

Fuzzy Regression Analysis

using Trapezoidal Fuzzy Numbers

Souhir Charfeddine
ENAC and UT2, Toulouse, France
Souhir.charfeddine@enac.fr

Karim Zbidi
ENAC and UT1, Toulouse, France
zbidi@recherche.enac.fr

Felix Mora-Camino
ENAC, Toulouse, France
Felix.mora-camino@enac.fr

Abstract

When it is question of prediction, no deterministic model can be totally efficient, especially when the output to be estimated is dependent imprecisely on many fluctuant variables measuring human behavior (cognitions, choices, consumption, etc.). Regressions based on fuzzy logic which combine statistics and expert's attitudes can be used to improve the estimation of such outputs. Those regressions are based on fuzzy logic which tries to traduce the human perception. In the literature, fuzzy linear regression has been developed following Tanaka's model (the pioneer of such models) but the majority of the works make use of triangular fuzzy numbers and symmetric ones. In this paper, an extension to these regressions using trapezoidal fuzzy numbers is displayed. This suggested method is intending to use optimally the available data and gives the decider the opportunity to intervene and to use his experience in order to improve the quality of predictions.

Keywords: Fuzzy Logic, Fuzzy Regression Analysis, Output Estimation.

1 Introduction

The purpose of regression analysis is to relate analytically the variation of a dependent variable Y in terms of explanatory variables x_1, \dots, x_N . An

estimation of Y denoted \hat{Y} in terms of $X (= [x_1 \dots x_N])$ can be obtained from data samples through for instance a linear statistical regression. The analysis of this latter has been much considered [3], where f is *naturally* taken as a crisp linear function such as:

$$f(X) = a_0x_0 + a_1x_1 + \dots + a_Nx_N \quad (1)$$

with $x_0 = 1$ where a_0, a_1, \dots, a_N are real values.

Defining $A = (a_0, a_1, \dots, a_N)$ and since in general the relationship between the input and the output cannot be known exactly, a random variable u which represents the disturbance or the error term can be added so it is possible to write for every pair of input output (X_i, y_i) :

$$y_i = AX_i + u_i \quad (2)$$

This disturbance term is a surrogate for the uncertainty due not only to the *a priori* linear form chosen for function f , but also to the omitted variables that affect the output. The vector of the parameters a_j is then estimated through well established methods [3] such as 'least square regressions'.

Then, given a set of predefined inputs X , a crisp estimation of Y will be given by:

$$\hat{Y} = AX \quad (3)$$

and to get some insight into the estimation error, strong assumptions related with the distribution of the data must be made (for example the values of the

error terms may be supposed mutually independent and identically distributed [3] along a centred normal distribution $N(0, \sigma)$.

But although statistical regressions have many applications, problems can occur for several situations: for example when the available historical set is small or when the assumptions related to this data are not gathered or when the relationship between input and output is vague. Fuzzy sets have been used to contain the uncertainty related with the input-output relationship through fuzzy regressions based on statistics. In some cases, the variables can be themselves fuzzy. Here we will focus on the models where the data is crisp and where only the relationship between the explanatory variables and the output is fuzzy.

Tanaka was the first to propose fuzzy regressions based on statistics [8], since then and built upon that model, several methods have been developed. Almost all of these methods are making use of triangular fuzzy numbers. In this paper, after a recall of the principles of such classical model, extensions using trapezoidal fuzzy numbers and trying to take into account the dispersion of the data samples are proposed. In the last section, these methods are illustrated via an example.

2 Tanaka's Model

2.1 Model exposition

In fuzzy linear regression (FLR) analysis [1], some of the assumptions of the classical statistical approach are relaxed and the uncertainty is traduced by a fuzzy relationship between the input and the output. Such a relationship is given by a fuzzy function \tilde{f} . The present paper considers first the model of Tanaka [8] which is a pioneer for such models.

The basic Tanaka's model assumes a linear fuzzy function:

$$\tilde{f}(X) = \tilde{A}_0 x_0 + \tilde{A}_1 x_1 + \dots + \tilde{A}_N x_N = {}^t \tilde{A} X \quad (4)$$

with $x_0 = 1$ where \tilde{A} is the fuzzy vector of the model's parameters.

