Computers and Mathematics with Applications 62 (2011) 3353–3365

Contents lists available at SciVerse ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

# Fuzzy logistic regression based on the least squares approach with application in clinical studies

Saeedeh Pourahmad [a], Seyyed Mohammad Taghi Ayatollahi [a,*], S. Mahmoud Taheri [b], Zahra Habib Agahi [c]

[a] Department of Biostatistics, School of Medicine, Shiraz University of Medical Sciences, Shiraz 71345-1874, Iran
[b] Department of Mathematical Sciences, Isfahan University of Technology, Isfahan 84156-83111, Iran
[c] Department of Rheumatology, School of Medicine, Shiraz University of Medical Sciences, Shiraz 71345-1874, Iran

ARTICLE INFO

*Article history:*
Received 6 November 2009
Received in revised form 23 May 2011
Accepted 22 August 2011

*Keywords:*
Fuzzy logistic regression
Possibility of success
Linguistic term
Possibilistic odds
Capability index
Clinical vague status

ABSTRACT

To model fuzzy binary observations, a new model named *"Fuzzy Logistic Regression"* is proposed and discussed in this study. In fact, due to the vague nature of binary observations, no probability distribution can be considered for these data. Therefore, the ordinary logistic regression may not be appropriate. This study attempts to construct a fuzzy model based on *possibility* of success. These possibilities are defined by some linguistic terms such as …, *low, medium, high*…. Then, by use of the Extension principle, the logarithm transformation of *"possibilistic odds"* is modeled based on a set of crisp explanatory variables observations. Also, to estimate parameters in the proposed model, the least squares method in fuzzy linear regression is used. For evaluating the model, a criterion named the *"capability index"* is calculated. At the end, because of widespread applications of logistic regression in clinical studies and also, the abundance of vague observations in clinical diagnosis, the suspected cases to Systematic Lupus Erythematosus (SLE) disease is modeled based on some significant risk factors to detect the application of the model. The results showed that the proposed model could be a rational substituted model of an ordinary one in modeling the clinical vague status.



## 1. Introduction

### 1.1. Categorical variables and vague observations

Nowadays, vague or non-precise observations include a large amount of research data. Depending on the kind of variables, vague observations arise for different reasons. In quantitative variables which are measured in terms of real numbers, no suitable or developed measuring instruments may lead to non-precise observations. Also, sometimes, unavailability of original characteristics may cause approximate measures.

In contrast, qualitative variables which express a qualitative attribute, take on a finite number of codes. These codes do not imply numerical properties and refer to the distinct categories of the variable. The definition of categories is very important in categorical variables. Ambiguous definition may cause confusion in classification. Indeed, borderlines of categories are not crisp and cases near the categories' borderline have a vague status. In addition, some categorical variables are inherently measured by fuzzy scale. For example, the observations are described by linguistic terms such as *large, heavy*,

* Corresponding author. Tel.: +98 7112349330; fax: +98 7112349330.
*E-mail address:* ayatolahim@sums.ac.ir (S.M.T. Ayatollahi).

0898-1221/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2011.08.050

or *approximately equal to five*. The best description of these kinds of observations is that they are fuzzy outputs. Modeling the relationship between these observations and making a prediction under the fuzzy environment is a challenge for the classical modeling analysis.

### 1.2. Motivation of study

Logistic regression analysis is one of the famous non-linear methods used to model the binary response variable based on ordinary explanatory variables. This method is particularly appropriate for models involving disease state (diseased/healthy), patient survival (alive/dead) and decision making (yes/no). So, it is widely used in studies in the health sciences [1,2].

Problems arise in classical logistic regression in situations such as

(1) Contravention of distribution assumptions (Bernoulli probability distribution for the binary response variable, uncorrelated explanatory variables, independency and identically distribution of error terms … ).

(2) Low sample size.

(3) Vagueness in the relationship between variables which do not follow the random error patterns in logistic regression models, and

(4) Non-precise observations.

In fact, non-precise or vague observations, which occur frequently in practice, may cause all other difficulties.

Consider clinical researches as an example; for some diseases, there are no biological examinations and the disease is diagnosed by some defined and wholly accepted criteria. To distinguish patients in these diseases, cases which have some of those defined criteria (not all of them) have a vague status. Lupus[1] and Behcet[2] are the examples in this field [3]. Also, in hypertension, determining a point[3] for blood pressure as a crisp borderline to detect the patients is not rational. Furthermore, some variables such as pain severity or disease severity are described by linguistic terms such as low, medium and high.

Widespread applications of logistic regression in clinical studies and also, the abundance of vague observations in clinical diagnosis, motivate us to think about a proper substituted model in a fuzzy environment. Section 2 makes a brief review on preliminary theory rules which are useful for introducing the new model. Section 3 discusses the proposed model in details. A goodness- of- fit criterion is used in this section. To detect the application of our proposed model, a real numerical example in clinical research is used in Section 4. Finally in Section 5, a discussion on the presented model is drawn.

## 2. Underlying preliminary theory for the proposed model

### 2.1. Fuzzy linear regression model

Regression analysis is a powerful and comprehensive methodology to model the relationships between a response variable, called the dependent variable, and one or more explanatory variables called independent variables.

By the classical statistical technique, the observations, either the response variable or the explanatory variables, are required to follow certain probability distributions [4]. In addition, deviations between the observed and the estimated values are assumed to be due to random errors. But, by the fuzzy observations, the distribution assumptions do not hold. Also, in this case, one cannot assign all the uncertainty of model to the randomness aspect of variables.

In the analysis of fuzzy data, the study of the relationship between a set of (crisp or fuzzy) independent variables and one or more dependent fuzzy variables is an important issue. Such a relationship can be studied, for instance, through fuzzy linear regression techniques [5]. Fuzzy regression analysis is an extension of (or an alternative for) the classical regression analysis in which some elements of the model are represented by fuzzy numbers. The uncertainty in this type of regression model becomes fuzziness, not randomness. This aspect of uncertainty is called ''possibility'' [6]. The article by Tanaka et al. [7] is probably the first research on fuzzy linear regression. Fuzzy regression methods have been successfully applied to various problems such as forecasting [8,9] and engineering [10].

There are two general categories of fuzzy regression analysis [11]; the first is a possibilistic regression analysis which is based on possibility concepts. Possibilistic regression analysis uses a fuzzy linear system as a regression model whereby the total vagueness of the estimated values for the dependent variables is minimized. It was first proposed by Tanaka et al. [7].

The second category of fuzzy regression analysis adopts the Fuzzy Least Squares Method (FLSM) for minimizing errors between the given outputs and the estimated outputs [12]. The advantage of Tanaka's possibilistic model is in its simplicity in programming and computation, while FLSM in its minimum degree of fuzziness between observed and estimated values [7].

### 2.2. Ordinary logistic regression model

Logistic regression is a mathematical modeling approach that is used to describe the relationship between a binary response variable and one or more explanatory variables. Explanatory variables may be continuous, discrete, and binary

[1] Systemic lupus erythematosus is a chronic autoimmune connective tissue disease that can affect any part of the body.

[2] Behcet disease can be interpreted as a chronic disturbance in the body's immune system.

[3] Systolic blood pressure $\leq$ 120 mg/dl and diastolic blood pressure $\leq$ 80 mg/dl.