Quintile multiple regression with fuzzy coefficients and initial Z-information

Olga Poleschchuk

1 Bauman Moscow State Technical University, 5/1, 2ya Baumanskaya st., 105005 Moscow, Russia

Abstract. The goal of the paper is to develop a regression model under the initial Z-information based on an alternative method to the least squares method, and free from the assumptions regarding probability distributions of initial data. Formalization of input and output information is carried out on the basis of Z-numbers and linguistic variables, followed by the construction of a multidimensional quintile regression model with fuzzy coefficients. The optimization function is defined as the sum of the loss functions for the differences between the weighted output fuzzy numbers and the weighted model fuzzy numbers. To determine the parameters of the unknown regression coefficients, a linear programming problem is solved to find the minimum of the optimization function. The developed Z-regression is free from the shortcomings of existing models and provides new opportunities for solving tasks in problem areas with the active participation of experts, taking into account the reliability of information received from them.

1 Introduction

With the development of the fuzzy set theory, it necessary became to process fuzzy information based on the apparatus of regression analysis, which became the reason for the development of methods of fuzzy regression analysis [1].

When constructing fuzzy regression models, various optimization criteria and methods were used, but the most commonly used was the well-known least-squares method [2, 3]. One of the disadvantages of this method is the sensitivity to outliers, since their squares are present, which significantly increase the total sum. In this regard, other approaches and methods were sought for constructing fuzzy regressions.

An alternative method for constructing fuzzy regression models, which is more robust to outliers than the least squares method, was proposed in [4]. Moreover, the examples given demonstrate its effectiveness.

However, after the introduction of the definition of Z-number [5], all the developed methods of fuzzy regression analysis turned out to be unable to analyze Z-information (information with Z-numbers), and therefore it necessary became to develop such methods that would provide such an opportunity.

* Corresponding author: olga.m.pol@yandex.ru

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).
To begin with, it was necessary to develop operations on Z-numbers. The basis of such operations is developed in [6], and further successful developments are presented in papers [7-11].

Distances between Z-numbers are determined in [12-14] and used for a number of practical tasks [15, 16].

In order to construct dependencies and predict output data under Z-information, methods of regression analysis are needed. The first regression model under Z-information is developed in [17] based on operations with fuzzy numbers and probability distributions, followed by the use of the Jaccard measure to formulate the optimization problem. The coefficients of the regression model are ordinary numbers.

In [18], a linear Z-regression with fuzzy coefficients is developed. For construction, an aggregating segment of a Z-number is determined in the form of weighted segments of the product of both of its components. The optimization function is defined as the sum of the squared distances between the aggregating segments of the initial output Z-numbers and the model Z-numbers, as well as between the aggregating segments of the first components of the initial output Z-numbers and the first components of the model Z-numbers.

In [19], a nonlinear Z-regression with crisp coefficients is developed. Since, in fact, both of the [18, 19] regressions are based on the least squares method, the drawback of which was mentioned above, the goal of the paper is to develop a regression model under the initial Z-information based on an alternative method to the least squares method, and free from the assumptions of paper [17] regarding probability distributions of initial data.

The second section contains the necessary definitions. The third section contains a regression model under Z-information. The fourth section contains conclusions on the paper.

2 Basic concepts and definitions

Z-number is defined in [5] as a pair \( Z = (\mathcal{C}, \mathcal{R}) \), where \( \mathcal{C}, \mathcal{R} \) are fuzzy numbers and \( \mathcal{R} \) is reliability of \( \mathcal{C} \).

A linguistic variable \( X \) with a term-set \( T(X) = \{X_l, l = \overline{1,m}\} \) is called a collection \( \{X,T(X),U,V,S\} \). Value names \( X_l, l = \overline{1,m} \) are determined by the rule \( V \) and the corresponding fuzzy sets of \( U \) for fuzzy variables \( \tilde{X}_l, l = \overline{1,m} \) are determined by the rule \( S \) [20].

Consider the LR fuzzy number \( \tilde{A} = (a_1, a_2, a_L, a_R)_{LR} : \)

\[
\tilde{A} = \begin{cases} 
L \left( \frac{a_1 - x}{a_L} \right), & x \leq a_1 \\
1, & a_1 \leq x \leq a_2 \\
R \left( \frac{x - a_2}{a_R} \right), & x \geq a_2 
\end{cases}
\]

where \( L(x), R(x) \) are strictly decreasing functions from \( R^+ \) to \([0,1]\), and \( L(0) = R(0) = 1 \).

