Application of the theory of fuzzy sets for a qualitative analysis of mathematical models

D T Muhamediyeva
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi,
Amir Temur street, 108, 100200, Tashkent, Uzbekistan

Abstract. The article discusses the estimation of fuzzy regression parameters when specifying various membership functions, since the effectiveness of the least squares method dramatically reduces when a number of prerequisites for its use are violated, in particular, when the sample being processed contains observations that are poorly consistent with the others. In these cases, one can use estimation methods that are less sensitive than the least squares method to specification errors and allow one to obtain so-called fuzzy estimates. Among such methods is the method of smallest modules, the implementation of which leads to the problem of linear programming.

1. Introduction
When constructing models of a statistical type, a situation often arises when the data to be processed at the stage of parameter estimation have gaps in individual elements. This may be caused by a temporary malfunction (malfunction) of the measuring equipment during the removal of various technical characteristics or by the negligence of the statistical services in fixing the reporting indicators.

The presence of gaps in the data when evaluating the parameters of regression equations means that the observation matrix of independent variables X has elements that are not defined in any way. Some researchers propose in such cases to exclude the relevant observations from consideration - or to process the variables and the remaining “complete” sample in a traditional way. However, under the conditions of the so-called small or medium samples (namely, they most often have to be dealt with), such artificial “truncation” of the sample is too wasteful from the point of view of losing useful information.

To date, a large number of methods for filling data gaps have been developed. The most famous of them should be considered the method, the main components, the "fixed point" [1-2], filling with unconditional and conditional means [3], local filling [2], etc. Some of them are probabilistic, others suggest the use of a nonparametric approach. As a rule, the methods for filling data gaps are based on various heuristic procedures, often very complex and iterative, which, combined with the known difficulty in introducing a formalized measure of the quality of recovery of missing elements, makes these methods easily vulnerable to reasonable criticism. Among the criticized include [2] mainly the following disadvantages of "complexing" procedures:

- Calculation of the parameters of the algorithm for filling gaps according to the data present, which makes the relationship between observations;
- deterioration in the quality of estimates with an increase in the share of omissions;
- the impossibility of conducting a rigorous study of the properties of the algorithm;
- distortion of the nature of the data and the nature of the conclusions.
While acknowledging in many respects the validity of this criticism, one cannot but admit that the problem of filling in the gaps in the data is very complex and hardly has a theoretically correct and only right solution that does not have any subjective elements in its search. Therefore, it is very likely that each new approach to solving this problem will differ from the ones already known by new heuristics and raise doubts in this connection about their validity, consistency, sufficiency and correct application [4-5].

A fundamentally different approach from the above methods consists in processing data with gaps based on methods not directly related to filling the latter.

Indeed, when constructing statistical dependences on the basis of a sample with missing elements, it is important not only to correctly fill in the gaps, but to estimate the parameters as accurately as possible, taking into account all the information, including those containing observation gaps. The following is a possible solution to this problem.bank” of replaced individuals [6-7].

2. Data processing with gaps.
We introduce some additional notation. Denote by K the set (1, 2, ..., n) of observation numbers. We divide the entire available sample of the researcher (X, y) into two parts – the complete (X1, y1) with the observations (Xk, yk), k ∈ K1, and the incomplete (X2, y2) with the observations (Xj, yj), l ∈ K2, where K1 is the set of numbers of complete observations, K2 is the set of observation numbers, containing omissions K = K1 ∩ K2, K1 ∩ K2 = Ø; Xk, – the k-th row of the matrix X. We will assume that only the matrix X contains the missing elements.

Let the researcher have at his disposal h ways (as simple as possible) to fill in the gaps in the data. This means that for any gap in the matrix X in the k-th row at the i-th place, h of its possible filling values xki1, xki2,..., xkih can be indicated. We find among them the minimum xki and maximum xki values. Thus, each gap in the matrix X is replaced by an interval, and this matrix is replaced by an interval [X−, X+].

Those elements that were not gaps in the original formulation have matching upper and lower boundaries of the corresponding intervals. Replacing gaps with intervals may not be necessary using the method mentioned above. This can be done on the basis of a meaningful analysis of the data with the help of expert knowledge or “soft” procedures to ensure that the missing value falls into the designated interval, which can be quite wide. Such filling in the gaps is based on the idea of giving part of the sample, including them, an auxiliary, corrective role in relation to the complete information.

