Confidence intervals for output variable in fuzzy linear regression model

O M Poleschchuk

Bauman Moscow State Technical University, Moscow, Russia
E-mail: olga.m.pol@yandex.ru

Abstract. An analogue of confidence intervals with a given level of possibility for output variable in a fuzzy linear regression model has been developed in the paper. The methods of fuzzy regression analysis extend the methods of classical regression analysis and allow to solve different problems in conditions of fuzzy and incomplete initial information without the limits of the probabilities methods. The developed approach opens up new opportunities for predicting of output fuzzy variable.

1. Introduction
The general problem of studies dependencies within the framework of classical regression analysis can be formulated as follows: based on the values \( x_{i1}, \ldots, x_{ik}, y_i, i = 1, n \) of variables \( X_1, \ldots, X_k, Y \), construct a function \( f(X_1, \ldots, X_k) \) that would restore (in a certain sense) the values of the variable \( Y \) with respect to the values of the variables \( X_1, \ldots, X_k \). In the classical linear regression model, the model can be written in the general form as:

\[
y_i = a_1 x_{i1} + \ldots + a_k x_{ik} + b + \varepsilon_i, i = 1, n.
\]

A special case of this model is a linear regression model for two variables (sample regression model):

\[
y_i = ax_i + b + \varepsilon_i,
\]

where \( x_i \) - non-random variables, \( y_i, \varepsilon_i \) - random variables, \( \varepsilon_i \) - regression errors, standard normal random variables, the vector \( (x_1, x_2, \ldots, x_n) \) is not collinear to the vector \( (1,1,\ldots,1) \),

\[
M(\varepsilon_i) = 0, \; D(\varepsilon_i) = \sigma^2, \; i = 1, n, \; M(\varepsilon_i, \varepsilon_j) = 0, i \neq j.
\]

According to the theorem Gauss – Markov [1], under these assumptions, unknown model parameters, obtained by the least squares method, are effective, that is, have the smallest variance in the class of all unbiased estimates.

To analyze relations between qualitative characteristics and the prediction of their values the methods of fuzzy regression analysis are used, which being actively developed have already considerably expanded boundaries of application of classical regression analysis methods, i.e. they allow to construct the regression relations on the basis of fuzzy initial information. Besides, this information can be of both quantitative and qualitative nature, thus, making possible application of methods of fuzzy regression analysis in the theory of expert evaluations and ensuring practical applications in various spheres of human activity [2, 3].

Methods of fuzzy regression analysis are used to study behavior of complex engineering, ecological and other systems with output indexes depending on a great many of parameters [2]. These methods
are applied to construct regression models not only within the limits of the fuzzy initial information, but also within the limits of the definite information. In this case the predicted output values are provided as fuzzy numbers. Such representation is explained by the fact that the real system is always more complex than any of its model not capable of combining all input variables on which the output variable depends.

The first fuzzy linear regression model [4] excited interest in contributors, thus resulting in occurrence of new fuzzy regression models based on various optimizing criteria. Today, a number of linear fuzzy regression models [5-11] is developed, and approaches to building of nonlinear fuzzy regression models [2] are outlined.

There appears to be three different approaches under the heading of «Fuzzy Regression»:

(a) Methods that were proposed by H. Tanaka [4] and further elaborated in current literature, where the coefficients of input variables are assumed to be fuzzy numbers. These fuzzy regression models are based on the possibility theory instead of the probability theory or they are based on both possibility and probability theories.

(b) Method proposed by R.J. Hathaway and J.C. Bezdek [5] where first the fuzzy clusters determined by an fuzzy c-means clustering algorithm define how many ordinary regressions are to be constructed, one for each cluster. Next each fuzzy cluster is used essentially for switching purposes to determine the most appropriate ordinary regression that is to be applied for a new input from amongst a number of ordinary regressions determined in the first place.

(c) Methods proposed by I.B.Turksen, A. Celikyilmaz, where the fuzzy functions approach to system modeling was developed. The new fuzzy functions approach augments the membership values together with their transformations to form a new input variable to find local functions. First the given system domain is fuzzy partitioned into c clusters using fuzzy c-means clustering algorithm. Then, one regression function is calculated to model the behavior of each partition. In [9] linear regression function to estimate the parameters of each function is proposed. A new fuzzy system modeling approach that identifies the fuzzy functions using support vector machines is proposed in [10]. This new approach is structurally different from the fuzzy rule base approaches and fuzzy regression methods. Method support vector machines is applied to determine the support vectors for each fuzzy cluster obtained by fuzzy c-means clustering algorithm. Original input variables, the membership values obtained from the fuzzy c-means clustering algorithm together with their transformations form a new augmented set of input variables.

