ELSEVIER

Fuzzy Sets and Systems 130 (2002) 33–42

FUZZY sets and systems

www.elsevier.com/locate/fss

A new approach to fuzzy regression models with application to business cycle analysis

Berlin Wu[a,*], Neng-Fang Tseng[b]

[a] *Department of Mathematical Sciences and Statistics, National Chengchi University, Taipei, Taiwan*
[b] *Department of Statistics, National Chengchi University, Taipei, Taiwan*

Received 31 March 1999; received in revised form 1 August 2001; accepted 9 August 2001

Abstract

Recently, fuzzy regression analysis has been largely applied in the modeling of economic or financial data. However, those data often exhibit certain kinds of linguistic terms, for instance: very good, a little reclining or stable, in the business cycle or the growth rate of GDP, etc. The goal of this paper is to construct a fuzzy regression model by fuzzy parameters estimation using the fuzzy samples. It deals with imprecise measurement of observed variables, fuzzy least square estimation and nonparametric methods. This is different from the assumptions as well as the estimation techniques of the classical analysis. Empirical results demonstrate that our new approach is efficient and more realistic than the traditional regression analysis.

Keywords: Fuzzy regression; Fuzzy parameter; Triangular membership function; *h*-cut; Methods of least square

1. Introduction

Regression analysis has been a very popular method with many successful applications. The problem of parameter estimation in the linear regression models has been an important research topic for statisticians. Conventional study on the regression analysis is based on the conception that the observed data are random with certain measurement errors or noise. However, in the empirical study those assumptions may hardly be realized, since there are many observations that experience linguistic or vague data inside the classical type.

For example, the official record of exchange rate for Japanese Yen to US dollar in January 1999 is 118.4. However, this exchange rate only accounts for the last exchange data, it cannot exactly display the variation of exchange rate (Japanese Yen to US dollar) during January 1999. Under such a situation, there may be a great chance of being misled if we try to apply these inaccurate data to fit a regression model.

Tanaka et al. [5] proposed the study in linear regression analysis using fuzzy set theory. They consider the linear interval regression model as

$$Y = A_0 + A_1x_1 + A_2x_2 + \cdots + A_px_p, \tag{1.1}$$

* Corresponding author. Tel.: +886-2-2938-7387; fax: +886-2-2939-0005.
E-mail address: berlin@math.nccu.edu.tw (B. Wu).

0165-0114/02/$ - see front matter
PII: S0165-0114(01)00175-0

where parameters A_i are triangular fuzzy numbers and the explained variables $x_1, x_2, \ldots, x_p$ are real value numbers. Therefore, the estimated value Y is also a fuzzy number. Tanaka et al. designed a useful technique to solve the estimation problem by transforming the optimization problem of estimation into a linear programming scheme. However, their method is a little complicated. Isshibuchi and Tanaka [2] presented an interval regression analysis based on the back-propagation neural networks. Their method is to obtain a nonlinear interval regression model by identifying the upper bound and the lower bound of the data interval. Recently, Yang and Ko [6] proposed a cluster-wise fuzzy in two approaches: the two-stage weighted fuzzy regression and the one-stage generalized fuzzy regression. The two-stage procedure extends the results of Jajuga [3] and Diamond [1]. The one stage is created by embedding fuzzy clustering into the fuzzy regression model fitting at each step of the procedure.

The regression analysis dealing with fuzzy data is usually called fuzzy regression analysis, while a linear interval regression model with fuzzy parameters is called fuzzy regression model. One advantage of using fuzzy regression analysis is that it can process the fuzzy sample data such as $(\mathbf{x}_i, Y_i)$, where Y_i is a fuzzy number and $\mathbf{x}_i$ is the vector of the explained variables, in a way which is closer to the reality. However, the estimation of fuzzy regression coefficients about fuzzy regression models has not been studied very much. That is, when the parameters A_i in Eq. (1.1) exhibit a linguistics form, such as A_i contains five linguistic values {very low, low, medium, high, very high}. In order to get an appropriate model to exhibit the real case, we had better apply the concept of fuzzy theory as well as the membership functions for these fuzzy sample data.

In this paper we propose a new approach to fuzzy regression models by using fuzzy number and method of least square. It is connected with imprecise measurement of observed variables, fuzzy least-squares estimation and nonparametric methods. This is different from the assumptions as well as the estimation techniques of the classical analysis. A generalized least-squares method with nonparametric statistics estimating the regression coefficients is derived. Empirical results demonstrate that our new approach is efficient and more realistic than the traditional regression analysis.

2. Fuzzy regression models

Since all information contained in a fuzzy set is described by its membership function, it is necessary to develop a lexical term to describe fuzzy number.

