Using fuzzy logistic regression for modeling vague status situations: Application to a dietary pattern study

S. Mahmoud Taheri\textsuperscript{a}, Alireza Abadi\textsuperscript{b}, Mahshid Namdari\textsuperscript{c,d,*}, Ahmad Esmailzadeh\textsuperscript{e} and Parvin Sarbakhsh\textsuperscript{f}
\textsuperscript{a}Faculty of Engineering Science, College of Engineering, University of Tehran, Tehran, Iran
\textsuperscript{b}Department of Health and Community Medicine, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran
\textsuperscript{c}Department of Community Oral Health, Dental School, Shahid Beheshti University of Medical Sciences, Tehran, Iran
\textsuperscript{d}Department of Biostatistics, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
\textsuperscript{e}Department of Community Nutrition, School of Nutrition and Food Science, Isfahan University of Medical Sciences, Isfahan, Iran
\textsuperscript{f}Department of Statistics and Epidemiology, School of Public Health, Tabriz University of Medical Sciences, Tabriz, Iran

Abstract. In some practical situations, it is not possible to categorize samples into one of two response categories because of the vague nature of the response variable. Statistical logistic regression models are, therefore, not appropriate for modeling such response variables. Moreover, the small sample size in most cases limits the use of statistical logistic regression models. Fuzzy logistic regression models, instead, can overcome these problems. In order to investigate the use of fuzzy logistic regression, the present study is designed and implemented to evaluate the relationship between dietary pattern and a set of risk factors of interest. Since it is not possible to define a healthy dietary pattern precisely, therefore, the possibility of having the healthy diet is reported for each subject as a number between zero and one. The conventional logistic model is not appropriate and fails in dealing with such imprecise data; hence, a possibilistic approach is used to model the available data and to estimate the fuzzy parameters of the model. For evaluating the model, a goodness-of-fit index and an appropriate predictive capability criterion with cross validation technique is developed. The logistic model investigated here is found to be general and inclusive enough to be recommended for modeling vague observations or ambiguous relations in any field of medical sciences.

Keywords: Fuzzy logistic regression, possibilistic odds, binary response, dietary pattern, goodness-of-fit, cross validation method

1. Introduction

In classical regression, a crisp relationship is established between crisp explanatory and response variables based on a statistical viewpoint. However, if we have imprecise observations or inexact relationships between variables, it is more natural to seek a fuzzy functional form \cite{21,26}. There are some aspects of uncertainty that measure the vagueness of the phenomena and cannot be summarized in random terms. In fuzzy regression, this kind of uncertainty is evaluated by a measure called “possibility”.

*Corresponding author: Mahshid Namdari, Department of Community Oral Health, Dental School, Shahid Beheshti University of Medical Sciences, Tehran, Iran. Tel.: +98 21 22439936; E-mail: namdari_mahshid@yahoo.com.
Over the last decades, many studies have been devoted to fuzzy regression modeling. Generally speaking, there are two main approaches to fuzzy regression analysis. The first, called 'possibilistic regression', is based on the concepts of the possibility theory [5]. Possibilistic regression analysis was initially proposed by Tanaka et al. [24]. It uses a fuzzy linear system as a regression model whereby the total vagueness of the estimated values for the dependent variable is minimized. The advantage of Tanaka’s possibilistic approach lies in the fact that it is totally derived from the concepts of fuzzy set theory not a usual extension of statistical methods in fuzzy environment. The second approach adopts the method of fuzzy least squares for minimizing errors between given and estimated outputs [34]. A review of some fuzzy linear regression methods can be found in Taheri [22] and Kelkkinna and Taheri [10].

Almost all previous studies on fuzzy regression have focused on linear models with only a few articles devoted to fuzzy nonlinear regression. A popular nonlinear regression model is the logistic regression which is appropriate for modeling categorical response variables [11]. Of extreme importance in this model is the definition of categories of response variables since ambiguous definitions often lead to confusion in classification [19]. A useful subtype of logistic regression is the binary one that is capable of describing the relationship between a binary response variable and a set of explanatory variables. In binary logistic regression, no assumptions are considered for the explanatory variables but the binary response should follow the Bernoulli probability distribution [11].

Logistic regression is frequently used in medical studies. Such studies are however faced with certain limitations such as imprecise observations. For example, it is not rational to set a crisp cut-off point as the reference for defining hypertension. In other words, since cases near the borderline have a vague status, it will be irrational to define an exact borderline. Moreover, the small size of the sample in rare diseases may often contravene the underlying distributional assumptions of statistical models or pose difficulties to check their validity. It is, therefore, natural to use a substitute model for the logistic regression in fuzzy environments. The fuzzy logistic regression introduced by Taheri and Yeganeh [23] and expounded in Pourrahmad et al. [20] seems to be the appropriate model for situations in which the real data do not match perfectly the underlying distributional assumptions of the statistical logistic models.

In this study, a fuzzy logistic regression is investigated for modeling a vague phenomenon in the field of nutrition. The possibility of having a healthy diet is reported as a number between zero and one and the model is expected to evaluate the relationship between crisp explanatory variables (such as anthropometric and biological determinants) and having a healthy dietary pattern.

The rest of this paper is organized as follows: for the sake of clarity, some preliminaries of the fuzzy set theory and fuzzy numbers are presented in Section 2. The logistic regression model and the motivation for presenting its fuzzy form accompanied by a detailed description of fuzzy logistic regression formulation and two goodness-of-fit indices are presented in Section 3. A brief description of dietary patterns is given in Section 4. Section 5 explains the application of the investigated model to the dietary patterns data. Finally, conclusions are presented in Section 6.

