JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering,
Vol. 7, No. 4, December 2013, pp. 263-271

Robust Fuzzy Varying Coefficient Regression Analysis with Crisp Inputs and Gaussian Fuzzy Output

Zhihui Yang and Yunqiang Yin
College of Sciences, East China Institute of Technology, Fuzhou, Jiangxi, China
zhhyang75@gmail.com, yunqiangyin@gmail.com

Yizeng Chen*
School of Management, Shanghai University, Shanghai, China
mfcyz@shu.edu.cn

Abstract
This study presents a fuzzy varying coefficient regression model after deleting the outliers to improve the feasibility and effectiveness of the fuzzy regression model. The objective of our methodology is to allow the fuzzy regression coefficients to vary with a covariate, and simultaneously avoid the impact of data contaminated by outliers. In this paper, fuzzy regression coefficients are represented by Gaussian fuzzy numbers. We also formulate suitable goodness of fit to evaluate the performance of the proposed methodology. An example is given to demonstrate the effectiveness of our methodology.

Category: Smart and intelligent computing

Keywords: Gaussian fuzzy number; Goodness of fit; Outlier; Fuzzy varying coefficient regression

I. INTRODUCTION

As an important statistical analysis tool, the regression analysis model is often utilized to describe the statistical functional relationship between a response variable and a set of explanatory variables, so that the response variable can be predicted accordingly. The traditional regression model is anchored on binary logic. The sampling data used in traditional regression analysis has some strict assumptions: every observation is independent of others, the sampling data has a certain probability distribution, and so on. In actual practice, however, the description of the observations is often vague, and the data are often influenced by subjective judgment, or described in linguistic terms.

In recent years, fuzzy logic has become more widely used in statistical analysis. In their pioneering work, Tanaka et al. [1] employed fuzzy input data to establish a fuzzy regression analysis model. From then on the fuzzy regression model and its applications have attracted considerable attention from many fields, such as engineering, economics, management science, and environmental science. In the traditional regression model, the deviation between the experimental data and the model is interpreted as arising from the error of observation, but the fuzzy regression model views this kind of error as fuzziness of the structure in the system and the regression parameters.

Compared with the traditional regression model, the fuzzy linear regression model is inferior to the traditional

Open Access http://dx.doi.org/10.5626/JCSE.2013.7.4.263 http://jcse.kiise.org

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 16 May 2013, **Accepted** 20 August 2013
*Corresponding Author

Copyright © 2013. The Korean Institute of Information Scientists and Engineers pISSN: 1976-4677 eISSN: 2093-8020

linear regression model in terms of predictive capability, whereas their comparative descriptive performance depends on various factors associated with the data set and proper specificity of the model, especially for large sample data [2]. However, fuzzy linear regression performance becomes relatively better, as the size of the data set diminishes, and the aptness of the regression model deteriorates. Fuzzy linear regression may be used as an alternative to traditional linear regression in estimating regression parameters when the data is insufficient. Existing fuzzy regression models are mainly based on triangular fuzzy number, or trapezoidal fuzzy number. However, due to the nonlinear and complex nature of the relationship among variables in the system, the estimated effect of these fuzzy regression models needs to be further improved.

Following Tanaka et al. [1], many scholars have proposed all kinds of approaches to solve fuzzy regression model. There are two kinds of approaches to solving a fuzzy regression model. 1) The interval regression method: minimize the total spread of the fuzzy parameters, with the constraint that the membership of the estimate is not less than a predefined value [3-5]. This method essentially transforms a fuzzy regression problem into an optimization problem. 2) The fuzzy least square method: minimize the total square of the distances between the observation and estimated values of the response variables, and induce some equations similar to the normal equations of traditional regression analysis to obtain the fuzzy parameters [6-11].

Besides the above-mentioned two kinds of approaches, there are some other types of approaches to solve a fuzzy regression model. For example, the Monte Carlo method can be applied to the fuzzy regression model to obtain the optimal solution within a predetermined error bound [12, 13]. As a new classification technique proposed by Vapnik [14], the support vector machine (SVM) has been successful in solving pattern recognition and function estimation problems. Hong and Hwang [15] studied the convex optimization problem of a multi-fuzzy linear regression model via SVM. Hao and Chiang [16] employed fuzzy set theory to SVM in which the parameters, such as the components within the weight vector and the bias term, were set to be fuzzy numbers. By using different kernel functions, their method can achieve automatic accuracy control in the fuzzy regression analysis task. Assigning fuzzy membership values to data samples, Khemchandani et al. [17] proposed an approach to fuzzy support vector regression for financial time series forecasting. Wu and Law [18] proposed a new fuzzy SVM with the ability to penalize Gaussian noises in triangular fuzzy number space. Lin and Pai [19] employed support vector regressions to calculate fuzzy upper and lower bounds, to formulate a fuzzy SVM model for forecasting the indices of business cycles.

Previous fuzzy regression models find difficulty in dealing with input data varying with a covariate. Shen et al. [20] proposed a fuzzy varying coefficient model, where the fuzzy coefficients are allowed to vary with a covariate. The fuzzy varying coefficient regression model is an extension of the fuzzy linear regression model. Their method can improve the feasibility and adaptability of the fuzzy linear model. The procedure of their approach includes the following two steps: 1) After the fuzzy varying coefficient regression model is given, the proper kernel function is selected based on the distance of fuzzy numbers, and the cross-validation method is utilized to determine smooth parameters. 2) According to the definition of the distance between fuzzy numbers, the objective function is determined, and the estimate of response variables is obtained by the least squares method.