After replacing the observation matrix of independent variables X with gaps with an interval one, the problem of estimating linear regression parameters reduces to the problem of constructing a set of estimates.

Because in this case y− = y+ = y. The following conditions are met:

\[ X_1^+ a - X_1^- a^2 \leq y, \]
\[ X_1^- a - X_1^+ a^2 \geq y, \]
given that, the equality \( X_1^- = X_1^+ = X_1 \) must hold \( X_1 a = y \). But if all the vectors \( (x_{k1}, x_{k2},..., x_{km}, y_k) \), \( k \in K_1 \) are linearly independent and \( \dim K > m \), then such that

\[ \sum_{i=1}^{m} a_i x_{ki} \neq y_k. \]

Therefore, one should look for an estimate of these sets. In this case, it is necessary to take into account the above-mentioned consideration that the complete part of the sample is the main part, and the incomplete part is the auxiliary, corrective one.

Let us consider in detail all the cases that arise when solving the problem of estimating the parameters of a linear equation based on a sample with gaps replaced by intervals.
We assume that rank $X \geq m$. Based on this assumption, it is possible to estimate the parameters of linear regression using a complete sample. We denote by $\hat{a}$ the estimate defined on the sample $(X, y)$, and by $\overline{R}(X, y)$ the set of parameter estimates corresponding to the interval sample $([X^-, X^+], y)$. By $J_1(\hat{a})$ we denote the value of the loss function for estimating $a$, i.e.

$$J_1(\hat{a}) = \sum_{k \in K_1} \left| y_k - \sum_{i=1}^m \hat{a}_k x_{ki} \right|.$$ 

Consider the possibility of adjusting the score according to the results of its correlation with the set $\overline{R}(X, y)$.

a) $\overline{R}(X, y)$ -empty set.

This situation is especially common in practice. Evaluation is a solution to the following linear programming problem:

$$X_k a + r_k - s_k = y_k, \quad k \in K_1,$$

$$\overline{X}^- a - u^l \leq y_l,$$

$$\overline{X}^+ a + v^l \geq y_l, \quad l \in K_2,$$

$$a_i \geq 0 \quad \text{if} \quad \hat{a}_i \geq 0,$$

$$a_i \leq 0 \quad \text{if} \quad \hat{a}_i \leq 0, \quad i = 1, m,$$

$$r \geq 0, \quad s \geq 0, \quad u_{kl} \geq 0, \quad v_{kl} \geq 0,$$

$$\sum_{k \in K_1} (r_k + s_k) + p \sum_{l \in K_2} (u^l + v^l) \rightarrow \min, \quad 0 < p \leq 1,$$

where the matrices $X^-_2$ and $X^+_2$ are formed by analogy with $X^-$ and $X^+$. The value of $p$ is less, the less the researcher is sure that the gaps in the given intervals are correctly filled.

b) $\hat{a} \notin \overline{R}(X, y)$ -not an empty set.

This case indicates a discrepancy found in sample $(X, y)$. Estimates of the interval sample $(X, y)$, which, if there are sufficiently wide intervals in place of the gaps, are of particular interest. To assess the degree of this non-compliance, the following problem should be solved:

$$\min_{a \in \overline{R}(X, y)} \rho(\hat{a}, a),$$

which, if we take the sum of the absolute deviations of the coordinates of the vectors $a$ and $\hat{a}$ as the metric $\rho$, reduces to a linear programming problem of the form

$$\overline{X}^- a \leq y_l,$$

$$\overline{X}^+ a \geq y_l, \quad l \in K_2,$$

$$a + u - v = \hat{a},$$

$$a_i \geq 0 \quad \text{if} \quad \hat{a}_i \geq 0,$$

$$a_i \leq 0 \quad \text{if} \quad \hat{a}_i \leq 0, \quad i = 1, m,$$

$$u \geq 0, \quad v \geq 0,$$

$$J(u, v) = \sum_{i=1}^m (u_i + v_i) \rightarrow \min.$$
If the vector \((a^*, u^*, v^*)\) is a solution to this problem, then the quantity \(J(u^*, v^*)\) is an estimate of the distance from \(\hat{a}\) to the set \(\tilde{R}(X_2, y_2)\). In order to give this distance a relative character that does not depend on the scale of the individual components of the vectors \(a^*\) and \(\hat{a}\), it is necessary, by analogy with the accuracy of approximation \(\lambda(\tilde{R})\), to calculate the value
\[
\lambda(a^*, \hat{a}) = \frac{1}{m} \sum_{i=1}^{m} \left( \frac{u_i + v_i}{|\hat{a}_i|} \right) 100\% .
\]

