International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:2, No:10, 2008

Fuzzy Logic Approach to Robust Regression Models of Uncertain Medical Categories

Arkady Bolotin

***Abstract*—**Dichotomization of the outcome by a single cut-off point is an important part of various medical studies. Usually the relationship between the resulted dichotomized dependent variable and explanatory variables is analyzed with linear regression, probit regression or logistic regression. However, in many real-life situations, a certain cut-off point dividing the outcome into two groups is unknown and can be specified only approximately, i.e. surrounded by some (small) uncertainty. It means that in order to have any practical meaning the regression model must be robust to this uncertainty. In this paper, we show that neither the beta in the linear regression model, nor its significance level is robust to the small variations in the dichotomization cut-off point.

As an alternative robust approach to the problem of uncertain medical categories, we propose to use the linear regression model with the fuzzy membership function as a dependent variable. This fuzzy membership function denotes to what degree the value of the underlying (continuous) outcome falls below or above the dichotomization cut-off point. In the paper, we demonstrate that the linear regression model of the fuzzy dependent variable can be insensitive against the uncertainty in the cut-off point location.

In the paper we present the modeling results from the real study of low hemoglobin levels in infants. We systematically test the robustness of the binomial regression model and the linear regression model with the fuzzy dependent variable by changing the boundary for the category ***Anemia*** and show that the behavior of the latter model persists over a quite wide interval.

***Keywords*—**Categorization, Uncertain medical categories, Binomial regression model, Fuzzy dependent variable, Robustness.

I. Introduction

CHANGING a dependent variable from continuous to categorical form is a common part of many medical data analyses, since categorization makes it easier for clinicians to use information about the relationship between a dependent variable and explanatory variables in medical decision-making models [9,11].

Often a single cut-off point that divides a dependent variable into just two categories is sought. The resulted binary dependent variable may be used for making treatment recommendations or determining study eligibility.

Among the most popular techniques, employed to analyze the relationship of a dichotomized dependent variable with explanatory variables, are linear regression, probit regression and logistic regression models.

A. Bolotin is with the Ben-Gurion University of the Negev, Beersheba 84105, Israel; (e-mail: arkadyv@bgumail.bgu.ac.il).

These techniques are well known and their procedures are well established. Troubles start when categories are becoming uncertain. Many clinical categories such as ***high***, ***low***, and the like, are linguistic ones, and therefore they do not suggest certain cut-off points, which divide the observations into two groups. Besides, in many real-life situations the boundary between categories can be identified only approximately, i.e. in some intervals. For that reason, when dichotomizing a continuous dependent variable, one can select any cut-off point within these intervals.

It follows then, that the linear regression model of such "approximately" dichotomized dependent variable (as well as the probit or logit model) must be robust over these intervals, that is, insensitive to uncertainties in the category boundary.

Whether the linear regression models are robust to small variations in the dichotomization cut-off point is needed to test theoretically.

The principle alternative to the linear regression model of a dichotomous dependent variable on explanatory variables would be the linear regression model of a fuzzy dependent variable on the same explanatory variables.

Indeed, a dichotomized dependent variable may be considered as some kind of crisp membership function denoting whether it is true that the observed value of the dependent continuous variable falls into the given category. In the same way, one can introduce into analysis a fuzzy dependent variable, which will be a fuzzy membership function denoting to what degree the observed value of the dependent continuous variable falls into the given category. Using this fuzzy membership function as a dependent variable in a linear regression model is another way to deal with uncertain categories in medical data analysis [1-3].

Whether this model is robust to small variations in the parameters of the fuzzy membership function is needed to test theoretically as well.

We test the robustness of the linear regression model to small variations in the dichotomization cut-off point in this paper. We also test the robustness of the linear regression model of a fuzzy dependent variable on explanatory variables to small variations in the parameters of the fuzzy membership function. To make theoretical conclusions more tangible we consider a practical example of the real study of iron deficiency anemia in infants.

II. Testing Robustness of Linear Regression Models

Let us consider a dataset containing N observations of the continuous dependent variable Y and, for the sake of simplicity, the only independent variable X. Let y_j denote the value of the variable Y for an observation j (j = 1,…, N), and let x_j be the observed value of the independent variable X for the same observation.

Suppose we have a category A, which corresponds to some values of the variable Y. We may think of the category A as a subset A of the set Y. We want to know the truth or falsity of the statement

$$y_j \in A \tag{1}$$

and how the value x_j of the independent variable X can be associated with it.

