



Hypotheses testing for fuzzy robust regression parameters

Kamile Şanlı Kula^{a,*}, Ayşen Apaydın^b

^a Ahi Evran University, Department of Mathematics, 40200 Kırşehir, Turkey

^b Ankara University, Department of Statistics, 06100 Ankara, Turkey

ARTICLE INFO

Article history:

Accepted 30 March 2009

ABSTRACT

The classical least squares (LS) method is widely used in regression analysis because computing its estimate is easy and traditional. However, LS estimators are very sensitive to outliers and to other deviations from basic assumptions of normal theory [Huynh H. A comparison of four approaches to robust regression. *Psychol Bull* 1982;92:505–12; Stephenson D. 2000. Available from: <http://folk.uib.no/ngbnk/kurs/notes/node38.html>; Xu R, Li C. Multidimensional least-squares fitting with a fuzzy model. *Fuzzy Sets and Systems* 2001;119:215–23.]. If there exists outliers in the data set, robust methods are preferred to estimate parameters values. We proposed a fuzzy robust regression method by using fuzzy numbers when x is crisp and Y is a triangular fuzzy number and in case of outliers in the data set, a weight matrix was defined by the membership function of the residuals. In the fuzzy robust regression, fuzzy sets and fuzzy regression analysis was used in ranking of residuals and in estimation of regression parameters, respectively [Şanlı K, Apaydın A. Fuzzy robust regression analysis based on the ranking of fuzzy sets. *Inernat. J. Uncertainty Fuzziness and Knowledge-Based Syst* 2008;16:663–81.]. In this study, standard deviation estimations are obtained for the parameters by the defined weight matrix. Moreover, we propose another point of view in hypotheses testing for parameters.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

Regression is a popular methodology for expressing the functional relationship on two or more related variables. With the mathematical form, the value of one variable can be predicted from the values of the others. It is this function that makes regression analysis one of the most useful techniques in operational research applications. Many areas including engineering, biology, business, economics etc. have wide applications [12].

The function form which is most frequently used for expressing the relationship is the linear form,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, \dots, n. \quad (1)$$

In Eq. (1), Y_i and X_{ij} , $j = 1, \dots, p$ are the observed response (or dependent) variable and the j th explanatory (or independent) variable in the i th case [12].

Fuzzy regression analysis is an extension of the classical regression analysis in which some elements of the model are represented by fuzzy number. Fuzzy regression methods have been successfully applied to various problems such as forecasting and engineering [12,14]. The fuzzy regression models analysis is generally divided into two categories. The first of them is based on the Tanaka et al. [18]'s linear programming approach and the second one is based on the fuzzy least squares approach.

Fuzzy linear regression was first introduced by Tanaka et al. [18]. The objective function is to minimize the total spread of the fuzzy parameters subject to the support of the estimated values to cover the support of the observed values for a certain

* Corresponding author.

E-mail addresses: sanli2004@hotmail.com (K.Ş. Kula), apaydin@science.ankara.edu.tr (A. Apaydın).

h -level. A generalization of the Tanaka approach for the general form of regression equations about LR-type fuzzy number is developed by Bardossy [2]. The Tanaka [19] approach is very complicated in solving the optimization problem. It is not clear what the relation is to a LS concept. The measure of best fit by the residuals is not presented in the Tanaka approach. Diamond [6] has suggested the fuzzy least squares method, to eliminate that disadvantage of Tanaka's model [14,15,22].

Celmins [3] criticized Tanaka's model (1987) because it produces crisp coefficient frequently. Later, Jozsef [11] showed that the minimization problem has a scale dependent for the model suggested by [18]. In response to Celmins's criticism (1987), Tanaka and Ishibuchi [20] developed a method to obtain interaction fuzzy coefficients. Their model is very sensitive to outliers [15].