From the viewpoint of robust analysis, the least square method above is not robust enough, that is, when data contains individual abnormal data called outliers, the least squares estimate will not be reliable, and in the worst case, the conclusion may be incorrect. Therefore, Watada and Yabuuchi [21] suggested that we should remove the irregular data or outliers, before constructing the fuzzy regression model. Combined with robust analysis, the fuzzy regression model will be free from the influence of outliers. In this article, we integrate the fuzzy varying coefficient regression model with robust analysis to improve the feasibility and effectiveness of the fuzzy regression model.

The rest of the paper is organized as follows. Section II illustrates the fuzzy varying coefficients regression model with its fuzzy regression coefficients estimation. Section III introduces the notion of goodness of fit (GOF) for evaluating the model and the distance of fuzzy numbers, and employed them for robust analysis. In Section IV, a numerical example is used to demonstrate the effectiveness of the proposed methodology. Finally, the conclusions in this work are summarized in Section V.

II. FUZZY VARYING COEFFICIENT REGRESSION MODEL

A. Distance of Gaussian Fuzzy Numbers

Definition 1. *A fuzzy number $\tilde{A}$ is called a Gaussian fuzzy number, denoted by $\tilde{A} = (a, \sigma)$, if its membership function can be formulated as*

$$\tilde{A}(x) = e^{-\left(\frac{x-a}{\sigma}\right)^2}, \; x \in R, \; \sigma > 0 \tag{1}$$

where, a, σ is the center and spread of $\tilde{A}$, respectively.

Let $\tilde{A} = (a, \sigma)$, $\tilde{B} = (b, \tau)$ be Gaussian fuzzy numbers; if $a = b$ and $\sigma = \tau$, $\tilde{A}$ is equal to $\tilde{B}$, denoted as $\tilde{A} = \tilde{B}$. According to Zadeh's extension principle [22], the Gaussian fuzzy number has the following linear operations.

Proposition 1. *Let* $\tilde{A} = (a, \sigma)$, $\tilde{B} = (b, \tau)$ *be Gaussian fuzzy numbers, then*

1) $\tilde{A} + \tilde{B} = (a+b, \sigma+\tau)$,

2) $k\tilde{A} = (ka, k\sigma), \forall k \in R, k>0$.

In order to characterize the nearness degree of Gaussian fuzzy numbers $\tilde{A}$ and $\tilde{B}$, we use the following distance, proposed by Xu [23]:

$$d(\tilde{A}, \tilde{B}) = \left(\int_0^1 f(\lambda) d^2(A_\lambda, B_\lambda) d\lambda\right)^{1/2} \tag{2}$$

where, $A_\lambda = [a_1(\lambda), a_2(\lambda)] = [a - \sigma\sqrt{-\ln\lambda}, a + \sigma\sqrt{-\ln\lambda}]$, (3)

$$B_\lambda = [b_1(\lambda), b_2(\lambda)] = [b - \tau\sqrt{-\ln\lambda}, b + \tau\sqrt{-\ln\lambda}], \tag{4}$$

$$\begin{aligned} d^2(A_\lambda, B_\lambda) &= (a_1(\lambda) - b_1(\lambda))^2 + (a_2(\lambda) - b_2(\lambda))^2 \\ &= 2((a-b)^2 - (\sigma-\tau)^2 \ln\lambda), \end{aligned} \tag{5}$$

$f(\lambda)$ is a monotonically increasing function on the interval [0,1], with $f(0) = 0$, and $\int_0^1 f(\lambda) d\lambda = \frac{1}{2}$. The degree of nearness and overlap between the λ-level sets A_λ and B_λ, denoted by $d(A_\lambda, B_\lambda)$, is weighted by $f(\lambda)$, to emphasize the contribution of the higher values of λ to the distance between $\tilde{A}$ and $\tilde{B}$.

In this article, we set $f(\lambda) = 2\lambda^3$, instead of $f(\lambda) = \lambda$, for emphasizing the contribution of λ; thus, we can draw the conclusion as follows.

$$d(\tilde{A}, \tilde{B}) = \left((a-b)^2 + \frac{1}{4}(\sigma-\tau)^2\right)^{1/2} \tag{6}$$

B. Fuzzy Varying Coefficient Regression Model

In order to describe the dynamic relationship between the response variable and a set of explanatory variables, Shen et al. [20] proposed a fuzzy varying coefficient regression model with its estimation. The procedure includes three steps as follows.

Step 1: Construct the fuzzy varying coefficient regression model.

Step 2: Determine the optimal value of smooth parameters by using the fuzzy cross-validation method.

Step 3: According to the distance of fuzzy numbers, determine the objective function and obtain the estimate of response variables by the restricted least squares method.

The fuzzy varying coefficient regression model proposed by Shen et al. [20] is formulated as follows:

$$\tilde{Y} = \tilde{A}_1(t)X_1 + \tilde{A}_2(t)X_2 + \cdots + \tilde{A}_m(t)X_m \tag{7}$$

where, X_j, $j = 1, 2, ..., m$ and t are explanatory variables (input variables) expressed by crisp numbers, $\tilde{A}_j(t) = (\alpha_j(t), \sigma_j(t))$, $j = 1, 2, ..., m$ are Gaussian fuzzy numbers varying with the variable t, and $\alpha_j(t)$, $\sigma_j(t)$ are the center and spread of $\tilde{A}_j(t)$, respectively. According to the linear operation of Gaussian fuzzy numbers, the response variable $\tilde{Y}$ is also a Gaussian fuzzy number, denoted by $\tilde{Y} = (Y, S)$, where $Y = \sum_{j=1}^m \alpha_j(t)X_j$, $S = \sum_{j=1}^m \sigma_j(t)X_j$. Generally, we take $X_1 \equiv 1$, to make the model include a fuzzy varying intercept.

