Transactions on Fuzzy Sets and Systems (TFSS)
URL: https://sanad.iau.ir/journal/tfss/
Online ISSN: 2821-0131
Vol.3, No.2, (2024), 23-36
DOI: http://doi.org/ 10.xxxxx/tfss.2022.yyyyy.zzzz

Article Type: Original Research Article
(Journal Pre-proof)

Fuzzy Logistic Regression Analysis Using the Least Squares Method

Zahra Behdani* , **Majid Darehmiraki**

Abstract. One of the most efficient statistical tools for modeling the relationship between a dependent variable and several independent variables is regression. In practice, observations relating to one or more variables, or the relationship between variables, may be vague or non-specific. In such cases, classic regression methods will not have enough capability to model data, and one of the alternative methods is regression in a fuzzy environment. The fuzzy logistic regression model provides a framework in the fuzzy environment to investigate the relationship between a binary response variable and a set of covariates. The purpose of this paper is to attempt to develop a fuzzy model that is based on the idea of the possibility of success. These possibilities are characterized by several linguistic phrases, including low, medium, and high, among others. Next, we use a set of precise explanatory variable observations to model the logarithm transformation of "possibilistic odds." We assume that the model's parameters are triangular fuzzy numbers. We use the least squares method in fuzzy linear regression to estimate the parameters of the provided model. We compute three types of goodness-of-fit criteria to evaluate the model. Ultimately, we model suspected cases of Systemic Lupus Erythematosus (SLE) disease based on significant risk factors to identify the model's application. We do this due to the widespread use of logistic regression in clinical studies and the prevalence of ambiguous observations in clinical diagnosis. Furthermore, to assess the prevalence of diabetes in the community, we will collect a sample of plasma glucose levels, measured two hours after a meal, from each participant in a clinical survey. The proposed model has the potential to rationally replace an ordinary model in modeling the clinically ambiguous condition, according to the findings.

AMS Subject Classification 2020: 62J86; 62J07
Keywords and Phrases: Least square, Distance measure, Logistic regression.

1 Introduction

Regression is one of the most efficient statistical tools for modeling the relationship between a dependent variable and one or more independent variables. Regression analysis primarily aims to identify the functional relationship between the dependent variable and the independent variable, enabling control over the dependent variable's values or future prediction. The standard model for statistical linear regression is as follows:

$$V_i = w_0 + w_1 u_{i1} + \cdots + w_n u_{in} + \epsilon_i, \quad i = 1, \cdots, p \tag{1}$$

where V_i is the dependent variable for the $i-$th observation and u_{ij} is the value of the $j-$th independent variable in the $i-$th sample observation and w_j are the coefficients of the independent variables in the regression function or model parameters. These parameters are based on a sample of observations and The

∗**Corresponding Author:** Zahra Behdani, Email: behdani@bkatu.ac.ir, ORCID: 0000-0002-0029-4833
Received: 16 May 2024; **Revised**: 20 August 2024; **Accepted**: 21 August 2024; **Available Online**: 8 September 2024; **Published Online**: 7 November 2024.

How to cite: Behdani Z, Darehmiraki M. Fuzzy logistic regression analysis using the least squares method. *Transactions on Fuzzy Sets and Systems.* 2024; 3(2): 23-36. DOI: http://doi.org/10.xxxxx/xxxx.yyyyy.zzzz

basis of statistical methods is estimated. In practice, observations of variables or their relationships can be vague or non-precise. In such cases, classical regression methods will not have enough capability to model the data. In such cases, one of the alternative methods of classical regression is fuzzy regression, or, in other words, regression in a fuzzy environment.

While linear regression models have dominated most existing studies in this field, sometimes the relationships between variables are too complex to model and analyse using a linear relationship. In the category of non-linear models, some models are inherently linear. In other words, appropriate transformations can linearise the relationship between model variables. One of these models is the logistic regression. We use this to model the relationship between a binary dependent variable and one or more independent variables. When defining the dependent variable's classes, we code the desired condition with the number one and the opposite class with the number zero. This model has many applications in various scientific fields, including health and medical studies. For instance, it models disease status (sick or healthy) and patient survival (death or survival).

Many scientific studies use imprecise observations, but the logistic regression model, like other statistical models, uses precise observations to fit the model. It is impossible to verify model assumptions with imprecise observations or small sample sizes. Any violation of these assumptions makes using the logistic model unreasonable. The previous discussion introduced us to the concept of modeling in a fuzzy environment. Because fuzzy modeling has more flexibility, it works well, especially when the sample size is small or the observations and relationships between variables are imprecise and approximate.

In 1965, Zadeh first introduced fuzzy sets [21]. Subsequently, Tanaka and his colleagues [1] engaged in a debate on the subject of fuzzy regression. Tanaka assumed that the data consisted of triangular fuzzy integers and proceeded to estimate the regression coefficients by minimising a fuzzy index. Tanaka based his work on mathematical programming methodologies. In the same year, Yager [19], with a different approach, predicted the value of the dependent variable, the simplest form of fuzzy regression, in her contract with fuzzy observations. Jajoga [9] calculated the linear regression coefficients using a generalised version of the least squares method, while until then most of the fuzzy regression models were analysed using the mathematical programming method. Celmins [3] proposed a method for fitting a multivariate fuzzy model by minimising a least squares objective function and presented a least squares method for fuzzy regression models. Diamond [5] introduced a distance measure on the set of fuzzy numbers and used it to define the least squares criterion. In general, there are three main methods for analysing fuzzy linear regression models:

- Fuzzy least squares methods,
- Mathematical programming methods,
- Numerical methods (simulation or iteration).

