Fuzzy linear regression
Application to the estimation of air transport demand

Souhir CHARFEDDINE
UTM, ENAC , Toulouse
Tel: (33) 5 62 17 46 95
e-mail: souhir.charfeddine@enac.fr

Félix MORA-CAMINO
ENAC, Toulouse
e-mail: felix.mora@enac.fr

Marc De Coligny
UTM, Toulouse
e-mail: coligny@univ-tlse2.fr

Abstract:
The demand for an air transport market is very sensitive to many factors. The vagueness of the impacts of all these factors makes the task of prediction of this demand by classical methods very hazardous, especially when this estimation is used afterwards for critical decisions such as those related with the definition of supply (frequency of flights, number of seats put on the market, trip price..). Then it appears that crisp methods are not able to take fully into account all the uncertainty making up the demand while possibilistic reasoning could be a way to catch it. Following this idea, it is shown in this communication how regressions based on fuzzy logic which combine statistics and experts' attitudes can be used to improve the estimation for air transport demand.
In the first section of the communication, following Tanaka's model, fuzzy linear regression is introduced. Then in the second part an extension using trapezoidal fuzzy numbers is displayed. Finally, in the last section, the application of the proposed fuzzy linear regression to the estimation of air transport demand is considered.

Keywords: Fuzzy Logic, Fuzzy Regression Analysis, Demand Estimation

1. Introduction:

The purpose of regression analysis is to relate analytically the variation of a dependent variable Y in terms of explanatory variables $x_1,...,x_N$. An estimation of Y denoted $\hat{Y}$ in terms of $X(={}^t[1\ x_1\ ...\ x_N])$ can be obtained from data samples (see table 1) through a linear statistical regression. The analysis of this latter has been much considered [2], where *f* was *naturally* taken as a crisp linear function such as:

$$f(X) = a_0 x_0 + a_1 x_1 + ... + a_N x_N \qquad \text{with} \qquad x_0 = 1$$

(1)

where $a_0, a_1,..., a_N$ are real values. Defining $A={}^t(a_0, a_1,..., a_N)$, we can write: $f(X)={}^tAX$.

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 to the right side of *(1)*:

$$y_i = f(X_i) + u_i$$

(2)

This disturbance term is a surrogate for the uncertainty due not only to the *a priori* affine 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 a least square regression as: $\hat{A} = ({}^tTT)^{-1}\,{}^tT\Upsilon$ where T is the matrix composed by the inputs samples and Y is the vector of the output samples:

$$T = \begin{bmatrix} 1 & x_{11} & .. & x_{1N} \\ 1 & x_{21} & .. & x_{2N} \\ . & . & .. & . \\ . & . & .. & . \\ 1 & x_{M1} & .. & x_{MN} \end{bmatrix} \quad \text{and} \quad \Upsilon = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_M \end{bmatrix}$$

with the condition that the matrix ${}^tTT$ is non singular.

sample	Output	Inputs
1	y_1	$x_{11},...,x_{1N}$
.	.	.
.	.	.
i	y_i	$x_{i1},...,x_{iN}$
.	.	.
M	y_M	$x_{M1},...,x_{MN}$

Table 1: Observed input output data

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

$$\hat{Y}={}^{t}\hat{A}X \qquad (3)$$

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 can be supposed mutually independent and identically distributed [2] along a centred normal distribution $N(0,\sigma)$). In the following, fuzzy sets are used to contain the uncertainty related with the inputs-output relationship.

2. Tanaka's model:

2.1. Model's 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 [5] which is a pioneer for such models.

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

$$\tilde{f}(X)=\tilde{A}_0x_0+\tilde{A}_1x_1+\ldots+\tilde{A}_Nx_N={}^{t}\tilde{A}X \qquad \text{With} \quad x_0=1 \qquad (4)$$

Where $\tilde{A}$ is the fuzzy vector of the model's parameters.

For every $j\in\{0,1,\ldots,N\}$, $\tilde{A}_j$ 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:

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

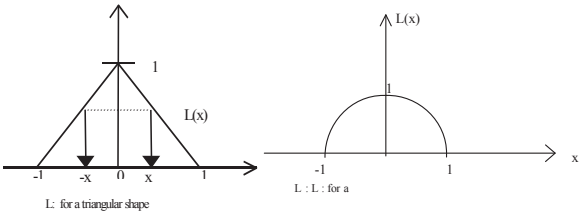


Fig.1: Examples of reference membership functions

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 $w_j > 0$.

