IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 12, December 2016
ISSN (Online) 2348 – 7968 | Impact Factor (2015) - 4.332
www.ijiset.com

The use of logistic regression in a human health problem.

Erjola Cenaj[1], Raimonda Dervishi[2], Zhifka Muka[3]

[1] Department of Mathematics, Polytechnic University of Tirana/ Mathematics and Physics Engineering Faculty, Tirana/Albania
[2] Department of Mathematics, Polytechnic University of Tirana/ Mathematics and Physics Engineering Faculty, Tirana/Albania
[3] Department of Mathematics, Polytechnic University of Tirana/ Mathematics and Physics Engineering Faculty, Tirana/Albania

Abstract

Logistic regression analysis is important statistical method for analyzing the medical records. This paper analyzes the data collected in two zones: one cold and the non cold. These data are considered as categorical variables. The use of regression analysis in modeling the data is discussed and a regress model for the categorical response disease is defined. The use of chi-square statistic to test the dependence of disease from the zone and gender.

Keywords: *Logistic regression, categorical data, dependence of variable, disease.*

1. Introduction

Logistic regression is a statistical modeling technique which may be applied to estimate the simultaneous effect of a set of predictors (age, gender) on the risk of a certain outcome variable (disease) which can take either one of two possible values (yes/no). This paper describes the study of the effects that the cold climate has in human health, as well as in regression relation between the disease, the gender, and age. The data are collected in two zones one in the cold zone and in an zone not cold. The Pearson's chi-square statistic is used to test the dependence of the disease from the zone. The logistic regression analysis is use in modeling the data of the cold zone for the categorical response disease.

2. Materials and Methods

We have investigated the data of two different environments in the north of Albania, in Kukes (cold climate) and we lezhe (with non cold climate). The database contains an excel file age in years (AGE), the gender (GND) Absence or presence of the disease Rheumatism (RHM) for 80 individuals: 60 from the cold zone and 25 from zone no cold Which Were selected to participate in the study. In the first part of the paper we were focused in testing the dependence of the RHM from the zone using the Pearson's chi-square statistic. In the Table 1 the number of the individuals with the absence or presence of the RHM for the cold and non cold zone is presented. In the second part of the paper we show the dependence of the disease by age and gender.

	0	1	Total
Cold zone	28	32	60
Not cold zone	21	4	25
Total	49	36	85

Table 1: Data of rheumatism (RHM) to cold zone and not cold zone.

The Result of Pearson Chi- square test χ^2=9,1036 shows that there is evidence at significant level of α = 0,05 that the RHM depends on the zone. Our main purpose is to study the relationship between AGE, GND and the presence or absence of RHM for the data of the cold zone. To know more about the relationship of the response RHM and one of the independent variables AGE we form a scatter plot presented in Figure1.

In the Scatter plot of the Figure 1 all points fall on one of two parallel lines representing the absence and the presence of RHM variable. As it is seen there is some tendency for the individuals with no evidence of RHM to be younger than those with the evidence of RHM.

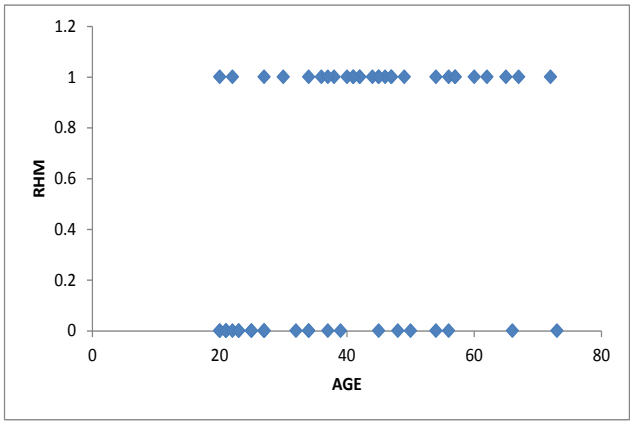

Figure 1: Scatterplot of RHM by AGE for 60 individuals

We know that the functional relationship between RHM and AGE is difficult to be described because of the large variability of RHM. To remove some variation and still maintaining the structure of the relationship between RHM variable and AGE variable we create intervals for the independent variable and compute the mean of RHM with in each group.

