

Fuzzy regression model with fuzzy input and output data for manpower forecasting

Hong Tau Lee*, Sheu Hua Chen

*Department of Industrial Engineering and Management, National Chin-Yi Institute of Technology and Commerce, Taiping,
Taichung Country 41111, Taiwan, People's Republic of China*

Received January 1998; received in revised form August 1998

Abstract

In modeling a fuzzy system with fuzzy linear functions, the vagueness of the fuzzy output data may be caused by both the indefiniteness of model parameters and the vagueness of the input data. This situation occurs as the input data are envisaged as facts or events of an observation which are uncontrollable or uninfluenced by the observer rather than as the controllable levels of factors in an experiment. In this research, we concentrate on such a situation and refer to it as a generalized fuzzy linear function. Using this generalized fuzzy linear function, a generalized fuzzy regression model is formulated. A nonlinear programming model is proposed to identify the fuzzy parameters and their vagueness for the generalized regression model. A manpower forecasting problem is used to demonstrate the use of the proposed model. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Fuzzy regression; Measure of vagueness; Forecasting

1. Introduction

When human judgments are involved with a system or the data obtained from the system is scarce or insufficiently precise, the vagueness of such a system considered must be dealt with in the system modeling. The fuzzy regression model developed by Tanaka et al. [10,11] using fuzzy parameters has been applied to modeling systems involving vague or imprecise phenomena. As one example, Heshmaty and Kandel [3] applied fuzzy regression to forecast the computer sales in the United States in an uncertain environment. In addition, other researchers have devoted their efforts to improve the application capability of this fuzzy regression methodology. In this direction, Moskowitz and Kim [9] studied the relationship among the H value, membership function shape, and the spreads of fuzzy parameters in fuzzy linear regression models, also developing a systematic approach to assess the proper H parameter values. Kim et al. [5] and Kim and Chen [6] made a comparison of fuzzy and nonparametric linear regression and concluded that when the

* Corresponding author. Fax: +886 4 393 4620.

size of a data set is small, error terms have small variability, and the relationships among variables are not well specified, fuzzy linear regression outperforms nonparametric linear regression with respect to descriptive capability which is concerned with how close the estimated model parameters are to the true parameter values. Although the fuzzy regression models assume that the residuals depend on the indefiniteness of the model parameters that contracts the conventional least squares considered it to be measurement errors. From the view point of the system of relationship between input and output data sets, the deviations between the estimated and the observed values of output data may be caused not only by the vagueness of the parameters of the model, but also by the vagueness of the input data. This situation may occur when the input data are envisaged as facts or events of an observation which are uncontrollable or uninfluenced by the observer rather than as the controllable levels of factors in an experiment. In this research, we concentrated on a fuzzy linear regression model that assumes the residuals are caused by the vagueness, both of the parameters of the model and of the input data simultaneously. In the next section, the original definition of the fuzzy regression by Tanaka [10] is described again for reading convenience. In Section 3, the membership function is conducted for the generalized fuzzy linear model that we are considering. In Section 4, a manpower forecasting example is proposed as an illustration of the generalized model. Finally, Section 5 is a conclusion.

2. The fuzzy linear model

First of all, we assume that a fuzzy phenomenon can be presented as a fuzzy system of equations which, in turn, can be described by the fuzzy function $\tilde{Y} = \tilde{\beta}_1 X_1 \oplus \tilde{\beta}_2 X_2 \oplus \dots \oplus \tilde{\beta}_n X_n$, in which the fuzzy parameters $\tilde{\beta}_j$ with membership functions are as follows:

$$U_{\tilde{\beta}_j}(\beta_j) = \begin{cases} \frac{1}{b_j - a_j} (\beta_j - a_j), & a_j \leq \beta_j \leq b_j, \\ \frac{1}{b_j - c_j} (\beta_j - c_j), & b_j \leq \beta_j \leq c_j, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Then the membership function of the fuzzy linear model can be obtained from the following proposition.

Proposition 1. *Given fuzzy parameters $\tilde{\beta}_j$ with membership functions as function (1), the membership function of the fuzzy linear function $\tilde{Y} = \tilde{\beta}_1 X_1 \oplus \tilde{\beta}_2 X_2 \oplus \dots \oplus \tilde{\beta}_n X_n$ is obtained as the following:*

$$U_{\tilde{Y}}(y) = \begin{cases} \frac{1}{C_3 - C_1} (y - C_3), & C_1 \leq y \leq C_3, \\ \frac{1}{C_3 - C_2} (y - C_2), & C_3 \leq y \leq C_2, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

in which $C_1 = \sum_{j=1}^n a_j x_j$, $C_2 = \sum_{j=1}^n c_j x_j$, $C_3 = \sum_{j=1}^n b_j x_j$.

