Journal of Applied Mathematics, Statistics and Informatics (JAMSI), **1** (2005), No. 2

Classical regression models versus fuzzy regression models

ŠTEFAN VARGA

Abstract

In this paper are presented classical and robust estimators of unknown fuzzy parameters in the fuzzy regression model and investigated the relationship between estimators in the classical regression model and in the fuzzy regression model.

Mathematics Subject Classification 2000: 62J12

Additional Key Words and Phrases: Fuzzy regression model, estimations, predictions

1. CLASSICAL REGRESSION MODELS

The classical regression model (linear in parameters) is studied in the form

$$y = a_1 f_1(x) + a_2 f_2(x) + \cdots + a_m f_m(x)$$

where $f_i(x)$ $(i = 1, 2, \ldots, m)$ are known functions of the input variable x (predictor), y is an output variable (response) and $a = (a_1, a_2, \ldots, a_m)^T$ is the vector of unknown parameters. An observed value

$$y_i = a_1 f_1(x_i) + a_2 f_2(x_i) + \cdots + a_m f_m(x_i) + e_i$$

measured in the point x_i with the error e_i $(i = 1, 2, \ldots, n)$ is a random variable with some probability distribution (the most frequently normal distribution). The uncertainty of the value y_i $(i = 1, 2, \ldots, n)$ is expressed by a probability distribution or at least by the expectation

$$E(y_i) = a_1 f_1(x_i) + a_2 f_2(x_i) + \cdots + a_m f_m(x_i)$$

$(E(e_i) = 0)$ and the variance

$$D(y_i) = D(e_i) = \sigma_i^2$$

Practically all types of estimators of the vector of unknown parameters
$a = (a_1, a_2, \ldots, a_m)^T$
in the classical regression model are functions of residuals r_i (distances between observed values y_i and estimated values $est\ y_i$; $i = 1, 2, \ldots, n$)

$$r_i = y_i - est\ y_i$$

The most known estimators of the vector of unknown parameters $a = (a_1, a_2, \ldots, a_m)^T$ in the studied model are

—least square estimator

$$est_{LS}\ a = \arg\min_{a\in\mathbb{R}^m} \sum_{i=1}^{n} (y_i - est\ y_i)^2$$

minimizes the sum of squares of the residuals $r_i = y_i - est\ y_i$. The estimators of the response variable y are functions of the estimators of the vector of unknown parameters

$$est\ y_i = est\ a_1 f_1(x_i) + est\ a_2 f_2(x_i) + \cdots + est\ a_m f_m(x_i)$$

—weighted least square estimator

$$est_{WLS}\ a = \arg\min_{a\in\mathbb{R}^m} \sum_{i=1}^{n} w_i (y_i - est\ y_i)^2$$

minimizes the weighted sum of squares of the residuals $r_i = y_i - est\ y_i$. If the variance of the observed value y_i $(i = 1, 2, \ldots, n)$ is

$$D(y_i) = D(e_i) = \sigma_i^2 = h_i \sigma^2$$

where h_i is its known and σ^2 unknown part, the weight $w_i = 1/h_i$.

—least trimmed square estimator

$$est_{LTS}\ a = \arg\min_{a\in\mathbb{R}^m} \sum_{i=1}^{k} (y_i - est\ y_i)^2_{(i)}$$

minimizes the sum of k least squares of the residuals $r_i = y_i - est\ y_i$. This is a robust estimator of the vector of the unknown parameters $a = (a_1, a_2, \ldots, a_m)^T$. The residuals in the sum for minimization are ordered and the number of the residuals $k \in [n/2, n]$.

The least square estimator is BLUE (best linear unbiased estimator) in the classical regression model if the observations of the response variable y are independent and with the same variances (the weights $w_i = 1/h_i = 1$). On the other hand, if the observations of the response variable y are independent but with different variances ($w_i = 1/h_i \neq 1$) , the weighted least square estimator is BLUE in the classical regression model. Statistical properties of the mentioned two estimators are expressed by BLUE (best linear unbiased estimator).

The least trimmed square estimator is a robust estimator. It is suitable to use this estimator, if there are some outliers among the observed values in the classical regression model. Statistical properties of the least trimmed square estimator are expressed by the breakdown point. For example, if $k = [n/2] + [(m+1)/2]$, then the breakdown point $\varepsilon^* = 0.5$, it means, that approximately fifty percent of observation can be arbitrarily changed and the estimator will be approximately the same.