A. Linear Probability Model

Assume that the category A is defined as a mapping between elements of the set Y and elements of the set {0, 1},

$$A:\ y_j \ \Rightarrow\ \{0,1\} \quad , \tag{2}$$

where the value zero represents non-membership, and the value one represents membership. This mapping can be described as a binary function, for example, as the function

$$\delta_A(y_j,c) = \begin{cases} 1, & y_j \ge c \\ 0, & y_j < c \end{cases} \quad , \tag{3}$$

where c is some cut-off point, or hurdle value. Thus, the statement (1) is true if the function $\delta_A(y_j,c)$ is equal to 1, and the statement (1) is false if $\delta_A(y_j,c)$ is 0.

Let us convert the continuous dependent variable Y to the binary dependent variable $\delta_A(Y,c)$ by the use of the formula (3).

Let β be the regression coefficient on the variable X, and let α be the intercept. The linear regression of the binary dependent variable $\delta_A(Y,c)$ on the independent variable X (*the linear probability model*) takes the form:

$$\delta_A(y_j,c) = \alpha + \beta x_j \quad . \tag{4}$$

The regression coefficient β shows how a change in the independent variable X affects ***the probability of truth*** of the statement (1).

The values $\hat{\alpha}$ and $\hat{\beta}$ that minimize the sum of squares

$$L(\alpha,\beta) = \sum_{j=1}^{N} \left[\delta_A(y_j,c) - \alpha - \beta x_j\right]^2 \tag{5}$$

are defined to be the least-squares estimators of α and β. Substituting the definition (3) for $\delta_A(y_j,c)$ one finds

$$\hat{\beta} = \frac{\sum_{y_j \ge c} x_j - N_{Y\ge c}\cdot \bar{x}}{\sum_{j=1}^{N}(x_j-\bar{x})^2}, \qquad \hat{\alpha} = \frac{N_{Y\ge c}}{N} - \hat{\beta}\bar{x}\,, \tag{6}$$

where $N_{Y\ge c}$ is the number of observations falling in the category A (i.e. above the cut-off point c).

Let us suppose that the unknown parameters α and β of the model (4) have been estimated by $\hat{\alpha}$ and $\hat{\beta}$.

Imagine now that the cut-off point c of the binary dependent variable $\delta_A(Y,c)$ is changing from c to $c+\Delta c$, where $\Delta c > 0$ is a small increment. This changes the model (4)

$$\delta_A(y_j, c+\Delta c) = \alpha' + \beta' x_j \tag{7}$$

and hence the estimators $\hat{\alpha}$ and $\hat{\beta}$

$$\hat{\beta}' = \hat{\beta} + \Delta\hat{\beta}, \qquad \hat{\alpha}' = \hat{\alpha} + \Delta\hat{\alpha}\,, \tag{8}$$

where

$$\Delta\hat{\beta} = -\frac{\sum_{y_j\in(c,c+\Delta c)} x_j - N_{(c,c+\Delta c)}\cdot\bar{x}}{\sum_{j=1}^{N}(x_j-\bar{x})^2} \quad , \tag{9}$$

$$\Delta\hat{\alpha} = -\frac{N_{(c,c+\Delta c)}}{N} + \Delta\hat{\beta}\cdot\bar{x} \quad , \tag{10}$$

and $N_{(c,c+\Delta c)}$ is the number of observations which fall into the interval $(c, c+\Delta c)$. Noting that

$$N_{(c,c+\Delta c)} = N\cdot p(y_m)\cdot\Delta c \quad , \tag{11}$$

$$\sum_{y_j\in(c,c+\Delta c)} x_j \approx N\cdot p(y_m)\cdot x_m\cdot\Delta c \quad , \tag{12}$$

where $p(y_m)$ is the density function of the probability distribution of Y values at some interior point y_m of the interval $(c, c+\Delta c)$, and also setting

$$b = -\frac{(x_m-\bar{x})}{\sum_{j=1}^{N}(x_j-\bar{x})^2}, \qquad a = -\frac{1}{N} - b\bar{x}\,, \tag{13}$$

we can write (8) as

$$\hat{\beta}' = \hat{\beta} + bN\cdot p(y_m)\cdot\Delta c \quad , \tag{14}$$

$$\hat{\alpha}' = \hat{\alpha} + aN\cdot p(y_m)\cdot\Delta c \quad . \tag{15}$$

It is readily seen, that the increments $\Delta\hat{\beta}$ and $\Delta\hat{\alpha}$ in the model estimators $\hat{\beta}$ and $\hat{\alpha}$ due to the increment Δc are of ***the same order*** as Δc. It means that strictly speaking, the linear probability model (4) of the binary dependent variable $\delta_A(Y,c)$ cannot be robust to the small variations in the model cut-off point c.