Definition 2.1 (fuzzy number). A fuzzy number A defined on the universe set X is a fuzzy subset in support R (the real number) which is both normal and convex where $supp(A) = \{x \in R \mid \mu_A(x) > 0\}$. Suppose we have an exhaustive collection of individual elements $\{x_1, x_2, \ldots, x_n\}$, which make up a universe of information or discourse X. Let A be a fuzzy number defined on the universe set X which contains $x_1, x_2, \ldots, x_n$ as its elements and $u(x_i)$ denote the grade of membership of x_i in A for $i = 1, 2, \ldots, n$. Then the fuzzy number A is written as

$$u_A(X)/A = u(x_1)/x_1 + u(x_2)/x_2 + \cdots + u(x_n)/x_n. \tag{2.1}$$

Example 2.1 (How many hours do you sleep in one day?). Let A be the fuzzy number of hours you sleep. Assume X to be the universe set with integral numbers. That is, $X = \{0, 1, 2, 3, 4, 5, \ldots, 24\}$. If a person's sleeping hours exhibit the following membership function A:

$$u_A(X) = \{u(6) = 0.1, u(7) = 0.2, u(8) = 0.4, \\ u(9) = 0.2, u(10) = 0.1\}$$

then the fuzzy number of sleeping hours can be written as

$$X = u_A(X)/A \\ = 0.1/6 + 0.2/7 + 0.4/8 + 0.2/9 + 0.1/10.$$

A general linear fuzzy regression model can be written as follows

$$\tilde{Y}(\mathbf{x}_i) = A_0 + A_1 x_{1i} + A_2 x_{2i} + \cdots + A_p x_{pi}, \tag{2.2}$$

where $\mathbf{x}_i = (1, x_{1i}, x_{2i}, \ldots, x_{pi})'$ is the vector of explanatory variables, $\tilde{Y}(\mathbf{x}_i)$ is the fuzzy sample dependent on $\mathbf{x}_i$ and A_m $(m = 1, 2, \ldots, p)$ is a fuzzy parameter. Usually, symmetric triangular membership functions are designed for the fuzzy parameter A_m. Hence, the membership function of $\tilde{Y}(\mathbf{x}_i)$ is also of symmetric triangular type.

Traditionally, we use the linear programming optimization method to estimate fuzzy parameter A_m as well as its support of $\tilde{Y}(\mathbf{x}_i)$. That is, one is to find a support of $\tilde{Y}(\mathbf{x}_i)$ such that the total length of interval has the minimized value and the support of $\tilde{Y}(\mathbf{x}_i)$ can cover the fuzzy sample Y_i. Therefore, we can derive the membership function of $\tilde{Y}(\mathbf{x}_i)$, and then find the estimators of fuzzy parameters A_m.

However, there still exist certain drawbacks in the estimation of fuzzy parameters by the use of linear programming. For instance, (i) Why should the fuzzy sample Y_i be covered by the interval estimations of $\tilde{Y}(\mathbf{x}_i)$? (ii) Are the fuzzy regression coefficients derived from the linear programming the best estimation? In this research, we will free the assumption, that is the fuzzy sample which falls outside the interval estimation $\tilde{Y}(\mathbf{x}_i)$ can be tolerated.

Although Savic and Pedrycz [4] have proposed least-squares method to estimate the fuzzy parameters, the main estimation procedure still uses the concept of the linear programming. For the sake of being realistic and efficient, we would propose an alternative new estimated procedure, which is different from the linear programming scheme.

In conventional estimation of the fuzzy parameter A_m for Eq. (2.2), people used to assume that the membership function of A_m is of triangle type, i.e.

$$\mu_{A_m}(t) = \max\left\{1 - \frac{|t - c_m|}{s_m}, 0\right\}, \quad -\infty < t < \infty, \tag{2.3}$$

where c_m and s_m are the center and radius on the triangular membership function, respectively. Their membership of the fuzzy parameter A_m is shown in Fig. 1.

Applying the fuzzy logic in the above Eq. (2.3), the membership function for fuzzy output $\tilde{Y}(\mathbf{x}_i)$ can be obtained as

$$\mu_{\tilde{Y}(\mathbf{x}_i)}(t) = \max\left\{1 - \left|t - \sum_{m=0}^{p} c_m x_{mi}\right| \Big/ \sum_{m=0}^{p} s_m |x_{mi}|, 0\right\}, \quad -\infty < t < \infty. \tag{2.4}$$

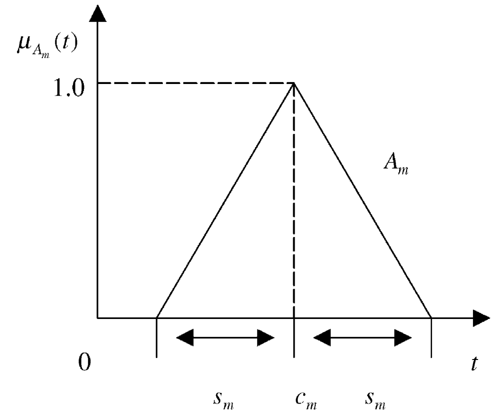

Fig. 1. Membership function for the fuzzy parameter A_m.