	Absence	Present	Mean
18-30	13	4	0,23
31-43	7	10	0,58
44-56	6	10	0,62
57-69	2	8	0,8
Total	28	32	2,23

Table 2: Frequency table of AGR by RHM

In Table 2 we present the age group (width equal) variable, AGR which group in intervals the age data in years, the frequency of occurrence of each RHM value and the mean for each group, which provide the estimated values of E(Y/X). We will assume that these values are close to the real values of E(Y/X) to have an estimation of the relationship between RHM and AGE.

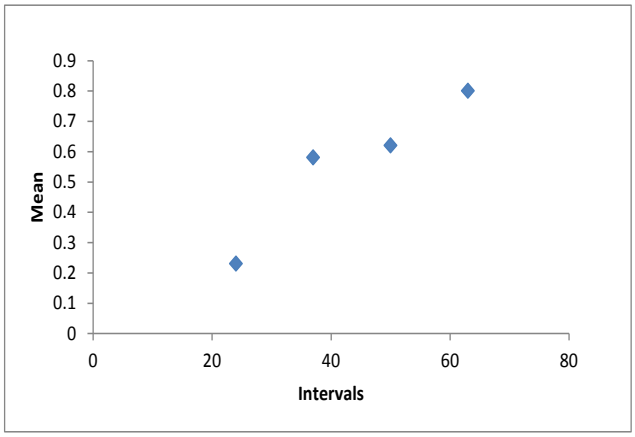

Figure 2: Plot of the proportion of individuals of RHM versus age group.

Gender (GND) is the other independent variable presented in the data.

	0	1	Total	Odds
Female	13	21	34	1,61
Male	15	11	26	0,73
Total	28	32	60	

Table 3: Female male number of people affected by the disease

The Result of Pearson Chi- square test χ^2=3,3436 shows that there is evidence at significant level of α = 0,05 that the RHM depends on the gender.

3. The model

Let us denote with *X* the vector of predictors $\{X_1, X_2, \ldots, X_k\}$, to model the response, we might use the ordinary linear regression model:

$E\{Y/X\} = X\beta$ (1)

Since the expectation of a binary variable *Y* is *P(Y= 1)*. The model (1) cannot fit the data over the whole range of the predictors since a linear model

$E\{Y/X\} = P(Y = 1/X) = X\beta$

can allow P(Y= 1)to exceed 1or fall below 0.

The statistical model that is generally used for the analysis of binary response is the binary logistic regression model,

stated in terms of the probability of *Y = 1* given X, the values of predictors:

$$P(Y = 1/X) = \frac{1}{1+e^{-x\beta}} = \frac{e^{x\beta}}{1+e^{x\beta}} \quad (2)$$

The function $\pi = \frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}}$
called the logistic function has unlimited range for x
while π is restricted to range from 0 to 1. Solving the equation 2:

$$x = \log(\pi/(1-\pi)) =$$

$$= \log[odds\ that\ Y = 1\ occurs]$$
$$= logit(Y = 1)$$

The logistic model assumption are understood by transforming *P(Y= 1)* to make a model
that is linear in $X\beta$.

$$\text{logit}\{Y = 1/X\} = \text{logit}(\pi)$$

$$= \text{logit}(\pi/(1-\pi))$$

$$= X\beta. \quad (3)$$

The model (3) is a linear regression model in the log odds that *Y = 1 since* logit(π) is a weighted sum of the X_k. So the model (3) assumes that for every predictor X_j

$$\text{logit}\{Y = 1/X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (4)$$

The parameters β_j of the model (4) are estimate using the maximum likehood method. The fitted logistic model for relating GND and AGE to RHM is given below.

The fitted logistic model is:

$$\text{logit}\{RHM = 1/AGE, GEN\} =$$

$$= -4.012 + 0.091AGE + 2.41GND \quad (5)$$

The p-values respectively for GND, AGE and Constant of the model (5) show that there is sufficient evidence that the parameters are not 0 using significant level
of $\alpha = 0{,}05$.

References

[1] Alan Agresti Categorical data analysis. 2002, Wiley –Inter Science.

[2] David W. Hosmer, Stanley Lemeshow Applied Logistic Regression Second edition, A. Wiley Interscience Publication.

[3] L. Prifti, Sh. Shehu, " Modeling Categorical data in a human health problrm" First International Conference on Applied Sciences, "New Technologies and Applications in Medicine", 7-8 Novembre 2014.