Nevertheless, let us consider whether the model (4) can be relatively robust, i.e. robust in terms of persistence of significance level of the model estimators $\hat{\alpha}$ and $\hat{\beta}$ over the small variations of the model cut-off point c.

Using the estimator of the variance σ^2 based on the least-squares estimators $\hat{\alpha}$ and $\hat{\beta}$

$$\hat{\sigma}^2 = (N-2)^{-1}\cdot[N_{Y\ge c} - 2\sum_{y_j\ge c}(\hat{\alpha}+\hat{\beta}x_j) + \sum_{j=1}^{N}(\hat{\alpha}+\hat{\beta}x_j)^2] \quad , \tag{16}$$

we will get the variances for $\hat{\alpha}$ and $\hat{\beta}$

$$\sigma_{\hat{\alpha}}^{\,2} = \frac{\hat{\sigma}^2\cdot\sum_j^N x_j^{\,2}}{N\sum_j^N(x_j-\bar{x})^2}, \qquad \sigma_{\hat{\beta}}^{\,2} = \frac{\hat{\sigma}^2}{\sum_j^N(x_j-\bar{x})^2}, \tag{17}$$

and the t statistics

$$t_{\hat{\alpha}} = \frac{\hat{\alpha}}{\sigma_{\hat{\alpha}}}, \qquad t_{\hat{\beta}} = \frac{\hat{\beta}}{\sigma_{\hat{\beta}}}\,. \tag{18}$$

Giving the cut-off point c a small increment $\Delta c > 0$, we will eventually find

$$\Delta t_{\hat{\alpha}} = \left[-\frac{S \cdot t_{\hat{\alpha}} \sum_j^N x_j^{\,2}}{2\sigma_{\hat{\alpha}}(N-2)\sum_j^N (x_j - \bar{x})^2} + aN \right] \frac{p(y_m)\Delta c}{\sigma_{\hat{\alpha}}} + \\ + O(\Delta c) \quad , \tag{19}$$

$$\Delta t_{\hat{\beta}} = \left[-\frac{N \cdot S \cdot t_{\hat{\beta}}}{2\sigma_{\hat{\beta}}(N-2)\sum_j^N (x_j - \bar{x})^2} + bN \right] \frac{p(y_m)\Delta c}{\sigma_{\hat{\beta}}} + \\ + O(\Delta c) \quad , \tag{20}$$

where

$$S = -1 + 2(\hat{\alpha} + \hat{\beta} x_m) - \\ - 2\sum_{y_j \geq c} (a + b x_j) + \\ + 2\sum_{j=1}^N (\hat{\alpha} + \hat{\beta} x_j)(a + b x_j) \tag{21}$$

and $O(\Delta c)$ denotes the terms of higher order than Δc.

Suppose that before the change in the cut-off point c, the coefficient $\hat{\beta}$ was significant, namely $\left| t_{\hat{\beta}} \right| = t_\varepsilon$, where ε is the chosen significance level. After the change, the t statistic for the coefficient $\hat{\beta}$ turns into $t'_{\hat{\beta}} = t_{\hat{\beta}} \pm \Delta t_{\hat{\beta}}$, where

$$\Delta t_{\hat{\beta}} \sim \Delta c \quad , \tag{22}$$

and we get the relation $\left| t'_{\hat{\beta}} \right| = t_\varepsilon \pm \Delta t_{\hat{\beta}}$, which means that the regression coefficient $\hat{\beta}$ may become insignificant.

Hence, neither the regression coefficient $\hat{\beta}$, nor its significance level is robust to the small variations in the model cut-off point c.