For simplicity, A_m can be represented by its center c_m and radius s_m as

$$A_m = \langle c_m, s_m \rangle. \tag{2.5}$$

Then, from (2.5), the fuzzy regression model (2.2) can also be represented as

$$\tilde{Y}(\mathbf{x}_i) = \langle c_0, s_0 \rangle + \langle c_1, s_1 \rangle x_{1i} + \langle c_2, s_2 \rangle x_{1i} + \cdots + \langle c_p, s_p \rangle x_{pi}. \tag{2.6}$$

However, in constructing fuzzy regression model, the fuzzy parameter A_m is unknown. In fact this is also what we are concerned with in this paper. But from the previous assumption of a triangle type, as long as we find its center c_m and the radius, we can get the estimation of membership functions A_m.

3. Estimation for fuzzy parameter

When the human subjective decision, measurement error or incomplete data affected the output data, we had better not look at those output data as an accurate numerical value. Therefore, we would like to use the fuzzy number as an exhibition of output. For observations $(\mathbf{x}_i, Y_i)$, where $\mathbf{x}_i = (x_{1i}, x_{2i}, \ldots, x_{pi})$ is the ith input vector and Y_i is its corresponding fuzzy sample, the fuzzy sample Y_i may be considered as a symmetric triangular membership function as follows:

$$\mu_{Y_i}(t) = \max\left\{1 - \frac{|t - y_i|}{r_i}, 0\right\}, \quad -\infty < t < \infty, \tag{3.1}$$

where y_i and r_i are the center and radius on the symmetric triangular membership function, respectively.

In order to estimate the middle point c_m and the radius s_m in Eq. (2.3), we have to consider the concepts of fuzzy number, which will be concerned with the concept of h-cut. An h-cut $[A]_h$ for a fuzzy number A on X is a crisp set denoted as

$$[A]_h = \{x \mid \mu_{A_i}(x) \geqslant h\}.$$

That is, the set $[A]_h$ contains all the elements of the universal set X such that the membership degree in A is greater than or equal to the specified value of h. Having the concept of h-cut, we give the definition of significant level h, the degree of fitness between the fuzzy sample Y_i and output $\tilde{Y}(\mathbf{x}_i)$.

Definition 3.1. Assume that $(\mathbf{x}_i, Y_i)$ is fuzzy sample and $\tilde{Y}(\mathbf{x}_i)$ is fuzzy output. Let h_i be the degree of fitness between the fuzzy sample Y_i and output $\tilde{Y}(\mathbf{x}_i)$. If $[Y_i]_h$ and $[\tilde{Y}(\mathbf{x}_i)]_h$ are the h-cut for Y_i and $\tilde{Y}(\mathbf{x}_i)$, respectively, i.e. $[Y_i]_h = \{t \mid \mu_{Y_i}(t) \geqslant h\}$ and $[\tilde{Y}(\mathbf{x}_i)]_h = \{t \mid \mu_{\tilde{Y}(\mathbf{x}_i)}(t) \geqslant h\}$, then, under the ith observation,

$$h_i = \max\{h \mid [Y_i]_h \subset [\tilde{Y}(\mathbf{x}_i)]_h\}. \tag{3.2}$$

The h_i is an important index in measuring the degree of fitness between the fuzzy sample Y_i and output $\tilde{Y}(\mathbf{x}_i)$. The larger the h_i is, the more fit the fuzzy sample Y_i and output $\tilde{Y}(\mathbf{x}_i)$. Note that the measurement of fitness for traditional regression analysis is judged by the sum of square error, while in the case of fuzzy sample the concept of distance measurement is more sophisticated. Therefore, we use h_i instead of sum of square error to measure the degree of fitness between the fuzzy sample Y_i and output $\tilde{Y}(\mathbf{x}_i)$. Fig. 2 illustrates the relationships among Y_i, $\tilde{Y}(\mathbf{x}_i)$ and h_i.

Under the condition of Definition 3.1, the degree of fitness h_i can be written as

$$h_i = 1 - \frac{|y_i - \sum_{m=0}^{p} c_m x_{mi}|}{\sum_{m=0}^{p} s_m |x_{mi}| - r_i}. \tag{3.3}$$

In order to get an appropriate estimation for c_m and s_m, we need to restrict all the values of h_i which are larger than a specific real value H, i.e. $h_i \geqslant H$, for all i, such that the degree of fitness between fuzzy sample Y_i and output $\tilde{Y}(\mathbf{x}_i)$ reaches a significant level H. We would like to guarantee that the degree of fitness

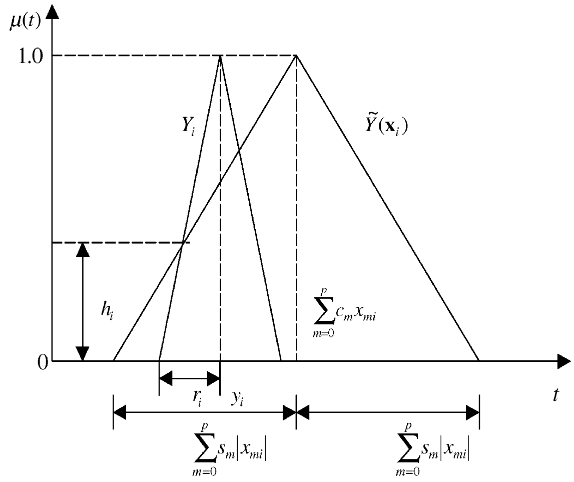

Fig. 2. Relationships among Y_i, $\tilde{Y}(\mathbf{x}_i)$ and h_i.