Then \( \alpha \)-cut of \( \tilde{A} \) is \( A_\alpha \) such that:

\[
A_\alpha = \{ x \in R : \mu_A(x) \geq \alpha \} = [a - L^{-1}(\alpha)a_L, a + R^{-1}(\alpha)a_R], \alpha \in [0,1].
\]

In [3], based on \( \alpha \)-cuts a weighted segment \([\beta_1, \beta_2] \) for the LR fuzzy number \( \tilde{A} = (a_1, a_2, a_L, a_R)_{LR} \) is determined:
\[
\beta_1 = \int_{0}^{1} \frac{2a_1 - L^{-1}(\alpha)a_l}{2} 2\alpha d\alpha = a_1 - \frac{1}{6} a_l, \\
\beta_2 = \int_{0}^{1} \frac{2a_2 + R^{-1}(\alpha)a_R}{2} 2\alpha d\alpha = a_2 + \frac{1}{6} a_R.
\]

If \(\hat{A} = (a_1, a_2, a_L, a_R)_{LR}\) is trapezoidal fuzzy number \(\bar{A} = (a_1, a_2, a_L, a_R)\), then (1) looks like this:

\[
\beta_1 = \int_{0}^{1} \frac{2a_1 - (1 - \alpha)a_l}{2} 2\alpha d\alpha = a_1 - \frac{1}{6} a_l, \\
\beta_2 = \int_{0}^{1} \frac{2a_2 - (1 - \alpha)a_R}{2} 2\alpha d\alpha = a_2 - \frac{1}{6} a_R.
\]

We define a weighted point \(\beta\) for a fuzzy number \(\bar{A} = (a_1, a_2, a_L, a_R)\) according to (2) as follows:

\[
\beta = \frac{\beta_1 + \beta_2}{2} = \frac{a_1 + a_2 + a_R - a_L}{12}.
\]

In [4], the loss function \(\rho_r(\bar{A})\) for the LR fuzzy number \(\bar{A} = (a_1, a_2, a_L, a_R)_{LR}\) is defined:

\[
\rho_r(\bar{A}) = \frac{1}{3} [\rho_r(a) + \rho_r(a - l a_l) + \rho_r(a + r a_R)],
\]

where

\[
l = \int_{0}^{1} L^{-1}(\alpha) d\alpha, r = \int_{0}^{1} R^{-1}(\alpha) d\alpha, \rho_r(a) = a(\tau - I_{\{\tau < 0\}})
\]

is the classical quantile loss function, \(I_{\{\tau < 0\}}\) is the indicator function of the set \(\{\tau < 0\}\).

### 3 Problem formulation and solution

Let \(Z_{ij} = (X_{ij}, R_{ij}), i = 1, m, j = 1, n\) and \(Z_j = (\bar{Y}_j, \bar{R}_j), j = 1, n\) are the input and output Z-numbers accordingly, \(X_{ij} = (x_{ij}^1, x_{ij}^2, x_{ij}^3, x_{ij}^4), i = 1, m, j = 1, n\), \(Y_j = (y_j^1, y_j^2, y_j^3, y_j^4), j = 1, n\). To evaluate the reliability of initial information the scale with levels \(R_v, V = 1, V\) is used. We construct a linguistic variable with name “Reliability” and terms \(R_v, V = 1, V\). Denote the fuzzy numbers that formalize the terms \(R_v, V = 1, V\) through \(R_v, V = 1, V\). In [21, 22], methods for constructing of linguistic variables are described in detail. Each of the fuzzy numbers \(\bar{R}_{ij} = (r_{ij}^1, r_{ij}^2, r_{ij}^3, r_{ij}^4), \bar{R}_j = (r_j^1, r_j^2, r_j^3, r_j^4), i = 1, m, j = 1, n\) equals to one of fuzzy numbers \(R, V = 1, V\).

Let construct a regression model in the form:

\[
Z = \bar{a}_1 Z_1 + \bar{a}_2 Z_2 + \cdots + \bar{a}_m Z_m,
\]

where \(\bar{a}_i = (a_i, a_i^1, a_i^2), i = 1, m\) are triangular fuzzy numbers.