For every $j \in \{0, 1, \dots, N\}$, \tilde{A}_j is a symmetric fuzzy number presented by (c_j, w_j) where c_j and w_j are respectively its centre and its width. The reference membership function of these numbers is denoted L and is such as:

- $L(0) = 1$
- L is decreasing on $[0, 1[$
- $L(x) = 0$ when $x \in [1, +\infty[$
- L is concave on $] -1, 1[$

Two simple shapes are drawn in fig. 1.

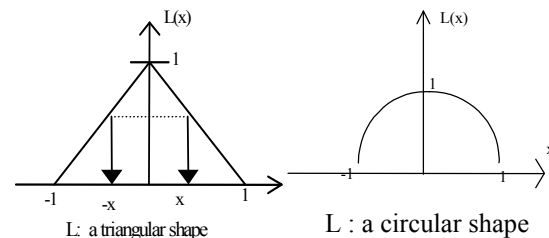


Figure 1: Examples of reference membership functions

The membership function $\mu_{\tilde{A}_j}$ is deduced from L as $\mu_{\tilde{A}_j}(a_j) = L((a_j - c_j) / w_j)$ when $w_j > 0$. (5)

A particular case is when the \tilde{A}_j are triangular, this case is the most developed in the literature. here:

$$L(x) = \begin{cases} 1 - |x| & \text{if } -1 \leq x \leq 1 \\ 0 & \text{if not} \end{cases} \quad (6)$$

and

$$\mu_{\tilde{A}_j}(a_j) = \begin{cases} 1 - |c_j - a_j| / w_j & \text{if } c_j - w_j \leq a_j \leq c_j + w_j \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

It can be shown (see [8]) that when the \tilde{A}_j are triangular fuzzy numbers, then the resulting Y is a triangular fuzzy number as well. The centre of Y is then ${}^t C X$ and its width is the sum of the widths of

all the terms: ${}^tW|X|$, where C is the vector of the centres of the \tilde{A}_j and W is the one of their widths . The membership function of Y is then given by:

$$\mu_Y(y) = \begin{cases} \text{Max}(0, |y - {}^tCX| / {}^tW|X|), & \text{if } X \neq 0 \\ 1 & \text{if } X = 0, y \neq 0 \\ 0 & \text{if } X = 0, y = 0 \end{cases} \quad (8)$$

Then the uncertainty about Y is illustrated by the width of the membership function of the resulting fuzzy number. Given a set of data samples D, it appears to be of interest to minimise the total vagueness resulting from the fuzzy regression through the tuning of its parameters.

Given a threshold number h ($0 \leq h \leq 1$), let us define a reduced data set D_h where the sample i is retained if y_i has a membership degree greater than h (see fig.2):

$$\forall i \in \{1, 2, \dots, M_h\}, \mu_{\tilde{y}_i}(y_i) \geq h \quad (9)$$

where M_h is the size of D_h . This can be written:

$$L(|y_i - {}^tCX_i| / {}^tW|X_i|) \geq h \quad (10)$$

and since L is decreasing over $[0, 1[$ then:

$$|y_i - {}^tCX_i| \leq L^{-1}(h) \cdot {}^tW|X_i| \quad (11)$$

Observe that in the case of triangular fuzzy numbers, $L^{-1}(h) = 1 - h$. Let us estimate the total vagueness associated to D_h and W as:

$$\sum_{i=1}^M \left(\sum_{j=0}^N w_j |x_{ij}| \right) = \sum_{j=0}^N \left(\sum_{i=1}^M |x_{ij}| \right) w_j \quad (12)$$

Then a linear program can be formulated to minimise this total vagueness index under h-degree membership constraints over D_h :

$$\begin{cases} \delta_L^h = \underset{W, C}{\text{Min}} \sum_{j=0}^N \left(\sum_{i=1}^M |x_{ij}| \right) w_j \\ \sum_{j=0}^N c_j x_{ij} + |L^{-1}(h)| \sum_{j=0}^N w_j |x_{ij}| \geq y_i \quad \forall i = 1, \dots, M_h \\ \sum_{j=0}^N c_j x_{ij} - |L^{-1}(h)| \sum_{j=0}^N w_j |x_{ij}| \leq y_i \quad \forall i = 1, \dots, M_h \\ W \geq 0, C \in \mathfrak{R}^N, x_{i0} = 1; i = 1, \dots, M_h. \end{cases} \quad (13)$$

The resulting linear fuzzy regression model will be denoted F_L^h .