Pourahmad et al. [14, 15] investigated fuzzy logistic regression from two perspectives: possibility and least squares. Namdari et al. [13] conducted a study on using fuzzy logistic regression models to analyze data with crisp input and fuzzy output. The study assessed the imprecision of the dependent variable using linguistic terminology. Their study primarily focused on the development of the least absolute deviations approach for modeling, followed by a comparison of the obtained findings to those derived from the least squares estimate method. In their work, the authors of reference [8] proposed a method for calculating the integral distance of cut sets. Additionally, they introduced a fuzzy adjustment term to reduce the likelihood of significant fuzzy errors in the fuzzy output, mainly when representing the independent variables as crisp integers. The least squares approach yields the parameters of the fuzzy logistic regression model. Mustafa et al. [12] proposed a fuzzy probabilistic logistic model that utilizes trapezoidal membership functions. Salmani et al. [17] proposed a fuzzy regression model that integrates fuzzy covariates to address the issue of erroneous binary-based response variables. The researchers used a least-squares methodology to estimate the model's parameters,

and then used a bootstrap technique to compute confidence intervals and test hypotheses about the model parameters. Salmani et al. [16] suggested three ways to measure the goodness-of-fit in logistic regression models: the Mean Squared Error (MSE), the Akaike Information Criterion (AIC), and C_p. The authors created a forward model selection method for fuzzy logistic regression that takes into account fuzzy sets' efficiency level and mean squared error.

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 or healthy), patient survival (alive or dead), and decision-making (yes or no). Therefore, studies in the health sciences widely use it (for more details refer to Bagley et al. [2]). Classical logistic regression encounters problems such as (1) Violation of distribution assumptions (Bernoulli probability distribution for the binary response variable, uncorrelated explanatory variables, independence, and identically distributed error terms). (2) Low sample size. (3) Vagueness in the relationship between variables that 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 the other difficulties. Take clinical research as an example; certain diseases lack biological examinations, and their diagnosis relies on well-defined and widely accepted criteria. To distinguish patients with these diseases, cases with some of those defined criteria (but not all of them) have a vague status. Lupus 1 and Behcet 2 are examples in this field [10]. In the case of hypertension, it is not rational to use a blood pressure threshold of 3 as a precise borderline to identify the patient. Furthermore, linguistic terms such as low, medium, and high describe some variables, such as pain severity or disease severity.

The main contributions of this paper are the creation of a fuzzy multiple linear least squares logistic regression model, the sharing of computational formulas for figuring out regression parameters, and the addition of a similarity measure between LR-type fuzzy numbers to test how well the proposed model works. We structure the remaining sections of this paper as follows: In Section 2, we provide some established findings about LR-type fuzzy numbers. We talk in depth about the suggested distance measurements between LR-type fuzzy numbers and show how to use computers to find regression parameters in Section 3. In Section 4, we show how the suggested model performs with two numerical instances. In the last section, we briefly summarize our results and provide directions for further research.

2 Preliminaries of Fuzzy Arithmetic

In 1965, Professor Zadeh proposed the concept of fuzzy sets and partial membership for sets whose boundaries are not completely clear. He introduced the concept of a fuzzy set as a collection of objects that belong to the set with a degree between 0 and 1, where degree 1 indicates complete membership and degree 0 indicates complete non-membership in the set. The membership function, which assigns a number from the interval $[0, 1]$ to each object, served as the basis for this definition.

Definition 2.1. The fuzzy set $\tilde{A}$ of $\mathbf{R}$ is called a fuzzy number if it applies in the following three conditions:

- $\tilde{A}$ is normal, it means that there exists exactly one $x \in \mathbf{R}$ such that $\tilde{A}(x) = 1$.
- $\tilde{A}$ is upper semicontinuous, that is, all $\alpha-$cuts of that interval are closed.
- The support $\tilde{A}$ is bounded.

Definition 2.2. A fuzzy number $\tilde{A}$ is called an LR fuzzy number if the membership function of $\tilde{A}$ is as follows:

$$A(x) = \begin{cases} L(\frac{m-x}{s_l}), & x < m, \\ R(\frac{x-m}{s_r}), & x \geq m \end{cases}$$

where s_l, $s_r > 0$ and $L, R : [0, \infty) \to [0, 1]$ are continuous, decreasing and invertible functions on $[0, 1]$ and also $L(0) = R(0) = 1$ and $L(1) = R(1) = 0$. We call m, s_l and s_r the center, left width, and right width of the fuzzy number $\tilde{A}$, respectively. For simplicity, we denote $\tilde{A}$ by $\tilde{A} = (m, s_l, s_r)$. If $L = R$ and $s = s_l = s_r$, $\tilde{A}$ is called a symmetric fuzzy number and we denote it by $\tilde{A} = (m, s_l, s_r)$. A fuzzy number with reference functions $L(x) = R(x) = max\{0, 1 - x\}$ is called a triangular fuzzy number and we denote it by $\tilde{A} = (m, s_l, s_r)$.

Definition 2.3. For two fuzzy numbers $\tilde{A} = (m, s_l, s_r)$ and $\tilde{B} = (n, t_l, t_r)$, we will have:

- $\tilde{A} + \tilde{B} = (m + n, s_l + t_l, s_r + t_r)$.
-

$$\lambda\tilde{A} = \begin{cases} (\lambda m, \lambda s_l, \lambda s_r), & \lambda > 0, \\ (\lambda m, \lambda s_r, \lambda s_l), & \lambda < 0. \end{cases}$$

Since one of the methods of solving regression models is to use the least squares estimator, and in this method we need to calculate the distance between two fuzzy numbers, we must define the measure of the distance between two fuzzy numbers. Researchers in this field have so far expressed different measures to calculate the distance between two fuzzy numbers, which can be referred to [11] for further study. Here, we improve the distance measure that Li et al. [11] stated so that the distance between two fuzzy numbers can be calculated at different levels of decision-making. One of the benefits of this improved interval is that we can have a model suitable for the same level of decision-making for the problem data by choosing the appropriate parameter values. The distance measure that was defined by Li et al. [11] for two fuzzy numbers $\tilde{A} = (m, s_l, s_r)$ and $\tilde{B} = (n, t_l, t_r)$,is as follows:

$$D(\tilde{A}, \tilde{B})^2 = \alpha_0(m - n)^2 + \alpha_1(s_l - t_l)^2 + \alpha_2(s_r - t_r)^2 + 2(m - n)(\alpha_3(s_r - t_r) - \alpha_4(s_l - t_l)), \tag{2}$$