A major difference between fuzzy regression and classical regression is in dealing with errors as fuzzy variables in fuzzy regression modeling, and in dealing with errors as random variables in classical regression modeling. The researchers have tried to integrate both fuzziness and randomness into regression model. As a result of this the hybrid fuzzy least-squares regressions were developed [2, 3, 11].

The problem of confidence intervals constructing for the output variable in fuzzy regression models was practically not considered, therefore the research topic is new and relevant.

2. Necessary information of fuzzy sets theory

According to [12], the set of pairs \( \{ x, \mu_\alpha(x) : x \in X \} \) is referred to as a fuzzy set \( \tilde{A} \), where \( \mu_\alpha(x) : X \rightarrow [0,1] \). Function definition area of \( \mu_\alpha(x) \) is referred to as universal set \( X \) of fuzzy set \( \tilde{A} \).

A subset of universal set, defined as \( A_\alpha = \{ x \in X : \mu_\alpha(x) \geq \alpha \}, \alpha \in [0,1] \)

is referred to as set of \( \alpha - \) cut ( \( \alpha \) -level) of fuzzy set \( \tilde{A} \) with membership function \( \mu_\alpha(x) \).

Fuzzy number \( \tilde{A} \) is referred to as fuzzy subset of set of real numbers \( R \) possessing membership function

\[ \mu_\alpha : R \rightarrow [0,1] \]
The subset \( S_A \subset R \) is referred to as the support of fuzzy number \( \tilde{A} \), if
\[
S_A = \sup \{ x : \mu_A(x) > 0 \}.
\]
Fuzzy number \( \tilde{A} \) is referred to as positive, if \( \forall x \in S_A, x > 0 \), and negative, if \( \forall x \in S_A, x < 0 \).
A subset of real line \( R \), defined as
\[
A_\alpha = \{ x \in R : \mu_A(x) \geq \alpha, \alpha \in [0,1] \}
\]
is referred to as set of \( \alpha \) - cut (\( \alpha \) -level) of fuzzy number \( \tilde{A} \) with membership function \( \mu_A(x) \).
Fuzzy number \( \tilde{A} \) with membership function \( \mu_A(x) \) is referred to as normal, if
\[
\max_x \mu_A(x) = 1, x \in R.
\]
A normal triangular fuzzy number \( \tilde{A} = (a, a_L, a_R) \) is fuzzy number with membership function
\[
\mu_A(x) = \begin{cases} 
1 - \frac{a-x}{a_L}, & 0 < \frac{a-x}{a_L} \leq 1, a_L > 0 \\
1 - \frac{x-a}{a_R}, & 0 < \frac{x-a}{a_R} \leq 1, a_R > 0 \\
1, & x = a \\
0, & x < a - a_L, \text{ or } x > a + a_R 
\end{cases}
\]
In that case \( A_\alpha = [A_\alpha^1, A_\alpha^2] = [a - (1 - \alpha)a_L, a + (1 - \alpha)a_R] \quad \alpha \in [0,1] \).

The definition of weighted point \( A \) for a normal triangular number \( \tilde{A} = (a, a_L, a_R) \) was given in [13]:
\[
A = \int_{0}^{1} \left( \int_{0}^{1} \frac{A_\alpha + A_\alpha^2}{2} \right) d\alpha = \int_{0}^{1} \left( \int_{0}^{1} \frac{A_\alpha^1 + A_\alpha^2}{2} \right) d\alpha = \frac{1}{6} (a - (1 - \alpha)a_L + a + (1 - \alpha)a_R) x d\alpha = a + \frac{1}{6} (a_R - a_L).
\]
The weighted interval \( [A_1, A_2] \) for the normal triangular number \( \tilde{A} = (a, a_L, a_R) \) was given in [13], such as \( A_1 = a - \frac{1}{6} a_L \), \( A_2 = a + \frac{1}{6} a_R \).