B. Linear Regression of Fuzzy Dependent Variable

Assume that the category A is defined as a mapping between elements of Y and values of the interval [0, 1],

$$A: \; y_j \;\Rightarrow\; [0, 1] \quad , \tag{23}$$

where the value 0 represents complete non-membership, the value 1 represents complete membership, and the values in between represent transitional degrees of membership. This mapping can be described as a fuzzy membership function, for example as the function

$$\mu_A(y_j, c_1, c_2) = \begin{cases} 1 & , \quad y_j \geq c_2 \\ \dfrac{y_j - c_1}{c_2 - c_1} & , \quad c_1 < y_j < c_2 \\ 0 & , \quad y_j \leq c_1 \end{cases} \quad , \tag{24}$$

where c_1 and c_2 are some edge points ($c_2 > c_1$), which isolate the transitional area where the function $\mu_A(y_j, c_1, c_2)$ varies between 0 and 1. Thus, the degree to which the statement (1) is true is determined by the value of the function $\mu_A(y_j, c_1, c_2)$.

Let us convert the continuous dependent variable Y to the fuzzy dependent variable $\mu_A(Y, c_1, c_2)$ by means of the formula (24).

The linear regression model of the fuzzy dependent variable $\mu_A(Y, c_1, c_2)$ on the independent variable X takes the form:

$$\mu_A(y_j, c_1, c_2) = \alpha + \beta x_j \quad . \tag{25}$$

The regression coefficient β on the variable X is assumed to be interpreted as a change in the degree of membership in the category A for a given change in X. In other words, the coefficient β determines how a change in the variable X affects ***the degree of truth*** of the statement (1).

Let us suppose that the unknown parameters α and β of the model (25) have been estimated by the least-squares estimators $\hat{\alpha}$ and $\hat{\beta}$

$$\hat{\beta} = \frac{\sum_{j=1}^N \mu_A(y_j, c_1, c_2) x_j - \bar{x} \sum_{j=1}^N \mu_A(y_j, c_1, c_2)}{\sum_{j=1}^N (x_j - \bar{x})^2} \quad , \tag{26}$$

$$\hat{\alpha} = \overline{\mu_A(Y, c_1, c_2)} - \hat{\beta}\bar{x} \quad . \tag{27}$$

Giving the left edging c_1 a small increment Δc_1 and the right edging c_2 a small increment Δc_2, and then assuming that $\Delta c_1 \sim \Delta c_2 \sim \Delta c > 0$ we will find the corresponding increments in the estimators $\hat{\alpha}$ and $\hat{\beta}$

$$\Delta\hat{\beta} = -\frac{\Delta c}{c_2 - c_1} \cdot \frac{\sum_{y_j \in (c_1, c_2)} (x_j - \bar{x})}{\sum_{j=1}^N (x_j - \bar{x})^2} + O(\Delta c) \quad , \tag{28}$$

$$\Delta\hat{\alpha} = -\frac{N_{(c_1, c_2)}}{N} \cdot \frac{\Delta c}{c_2 - c_1} - \Delta\hat{\beta}\bar{x} + O(\Delta c) \quad , \tag{29}$$

where $N_{(c_1, c_2)}$ is the number of observations falling into the transitional area (c_1, c_2). If the transitional area (c_1, c_2) is large enough to hold the relation $N_{(c_1, c_2)} \sim N$, the 1st sample moment about the mean $\bar{x}$ will be very close to zero

$$\sum_{y_j \in (c_1, c_2)} (x_j - \bar{x})^1 \approx \sum_{j=1}^N (x_j - \bar{x}) = 0 \quad , \tag{30}$$

and so, instead of (28) and (29), we will get

$$\Delta\hat{\beta} \cong 0 + O(\Delta c), \qquad \Delta\hat{\alpha} \cong -\frac{\Delta c}{c_2 - c_1} + O(\Delta c) . \tag{31}$$

Computing $\Delta\hat{\sigma}^2$ and substituting its value in the t statistics for estimators $\hat{\alpha}$ and $\hat{\beta}$ we will get

$$t'_{\hat{\beta}} \cong t_{\hat{\beta}} + O(\Delta c) \quad , \tag{32}$$

$$t'_{\hat{\alpha}} \cong t_{\hat{\alpha}} - \frac{1}{\sigma_{\hat{\alpha}}} \cdot \frac{\Delta c}{c_2 - c_1} + O(\Delta c) \quad . \tag{33}$$

It follows, then, that the regression coefficient $\hat{\beta}$ along with its significance level is insensitive to the increment Δc. Thus, unlike the linear regression model (4) of the binary dependent variable $\delta_A(Y, c)$, the linear regression model of the fuzzy dependent variable $\mu_A(Y, c_1, c_2)$ may be robust to the small variations in the model edge points c_1 and c_2.