2. Prerequisites

In this section, the requirements necessary for the explanation of fuzzy logistic regression model are briefly described.

2.1. Fuzzy numbers

Let $X$ be a collection of objects denoted by $x$. Then, a fuzzy set $\tilde{A}$ in $X$ is defined as a set of ordered pairs: $\tilde{A} = \{(x, \tilde{A}(x)) \mid x \in X\}$, where $0 \leq \tilde{A}(x) \leq 1$ is called the amount of membership function at $x$ and it is the grade of membership of $x$ in the fuzzy set $\tilde{A}$. If $m = x \tilde{A}(x) = 1$, then $\tilde{A}$ is called normal. The complement set of fuzzy set $\tilde{A}$ is a fuzzy set $\tilde{B}$ with the membership function $\tilde{B}(x) = 1 - \tilde{A}(x)$, $\forall x \in X$. A fuzzy set of $\tilde{A}$ of $\mathbb{R}$ (real line) is said to be convex if

$$\tilde{A}(\lambda x_1 + (1 - \lambda) x_2) \geq \min(\tilde{A}(x_1), \tilde{A}(x_2)), \quad x_1, x_2 \in X, \lambda \in [0, 1].$$

The fuzzy set $\tilde{M}$ of $\mathbb{R}$ is called a fuzzy number if it is a normal convex fuzzy set of $\mathbb{R}$. This will be of LR-type if its membership function is as follows

$$\tilde{M}(x) = \begin{cases} L \left( \frac{m - x}{\alpha} \right), & x \leq m, \\ R \left( \frac{x - m}{\beta} \right), & x > m. \end{cases} \quad (1)$$

where, $L$ and $R$ are decreasing shape functions from $\mathbb{R}^+$ to $[0, 1]$ with $L(0) = 1$, $L(x) < 1$ for all $x > 0$, and $R(0) = 1$, $R(x) < 1$ for all $x > 0$. 

The advantage of Tanaka's possibilistic approach lies in the fact that it is totally derived from the concepts of the possibility theory [5].
\[ L(x) > 0 \text{ for all } x < 1 \text{ and } L(1) = 0; \text{ (or } L(x) > 0 \text{ for all } x \text{ and } L(+\infty) = 0). \]

Similar conditions hold for \( R \).

The real number \( m \) is called the mean value for \( M \), and \( \alpha \) and \( \beta \) (positive numbers) are called the right and left spreads, respectively. Symbolically, \( \tilde{M} \) is denoted by \((m, \alpha, \beta)_{LR}\). In a special case where \( L(x) = R(x) = \max(0, 1 - |x|) \), \( M \) is called a triangular fuzzy number denoted by \((m, \alpha, \beta)_{TR}\) whose membership function is:

\[
\tilde{M}(x) = \begin{cases} 
1 - \frac{m-x}{\alpha}, & m - \alpha \leq x \leq m, \\
1 - \frac{x-m}{\beta}, & m < x \leq m + \beta, \\
0, & \text{otherwise}
\end{cases} \tag{2}
\]

If \( \alpha = \beta \), then \( \tilde{M} \) is denoted by \((m, \alpha)_{TR}\) and is called a symmetric triangular fuzzy number. Algebraic operations on fuzzy numbers are defined based on the extension principle.

**Extension principle:** Let \( X \) be the Cartesian product of universes \( X_1 \times \ldots \times X_n \) and \( A_1, A_2, \ldots, A_n \) be \( n \) fuzzy sets in \( X_1, \ldots, X_n \), respectively. Suppose that \( f \) is a mapping from \( X \) to a universe \( Y \), \( y = f(x_1, \ldots, x_n) \). Then, the extension principle allows us to define a fuzzy set \( \tilde{B} \) in \( Y \) by \( \tilde{B} = \{ (y, \tilde{B}(y)) \mid y = g(x_1, x_2, \ldots, x_n), (x_1, x_2, \ldots, x_n) \in X \} \) where,

\[
\tilde{B} = \begin{cases} 
\sup_{(x_1, x_2, \ldots, x_n) \in f^{-1}(y)} \min\{\tilde{A}_1(x_1), \tilde{A}_2(x_2), \ldots, \tilde{A}_n(x_n)\}, & f^{-1}(y) \neq \emptyset, \\
0, & \text{otherwise}
\end{cases} \tag{3}
\]

in which, \( f^{-1}(y) \) is the inverse image of \( y \), i.e. \( f^{-1}(y) := \{ x \in X : f(x) = y \} \).

For the purposes of the present study, only two results on fuzzy arithmetic are needed, which are obtained based on the extension principle.

Let \( \tilde{M} = (m, \alpha, \beta)_{TR} \) be a triangular fuzzy number and \( \lambda \in \mathbb{R} \). Then, based on the extension principle, the extended scalar multiplication is obtained as

\[
\lambda \tilde{M} = \begin{cases} 
(\lambda m, \lambda \alpha, \lambda \beta), & \lambda > 0, \\
(\lambda m, -\lambda \beta, -\lambda \alpha), & \lambda < 0
\end{cases} \tag{4}
\]

Let \( \tilde{M} = (m, \alpha, \beta)_{TR} \) and \( \tilde{N} = (n, \gamma, \delta)_{TR} \) be two triangular fuzzy numbers. Then, using the extension principle, the extended sum of \( \tilde{M} \) and \( \tilde{N} \) is given by \( \tilde{M} \oplus \tilde{N} = (m, \alpha, \beta)_{TR} \oplus (n, \gamma, \delta)_{TR} = (m + n, \alpha + \gamma, \beta + \delta)_{TR} \). More details on fuzzy arithmetic can be found in e.g. [29].