between the estimated $\tilde{Y}(\mathbf{x}_i)$ and Y_i will be better. It is natural for us to make the assumption that h_i should be larger than a significant level H, i.e. $h_i \geqslant H$ for all i. Given the significant level H, we can get the following two equations:

$$y_i \geqslant \sum_{m=0}^{p} c_m x_{mi} - (1-H)\sum_{m=0}^{p} s_m |x_{mi}| + (1-H) r_i$$

and

$$y_i \leqslant \sum_{m=0}^{p} c_m x_{mi} + (1-H)\sum_{m=0}^{p} s_m |x_{mi}| - (1-H) r_i$$

for all i. (3.4)

Moreover, we want to minimize the sum of each radius of triangle memberships for the fuzzy output $\tilde{Y}(\mathbf{x}_i)$, see Fig. 2, such that the estimated output will get the maximum likelihood information from the data. Following the above concept, we give the following definition as the objective function.

Definition 3.2. Let $\sum_{m=0}^{p} s_m |x_{mi}|$ be the radius of the ith $(i = 1, \ldots, n)$ output $\tilde{Y}(\mathbf{x}_i)$. The objective function O_{f} for the estimated parameter of fuzzy regression is denoted by

$$O_{\mathrm{f}} = \sum_{i=1}^{n} \sum_{m=0}^{p} s_m |x_{mi}|. \tag{3.5}$$

Therefore, under the condition of (3.4), we aim to find c_m and s_m such that O_f be minimized by the linear programming method.

In the above linear programming estimation process, the concept of h_i is used to estimate output $\tilde{Y}(\mathbf{x}_i)$ such that $[Y_i]^h \subseteq [\tilde{Y}(\mathbf{x}_i)]^h$. Therefore the output for a fuzzy regression is to find a set of distributed intervals which will cover all the samples. From the above estimation process, we may ask: is it necessary to use the fitted value h_i for sample Y_i and output $\tilde{Y}(\mathbf{x}_i)$? If we use h_i then it means that $[Y_i]^h$ can only fall in the interior of $[\tilde{Y}(\mathbf{x}_i)]^h$ because $[Y_i]^h \subseteq [\tilde{Y}(\mathbf{x}_i)]^h$. But if there exist certain outliers, according to the linear programming estimation technique, we may obtain a huge but inefficient estimated interval. We would like to believe that Y_i may reflect most properties of output, even that $[Y_i]^h \not\subset [\tilde{Y}(\mathbf{x}_i)]^h$. Hence, the assumption of the linear programming method does not seem robust. A more precise explanation should be that the h cut $[Y_i]^h$ for the representative sample Y_i should be close to $[\tilde{Y}(\mathbf{x}_i)]^h$, and as near as possible, i.e $[Y_i]^h \approx [\tilde{Y}(\mathbf{x}_i)]^h$.

Since the fuzzy sample $Y_i = \langle y_i, r_i \rangle$ belongs to a triangle membership function, the membership of Y_i is distributed on the interval $[y_i - r_i, y_i + r_i]$, where $y_{\mathrm{L}i} = y_i - r_i$ is the lower bound and $y_{\mathrm{U}i} = y_i + r_i$ is the upper bound. Though we require both bounds of Y_i to be very close to the support of $\tilde{Y}(\mathbf{x}_i)$, we cannot claim which one is contained in the other or whether both of their supports are equal.

We think that the memberships of lower and upper bounds, $y_{\mathrm{L}i}$ and $y_{\mathrm{U}i}$ for $\tilde{Y}(\mathbf{x}_i)$ reach a certain degree of significant level. Given the significant level H, we have

$$\mu_{\tilde{Y}_{(\mathbf{x}_i)}}(y_i - r_i) \approx H, \quad \mu_{\tilde{Y}_{(\mathbf{x}_i)}}(y_i + r_i) \approx H, \tag{3.6}$$

where $\mu_{\tilde{Y}_{(\mathbf{x}_i)}}(y_i - r_i)$ and $\mu_{\tilde{Y}_{(\mathbf{x}_i)}}(y_i + r_i)$ are the degrees of membership of the lower bound $y_{\mathrm{L}i}$ and upper bound $y_{\mathrm{U}i}$, respectively. In order to let the $\mu_{\tilde{Y}_{(\mathbf{x}_i)}}(y_i - r_i)$ and $\mu_{\tilde{Y}_{(\mathbf{x}_i)}}(y_i + r_i)$ get close to H, we may adjust their values instead of asking both of them to be larger than H.

From the above explanation and condition (3.6), we will present a new method in estimating fuzzy parameters. The method has the advantage of efficiency and it is easy to run.

