Modeling Vague Status by Fuzzy Logistic Regression: Application in Evaluating the Effect of Folic Acid on Child's Appetite status

Mahshid Namdari[1], S. Mahmoud Taheri[2], Alireza Abadi[3*], Mansour Rezaei[4], Naser Kalantari[5]

[1]Department of Biostatistics, Student Research committee, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran

[2]Department of Engineering Sciences, College of Engineering, University of Tehran, Tehran, Iran, and Department of Mathematical Sciences, Isfahan University of Technology, 8415683111 Isfahan, Iran

[3*]Department of Health and Community Medicine, School of Medicine, Shahid Beheshti University of Medical Sciences, Tehran, Iran, e-mail: alirezaabadi@gmail.com

[4]Chronic Respiratory Diseases Research Center, National Research Institute of Tuberculosis and Lung Disease (NRITLD), Shahid Beheshti University of Medical Sciences, Tehran, Iran

[5]Department of Community nutrition, Faculty of Nutrition Sciences and Food Technology, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Abstract

Statistical logistic regression is used for modeling a binary response variable with two exact categories based on a set of explanatory variables. In practice, the state of a binary response variable may be described in linguistic terms rather than in exact ones. In these situations, the borders between linguistic terms are vague and it is not possible to categorize the samples in one of two response categories. So, no usual probability distribution can be considered for such binary response variables and, therefore, statistical logistic regression is not appropriate for modeling. In this paper the state of child's appetite is described by linguistic terms, and then a set of crisp explanatory variables which are supposed to be related with the state of appetite are collected. Fuzzy logistic regression, based on a least squares method, is used for modeling child's appetite via a set of explanatory variables. Finally, the obtained model is evaluated by the means of a goodness-of-fit index.

Keywords: appetite, fuzzy logistic regression, least squares method, linguistic variable

1. Introduction

Many parents complain of their young children's poor appetite and some pediatricians have used folic acid as an appetite-enhancing drug for low-weight children with poor (Hatamizadeh et al. (2010)). It is realistic and suitable to evaluate the appetite status by linguistic terms, such as: very little, little, average, much and very much. Traditionally, symbolic numbers are used to represent qualitative terms for a linguistic variable, e.g. number 5 for "very much", 4 for "much", 2 for "little", and 1 for "very little". Such an oversimplification of data could leave out important information for modeling (Chang et al. (2001)). Moreover, the borderlines of

categories of linguistic variables are not crisp and have a vague status. These variables are inherently measured by fuzzy scale. It seems that fuzzy regression models can be suitable for modeling such an imprecise data (Pourahmad et al. (2011a, b)).

Logistic regression analysis is one of the famous non-linear methods which is used to model a binary response variable based on ordinary explanatory variables (Agresti (2002)). Like other statistical models, it depends heavily on its assumptions, such as, distribution assumptions (Bernoulli probability distribution for the response variable), adequate sample size, and exact observations. These assumptions impose some limitations in practice (Pourahmad et al. (2011b)). In this study, for detecting the relationship between folic acid and child appetite status fuzzy logistic regression is implemented for modeling.

2. Materials and Methods

The study sample was consisted of 26 girls, 3-4 years old who were selected in Tehran, during the year 2011. The study was approved by the Medical Ethics Committee of National Nutrition and Food Technology Research Institute. Body mass index (BMI) was calculated as weight in kilograms divided by the square of height in meters (kg/m^2). The serum concentration of folate was measured. Child's appetite was assessed through asking their mothers the following question: "How do you describe the amount of food that normally has been eaten by your child in the last few days: "very little", "little", "average", "much" or "very much".