2. FUZZY REGRESSION MODELS

Very natural generalization of the classical regression model is the fuzzy regression model studied in the form

$$Y = A_1 f_1(x) + A_2 f_2(x) + \cdots + A_m f_m(x)$$

where the input variable x (predictor) is a crisp (real) variable, $f_i(x)$ $(i = 1, 2, \ldots, m)$ are known real functions of the variable x, Y is an output fuzzy variable (response) and $A = (A_1, A_2, \ldots, A_m)^T$ is the vector of unknown fuzzy parameters. It is easy to see that the fuzzy numbers Y, A_i $(i = 1, 2, \ldots, m)$ are crisp (real number is a special case of fuzzy number), the fuzzy regression model is equal to the classical regression model.

The uncertainty of an observation Y_i in the point x_i $(i = 1, 2, \ldots, n)$ is expressed by a membership function μ_{Y_i} of the fuzzy number Y_i. We do not have any probability distribution, any expectation and any variance of the observed value Y_i $(i = 1, 2, \ldots, n)$.

The principle question is how to estimate the vector of unknown fuzzy parameters in the fuzzy regression model and how to define a quality of the estimator. One eventuality could be to generalize not only model, but to generalize the estimators defined in the classical regression model to the estimators in the fuzzy regression model too. What does it mean? It means that, for example,

$$est_{LTS}\ A = est_{LTS}(A_1, A_2, \ldots, A_m)^T$$

is the least trimmed square estimator of the vector of unknown fuzzy parameters in the fuzzy regression model, if it is equal to the least trimmed square estimator in the classical regression model

$$est_{LTS}\ a = est_{LTS}(a_1, a_2, \ldots, a_m)^T = \arg\min_{a \in \mathbb{R}^m} \sum_{i=1}^{k} (y_i - est\ y_i)^2_{(i)}$$

in the case that the observation Y_i $(i = 1, 2, \ldots, n)$ in the fuzzy regression model is a crisp (real) number (as a special case of fuzzy number) with the membership function

$$\mu_{Y_i}(x) = \begin{cases} 1 & x = Y_i \\ 0 & x \neq Y_i \end{cases}$$

Because the difference of two fuzzy numbers is a fuzzy number, we will not minimize a sum of squares of differences $Y_i - est\ Y_i$ between observed and estimated fuzzy values, but distances between them that can be defined as crisp numbers.

The most commonly used in practice are symmetric triangular fuzzy numbers

$$A = \langle a, s \rangle$$

where a is a center and s a spread of the fuzzy number A. This fuzzy number is "about a". Its membership function is

$$\mu_A(x) = \begin{cases} 1 - \frac{|x-a|}{s} & a - s \leq x \leq a + s \\ 0 & \text{otherwise} \end{cases}$$

For addition of two fuzzy numbers $A = \langle a, s_1 \rangle$, $B = \langle b, s_2 \rangle$ we can use

$$A + B = \left\langle a + b,\ \sqrt[w]{s_1^w + s_2^w} \right\rangle$$

and for multiplication of the fuzzy number $A = \langle a, s_1 \rangle$ with the real number k

$$kA = \left\langle ka,\ \sqrt[w]{|k|} s_1 \right\rangle$$

where the parameter $w \in [1, \infty]$. We have the set of arithmetic, but the most interesting are the limit situations. For $w = 1$

$$A + B = \langle a + b, s_1 + s_2 \rangle, \quad kA = \langle ka, |k| s_1 \rangle$$

and for $w = \infty$

$$A + B = \langle a + b, \max\{s_1, s_2\} \rangle, \quad kA = \langle ka, s_1 \rangle$$

The distance of two fuzzy numbers that is a real number and that is a generalization of the Euclidean distance of two real numbers is the Diamond distance defined for two fuzzy numbers $A = \langle a, s_1 \rangle$, $B = \langle b, s_2 \rangle$ by the formula

$$d_D^2(A, B) = (a - b)^2 + \frac{2}{3}(s_1 - s_2)^2$$

Now when we have defined arithmetic and distance for fuzzy numbers we can specify the studied fuzzy regression model and define suitable estimators for unknown fuzzy parameters.

