BRIEF NOTES ON FUZZY REGRESSION*

Milan Mareš

ÚTIA AV ČR, P. O. Box 18, 182 08 Praha 8
Czech Republic
mares@utia.cas.cz

The linear regression belongs to significant characteristics of the development of homogeneous series of data. If the data are vague then it is important to find adequate approach to their regression which could reflect both types, stochastic and fuzzy, of uncertainty which are included in them. In the following sections we briefly discuss three of such possible methods. The first one simply repeats the classical statistical methods for processing fuzzy quantities. The second method simply (and rather voluntarily) processes the empirical fuzzy data, and the third one suggests the application of statistical formulas to important values of trapezoidal fuzzy quantities.

1 Uncertainty in Regression – Heuristic Introduction

The formal quantification of dependence between two (or more) series of data belongs to significant problems of applied mathematics. The classical statistical regression model is well elaborated and its transmission to the fuzzy environment could seem to be natural. As follows, e. g., from the contributions selected in edited volume [2] the practical realization of this task need be neither simple nor easily interpretable.

The aim of this paper is to contribute, at least briefly, to the problem of practical calculation of some analogies of the statistical regression model. The stochastic counterpart of our considerations will be the basic model of linear regression between two random quantities. Let us remember that for two sequences of numerical values $(\xi_1, \xi_2, \ldots, \xi_n)$, $(\eta_1, \eta_2, \ldots, \eta_n)$ representing n realizations of random quantities Ξ and H the linear regression relation is

$$H = \boldsymbol{a} + \boldsymbol{b}\,\Xi \tag{1}$$

where

$$\boldsymbol{b} = \frac{\sum_{i=1}^{n}(\xi_i - \overline{\xi})\cdot(\eta_i - \overline{\eta})}{\sum_{i=1}^{n}(\xi_i - \overline{\xi})^2}, \quad \boldsymbol{a} = \overline{\eta} - \boldsymbol{b}\,\overline{\xi}, \quad \overline{\xi} = \frac{\sum_{i=1}^{n}\xi_i}{n}, \quad \overline{\eta} = \frac{\sum_{i=1}^{n}\eta_i}{n}. \tag{2}$$

In the following sections we assume that some of the values existing in the above formulas are fuzzy. Namely, we assume only the vagueness of the values $\eta_1, \ldots, \eta_n$ meanwhile the numbers $\xi_1, \ldots, \xi_n$ are exactly known realizations of the random variable Ξ (for example, they denote time values in the model of time-series).

Here, we would briefly discuss the sources and types of uncertainty existing in Ξ and, especially, in H. It is not realistic to assume that the single one is the vagueness, it means that the values $\eta_1, \eta_2, \ldots, \eta_n$ are essentially determined but not exactly known or specified by vague verbal expressions. In such case the linear relation (1) would be fulfilled in certain sense deterministically, where the fuzzy values fulfil (1) regularly (e. g. their modal values are located on a straight line). It is more realistic to assume that the uncertainty connected with the values

*The research summarized in this paper was partly supported by the Key project of the Academy of Sciences of the Czech Republic No. K 1075601, Grant Agency of the Academy of Sciences No. A 1075905, Grant Agency of the Czech Republic No. 402/99/0032, and the Ministry of Education, Youth and Sports of the Czech Republic project No. VS 96063.

$\eta_1, \eta_2, \ldots, \eta_n$ combines randomness and vagueness. They represent a sequence of vaguely known realizations of random variable and in this sense they display both, probabilistic and fuzzy, features. It means that it has a sense to combine statistical and fuzzy set theoretical approaches to their processing.

In the following sections we formulate and discuss three of such combined approaches.

2 Conservative Computation

The first of the combined approaches to stochastical-fuzzy regression means a quite mechanical reproduction of the usual statistical procedures (2) applied to fuzzy quantities $\eta_1, \ldots, \eta_n$. Let us stress that values $\xi_1, \ldots, \xi_n$, as well as $\overline{\xi}$, $(\xi_i - \overline{\xi})$, $(\xi - \overline{\xi})^2$, are crisp results of stochastic procedures generating their values. On the other hand $\eta_1, \ldots, \eta_n$ are fuzzy quantities (with randomly distributed possible values) which means that $\overline{\eta}$, $(\eta_i - \overline{\eta})$ and also $\boldsymbol{a}$, $\boldsymbol{b}$, are fuzzy quantities. Then we can proceed using one of the following methods.