### 3. Fuzzy linear regression model

Let us consider a problem of fuzzy linear regression construction. Relation between input and output data \( x_i, y_i, i = 1,n \) of variables \( X, Y \) is determined as
\[
Y = \tilde{a}X + \tilde{b},
\]
where \( \tilde{a} = (a, a_L, a_R) \), \( \tilde{b} = (b, b_L, b_R) \) are unknown coefficients of the regression model, which are determined as normal triangular fuzzy numbers.

The method of regression’s creation is based on the transformation of the input and output fuzzy numbers into weighted intervals.

If \( x_i \geq 0 \), then the weighted interval of \( \tilde{a}x_i + \tilde{b} \) has the left border equals to \( x_i \left( a - \frac{1}{6} a_L \right) + \left( b - \frac{1}{6} b_L \right) \) and has the right border equals to \( x_i \left( a + \frac{1}{6} a_R \right) + \left( b + \frac{1}{6} b_R \right) \). If \( x_i < 0 \),
then the weighted interval of \(\tilde{a}x_i + \tilde{b}\) has the left border equals to \(x_i\left(a + \frac{1}{6}a_L\right) + \left(b - \frac{1}{6}b_R\right)\) and has the right border equals to \(x_i\left(a - \frac{1}{6}a_L\right) + \left(b + \frac{1}{6}b_R\right)\).

Let us define an affinity measure for two numbers \(\hat{A}, \hat{B}\) with the weighed intervals \([A_i, A_j], [B_i, B_j]:\)

\[
f(\hat{A}, \hat{B}) = \sqrt{(A_i - B_i)^2 + (A_j - B_j)^2}.
\]

Let us consider a functional

\[
F = \sum_{i=1}^{n} f^2(Y_i, \tilde{Y}_i),
\]

which characterizes an affinity measure between initial \(Y_i\) and model output data \(\tilde{Y}_i\).

The optimization problem is set as follows:

\[
F(a, a_L, a_R, b, b_L, b_R) = \sum_{i=1}^{n} f^2(Y_i, \tilde{Y}_i) \rightarrow \min,
\]

\[
a_L \geq 0, a_R \geq 0, b_L \geq 0, b_R \geq 0.
\]

If \(x_i \geq 0\), then \(f^2(Y_i, \tilde{Y}_i)\) equals to

\[
\left(y_i - x_i\left(a - \frac{1}{6}a_L\right) - \left(b - \frac{1}{6}b_L\right)\right)^2 + \left(y_i - x_i\left(a + \frac{1}{6}a_R\right) - \left(b + \frac{1}{6}b_R\right)\right)^2, \quad i = 1, n.
\]

If \(x_i < 0\), then \(f^2(Y_i, \tilde{Y}_i)\) equals to

\[
\left(y_i - x_i\left(a + \frac{1}{6}a_L\right) - \left(b + \frac{1}{6}b_L\right)\right)^2 + \left(y_i - x_i\left(a - \frac{1}{6}a_R\right) - \left(b - \frac{1}{6}b_R\right)\right)^2, \quad i = 1, n.
\]

Since sum of \(f^2(Y_i, \tilde{Y}_i)\) is piecewise linear functions in the field \(a_L \geq 0, a_R \geq 0, b_L \geq 0, b_R \geq 0\), then \(F\) is piecewise differentiable function, and solutions of an optimization problem are determined by means of known methods [14].

4. Confidence intervals for output variable in linear regression models

Let us consider a statistic \(Y - (a + bX) / \sigma_y\), that belongs to the standard normal distribution with known \(\sigma_y, a, b\). In that case we would use Laplace function for prediction output variable \(Y\) in the form of confidence intervals. Since \(\sigma_y, a, b\) are unknown, we will use their estimates and in particular

\[
S^2_y = \sigma^2\left(\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right).
\]

Accordingly for constructing of confidence intervals we will use Student distribution.

Since \(\sigma^2\) is unknown, we will use their estimate

\[
S^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2,
\]

Then

\[
S^2_y = S^2\left(\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right).
\]
Then random variable \( \frac{Y - (a + bX)}{S_y} \) has Student distribution \( t_{n-2} \) with the number of degree of freedom - \( n - 2 \), \( P(t_{n-2} > t_{n-2,\alpha}) = \alpha \). Then:

\[
a + bX - t_{n-2,\alpha}S_y \leq Y \leq a + bX + t_{n-2,\alpha}S_y.
\]

The developed in [13] method of regression’s creation, based on the transformation of the input and output fuzzy numbers into intervals, allows to construct an analogue of confidence intervals for output variable in a fuzzy linear regression model. It should be emphasized that this does not require knowledge of the distribution of regression errors and any conditions on the initial information.