### 3. Fuzzy logistic regression

In traditional statistics, binary logistic regression is appropriate for regressing a binary response variable with two categories on a set of explanatory variables \( X = (x_1, x_2, \ldots, x_p) \). In binary logistic regression, the response variable \( y_i = 0, 1, i = 1, 2, \ldots, n \) has a binomial distribution with \( E(Y_i) = P(Y_i = 1) = p_i, 0 < p_i < 1 \); so a function of mean response named “logit” \( \ln(p_i/(1-p_i)) \) is considered for modeling a linear combination of explanatory variables \( 1111 \). Since \( y_i = 0, 1 \), the expression \( p/(1-p) \) is called the odds of characteristic 1.

Like other statistical models, logistic regression depends heavily on its assumptions [19], such as:

1. Distribution assumptions (Bernoulli probability distribution for the response variable, uncorrelated explanatory variables, independent and identically distributed error terms),
2. Adequate sample size, and
3. Exact observations.

These assumptions impose some limitations in practice. Sometimes, lack of suitable instruments or well-defined criteria leads to doubts in determining the state of response variable (0 or 1) so that the individual samples cannot be categorized in one of two response categories. In such situations, the Bernoulli probability distribution cannot be considered for the response variable due to the vague status of cases relative to the response categories. The probability of the response variable belonging to category 1 \( P(Y = 1) \) cannot be computed and, therefore, the probability odds cannot be modeled [19][20]. In cases where all the observations of the binary response variable are vague with unclear status with respect to their response category, another type of uncertainty, namely possibility, is considered that is not related to randomness or probability but considers the possibility of success rather than the probability of success; so that we may model the data based on the so-called “possibilistic odds” [20]. The concept of possibilistic odds proposed in Taheri and Mirzaei Yeganeh [23] and Pourahmad et al. [20] will be discussed in the next section as the basis for the current study.

### 3.1. Formulation of fuzzy logistic regression

Based on consultation with an expert and according to the consistency degree with the known characteristic of category 1 of the response variable (1 = success and 0 = failure), a real number is determined in
the interval (0,1) for each of the fuzzy cases. This real number can be represented by \( \mu_i, i = 1, 2, \ldots, n \). So, \( \mu_i = \text{Poss}(Y_i = 1) \) is the possibility of success (i.e. having the defined characteristic of category 1). The ratio \( \mu_i/(1 - \mu_i) \) is, therefore, considered as the possibilistic odds of the \( i \)-th case. It detects the possibility of success relative to the possibility of non-success for the \( i \)-th case.

For modeling the possibilistic odds based on a set of crisp explanatory variables, the observations for the \( i \)-th case are considered as \( (x_{i1}, x_{i2}, \ldots, x_{ip}, \mu_i/(1 - \mu_i)) \). A regression model with fuzzy coefficients could be suitable for this problem. In order to be able to use the Tanaka’s possibilistic approach for estimating the parameters \([24]\), the logarithmic transformation of possibilistic odds, \( w_i = \ln(\mu_i/(1 - \mu_i)) \), can be used as the response variable. The fuzzy logistic regression model will, therefore, be as follows

\[
\tilde{W}_i = \tilde{b}_0 \oplus \tilde{b}_1 x_{i1} \oplus \ldots \oplus \tilde{b}_p x_{ip}, \quad i = 1, \ldots, n. \quad (5)
\]

where, \( \tilde{W}_i \), is the estimate of the logarithmic transformation of observed possibilistic odds \( (w_i = \ln(\mu_i/(1 - \mu_i)) \) based on the model. The fuzziness of the relationships is hidden in fuzzy coefficients. For simplicity in computation and interpretation, we assume that the parameters of the model are symmetric triangular fuzzy numbers \( \tilde{b}_j = (a_j^c, s_j) \), \( j = 0, 1, 2, \ldots, p \).

Based on fuzzy arithmetic, it is obvious that \( \tilde{W}_i \) will be a symmetric triangular fuzzy number that can be displayed with \( \tilde{W}_i = \{f_i^c, f_i^s, f_i^t\} \), where, \( f_i^c(x) = a_0^c + a_0^c x_1 + \ldots + a_0^c x_p \) and \( f_i^s(x) = s_0 + s_1 x_1 + \ldots + s_p x_p \). The membership function of \( \tilde{W}_i \) can be shown in the following way

\[
\tilde{W}_i(w_i) = \begin{cases} 
1 - \frac{f_i^c(x) - w_i}{f_i^t(x)}, & w_i \leq f_i^c(x), \\
\frac{1 - f_i^c(x) - f_i^s(x)}{f_i^t(x)}, & f_i^c(x) < w_i \leq f_i^c(x) + f_i^s(x), \\
0, & \text{otherwise}.
\end{cases} 
\]

(6)