3.1. Least-squares method for fuzzy parameter estimation

For a linear regression line, let the *lower regression line* be constructed by the least-squares method with the lower bound of $\{(\mathbf{x}_i, y_{\mathrm{L}i}) \mid i = 1, 2, \ldots, n\}$, which is written as

$$y_{\mathrm{L}}(\mathbf{x}_i) = \hat{L}_0 + \sum_{m=1}^{p} \hat{L}_m x_{mi}, \quad i = 1, 2, \ldots, n, \tag{3.7}$$

where L_m is a regression coefficient. Similarly, a linear regression line which is constructed by the least-squares method with the upper bound of $\{(\mathbf{x}_i, y_{\mathrm{L}i}) \mid i = 1, 2, \ldots, n\}$ is called an *upper regression line*, which is written as

$$y_{\mathrm{U}}(\mathbf{x}_i) = \hat{U}_0 + \sum_{m=1}^{p} \hat{U}_m x_{mi}, \quad i = 1, 2, \ldots, n. \tag{3.8}$$

Now we want to use the two functions to estimate the output $\tilde{Y}(\mathbf{x}_i)$ such that its degree of membership function approaches H. That is

$$[\tilde{Y}(\mathbf{x}_i)]^H = \left[\hat{L}_0 + \sum_{m=1}^{p} \hat{L}_m x_{mi}, \hat{U}_0 + \sum_{m=1}^{p} \hat{U}_m x_{mi}\right]. \tag{3.9}$$

By a simple calculation, we can get

$$\mu_{\tilde{Y}(\mathbf{x}_i)}(t) = \max\left\{1 - \left| t - \sum_{m=0}^{p} \left(\frac{\hat{U}_m + \hat{L}_m}{2}\right) x_{mi} \right| \Big/ \sum_{m=0}^{p} \left(\frac{\hat{U}_m - \hat{L}_m}{2(1-H)}\right) |x_{mi}|, 0\right\}, \tag{3.10}$$

where $x_{i0} = 1$. Moreover, the memberships of the fuzzy parameter A_m can be written as

$$\mu_{A_m}(t) = \max\left\{1 - \frac{|t - ((\hat{U}_m + \hat{L}_m)/2)|}{((\hat{U}_m - \hat{L}_m)/2(1-h))}, 0\right\} \tag{3.11}$$

or equivalently

$$A_m = \left\langle \frac{\hat{U}_m + \hat{L}_m}{2}; \quad \frac{\hat{U}_m - \hat{L}_m}{2(1-H)} \right\rangle, \tag{3.12}$$

where $(\hat{U}_m + \hat{L}_m)/2$ is the center of the support and $(\hat{U}_m - \hat{L}_m)/2(1-H)$ is the distributed radius. Finally, we get the following fuzzy regression model

$$\tilde{Y}(\mathbf{x}_i) = \left\langle \frac{\hat{U}_0 + \hat{L}_0}{2}, \frac{\hat{U}_0 - \hat{L}_0}{2(1-H)} \right\rangle + \left\langle \frac{\hat{U}_1 + \hat{L}_1}{2}, \frac{\hat{U}_1 - \hat{L}_1}{2(1-H)} \right\rangle x_{1i} + \left\langle \frac{\hat{U}_2 + \hat{L}_2}{2}, \frac{\hat{U}_2 - \hat{L}_2}{2(1-H)} \right\rangle x_{2i} + \cdots + \left\langle \frac{\hat{U}_p + \hat{L}_p}{2}, \frac{\hat{U}_p - \hat{L}_p}{2(1-H)} \right\rangle x_{pi}. \quad (3.13)$$

From the above process, the estimation of fuzzy regression model based on the least square method is simpler than the linear programming method.

Conventional estimation for the fuzzy parameters is based on the linear programming with the restriction that the sample must not fall outside of a bound area. Meanwhile, in order to make the estimation procedure more applicable, the new method we proposed is to free this restriction such that the data that fall outside the bound area are acceptable.

In fact, in the real case when the sample exhibits many outliners, the traditional LSE may not accurately estimate the fuzzy parameters $A_m = \langle c_m, s_m \rangle$, since the estimators induced by LSE will easily be influenced by the outliers. On the other hand, it is well known that by the application of nonparametric regression methods, we can find a more robust estimator. Therefore, we will combine the technique with fuzzy statistical method to get a more appropriate estimation of fuzzy parameters $A_m = \langle c_m, s_m \rangle$.

Let the lower boundary of H-cut be

$$y_{\mathrm{L}}(\mathbf{x}_i) = L_0 + \sum_{m=1}^{p} L_m x_{mi}, \quad i = 1, 2, \ldots, n. \quad (3.14)$$

Using the Gram–Schmidt orthogonalization procedure we can transform the endogenous variable x_{mi} into z_{mi}. The procedure of orthogonalization is

$$z_{1i} = x_{1i}, \quad 1, 2, \ldots, n,$$

$$z_{mi} = x_{mi} - \sum_{k=1}^{m-1} r_{mk} z_{ki} \quad \text{for } m > 1, \ 1, 2, \ldots, n,$$

$$r_{mk} = \left(\sum_{i=1}^{n} x_{mi} z_{ki} \right) \Big/ \left(\sum_{i=1}^{n} z_{ki}^2 \right). \quad (3.15)$$

Hence we have the following regression models:

$$y_{\mathrm{L}}(\mathbf{z}_i) = L'_0 + \sum_{m=1}^{p} L'_p z_{mi}, \quad i = 1, 2, \ldots, n, \quad (3.16)$$

Then we will find the L'_m by the transformed regression model and later through a simple transformation we will get the estimator L'_m of regression model (3.14). A detailed algorithm is described below.