Its specific modes

$$\alpha_0 = 3,\ \alpha_1 = \lambda^2,\ \alpha_2 = \rho^2,\ \alpha_3 = \rho\, and\, \alpha_4 = \lambda$$

and

$$\alpha_0 = 1,\ \alpha_1 = \lambda_2,\ \alpha_2 = \rho_2,\ \alpha_3 = \rho_1\, and\, \alpha_4 = \lambda_1,$$

that results in the measures of the distance defined by Yang and Ko [20] and Diamond and Korner [6], respectively. where

$$\lambda = \int_0^1 L^{-1}(q)dq, \qquad \lambda_1 = \frac{1}{2}\int_0^1 |L^{-1}(q)|dq, \qquad \lambda_2 = \frac{1}{2}\int_0^1 |L^{-1}(q)|^2 dq,$$
$$\rho = \int_0^1 R^{-1}(q)dq, \qquad \rho_1 = \frac{1}{2}\int_0^1 |R^{-1}(q)|dq, \qquad \rho_2 = \int_0^1 |R^{-1}(q)|^2 dq$$

The least squares method uses minimizing the sum of squared errors as a fit criterion. Fuzzy least squares methods are also based on the lowest degree of difference between the observed values and the fitted values. In the following, we use the least squares method to estimate the parameters of the logistic regression model. In this method, we use the meter introduced in relation 2 to measure the error sentences and the distance between the observed and fitted fuzzy numbers. This meter is an extended version of the previous meters Yang and Ko [20] and Diamond and Krner [6] talked about here.

3 Fuzzy Logistic Regression

Consider a regression model in which the dependent variable has a binary state, such as illness or health, death or life, buying or not buying, going bankrupt or not going bankrupt, etc. Initially, medical applications utilized this model primarily to predict the likelihood of a disease's occurrence. Today, it finds widespread use across all scientific fields. Logistic regression can be a suitable model for such situations.
The logistic regression model can be considered a generalized linear model that uses the logit function as a link function, and its error follows a polynomial distribution. When the response variable follows a binomial distribution, we use binary logistic regression as a statistical method. This approach models a linear combination of explanatory variables using a function known as the "logit". The logit function is defined as the natural logarithm of the ratio of the probability of success (π) to the probability of failure ($1-\pi$). The following mathematical representation can express the association between the independent and dependent variables in the context of logistic regression:

$$V_i = logit(\pi_i) = Ln\left(\frac{\pi_i}{1-\pi_i}\right) = w_0 + w_1u_{i1} + \cdots + w_nu_{in}, \quad i = 1, \cdots, p \tag{3}$$

This study will primarily examine a scenario where the explanatory factors represent crisp values, but the dependent variable is imprecise and quantified using language phrases. The definition of "possibilistic odds," as provided by Pourahmad et al. [14], will be presented in the subsequent definition.

Definition 3.1. Let μ_i represent the probability of seeing feature 1 or success, denoted as $V_i = 1$, for the ith example in a sample of size n. The eventuality of achieving success for the selected feature is determined by a linguistic word, $\mu_i \in \{\cdots, low, average, high, \cdots\}$. We can use expert-defined fuzzy numbers to accurately represent each term of a linguistic variable. It is important to provide precise definitions for these words in a manner that ensures the collective range of their respective supports encompasses the whole of the interval $(0, 1)$. The ratio $\frac{\mu_i}{1-\mu_i}$ is regarded as the possibilistic odds of the ith scenario, indicating the eventuality of success in relation to the eventuality of failure.

For instance, triangular fuzzy numbers, which are designed to represent the eventuality of success as $\mu = (Verylow, Low, Medium, High, Veryhigh)$, are provided in equation (4) and visually shown in Figure 1.

$$\begin{aligned} &Very\,low = (0.01, 0.02, 0.18),\ Low = (0.1, 0.25, 0.40),\ Medium = (0.35, 0.50, 0.65), \\ &High = (0.6, 0.75, 0.90),\ Very\ High = (0.8, 0.98, 0.90) \end{aligned} \tag{4}$$

3.1 Introducing the Model

The logistic regression model is a generalized linear model that uses the logit function as the dependent variable and a binomial distribution for the error sentences. The following diagram illustrates this model:

$$v_i = w_0 + w_1u_{i1} + \cdots + w_ku_{ik} + \epsilon_i, \quad i = 1, \cdots, n \tag{5}$$

Remark 3.2. According to the second part of Definition 2.3, there are differences when the coefficients of fuzzy numbers are positive or negative. Therefore, according to this definition and applying changes, calculations for negative coefficients can also be considered, but in this article, calculations based on positive coefficients are considered.

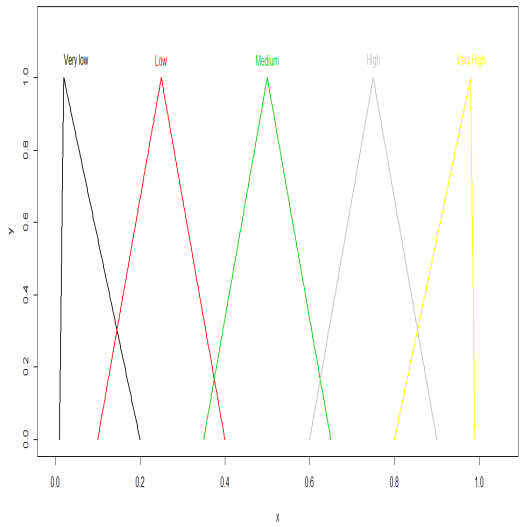

Figure 1: The membership functions of triangular fuzzy numbers represent the eventuality of success as $\mu = (Verylow, Low, Medium, High, Veryhigh)$