Proof. By the extension principle of fuzzy sets [12,13] and the definition of the triangular fuzzy number [2], a triangular fuzzy number multiplied by a positive real number is still a triangular fuzzy number. For example, if a triangular fuzzy number $\tilde{A}' = (a, b, c)$ then $\tilde{A} = \tilde{A}' \otimes x = (ax, bx, cx)$. Another extension principle is that

the summation of triangular fuzzy numbers is also a triangular fuzzy number. For example, if $\tilde{A} = (a, b, c)$ and $\tilde{B} = (d, e, f)$ then $\tilde{C} = \tilde{A} \oplus \tilde{B} = (a + d, b + e, c + f)$. By the two extension principles, the membership function of the fuzzy linear function is available. \square

From Proposition 1, the fuzzy linear regression model $Y^* = \beta_1^* X_1 \oplus \beta_2^* X_2 \oplus \dots \oplus \beta_n^* X_n$ can be conceived as summation of the triangular fuzzy numbers multiplied by positive real numbers. In a fuzzy linear regression model, the bases of the triangular membership function of the fuzzy parameters are usually assumed to have the same spread width at both sides of their center values, such that $b_j = \alpha_j$, $a_j = b_j - l_j$, $c_j = b_j + l_j$, where α_j is the center value of β_j^* , and l_j is the spread width of parameter β_j^* . Consequently, the membership function of β_j^* is as follows:

$$U_{\beta_j^*}(\beta_j) = \begin{cases} 1 + \frac{1}{l_j}(\beta_j - \alpha_j), & \alpha_j - l_j \leq \beta_j \leq \alpha_j, \\ 1 - \frac{1}{l_j}(\beta_j - \alpha_j), & \alpha_j \leq \beta_j \leq \alpha_j + l_j, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

By Proposition 1, the membership function of the fuzzy linear regression function $Y^* = \beta_1^* X_1 \oplus \beta_2^* X_2 \oplus \dots \oplus \beta_n^* X_n$ will be

$$U_{Y^*}(y) = \begin{cases} 1 + \frac{1}{\sum_{j=1}^n l_j x_j} \left(y - \sum_{j=1}^n \alpha_j x_j \right), & C_1 \leq y \leq C_3, \\ 1 - \frac{1}{\sum_{j=1}^n l_j x_j} \left(y - \sum_{j=1}^n \alpha_j x_j \right), & C_3 \leq y \leq C_2, \\ 0 & \text{otherwise,} \end{cases} \quad (4)$$

in which $C_1 = \sum_{j=1}^n (\alpha_j - l_j) x_j$, $C_2 = \sum_{j=1}^n (\alpha_j + l_j) x_j$, $C_3 = \sum_{j=1}^n \alpha_j x_j$.

The membership function of the output data y also should have the same special form of the previous fuzzy parameters such as

$$U_Y(y) = \begin{cases} 1 + \frac{1}{w} (y - q), & q - w \leq y \leq q, \\ 1 - \frac{1}{w} (y - q), & q \leq y \leq q + w, \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

where q and w represent the center value and spread width of y , respectively. The degree of the fitting of the estimated fuzzy linear regression model $Y^* = \beta_1^* X_1 \oplus \beta_2^* X_2 \oplus \dots \oplus \beta_n^* X_n$ to the given data set $Y = (q, w)$ is measured by the following index \bar{h} which maximizes \bar{h} subject to $Y^h \subset Y^{*h}$, where $Y^h = \{y | U_Y(y) \geq h\}$, $Y^{*h} = \{y | U_{Y^*}(y) \geq h\}$, which are h -level sets. This index \bar{h} is illustrated in Fig. 1. This degree of the fitting of the fuzzy regression model to the data set $Y = (q, w)$ is defined by minimizing \bar{h} for each element of this data in set $Y = (q, w)$. The vagueness of the fuzzy regression model $Y^* = \beta_1^* X_1 \oplus \beta_2^* X_2 \oplus \dots \oplus \beta_n^* X_n$ is defined by $V = \sum_{j=1}^n l_j$. The problem is explained as obtaining fuzzy parameters β^* which minimize V subject to

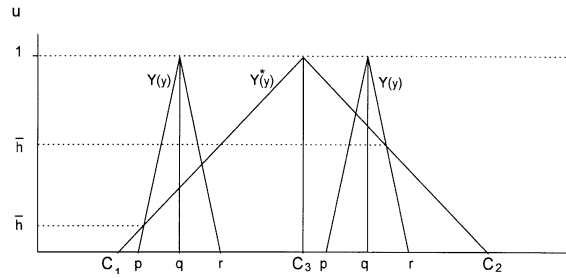


Fig. 1. Degree of fitting of y^* to a given fuzzy data y .