Suppose that there are the experimental data set of the response variable $\tilde{Y}$ and the set of explanatory variables $X_j, j = 1, 2, ..., m$: $\{(y_1, s_1, t_1, \vec{x}_1), (y_2, s_2, t_2, \vec{x}_2), ..., (y_i, s_i, t_i, \vec{x}_i), ..., (y_n, s_n, t_n, \vec{x}_n)\}$, where, $\vec{x}_i = (x_{i1}, x_{i2}, ..., x_{ij}, ..., x_{im})$ is the i-th experimental data set of the explanatory variable presented by crisp numbers; the Gaussian fuzzy number (y_i, s_i) is the i-th observation of the response variable, $i = 1, 2, ..., n$.

Therefore, the sample form of the model Eq. (7) is

$$\begin{cases} y_i = \sum_{j=1}^m \alpha_j(t_i) x_{ij} \\ s_i = \sum_{j=1}^m \sigma_j(t_i) x_{ij} \end{cases}, \; i = 1, 2, ..., n \tag{8}$$

Now we should estimate the fuzzy regression coefficient $\tilde{A}_j(t_0)$, $j = 1, 2, ..., m$, for any $t = t_0$. On the basis of the distance proposed above and the principle of kernel smoothing in statistics, the restricted weighted least-squares problem is formulated as follows:

$$\begin{aligned} &g(\alpha_1(t_0), \alpha_2(t_0), ..., \alpha_m(t_0); \sigma_1(t_0), \sigma_2(t_0), ..., \sigma_m(t_0)) \\ &= \sum_{i=1}^n d^2\left((y_i, s_i), \left(\sum_{j=1}^m \alpha_j(t_0)x_{ij}, \sum_{j=1}^m \sigma_j(t_0)x_{ij}\right)\right) K_h(t_i - t_0) \\ &= \sum_{i=1}^n \left(\left(y_i - \sum_{j=1}^m \alpha_j(t_0)x_{ij}\right)^2 + \frac{1}{4}\left(s_i - \sum_{j=1}^m \sigma_j(t_0)x_{ij}\right)^2\right) K_h(t_i - t_0) \\ &= \sum_{i=1}^n \left(y_i - \sum_{j=1}^m \alpha_j(t_0)x_{ij}\right)^2 K_h(t_i - t_0) + \frac{1}{4}\sum_{i=1}^n \left(s_i - \sum_{j=1}^m \sigma_j(t_0)x_{ij}\right)^2 K_h(t_i - t_0) \\ &\sigma_j(t_0) \geq 0, j = 1, 2, ..., m \end{aligned} \tag{9}$$

where, $K_h(t) = K(t/h)/h$, with $K(x) = (2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}x^2)$ being a Gaussian kernel function, and h being the smoothing parameter determined by the fuzzy cross-validation procedure.

The restricted weighted least-squares problem above is equivalent to minimizing two formulas as follows:

$$g_1 = \sum_{i=1}^n \left(y_i - \sum_{j=1}^m \alpha_j(t_0)x_{ij}\right)^2 K_h(t_i - t_0) \tag{10}$$

and

$$\begin{cases} \sum_{i=1}^n \left(s_i - \sum_{j=1}^m \sigma_j(t_0)x_{ij}\right)^2 K_h(t_i - t_0) \\ \sigma_j(t_0) \geq 0, j = 1, 2, ..., m \end{cases} \tag{11}$$

In order to minimize Eq. (10), we set to zero the partial derivatives of g_1 with respect to α_k as follows:

$$\begin{aligned}
&\frac{\partial}{\partial \alpha_k}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{m}\alpha_j(t_0)x_{ij}\right)^2 K_h(t_i-t_0)\\
&=\frac{\partial}{\partial \alpha_k}\sum_{i=1}^{n}(y_i-\alpha_1(t_0)x_{i1}-\cdots-\alpha_m(t_0)x_{im})^2K_h(t_i-t_0) \quad , k=1,2,...,m,\\
&=2\sum_{i=1}^{n}(-x_{ik})(y_i-\alpha_1(t_0)x_{i1}-\cdots-\alpha_m(t_0)x_{im})K_h(t_i-t_0)\\
&=0 \qquad (12)
\end{aligned}$$

The equations above are equivalent to the following equations:

$$\begin{cases}
\sum_{k=1}^{m}\sum_{i=1}^{n}x_{i1}x_{ik}K_h(t_i-t_0)\alpha_k(t_0)=\sum_{i=1}^{n}x_{i1}K_h(t_i-t_0)y_i\\
\sum_{k=1}^{m}\sum_{i=1}^{n}x_{i2}x_{ik}K_h(t_i-t_0)\alpha_k(t_0)=\sum_{i=1}^{n}x_{i2}K_h(t_i-t_0)y_i\\
\cdots\cdots\\
\sum_{k=1}^{m}\sum_{i=1}^{n}x_{im}x_{ik}K_h(t_i-t_0)\alpha_k(t_0)=\sum_{i=1}^{n}x_{im}K_h(t_i-t_0)y_i
\end{cases} \qquad (13)$$

Let $X=(x_{ij})_{n\times m}$, $Y=(y_1, y_2, ..., y_n)^T$, $S=(s_1, s_2, ..., s_n)^T$,
$W(t_0)=diag(K_h(t_1-t_0), K_h(t_2-t_0), ..., K_h(t_n-t_0))$,
$\alpha(t_0)=(\alpha_1(t_0), \alpha_2(t_0), ..., \alpha_m(t_0))^T$,
$\sigma(t_0)=(\sigma_1(t_0), \sigma_2(t_0), ..., \sigma_m(t_0))^T$.