1. For each m, let $L'_m = 0$, $m = 1, 2, \ldots, p$,
2. let $\delta L'_m = med \left\{ b_{ij}(m) = \dfrac{y_{\mathrm{L}j} - y_{\mathrm{L}i}}{z_{mj} - z_{mi}}; \quad z_{mi} < z_{mj}, \quad 1 \leqslant i < j \leqslant n \right\}$, (3.17)
3. $L'_m \leftarrow L'_m + \delta L'_m$,
4. $y_{\mathrm{L}i} \leftarrow y_{\mathrm{L}i} - \delta L'_m z_{mi}$,
5. if the L'_m converges then go to Step 6, otherwise let $m = m + 1$, go to Step 3,
6. decide $\tilde{L}'_0 = med\{y_{\mathrm{L}i}\}$.

From the above procedure we can get $\tilde{L}'_0, \tilde{L}'_1, \tilde{L}'_2 \ldots, \tilde{L}'_p$. The final estimation of $\tilde{L}_m$ can be written as

$$\tilde{L}_p = \tilde{L}'_p,$$

$$\tilde{L}_{p-m} = \tilde{L}'_{p-m} - \sum_{i=0}^{m-1} r_{p-i,p-m} \tilde{L}_{m-i}, \quad m = 1, 2, \ldots, p-1,$$

$$\tilde{L}_0 = \tilde{L}'_0. \quad (3.18)$$

Inserting $\tilde{L}_0, \tilde{L}_1, \ldots, \tilde{L}_p$ into Eq. (3.14), we can get the *H-cut* of lower bound of $\tilde{Y}(\mathbf{x}_i)$. Similarly, we can find the *H-cut* of upper bound of $\tilde{Y}(\mathbf{x}_i)$. Lastly, puting the *H-cut* of lower and upper bounds of $\tilde{Y}(\mathbf{x}_i)$ into Eq. (3.13), we can derive a fuzzy regression model

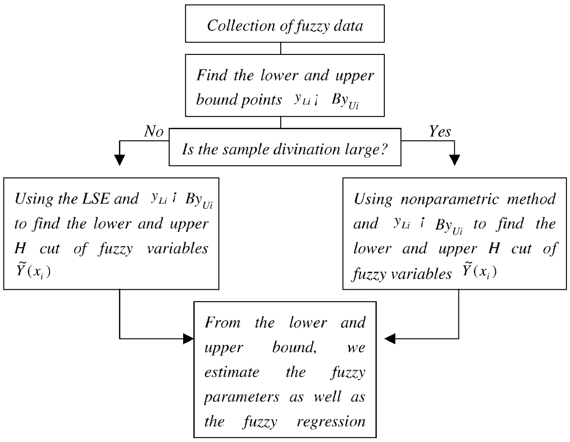

Fig. 3. The integrated flowchart of the estimation procedure.

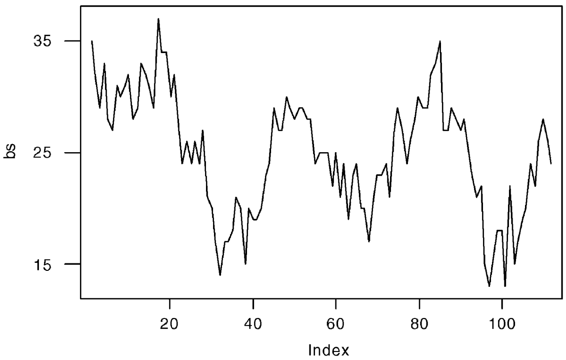

Fig. 4. Index of Taiwan business cycle.

under the nonparametric concept. Fig. 3 gives the integrated flowchart of the estimation procedure.

4. Empirical study

Fig. 4 is a plot of the monthly Taiwan Business Index from November 1987 to February 1997. These data come from the Council of Economic Planning and Development, Taiwan. Several nonregular business cycles in the 112 time series as well as certain structural change are exhibited. It is difficult to identify the structural change in this case. Fig. 4 also presents a large fluctuation and it is hard to construct a suitable model. First, we inspect the tendency of the Taiwan business in Fig. 4. We find large fluctuations in the beginning and the end of the Taiwan business cycle index which shows that there is larger variation in these periods.

In the data collection, because of the time lag, the published data cannot exhibit the exact value during the period. Some volume has been shifted to the next or the last time, for instance, the GDP, exchange rate, business cycles etc. The historical data record, rough average value, close market price or the mod of the price during the time period. They seldom described the variation during the period. To apply those data, statistical analysis might meet the danger of overestimation or overexplanation. Hence, we will use the concept of fuzzy statistics in constructing the regression models.