An interesting case is when the $\tilde{A}_j$ are triangular. It is the case the most developed in the literature; where:

$$L(x) = \begin{cases} 1-|x| & \text{if } -1 \le x \le 1 \\ 0 & \text{if not} \end{cases} \qquad (5)$$

and

$$\mu_{\tilde{A}_j}(a_j) = \begin{cases} 1-|c_j - a_j|/w_j & \text{if } c_j - w_j \le a_j \le c_j + w_j \\ 0 & \text{otherwise.} \end{cases} \qquad (6)$$

It can be shown (see [5]) that when the $\tilde{A}_j$ are triangular fuzzy numbers, then the resulting Y *(7)* is a triangular fuzzy number as well. The centre of Y is then ${}^tCX$ 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:

$$\mu_Y(y) = \begin{cases} \text{Max}\,(0, 1-\dfrac{|y - {}^tCX|}{{}^tW|X|}), & \text{if } X \ne 0 \\ 1 & \text{if } X = 0, y \ne 0 \\ 0 & \text{if } X = 0, y = 0 \end{cases} \qquad (7)$$

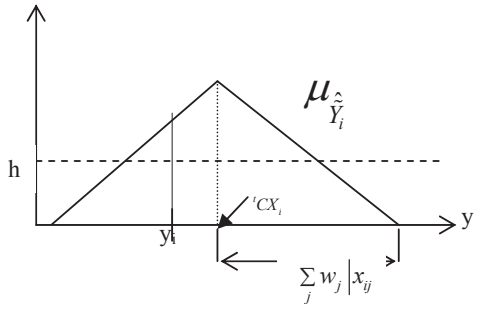


Fig.2. Membership function of $\tilde{Y}_i$

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 ($0 \le h \le 1$), 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:

$$\forall i \in \{1,2,...,M_h\}, \mu_{\tilde{Y}_i}(y_i) \ge h$$

(8)

where M_h is the size of D_h

This can be written: $L(|y_i - {}^tCX_i| / {}^tW|X_i|) \ge h$ and since L is decreasing over $[0,1[$ then:

$$|y_i - {}^tCX_i| \le L^{-1}(h).{}^tW|X_i| \qquad (9)$$

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:

$$\sum_{i=1}^{M}(\sum_{j=0}^{N} w_j |x_{ij}|) = \sum_{j=0}^{N}(\sum_{i=1}^{M} |x_{ij}|)w_j \qquad (10)$$

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

$$\begin{cases} \delta_L^h = \underset{W,C}{Min} \quad \sum_{j=0}^{N}(\sum_{i=1}^{M}|x_{ij}|)w_j & (a) \\ \text{st} \quad \sum_{j=0}^{N} c_j x_{ij} + |L^{-1}(h)| \sum_{j=0}^{N} w_j |x_{ij}| \ge y_i \quad \forall i = 1,...M_h & (b) \\ \sum_{j=0}^{N} c_j x_{ij} - |L^{-1}(h)| \sum_{j=0}^{N} w_j |x_{ij}| \le y_i \quad \forall i = 1,...M_h & (c) \\ W \ge 0, C \in \Re^N, x_{i0} = 1; i = 1,...,M_h. & (d) \end{cases} \qquad (11)$$

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

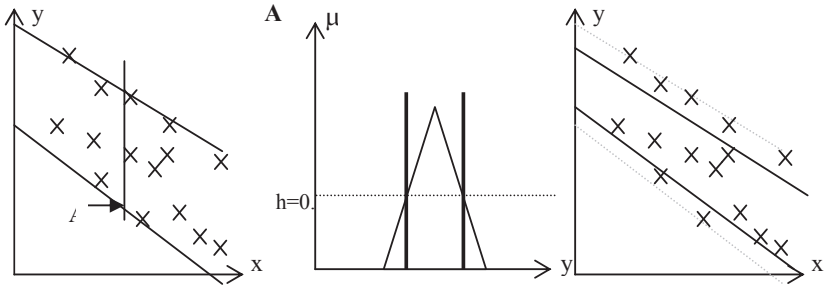


Fig.4.2 .Interval estimation and membership

2.2. Analysis of model:

If $C_h^* = {}^t(c_0^*, c_1^*, ..., c_N^*)$ and $W_h^* = {}^t(w_0^*, w_1^*, ..., w_N^*)$ compose the optimal solution of problem *(11)*, then the vector of estimated parameters resulting from the regression F_L^h is:

$$\hat{\tilde{A}}_L^h = (C_h^*, W_h^*)_L . \qquad (12)$$

When another membership degree h' ($h' \neq h$) is considered, it is easy to show that the resulting linear fuzzy regression $F_L^{h'}$ is given by: $\hat{\tilde{A}}_L^{h'} = (C_h^*, [L^{-1}(h)/L^{-1}(h')]W_h^*)_L$. 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 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. Extension of the Tanaka's model:

The proposed extension makes use of *level fuzzy functions* in the sense of Zimmermann[6]. 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 $[X_{\min}, X_{\max}]$ ($= [(x_1)_{\min}, (x_1)_{\max}] \times ... \times [(x_N)_{\min}, (x_N)_{\max}]$):

$$\forall x \in [X_{\min}, X_{\max}], f_a(x) \leq f_b(x) \leq f_c(x) \leq f_d(x)$$

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

$$\mu_{\tilde{f}}(f(x)) = \begin{cases} (f(x)-f_a(x))/(f_b(x)-f_a(x)) & \text{if} \\ f_a(x) \le f(x) \le f_b(x) & (13) \\ 1 & \text{if} \\ f_b(x) \le f(x) \le f_c(x) & \\ (f_d(x)-f(x))/(f_d(x)-f_c(x)) & \text{if} \quad f_c(x) \le f(x) \le f_d(x) \\ 0 & \\ \text{otherwise} & \end{cases}$$

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 *(11)* for $h = 0$:

$$\begin{cases} \quad Min \quad \sum_{j=0}^{N} w_j \sum_{i=1}^{M} \left| x_{ij} \right| & (a) \\ \text{st} \quad \sum_{j=0}^{N} c_j x_{ij} + \sum_{j=0}^{N} w_j \left| x_{ij} \right| \ge y_i \quad \forall i = 1,...M & (b) \\ \quad \sum_{j=0}^{N} c_j x_{ij} - \sum_{j=0}^{N} w_j \left| x_{ij} \right| \le y_i \quad \forall i = 1,...M & (c) \\ W \ge 0 \, , C \in \Re^N , x_{i0} = 1; i = 1,...,M. & (d) \end{cases} \qquad (14)$$

giving: $f_a(X) = \sum_{j=0}^{N} c_j^* x_j - \sum_{j=0}^{N} w_j^* \left| x_j \right|$ and $f_d(X) = \sum_{j=0}^{N} c_j^* x_j + \sum_{j=0}^{N} w_j^* \left| x_j \right|$

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.

a) 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 *(11)* where L is the triangular reference membership function and h is a chosen number in $[0,1[$.