3. ESTIMATIONS IN FUZZY REGRESSION MODELS

The fuzzy regression model is studied in the form

$$Y = A_1 f_1(x) + A_2 f_2(x) + \cdots + A_m f_m(x)$$

where the input variable x (predictor) is a crisp variable, $f_i(x)$ are known real functions of the variable x, Y is an output fuzzy variable (response), the observation Y_i $(i = 1, 2, \ldots, n)$ is a symmetric triangular fuzzy number (y_i is a center and z_i is a spread)

$$Y_i = \langle y_i, z_i \rangle$$

$(y_i \in \mathbb{R}, z_i \in \mathbb{R}^+)$ and $A = (A_1, A_2, \ldots, A_m)^T$ is the vector of unknown symmetric triangular fuzzy parameters

$$A_i = \langle a_i, s_i \rangle$$

$(a_i \in \mathbb{R}, s_i \in \mathbb{R}^+)$.

Definition 1. Least square estimator of the vector of unknown fuzzy parameters $A = (A_1, A_2, \ldots, A_m)^T$ in the fuzzy regression model is

$$est_{LS}\ A = est_{LS}(A_1, A_2, \ldots, A_m)^T = \arg \min_{a_j \in \mathbb{R}, s_j \in \mathbb{R}^+} \sum_{i=1}^{n} d_D^2(Y_i, est\ Y_i)$$

Definition 2. Weighted least square estimator of the vector of unknown fuzzy parameters $A = (A_1, A_2, \ldots, A_m)^T$ in the fuzzy regression model is

$$est_{WLS}\ A = est_{WLS}(A_1, A_2, \ldots, A_m)^T = \arg \min_{a_j \in \mathbb{R}, s_j \in \mathbb{R}^+} \sum_{i=1}^{n} w_i d_D^2(Y_i, est\ Y_i)$$

where w_i $(i = 1, 2, \ldots, n)$ is the weight of the observation $Y_i = \langle y_i, z_i \rangle$.

Definition 3. Least trimmed square estimator of the vector of unknown fuzzy parameters $A = (A_1, A_2, \ldots, A_m)^T$ in the fuzzy regression model is

$$est_{LTS}\ A = est_{LTS}(A_1, A_2, \ldots, A_m)^T = \arg \min_{a_j \in \mathbb{R}, s_j \in \mathbb{R}^+} \sum_{i=1}^{k} d_D^2(Y_i, est\ Y_i)_{(i)}$$

where the distances between observed and estimated values are ordered ($k \in [n/2, n]$)

$$d_D^2(Y_i, est\ Y_i)_{(1)} \leq d_D^2(Y_i, est\ Y_i)_{(2)} \leq \ldots$$

To estimate the vector of unknown fuzzy parameters $A = (A_1, A_2, \ldots, A_m)^T$ means, to estimate the vector of all centers $a = (a_1, a_2, \ldots, a_m)^T$ and the vector of all spreads $s = (s_1, s_2, \ldots, s_m)^T$ of the parameters.

Theorem 1. The least square estimator of the vector of the centers a and the vector of the spreads s of the unknown fuzzy parameters $A = (A_1, A_2, \ldots, A_m)^T$ in the fuzzy regression model is

$$\begin{aligned} est_{LS}(a, s) &= est_{LS}(a_1, \ldots, a_m, s_1, \ldots, s_m) = \\ &= \arg \min_{a_j \in \mathbb{R}, s_j \in \mathbb{R}^+} \sum_{i=1}^{n} \left[(y_i - a^T f_i)^2 + \frac{2}{3} (z_i - \sqrt[w]{s^w |f_i|})^2 \right] \end{aligned}$$

where new elements in the formula are two column vectors $f_i = \big(f_1(x_i), \ldots, f_m(x_i)\big)^T$, $|f_i| = \big(|f_1(x_i)|, \ldots, |f_m(x_i)|\big)^T$ and one row vector $s^w = (s_1^w, \ldots, s_m^w)$.