3. Data analysis

3.1 Method of fuzzy logistic regression

In traditional statistics, in order to regress a binary response variable with two categories on a set of explanatory variables $X = (x_1, x_2, ..., x_p)$, a binary logistic regression model can be used. Since, in binary logistic regression the response variable $y_i = 0, 1$ has binomial distribution, with $E(Y_i) = P(Y_i = 1) = \pi_i$; $0 < \pi_i < 1$; i=1,2, ..., n; therefore a function of mean response named "logit" $(\ln(\pi/1-\pi))$ is considered for modeling a linear combination of explanatory variables (Agresti (2002)). The form of logistic regression model is as follows

$$\ln(\frac{\pi}{1-\pi}) = b_0 + b_1 x_1 + ... + b_p x_p \qquad (1)$$

in which the expression $(\pi/1-\pi)$ is called the probabilistic odds of characteristic 1.

In some situations the response variable is measured by linguistic terms such as (very low, low, average, high, very high). These linguistic terms detects the status of each case relative to the binary response categories (having the mentioned characteristics or not). In these cases, the binary response cannot be defined precisely and therefore, Bernoulli probability cannot be assumed. So, the probability of success (fully having the mentioned characteristic) cannot be calculated. One strategy to this problem, which was initially proposed by Taheri and Mirzaei Yeganeh (2009) and pourahmad et al. (2011b), is to use possibility of success instead of probability of success. Degree of possibility, measure the consistency degree of each case to the accepted criteria of category 1 of the response variable. So, when the response variable measure a subjective quality by linguistic terms, the possibility of success for each observation can be numerized by defining a proper fuzzy number for each terms of the linguistic variable (μ_i). These fuzzy numbers should be defined in such a way that their support cover the whole range of (0, 1). A brief description on fuzzy numbers can be found in Appendix. To establish the fuzzy logistic regression model, we need to model the logarithmic transformation of possibilistic odds $\ln(\mu_i/1-\mu_i)$, which is linearly dependent to explanatory variables. So, the proposed model is as follows

$$\tilde{W}_i = (\ln\frac{\mu_i}{1-\mu_i}) = \tilde{a}_0 + \tilde{a}_1 x_{i1} + ... + \tilde{a}_p x_{ip} \quad \text{i=1, 2, ..., n,} \qquad (2)$$

in which $\tilde{a}_0, \tilde{a}_1, ..., \tilde{a}_p$ are fuzzy coefficients. After defining a proper fuzzy number for each term of linguistic variable, by the use of extension principle, we have: $\tilde{w}_i(y) = \sup_{\forall x:\ \ln\frac{x}{1-x}=y} \mu_i(x),\ 0<x<1$. Since $\ln(x/1-x)$ is a one to one function, so there is one and only one $x \in (0,1)$, such that $\ln(x/1-x) = y$. Therefore, the membership function of $\ln(\mu_i/1-\mu_i)$ can be computed as follows

$$\tilde{w}_i = (y = \ln\frac{\mu_i}{1-\mu_i}) = \mu_i(\frac{\exp(x)}{1+\exp(x)}). \qquad (3)$$

In the current research, child's appetite is a subjective quality which is measured by linguistic terms. The triangular fuzzy numbers that are defined for terms of linguistic variable $\mu_i = (\text{Very little, Little, Average, Much, Very much})$, are shown in Fig. 1.

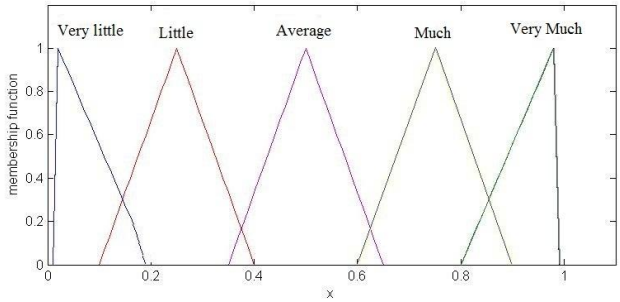

Fig.1: Membership functions of the linguistic terms for child's appetite

$$\text{Very little(x)}=\begin{cases} 1-\frac{0.02-x}{0.01} & 0.01 \le x \le 0.02 \\ 1-\frac{x-0.02}{0.18} & 0.02 < x \le 0.18 \end{cases} \qquad \text{Much(x)}=\begin{cases} 1-\frac{0.75-x}{0.15} & 0.60 \le x \le 0.75 \\ 1-\frac{x-0.75}{0.15} & 0.75 < x \le 0.9 \end{cases}$$