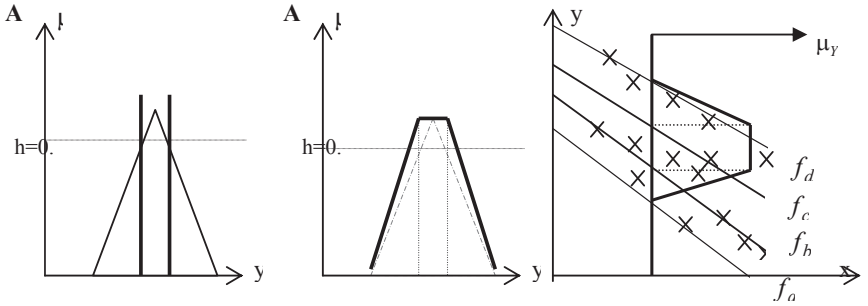


Fig.5 . Interval estimation and construction of trapezoidal

b) Mixed method

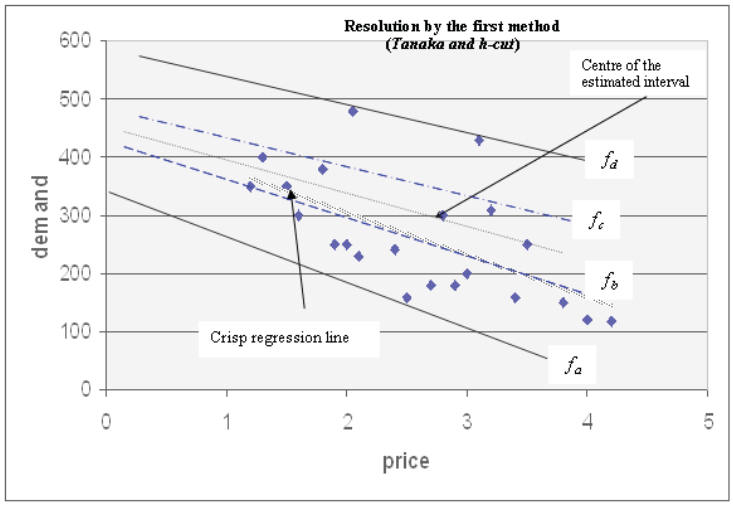
Since the methods presented above do not pay a direct attention to how the data samples are dispersed in the interval, with the mixed method, the level functions f_a and f_d are obtained through a 0-cut using Tanaka's model while a least square regression is used to determine the central level functions f_b and f_c. From the resulting statistical regression model $\hat{f}(X)$ and standard deviation σ, f_b and f_c are given by $f_b = \hat{f} - \lambda\sigma$ and $f_c = \hat{f} + \lambda\sigma$ where λ is a positive constant chosen by an expert depending about his opinion about the representativeness of the proposed samples. A large λ means that he has a poor opinion about their representativeness.

These two methods define trapezoidal fuzzy numbers taking into account all the data samples for the definition of their limits since they are effective realisations. 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.

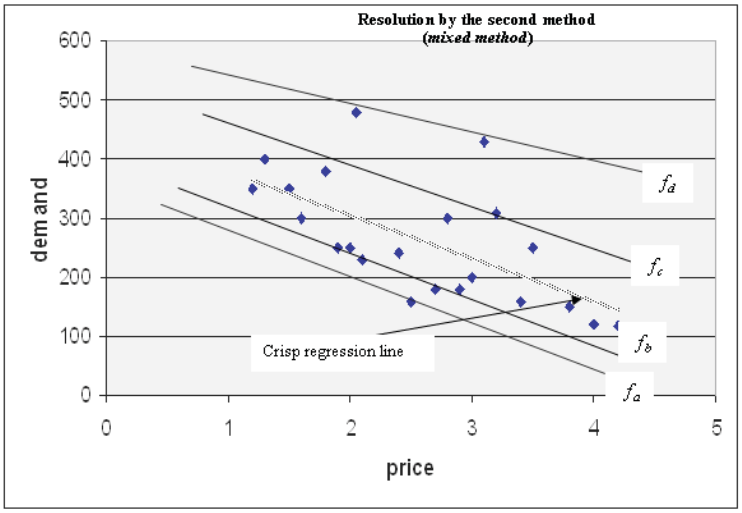
4. Example

In the air transport market uncertainty is very frequent so let us consider a simple example of air transport demand estimation as function of ticket price. The table below provides a set of data samples:

price(u.m)	2	1,5	3	4	2,5	1,8	2,4	3,5	2,8	3.1	3,2
Demand	250	350	200	120	160	380	240	250	300	430	310
price(u.m)	3,8	1,6	1,3	2,1	1,9	3,4	2,7	1,2	4,2	2.05	2,9
Demand	150	300	400	230	250	160	180	350	118	480	180



Tanaka's method in this example provides an estimate of the demand function centred in the interval (h=0), a large number of observations is then excluded, contrarily to the proposed method to build trapezoidal fuzzy numbers which keeps all these data samples. In this case, it can be observed that the concentration of the data is rather bellow the centre of the fuzzy estimation.



With the second trapezoidal fuzzy numbers method, the crisp linear regression estimate remains in the centre of the base of the trapezoidal fuzzy estimation.

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 compare these methods.

References

[1] CHEN T., WANG M.J. *Forecasting methods using fuzzy concepts*. Fuzzy sets and systems. 105 (1999) 339-352.

[2] DORMONT B. *Introduction à l'économétrie*. Montchrestien, EJA., 1999. PP 450. ISBN : 2.7076.1020.8

[3] PAPADOPOULOS B.K., SIRPI M.A. Similarities in Fuzzy Regression Models. Journal of Optimization Theory and Applications, Vol 102, No 2. pp. 373-383. August 1999.

[4] PETERS G. *Fuzzy linear regression with fuzzy intervals.* Fuzzy Sets and Systems 63 (1994), pp45-55.

[5] TANAKA H., UEJIMA S. et ASAI. *Linear Regression Analysis with Fuzzy Model,* IEEE Transactions Systems, Man and Cybernetics 12 (1982) 903-907.

[6] ZIMMERMANN Hans J. *Fuzzy Set Theory- and Its Applications*. Edition : Hardcover. PP.544.(1991). ISBN : 0792374355