Taiwan Business Index, published monthly by the Economics Counsel, Execute Yen, consists of nine variables: M_1b, *Loan*, *Amount of Bills Exchange*, *Stock Market Index*, *New Order for Manufacture Department*, *Export Volume*, *Index of Industry Manufacture Department*, *Inventory of Manufacture Department*, *the Employment Rate of Non-Agriculture Department*. Summarizing the above 9 variables we can get the index of business cycle. There are 5 *Business Lights* corresponding to the linguistic term, that is: (1) *red light* means that the business activity is excellent, (2) *yellow–red light* means that the business activity is good, (3) *green light* means that the business activity is medium, (4) *yellow–blue light* means that the business activity is not good, (5) *blue light* means that the business activity is retrieving.

Since the index of business exhibits fuzzy characteristics, we apply the fuzzy parameter estimation to construct a fuzzy regression linear model. Among the nine factors, we choose the most essential exogenous variables (monthly): (1) increasing rates of $M_1B(X_1)$; (2) increasing rate of export volumes (X_2); (3) increasing rate of industry production (X_3); and also take the index of business cycle as our fuzzy endogenous variable. There are 88 data for each variable starting from October 1990 to January 1998.

The index of business cycle is composed of nine exogenous variables, thus we need to investigate the relationship among those variables. Moreover, we want to (1) examine whether fewer variables can better exhibit the index of business cycle, (2) construct a linear regression model under the point of fuzzy statistical view, (3) identify the linguistic category of the business lights (red, yellow–red, green, Yellow–blue, blue) among the main variables.

Table 1
The official record for index of business cycle

Light	Index of business cycle
Red	38–45
Yellow–Red	32–37
Green	23–31
Blue–Yellow	17–22
Blue	9–16

Table 2
The statistics of parameters

Parameter	Estimate of $\hat{L}_i$	$\hat{\sigma}(\hat{L}_j)$	T-Value	P-Value
L_0	14.500	0.6461	22.45	0.000
L_1	0.448	0.0500	8.96	0.000
L_2	0.116	0.0321	3.63	0.000
L_3	0.182	0.0711	2.56	0.012
$\hat{\sigma} = 3.057$	$R^2 = 61.0\%$			

Let the regression model $\tilde{Y}(\mathbf{x}_i)$ for the business cycle be

$$\tilde{Y}(\mathbf{x}_i) = A_0 + A_1 x_{1i} + A_2 x_{2i} + A_3 x_{3i}, \tag{4.1}$$

where x_{1i} is the $\mathrm{M_1b}$, x_{2i} is the annual rate of export, x_{3i} is the annual rate of industrial production, and $A_m = \langle c_m, s_m \rangle$ are the regression coefficients. For this 88 data size, we assume the index of business cycle Y_i as the fuzzy sample. The official record, using color as the symbol of index of business cycle was illustrated in Table 1.

We translate the rank of light into several intervals such that we can get both boundary points $y_{\mathrm{L}i} \square y_{\mathrm{U}i}$ for each sample point.

Under the significant level $H = 0.3$, for all the lower points of sample $(\mathbf{x}_i, y_{\mathrm{L}i})$, we get the following lower regression model by the use of least square error method:

$$y_{\mathrm{L}}(\mathbf{x}_i) = \sum_{m=0}^{3} L_m x_{mi} = 145 + 0.45 x_{1i} + 0.12 x_{2i} + 0.18 x_{3i}, \tag{4.2}$$

where the statistics of parameters are illustrated in Table 2.

Table 3
The statistics of parameters

Parameter	Estimate of U_m	$\hat{\sigma}(\hat{U}_m)$	T-Value	P-Value
U_0	20.200	0.6779	29.85	0.000
U_1	0.550	0.0524	10.48	0.000
U_2	0.125	0.0336	3.70	0.000
U_3	0.211	0.0745	2.83	0.006
$\hat{\sigma} = 3.208$	$R^2 = 67.0\%$			

Similarly, under the significant level $H = 0.3$, for all the upper points of sample $(\mathbf{x}_i, y_{\mathrm{U}i})$, we get the following upper regression model by the use of least square error method:

$$y_{\mathrm{U}}(\mathbf{x}_i) = \sum_{m=0}^{3} U_m x_{mi} = 20.2 + 0.550 x_{1i} + 0.125 x_{2i} + 0.211 x_{3i}, \tag{4.3}$$

where the statistics of parameters are illustrated in Table 3.