The regression model under Z-information is shown in Fig. 1.
We define a weighted fuzzy number $\tilde{A}$ for the Z-number $Z = (\bar{A}, \bar{R})$, $\bar{A} = (a_1, a_2, a_L, a_R)$, $\bar{R} = (r_1, r_2, r_L, r_R)$ as follows: $\tilde{A} = (\gamma a_1, \gamma a_2, \gamma a_L, \gamma a_R)$, where
\[
\gamma = \frac{r_1 + r_2}{2} + \frac{r_R - r_L}{12}, \gamma > 0
\]
is a weighted point for the fuzzy number $\bar{R} = (r_1, r_2, r_L, r_R)$.

Let us define, based on [4], the loss function for the number $\hat{A} = (a_1, a_2, a_L, a_R)$ as follows:
\[
\rho_\tau(\hat{A}) = \frac{1}{3} \left[ \frac{\rho_\tau(a_1 + a_2)}{2} + \rho_\tau\left(a_1 - \frac{a_L}{2}\right) + \rho_\tau\left(a_2 + \frac{a_R}{2}\right) \right].
\]

Based on (4) and the subtraction operation for fuzzy numbers $\bar{A} = (a_1, a_2, a_L, a_R), \bar{B} = (b_1, b_2, b_L, b_R)$, we define the loss function for the difference of fuzzy numbers:
\[
\rho_\tau(\bar{A} - \bar{B}) = \frac{1}{3} \left[ \frac{\rho_\tau(a_1 - b_2)}{2} + \rho_\tau\left(a_2 - b_1\right) + \rho_\tau\left(a_1 - b_2 - \frac{a_L}{2} - \frac{b_R}{2}\right) + \rho_\tau\left(a_2 - b_1 + \frac{a_R}{2} + \frac{b_L}{2}\right) \right].
\]

Define weighted input fuzzy numbers $\tilde{X}_{ij}, i = 1, m, j = 1, n$ for input Z-numbers $Z_{ij} = (\bar{X}_{ij}, \bar{R}_{ij}), i = 1, m, j = 1, n$, $\bar{X}_{ij} = (x_{ij}^1, x_{ij}^2, x_{ij}^L, x_{ij}^R), i = 1, m, j = 1, n$, $\bar{R}_{ij} = (r_{ij}^1, r_{ij}^2, r_{ij}^L, r_{ij}^R), i = 1, m, j = 1, n$, $\tilde{X}_{ij} = (\gamma_{ij} x_{ij}^1, \gamma_{ij} x_{ij}^2, \gamma_{ij} x_{ij}^L, \gamma_{ij} x_{ij}^R)$. 

Predictions of Y with some level of reliability

Fig. 1. $Z$–regression.
Define weighted output fuzzy numbers $\tilde{Y}_{i,j} = \frac{1}{n_i}$ for the output Z-numbers

$$Z_j = (\tilde{Y}_j, R_j), j = \frac{1}{n}, \quad \tilde{Y}_{j} = (y_j y^1_j, y_j y^2_j, y_j y^3_j), \quad y_j = \frac{r_j^1 + r_j^2}{2} + \frac{r_j^R - r_j^L}{12}, \quad j = \frac{1}{n}.$$ 

Define weighted model fuzzy numbers $\tilde{Y}_{M,j} = \frac{1}{n}$ for the model output Z-numbers

$$Z_j = \sum_{i=1}^{m} \alpha_i Z_{ij}, j = \frac{1}{n}, \quad \tilde{Y}_{M,j} = \left( \sum_{i=1}^{m} \beta_i y_{ij} x_{ij}^1, \sum_{i=1}^{m} \beta_i y_{ij} x_{ij}^2, \sum_{i=1}^{m} \beta_i y_{ij} x_{ij}^3 \right).$$

$\beta_i > 0, i = \frac{1}{m}, m = \frac{1}{n}, \beta_i = a_i + \frac{a_i^R - a_i^L}{12},$ \quad $\tilde{Y}_{M,j} = \left( \sum_{i=1}^{m} \beta_i y_{ij} x_{ij}^1, \sum_{i=1}^{m} \beta_i y_{ij} x_{ij}^2, \sum_{i=1}^{m} \beta_i y_{ij} x_{ij}^3 \right), \beta_i < 0, i = \frac{1}{m}, m = \frac{1}{n}.$