Proof. It is enough to prove that the square of the Diamond distance of the observed value Y_i and the estimated value $est\ Y_i$ in the definition 1 is

$$d_D^2(Y_i, est\ Y_i) = \big(y_i - a^T f_i\big)^2 + \frac{2}{3} \Big(z_i - \sqrt[w]{s^w |f_i|}\Big)^2$$

The observation $Y_i = \langle y_i, z_i \rangle$ but what is the fuzzy number $est\ Y_i$? Using arithmetic presented in this paper we have

$$\begin{aligned} est\ Y_i &= \langle a_1, s_1 \rangle \cdot f_1(x_i) + \cdots + \langle a_m, s_m \rangle \cdot f_m(x_i) \\ est\ Y_i &= \langle a_1 f_1(x_i) + \cdots + a_m f_m(x_i),\ \sqrt[w]{s_1^w |f_1(x_i)| + \cdots + s_m^w |f_m(x_i)|} \rangle \\ est\ Y_i &= \langle a^T f_i,\ \sqrt[w]{s^w |f_i|} \rangle \end{aligned}$$

and the square of the distance of the fuzzy numbers Y_i, $est\ Y_i$ is

$$d_D^2(Y_i, est\ Y_i) = \big(y_i - a_T f_i\big)^2 + \frac{2}{3} \Big(z_i - \sqrt[w]{s^w |f_i|}\Big)^2$$

Theorem 2. The weighted least square estimator of the vector of the centers a and the vector of the spreads s of the unknown fuzzy parameters $A = (A_1, A_2, \ldots, A_m)^T$ in the fuzzy regression model is

$$\begin{aligned} est_{WLS}\ (a, s) &= est_{WLS}\ (a_1, \ldots, a_m, s_1, \ldots, s_m) = \\ &= \arg \min_{a_j \in \mathbb{R}, s_j \in \mathbb{R}^+} \sum_{i=1}^{n} w_i \left[(y_i - a^T f_i)^2 + \frac{2}{3} (z_i - \sqrt[w]{s^w |f_i|})^2 \right] \end{aligned}$$

where w_i $(i = 1, 2, \ldots, n)$ is the weight of the observation $Y_i = \langle y_i, z_i \rangle$.

Theorem 3. The least trimmed square estimator of the vector of the centers a and the vector of the spreads s of the unknown fuzzy parameters $A = (A_1, A_2, \ldots, A_m)^T$ in the fuzzy regression model is

$$est_{LTS}\ (a, s) = est_{LTS}\ (a_1, \ldots, a_m, s_1, \ldots, s_m) =$$
$$= \arg \min_{a_j \in \mathbb{R}, s_j \in \mathbb{R}^+} \sum_{i=1}^{k} \left[(y_i - a^T f_i)^2 + \frac{2}{3} \left(z_i - \sqrt[w]{s^w |f_i|} \right)^2 \right]_{(i)}$$

where the distances between observed and estimated values are ordered and $k \in [n/2, n]$.

Proofs of theorem 2 and theorem 3 are very simple modifications of the proof of theorem 1.

Theorem 4. If the observations $Y_i = \langle y_i, z_i \rangle$; $(i = 1, 2, \ldots, n)$ in the fuzzy regression model are crisp $(z_i = 0; i = 1, 2, \ldots, n)$ then the estimators of the unknown fuzzy parameters presented in theorems 1, 2, 3 are crisp too and equal to the analogous estimators in the classical regression model.

Proof. The spreads of all observations $z_i = 0$ $(i = 1, 2, \ldots, n)$ and therefore

$$\sum_{i=1}^{k} \left[(y_i - a^T f_i)^2 + \frac{2}{3} \left(z_i - \sqrt[w]{s^w |f_i|} \right)^2 \right]_{(i)} = \min$$

if all elements of the vector $s^w = (s_1^w, \ldots, (s_m^w)$ are zero. It means that $s_i = 0$ $(i = 1, 2, \ldots, m)$ thus the fuzzy parameters $A_i = \langle a_i, 0 \rangle$ are crisp $(i = 1, 2, \ldots, m)$ and

$$\sum_{i=1}^{k} \left[(y_i - a^T f_i)^2 + \frac{2}{3} \left(z_i - \sqrt[w]{s^w |f_i|} \right)^2 \right]_{(i)} = \sum_{i=1}^{k} (y_i - a^T f_i)^2 =$$
$$= \sum_{i=1}^{k} \left[y_i - \left(a_1 f_1(x_i) + \cdots + a_m f_m(x_i) \right) \right]_{(i)}^2 = \sum_{i=1}^{k} (y_i - est\ y_i)_{(i)}^2$$

what is the formula for the least trimmed square estimator in the classical regression model. The proofs for others two estimators are analogous.