The role of each observation should be given importance and the data should be tested in detail when data is analyzed. Because, sometimes even only one observation can change the values of the parameter estimates and omitting this observation from the data may lead to totally different estimates. This kind of observations which has a bigger residual value than the others is called outlier [1,16]. In the event that outliers in the data set, robust methods are preferred to estimate parameter values.

In this study, for the parameters of fuzzy robust regression, standard deviation estimations are calculated by the defined weight matrix and hypotheses testing have been applied for the parameters.

In the second part of the study, widely used robust methods are given and in the following part, the fuzzy regression method is discussed. In the fourth part, we give the comparison in case that observed Y_i and estimated \hat{Y}_i are symmetrical triangular fuzzy number. In the last part, hypotheses testing for parameters are obtained via the LS method, the M methods of Huber, Hampel, Andrews and Tukey, and the fuzzy robust regression method and numerical result of hypotheses testing for parameters are compared.

2. Robust methods

A linear regression model is given by matrix notation as,

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$$

$$\underline{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \quad \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \text{ and } \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where \underline{Y} is the dependent variable vector, X is independent variables matrix, $\underline{\beta}$ is the regression coefficients vector and $\underline{\varepsilon}$ is random error vector.

The LS estimates for the regression coefficients are obtained by minimizing equation

$$\underline{\varepsilon}'\underline{\varepsilon} = (\underline{Y} - X\underline{\beta})'(\underline{Y} - X\underline{\beta}),$$

so we have [10,15]

$$\hat{\underline{\beta}}_{LS} = (X'X)^{-1}X'\underline{Y}.$$

Standard deviation of the estimated LS coefficients is square root of diagonal components of matrix with $(p \times p)$ dimension following as

$$\frac{1}{n-p} \left(\sum_{i=1}^n e_i^2 \right) (X'X)^{-1}$$

[1].

The M estimator used minimizing of a function of the residuals much than minimizing the sum of the squared residuals. Regression coefficients are obtained by the minimizing the following sum

$$\sum_{i=1}^n \rho \left[\left(y_i - \sum_{j=1}^p x_{ij} \hat{\beta}_j \right) / d \right]. \quad (2)$$

Differentiating the sum in Eq. (2) with respect to the coefficients β_j , and setting the partial derivatives to zero, it may be found regression coefficient that p equations

$$\sum_{i=1}^n x_{ij} \psi \left[\left(y_i - \sum_{j=1}^p x_{ij} \hat{\beta}_j \right) / d \right] = 0, \quad j = 1, 2, \dots, p,$$

where $\psi(z) = \rho'(z)$ be the derivative of ρ and $r_i = y_i - \sum_{j=1}^p x_{ij}\hat{\beta}_j$. When the data contain outliers, standard deviations are not good measures of variability. Hence, other robust measures of variability are required. One robust measure of variability is d . In case of r_i is the residual of i th observation, $d = \text{median}|r_i - \text{median}(r_i)|/0.6745$, $i = 1, 2, \dots, n$. Therefore, the standardized residuals may be defined as $z_i = r_i/d$.

Standard deviation of the estimated robust regression coefficients is square root of diagonal components of matrix with $(p \times p)$ dimension in Eq. (3)

$$\frac{\frac{d^2}{n-p} \sum (\psi(z_i))^2}{\left(\frac{1}{n} \sum \psi'(z_i)\right)^2} (X'X)^{-1}, \tag{3}$$

where ψ' is derivative of ψ [7–10].

3. Fuzzy regression analysis

Tanaka et al. proposed the study in linear regression analysis using fuzzy set theory. They consider the fuzzy linear regression model as, $Y = A_0 + A_1x_1 + A_2x_2 + \dots + A_px_p$, where the independent variables, x_1, x_2, \dots, x_p , are non-fuzzy numbers while the dependent variable is a symmetric triangular fuzzy number. For i th observation, the estimated value $\tilde{Y}_i = (\bar{y}_i, e_i)$ is a fuzzy number where \bar{y}_i is the central value of \tilde{Y}_i and e_i is the spread value. The parameters estimations of fuzzy linear regression model are symmetrical triangular fuzzy numbers.