III. Analysis of Anemia in Children

To make these theoretical conclusions more tangible let us consider the following practical example. In the example, we will use the data obtained from the real study of iron deficiency anemia in 618 infants (children of age 0 to 12 months) living in the Negev desert area, Israel. (The study was done by Prof. D. Fraser of Ben-Gurion University of the Negev and her colleagues in 2002-2005 years [7].)

The study dataset comprises questionnaire information and laboratory findings. The dependent continuous variable of the

dataset is an infant's hemoglobin level (Hgb, measured in g/dl). The summary statistics calculated for the Hgb variable are shown in the Table I.

TABLE I
THE UNIVARIATE SUMMARY STATISTICS FOR THE HGB VARIABLE

Parameters	Values
Nonmissing observations	200
Mean	10.88
Std. Dev.	0.96
Variance	0.92
Skewness	-0.04
Kurtosis	3.78
1% percentile	8.20
50% percentile	10.85
99% percentile	13.40

The dataset independent (explanatory) variables of our interest along with their summary statistics are presented in the Table 2.

TABLE II
THE INDEPENDENT VARIABLES AND THEIR SUMMARY STATISTICS

#	Variable	Non-missing obser-vations	Mean ± S.D.	Min	Max
1	Infant's age (*months*)	611	4.96 ± 1.33	1.60	11.77
2	Ethnicity	618	binary variable: Jews = 212, Bedouins = 406		
3	Breastfeeding (BF) (*times a day*)	252	3.12 ± 3.08	0	9
4	Intake of infant formula *(times a day)*	252	2.81 ± 2.00	0	8
5	Intake of cow's milk (*times a day*)	252	0.11 ± 0.51	0	4
6	Infant's age (*months*) when BF stopped	252	5.05 ± 1.88	0	7

The age-matched normal values for hemoglobin are listed in the Table 3 [12].

TABLE III
AGE-SPECIFIC HEMOGLOBIN VALUES

Infant's age	Hgb normal values	Mean Hgb
< 1 month	10.0÷20.0	13.9
2÷6 months	9.5÷14.0	12.6
6÷24 months	10.5÷13.5	12.0

Iron deficiency anemia is defined as a ***low*** level of hemoglobin [5]. In view of that, our task is to analyze whether the given independent variables can influence low levels of hemoglobin in infants.

A. Linear Probability Model of Pediatric Anemia

According to the CDC Guidelines [12], a level of Hgb less than 11 g/dl is the criteria for anemia in infants and children under five. Therefore, we can start by generating the binary dependent variable $\delta_{Anemia}(Hgb, 11)$

$$\delta_{Anemia}(Hgb_j, 11) = \begin{cases} 1, & Hgb_j < 11 \\ 0, & Hgb_j \geq 11 \end{cases}, \quad (34)$$

which determines that an infant's hemoglobin is ***low*** if the Hgb is less than 11 g per dl. Now we can fit the linear probability model of $\delta_{Anemia}(Hgb, 11)$ on the independent variables X_i (i = 1,…,6) to our dataset

$$\delta_{Anemia}(Hgb_j, 11) = \alpha + \sum_{i=1}^{6} \beta_i x_{ij} \quad . \quad (35)$$

Obviously, the CDC Guidelines are approximate. This means, that the boundary for the category ***Anemia*** is not precisely 11 but rather any value c_0 within the interval $11 - \Delta c$ to $11 + \Delta c$, where Δc is some small uncertainty. It follows then, that the model (35) must be very close (in terms of the regression coefficients β_i and their significance levels) to any of the models

$$\delta_{Anemia}(Hgb_j, 11 \pm \Delta c) = \alpha' + \sum_{i=1}^{6} \beta_i' x_{ij} \quad , \quad (36)$$

We have tested the models (36) by changing the value c_0 within the interval 11±0.4; the results are presented in the Table 4.