For the fuzzy model, consider the set of observations $U_i = (u_{i1}, u_{i2}, ..., u_{ik})$, where U_i is the non-fuzzy observation vector of covariates for the ith case. We indicate the observation of the corresponding answer with v_i, which is a number between 0 and 1, and it shows the possibility of having a desirable characteristic for the ith case. Consequently, we present the fuzzy logistic regression model with fuzzy coefficients as follows:

$$\tilde{v}_i = \tilde{w}_0 + \tilde{w}_1 u_{i1} + \cdots + \tilde{w}_k u_{ik} + \epsilon_i, \quad i = 1, \cdots, n \tag{6}$$

$\tilde{w}_j$, $j = 0, 1, ..., k$ are the parameters of the model, which are assumed to be triangular in fuzzy number $\tilde{w}_j = (w_j, l_j, r_j)_T$ calculations for simplicity. $\tilde{v}_i = ln\dfrac{\mu_i}{1-\mu_i}$ is the probability logarithmic transformation estimator, so based on the properties of addition and subtraction of triangular fuzzy numbers, $\tilde{V}_i$ will also be a triangular fuzzy number in the form of $\tilde{V}_i = (f_{ic}(u), f_{il}(u), f_{ir}(u))$, which:

$$\begin{aligned} f_{ic}(u) &= w_0 + w_1 u_{i1} + \cdots + w_k u_{ik}, \\ f_{il}(u) &= l_0 + l_1 u_{i1} + \cdots + l_k u_{ik}, \\ f_{ir}(u) &= r_0 + r_1 u_{i1} + \cdots + r_k u_{ik}, \end{aligned} \tag{7}$$

Therefore, the fuzzy estimated output membership function is obtained as follows:

$$\tilde{V}_i(v_i) = \begin{cases} 1 - \dfrac{f_{ic}(u) - v_i}{f_{il}(u)}, & f_{ic}(u) - f_{il}(u) \le v_i \le f_{ic}(u) \\ 1 - \dfrac{v_i - f_{ic}(u)}{f_{ir}(u)}, & f_{ic}(u) \le v_i \le f_{ic}(u) + f_{ir}(u) \end{cases} \tag{8}$$

As mentioned, $\tilde{V}_i$ is the natural logarithm of the probability of having the desired property for the observed ith case. According to the expansion principle, if $\tilde{N}$ is a fuzzy number with the membership function $\tilde{N}(x)$

and $f(x) = \exp(x)$, then $f(\tilde{N}) = \exp(\tilde{N})$ is a fuzzy number with the following membership function:

$$\exp(\tilde{N}(x)) = \begin{cases} \tilde{N}(ln(x)), & x > 0 \\ 0, & o.w. \end{cases} \tag{9}$$

Therefore, after estimating the coefficients of the model, the probability membership function of $\exp(\tilde{V}_i(x))$, $x > 0$ can be defined as follows:

$$\exp(\tilde{V}_i(u)) = \tilde{V}_i(ln(u)) = \begin{cases} 1 - \dfrac{f_{ic}(u) - ln(u)}{f_{il}(u)}, & f_{ic}(u) - f_{il}(u) \le ln(u) \le f_{ic}(u) \\ 1 - \dfrac{ln(u) - f_{ic}(u)}{f_{ir}(u)}, & f_{ic}(u) \le ln(u) \le f_{ic}(u) + f_{ir}(u) \end{cases} \tag{10}$$

Therefore, for a new fuzzy observed case, its probability is predicted as a fuzzy number using the odds model.

3.2 Estimation of Parameters

We consider the regression model to be a logistic model with fuzzy output, regression coefficients, and a non-fuzzy input vector (independent variables). To estimate the coefficients, we use the least squares error method, which uses the distance measure introduced in Equation 2. To achieve this goal, we will estimate the parameters by minimizing the following relationship:

$$S(\tilde{w}) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} D^2(\tilde{v}_i, \tilde{V}(u_i)) \tag{11}$$

which is defined as following:

$$\sum_{i=1}^{n} \left(\alpha_0 \left(v_i - w_0 - \sum_{j=1}^{k} w_j u_{ji} \right)^2 + \alpha_1 \left(v_{l_i} - l_0 - \sum_{j=1}^{k} l_j u_{ji} \right)^2 + \alpha_2 \left(v_{r_i} - r_0 - \sum_{j=1}^{k} r_j u_{ji} \right)^2 \right)$$
$$+ \sum_{i=1}^{n} \left(2 \left(v_i - w_0 - \sum_{j=1}^{k} w_j u_{ji} \right) \left(\alpha_3 \left(v_{r_i} - r_0 - \sum_{j=1}^{k} r_j u_{ji} \right) - \alpha_4 \left(v_{l_i} - l_0 - \sum_{j=1}^{k} l_j u_{ji} \right) \right) \right)$$

In order to minimize $S(\tilde{w})$, the partial derivatives of S with respect to the primal variables w_j, r_j, l_j, $j = 1, 2, ..., k$ have to vanish for optimality. To compress the above relationships, we use the matrix symbol as follows.

$$\begin{aligned} S(\tilde{w}) &= \sum_{i=1}^{n} \epsilon_i^2 = \epsilon' \epsilon \\ &= \left(\alpha_0 (V - XW)'(V - XW) + \alpha_1 (L - XS)'(L - XS) + \alpha_2 (R - XP)'(R - XP) \right) \\ &+ \left(2(V - XW)' \left(\alpha_3 (R - XP) - \alpha_4 (L - XS)' \right) \right) \end{aligned}$$

where in

$$V = \begin{bmatrix} V_1 \\ V_2 \\ . \\ . \\ . \\ V_n \end{bmatrix} \quad R = \begin{bmatrix} r_1 \\ r_2 \\ . \\ . \\ . \\ r_n \end{bmatrix} \quad L = \begin{bmatrix} l_1 \\ l_2 \\ . \\ . \\ . \\ l_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ . & . & . & \cdots & . \\ . & . & . & \cdots & . \\ . & . & . & \cdots & . \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$W = \begin{bmatrix} w_0 \\ w_1 \\ . \\ . \\ . \\ w_k \end{bmatrix} \quad S = \begin{bmatrix} w_{l_0} \\ w_{l_1} \\ . \\ . \\ . \\ w_{l_k} \end{bmatrix} \quad P = \begin{bmatrix} w_{r_0} \\ w_{r_1} \\ . \\ . \\ . \\ w_{r_k} \end{bmatrix}$$