In the model developed by Tanaka et al., the fuzzy parameters are estimated according to certain conditions. One of them is that for a certain h -level, $0 \leq h \leq 1$, the support of the estimated values from the model should include support of the observed values. To determine the fuzzy coefficients while minimizing the total sum of the spreads of the estimated values for a certain h -level, formulated a linear programming problem, as follows

$$\begin{aligned} \text{Min} \quad & \sum_{i=1}^n \sum_{j=0}^p \alpha_j |X_{ij}| \\ \text{subject to} \quad & \sum_{j=0}^p a_j X_{ij} + (1-h) \sum_{j=0}^p \alpha_j |X_{ij}| \geq y_i + (1-h)e_i, \quad i = 1, 2, \dots, n, \\ & - \sum_{j=0}^p a_j X_{ij} + (1-h) \sum_{j=0}^p \alpha_j |X_{ij}| \geq -y_i + (1-h)e_i, \quad i = 1, 2, \dots, n, \\ & a_j \in R, \alpha_j \geq 0, \quad j = 0, 1, \dots, p. \end{aligned}$$

In this model, the constraints guarantee the support of the estimated values from the model includes the support of the observed values [15].

The model given above does not allow negative spreads. If the spreads do not take negative values, this causes wrong interpretation of the spreads and the centers. Chang and Lee defined the model, which allows negative spreads to overcome this misinterpretation [4].

Triangular fuzzy numbers are defined as $X = (m, \underline{m}, \bar{m})$, where m is the central value of X , \underline{m} is left spreads and \bar{m} is right spreads. When $X_i = (x_i, \underline{\xi}_i, \bar{\xi}_i)$ and $Y_i = (y_i, \underline{\eta}_i, \bar{\eta}_i)$, $i = 1, 2, \dots, n$ are triangular fuzzy numbers, fuzzy regression model is given by the equation

$$Y = a + bX, \tag{4}$$

where a and b are crisp numbers. When parameters are crisp, the least squares optimization problem is defined as

$$\text{Minimize } r(a, b) = \sum d(a + bX_i, Y_i)^2. \tag{5}$$

There are two situations according as $b \geq 0$ or $b < 0$ in Eq. (5). In case of $b \geq 0$, $d(a + bX_i, Y_i)^2$ is given by the equation

$$d(a + bX_i, Y_i)^2 = [a + bx_i - y_i - (b\underline{\xi}_i - \underline{\eta}_i)]^2 + [a + bx_i - y_i + (b\bar{\xi}_i - \bar{\eta}_i)]^2 + (a + bx_i - y_i)^2. \tag{6}$$

In Eq. (6), the parameters a, b are derived via $\frac{\partial r}{\partial a} = 0$ and $\frac{\partial r}{\partial b} = 0$.

For $i = 1, 2, \dots, n$, when x_i is crisp and $Y_i = (y_i, \underline{\eta}_i, \bar{\eta}_i)$ is triangular fuzzy number, fuzzy regression model is

$$Y = A + xB, \tag{7}$$

where $A = (a, \underline{\alpha}, \bar{\alpha})$ and $B = (b, \underline{\beta}, \bar{\beta})$ are fuzzy parameters. The parameters are estimated through the following equation, which is the fuzzy least squares optimization problem

$$\text{Minimize } r(A, B) = \sum d(A + x_i B, Y_i)^2, \tag{8}$$

where

$$d(A + x_i B, Y_i)^2 = (a + bx_i - y_i)^2 + (a + bx_i - \underline{\alpha} - \underline{\beta}x_i - y_i + \underline{\eta}_i)^2 + (a + bx_i + \bar{\alpha} + \bar{\beta}x_i - y_i - \bar{\eta}_i)^2.$$

Here, for $\frac{\partial r}{\partial a} = 0$ and $\frac{\partial r}{\partial b} = 0$, the parameters a and b obtained and similarly, for $\frac{\partial r}{\partial \alpha} = 0$, $\frac{\partial r}{\partial \beta} = 0$, parameters α and β are easily obtained [6,13,21,23].