In order to estimate the regression coefficients based on the Tanaka’s possibilistic approach, a linear programming problem should be solved in which the fuzziness of the specified model is minimized by minimizing the objective function as the sum of the spreads of the fuzzy outputs. Its minimization is performed under the constraint that all the observations of \( w_i, i = 1, \ldots, n \) have a membership degree as big as the so-called credit level in the function of the estimated fuzzy output \( \tilde{W}_i(w_i) \geq h \). The following optimization problem will then need to be solved

\[
\min Z = 2ns_0 + \sum_{j=1}^{p} \left( 2s_j \sum_{i=1}^{n} x_{ij} \right)
\]

s.t.

\[
1 - \frac{f_i^c(x) - w_i}{f_i^t(x)} \geq h \Rightarrow (1 - h)s_0 + (1 - h) \\
\sum_{j=1}^{p} s_j x_{ij} - a_0^c - \sum_{j=1}^{p} a_0^c x_{ij} \\
\geq - w_i = - \ln \frac{\mu_i}{1 - \mu_i},
\]

\[
1 - \frac{f_i^c(x) - w_i}{f_i^t(x)} \geq h \Rightarrow (1 - h)s_0 + (1 - h) \\
\sum_{j=1}^{p} s_j x_{ij} + a_0^c + \sum_{j=1}^{p} a_0^c x_{ij} \\
\geq w_i = \ln \frac{\mu_i}{1 - \mu_i},
\]

(7)

It should be noted that a suitable value must be selected for \( h \). The above linear programming problem can be solved using an efficient software such as Lingo.[12]

3.2. Estimating the possibilistic odds for a new case

According to the extension principle, if \( \tilde{M} \) is a fuzzy number and further \( f(x) = \text{exp}(x) \), then \( f(\tilde{M}) = \text{exp}(\tilde{M}) \) is a fuzzy number with the following membership function:

\[
\text{exp}(\tilde{M}(x)) = \begin{cases} 
\tilde{M}(\ln x), & x > 0, \\
0, & \text{otherwise}.
\end{cases}
\]

(8)

Once model coefficients are estimated, the logarithm of the possibilistic odds for each subject (\( \tilde{W}_i \)) can be estimated using Relation Eq. (6). The estimated value can then be used to calculate the membership function of the possibilistic odds, \( \text{exp}(\tilde{W}_i(x)) \), as follows.
dicted one is calculated of membership in the membership function of the pre-
model. The model thus obtained is then used to pre-
ing observations are used to develop a fuzzy regression
each time left out from the data set while the remain-
i the degree of membership based on the cross-validation
3.3.2. Predictive ability
For investigating the predictive ability of the model,
the membership degree of the observed values in the mem-
membership function of the estimated one. A large
the degree of membership of the i-th observation in
\begin{equation}
\exp(W_i(x)) = \tilde{W}_i(\ln(x))
\end{equation}
\begin{equation}
1 - \frac{f_i^s(x) - \ln(x)}{f_i^c(x)},
\end{equation}
\begin{equation}
1 - \frac{\ln(x) - f_i^s(x)}{f_i^c(x)},
\end{equation}
\begin{equation}
f_i^s(x) < \ln(x) \leq f_i^c(x) + f_i^s(x)
\end{equation}
\begin{equation}
0, \text{ otherwise}.
\end{equation}
For a new case with a crisp vector of explanatory ob-
the logarithm of the possibilistic odds, and
\begin{equation}
MMPA = \frac{1}{n} \sum_{i=1}^{n} PA_i
\end{equation}
where, \( PA_i = \tilde{W}_i^{(i)}(w_i) = \exp(\tilde{W}_i^{(i)}(\mu_i/(1-\mu_i))) \)
is the degree of membership for the i-th observed odds in the membership function of the predicted one. The membership function is predicted from the model which i-th observation was left out.
4. Dietary patterns
Dietary status directly relates to health and plays an important role in determining the leading causes of morbidity and mortality [28]. Many researchers have proposed dietary status to be analyzed as dietary pat-
terns. Dietary intake is assessed by using a 168-item semiquantitative food frequency questionnaire (FFQ) that consists of a list of foods with standard servings. Participants report their frequency of consumption of a given serving of each food item during the previous year on a daily (e.g. bread), weekly (e.g. rice or meat), or monthly (e.g. fish) basis. The reported fre-
regular dietary patterns are some of the well-known patterns. The healthy pattern is high in food groups such as:
fruit, vegetables, tomatoes, poultry, legumes, green leafy vegetables, tea, fruit juices, and whole grains.
The western dietary pattern is high in refined grains, red meat, butter, processed meat, high-fat dairy prod-
cuits, sweets and desserts, pizza, potatoes, eggs, hydrogenated fats, and soft drinks but low in other vegeta-
bles and low-fat dairy products. The traditional dietary pattern is supposed to be high in refined grains, pota-
toes, tea, whole-grains, hydrogenated fats, legumes, and broth [6]. For each person, the factor score for each of the dietary patterns can be calculated by summing intakes of individual food groups weighted by loadings derived from the factor analysis. So, each subject can
Fig. 1. Healthy factor score with different cut-off points.

be assigned with a factor score for each of the dietary patterns. For example, in the healthy pattern, a higher score indicates a healthier diet.