2.1 Extension Principle

The first selfevident possibility is to use the classical extension principle (see [1]) for the definition of arithmetic operations with fuzzy quantities. In our case we need only the linear operations – the sum of fuzzy quantities and the product of fuzzy quantity and crisp real number. Their properties can be found, e. g., in [1] or [4]. If $r \in R$ is a real number, s, t are fuzzy quantities with membership functions μ_s, μ_t, then also the sum $s \oplus t$ and the product $r \cdot s$ are fuzzy quantities with membership functions

$$\mu_{s\oplus t}(x) = \sup_{y\in R} \left(\min(\mu_s(y), \mu_t(x-y))\right), \tag{3}$$

$$\mu_{r\cdot s}(x) = \mu_s(x/r), \quad \text{if } r \neq 0,\ x \in R. \tag{4}$$

It is relatively easy, using (3) and (4), to compute $\overline{y}$ as defined by (2) (n is a crisp number). The difference $(\eta_i - \overline{\eta})$ can be computed using (3) in $\eta_i \oplus (-\overline{\eta})$ where for any s

$$\mu_{-s}(x) = \mu_s(-x), \quad x \in R \tag{5}$$

and also the product $(\xi_i - \overline{\xi}) \cdot (\eta_i - \overline{\eta})$, $i = 1, \ldots, n$, the sum

$$(\xi_1 - \overline{\xi}) \cdot (\eta_1 - \overline{\eta}) \oplus (\xi_2 - \overline{\xi}) \cdot (\eta_2 - \overline{\eta}) \oplus \cdots \oplus (\xi_n - \overline{\xi}) \cdot (\eta_n - \overline{\eta}),$$

and the product of its fuzzy result with crisp

$$\frac{1}{\sum_{i=1}^{n} (\xi_i - \overline{\xi})^2}$$

can be computed. In this way we derive fuzzy quantities $\boldsymbol{a}$, $\boldsymbol{b}$ from (1) and the fuzzy regression formula is complete. The associativity and distributivity of the above operations, $r \cdot (s \oplus t) = r \cdot s \oplus r \cdot t$, means that there are no theoretical problems connected with this procedure.

Having completed the regression formula (1) with fuzzy coefficients we can theoretically determine for any realization ξ of Ξ the membership function of "typical" fuzzy value η which corresponds with the given value ξ.

There exists one methodological problem connected with the application of the extension principle. Its multiple application, namely of the summation formula (3), enormously increases the extent of the support of the membership function of the result. It means also enormous

growth of the range of possible values of, e. g., the fuzzy coefficients $\boldsymbol{a}$, $\boldsymbol{b}$. If these fuzzy coefficients are "too fuzzy" then the informative value of formula (1) with such coefficients is rather low. The fuzzy quantity $\eta = \boldsymbol{a} + \boldsymbol{b}\xi$ for typical realization ξ of Ξ covers such a wide interval of possible "typical" values that it offers only very little information about the properties of the modelled process.

In such case it is desirable to find another method for processing fuzzy quantities which could lead to more "concentrated" values of the result.

2.2 Generated Fuzzy Quantities

A model of fuzzy quantities and their processing which need not lead to enormous increase of formal uncertainty was suggested in the literature. It is treated, e. g., in [5] and [6] and it is based on the concept of generation of fuzzy quantities. Due to this concept every fuzzy quantity, let us say η_i in our case, consists of a *crisp numerical value* η_i^* combined with a "normalized" representative of fuzziness called *shape function* which represents the form of uncertainty included in its (usually verbal) characterization, and which we denote $\varphi : R \to [0, 1]$,

$$\varphi(0) = 1, \qquad \varphi(x) \text{ increases for } x < 0, \qquad \varphi(x) \text{ decreases for } x > 0.$$