Example. Let the observed data Y in the real (crisp) points x are fuzzy numbers

x	1.0	2.0	2.4	2.6	3.0	4.0
Y	$\langle 2.2, 0.3 \rangle$	$\langle 3.5, 0.2 \rangle$	$\langle 3.8, 0.1 \rangle$	$\langle 4.2, 0.1 \rangle$	$\langle 4.6, 0.2 \rangle$	$\langle 4.7, 0.3 \rangle$

Find the least square estimators $(w = 1)$ of the fuzzy parameters $A_1 = \langle a_1, s_1 \rangle$, $A_2 = \langle a_2, s_2 \rangle$ of the dependence

$$Y = A_1 + A_2 \ln x$$

Solution. Using the theorem 1, $f_1(x) = 1$, $f_2(x) = \ln x$, $f_i^T = (1, \ln x_i)$, $|f_i^T| = (1, |\ln x_i|)$ $i = 1, 2, \ldots, 6$, $a^T = (a_1, a_2)$, $s^w = (s_1, s_2)$ and we have minimized the

expression

$$\sum_{i=1}^{n}\Big[(y_i - a^T f_i)^2 + \frac{2}{3}\big(z_i - \sqrt[w]{s^w|f_i|}\big)^2\Big] = (2.2 - a_1 - a_2\ln 1)^2 +$$

$$+\frac{2}{3}(0.3 - s_1 - s_2\ln 1)^2 + (3.5 - a_1 - a_2\ln 2)^2 + \frac{2}{3}(0.2 - s_1 - s_2\ln 2)^2 + \cdots$$

$$\cdots + (4.7 - a_1 - a_2\ln 4)^2 + \frac{2}{3}(0.3 - s_1 - s_2\ln 4)^2$$

under the conditions $s_1 \geq 0$, $s_2 \geq 0$.

Results: The minimum 0.166530 was obtained for

$$a_1 = 2.223223, \quad s_1 = 0.199897$$
$$a_2 = 1.928391, \quad s_2 = 0.000037$$

Estimators of the unknown fuzzy parameters of the dependence $Y = A_1 + A_2\ln x$ are

$$A_1 = \langle 2.223223, 0.199897\rangle$$
$$A_2 = \langle 1.928391, 0.000037\rangle$$

The *observed* and *estimated* values of the output fuzzy variable Y:

x	Y-observed	Y-estimated
1.0	(1.9, 2.2, 2.5)	(2.023, 2.223, 2.423)
2.0	(3.3, 3.5, 3.7)	(3.360, 3.560, 3.760)
2.4	(3.7, 3.8, 3.9)	(3.712, 3.911, 4.111)
2.6	(4.1, 4.2, 4.3)	(3.866, 4.066, 4.266)
3.0	(4.4, 4.6, 4.8)	(4.142, 4.342, 4.542)
4.0	(4.4, 4.7, 5.0)	(4.697, 4.897, 5.096)

REFERENCES

J. Anděl: *Matematická statistika*, SNTL / Alfa Praha, 1985.

A. Bárdossy, L. Duckstein: *Fuzzy Rule - Based Modeling with Applications to Geophysical, Biological and Eng. Systems*, CRC Press, Boca Raton, 1995.

G. J. Klir, B. Yuan: *Fuzzy Sets and Fuzzy Logic – Theory and Applications*, Prentice - Hall PTR, Upper Saddle River, NJ, USA, 1995.

M. Šabo: *On T-reverse of T-norms*, Tatra Mt. Math. Pub. 12 (1997), 35-40.

Varga, Š.: *Robust estimations in fuzzy linear regression models. Quo Vadis Computational Intelligence*, Physica-Verlag, Heidelberg 2000, 239-246.

Varga, Š., Šabo, M. : *Linear regression with fuzzy variables. The State of the Art in Computational Intelligence*, Physica-Verlag, Heidelberg 2000, 99-103.

ŠTEFAN VARGA,
Faculty of Chemical and Food Technology,
Department of Mathematics,
Slovak University of Technology,
Radlinského 9, 812 34 Bratislava, Slovak Republic,
e-mail: stefan.varga@stuba.sk

Received Jun 2005