$$\text{little(x)}=\begin{cases} 1-\frac{0.25-x}{0.15} & 0.10 \le x \le 0.25 \\ 1-\frac{x-0.25}{0.15} & 0.25 < x \le 0.40 \end{cases} \qquad \text{Very much(x)}=\begin{cases} 1-\frac{0.98-x}{0.18} & 0.80 \le x \le 0.98 \\ 1-\frac{x-0.98}{0.01} & 0.98 < x \le 0.99. \end{cases}$$

$$\text{Average(x)}=\begin{cases} 1-\frac{0.50-x}{0.15} & 0.35 \le x \le 0.50 \\ 1-\frac{x-0.50}{0.15} & 0.50 < x \le 0.65 \end{cases}$$

Consider $\mu_i = "low"$, to derive the membership function of the possiblistic odds for this observation by Eq. (3) we have,

$$\tilde{w}_i = \begin{cases} 1-\frac{0.25-\frac{\exp(x)}{1+\exp(x)}}{0.15} & 0.1 \le \frac{\exp(x)}{1+\exp(x)} \le 0.25 \Rightarrow -2.197 \le x \le -1.099 \\ 1-\frac{\frac{\exp(x)}{1+\exp(x)}-0.25}{0.15} & 0.25 < \frac{\exp(x)}{1+\exp(x)} \le 0.40 \Rightarrow -1.099 < x \le -0.405. \end{cases}$$

Without loss of generality and for simplicity, we assume that the coefficients in Eq. (2) are symmetric triangular fuzzy numbers $(\tilde{a}_j = (a_j, s_j)_T$ j=0, 1, ..., p). For estimating the regression coefficients in Eq. (2), a fuzzy least squares method will be used. In this method, the sum of squared errors (SSE), which is the sum of distances between fuzzy observations $(\tilde{w}_i)$ and their estimations $(\tilde{W}_i)$, is minimized. The distance function that is defined by Xu and Li (2001) is used here. The details can be found in Appendix (Def. 3). The sum of square errors between observed and estimated outputs based on the distance function, (Def. 3), would be as follows

$$SSE = \sum_{i=1}^{n} (d(\tilde{w}_i, \tilde{W}_i))^2 = \sum_{i=1}^{n} ([\int_0^1 f(\alpha) d^2((\tilde{w}_i)_\alpha, (\tilde{W}_i)_\alpha) d\alpha]^{\frac{1}{2}})^2 \quad (4)$$

Since we have supposed that the regression coefficients are symmetric triangular fuzzy numbers, the estimated outputs can be shown as $\tilde{W}_i = (f_i(a), f_i(s))_T$, in which $f_i(a) = a_0 + a_1 x_{i1} + ... + a_p x_{ip}$, $f_i(s) = s_0 + s_1 x_{i1} + ... + s_p x_{ip}$. In order to obtain the SSE, the $\alpha - cuts$ for the estimated logarithm of odds would be as $(\tilde{W}_i)_\alpha = [(\alpha - 1) f_i(s) + f_i(a),\ (1-\alpha) f_i(s) + f_i(a)]$. If $(\tilde{\mu}_i)_\alpha = [b_{i1}, b_{i2}]$, then the $\alpha - cuts$ of the logarithmic transformation of the observed odds would be $(\tilde{w}_i)_\alpha = \left[\ln \frac{b_{i1}}{1-b_{i1}}, \ln \frac{b_{i2}}{1-b_{i2}} \right]$. By substituting $(\tilde{W}_i)_\alpha$ and $(\tilde{w}_i)_\alpha$ in distance function and computing the SSE according to Eq. (4), it leads to the following structure for SSE

$$SSE = \sum_{i=1}^{n} \left(\int_0^1 f(\alpha). \left[\left[\ln \frac{b_{i1}}{1-b_{i1}} - (\alpha - 1) f_i(s) - f_i(a) \right]^2 + \left[\ln \frac{b_{i2}}{1-b_{i2}} - (1-\alpha) f_i(s) - f_i(a) \right]^2 \right] d\alpha \right).$$