$\bar{h} > H$ for all data in $Y = (q, w)$, where H is chosen as the degree of the fitting of the fuzzy linear model by the decision maker. The following linear programming problem expresses the above situation:

$$\begin{aligned} \min_{\alpha_j, l_j} \quad & V = l_1 + l_2 + \dots + l_n \\ \text{s.t.} \quad & \sum_{j=1}^n \alpha_j x_j + (1 - H) \sum_{j=1}^n l_j x_j \geq q + (1 - H)w, \quad q \leq C_3, \\ & - \sum_{j=1}^n \alpha_j x_j + (1 - H) \sum_{j=1}^n l_j x_j \geq -q + (1 - H)w, \quad q \geq C_3, \\ & l_j \geq 0 \quad \text{for } j = 1, \dots, n, \\ & w \geq 0, \end{aligned}$$

the set of constraints must hold for every data y in set $Y = (q, w)$.

3. The fuzzy linear model with fuzzy input and output data

In the previous section, we discussed the original fuzzy linear function which assumes the deviation of the output data between its observed value and estimated value is only caused by the indefiniteness of the model parameters. This interpretation can be extended if the fuzzy linear function represents the relationship between the input data and output data. From this point of view, the vagueness of the output data may be caused not only by the indefiniteness of the model parameters but also by the vagueness of the input data. In this section, the general situation of the fuzzy linear function is discussed. Suppose the input data X_j is a fuzzy number with the relationship function as following:

$$U_{X_j(x_j)} \cong \begin{cases} \frac{1}{e_j - d_j} (x_j - d_j), & d_j \leq x_j \leq e_j, \\ \frac{1}{e_j - f_j} (x_j - f_j), & e_j \leq x_j \leq f_j, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

The membership function for the generalized fuzzy linear function $\tilde{Y} = \tilde{\beta}_1 X_1 \oplus \tilde{\beta}_2 X_2 \oplus \dots \oplus \tilde{\beta}_n X_n$ is obtained from the following Proposition 2,

Proposition 2. Given fuzzy parameters $\tilde{\beta}_j$ with membership functions as function (1) and fuzzy input data X_j with membership function as function (6), the membership function of the generalized fuzzy linear function $\tilde{Y} = \tilde{\beta}_1 X_1 \oplus \tilde{\beta}_2 X_2 \oplus \dots \oplus \tilde{\beta}_n X_n$ is obtained as the following:

$$U_{\tilde{Y}(y)} \cong \begin{cases} \frac{-B_1}{2A_1} + \left[\left(\frac{B_1}{2A_1} \right)^2 - \frac{C_1 - y}{A_1} \right]^{1/2}, & C_1 \leq y \leq C_3, \\ \frac{B_2}{2A_2} - \left[\left(\frac{B_2}{2A_2} \right)^2 - \frac{C_2 - y}{A_2} \right]^{1/2}, & C_3 \leq y \leq C_2, \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

in which,

$$\begin{aligned} A_1 &= \sum_{j=1}^n (b_j - a_j)(e_j - d_j), & A_2 &= \sum_{j=1}^n (c_j - b_j)(f_j - e_j), \\ B_1 &= \sum_{j=1}^n (a_j(e_j - d_j) + d_j(b_j - a_j)), & B_2 &= \sum_{j=1}^n (c_j(f_j - e_j) + f_j(c_j - b_j)), \\ C_1 &= \sum_{j=1}^n a_j d_j, & C_2 &= \sum_{j=1}^n c_j f_j, & C_3 &= \sum_{j=1}^n b_j e_j. \end{aligned}$$

Proof. From the relationship functions of the fuzzy number $\tilde{\beta}_j$ and X_j , the α -cut interval sets for the two fuzzy numbers are $\tilde{\beta}_j^\alpha = [a_j + \alpha(b_j - a_j), c_j + \alpha(b_j - c_j)]$ and $X_j^\alpha = [d_j + \alpha(e_j - d_j), f_j + \alpha(e_j - f_j)]$, respectively. By the extension principle of fuzzy sets, the multiplication of two α -cut interval sets causes a α -cut interval set as $Z_j^\alpha = \tilde{\beta}_j^\alpha * X_j^\alpha = \{[a_j + \alpha(b_j - a_j)] * [d_j + \alpha(e_j - d_j)], [c_j + \alpha(b_j - c_j)] * [f_j + \alpha(e_j - f_j)]\}$. The membership function of the fuzzy number Z_j^α is facile. \square

Although the membership function of Z_j^α is obtained from Proposition 2, it is not again a triangular fuzzy number. Hence, the summation of the n fuzzy number derived a fuzzy number with the approximated relationship function as function (7) above. Fig. 2 describes the shape of the approximated function.