Assuming that the inverse of $X^TW(t_0)X$ always exists for any t_0, the estimate of $\alpha(t_0)$ will be solved as follows:

$$\begin{aligned}
\hat{\alpha}(t_0)&=(\alpha_1(t_0), \alpha_2(t_0), ..., \alpha_m(t_0))^T\\
&=(X^TW(t_0)X)^{-1}X^TW(t_0)Y
\end{aligned} \qquad (14)$$

Similarly, the estimation of $\sigma(t_0)$ will be obtained without positive restriction, as follows:

$$\begin{aligned}
\hat{\sigma}(t_0)&=(\sigma_1(t_0), \sigma_2(t_0), ..., \sigma_m(t_0))^T\\
&=(X^TW(t_0)X)^{-1}X^TW(t_0)S
\end{aligned} \qquad (15)$$

Considering the positive restriction of the spread, Shen et al. [20] suggested that we utilize the method proposed by D'Urso et al. [24].

Let $X_i=(x_{i1}, x_{i2}, ..., x_{im})^T$, performing the estimate procedure above at $t_0=t_1, t_2, ..., t_n$ respectively, we can obtain the following fitted values of the center and spread of $\tilde{Y}$:

$$\begin{aligned}
&\hat{y}_i=X_i^T(X_i^TW(t_i)X_i)^{-1}X_i^TW(t_i)Y\\
&\hat{s}_i=X_i^T(X_i^TW(t_i)X_i)^{-1}X_i^TW(t_i)S\\
&i=1, 2, ..., n
\end{aligned} \qquad (16)$$

C. Determination of the Smoothing Parameter

The role of the smoothing parameter h is to adjust the degree of smoothness of the estimates of the center and spread of the fuzzy regression coefficients. The fuzzy cross-validation procedure can be used to select the optimal value of the smoothing parameter: suppose the number of data is n, for each $i=1, 2, ..., n$, remove the i-th observation $\tilde{y}_i=(y_i, s_i)$, and compute the estimates of the center and spread of the fuzzy coefficients $\tilde{A}_j(t)$ $(j=1, 2, ..., n)$ at $t=t_i$, according to the procedure described above. Let $\hat{\alpha}_j^{(-i)}(t_i, h)$ and $\hat{\sigma}_j^{(-i)}(t_i, h)$ be the estimates of the centers and spreads of the fuzzy coefficients under h, then the predicted values of the center and spread of the fuzzy response $\tilde{Y}$ at t_i can be obtained by the following formulations, respectively:

$$\hat{y}_{(-i)}(h)=\sum_{j=1}^{m}\hat{\alpha}_j^{(-i)}(t_i, h)x_{ij}, \ \hat{s}_{(-i)}(h)=\sum_{j=1}^{m}\hat{\sigma}_j^{(-i)}(t_i, h)x_{ij}, \qquad (17)$$

The cross-validation (CV) score is formulated as follows:

$$\begin{aligned}
CV(h)&=\sum_{i=1}^{n}d^2((y_i, s_i), (\hat{y}_{(-i)}(h), \hat{s}_{(-i)}(h)))\\
&=\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{m}\hat{\alpha}_j^{(-i)}(t_i, h)x_{ij}\right)^2+\frac{1}{4}\sum_{i=1}^{n}\left(s_i-\sum_{j=1}^{m}\hat{\sigma}_j^{(-i)}(t_i, h)x_{ij}\right)^2
\end{aligned} \qquad (18)$$

Then, we will select h_0 as the optimal value of the smoothing parameter, such that

$$CV(h_0)=\min_{h>0}CV(h) \qquad (19)$$

There are two main advantages to using the fuzzy varying coefficient regression model: 1) Because the fuzzy regression coefficients may vary with another explanatory variable, the flexibility and adaptability of the fuzzy regression model are enhanced. 2) The model can deal with data that vary with a time variable, and establish a dynamic relationship between a response variable and a set of explanatory variables.

III. ROBUST ANALYSIS AND GOF OF GAUSSIAN FUZZY NUMBERS

A. Robust Analysis

From the robustness point of view, the least square method is not robust, that is, when data contains an individual abnormal value or outliers, the least square estimate will not be reliable and a wrong conclusion may even be drawn. Therefore, diagnostic checks should be

done on data, before applying the least square method.