For deriving the model parameters, the partial derivatives of SSE with respect to a_j and s_j must be set to 0. After some algebraic computations the following equations are derived

$$\begin{aligned} &a_0 \sum_{i=1}^{n} x_{i0} x_{ij} + a_1 \sum_{i=1}^{n} x_{i1} x_{ij} + ... + a_p \sum_{i=1}^{n} x_{ip} x_{ij} = \sum_{i=1}^{n} z_i x_{ij}, \quad \text{j=0, 1, ..., p,} \\ &s_0 \sum_{i=1}^{n} x_{i0} x_{ij} + s_1 \sum_{i=1}^{n} x_{i1} x_{ij} + ... + s_p \sum_{i=1}^{n} x_{ip} x_{ij} = \sum_{i=1}^{n} k_i x_{ij}, \quad \text{j=0, 1, ..., p,} \end{aligned} \quad (5)$$

where

$$z_i = \int_0^1 \left(\alpha . \ln \frac{b_{i1}}{1-b_{i1}} + \alpha . \ln \frac{b_{i2}}{1-b_{i2}} \right) d\alpha,\ k_i = 6 \int_0^1 \left(\alpha(1-\alpha) . \ln \frac{b_{i2}}{1-b_{i2}} - \alpha(1-\alpha) . \ln \frac{b_{i1}}{1-b_{i1}} \right) d\alpha \text{ and } x_{i0} = 1.$$

The expressions in Eq. (5) can be written as the following matrix form,

$$(X'X)a = Z \quad (X'X)s = K \quad (6)$$

In which,

$$X = \begin{pmatrix} 1 & x_{11} & ... & x_{1p} \\ 1 & x_{21} & ... & x_{2p} \\ ... & ... & ... & ... \\ 1 & x_{n1} & ... & x_{np} \end{pmatrix}_{n \times (p+1)}$$

$$a = (a_0, a_1, ..., a_p)^T,\ \mathrm{Z} = (\sum_{i=1}^{n} z_i x_{i0}, \sum_{i=1}^{n} z_i x_{i1}, ..., \sum_{i=1}^{n} z_i x_{ip})^T$$

$$s = (s_0, s_1, ..., s_p)^T,\ \mathrm{K} = (\sum_{i=1}^{n} k_i x_{i0}, \sum_{i=1}^{n} k_i x_{i1}, ..., \sum_{i=1}^{n} k_i x_{ip})^T.$$

If matrix $X'X$ is positive definite, the above system has a unique solution which can be computed by the following expressions: $a = (X'X)^{-1} Z$, $\mathrm{s} = (X'X)^{-1} K$ (Xu and Li (2001)).

3.2 Evaluation of model: Capability Index

In fuzzy logistic regression, the Mean of Capability Index between the observed $(\tilde{w}_i)$ and the estimated values $(\tilde{W}_i)$ can be defined as $MCI = \frac{1}{n} \sum_{i=1}^{n} I(\tilde{w}_i, \tilde{W}_i)$, where $I(\tilde{w}_i, \tilde{W}_i) = Card(\tilde{w}_i \cap \tilde{W}_i) / Card(\tilde{w}_i \cup \tilde{W}_i)$ and $Card(\tilde{w}_i) = \int_t \tilde{w}_i(t) dt$. The "min" operator for the intersection and "max" operator for the union can be used. The MCI's maximum value is 1 and its minimum value is 0, a larger MCI confirms that the model supports the data well.

4. Results

Our sample included 26 girls whose age range from 3 to 4 years. A part of participants' data is shown in Table 1.