According to Eq. (3.12), the fuzzy parameter of A_m is

$$A_0 = \langle 17.3687, 4.0919 \rangle, \quad A_1 = \langle 0.4992, 0.0727 \rangle,$$

$$A_2 = \langle 0.1205, 0.0057 \rangle, \quad A_3 = \langle 0.1963, 0.0206 \rangle.$$

Then we get the regression fuzzy model

$$\tilde{Y}(\mathbf{x}_i) = \langle 17.3687, 4.0919 \rangle + \langle 0.4992, 0.0727 \rangle x_{1i} + \langle 0.1205, 0.0057 \rangle x_{2i} + \langle 0.1963, 0.0206 \rangle x_{3i}. \tag{4.4}$$

We use the fuzzy regression model to explain the input–output variables among index of business cycle $(\tilde{Y}(\mathbf{x}_i))$, growth rate of $\mathrm{M_1b}$ (x_1), growth rate of export volume and growth rate of industry production. Under this model, the explanation of business cycle will not be limited to the *Business lights*. We will have more flexible methods to explain the business cycle. For instance, the increasing rate of $\mathrm{M_1b}$ (x_1) from October 1997 is 12.87%, the increasing rate of export volume x_2 is 5.9%, and the industry production (x_3) is 11.05%. The official recorded *green light*. If we put the above three exogenous variables in model (4.4), we get the

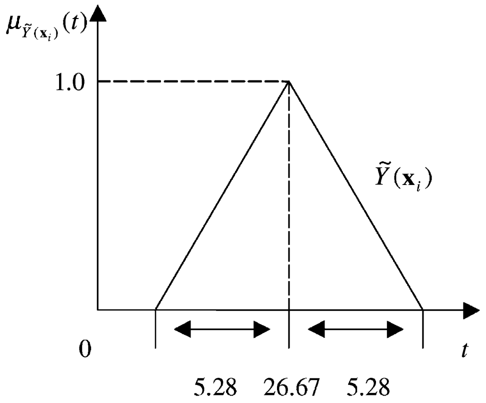

Fig. 5. The memberships of business cycle of Taiwan in October 1997.

memberships of business cycle $\tilde{Y}(\mathbf{x}_i)$ as

$$\mu_{\tilde{Y}(\mathbf{x}_i)}(t) = \max\{1 - |t - 26.27|/5.28, 0\}. \quad (4.5)$$

The memberships are illustrated in Fig. 5, where, instead of the traditional business light, t is the index value which stands for the state of the whole period. This kind of representation seems to give us more information about the period of time.

In the analysis of business cycle, we can find the correlation among the business index and the exogenous variables from the estimated regression model parameters. In the empirical study, it is easy to find that the fuzzy regression model renders a better explanation for the real case than the traditional one. Moreover, under the fuzzy set theory we can make use of all the information and demonstrate the data which people often ignore in the traditional analysis.

5. Conclusion

In this research we proposed a new method for parameters estimation of linear regression models integrated with statistical theory and the concept of fuzzy logic. The presented procedure to find the α-level of fuzzy parameter for a set of regression data is carefully discussed. Experimental results show that the proposed method of estimated fuzzy parameters is efficient and practical in explanation of the real data with the significant α-level. The results of this research present a feasible application and a new promising area for constructing regression models.

The conventional method for parameters estimation of linear regression rested on the concept of linear programming. Hence, if we use fuzzy statistical concept to analyze the relationships among the endogenous variables Y and exogenous variables X, we will get a better explanation for those variables. Especially from the statistical point of view, the least square method to estimate the boundaries of fuzzy variable is more appropriate and its computation is also more efficient than the traditional linear programming method. Though the least square method can be infected by certain outliers, we used the nonparametric technique to reduce the influence of those data and make the estimator more robust.

Finally, linguistic value estimation by the use of regression data is very complicated, involving entities with many features and parts which interact with each other and their environment in intricate ways. The proposed method is also suitable for dealing with historical data, which include linguistic values. The method of fuzzy parameter estimation does not require precise knowledge about the structure in the data and can take full advantage of the model-free approach. However, there still remain many problems for future studies, such as the following:

(i) In the traditional regression models construction, the estimation based on the least square is best linear unbiased estimate (BLUE). Is our estimator BLUE? How to define the BLUE from the fuzzy statistics point of view?
(ii) How to identify the fuzzy endogenous and exogenous variables as well as detect the intervention among them?
(iii) How to extend the estimated parameter methods to the seasonal regression models?
(iv) How much precision is required in constructing the fuzzy parameters under the significant α-level?

References

[1] P. Diamond, Fuzzy least squares, Inf. Sci. 46 (1988) 141–157.
[2] H. Isshibuchi, H. Tanaka, Fuzzy regression analysis using neural networks, Fuzzy Sets and Systems 50 (1992) 257–265.

[3] K. Jajuga, Linear fuzzy regression, Fuzzy Sets and Systems 20 (1986) 343–353.
[4] D.A. Savic, W. Pedrycz, Evaluation of fuzzy linear regression models, Fuzzy Sets and Systems 23 (1991) 51–63.
[5] H. Tanaka, H. Ishibuchi, Possibilistic regression analysis based on linear programming, in: J. Kacprzyk, M. Fedrizzi (Eds.), Fuzzy Regression Analysis, Omnitech Press, Warsaw and Physica-Verlag, Heidelberg, 1992.
[6] M. Yang, C. Ko, On cluster-wise fuzzy regression analysis, IEEE Trans. Systems Man Cybernet. 27 (1997) 1–13.