Definition 2. *An observation that has a bigger residual value than the others is called an outlier.*

The least square estimator is sensitive to outliers. Therefore, the estimation results are directly affected by each observation, and the data should be analyzed in detail. Sometimes, even a single observation may dramatically influence the value of the parameter estimates, and omitting this observation from the data may lead to totally different results. To handle the outlier problem, Hung and Yang [25] proposed an omission approach for Tanaka's linear programming method. This approach has the capability to examine the behavior of value changes in the objective function of fuzzy regression models, when observations are omitted. On the basis of the Least Median Squares-Weighted Least Squares estimation procedure, D'Urso et al. [24] proposed a robust fuzzy linear regression model to deal with data contaminated by outliers. To overcome the higher order or interactive terms, and the influence of outliers existing in the manufacturing process data, Chan et al. [26] integrated genetic programming with the fuzzy regression model.

Next, we introduce the M estimation methods introduced by Huber [27], which are widely used for robust regression. M estimation can be regarded as a generalization of maximum-likelihood estimation. The general M estimator minimizes the following objective function

$$\upsilon = \sum_{i=1}^{n} \rho(y_i - \textstyle\sum_{j=1}^{m} x_{ij}\hat{\alpha}_j)/d) \tag{20}$$

where, the function ρ gives the contribution of each residual to the objective function.

By setting to zero the partial derivative of υ with respect to $\hat{\alpha}_j$, we have *m* equations as follows:

$$\sum_{i=1}^{n} x_{ij}\psi(y_i - \textstyle\sum_{j=1}^{m} x_{ij}\hat{\alpha}_j)/d) = 0\,,\ j = 1,\ 2,\ \ldots,\ m\,. \tag{21}$$

where, $\psi(z) = \rho'(z)$ is the derivative of ρ. The standardized residuals may be defined as $z = r_i/d$, where $r_i = y_i - \sum_{j=1}^{m} x_{ij}\hat{\alpha}_j$. d is a robust estimate of scale. Under normality, the expected value of d is the standard error of estimate in the population.

In this article, we will integrate robustness analysis with the fuzzy varying coefficient regression model. The main procedure is as follows:

Step 1: Utilize the robustfit function of Matlab toolbox to perform robust analysis and remove the outliers, according to the result of the robustness analysis.

Step 2: Construct a fuzzy varying coefficient regression model and use the fuzzy cross-validation method to determine the optimal value of the smoothing parameter.

Step 3: Determine the objective function according to the definition of distance Eq. (6), and solve the least squares problem for the parameter estimation of response variables.

Step 4: Evaluate our approach, by comparing it with Shen et al.'s approach [20] according to AGOF, which will be introduced in Subsection III-B.

B. GOF of Gaussian Fuzzy Numbers

In order to evaluate the fit performance of our approach, we next introduce an index named GOF for describing the nearness and overlap of two Gaussian fuzzy numbers. For two Gaussian fuzzy numbers $\tilde{A} = (a,\ \sigma)$, $\tilde{B} = (b,\ \tau)$, the GOF for them denoted by $GOF(\tilde{A},\ \tilde{B})$ should be a strict monotonically decreasing function with respect to the value of $|b-a|$, and be equal to 1 if and only if $\tilde{A} = \tilde{B}$.

We utilize the proportion of the shadow area to the maximum of the area shaped by $\tilde{A}(x)$ and $\tilde{B}(x)$ to be $GOF(\tilde{A},\ \tilde{B})$. $\int_{-\infty}^{+\infty} e^{-\left(\frac{x-a}{\sigma}\right)^2} dx = \sqrt{\pi}\sigma$, $\int_{-\infty}^{+\infty} e^{-\left(\frac{x-b}{\tau}\right)^2} dx = \sqrt{\pi}\tau$ are the areas under the curve of membership functions $\tilde{A}(x)$ and $\tilde{B}(x)$, respectively. Therefore, we define $GOF(\tilde{A},\ \tilde{B})$ as follows.

Definition 3. *Let* $\tilde{A} = (a,\ \sigma)$, $\tilde{B} = (b,\ \tau)$ *be Gaussian fuzzy numbers; the GOF for* $\tilde{A}$ *and* $\tilde{B}$ *is defined as*

$$GOF(\tilde{A},\ \tilde{B}) = \frac{(\sigma+\tau)(1-\phi(\sqrt{2}|b-a|/(\sigma+\tau)))}{\max(\sigma,\ \tau)} \tag{23}$$

where, $\phi(\cdot)$ is the standard normal distribution function.

For simplicity of calculation, we define another GOF for $\tilde{A}$ and $\tilde{B}$ used in Section IV as follows:

Definition 4. *Let* $\tilde{A} = (a,\ \sigma)$, $\tilde{B} = (b,\ \tau)$ *be Gaussian fuzzy numbers; the GOF for* $\tilde{A}$ *and* $\tilde{B}$ *is defined as*

$$GOF(\tilde{A},\ \tilde{B}) = \begin{cases} \exp\left(-\left(\dfrac{a-b}{\sigma+\tau}\right)^2\right), & a \neq b, \\ \dfrac{\min\{\sigma,\ \tau\}}{\max\{\sigma,\ \tau\}}, & a = b. \end{cases} \tag{24}$$

In Section IV, for the purpose of evaluating the performance of our approach, we will use AGOF (average of all GOF for the observations with their estimates) to compare our approach with the approach proposed by Shen et al. [20].