The functional forms that are used by researchers for finding the independent variables that have strong associations with dietary patterns are linear or logistic regression models. In linear regression, a linear relation is regressed between the derived factor score and the explanatory variables. Since, the scale of the factor scores is not meaningful, the results of linear regression are difficult to interpret and the interpretation of the estimated coefficients is not clinically straightforward [18]. Thus, the results of logistic regression are better interpreted and this type of regression is, therefore, preferred to the linear one. On the other hand, in the classical logistic regression, participants are divided into two groups (namely, healthy or unhealthy dietary groups) according to a crisp cut-off point (i.e. median, first, or third quartiles of the healthy factor scores). Then, the odds ratio (OR) for being above the median score or any other pre-specified cut-off point is calculated by logistic regression [18]. Dichotomizing participants according to a subjective cut-off point seems to be irrational and cases near the cut-off point may have a vague status. Moreover, since different quartiles can be selected as a cut-off point, defining an exact crisp cut-off point is a controversial issue and arbitrary. For example, Park et al. [18] calculated the odds ratios for being above the median score while others like Tseng et al. [25] and Mizoue et al. [16] calculated the odds ratios relative to the lowest quartile. Hence, categories can overlap and different results may be obtained with different cut-off points. A detailed description of the problem can be found in Fig. 1.

In practice, factor analysis is criticized for detecting patterns that may not be reproducible across populations. Moreover, its results are affected by subjective decisions, including the assignment of food items into food groups, the number of factors to be extracted, the method of rotation, and even the labeling of components [14]. We, therefore, believe that there is an inexact relationship between the state of a dietary pattern and the covariates of interest, which cannot be modeled by a crisp statistical model. Furthermore, due to the vague nature of binary observations (having either a healthy or an unhealthy dietary pattern), no probability distribution can be considered. It is, therefore, suggested that an ordinary statistical logistic model which is based on exact observations is not adequate for modeling the relationship between having a healthy dietary pattern and the explanatory variables, particularly as the probabilistic assumptions of the logistic model will not be met. This gives rise to the more natural solution of seeking a fuzzy functional relationship for modeling such cases.

In this study, the explanatory variables (such as anthropometric and biological determinants) are taken as crisp variables, and the value of the binary response variable is reported as a number between zero and one to indicate the possibility of having a healthy dietary pattern.

5. Applied example

5.1. Description of data

In order to determine the healthy dietary pattern and to evaluate the impacts of important factors, from a cross-sectional study on a representative sample of Isfahani female nurses, 16 women aged 40–60 who smoked and had no prior history of metabolic abnormalities in their relatives were selected for the study. The usual dietary intake was assessed using a validated 168-item semiquantitative food frequency questionnaire (FFQ) which consisted of a list of foods with standard serving sizes commonly consumed by Iranians [6]. Participants were asked to report their frequency of consumption of a given serving of each food item. The reported frequency for each food item was then converted to a daily intake. Portion sizes of the consumed foods were converted to grams using household measures.

To identify major dietary patterns based on the food groups of similar nutrients, factor analysis was used with the factors rotated by an orthogonal transformation. Natural interpretations of the factors in conjunction with eigenvalues \( \geq 1 \) and the scree test was used to determine whether a factor should be retained. The derived factors (dietary patterns) were subsequently labeled on the basis of our own interpretation of the data as well as reports in the literature [6,7]. Thus, three major healthy, western, and traditional dietary patterns...
were detected. This study focuses on the healthy pattern which is high in fruits, vegetables, tomatoes, poultry, legumes, green leafy vegetables, tea, fruit juices, and whole grains. The factor score for the healthy pattern was calculated for all the participants by summing up intakes of individual food groups weighted by loadings that were derived from the healthy factor. To assign a possibility of having a healthy diet, each participant’s percentile was calculated from a larger set of factor scores. The percentile which was a real number in (0,1) represented the degree of possibility or consistency for that person to have the healthy diet.

In order to predict the possibilistic odds of having a healthy diet, such information as participant’s age, BMI [weight (kg)/height (m)]², systolic and diastolic blood pressure, serum triacylglycerol concentration, and HDL cholesterol level were assayed. These variables were shown to have significant relations with dietary patterns. The participants’ data are presented in Table 1.

### 5.2. Implementation of the fuzzy logistic regression

For modeling the possibilistic odds of having a healthy dietary pattern with the explanatory variables, i.e., age, body mass index (BMI), systolic blood pressure (sbp), diastolic blood pressure (dbp), triacylglycerol level (TG), and HDL, the following possibilistic model is fitted

\[
\tilde{W} = b_0 \oplus b_1 \text{age} \oplus b_2 \text{BMI} \oplus b_3 \text{sbp} \oplus b_4 \text{dbp} \oplus b_5 \text{TG} \oplus b_6 \text{HDL}
\]

where, the regression coefficients \( \tilde{b}_j = (a_j, s_j) \tau \), \( j = 0, 1, \ldots, 6 \) are considered as symmetric triangular fuzzy numbers. In order to obtain the optimal model, the objective function and the restrictions of the optimization problem are formulated as follows

\[
Z = 2 \left( 16 s_0 + \sum_{i=1}^{16} a_{i_1} + s_2 \sum_{i=1}^{16} \text{BMI}_i \right. \\
\left. + \sum_{i=1}^{16} s_3 \text{sbp}_i + s_4 \sum_{i=1}^{16} \text{dbp}_i + s_5 \sum_{i=1}^{16} \text{TG}_i \right. \\
\left. + s_6 \sum_{i=1}^{16} \text{HDL}_i \right)
\]

s.t. (1 - \( h \)) [s0 + 40s1 + 24.78s2 + 111s3 + 69s4 + 167s5 + 42s6] - [a0 + 40a1 + 24.78a2 + 111a3 + 69s4a5 + 167a5 + 42a6] \( \geq 0.78 \)