4. The comparison when observed Y_i and estimated \hat{Y}_i are symmetrical triangular fuzzy number

In this part, the comparison of fuzzy numbers in [5] is adapted to fuzzy regression analysis. Observed $Y_i = (y_i, \underline{\eta}_i, \bar{\eta}_i)$ and estimated $\hat{Y}_i = (\hat{y}_i, \underline{\hat{\eta}}_i, \bar{\hat{\eta}}_i)$ are considered as two fuzzy numbers and the comparison of fuzzy numbers is defined for regression analysis.

When Y_i and \hat{Y}_i are symmetrical triangular fuzzy numbers, for the condition $\chi_1(w) + \chi_2(w) = 1$ and $\chi_1(w), \chi_2(w) \in (0, 1]$ if $\chi_1(w) = \chi_2(w) = \chi$ then according to index OM we obtained,

$$d(Y_i, \hat{Y}_i) = \int_0^{w_{\text{hgt}}} g_Y(\{\mu_Y^{-1}(w)\})dw - \int_0^{w_{\text{hgt}}} g_{\hat{Y}}(\{\mu_{\hat{Y}}^{-1}(w)\})dw = 2\chi(y_i - \hat{y}_i) \quad (9)$$

[17].

5. Hypotheses testing for fuzzy robust regression parameters

In regression analysis, many studies were done about parameter estimation in the event that outliers and were defined the robust estimators. In the last years, fuzzy estimations have been studied in most papers. The aim of this study is to make hypotheses testing for fuzzy robust regression parameters.

5.1. Fuzzy regression analysis when x_i is crisp and $Y_i = (y_i, \underline{\eta}_i, \bar{\eta}_i)$ is a triangular fuzzy number

The studies carried out when data are fuzzy are generally related to simple linear regression model estimates consisting one independent variable. A multi linear regression model consisting of more than one independent variable is dealt with and the optimization problem given in Eq. (7) is generalized for multi variable. In case that x is crisp and $Y = (y, \underline{\eta}, \bar{\eta})$ is a symmetrical triangular fuzzy number, the regression model with multi variables is $Y = A + B_1x_{1i} + \dots + B_px_{pi}$. In this case, optimization problem is defined as

$$P: \quad \text{Min } r(A, B_1, B_2, \dots, B_p) = \sum d(A + B_1x_{1i} + \dots + B_px_{pi}, Y_i)^2.$$

In multi regression analysis, in case those independent variables are crisp and the dependent variable is a symmetrical triangular fuzzy number and there is outlier in the data set, the steps of algorithm of the fuzzy robust regression method for estimated regression model are as follows:

Step 1: When independent variables x_i are crisp and the dependent variable $Y_i = (y_i, \underline{\eta}_i, \bar{\eta}_i)$ is a symmetrical triangular fuzzy number, initial fuzzy regression model estimation is found from problem P.

Step 2: Estimated \hat{y}_i and the residuals r_i are calculated. The residuals are found from Eq. (9).

Step 3: Median is determined with respect to the absolute residuals values and distances D_i are calculated,

$$D_i = \|\text{abs}(r_i) - \text{median}(\text{abs}(r_i))\|, \quad i = 1, 2, \dots, n,$$

where $\|\cdot\|$ is Euclidean distance.

Step 4: The membership function is defined as follows,

$$\mu(r) = \begin{cases} 1, & \text{abs}(r) \leq a \\ \frac{b - \text{abs}(r)}{b - a}, & a < \text{abs}(r) < b \\ 0, & \text{otherwise,} \end{cases} \quad (10)$$

where

$$a = \text{median}(D_i),$$

$$b = \max(D_i) + d.$$

Step 5: From the membership function is defined in Eq. (10), the membership values are determined and the weight matrix is constituted. The weight matrix is a diagonal matrix which diagonal elements are consisting of the degree of membership. Weighted fuzzy least squares estimations are found via the weight matrix obtained.