The crisp value η^* localizes the fuzzy quantity η_i (more exactly – its membership function) on the real line and its actual membership function depends on the standard "normalized" shape φ and on this localization. This dependence is arranged by means of a real-valued function f called *scale*,

$$f(0) = 0, \qquad f(x) \text{ is strictly increasing.}$$

Exactly, the membership function μ_i of η_i is defined by

$$\mu_i(x) = \varphi\left(f(x) - f(\eta_i^*)\right), \quad x \in R. \tag{6}$$

If, for example,

$$\varphi(x) = \max(0,\, 1 - |x|)$$

and

$$\begin{aligned} f(x) &= x && \text{for} \quad x < 5, \\ &= \frac{x}{2} + \frac{5}{2} && x \geq 5, \end{aligned}$$

then for $\eta_i^* = 2$

$$\begin{aligned} \mu_i^{(2)}(x) &= 0 && \text{if} \quad x \leq 1 \\ &= x - 1 && \text{if} \quad x \in (1, 2) \\ &= 3 - x && \text{if} \quad x \in [2, 3) \\ &= 0 && \text{if} \quad x \geq 3 \end{aligned}$$

meanwhile for $\eta_i^* = 8$

$$\begin{aligned} \mu_i^{(8)}(x) &= 0 && \text{if} \quad x \leq 6 \\ &= \frac{x}{2} - 3 && \text{if} \quad x \in (6, 8) \\ &= 5 - \frac{x}{2} && \text{if} \quad x \in [8, 10) \\ &= 0 && \text{if} \quad x \geq 10. \end{aligned}$$

More details about this topic can be found in [5, 6] and some other papers.

The main effect of the concept of generated fuzzy quantity is the possibility to separate the calculation with quantitative numerical values of the quantities from the logical or semantical processing of their "normalized" uncertainties. In our case, we can apply formulas (2) for usual algebraical processing of crisp values ξ_i and (also crisp) values η_i^*, $i = 1, \ldots, n$. In this way we compute the crisp values a^* and b^* of the fuzzy quantities $\boldsymbol{a}$ and $\boldsymbol{b}$. Parallelly, we can derive the "normalized" shapes φ_a and φ_b from the shapes $\varphi_1, \ldots, \varphi_n$ of $\eta_1, \ldots, \eta_n$ by means of fuzzy logical principles, e. g.

$$\varphi_a(x) = \max(\varphi_1(x), \ldots, \varphi_n(x)) \quad \text{or} \quad \varphi_b(x) = \varphi_1(x) \cdot \ldots \cdot \varphi_n(x).$$

(Let us note that there exist numerous formulas for deriving φ_a and φ_b from $\varphi_1, \ldots, \varphi_n$. The extension principle (3) and formula (4) belong to them and it is worth mentioning that (3) belongs to the patterns which enormously extend the range of the result). The choice of actual procedure for the specification of φ_a and φ_b depends on the logical relations between the vague quantities in the sequence $\eta_1, \ldots, \eta_n$. Usually, the operations of maximum or minimum are quite satisfactory. The shapes φ_a, φ_b and the computed crisp values a^*, b^* can be combined with the scale f and, using (6), the membership functions μ_a and μ_b of $\boldsymbol{a}$ and $\boldsymbol{b}$, respectively, can be derived.

The procedure described above has some methodological advantages consisting in the separation of the quantitative and qualitative component of the fuzzy quantity, mentioned above, and it has also pleasant practical consequences. The supports of μ_a, μ_b are usually significantly narrower than in the case when the classical extension principle is used in the calculation. It means that the idea about the "outputs" of the regression with respect to given "input" crisp data $\xi_1, \ldots, \xi_n$ and fuzzy data $\eta_1, \ldots, \eta_n$, (where "outputs" means the linear relation (1)) is more reliable and more concentrated than if the extension principle is used. Both methods described in this section respect the theoretical paradigms of the fuzzy set theory.

3 Empirical Estimation

The values of fuzzy coefficients $\boldsymbol{a}$ and $\boldsymbol{b}$ in (1) can be estimated also much more empirically without deep theoretical background. One of such approaches is briefly mentioned in [3]. Its reliability depends on the existence of a qualified expert whose view is combined with the statistical (and fuzzy) empirical data.