The membership function of the output data y is also supposed to have the same special form of triangular described as function (5) in Section 2 above. The degree of the fitting of the estimated generalized fuzzy linear regression model $Y^* = \beta_1^* X_1^* \oplus \beta_2^* X_2^* \oplus \dots \oplus \beta_n^* X_n^*$ to the given data set $Y = (q, w)$ is measured, as the same as the linear regression model in Section 2, by the index \bar{h} which maximizes h subject to $Y^h \subset Y^{*h}$, where $Y^h = \{y | U_Y(y) \geq h\}$, $Y^{*h} = \{y | U_{Y^*}(y) \geq h\}$. This index \bar{h} is illustrated in Fig. 3.

The degree of the fitting of the fuzzy regression model to the data set $Y = (q, w)$ is defined by minimizing \bar{h} for each of this data in set $Y = (q, w)$. The vagueness of the fuzzy regression model $Y^* = \beta_1^* X_1^* \oplus \beta_2^* X_2^* \oplus \dots \oplus \beta_n^* X_n^*$ is defined by $V = \sum_{j=1}^n l_j$. The problem is explained as obtaining fuzzy parameters β^* which minimizes V subject to $\bar{h} > H$ for all data in $Y = (q, w)$, where H is chosen as the degree of the fitting of the fuzzy

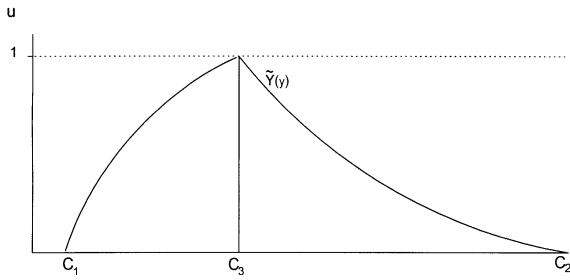


Fig. 2. Shape of the approximated membership Function 7.

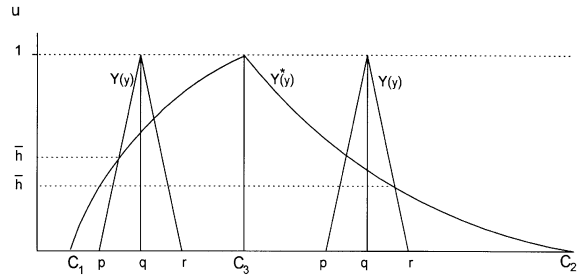


Fig. 3. Degree of fitting of y^* to given fuzzy data y .

linear model by the decision maker. The above situation is expressed by the following nonlinear programming problem:

$$\min_{z_j, l_j} V = l_1 + l_2 + \dots + l_n$$

s.t.

$$h = \frac{1}{w} \left(- \left(\frac{wB_1}{2A_1} - p - \frac{w^2}{2A_1} \right) + \left(\left(\frac{wB_1}{2A_1} - p - \frac{w^2}{2A_1} \right)^2 - \left(p^2 - \frac{wpB_1}{A_1} + \frac{w^2C_1}{A_1} \right) \right)^{1/2} - p \right) \geq H$$

if $q \leq C_3$,

in which $w = q - p$, and

$$h = - \frac{1}{w} \left(- \left(r - \frac{wB_2}{2A_2} + \frac{w^2}{2A_2} \right) + \left(\left(r - \frac{wB_2}{2A_2} + \frac{w^2}{2A_2} \right)^2 - \left(r^2 - \frac{wrB_2}{A_2} + \frac{w^2C_2}{A_2} \right) \right)^{1/2} - r \right) \geq H$$

if $q \geq C_3$,

in which $w = r - q$.

The set of constraints must hold for every data y in set $Y = (q, w)$.