We get,

$$\tilde{Y}_{M,j} = \left( \sum_{i=1}^{m} \beta_i y_{ij} x_{ij}^{p_i}, \sum_{i=1}^{m} \beta_i y_{ij} x_{ij}^{q_i}, \sum_{i=1}^{m} \beta_i y_{ij} x_{ij}^{M_{p_i}}, \sum_{i=1}^{m} \beta_i y_{ij} x_{ij}^{M_{q_i}} \right), i = \frac{1}{m}, n = \frac{1}{n},$$

where

$$p_i = \begin{cases} 1, & \beta_i > 0 \quad q_i = \begin{cases} 1, & \beta_i < 0 \quad M_{p_i} = \begin{cases} L, & p_i = 1 \quad M_{q_i} = \begin{cases} L, & q_i = 1 \quad R, & p_i = 2 \quad R, & q_i = 2 \end{cases} \end{cases} \end{cases}.$$ 

Define the loss function for the difference between the initial output Z-number and the model output Z-number as the loss function between the weighted output fuzzy number and the weighted model fuzzy number:

$$\rho_{\tau} = \left( \tilde{Y}_j - \tilde{Y}_{M,j} \right) = \frac{1}{3} \rho_{\tau}(y_j y^1_j - \overline{y}_{ij} x_{ij}^1 + p_i) + \rho_{\tau}(y_j y^2_j - \overline{y}_{ij} x_{ij}^2 + q_i) + \rho_{\tau}(y_j y^3_j - \overline{y}_{ij} x_{ij}^3 + M_{p_i}).$$

$$\rho_{\tau} \left( \rho_{\tau} - \tilde{Y}_{M,j} \right) = \frac{1}{3} \rho_{\tau}(y_j y^1_j - \overline{y}_{ij} x_{ij}^1 + p_i) + \rho_{\tau}(y_j y^2_j - \overline{y}_{ij} x_{ij}^2 + q_i) + \rho_{\tau}(y_j y^3_j - \overline{y}_{ij} x_{ij}^3 + M_{p_i}).$$

We introduce the following notation:

$$\eta_j = y_j y^1_j - \sum_{i=1}^{m} \beta_i y_{ij} x_{ij}^{p_i}, \delta_j = y_j y^2_j - \sum_{i=1}^{m} \beta_i y_{ij} x_{ij}^{q_i},$$

$$\xi_j = y_j y^3_j - \sum_{i=1}^{m} \beta_i y_{ij} x_{ij}^{M_{p_i}} - \frac{\sum_{i=1}^{m} \beta_i y_{ij} x_{ij}^{M_{q_i}}}{2},$$

$$\zeta_j = y_j y^5_j - \sum_{i=1}^{m} \beta_i y_{ij} x_{ij}^{p_i} + \frac{y_j y^R_j}{2} - \frac{\sum_{i=1}^{m} \beta_i y_{ij} x_{ij}^{M_{p_i}}}{2}.$$

Then:

$$\rho_{\tau} \left( \tilde{Y}_j - \tilde{Y}_{M,j} \right) = \frac{1}{3} \rho_{\tau}(\eta_j) + \rho_{\tau}(\zeta_j) + \rho_{\tau}(\xi_j) + \rho_{\tau}(\delta_j).$$

Using definition (4) and definition of the classical quantile loss function, we obtain:

$$\rho_{\tau} \left( \tilde{Y}_j - \tilde{Y}_{M,j} \right) = \frac{1}{6} \left( \tau \eta_j^+ + (1 - \tau) \eta_j^- + \tau \delta_j^+ + (1 - \tau) \delta_j^- \right) + \frac{1}{3} \left( \tau \xi_j^+ + (1 - \tau) \xi_j^- + \tau \zeta_j^+ + (1 - \tau) \zeta_j^- \right).$$
where \( \eta_j^+ = \max(y_j, 0), \eta_j^- = -\min(y_j, 0), \eta_j = \eta_j^+ - \eta_j^- \),
\( \delta_j^+ = \max(\delta_j, 0), \delta_j^- = -\min(\delta_j, 0), \delta_j = \delta_j^+ - \delta_j^- \),
\( \xi_j^+ = \max(\xi_j, 0), \xi_j^- = -\min(\xi_j, 0), \xi_j = \xi_j^+ - \xi_j^- \),
\( \zeta_j^+ = \max(\zeta_j, 0), \zeta_j^- = -\min(\zeta_j, 0), \zeta_j = \zeta_j^+ - \zeta_j^- \).