where W and S, P are the central values and the left and right bounds for the unknown parameters. The X-matrix is a matrix of observations, and the first column is one. We have also observed the V, L, and R vectors, which represent central values, left width, and right width, respectively, for possible odds values. The following is the estimate of the least squares of the unknown parameters obtained by deriving the above expression:

$$\begin{aligned} \hat{P} &= (X'X)^{-1}X' \left(d(\alpha_2 - \alpha_1\alpha_3)R + ((\alpha_1\alpha_0 - \alpha_4^2) - d\alpha_3)Y\right)/b \\ \hat{W} &= (X'X)^{-1}X' \left((\alpha_2 R - \alpha_3 Y) - \alpha_2\hat{P}\right)/ - \alpha_3 \\ \hat{S} &= (X'X)^{-1}X' \left((\alpha_1 L + \alpha_4 Y) - \alpha_4\hat{W}\right)/\alpha_1 \end{aligned} \tag{12}$$

where $d = (\alpha_1\alpha_0 - \alpha^2)/\alpha_3$ and $b = \alpha_2 d - \alpha_3\alpha_1$.

3.3 Goodness of Fit Criteria

To evaluate the model, there are many criteria for the goodness of fit. In this article, we will use the following two criteria to evaluate the model:

$$S = \frac{1}{n}\sum_{i=1}^{n} \frac{\int min\{\hat{v}_i(t), \tilde{\hat{v}}_i(t)\}dt}{\int max\{\hat{v}_i(t), \tilde{\hat{v}}_i(t)\}dt}, \quad E_1 = \frac{1}{n}\sum_{i=1}^{n} \int |\hat{v}_i(t), \tilde{\hat{v}}_i(t)|dt, \quad E_2 = \frac{1}{n}\sum_{i=1}^{n} \frac{\int |\hat{v}_i(t), \tilde{\hat{v}}_i(t)|dt}{\int \hat{v}_i(t)dt}. \tag{13}$$

Many authors (e.g., [4, 7, 18]) commonly use these criteria for model evaluation. That S is the similarity criterion and the closer it is to one, the better. For the two criteria E_1 and E_2, the smaller value indicates a better model.

4 Numerical Example

This part uses two real-life examples from the field of medicine and clinical issues to show how well the suggested method works for estimating parameters, testing hypotheses, and figuring out confidence intervals in fuzzy logistic regression models.

Example 4.1. The data includes information about 15 people suspected of having lupus who are aged 18 to 40 years. Lupus is a chronic disease where the body's immune system, for unknown reasons, produces antibodies While the body defends itself against bacteria and viruses, it also targets its healthy organs. These attacks cause symptoms such as pain and muscle cramps. Several body organs, such as the skin, joints, kidneys, heart, and nervous system, are involved in this type of disease at the same time [13]. This disease takes several months or even several years to show its symptoms. Therefore, there is no specific test to identify it. Doctors must gather the required information from various sources, such as a person's medical history, laboratory test results, and some external symptoms. This disease is diagnosed based on its symptoms. Early detection accelerates treatment and prevents disease progression. Generally, lupus disease is defined as a set of 11 symptoms, and a person with at least four symptoms is considered a patient. Here, we categorize the degree of illness in the patient group based on the quantity of symptoms.

This study aims to model the status of people suspected of having lupus based on several important risk factors. The fitted model estimates each person's potential risk of contracting the disease. Past research has identified risk factors such as exposure to sunlight, family history, and various laboratory tests like ANA and DNA-Anti. In addition, in ESR, we use these special blood tests to diagnose lupus. We can summarise the introduction of these tests by stating that the nucleus of living cells consists of a significant quantity of chemicals known as RNA and DNA. The term ANA, or anti-nuclear antibody, literally translates to "anti-nuclear substance of the cell." These substances can damage and destroy cells and tissues. DNA-Anti also means special anti-DNA immune cells. For these two tests, the unit of measurement is defined by the number of these substances per millilitre of blood (ml/u). Their normal value is also considered to be less than 25 ml/u. In addition, ESR is a sign of inflammation. It is uncertain whether ESR increases in lupus patients, and it may be higher in women and elderly individuals. About 95 to 98% of lupus patients have a high value in the ANA test, and the amount of DNA - Anti in the blood of lupus patients increases. However, the high results of these tests alone do not indicate the presence of disease [15]. In order to model the relationship

Table 1: Doubtful cases of lupus and its risk factors

No.	Family History	Sun exposure	ANA test	Anti DNA test	ESR	Possibility of disease
1	1	1	112	105	1	High
2	0	1	80	23	0	Medium
3	0	1	115	15	0	High
4	0	1	105	107	1	High
5	0	0	89	150	1	Medium
6	1	1	160	110	1	Very High
7	0	1	100	23	0	Medium
8	0	0	100	85	1	High
9	0	1	48	83	0	Low
10	1	0	15	19	1	Very Low
11	0	0	50	91	0	Low
12	0	1	59	200	1	Medium
13	0	1	83	20	1	Low
14	0	0	15	200	0	Low
15	1	0	85	15	1	Medium

between the possibility of lupus and the risk factors mentioned in Table 1, the following model is used:

$$\tilde{y}_i = \ln \frac{\tilde{\pi}_i}{1-\tilde{\pi}_i} = \tilde{w}_0 + \tilde{w}_1 u_{i1} + \tilde{w}_2 u_{i2} + \tilde{w}_3 u_{i3} + \tilde{w}_4 u_{i4} + \tilde{w}_5 u_{i5},\ i = 1, 2, ..., 15. \tag{14}$$