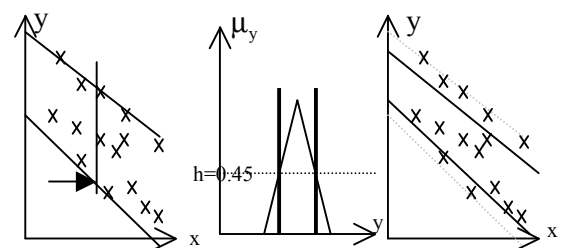


Figure 2: Interval estimation and membership threshold

2.2 Analysis of model:

If the vector $C_h^* = (c_0^*, c_1^*, \dots, c_N^*)$ and the vector $W_h^* = (w_0^*, w_1^*, \dots, w_N^*)$ compose the optimal solution of problem (13), then the vector of estimated parameters resulting from the regression F_L^h is:

$$\hat{A}_L^h = (C_h^*, W_h^*)_L \quad (14)$$

When another membership degree h' ($h' \neq h$ and $h' < 1$) is considered, it is easy to show that the resulting linear fuzzy regression $F_L^{h'}$ is given by:

$\hat{A}_L^{h'} = (C_h^*, [L^{-1}(h) / L^{-1}(h')] W_h^*)_L$. Then, once a given reference function L is adopted, the LFR associated to a threshold h can be deduced from the one corresponding to $h = 0$.

This model can be interpreted as an estimation of the interval of the dependent variable Y. At the beginning ($h = 0$) an interval containing all the observations is defined and when an effective threshold h is chosen a resulting narrower data interval is defined for the estimation. As some data samples located near the bounds of the current interval become outliers, they are removed from the refined data set. Some observations can be made

here about this method: it can be instructive to interpret the detected outliers samples instead of merely removing them. This method does not take fully into consideration the effective dispersion of the data samples within the learning interval. When rather large uncertainties are involved, L may be not a strictly decreasing function on $[0,1[$ (trapezoidal numbers can be of interest in this case) and the above approach is no more applicable.

3 Extensions of the Tanaka's model:

Despite the presentation of the model in a general case for the form of the reference membership function L, the literature has mainly treated the triangular shape and more precisely the symmetric one. In this section we intend to extend the fuzzy linear regressions to the trapezoidal fuzzy numbers.

The proposed extensions make use of *level fuzzy functions* in the sense of Zimmermann [9]. A level fuzzy function \tilde{f} is given by

- Four level crisp functions: f_a, f_b, f_c, f_d .
- f_b, f_c provide the curves for which the degree of membership reaches 1.
- f_a, f_d provide the curves for which the grade of membership starts from zero.

For consistency reasons, these four functions cannot intersect on the input domain given by $[X_{\min}, X_{\max}] (= [(x_1)_{\min}, (x_1)_{\max}] \times \dots \times [(x_N)_{\min}, (x_N)_{\max}])$: $\forall x \in [X_{\min}, X_{\max}] f_a(x) \leq f_b(x) \leq f_c(x) \leq f_d(x)$

Then a membership function can be attached to the level fuzzy function:

$$\mu_{\tilde{f}}(f(x)) = \begin{cases} \frac{(f(x) - f_a(x))}{f_b(x) - f_a(x)} & \text{if } f_a(x) \leq f(x) \leq f_b(x) \\ 1 & \text{if } f_b(x) \leq f(x) \leq f_c(x) \\ \frac{f_d(x) - f(x)}{f_d(x) - f_c(x)} & \text{if } f_c(x) \leq f(x) \leq f_d(x) \\ 0 & \text{otherwise} \end{cases} \quad (15)$$

A simple way to determine the extreme level functions f_a and f_d is to use the Tanaka's model considering the resolution of (13) for $h = 0$, giving:

$$f_a(X) = \sum_{j=0}^N c_j^* x_j - \sum_{j=0}^N w_j^* |x_j| \quad (16)$$

and

$$f_d(X) = \sum_{j=0}^N c_j^* x_j + \sum_{j=0}^N w_j^* |x_j| \quad (17)$$

The determination of the central functions f_b and f_c is not so straightforward. They provide the bounds of the certainty domain. There are many ways to define them. In the following, two methods are considered.

3.1 Method using an h-cut:

The h-cut considered in the Tanaka's model can be used here to define the bounds of the set of possibilities that will correspond to the certainty domain. It is assumed that any output value having a membership level higher than a given level h_1 ($h_1 \in]0,1[$) is in the certainty domain. So f_b and f_c are here defined by the resolution of (13) where L is the triangular reference membership function and h is a chosen number in $[0,1[$.