If, according to the researcher, the relative value of the distance \(\lambda(a^*, \hat{a})\) is too large, then this means that either the gaps in these intervals were filled with an error, or the complete and incomplete parts of the sample are mutually contradictory. In the first case, it is necessary to correct the intervals filling in the gaps, and in the second, it is necessary to find a compromise between the parameters \(a^*\) and \(\hat{a}\). If \(\lambda(a^*, \hat{a})\) indicates the proximity of the found estimates, any of them or their half-sum can be taken as the final regression estimate.

When searching for a compromise between \(a^*\) and \(\hat{a}\), the assessment should be done as follows. Let the researcher, in order to achieve the necessary compromise, be able to assign a value \(\Delta J_i(\hat{a})\) by which it is possible to increase the value of the loss function \(J_i(\hat{a})\) on the sample \((X_1, y_1)\) to reduce the distance \(\rho(\hat{a}, R(X_2, y_2))\). Then the desired compromise estimate \(\tilde{a}\) will be the solution of the following linear programming problem:

\[
X_k a + r_k - s_k = y_k, \quad k \in K_2,
\]
\[
\overline{X}_i a - u_i^l \leq y_i,
\]
\[
\overline{X}_i a + v_i^l \geq y_i, \quad l \in K_2,
\]
\[
\sum_{k \in K_1} (r_k + s_k) \leq J_i(\hat{a}) + \Delta J_i(\hat{a}),
\]
\[
a_i \geq 0 \text{ if } \hat{a}_i \geq 0,
\]
\[
a_i \leq 0 \text{ if } \hat{a}_i \leq 0, \quad i = 1, m,
\]
\[
r \geq 0, \quad s \geq 0, \quad u_i \geq 0, \quad v_i \geq 0,
\]
\[
\delta \sum_{k \in K_1} (r_k + s_k) + \sum_{l \in K_2} (u_i^l + v_i^l) \rightarrow \min.
\]

The presence in the objective function of the first term with a small positive factor provides fulfillment of the necessary condition \((r,s) = 0\).

If the researcher believes that the significance of the complete and incomplete samples in solving the problem of estimating the parameters of linear regression is approximately equal, for \(J_i(\hat{a})\) we should take the value of
\[
\frac{J_i(a^*) - J_i(\hat{a})}{a}
\]
where
\[
J_i(a^*) = \sum_{k \in K_i} y_k - \sum_{i=1}^{m} a_i^* x_{ki}.
\]

c) \(\hat{a} \in \tilde{R}(X_2, y_2)\) - not an empty set.

If the estimate \(a_1\), determined from the point sample \((X_1, y_1)\), is contained in the set of valid estimates for the interval sample \((X_2, y_2)\), we can conclude that the complete and incomplete
information is consistent. The latter in such a situation does not specify the estimate of $\hat{a}$, which can be taken as the desired one.

It seems important to once again emphasize the following circumstance. All of the situations listed above, which differ in the result of correlating the estimate of $\hat{a}$ with the set of $R$($X_2$, $y_2$) and realizing their tasks, are used here not to fill in the gaps in the data, but to evaluate the parameters - linear regression based on a combination of complete and incomplete information. Moreover, the proposed approach easily extends to the case of gaps in observations, both for independent and dependent variables.

If we remove the assumption about the rank of the matrix $X_1$, then in this case the whole sample ($X_1$, $[X_2^+,X_2^-]$, $y_1$, $[y_2^+,y_2^-]$) should be treated as an interval one.

Forecasting is a scientifically based assessment of future conditions of the object of interest to us [8-10].

Another area of research involves the use of the apparatus of fuzzy sets to process expert judgments about the values of the predicted indicator and the factors influencing it. This approach is very promising; however, it is focused on the case of static data that is uncertain in a special sense.