Step 6: If $|\hat{\beta}^{k+1} - \hat{\beta}^k| < \varepsilon$ then stop. Otherwise is going to Step 2. Where $\hat{\beta}$ is the estimates of regression model coefficients, k denote the iteration number and $\varepsilon > 0$ is a very small number [17]. Now, we define function ψ with respect to the weight function in Eq. (10) following as

$$\psi = \begin{cases} \text{abs}(r), & \text{abs}(r) \leq a \\ \frac{\text{abs}(r)b - (\text{abs}(r))^2}{b - a}, & a < \text{abs}(r) < b \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

Table 1

Data set.

i	X_1	X_2	X_3	(y_i, η_i)
1	7	26	6	(78.5, 6.9)
2	1	29	15	(74.3, 6.4)
3	11	56	8	(104.3, 9.4)
4	11	31	8	(87.6, 7.8)
5	7	52	6	(95.5, 8.6)
6	11	55	9	(109.2, 9.9)
7	3	71	17	(102.7, 9.3)
8	1	31	22	(60, 6.2)
9	2	54	18	(93.3, 8.3)
10	21	47	4	(115.9, 10.6)
11	1	40	23	(83.8, 7.4)
12	11	66	9	(113.3, 10.6)
13	10	68	8	(109.4, 9.9)

Table 2

Estimates of regression coefficients and standard deviations of coefficients.

Method	Constant	Regression coefficients		
		$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
LS	49.4442 (8.5282)	1.5253 (0.4458)	0.7248 (0.0964)	-0.1087 (0.4025)
Huber	47.9608 (4.4247)	1.7021 (0.2313)	0.6558 (0.0500)	0.2641 (0.2088)
Hampel	47.3509 (3.0715)	1.7757 (0.1606)	0.6249 (0.0347)	0.4268 (0.1450)
Tukey	47.4435 (3.1734)	1.7733 (0.1659)	0.6236 (0.0359)	0.4256 (0.1498)
Andrews	47.3269 (3.2869)	1.7811 (0.1718)	0.6255 (0.0372)	0.4277 (0.1551)
Diamond	(49.4442, 3.7962)	(1.5253, 0.1704)	(0.7248, 0.0668)	(-0.1087, 0.0234)
Fuzzy Robust Regression	(47.4154, 3.7376) (2.5812)	(1.7697, 0.1778) (0.1349)	(0.6308, 0.0638) (0.0292)	(0.4019, 0.0395) (0.1218)

Standard deviations for the estimated fuzzy robust regression model are obtained via robust standard deviations defined as Eq. (3) in Section 2.

6. Numerical examples

In this Section, we use an example, containing 13 observation, three independent variables and one dependent variable from Xu and Li [21], to make hypotheses testing on parameters. The 8th observation of the dependent variable is changed with 60 and obtained as an outlier. The data set is shown in Table 1. The M method and the fuzzy robust regression method were executed with a program written in Matlab. Estimates of regression coefficients and standard deviations of coefficients are located in Table 2.

In Table 2, it is given estimates of regression coefficients which are obtained via the LS method, the M method and the fuzzy robust regression method, and standard deviations of coefficients are enclosed in parentheses. Standard deviations for the estimated fuzzy robust regression coefficients have been calculated using the function in Eq. (11). The residuals in fuzzy robust regression method are calculated via OM index. When it is used OM index for ranking of the residuals, the each residual is equal to difference between centers of symmetrical triangular fuzzy numbers. Thus, standard deviations are calculated only for the centers. For the example, the data set contains 13 observation and 3 independent variables. In this case, for $\alpha = 0.05$, t -value is equal to 2.262. For the each parameter, calculated values are higher than this value. Thus, we reject null hypotheses at the level of significance α . According to standard deviations obtained via the function in Eq. (11), all parameters are significant.