The h-cut considered in the Tanaka's model can be used here to define the bounds of the set of possibilities that will correspond to the certainty domain. It is assumed that any output value having a membership level higher than a given level h_1 ($h_1 \in]0,1[$) is in the certainty domain. So f_b and f_c are here defined by the resolution of (13) where L is the triangular reference membership function and h is a chosen number in $[0,1[$ as illustrated by fig.3. The obtained fuzzy numbers are defined by symmetric trapezoidal fuzzy numbers.

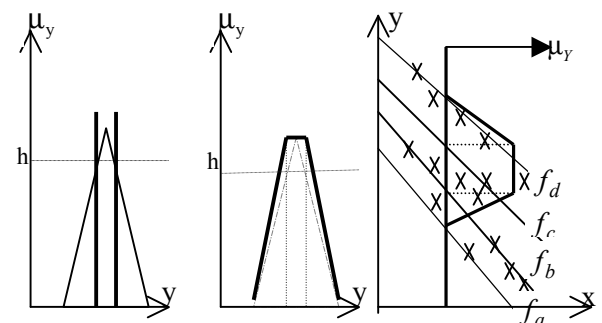


Figure 3: Interval estimation and construction of trapezoidal numbers

3.2 Mixed method

The methods presented above do not pay a direct attention to how the data samples are dispersed in the interval. With the mixed method, the level functions f_a and f_d are obtained through a 0-cut using Tanaka's model while a least square regression is used to determine the central level functions f_b and f_c . From the resulting statistical regression model $\hat{f}(X)$ and standard deviation σ (of the error u distribution), f_b and f_c are given by:

$$f_b = \hat{f} - \lambda\sigma \tag{18}$$

and

$$f_c = \hat{f} + \lambda\sigma \tag{19}$$

where λ is a positive constant chosen by an expert depending on his opinion about the representativeness of the proposed samples. For instance a large λ means that he has a poor opinion about their representativeness.

These two proposed methods define trapezoidal fuzzy numbers taking into account all the data samples for the definition of their limits since they are effective realizations. Besides that, experts can choose directly the criteria used to determine the central functions f_b and f_c . The possibilities above $f_c(X)$ can be interpreted as corresponding to optimistic scenarios and the ones under $f_b(X)$ can be associated to pessimistic conditions.

4 Case study

Here a simple case is considered: the output y depends only on one variable x . Table1 gives a set of data pairs to be used for the estimation.

Table1: simple data set

x_i	2	3	4	5	6	7	8	9	10	11	12	13	14	15
y_i	15	17	17	16	18	18	17	19	19	19	20	20	18	21

The purpose here is to estimate y by a fuzzy level function defined by the crisp linear levels: y_a, y_b, y_c, y_d , built as suggested in the last section: both given methods estimate the extreme levels y_a and y_d as Tanaka proposes to do (see fig.4).

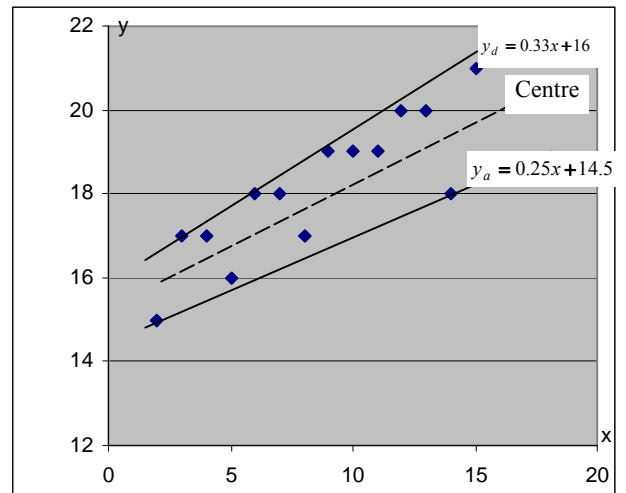


Figure 4: Interval estimation
Upper and lower lines estimation

Then the first proposed method uses a h-cut to build the central level linear functions as illustrated by fig.5.