IV. AN ILLUSTRATED EXAMPLE

Gross domestic product (GDP) is the market value of all officially recognized final goods and services produced within a country in a given period. In economics,

GDP is a sum of consumption, investment, government spending, and exports. In this article, we think that the dominant explanatory variables to determine GDP are the total amount of urban and rural savings deposit, the total financial expenditure, fixed assets investment, and sum of imports and exports, which are denoted by X_1, X_2, X_3, X_4, respectively. By using the fuzzy varying coefficient regression model combined with robustness analysis, we remove the outliers by utilizing the robustfit function in Matlab, and construct the fuzzy regression varying regression model for fitting the response (GDP).

The dataset shown in Table 1 consists of 29 observations of the total amount of urban and rural savings deposit, the total financial expenditure, fixed assets investment, sum of imports and exports, and GDP of China from 1981 to 2009, in which the GDP is assumed to be a Gaussian fuzzy number.

First, we utilize the robustfit function in Matlab to analyze the robustness of this dataset; the diagram of residual errors is given in Fig. 1.

The residual of the 27th observation is 3.54×10^4, which is obviously much larger than that of the other observations, so we think it is an outlier and should be removed. Next we will construct the fuzzy varying coef-

Table 1. Observations of gross domestic product (GDP) from 1981 to 2009 (unit: CNY 10,000)

Year	GDP	X_1	X_2	X_3	X_4
1981	(4891.6, 407.63)	523.3	1138.41	961	735.3
1982	(5323.4, 443.62)	675.4	1229.98	1230.4	771.3
1983	(5962.7, 496.89)	892.5	1409.52	1430.1	860.1
1984	(7208.1, 600.68)	1214.7	1701.02	1832.9	1201
1985	(9016.04, 751.33)	1622.60	2004.25	2543.20	2066.70
1986	(10275.18, 856.26)	2237.60	2204.91	3120.60	2580.40
1987	(12058.62, 1004.88)	3073.30	2262.18	3791.70	3084.20
1988	(15042.82, 1253.56)	3801.50	2491.21	4753.80	3821.80
1989	(16992.32, 1416.02)	5146.90	2823.78	4410.40	4155.90
1990	(18667.82, 1555.65)	7119.80	3083.59	4517.00	5560.10
1991	(21781.50, 1815.12)	9241.60	3386.62	5594.50	7225.80
1992	(26923.48, 2243.62)	11759.40	3742.20	8080.10	9119.60
1993	(35333.92, 29611.16)	15203.50	4642.30	13072.30	11271.00
1994	(48197.86, 4016.48)	21518.80	5792.62	17042.10	20381.90
1995	(60793.73, 5066.14)	29662.30	6823.72	20019.30	23499.90
1996	(71176.59, 5931.38)	38520.80	7937.55	22913.50	24133.80
1997	(78973.03, 6581.08)	46279.80	9223.56	24941.10	26967.20
1998	(84402.28, 7033.52)	53407.50	10798.18	28406.20	26849.70
1999	(89677.05, 7473.08)	59621.80	13187.67	29854.70	29896.30
2000	(99214.55, 8267.87)	64332.40	15886.50	32917.70	39273.20
2001	(109655.17, 9137.93)	73762.40	18902.58	37213.50	42183.60
2002	(120332.69, 10027.72)	86910.60	22053.15	43499.90	51378.20
2003	(135822.76, 11318.56)	103617.30	24649.95	55566.60	70483.50
2004	(159878.34, 13323.19)	119555.39	28486.89	70477.40	95539.10
2005	(184937.37, 15411.44)	141051.00	33930.28	88773.60	116921.80
2006	(216314.43, 18026.20)	161587.30	40422.73	109998.20	140971.45
2007	(265810.31, 22150.85)	172534.19	49781.35	137323.9	166740.19
2008	(314045.43, 26170.45)	217885.35	62592.66	172828.4	179921.47
2009	(340506.87, 28375.57)	260771.66	76299.93	224598.8	150648.06

ficient regression model after removing the 27th observation, and estimate the parameters.

Next, we set the variable t to be the time order of each year and consider the following fuzzy varying coefficient model:

$$\tilde{Y} = \alpha_0(t) + \alpha_1(t)X_1 + \alpha_2(t)X_2 + \alpha_3(t)X_3 + \alpha_4(t)X_4 . \quad (25)$$

With the selected optimal value of $h = 2.16$ obtained by the fuzzy cross-validation procedure, we calibrate model Eq. (25) with the restricted weighted least-squares procedure.

For comparison, the fit values of Shen et al.'s method and our method are shown in the third and fifth column in Table 2, respectively, with their GOF shown in the fourth and sixth column, respectively. As is noticed, the AGOF of Shen et al.'s method is 0.8943, while the AGOF of our method is 0.9934. The larger value of AGOF for our method demonstrates that the fit performance of our approach is better than the approach proposed by Shen et al. [20].

Table 2. Comparison between the observations and estimates of GDP (unit: CNY 10,000)