(1 - \( h \)) [s0 + 40s1 + 24.78s2 + 111s3 + 69s4 + 167s5 + 42s6] - [a0 + 40s1 + 24.78s2 + 111a3 + 69s4a5 + 167a5 + 42a6] \( \geq -0.78 \)

\[
\vdots
\]

(1 - \( h \)) [s0 + 60s1 + 30.09s2 + 121s3 + 79s4 + 140s5 + 46s6] - [a0 + 60s1 + 30.09a2 + 140a3 + 46a6] \( \geq -1.19 \)

\[
\vdots
\]

### Table 1

Characteristics of the sample data

<table>
<thead>
<tr>
<th>No.</th>
<th>Age</th>
<th>BMI</th>
<th>Systolic blood pressure</th>
<th>Diastolic blood pressure</th>
<th>TG</th>
<th>HDL</th>
<th>( \mu_1 )</th>
<th>( \mu_1 )</th>
<th>( w_i = \ln \frac{\mu_1}{1 - \mu_1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40</td>
<td>24.78</td>
<td>111</td>
<td>69</td>
<td>167</td>
<td>42</td>
<td>0.32</td>
<td>0.46</td>
<td>-0.78</td>
</tr>
<tr>
<td>2</td>
<td>41</td>
<td>25.15</td>
<td>143</td>
<td>73</td>
<td>113</td>
<td>39</td>
<td>0.18</td>
<td>0.22</td>
<td>-1.51</td>
</tr>
<tr>
<td>3</td>
<td>42</td>
<td>23.62</td>
<td>99</td>
<td>66</td>
<td>120</td>
<td>39</td>
<td>0.28</td>
<td>0.38</td>
<td>-0.96</td>
</tr>
<tr>
<td>4</td>
<td>42</td>
<td>25.77</td>
<td>153</td>
<td>86</td>
<td>189</td>
<td>42</td>
<td>0.08</td>
<td>0.09</td>
<td>-2.44</td>
</tr>
<tr>
<td>5</td>
<td>43</td>
<td>29.52</td>
<td>99</td>
<td>71</td>
<td>110</td>
<td>42</td>
<td>0.60</td>
<td>1.47</td>
<td>0.39</td>
</tr>
<tr>
<td>6</td>
<td>44</td>
<td>30.12</td>
<td>112</td>
<td>80</td>
<td>61</td>
<td>46</td>
<td>0.79</td>
<td>3.69</td>
<td>1.31</td>
</tr>
<tr>
<td>7</td>
<td>45</td>
<td>32.86</td>
<td>130</td>
<td>77</td>
<td>162</td>
<td>35</td>
<td>0.31</td>
<td>0.44</td>
<td>-0.82</td>
</tr>
<tr>
<td>8</td>
<td>45</td>
<td>31.24</td>
<td>113</td>
<td>72</td>
<td>260</td>
<td>35</td>
<td>0.28</td>
<td>0.39</td>
<td>-0.95</td>
</tr>
<tr>
<td>9</td>
<td>46</td>
<td>26.99</td>
<td>120</td>
<td>80</td>
<td>326</td>
<td>39</td>
<td>0.44</td>
<td>0.78</td>
<td>-0.25</td>
</tr>
<tr>
<td>10</td>
<td>46</td>
<td>29.27</td>
<td>142</td>
<td>72</td>
<td>111</td>
<td>39</td>
<td>0.16</td>
<td>0.20</td>
<td>-1.62</td>
</tr>
<tr>
<td>11</td>
<td>49</td>
<td>33.71</td>
<td>109</td>
<td>73</td>
<td>247</td>
<td>48</td>
<td>0.25</td>
<td>0.33</td>
<td>-1.10</td>
</tr>
<tr>
<td>12</td>
<td>52</td>
<td>24.00</td>
<td>103</td>
<td>69</td>
<td>156</td>
<td>42</td>
<td>0.94</td>
<td>16.04</td>
<td>2.77</td>
</tr>
<tr>
<td>13</td>
<td>55</td>
<td>28.44</td>
<td>156</td>
<td>106</td>
<td>283</td>
<td>46</td>
<td>0.94</td>
<td>14.86</td>
<td>2.70</td>
</tr>
<tr>
<td>14</td>
<td>56</td>
<td>29.97</td>
<td>108</td>
<td>75</td>
<td>197</td>
<td>39</td>
<td>0.41</td>
<td>0.68</td>
<td>-0.38</td>
</tr>
<tr>
<td>15</td>
<td>58</td>
<td>36.59</td>
<td>142</td>
<td>96</td>
<td>127</td>
<td>45</td>
<td>0.95</td>
<td>19.00</td>
<td>2.94</td>
</tr>
<tr>
<td>16</td>
<td>60</td>
<td>30.09</td>
<td>121</td>
<td>79</td>
<td>140</td>
<td>46</td>
<td>0.77</td>
<td>3.30</td>
<td>1.19</td>
</tr>
</tbody>
</table>
was obtained function (Fig. 2). Finally, the following optimal model of different values of $\mu_i$ was selected by comparing the effects $h$ and the value of the objective function $Z$ will be affected by changing $h$ values.

Value of 0.6 was selected by comparing the effects of different values of $h$ on the value of the objective function ($Z$), will be affected by changing $h$ values.