We estimate the model's parameters using the least squares method with the meter introduced in the previous section. Ultimately, we calculate the parameter estimation as follows: (with $\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4) = (3, 0.25, 0.25, 0.5, 0.5)$)

$$\begin{aligned}\hat{\tilde{y}}_i =& (-4.0885, -3.3528, -0.5554)_T + (-0.7017, -0.3554, 0.4145)_T\, ESR \\ &+ (0.01033, 0.0042, -0.0045)_T\, Anti\, DNA\, test + (0.0451, 0.0168, -0.0123)_T\, ANA\, test \\ &+ (-0.1405, -0.0853, -0.0235)_T\, Sun.. + (0.2975, -0.4653, -0.8175)_T\, Fam.\end{aligned}$$

The outputs of the fitted log-odds model estimate each suspect's probability of developing lupus. We can calculate the possibility of infection for each suspected person using the principle of expansion of possible

odds. For example, for the 6th person studied with the variables $Family\,History = 1$, $Sun\,exposure = 1$, $ANA\,test = 160\,Anti\,DNA\,test = 23\,ESR = 0$ based on the estimated model, we calculate the logarithm of potential odds as follows:

$$\begin{aligned}\hat{\tilde{y}}_i =&(-4.0885, -3.3528, -0.5554)_T + (0.01033, 0.0042, -0.0045)_T\,23 + (0.0451, 0.0168, -0.0123)_T 160\\ &+ (-0.1405, -0.0853, -0.0235)_T + (0.2975, -0.4653, -0.8175)_T\end{aligned}$$

So we will have, $\hat{\tilde{V}}_6(3.72, -1.11, -3.45)_T$. This implies that the model has calculated the likelihood of lupus disease in the sixth individual as follows:

$$(0.98, 0.25, 0.03)_T$$

This is extremely close to the table's actual value. We can also use this model to predict the likelihood of disease in a new case. For example, if a person suspected of lupus presents with the following information, we can use the model to predict the likelihood of disease: She (he) has a family history of $u5 = 1$; she (he) has not been exposed to sunlight $u_4 = 0$; the results of the ANA, Anti-DNA, and ESR tests for this person were $u_3 = 110$, $u_2 = 87$, $u_1 = 0$, respectively. Given the characteristics as mentioned above, we estimate the likelihood of a disease and calculate the logarithm of its potential odds as follows: $\hat{\tilde{V}}_{new}(2.07, -1.61, -3.12)_T$ and $(0.89, 0.17, 0.04)_T$.
The fitted values for the possibility of disease, as well as the logarithm of odds, were calculated and recorded in Table 2 using the estimated model. The values of the goodness of fit indices introduced in Equation 13

Table 2: Prediction of the logarithm values of the odds possibility and the possibility of contracting lotus disease for the data in Table 1.

No.	Possibility of disease	The predicted of the logarithm odds disease	The predicted of possibility of lupus
1	High	$(1.50, -1.94, -2.83)_T$	$(0.82, 0.12, 0.05)_T$
2	Medium	$(-0.38, -2.00, -1.67)_T$	$(0.40, 0.12, 0.16)_T$
3	High	$(1.11, -1.44, --2.06)_T$	$(0.75, 0.20, 0.11)_T$
4	High	$(0.91, -1.58, -1.94)_T$	$(0.71, 0.17, 0.12)_T$
5	Medium	$(0.77, -1.59, -1.91)_T$	$(0.68, 0.17, 0.13)_T$
6	Very High	$(3.72, -1.11, -3.45)_T$	$(0.98, 0.25, 0.03)_T$
7	Medium	$(0.52, -1.66, -1.91)_T$	$(0.63, 0.16, 0.13)_T$
8	High	$(0.60, -1.67, -1.75)_T$	$(0.64, 0.16, 0.15)_T$
9	Low	$(-1.21, -2.29, -1.54)_T$	$(0.23, 0.09, 0.18)_T$
10	Very Low	$(-3.62, -3.84, -1.23)_T$	$(0.026, 0.02, 0.23)_T$
11	Low	$(-0.89, -2.13, -1.58)_T$	$(0.29, 0.10, 0.17)_T$
12	Medium	$(-0.20, -1.97, -1.79)_T$	$(0.45, 0.12, 0.14)_T$
13	Low	$(-0.98, -2.32, -1.27)_T$	$(0.27, 0.09, 0.22)_T$
14	Low	$(-1.34, -2.27, -1.64)_T$	$(0.21, 0.09, 0.16)_T$
15	Medium	$(-0.50, -2.68, -2.07)_T$	$(0.38, 0.06, 0.11)_T$

are equal to:

$$S = 0.8156 \quad E1 = 0.0194 \quad E2 = 0.1725.$$

These criteria have been calculated by using the estimated values and actual values for the response variable and placing them in Equation 13.

Example 4.2. We will use a sample of each community member's two-hour postprandial plasma glucose levels from a clinical survey to assess their diabetes condition. We discovered that 15 instances fell within the range of 140200 (mg/dl), using a cut-off point of 200 (mg/dl). To guess how likely it was that these people had diabetes, we added extra information like their gender (female), age (in years), BMI (body mass index, which is weight in kilograms divided by height in meters squared), family history (including father, mother, sister, and brother), and two-hour plasma glucose levels (measured in milligrams per decilitre), all of which have been linked to a higher risk of diabetes (see Table 4). We asked an expert to assign a probability of illness to each instance. Two-hour postprandial plasma glucose (THPPG)

Table 3: The values of associated risk variables and fuzzy binary observations in SLE disease