Table 1: Characteristics of the sample data

No.	Age	BMI	Folic acid	Birth order	Number of family members	Mother's job[§]	Mother's education*	Father's education*	Appetite status
1	37	13.87	2	1	3	0	1	1	Little
2	37	14.33	3.1	2	4	1	0	0	Average
3	39	13.89	3.5	1	3	1	0	1	Average
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
26	48	15.00	4.7	1	3	0	0	1	Much

* Having academic degree

[§]House keeper

In order to evaluate the relationship between the odds of having a perfect appetite and the level of folic acid, we have decided to adjust the effect of some covariates that are supposed to have a significant impact on child's appetite by the means of regression modeling. These covariates are about child's age, body mass index (BMI), child's birth order, the number of family members who live with the child, an indicator variable for mother's job (house keeper=1 or employed=0) and an indicator variable for parents academic education. The proposed model is as follows

$$\tilde{w} = (\ln \frac{\mu}{1-\mu}) = \tilde{a}_0 + \tilde{a}_1 Age + \tilde{a}_2 BMI + \tilde{a}_3 Folic\ \text{Acid} + \tilde{a}_4 Birth\ \text{Order} + \tilde{a}_5 \text{Number of family members} + \tilde{a}_6 House\ \text{keeper} + \tilde{a}_7 Mother's\ \text{education} + \tilde{a}_8 Father's\ \text{education}. \quad (7)$$

By using the method that was described in section 3, the following solution is given

$$a = A^{-1}Z = (-4.8656\ \ 0.0326\ \ 0.2458\ \ 0.1549\ -1.1515\ \ 0.6064\ \ 0.2091\ -0.3333\ -0.2902)^T$$

$$s = A^{-1}K = (\ -0.0094\ \ 0.0032\ \ 0.0395\ -0.0057\ -0.1118\ \ 0.0579\ -0.0423\ \ 0.0184\ \ 0.0204)^T$$

It is possible to encounter conditions that $A^{-1}K < 0$, therefore the spreads of the fuzzy parameters will be negative. For this problem, we have decided to follow the procedure that is suggested in (Mohammadi and Taheri (2004)). By the means of that procedure, the spreads are estimated again and the results are obtained as $s = A^{*-1}K^* = (\ 0\ \ 0.0034\ \ 0.0375\ 0\ 0\ 0\ 0\ \ 0.0014\ \ 0.0707)^T$. The optimal model is obtained as

$$\tilde{W} = (\ln \frac{\mu}{1-\mu}) = -4.85 + (0.33, 0.003)_T\ Age + (0.25, 0.04)_T\ BMI + 0.16 Folic\ \text{Acid} - 1.15 Birth\ \text{Order} + 0.61 \text{Number of family members} + 0.21 House\ \text{keeper} + (-0.33, 0.001)_T\ Mother's\ \text{education} + (-0.29, 0.07)_T\ Father's\ \text{education}.$$

In summary, the estimated positive coefficient indicates that the corresponding variable is related to the increase in possibilistic odds of having a better appetite so, we can understand that having a higher level of folic acid is related to the increase in possibilistic odds of having a better appetite. To evaluate the model based on the goodness-of-fit index we get $MCI = 0.50$, which reveals fairly good fitting.

5. Discussion

In this paper, the detail of fuzzy logistic regression which was proposed in (Pourahamd et al. (2011b)) was investigated and its application in a real clinical situation in the field of nutrition is discussed. Fuzzy least squares approach was used for estimating the fuzzy coefficients of the model and for evaluating the model's goodness-of-fit, the Mean Capability Index was used.

Few studies are done for evaluating the relationship between folate and appetite level. According to a study in Iran (Hatamizadeh et al. (2007)) and another study on girls in India (Kanani and Poojara (2000)), iron and folic acid supplements were resulted in weight gain and an increase in perceived level of hunger. Since many

vague observations in clinical studies are measured by linguistic terms, the proposed model can be used in other research areas with similar situations.