The optimization problem can be solved using a suitable software such as Lingo [12]. The value of $\mu_i$ may be determined by fitting the model with different values of $h$. The spreads ($s_i$) and the value of the objective function ($\tilde{W}$), will be affected by changing $h$ values.

\begin{align}
(1 - h)[s_0 + 60s_1 + 30.09s_2 + 121s_3 + 79s_4 + 140s_5 + 46s_6] + [a_0 + 60a_1 + 30.09a_2 + 121a_3 + 79a_4 + 140a_5 + 46a_6] \geq 1.19 \\
\end{align}

(13)

The optimal model can be exploited to estimate the value of the objective function for different values of $h$.

\[ \tilde{W} = (-7.78,0)_T \oplus (0.14, 0.05)_T \text{age} \]
\[ \oplus (-0.1, 0)_T \text{BMI} \oplus (-0.05, 0)_T \text{sbp} \]
\[ \oplus (0.074, 0)_T \text{dbp} \oplus (-0.01, 0.003)_T \]
\[ \text{TG} \oplus (0.13, 0)_T \text{HDL} \]

(14)

In summary, estimated positive coefficients for an explanatory variable indicate that increasing the variable results in increased possibilistic odds of having a healthy diet while negative coefficients indicate that increasing the variable results in reduced possibilistic odds of having a healthy diet. It may be inferred from this model that being older and having a lower body mass index, a lower systolic blood pressure, a higher diastolic blood pressure, a lower triacylglycerol, and a higher HDL serum level are related to an increase in possibilistic odds of having a healthy dietary pattern.

For evaluating the obtained model, $MDM$ was calculated as follows

\[ MDM = \frac{1}{n} \sum_{i=1}^{n} \exp \left( \tilde{W}_i \left( \frac{\mu_i}{1 - \mu_i} \right) \right) = 0.72 \]

(15)

Table 2

<table>
<thead>
<tr>
<th>No.</th>
<th>$\mu_i$</th>
<th>$w_i$</th>
<th>$\tilde{W}_i$</th>
<th>$DM_i$</th>
<th>$PA_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.46</td>
<td>-0.78</td>
<td>(-0.69, 2.51)</td>
<td>0.97</td>
<td>0.91</td>
</tr>
<tr>
<td>2</td>
<td>0.22</td>
<td>-1.51</td>
<td>(-1.70, 2.44)</td>
<td>0.92</td>
<td>0.91</td>
</tr>
<tr>
<td>3</td>
<td>0.38</td>
<td>-0.96</td>
<td>(+0.04, 2.50)</td>
<td>0.60</td>
<td>0.22</td>
</tr>
<tr>
<td>4</td>
<td>0.09</td>
<td>-2.44</td>
<td>(-1.37, 2.67)</td>
<td>0.60</td>
<td>0.67</td>
</tr>
<tr>
<td>5</td>
<td>1.47</td>
<td>0.39</td>
<td>(+0.42, 2.53)</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>6</td>
<td>3.69</td>
<td>1.31</td>
<td>(+1.48, 2.47)</td>
<td>0.93</td>
<td>0.97</td>
</tr>
<tr>
<td>7</td>
<td>0.44</td>
<td>-0.82</td>
<td>(-1.92, 2.76)</td>
<td>0.60</td>
<td>0.51</td>
</tr>
<tr>
<td>8</td>
<td>0.39</td>
<td>-0.95</td>
<td>(-2.15, 3.00)</td>
<td>0.60</td>
<td>0.58</td>
</tr>
<tr>
<td>9</td>
<td>0.78</td>
<td>-0.25</td>
<td>(-1.36, 3.21)</td>
<td>0.66</td>
<td>0.67</td>
</tr>
<tr>
<td>10</td>
<td>0.20</td>
<td>-1.62</td>
<td>(-1.42, 2.69)</td>
<td>0.93</td>
<td>0.99</td>
</tr>
<tr>
<td>11</td>
<td>0.33</td>
<td>-1.10</td>
<td>(+0.17, 3.18)</td>
<td>0.60</td>
<td>0.83</td>
</tr>
<tr>
<td>12</td>
<td>16.04</td>
<td>2.77</td>
<td>(+1.53, 3.12)</td>
<td>0.60</td>
<td>0.67</td>
</tr>
<tr>
<td>13</td>
<td>14.86</td>
<td>2.70</td>
<td>(+1.27, 3.58)</td>
<td>0.60</td>
<td>0.46</td>
</tr>
<tr>
<td>14</td>
<td>0.68</td>
<td>-0.38</td>
<td>(+0.99, 3.43)</td>
<td>0.60</td>
<td>0.26</td>
</tr>
<tr>
<td>15</td>
<td>19.00</td>
<td>2.94</td>
<td>(+1.94, 3.36)</td>
<td>0.70</td>
<td>0.68</td>
</tr>
<tr>
<td>16</td>
<td>3.30</td>
<td>1.19</td>
<td>(+2.59, 3.50)</td>
<td>0.60</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Mean 0.72 0.69

Indicating good fitting of the model. The details are given in Table 2.

$MPA$ was used to evaluate the predictive ability of the model. The results are provided in Table 2. The value of $MPA$ obtained for the optimal model was equal to 0.69, indicating the good predictive ability of the model.