TABLE IV
THE BETAS AND T-STATISTICS
OF THE LINEAR PROBABILITY MODELS OF $\delta_{ANEMIA}(HGB, C_0)$

Terms in the equation and R-squared	Boundary c_0 for the category ***Anemia***				
	10.6	10.8	11.0	11.2	11.4
Infant's age (*months*)					
Ethnicity = Bedouins			0.28 *2.86*	0.23 *2.30*	0.26 *2.74*
Breastfeeding (BF) (*times a day*)			-0.04 *-2.71*	-0.04 *-2.39*	
Intake of infant formula (times a day)	-0.05 *-2.59*	-0.06 *-3.08*	-0.06 *-3.17*	-0.05 *-2.69*	-0.05 *-2.63*
Intake of cow's milk (*times a day*)	0.13 *2.64*	0.12 *2.41*			
Infant's age (*months*) when BF stopped					
α			0.56 *2.69*	0.69 *3.32*	0.87 *4.11*
R-squared	0.09	0.09	0.10	0.08	0.09

Key for terms: the upper value is the beta; the lower value is its t statistic.
Only coefficients, which are statistically significant at, at least, the 0.05 level, are posted in the table.

In perfect agreement with the theoretical conclusions of the previous chapter, the models presented in this table have different statistically significant predictors of anemia in infants. For example, we see that the variable *Breastfeeding* is statistically significant in the model of the $\delta_{Anemia}(Hgb, 11.0)$, but it is not in the model of the $\delta_{Anemia}(Hgb, 10.8)$ or in the model of the $\delta_{Anemia}(Hgb, 11.4)$. This means that the analysis of the given dataset based on linear probability modeling may be biased.

B. Regression of the Fuzzy Dependent Variable for Anemia

Therefore, now we will try the approach based on the regression analysis with a fuzzy dependent variable.

First, using the boundary c_0=11g/dl and the edge points

$$c_1 = c_0 - h, \qquad c_2 = c_0 + h \qquad (h = 2.5) \quad , \tag{37}$$

we will construct the fuzzy dependent variable

$$\mu_{Anemia}(Hgb_j, 11, 2.5) = \begin{cases} 0 & , \quad Hgb_j \geq 13.5 \\ (13.5 - Hgb_j)/5 & , \quad 8.5 \leq Hgb_j < 13.5 \\ 1 & , \quad Hgb_j < 8.5 \end{cases} , \tag{38}$$

then we will try to fit the linear regression model of the $\mu_{Anemia}(Hgb_j, 11, 2.5)$ to the dataset

$$\mu_{Anemia}(Hgb_j, 11, 2.5) = \alpha + \sum_{i=1}^{6} \beta_i x_{ij} \quad . \tag{39}$$

Since the boundary c_0=11g/dl for the category ***Anemia*** is only approximate, the model (39) must be very close to any of the models

$$\mu_{Anemia}(Hgb_j, 11 \pm \Delta c, 2.5) = \alpha' + \sum_{i=1}^{6} \beta'_i x_{ij} \quad . \tag{40}$$

As we can see from the Table 1, more than 95% of the observed values of Hgb fall into the transitional area (8.5, 13.5) of the fuzzy membership function (38). Hence, basing on the theoretical conclusions (31) and (32), we can anticipate that all the models (40) will be alike.

By analogy with the previous case, we have tested the models (40) by changing the value c_0 within the same range 11±0.4; the results are presented in the Table 5.

TABLE V
THE BETAS AND T-STATISTICS
OF THE LINEAR REGRESSION MODELS OF $\mu_{ANEMIA}(HGB, C_0, 2.5)$

Terms in the equation and R-squared	Boundary c_0 for the category ***Anemia***				
	10.6	10.8	11.0	11.2	11.4
Infant's age (*months*)					
Ethnicity= Bedouins	0.12 *2.94*	0.12 *2.93*	0.12 *2.92*	0.12 *2.91*	0.12 *2.90*
Breastfeeding (BF) (*times a day*)					
Intake of infant formula (times a day)	-0.03 *-3.41*	-0.03 *-3.41*	-0.03 *-3.41*	-0.03 *-3.40*	-0.03 *-3.40*
Intake of cow's milk (*times a day*)					
Infant's age (*months*) when BF stopped					
α	0.41	0.45	0.45	0.54	0.58
t_α	*4.98*	*5.51*	*5.51*	*6.62*	*7.17*
R-squared	0.15	0.15	0.15	0.14	0.14

Key for terms: the upper value is the beta; the lower value is its t statistic.
Only coefficients, which are statistically significant at, at least, the 0.05 level, are posted in the table.

Indeed, we see that the models presented in this table are quite similar. They all show the same statistically significant predictors of low hemoglobin in children. (Namely, the child's Bedouin ethnicity increases the degree of truth of the statement ***"the child has anemia"***, while the intake of infant formula decreases it.)