The present work also considers an approach to the joint use of statistical and expert information when evaluating model parameters. At the same time, however, it was possible to significantly expand the range of possible expert statements, limiting itself to solving arising problems with a rather simple linear programming apparatus. In addition, the scope of application of such statements in relation to other problems has also been expanded

$$\tilde{y} = \sum_i (\tilde{a}_i \tilde{x}_i + c_i [\tilde{x}_i]).$$

To find the parameters of the bell-shaped membership function, it is necessary to solve the following linear programming problem:

$$\begin{align*}
&\min \sum c_ik + c_2k \sqrt{\frac{1-\alpha}{\alpha}}, \\
&\quad \sum c_ik + c_2k \sqrt{\frac{1-\alpha}{\alpha}}, \\
&\quad y_k \leq \tilde{a}_k + c_2k \sqrt{\frac{1-\alpha}{\alpha}}, \\
&\quad y_k \geq \tilde{a}_k + c_1k \sqrt{\frac{1-\alpha}{\alpha}}.
\end{align*}$$

After finding the parameters $\tilde{a}_k$, $c_1k$, $c_2k$ the type of fuzzy logistic model is determined.

3. Computational experiment.

As an example, we will use various tasks located on the website http://www.ics.uci.edu/~mlearn/databases/. The parameters of the indicated model problem are given below (see Table 1).

| Training Sample Name | Characters numbers | Number of objects | Number of classes |
|----------------------|--------------------|------------------|------------------|
| IRIS                 | 4                  | 150              | 3                |
| HEART                | 13                 | 270              | 2                |

Table 1. Model task parameters.
Based on the proposed method, model problems were solved using various membership functions and the results obtained were compared (see Table 2).

**Table 2.** Comparative results of the proposed method in the case of various membership functions

| Methods  | Training Sample Name | Triangle | Gauss | Bell |
|----------|-----------------------|----------|-------|------|
| IRIS     | 100 %                 | 100 %    | 100 % |
| HEART    | 86 %                  | 86 %     | 86 %  |
| DIABETES | 82.5 %                | 82.5 %   | 82.5 %|

4. **Conclusion.**

Evaluation of the parameters of the regression equations included in the composition of a particular statistical model implies solving a problem or a system for evaluating the parameters of the estimation equations. If the system is not recursive, then the estimation of the parameters of the general system differs from the estimates calculated for each regression. The intermediate stage of structural assessment is the attraction of additional a priori assessment. Statistical models along with linear equations include non-linear regression equations and have a large dimension, and this in turn causes a number of problems in solving the problems of determining the model parameters as well as structural estimation methods.

We note two more main cases. When building a model, it becomes necessary to evaluate the parameters of the regression components.

The above considerations show the appropriateness of developing statistical models in the form of recursive systems. This proves the mathematical correctness of the estimation of the parameters of each equation separately, controls the preservation of the signs of the coefficients, which is very difficult from the point of view of implementation in the structural assessment.

**References**

[1] S A Ayvazyan, V M Buchstaber, I S Enyukov and L D Meshalkin 2002 *Applied statistics. Classification and reduction of dimension. Finance and statistics*, 607

[2] A N Gorban, B Kegl, D Wunsch and Zinovyev A Y 2008 *Principal Manifolds for Data Visualisation and Dimension Reduction* 58 340

[3] K Liu, B Wei and B Liu 2012 *Analysis model of slope deformation time series based on the genetic-adaptive neuron-fuzzy inference system* 36 56

[4] L Zadeh, N Neshat, A Kazemi and M. Saberi 2012 Predictive control of drying process using an adaptive neuro-fuzzy and partial least squares approach *International Journal of Advanced Manufacturing Technology* 58 585

[5] Norman Draper, Harry Smith 1981 *Applied Regression Analysis* (Wiley)

[6] A I Kobzar 2006 *Prikladnaya matematicheskaya statistika.*
[7] D. Rutkovskaya, M. Pilinsky and L. Rutkovsky 2004 *Neural networks, genetic algorithms and fuzzy systems* 452

[8] Muhamediyeva D.T. 2020 *Fuzzy cultural algorithm for solving optimization problems* 1441

[9] Muhamediyeva D.T. 2020 *Particle swarm method for solving the global optimization problem using the equilibrium coefficient* 1441

[10] D. Sotvoldiev, D.T. Muhamediyeva and Z. Juraev 2020 Deep learning neural networks in fuzzy modeling *Journal of Physics: Conference Series* 1441