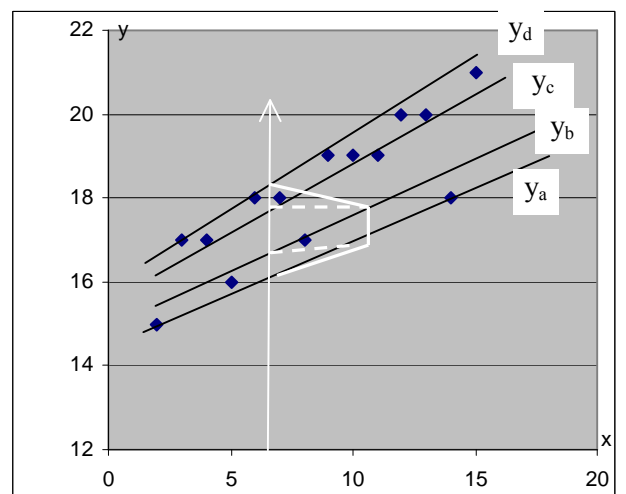


Figure 5 : Trapezoidal fuzzy numbers estimation using h-cut

Fig.6 draws an application of the mixed method to estimate trapezoidal fuzzy numbers.

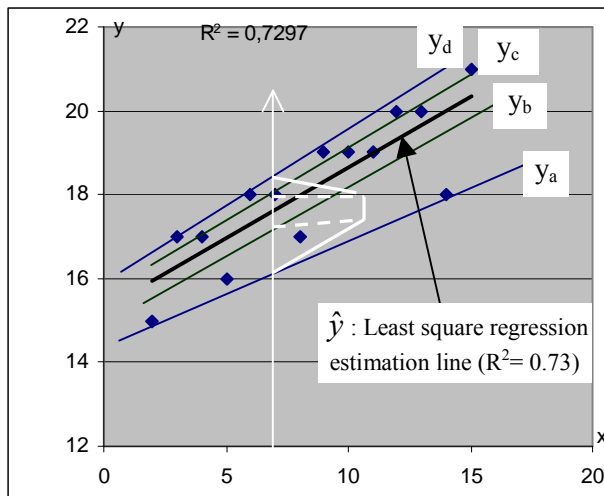


Figure 6 : Trapezoidal fuzzy numbers estimation using mixed method

Despite the high linear correlation between the used data, the dispersion is not to be neglected. An estimation with classical statistical methods could reduce the predictions to only one line (\hat{y}) but through the fuzzy regressions many values are possible, each with an associated degree of membership. The proposed methods allow to construct trapezoidal fuzzy numbers through an estimated fuzzy level function. Note that the last method takes into account the dispersion of the data in the interval.

5 Conclusion

In this paper, a new approach for fuzzy linear regression analysis has been introduced. This approach is inspired from the Tanaka's method. The target of this approach is to build trapezoidal fuzzy sets for the estimated variable. It tries also to take into account all the data samples and sometimes the dispersion of these latter. A simple example has been treated to illustrate these methods.

References

- [1] T. CHEN and M.J. WANG, "Forecasting methods using fuzzy concepts", *Fuzzy sets and systems*. 105 (1999) 339-352.
- [2] P. Diamond, "Fuzzy least squares", *Information sciences* 46(3), 141-157.
- [3] B. DORMONT, « *Introduction à l'économétrie* ». Montchrestien, EJA., 1999. PP 450. ISBN : 2.7076.1020.8
- [4] B.K. PAPADOPOULOS and M.A. SIRPI, "Similarities in fuzzy regression models", *Journal of Optimization Theory and Applications*, Vol 102, No 2. pp. 373-383. August 1999.
- [5] G. PETERS, "Fuzzy linear regression with fuzzy intervals", *Fuzzy Sets and Systems* 63 (1994), pp45-55.
- [6] V.A. PROFILLIDIS, "Econometric and fuzzy models for the forecast the forecast of demand in the airport of Rhodes" *Journal OF AIR TRANSPORT MANAGEMENT* 6(2000) 95-100.
- [7] A.F. Shapiro, "Fuzzy Regression and the term structure of interest rates revisited", *AFIR* 2004.
- [8] H. TANAKA, S. UEJIMA and ASAI. "Linear regression analysis with fuzzy model", *IEEE Transactions Systems, Man and Cybernetics* 12 (1982) 903-907.
- [9] H.J. ZIMMERMANN, "*Fuzzy Set Theory- and Its Applications*", edition: Hardcover. PP.544.(1991). ISBN : 0792374355