For every fuzzy quantity η_i, $i = 1, \ldots, n$, we denote by η_i^+ its *modal value*, i. e., such real number for which $\mu_i(\eta_i^+) = 1$ where μ_i is the membership function of η_i. If there are more than one modal values then η_i^+ represents the "mean" one of them, i. e.,

$$\eta_i^+ = \frac{\max(x \in R : \mu_i(x) = 1) - \min(x \in R : \mu_i(x) = 1)}{2}.$$

Modal values η_i^+ are crisp numbers, and we can use formulas (2) with inputs $\xi_1, \ldots, \xi_n$, η_1^+, $\ldots, \eta_n^+$ to compute crisp values a^+, b^+ of the coefficients $\boldsymbol{a}$, $\boldsymbol{b}$ in (1),

Then an expert would be able to extend the crisp values a^+, b^+ into fuzzy quantities, for example in such way that he constructs real numbers a_1, a_2, b_1, b_2 for which $a_1 < a_2$, $b_1 < b_2$ and forms the membership functions μ_a, μ_b such that

$$\begin{aligned}
\mu_a(x) &= 0 && \text{if } x \le a_1, & \mu_b(x) &= 0 && \text{if } x \le b_1,\\
&= \frac{x - a_1}{a^+ - a_1} && a_1 < x \le a^+, & &= \frac{x - b_1}{b^+ - b_1} && b_1 < x \le b^+,\\
&= \frac{a_2 - x}{a_2 - a^+} && a^+ \le x < a_2, & &= \frac{b_2 - x}{b_2 - b^+} && b^+ \le x < b_2,\\
&= 0 && x \ge a_2, & &= 0 && x \ge b_2.
\end{aligned}$$

The expert can state the numbers a_1, a_2, b_1, b_2, for example, so that all modal values η_i^+, $i = 1, \ldots, n$ are between the lines

$$a_1 + b_1 \xi, \qquad a_2 + b_2 \xi$$

for $\xi = \xi_i$, $i = 1, \ldots, n$, or so that the area between these lines covers some significant part of possible values of η_i, $i = 1, \ldots, n$.

The approach described in this section is simple and lucid but it is, perhaps, rather too much dependent on the subjective qualities (and prejudices) of the expert. This discrepancy is rather limited if the following method can be used.

4 Fuzziness in Probabilistic Model

In this section we suppose that all fuzzy quantities η_1, $\eta_2, \ldots, \eta_n$ are trapezoidal. It means that there exist quadruples of real numbers

$$\left(\eta_i^{(1)}, \eta_i^{(2)}, \eta_i^{(3)}, \eta_i^{(4)}\right), \qquad \eta_i^{(1)} < \eta_i^{(2)} \leq \eta_i^{(3)} < \eta_i^{(4)}, \tag{7}$$

(if $\eta_i^{(2)} = \eta_i^{(3)}$ then η_i is called triangular fuzzy quantity) such that the membership function μ_i of η_i is defined by

$$\begin{aligned}
\mu_i(x) &= 0 && \text{if } x \leq \eta_i^{(1)}, \\
&= \frac{x - \eta_i^{(1)}}{\eta_i^{(2)} - \eta_i^{(1)}} && \text{if } \eta_i^{(1)} < x < \eta_i^{(2)}, \\
&= 1 && \text{if } \eta_i^{(2)} \leq x \leq \eta_i^{(3)}, \\
&= \frac{\eta_i^{(4)} - x}{\eta_i^{(4)} - \eta_i^{(3)}} && \text{if } \eta_i^{(3)} < x < \eta_i^{(4)}, \\
&= 0 && \text{if } x \geq \eta_i^{(4)}.
\end{aligned}$$

The procedure of construction of fuzzy quantities $\boldsymbol{a}$ and $\boldsymbol{b}$ using these trapezoidal inputs is in certain sense rather similar to the one used in the previous section but without its subjectivity. Having four sequences $(\eta_1^{(1)}, \ldots, \eta_n^{(1)})$, $(\eta_1^{(2)}, \ldots, \eta_n^{(2)})$, $(\eta_1^{(3)}, \ldots, \eta_n^{(3)})$, $(\eta_1^{(4)}, \ldots, \eta_n^{(4)})$ of crisp numbers, we can use them in combination with the crisp sequence $(\xi_1, \ldots, \xi_n)$, and to compute by means of (2) quadruples of coefficients

$$\left(a^{(1)}, a^{(2)}, a^{(3)}, a^{(4)}\right), \quad \left(b^{(1)}, b^{(2)}, b^{(3)}, b^{(4)}\right).$$