Year	GDP	Shen et al.'s approach [20]		Our approach	
		Estimation	GOF	Estimation	GOF
1981	(4891.6, 407.63)	(4885, 405)	0.9989	(4880, 407)	0.9999
1982	(5323.4, 443.62)	(5341, 445)	0.9996	(5340, 445)	0.9996
1983	(5962.7, 496.89)	(60047, 501)	0.9972	(6000, 500)	0.9982
1984	(7208.1, 600.68)	(72018, 599)	0.9999	(7200, 600)	0.9999
1985	(9016.04, 751.33)	(88316, 632)	0.9824	(8830, 736)	0.9847
1986	(10275.18, 856.26)	(10555, 614)	0.9644	(10560, 880)	0.9743
1987	(12058.62, 1004.88)	(132880, 26132)	0.9606	(12290, 1024)	0.9868
1988	(15042.82, 1253.56)	(167876, 30033)	0.9798	(14510, 1209)	0.9541
1989	(16992.32, 1416.02)	(177028, 28661)	0.8454	(16560, 1380)	0.976
1990	(18667.82, 1555.65)	(196617, 29140)	0.9728	(18760, 1563)	0.9992
1991	(21781.50, 1815.12)	(224375, 31662)	0.9518	(22060, 1839)	0.9941
1992	(26923.48,2243.62)	(265967, 37339)	0.9828	(27420, 2285)	0.9881
1993	(35333.92,29611.16)	(330949, 46544)	0.9970	(35220, 2935)	0.9996
1994	(48197.86,4016.48)	(417044, 54488)	0.9957	(48910, 4076)	0.9923
1995	(60793.73, 5066.14)	(516352, 62273)	0.6246	(60360, 5030)	0.9981
1996	(71176.59, 5931.38)	(621982, 70021)	0.5181	(70030, 5836)	0.9906
1997	(78973.03, 6581.08)	(708413,74129)	0.6176	(79260, 6605)	0.9995
1998	(84402.28, 7033.52)	(796136, 79467)	0.7134	(84400, 7033)	0.9999
1999	(89677.05, 7473.08)	(853313, 75548)	0.9029	(91160, 7597)	0.9904
2000	(99214.55, 8267.87)	(903645, 72462)	0.9198	(98240, 8187)	0.9965
2001	(109655.17, 9137.93)	(101023, 74206)	0.7222	(108810, 9068)	0.9979
2002	(120332.69, 10027.72)	(116935, 82586)	0.7621	(120620, 10051)	0.9998
2003	(135822.76, 11318.56)	(141182, 10929)	0.9661	(136160, 11346)	0.9998
2004	(159878.34, 13323.19)	(165663, 13546)	0.9436	(158630, 13219)	0.9978
2005	(184937.37, 15411.44)	(197172, 16539)	0.9547	(184930, 15411)	0.9999
2006	(216314.43, 18026.20)	(228924, 19577)	0.8636	(216990, 18083)	0.9996
2007	**(265810.31, 22150.85)**	**(252604, 21771)**	**0.8936**	**(238820, 19902)**	**0.66**
2008	(314045.43, 26170.45)	(315815, 26736)	0.9136	(313960, 26163)	0.9999
2009	(340506.87, 28375.57)	(371546, 32964)	0.9889	(367480, 30623)	0.9998

GDP: gross domestic product, GOF: goodness of fit.

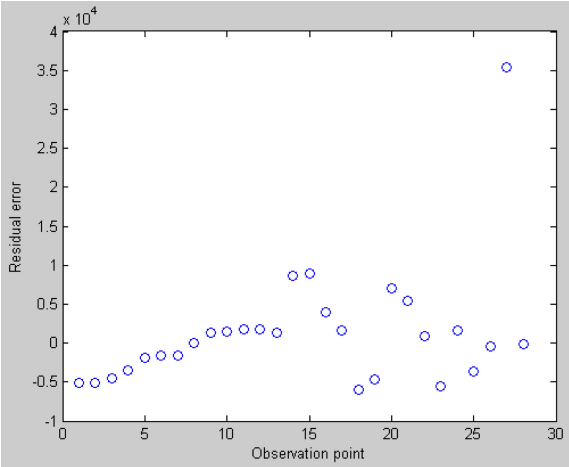

Fig. 1. Scatter diagram of robust analysis residual.

V. CONCLUSION

The aim of our study is to propose a methodology for dealing with the dynamic fuzzy function relationships between the response variable and a set of explanatory variables, while simultaneously avoiding the impact of outliers in the dataset. Our methodology can improve the feasibility and effectiveness of the fuzzy regression model. Specifically, the contribution of this work can be summarized in two points. First, it has provided a dynamic model to deal with the data varying with a covariate, especially for the sampling data having approximately a Gaussian contribution. Second, with the help of robust analysis, the proposed model is free from the irregular data or outliers, after removing the outliers. From the robustness standpoint, a suitable index for characterizing the outliers needs to be studied to improve the robustness of the fuzzy regression model in future work.

ACKNOWLEDGMENTS

The work described in this paper was supported by a grant from the National Nature Science Foundation of Chinese (project no. NSFC71272177), the funds of "Innovation Program of Shanghai Municipal Education Commission, China (project no. 12ZS101)."