From the perspective of the possibilities theory, the membership function of a fuzzy number is a function of the possibilities distribution, that a fuzzy number takes one or another of its values. If \( \mu_A(x) \geq 1 - \alpha \), then \( a - aa_L \leq x \leq a + aa_R \) with the possibility greater or equal to \( 1 - \alpha \).

Based on this, a confidence interval for the output variable with the possibility \( \mu_Y(x) \geq 1 - \alpha \) can be determined as follows:

\[
x(a - aa_L) + b - ab_L \leq y \leq x(a + aa_R) + b + ab_R \quad \text{with} \quad x \geq 0,
\]

\[
x(a + aa_R) + b - ab_L \leq y \leq x(a - aa_L) + b + ab_R \quad \text{with} \quad x < 0.
\]

**Conclusion**

An analogue of confidence intervals for output variable in a fuzzy linear regression model is developed in this paper. The methods of fuzzy regression analysis extend the possibilities of classical regression analysis as these methods allow to fit a model to fuzzy data, crisp data, and their mixture.

The problem of confidence intervals constructing for the output variable in fuzzy regression models was practically not considered, therefore the research topic is new and relevant. It should be emphasized that developed constructing does not require knowledge of the distribution of regression errors and does not require any conditions on the initial information. Based on this, an analogue of confidence intervals with a given level of possibility for output variable in a fuzzy linear regression model has been developed in the paper. The developed approach opens up new opportunities for predicting of output fuzzy variable.

**References**

[1] Gnedenko B V 1988 *Probability theory* (M: Nauka) p 448.
[2] Poleshchuk O, Komarov E 2011 *Expert fuzzy information processing*, Studies in Fuzziness and Soft Computing, vol. 268, pp 1-239.
[3] Poleshchuk O M, Komarov E G 2008 Multiple hybrid regression for fuzzy observed data, *Annual Conference of the North American Fuzzy Information Processing Society - NAFIPS* - 2008, p 4531224.
[4] H Tanaka, S Uejima, K Asai 1982 Linear regression analysis with fuzzy model, *IEEE. Systems, Trans. Systems Man Cybernet. - SMC-2*, pp 903-907.
[5] R J Hathaway, J C Bezdek 1993 Switching regression models and fuzzy clustering, *IEEE Transactions on fuzzy systems*, vol. 1, № 3, pp 195-203.
[6] D A Sabic, W Pedrycz 1991 Evaluation on fuzzy linear regression models, *Fuzzy Sets and Systems*, vol. 39, pp 51 - 63.
[7] Y-H O Chang 2001 Hybrid fuzzy least-squares regression analysis and its reliability measure, *Fuzzy Sets and Systems*, vol. 119, pp 225-246.
[8] Y-H O Chang, B M Ayyub 2001 Fuzzy regression methods – a comparative assessment, *Fuzzy Sets and Systems*, vol. 119, pp 187-203.
[9] I.B. Turksen 2008 Fuzzy functions with LSE, *Applied Soft Computing*, vol. 8 № 3, pp 1178-
[10] A. Celikyilmaz, I.B. Turksen 2007 Fuzzy functions with support vector machines, *Information Sciences*, vol. 177, pp 5163–5177.

[11] Poleshchuk, O.M., Komarov, E.G., Darwish, A. 2016 A fuzzy linear regression model with interval type-2 fuzzy coefficients, *Proceedings of the 19th International Conference on Soft Computing and Measurements, SCM - 2016*, pp 388-391.

[12] L.A. Zadeh 1975 The Concept of a linguistic variable and its application to approximate reasoning, *Information Sciences*, vol. 8 pp 199-249.

[13] Poleshuk O. M., Komarov E. G. 2008 New defuzzification method based on weighted intervals, *Annual Conference of the North American Fuzzy Information Processing Society – NAFIPS 2008*, p. 4531223.

[14] T.F. Coleman, Y. Li 1996 A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables, *SIAM J. Optim*, vol. 6, pp 1040-1058.