Please, note that these quadruples need not fulfil the inequalities in (7) and in this sense they need not be definitoric quadruples of somehow trapesoidal fuzzy quantities $\boldsymbol{a}$ and $\boldsymbol{b}$ in (1). Nevertheless, it is not difficult to verify that for any $\xi \in R$, fulfilling

$$\min(\xi_i : i = 1, \ldots, n) \leq \xi \leq \max(\xi_i : i = 1, \ldots, n) \tag{8}$$

the inequalities

$$a^{(1)} + b^{(1)} \cdot \xi \leq a^{(2)} + b^{(2)} \cdot \xi \leq a^{(3)} + b^{(3)} \cdot \xi \leq a^{(4)} + b^{(4)} \cdot \xi$$

hold. In this sense the segments of straight lines

$$a^{(k)} + b^{(k)} \cdot \xi, \quad k = 1, 2, 3, 4,$$

for ξ fulfilling (8) determine a trapezoidal area where for any ξ respecting (8), the quadruple

$$\left(a^{(1)} + b^{(1)} \cdot \xi,\, a^{(2)} + b^{(2)} \cdot \xi,\, a^{(3)} + b^{(3)} \cdot \xi,\, a^{(4)} + b^{(4)} \cdot \xi,\right) = \left(\eta^{(1)}, \eta^{(2)}, \eta^{(3)}, \eta^{(4)}\right)$$

defines a trapezoidal fuzzy quantity η which represents the probable and possible values of the random and vaguely known variable H.

5 Conclusive Remarks

In the previous sections we have discussed three principal approaches to the regression-like relations between the sequences of quantities where one of them consists of crisp values, the second one contains vague (fuzzy) quantities, elements of both sequences can be randomly generated.

It is useful to note that the computational methods treated in Section 2 can be used even if both sequences $(\xi_1, \ldots, \xi_n)$, $(\eta_1, \ldots, \eta_n)$ are sequences of randomly generated fuzzy quantities or, in an alternative formulation, sequences of vaguely known realizations of random variables (see [4]). The remaining two methods, summarized in Sections 3 and 4, are conditioned by the assumption that the values $\xi_1, \ldots, \xi_n$ are crisp, and their extension to the more general case would be difficult.

The above approaches differ also in the proportion between formal mathematical exactness and the relatively free subjective evaluation of the modelled situation. The methods shown in Section 2 display relatively high degree of respect to the mathematical formalism. Especially the approach in Subsection 2.1, based on the application of the extension principle, almost exclusively uses formal exact procedures. The application of generated fuzzy quantities already includes certain degree of subjectivism – at least in the selection of the shape functions and logically-semantically motivated procedure of their combination during the construction of φ_a and φ_b. The approach presented in Section 4 is also in its nature relatively objective and exact, using the classical regression formula even if in less standard way. Finally, methods mentioned in Section 3 are essentially connected with subjectivity and (qualified) opinion of the expert evaluating the situation.

Nevertheless, the present methods in their completeness allow to find a relatively balanced view of the regression in the environment of random-fuzzy phenomena.

References

[1] D. Dubois, H. Prade: Fuzzy numbers: An overview. In: Analysis of Fuzzy Information (J. C Bezdek (Ed.). CRC–Press, Boca Raton 1988, pp. 3–39.

[2] J. Kacprzyk, M. Fedrizzi (eds.): Fuzzy Regression Analysis. Physica Verlag, Heidelberg 1992.

[3] M. Mareš: Fuzzification of the regression model. In: Information Asymmetries on Capital Market (M. Vošvrda, ed.), ÚTIA, Praha 1998, XV.

[4] M. Mareš: Computation Over Fuzzy Quantities. CRC–Press, Boca Raton 1994.

[5] M. Mareš, R. Mesiar: Computation over verbal variables. In: Computing With Words in Systems Analysis (J. Kacprzyk, L. A. Zadeh, eds.). Physica Verlag. In print.

[6] M. Mareš, R. Mesiar: Vagueness of verbal variable. In: Soft Computing in Financial Engineering (J. Kacprzyk, R. Ribeiro, R. R. Yager, H.-J. Zimmermann, eds.). Physica Verlag, Heidelberg 1999, 3–20.