REFERENCES

1. H. Tanaka, S. Uejima, and K. Asai, "Linear regression analysis with fuzzy model," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 12, no. 6, pp. 903-907, 1982.
2. K. J. Kim, H. Moskowitz, and M. Koksalan, "Fuzzy versus statistical linear regression," *European Journal of Operational Research*, vol. 92, no. 2, pp. 417-434, 1996.
3. Y. Chen, J. Tang, R. Y. K. Fung, and Z. Ren, "Fuzzy regression-based mathematical programming model for quality function deployment," *International Journal of Production Research*, vol. 42, no. 5, pp. 1009-1027, 2004.
4. H. Tanaka, "Fuzzy data analysis by possibilistic linear models," *Fuzzy Sets and Systems*, vol. 24, no. 3, pp. 363-375, 1987.
5. H. Tanaka and J. Watada, "Possibilistic linear systems and their application to the linear regression model," *Fuzzy Sets and Systems*, vol. 27, no. 3, pp. 275-289, 1988.
6. P. Diamond, "Fuzzy least squares," *Information Sciences*, vol. 46, no. 3, pp. 141-157, 1988.
7. P. D'urso and A. Santoro, "Goodness of fit and variable selection in the fuzzy multiple linear regression," *Fuzzy Sets and Systems*, vol. 157, no. 19, pp. 2627-2647, 2006.
8. M. B. Ferraro, R. Coppi, G. Gonzalez Rodriguez, and A. Colubi, "A linear regression model for imprecise response," *International Journal of Approximate Reasoning*, vol. 51, no. 7, pp. 759-770, 2010.
9. S. M. Taheri and M. Kelkinnama, "Fuzzy linear regression based on least absolutes deviations," *Iranian Journal of Fuzzy Systems*, vol. 9, no. 1, pp. 121-140, 2012.
10. M. S. Waterman, "A restricted least squares problem," *Technometrics*, vol. 16, no. 1, pp. 135-136, 1974.
11. H. C. Wu, "The construction of fuzzy least squares estimators in fuzzy linear regression models," *Expert Systems with Applications*, vol. 38, no. 11, pp. 13632-13640, 2011.
12. A. Abdalla and J. J. Buckley, "Monte Carlo methods in fuzzy linear regression I," *Soft Computing*, vol. 11, no. 10, pp. 991-996, 2007.
13. A. Abdalla and J. J. Buckley, "Monte Carlo methods in fuzzy linear regression II," *Soft Computing*, vol. 12, no. 5, pp. 463-468, 2007.
14. V. N. Vapnik, *The Nature of Statistical Learning*, New York, NY: Springer, 1995.
15. D. H. Hong and C. Hwang, "Support vector fuzzy regression machines," *Fuzzy Sets and Systems*, vol. 138, no. 2, pp. 271-281, 2003.
16. P. Y. Hao and J. H. Chiang, "Fuzzy regression analysis by support vector learning approach," *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 2, pp. 428-441, 2008.
17. R. Khemchandani, Jayadeva, and S. Chandra, "Regularized least squares fuzzy support vector regression for financial time series forecasting," *Expert Systems with Applications*, vol. 36, no. 1, pp. 132-138, 2009.
18. Q. Wu and R. Law, "Fuzzy support vector regression machine with penalizing Gaussian noises on triangular fuzzy number space," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7788-7795, 2010.
19. K. P. Lin and P. F. Pai, "A fuzzy support vector regression model for business cycle predictions," *Expert Systems with Applications*, vol. 37, no. 7, pp. 5430-5435, 2010.
20. S. L. Shen, C. L. Mei, and J. L. Cui, "A fuzzy varying coefficient model and its estimation," *Computers and Mathematics with Applications*, vol. 60, no. 6, pp. 1696-1705, 2010.
21. J. Watada and Y. Yabuuchi, "Fuzzy robust regression analysis," in *Proceeding of the 3rd IEEE International Conference on Fuzzy Systems*, Orlando, FL, 1994, pp. 1370-1376.
22. L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8,

no. 3, pp. 338-353, 1965.
23. R. N. Xu, "A linear regression model in fuzzy environment," *Advance in Modelling Simulation*, vol. 27, pp. 31-40, 1991.
24. P. D'Urso, R. Massari, and A. Santoro, "Robust fuzzy regression analysis," *Information Sciences*, vol. 181, no. 19, pp. 4154-4174, 2011.
25. W. L. Hung and M. S. Yang, "An omission approach for detecting outliers in fuzzy regressions model," *Fuzzy Sets and Systems*, vol. 157, no. 23, pp. 3109-3122, 2006.
26. K. Y. Chan, C. K. Kwong, and T. C. Fogarty, "Modeling manufacturing processes using a genetic programming-based fuzzy regression with detection of outliers," *Information Sciences*, vol. 180, no. 4, pp. 506-518, 2010.
27. P. J. Huber, "Robust estimation of a location parameter," *Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73-101, 1964.

Zhihui Yang

Zhihui Yang is currently a Ph.D. student in the School of Management at Shanghai University and an associate professor at the East China Institute of China. He received his B.S. from Nanchang University, Nanchang, China, in 1996, and M.S. from the East China Institute of China, Fuzhou, Jiangxi, China, in 2004. He has worked at the East China Institute of Technology since 1996. His research interests are in fuzzy sets, fuzzy program modeling & optimization and its applications in industrial engineering. He has published about 20 papers.

Yizeng Chen

Yizeng Chen received his B.S. from the Institute of Shenyang Aerospace, Shenyang, China, in 1993, M.S. from Beihang University, Beijing, China, in 2000, and Ph.D. from Northeastern University, Shenyang, China, in 2003. He has worked at Shanghai University since 2008. His research interests are in fuzzy program modeling & optimization, and its applications in industrial engineering. He has published about 40 papers.

Yunqiang Yin

Yunqiang Yin received his B.S. from Shandong University of Science and Technology, Taian, China, in 2003, M.S. from Kunming University of Science and Technology, Kunming, China, in 2006, and Ph.D. from Beijing Normal University, Beijing, China, in 2009. He has worked at the East China Institute of Technology since 2009. His research interests are in discrete optimization, scheduling, algebraic hyperstructure theory, fuzzy sets, and rough sets. He has published about 100 papers, and a book on fuzzy hemirings.