5.3. Prediction for a new case

The optimal model can be exploited to estimate the odds of having a healthy diet based on the observed covariates for a new case with the following data: $age = 40$, $BMI = 30$, $sbp = 112$, $dbp = 80$, $TG = 61$, $HDL = 46$. According to Relation (14), the logarithm of the possibilistic odds for this case will be (0.92, 2.25)$T$. The membership function of the possibilistic odds for this case is calculated from Relation (9):

\[ \exp(W_{new}(x)) = \begin{cases} 
1 - 0.92 - \ln(x) & \text{if } 2.25 \\
-1.33 \leq \ln(x) \leq 0.92 (0.26 \leq x \leq 2.51), \\
1 - \ln(x) - 0.92 & \text{if } 2.25 \\
0.92 < \ln(x) \leq 3.17 (2.51 < x \leq 23.81). 
\end{cases} \]

It can be said that the possibilistic odds for this case is about 2.51.
Table 3

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>OR</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-437.88</td>
<td>0.998</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>3.27</td>
<td>26.40</td>
<td>0.999</td>
</tr>
<tr>
<td>BMI</td>
<td>-1.79</td>
<td>0.17</td>
<td>1</td>
</tr>
<tr>
<td>Systolic blood pressure</td>
<td>-0.63</td>
<td>0.53</td>
<td>0.999</td>
</tr>
<tr>
<td>Diastolic blood pressure</td>
<td>2.53</td>
<td>12.51</td>
<td>0.999</td>
</tr>
<tr>
<td>TG level</td>
<td>-0.69</td>
<td>0.50</td>
<td>0.997</td>
</tr>
<tr>
<td>HDL level</td>
<td>7.78</td>
<td>239</td>
<td>0.998</td>
</tr>
</tbody>
</table>

5.4. Implementation of the classical regression

In order to compare the results of the proposed model with those of the conventional statistical logistic regression model, the second quartile of healthy factor scores was considered as the cut-off point and participants were divided into healthy and unhealthy dietary groups accordingly. Thus, a binary response variable was created. The results of the conventional logistic regression are presented in Table 3.

A relatively large sample size is needed in both response categories for fitting a statistical logistic regression since small sample size in response categories may often result in imprecise and unreliable estimates with large standard errors. Furthermore, working with a small sample size would lead to the reduced distinction power of statistical logistic regression analysis. Unreliable estimates with large standard errors and large \( p \)-values in the conventional statistical approach may be due to the inadequacy of sample size for fitting statistical models.

5.5. Discussions of the clinical example

To the best of the authors’ knowledge, no similar study has been reported using a fuzzy approach in the field of dietary pattern studies. In most previous studies, the first, second, or third quartiles of factor scores were used for defining the subgroups and ordinary logistic regression was commonly used for modeling. Naturally, the final results differ greatly with different cut-off points.

In the present study, the concept of possibilistic odds and the fuzzy logistic regression were used for evaluating the impact of a set of crisp explanatory variables on the possibilistic odds of having the healthy diet due to the imprecision of the binary response. Tanaka’s possibilistic approach was also used to establish a fuzzy regression model and to estimate the fuzzy coefficients. The advantage of Tanaka’s possibilistic approach lies in its simplicity in programming and computation. It should be noted that small sample sizes in statistical models results in estimates with large standard errors; hence, statistical logistic models do not seem appropriate for such cases.

The present data was obtained from a cross-sectional study and some intermediary events, including elevated blood pressure, could have led to changes in diet. This can justify the positive association between diastolic blood pressure and healthy dietary pattern. The findings of the present study are in agreement with those of previous studies in that they show a protective effect of healthy dietary pattern against risk factors of chronic diseases, including heart diseases \[13\]. In accordance with the results of our study in a cross-sectional study in a British population \[27\] a dietary pattern characterized by high consumption of fruit and vegetables and low consumption of processed meat and fried foods was inversely associated with features of metabolic abnormalities which are obesity, high serum triacylglycerol concentration, low HDL cholesterol level and high blood pressure. Our healthy dietary pattern was also found similar to the dietary approaches adopted for stopping hypertension eating plans, which have been recommended for decreasing blood pressure \[17\] and improving features of the metabolic syndrome \[2\]. The results further support the idea that the right choice of foods is important in weight control. In agreement with the present results, Maskarinec et al. \[15\] found a negative association between BMI and healthy dietary patterns. In conclusion, our findings support the current healthy dietary recommendations characterized by high consumption of fruits, vegetables, poultry, and legumes, which is a preventive measure against metabolic syndromes.

6. Conclusions

In practice, medical and clinical studies are usually handicapped by such problems as low sample size, imprecise data, and vague relationships. In these situations, the use of ordinary statistical models is not rational as it leads to biased results. Fuzzy regression provides an appropriate alternative tool for modeling in a more flexible environment when the assumptions of the statistical models are not established.

In this study, the fuzzy logistic regression was described and its application in a real clinical example was investigated. Tanaka’s possibilistic approach was used for the estimation of regression coefficients. The Mean Degree of Memberships (MDM) was used to evaluate the model. Moreover, using the cross-
validation method, a criterion called Mean of Predictive Ability (MPA) was introduced and applied to justify the predictive ability of the model. These two indices reveal a good fitting of the proposed model on the given numerical example. Furthermore, the applicability of the model to predict the possibilistic odds for a new case was explored.

The logistic model investigated was found to be general and can be recommended for use in similar situations in any field of medical sciences.

Acknowledgments

The authors would like to express their gratitude to the Food Security Research Center, Isfahan University of Medical Sciences for their help in data collection. Dr. S. Pourahmad also deserves the authors acknowledgment for her kind support and assistance. The referees are especially appreciated for their valuable comments which greatly improved the final version of this paper.

References


