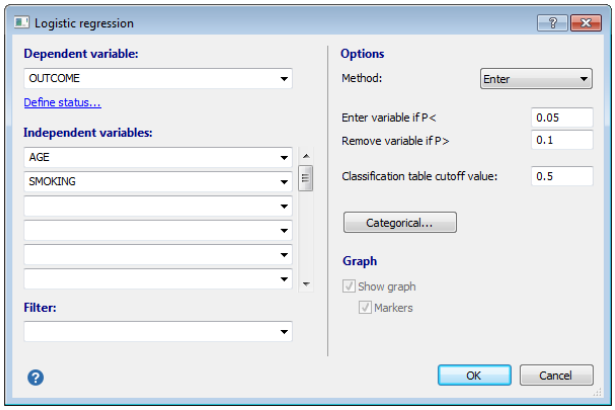

Figure 6: required input

Dependent variable

Binary variable or dichotomous is known as dependent variable. This is the variable which has to be predicted. It contains only value 0 or 1. To recode the data it can use define status [4].

Independent variables

It has to be selecting another variable that is expected to relevant to the dependent variable.

Filter

In the analysis to include only selected subgroup the data filter can enter. It is optionally used.

Graph

If there available only one single independent variable then the logistic regression curve has the option to plot a graph.

Here the figure7 shows that MedCalc plot a logistic regression curve by using only one single independent variable.

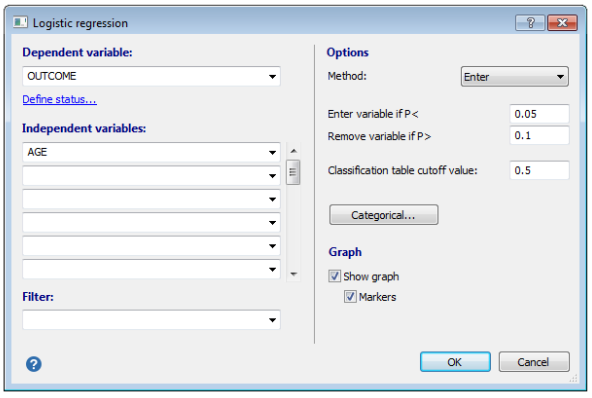

Figure 7: logistic regression curve with single independent variable

By using the above value a graph is created which shown in figure 8 [5]:

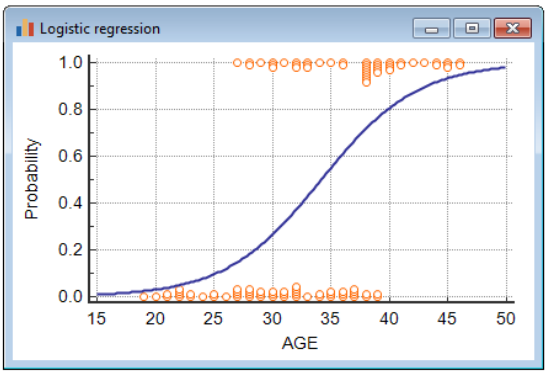

Figure 8: created graph

VI. RIGHT VARIABLES PICKING AND THE LOGISTIC REGRESSION VALIDATING

It is very necessary to pick the correct variable in logistic regression. Here picking the correct explanatory variables and model validation are linked together [8].

For a logistic regression picking the right predictors is parallel as picking the regressors. It is called as OLS (ordinary least squares) regression. It has one or two complications. To measure the different model it has to try several different models and to measure different model use diagnostics raft [9].

Few year back a review in medical research describe that show some required steps in validation have to study the different aspects [11].

A user can decide the regression analysis type which is depend on the nature of the outcome variable, it require giving the answer of a research question.

Table 1: Depend on outcome and predictor variable Choice of regression analysis

Outcome	Variable Predictor	Regression
Continuous	Continuous/ Categorical	Linear Regression
Categorical (Binary)	Continuous/ Categorical	Binomial Logistic Regression
Categorical (Multiple)	Continuous/ Categorical	Multinomial Logistic Regression
Categorical (Ordered)	Continuous/ Categorical	Ordinal Logistic Regression

VII. CONCLUSION

In an occasion when there are heaps of hazard factor then the logistic regression is utilized for foreseeing the likelihood. For paired and ordinal information the restorative scientist increment the utilization of logistic investigation. A few grouping issues like spam identification utilized calculated relapse. On the off chance that a client buys a particular item in Diabetes forecast or they will motivate with whatever other contender, regardless of whether client click on given commercial connection or not are some model. For two class arrangement the Logistic Regression is one of the most basic and basic AI calculations. For any paired arrangement issue it is extremely simple to use as a fundamental methodology.

Profound learning is likewise its central idea. The logistic regression in the field of health profiling is the current major issue to concern. The study of the research shows the logistic regression model to measure the value in the form of binary variables. And there is lots of scope to develop a better model to enhance the accuracy and efficiency in logistic regression model when it is using for monitoring the health issues.

ACKNOWLEDGMENT

This article has been written with financial support of RUSA–Phase 2.0 grant sanctioned vide Letter No. F. 24-51/2014-U, Policy (TNMulti-Gen), Dept. of Edn. Govt. of India, Dt. 09.10.2018.

REFERENCES

1. Adul-Aziz AR, Harris E, Munyakaza L, 2012. Risk factors in malaria mortality among children in northern Ghana. International Journal of Business and Social Research 2 (Issue 5):35-45.
2. Hosmer DW Jr, Lemeshow S, Sturdivant RX (2013) Applied Logistic Regression. Third Edition. New Jersey: John Wiley & Sons.
3. Long JS (1997) Regression Models for categorical and limited dependent variables. Thousand Oaks, CA: Sage Publications.
4. Pampel FC (2000) Logistic regression: A primer. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-132. Thousand Oaks, CA: Sage Publications.
5. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR (1996) A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology 49:1373-1379.
6. Boland A. Lawrence, 2007.The identification problems and the validity of economic models. South African Journal of Economics 36 (Issue 3):236-240.
7. Kleinschmidt I, Sharp BL, Clarke GPY, Curtis B, Fraser C, 2001. Use of Generalized Linear Mixed Models in the Spatial Analysis of Small-Area Malaria Incidence Rates in KwaZulu Natal, South Africa. American Journal of Epidemiology, 153 (Issue 12): 1213-1221.
8. Meulbroek L, 2001, "The Efficiency of Equity-Linked Compensation: Understanding the Full Cost of Awarding Executive Stock Options," Financial Management (Summer 2001), 5-30.
9. Pigeon JG, Heyse JF, 1999b. A cautionary note about assessing the fit of logistic regression of freedom Journal of Applied Statistical Science, 26 (Suppl 7): 847–853.
10. UNITED NATIONS, 1986. The role of the family in the development process. New York: United Nations (Department of International Economic and Social Affairs).
11. Boland A. Lawrence, 2007.The identification problems and the validity of economic models. South African Journal of Economics 36 (Issue 3):236-240.
12. Azzalini A, Bowman AW, Härdle W, 1989. On the use of nonparametric regression for model checking. Biometrika, 76 (Suppl 1): 1–11.