Appendix

A fuzzy set of the universal set X is defined as a set of ordered pairs: $\tilde{A}=\{(x,\tilde{A}(x)\,|\,x\in X\}$, where, $\tilde{A}(.)$ is called the membership function of A, and $\tilde{A}(x)$ is the grade of membership of x in the fuzzy set A.

Definition 1: A fuzzy set $\tilde{u}$ of $\Re$ (the real line) is called a triangular fuzzy number and is denoted by $(m,a,\beta)_T$, if its membership function is as follows

$$\tilde{u}(x)=\begin{cases}1-\dfrac{m-x}{a}, & m-a\le x\le m,\\ 1-\dfrac{x-m}{\beta}, & m<x\le m+\beta.\end{cases}$$

If $a=\beta$, then $\tilde{u}$ is denoted by $(m,a)_T$ and is called a symmetric triangular fuzzy number.

Definition 2: The (crisp) set of elements that belong to the fuzzy set $\tilde{A}$ at least to the degree α is called the α-cuts or α-level set: $A_\alpha=\{x\in X\,|\,\mu_{\tilde{A}}(x)\ge\alpha\}$, $0<\alpha\le 1$.

Definition 3: Let E denote a fuzzy number space, such that $u, v\in E$ are fuzzy numbers. The distance between u and v based on a function $f(\alpha)$ is $d(u,v)=[\int_0^1 f(\alpha)d^2((u)_\alpha,(v)_\alpha)d\alpha]^{\frac{1}{2}}$ in which $d^2((u)_\alpha,(v)_\alpha)=[a_1(\alpha)-b_1(\alpha)]^2+[a_2(\alpha)-b_2(\alpha)]^2$. $(u)_\alpha=[a_1(\alpha),\ a_2(\alpha)]$, $(v)_\alpha=[b_1(\alpha),\ b_2(\alpha)]$ are $\alpha-cuts$ of u and v respectively. $f(\alpha)$ is an increasing weighting function on $[0,1]$, where $f(0)=0$ and $\int_0^1 f(\alpha)d\alpha=0.5$ (Xu and Li (2001)). For more details on fuzzy arithmetic, see Viertl (1996).

References

Agresti, A. (2002) *Categorical Data Analysis*, Wiley, New York,.

Chang ,Y.O., and Ayyub, B.M. (2001) "Fuzzy regression methods - a comparative assessment," *Fuzzy Sets and Systems*, 119, 187-203.

Hatamizadeh, N., Eftekhar, H., Shafaghi, B., and Mohammad, K. (2007) "Effects of folic acid on preschool children's appetite: Randomized triple-blind clinical trial", *Pediatrics International*, 21, 49, 558-63.

Kanani, S.J., and Poojara, R.H. (2001) "Supplementation with iron and folic acid enhances growth in adolescent Indian girls," *Journal of Nutrition*, 130, S452- 55.

Mohammadi, J., and Taheri, S.M. (2004) "Pedomodels fitting with fuzzy least squares regression," *Iranian Journal of Fuzzy Systems*, 1, 45–61.

Pourahmad, S., Ayatollahi, S.M.T., and Taheri, S.M. (2011a) "Fuzzy logistic regression: A new possiblistic model and its application in clinical vague states", *Iranian Journal of Fuzzy Systems*, 8, 1-17.

Pourahmad, S., Ayatollahi, S.M.T., Taheri, S.M. and Agahi, Z.H. (2011b) "Fuzzy logistic regression based on the least squares approach with application in clinical studies," *Computers and Mathematics with Applications*, 62, 3353–3365.

Taheri, S.M., Mirzaei Yeganeh, S. (2009) "Logistic regression with non-precise data," in: Proc. of the 57th ISI (International Statistical Institute) Congress, South Africa, Durban, 1–9, 2009.

Viertl, R. (1996) *Statistical methods for non-precise data*, CRC Press, Boca Raton, Florida.

Xu, R., and Li, C. (2001) "Multidimensional least-squares fitting with fuzzy model," *Fuzzy Sets and Systems*, 119, 215–223.