No.	Sex	THPPG (mg/dl)	Age(year)	Family history	BMI(kg/m2)	π
1	1	145	40	0	24	$(0.1, 0.74)_T$
2	1	147	42	0	25	$(0.15, 0.74)_T$
3	0	150	45	1	21	$(0.35, 0.82)_T$
4	0	155	37	1	23	$(0.42, 0.83)_T$
5	0	157	59	1	25	$(0.49, 0.83)_T$
6	1	160	44	0	20	$(0.50, 0.72)_T$
7	1	160	38	1	26	$(0.60, 0.90)_T$
8	1	165	52	0	33	$(0.60, 0.77)_T$
9	0	182	50	0	31	$(0.70, 0.64)_T$
10	1	187	55	1	33	$(0.85, 0.91)_T$
11	0	190	53	1	35	$(0.90, 0.86)_T$
12	0	192	62	1	30	$(0.97, 0.85)_T$
13	0	195	57	0	32	$(0.95, 0.65)_T$
14	1	195	50	0	34	$(0.95, 0.77)_T$
15	1	196	60	1	35	$(0.99, 0.92)_T$

$$\begin{aligned}\hat{\hat{V}}_i = &(-16.566, 0.018)_T + (0.476, 0.581)_T \times 0 + 0.102 \times 150 + 0.031 \times 45 \\ &+ (0.680, 1.13)_T \times 1 + (-0.0727, 0.019)_T \times 21\end{aligned}$$

This means that $\hat{\hat{V}}_3 = (0.32, 0, 82)_T$, and the logarithm of possibilistic odds for case 3 is about $(-0.77, 1.54)_T$ This model is capable of estimating the possibility odds of diabetes in a case that is suspected of having the condition. Please be aware that the estimated probability odds for each case are provided in a fuzzy format. For example, suppose we want to predict the possible disease odds for the case number 3 in Table 4. We have:

$$\begin{aligned}\hat{\hat{V}}_i = &(-16.566, 0.018)_T + (0.476, 0.581)_T \, Sex + 0.102 \, THPPG + 0.031 \, Age \\ &+ (0.680, 1.13)_T Family\, history + (-0.0727, 0.019)_T \, BMI\end{aligned}$$

The predicted values for the possibility of disease, as well as the logarithm of odds, were calculated and recorded in Table 2 using this estimated model. To assess the model, we use the three criteria suggested in Section 3.3, namely, S, E_1, and E_2.

$$S = 0.9991 \quad E1 = 0.0011 \quad E2 = 0.00087$$

Table 4: The values of associated risk variables and fuzzy binary observations in SLE disease are significant.

No.	π	The predicted of the logarithm odds disease	The predicted of possibility of disease
1	$(0.1, 0.74)_T$	$(-1.85, 1.05)_T$	$(0.13, 0.74)_T$
2	$(0.15, 0.74)_T$	$(-1.66, 1.07)_T$	$(0.16, 0.74)_T$
3	$(0.35, 0.82)_T$	$(-0.77, 1.54)_T$	$(0.32, 0.82)_T$
4	$(0.42, 0.83)_T$	$(-0.65, 1.58)_T$	$(0.340.83)_T$
5	$(0.49, 0.83)_T$	$(0.08, 1.62)_T$	$(0.52, 0.83)_T$
6	$(0.50, 0.72)_T$	$(0.09, 0.97)_T$	$(0.52, 0.72)_T$
7	$(0.60, 0.90)_T$	$(0.14, 2.22)_T$	$(0.54, 0.90)_T$
8	$(0.60, 0.77)_T$	$(-0.10, 1.22)_T$	$(0.47, 0.77)_T$
9	$(0.70, 0.64)_T$	$(1.23, 0.60)_T$	$(0.77, 0.64)_T$
10	$(0.85, 0.91)_T$	$(2.90, 2.35)_T$	$(0.95, 0.91)_T$
11	$(0.90, 0.86)_T$	$(2.53, 1.81)_T$	$(0.93, 0.86)_T$
12	$(0.97, 0.85)_T$	$(3.37, 1.71)_T$	$(0.97, 0.85)_T$
13	$(0.95, 0.65)_T$	$(2.70, 0.62)_T$	$(0.94, 0.65)_T$
14	$(0.95, 0.77)_T$	$(2.81, 1.24)_T$	$(0.94, 0.77)_T$
15	$(0.99, 0.92)_T$	$(3.83, 2.39)_T$	$(0.98, 0.92)_T$

5 Conclusion

Typically, the actual conditions of the data do not fully align with the assumed distributional properties of theoretical statistical models. This encourages academics to use fuzzy models as a means of simulating data within a more adaptable framework that closely resembles the actual circumstances of the observations. Researchers have extensively researched these models and implemented them in various fields. Undoubtedly, fuzzy models are more intricate than conventional ones regarding computation and interpretation. However, the assumptions of standard statistical models limit their utility. When the data does not meet the model assumptions, applying standard procedures is not logical because it introduces bias in the findings. Note that you cannot substitute conventional and fuzzy models for one another because of their distinct uses. Typically, it is not possible to use both of these models on the same dataset concurrently, therefore making it impossible to compare their respective outcomes.

When observations are not accurate, we advise using fuzzy modeling methods. Clinical investigations frequently reveal these characteristics. Occasionally, clinical measurement devices may exhibit mistakes. Furthermore, this research includes some ethical issues. Often, the precise magnitude of variables remains unmeasurable in such instances, leading to the reporting of observations based on approximations. The diagnosis of illness, which determines a condition based on established criteria, presents another ambiguous scenario in clinical investigations. We classify an individual as a patient if they exhibit all the signs of an illness. On the other hand, we classify an individual as healthy if they show no symptoms. What is the outcome when an individual experiences only a subset of these symptoms? The physician is unsure whether to start treatment. Furthermore, clinical laboratory tests do not provide a clear-cut threshold to distinguish between patients and healthy individuals. It implies that all people near the cut-off point have ambiguous status. To identify the primary risk factors that contribute to the disease's progression of the disease, it is not logical to rely on vague observations in the typical modeling analysis. Disregarding or neglecting these observations in the analysis is not rational. For this situation, fuzzy models appear to be suitable methods. Fuzzy logistic regression provides a framework in a fuzzy environment for investigating the relationship between a binary response variable and a set of covariates. To date, researchers have presented two general

methods to estimate the parameters in fuzzy logistic regression models: the least squares error method and the probability method, both of which use the definition of probability to estimate the parameters. The term "possible odds" refers to the ratio between the possibility of having the desired feature and not having it. This paper presents a method for estimating the parameters of the fuzzy logistic regression model using the least squares method.