Accordingly, we can conclude that, the analysis of the given dataset based on linear regression modeling with fuzzy dependent variables has no bias related to the choice of the boundary for the category ***Anemia.***

IV. CONCLUSION

The linear regression models, which we considered in this paper, are in fact ***latent*** variable models. Indeed, the linear regression model of the binary dependent variable $\delta_A(Y, c)$ has the underlying variable *Y*, which is outside the model structure (i.e. it does not belong to the model variables). Therefore, the cut-off point *c*, which is placed on the *Y*-axis, is actually a ***latent model parameter***.

Binomial regression models have been studied well for robustness of their ***explicit parameters***, not the latent ones. For example, there exist robust parameter estimates that provide a good fit to the bulk of the data (i.e. the observed values of the variables in the regression equation) when the data contain outliers, as well as when the data are free of them [8,11].

That is why one of the goals of our study was to test sensitivity of the binomial regression model with the dependent variable $\delta_A(Y, c)$ to small variations in the latent model parameter – the dichotomization cut-off point *c*.

We found that neither the regression coefficient in this model, nor its significance level is robust to the uncertainty Δc in the model cut-off point $c \pm \Delta c$.

We also tested the linear regression with the simplest fuzzy membership function $\mu_A(Y, c_1, c_2)$ (which is nothing more than a collection of two edge points c_1 and c_2 placed on the *Y*-axis) as a model dependent variable. This straight line membership function $\mu_A(Y, c_1, c_2)$ has the advantage of being the simplest generalization of the step-like function $\delta_A(Y, c)$.

We found that unlike the binomial regression model of the $\delta_A(Y, c)$, the linear regression model of the fuzzy dependent variable $\mu_A(Y, c_1, c_2)$ can be insensitive to the uncertainty Δc in the position of the edge points $c_1 \pm \Delta c$ and $c_2 \pm \Delta c$. This makes the model of the $\mu_A(Y, c_1, c_2)$ more proper choice for real-life modeling.

To illustrate this theoretical conclusion we presented the model results from the real study of iron deficiency anemia in infants. We showed that the linear regression model with the fuzzy dependent variable depicted the category ***Anemia*** has a persistent behavior over a reasonably wide interval of hemoglobin levels.

REFERENCES

[1] Bolotin A., "Fuzzification of Linear Regression Models with Indicator Variables in Medical Decision Making", In: Proceedings of the CIMCA 2005, IEEE, 2006, Vol. 1, pp. 572-577.
[2] Bolotin A., "Replacing indicator variables by fuzzy membership functions in statistical regression models: Examples of epidemiological stud-

ies" In: Lecture Notes in Computer Science: Biological and Medical Data Analysis, Springer, 2004, pp. 251-258.
[3] Bolotin A., "Uncertain categories in medical data analysis", In: Proceedings of the IPMU 2006, Paris, 2006, to be printed.
[4] Gujarati, D., *Basic Econometrics* [Chapter 15: Regression on dummy-dependent variables]. McGraw-Hill, 2003.
[5] Irwin J., Kirchner, J., "Anemia in Children", Am. Fam. Physician, 64 (2001), 1379-86.
[6] Johnston, J., DiNardo, J., *Econometric Methods* [Chapter 13: Discrete and limited dependent variable models], McGraw-Hill. 1997.
[7] Levy A., Fraser D., Rosen S., Dagan R., Deckelbaum R., Coles C., Naggan L., "Anemia as a risk factor for infectious diseases in infants and toddlers: results from a prospective study", Eur J Epidemiol., 2005; 20(3): 277-84
[8] Maronna R., Martin D., Yohai V., *Robust Statistics: Theory and Methods*, Wiley, 2006.
[9] Mazumdar M., Glassman R., "Categorizing a Prognostic Variable: Review of Methods, Code for Easy Implementation and Applications to Decision-Making about Cancer Treatments", Statist. Med. 19 (2000), 113-132.
[10] Peracchi F., *Econometrics* [Chapter 15: M-Estimators], Wiley, 2001.
[11] Preisser J., Koch G., "Categorical Data Analysis in Public Health", Ann. Rev. Public Health. 18 (1997), 51–82.
[12] Siberry G., Iannone R., eds., *The Harriet Lane handbook.* 15th ed. St. Louis: Mosby, 2000..