7. Conclusion and discussion

In this study, when independent variables are crisp and the dependent variable is a symmetrical triangular fuzzy number and there is outlier in the data set, we suggested hypotheses testing for the fuzzy robust regression parameters. For this purpose, we define the function ψ with respect to the weight function, and calculated standard deviation estimations via the

function ψ . Moreover, significance of parameters is tested. When Table 2 is examined, it is seen that fuzzy robust regression model parameters are significant. Consequently, the defined function ψ can be used hypotheses testing for the fuzzy robust regression parameters when observations in multi regression analysis are fuzzy numbers and the outliers exist in the given data set.

References

- [1] Alpar R. Introduction to statistical methods with multi variables I. Bagirgan Publishing; 1997.
- [2] Bardossy A. Note on fuzzy regression. *Fuzzy Sets Syst* 1990;37:65–75.
- [3] Celmins A. Least squares model fitting to fuzzy vector data. *Fuzzy Sets Syst* 1987;22:245–69.
- [4] Chang PT, Lee ES. Fuzzy linear regression with spreads unrestricted in sign. *Comput Math Appl* 1994;28:61–70.
- [5] Chang PT, Lee ES. Ranking of fuzzy sets based on the concept of existence. *Comput Math Appl* 1994;27:1–21.
- [6] Diamond P. Fuzzy least squares. *Inform Sci* 1988;46:141–57.
- [7] Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA. Robust statistics. New-York: John-Wiley & Sons; 1986.
- [8] Hogg RV. Statistical robustness: one view of its use in applications today. *Am Stat* 1979;33:108–15.
- [9] Huber PJ. Robust statistics. John Willey & Son; 1981.
- [10] Huynh H. A comparison of four approaches to robust regression. *Psychol Bull* 1982;92:505–12.
- [11] Jozsef S. On the effect of linear data transformations in possibilistic fuzzy linear regression. *Fuzzy Sets Syst* 1992;45:185–8.
- [12] Kao C, Chyu CL. Least-squares estimates in fuzzy regression analysis. *Eur J Operational Res* 2003;148:426–35.
- [13] Modares M, Nasrabadi E, Nasrabadi MM. Fuzzy linear regression models with least square errors. *Appl Math Comput* 2005;163:977–89.
- [14] Nasrabadi MM, Nasrabadi E. A Mathematical-programming approach to fuzzy linear regression analysis. *Appl Math Comput* 2004;155:873–81.
- [15] Redden DT, Woddall WH. Further examination of fuzzy linear regression. *Fuzzy Sets Syst* 1996;79:203–11.
- [16] Stephenson D. 2000. Available from: <http://folk.uib.no/ngbnk/kurs/notes/node38.html>.
- [17] Şanlı K, Apaydin A. Fuzzy robust regression analysis based on the ranking of fuzzy sets. *Inernat. J. Uncertainty Fuzziness Knowledge-Based Syst* 2008;16:663–81.
- [18] Tanaka H, Uegima S, Asai K. Linear regression analysis with fuzzy model. *IEEE Trans Syst Man Cybern* 1982;12:903–7.
- [19] Tanaka H. Fuzzy data analysis by possibilistic linear models. *Fuzzy Sets Syst* 1987;24:363–75.
- [20] Tanaka H, Ishibuchi I. Identification of possibility linear systems by quadratic membership functions of fuzzy parameters. *Fuzzy Sets Syst* 1991;41:145–60.
- [21] Xu R, Li C. Multidimensional least-squares fitting with a fuzzy model. *Fuzzy Sets Syst* 2001;119:215–23.
- [22] Yang MS, Ko CH. On cluster-wise fuzzy regression analysis. *IEEE Trans Syst Man Cybern B Cybern* 1997;27:1–13.
- [23] Yun-Hsi OC. Hybrid fuzzy least-squares regression analysis and its reliability measures. *Fuzzy Sets Syst* 2001;119:225–46.