Acknowledgements: We would like to thank the reviewers for their thoughtful comments and efforts towards improving our manuscript.

Conflict of Interest: The authors declare no conflict of interest.

References

[1] Asai HTSUK, Tanaka S, Uegima K. Linear regression analysis with fuzzy model. *IEEE Transactions on Systems Man Cybernet*. 1982; 12(6): 903-907. DOI: https://doi.org/10.1109/TSMC.1982.4308925

[2] Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: Standards for use and reporting with particular attention to one medical domain. *Journal of Clinical Epidemiology*. 2001; 54(10): 979-985. DOI: https://doi.org/10.1016/S0895-4356(01)00372-9

[3] Celmins A. Least squares model fitting to fuzzy vector data. *Fuzzy Sets and Systems*. 1987; 22(3): 245-269. DOI: https://doi.org/10.1016/0165-0114(87)90070-4

[4] Chachi J, Taheri SM, D'Urso P. Fuzzy regression analysis based on M-estimates. *Expert Systems with Applications*. 2022; 187: 115891. DOI: https://doi.org/10.1016/j.eswa.2021.115891

[5] Diamond P. Least squares fitting of several fuzzy variables. *InPreprints of Second IFSA World Congress, Tokyo, Japan*, 1987, pp. 329-331.

[6] Diamond P, Krner R. Extended fuzzy linear models and least squares estimates. *Computers & Mathematics with Applications*. 1997; 33(9): 15-32. DOI: https://doi.org/10.1016/S0898-1221(97)00063-1

[7] D'Urso P, Massari R, Santoro A. Robust fuzzy regression analysis. *Information Sciences*. 2011; 181(19): 4154-4174. DOI: https://doi.org/10.1016/j.ins.2011.04.031

[8] Gao Y, Lu Q. A fuzzy logistic regression model based on the least squares estimation. *Computational and Applied Mathematics*. 2018; 37: 3562-3579. DOI: https://doi.org/10.1007/s40314-017-0531-0

[9] Jajuga K. Linear fuzzy regression. *Fuzzy Sets and Systems*. 1986; 20(3): 343-353. DOI: https://doi.org/10.1016/S0165-0114(86)90045-X

[10] Klippel JH, Stone JH, Crofford LJ, White PH. *Primer on the rheumatic diseases*. Springer; 2008.

[11] Li Y, He X, Liu X. Fuzzy multiple linear least squares regression analysis. *Fuzzy Sets and Systems*. 2023; 459: 118-143. DOI: https://doi.org/10.1016/j.fss.2022.06.012

[12] Mustafa S, Asghar S, Hanif M. Fuzzy logistic regression based on least square approach and trapezoidal membership function. *Iranian Journal of Fuzzy Systems*. 2018; 15(6): 97-106. DOI: https://doi.org/10.22111/ijfs.2018.4369

[13] Namdari M, Yoon JH, Abadi A, Taheri SM, Choi SH. Fuzzy logistic regression with least absolute deviations estimators. *Soft Computing*. 2015; 19: 909-917. DOI: https://doi.org/10.1007/s00500-014-1418-2

[14] Pourahmad S, Taghi Ayatollahi SM, Taheri SM. Fuzzy logistic regression: a new possibilistic model and its application in clinical vague status. *Iranian Journal of Fuzzy Systems*. 2011; 8(1): 1-17. DOI: https://doi.org/10.22111/ijfs.2011.232

[15] Pourahmad S, Ayatollahi SMT, Taheri SM, Agahi ZH. Fuzzy logistic regression based on the least squares approach with application in clinical studies. *Computers & Mathematics with Applications*. 2011; 62(9): 3353-3365. DOI: https://doi.org/10.1016/j.camwa.2011.08.050

[16] Salmani F, Taheri SM, Abadi A. A forward variable selection method for fuzzy logistic regression. *International Journal of Fuzzy Systems*. 2019; 21: 1259-1269. DOI: https://doi.org/10.1007/s40815-019-00615-z

[17] Salmani F, Taheri SM, Yoon JH, Abadi A, Alavi Majd H, Abbaszadeh A. Logistic regression for fuzzy covariates: Modeling, inference, and applications. *International Journal of Fuzzy Systems*. 2017; 19: 1635-1644. DOI: https://doi.org/10.1007/s40815-016-0258-x

[18] Takemura K. Fuzzy logistic regression analysis for fuzzy inputoutput data. *InProceedings of the joint 2nd International Conference on Soft Computing and Intelligent Systems and the 5th International Symposium on Advanced Intelligent Systems*, 2004. pp. 1-6.

[19] Yager RR. Fuzzy prediction based on regression models. *Information Sciences*. 1982; 26(1): 45-63. DOI: https://doi.org/10.1016/0020-0255(82)90043-3

[20] Yang MS, Ko CH. On a class of fuzzy c-numbers clustering procedures for fuzzy data. *Fuzzy Sets and Systems*. 1996; 84(1): 49-60. DOI: https://doi.org/10.1016/0165-0114(95)00308-8

[21] Zadeh LA. Fuzzy sets. *Information and Control*. 1965; 8(3): 338-353. DOI: https://doi.org/10.1016/S0019-9958(65)90241-X

Zahra Behdani
Department of Mathematics and statistics
Faculty of Data Sciences and Energy, Behabahan Khatam Alanbia University of Technology
Behbahan, Iran
E-mail: behdani@bkatu.ac.ir

Majid Darehmiraki
Department of Mathematics and statistics
Faculty of Data Sciences and Energy, Behabahan Khatam Alanbia University of Technology
Behbahan, Iran
E-mail: darehmiraki@bkatu.ac.ir