The optimization problem for finding the unknown coefficients
\( \bar{a}_i = (a_i, a_i^0, a_i^R), i = 1, \ldots, m \) is translated into the following linear programming problem:

\[
F(a_i, a_i^0, a_i^R) = \sum_{j=1}^{n} \rho_r(\bar{y}_j - \bar{y}_{M_j}) = \frac{1}{3} \sum_{j=1}^{n} \left[ \tau \left( \frac{\eta_j^+ + \delta_j^-}{2} \right) + \xi_j^+ + \zeta_j^+ \right] + \frac{1}{2} \sum_{j=1}^{n} \left[ (1 - \tau) \left( \frac{\eta_j^- + \delta_j^+}{2} \right) + \xi_j^- + \zeta_j^- \right] \to \min,
\]

\( \eta_j^+, \delta_j^+, \xi_j^+, \zeta_j^+ \geq 0, \eta_j^-, \delta_j^-, \xi_j^-, \zeta_j^- \geq 0, a_i \in \mathbb{R}, a_i^0 \geq 0, a_i^R \geq 0, i = 1, \ldots, m. \)

When unknown regression coefficients are found, it is easy to determine the first component of the model output Z-number. The question is how to determine the reliability (the second component) of this number? Define the weighted fuzzy number \( \bar{y}_{M_j}, j = 1, n \) of the model value \( Z_j = \bar{a}_1 Z_{1j} + \bar{a}_2 Z_{2j} + \cdots + \bar{a}_m Z_{mj}, j = 1, n \). and the weighted fuzzy numbers. \( \bar{y}_v, \nu = \mathbb{T}, \mathbb{V}, j = 1, n \) of Z-numbers \( (\bar{a}_1 \bar{x}_{1j} + \bar{a}_2 \bar{x}_{2j} + \cdots + \bar{a}_m \bar{x}_{mj}, \bar{R}_v), \nu = \mathbb{T}, \mathbb{V}, j = 1, n \), where \( \bar{R}_v, \nu = \mathbb{T}, \mathbb{V} \) are formalizations of the values of a linguistic variable with name “Reliability”. Let define \( \rho_r(\bar{y}_{M_j} - \bar{y}_v), \nu = \mathbb{T}, \mathbb{V}, j = 1, n \). If \( \rho_r(\bar{y}_{M_j} - \bar{y}_v) = \min \rho_r(\bar{y}_{M_j} - \bar{y}_v), \nu = \mathbb{T}, \mathbb{V}, j = 1, n, \) then the reliability of the model value is fuzzy number \( \bar{R}_v \) and accordingly s-th term of linguistic variable “Reliability”.

4 Conclusion

Fuzzy regression analysis has significantly expanded the capabilities of classical regression analysis and has become an indispensable tool for predicting and studying dependencies in conditions of fuzzy characteristics, statements and requirements, fuzzy estimates and goals. When it became possible to assess the reliability of the received fuzzy information and formalize it for further processing, it became clear that the known methods of regression analysis could not be used in such a situation and therefore the development of new methods is necessary.

It is quite obvious that one well-known regression model with crisp coefficients is not enough to solve most practical problems. Also, linear and quadratic Z-regressions with fuzzy coefficients developed with the help of the least-squares method, the disadvantage of which is their high sensitivity to outliers are not enough.

In this paper, we have developed a Z-regression that is free from the shortcomings of existing models, and the solution of the optimization problem is simply reduced to solving the well-known linear programming problem. The coefficients of Z-regression are triangular numbers.

For a Z-number, a weighted fuzzy number is defined as the result of multiplying the first component of Z-number by weighted point determined for the corresponding second component of Z-number. For a weighted fuzzy number, the definition of the loss function is given, which was applied to construct an optimization function based on the loss functions of the difference between the initial and model data.
The developed Z-regression provides new opportunities for solving tasks in problem areas with the active participation of experts, taking into account the reliability of information received from them.

The further research will be related to the solution of decision support problems based on the prediction of environmental monitoring indicators and their reliability.

References

16. O. M. Poleshchuk, LNEE., 729, 404-412 (2021)
19. O. M. Poleshchuk, LNEE., 857, 383-392 (2022)