For simplicity, in practice the approximation formula $U_{Y^*}(y) \cong (C_1, C_3, C_2)$ that represents triangular fuzzy number can be used. The above equation can be simplified as follows (see Fig. 4):

$$\min_{z_j, l_j} V = l_1 + l_2 + \dots + l_n$$

$$\text{s.t. } h = \frac{1}{C_3 - C_1} \left(\frac{C_1(q - p) - p(C_3 - C_1)}{(q - p) + (C_3 - C_1)} - C_1 \right) \geq H \quad \text{if } q < C_3,$$

$$h = \frac{1}{C_3 - C_2} \left(\frac{C_2(q - r) - r(C_3 - C_2)}{(q - r) - (C_3 - C_2)} - C_2 \right) \geq H \quad \text{if } q > C_3.$$

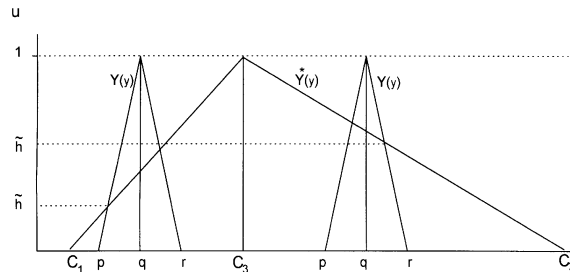


Fig. 4. Degree of fitting of y^* to given fuzzy data y for the simplified regression model.

4. Manpower forecasting problem

Most businesses and government agencies are engaged in some form of manpower forecasting, indicating that to forecast the future demand for manpower is one important area of manpower policy for individual enterprises and government. Although the aggregate demand of an entire industry is essentially the sum of the demands of individual firms, it is hardly possible to acquire the demand of every firm. One method is to study a sample of firms and make inferences to the whole population. However, consideration must be given to two major sources which influence the variance of the sample, making the forecasts unreliable. One factor is that most registered firms are considered as small businesses (in Taiwan, ROC approximately 98%) which will not last very long. Statistics indicate that in Taiwan, approximately 62% of these firms will close within five years operation (Medium and Small Business Consulting Center [1]). The other factor is that, due to the large number of firms, only a small portion can be investigated. In addition to the above two sources of variance of sample, the demand for manpower is typically concentrated in very large companies. To deal with these factors, Kao and Lee [4] proposed the idea of sampling from the more important companies to acquire reliable data than making suitable extrapolations to the less important small companies and using regression analysis by selecting approximate explanatory variables to make forecasts.

In this research, the method of Kao and Lee [4] is followed. A sample of 100 firms of the top 1000 firms according to their sales volume in Taiwan is extracted. To ensure that the correct data is obtained for analysis, site visits to the firms were undertaken. Each firm investigated is requested to provide the number of employees in the areas of industrial engineering and management (IE&M) planned to be hired in 1998. Table 1 shows the profile of the sample 100 firms, grouping the firms into the classes by size, there are 12 firms of the top 100 firms, and 7, 11, 8, 14, 11, 10, 11, 8 and 8 firms are in the subsequent classes of 100 firms, respectively. The means and variances for the ranks and IE&M employees planned to be hired corresponding to the classed firms in the sample are described in Table 1, too. These two values indicate that both of the employees with IE&M profession and ranks are variable with some percentage of vagueness. It can be seen, from Table 1 that there is a clear descending trend along the rank-classes in an average sense. Most of the employees in IE&M profession are needed in the 100 largest firms and its variance is relatively large compared to the other smaller firms. Fig. 5 also depicts this relationship in an exponential form. Consequently, we assume the fuzzy regression model is $Y = \beta_1 X^{\beta_2}$, in which, variables Y and X represent the number of IE&M employees planned to be hired and the rank of the firm, respectively. After implementing the natural log operation on both sides of the regression model, the equation $\ln(Y + 1) = \ln(\beta_1) + \beta_2 \ln(X)$ is derived. The reason for using $Y + 1$ instead of Y in the previous model is because some original data for Y are zero, and adding 1 to the original data Y make the logarithm operations meaningful. According to the property of the collected data set, the generalized fuzzy linear model is applied to identify the regression line. After transforming the original data to natural log form, the mean and one standard deviation value of the data are taken as the central and spread range values of variables $\ln(Y)$ and $\ln(X)$, respectively. The decision maker

Table 1
The profile for the 100 observations in the 1000 largest companies

Rank	Number of firms	Mean of the ranks	Standard error of the ranks	Mean of IE&M employees	Standard error of IE&M employees
1–100	12	52.83	26.66	6.83	4.53
101–200	7	140.00	31.15	2.14	0.90
201–300	11	255.73	34.61	1.18	0.60
301–400	8	370.50	24.56	0.88	0.64
401–500	14	450.36	26.71	0.71	0.73
501–600	11	561.09	29.95	0.27	0.47
601–700	10	633.10	21.58	0.70	0.67
701–800	11	741.45	22.54	0.27	0.47
801–900	8	848.00	22.27	0.50	0.53
901–1000	8	969.75	19.54	0.50	0.53

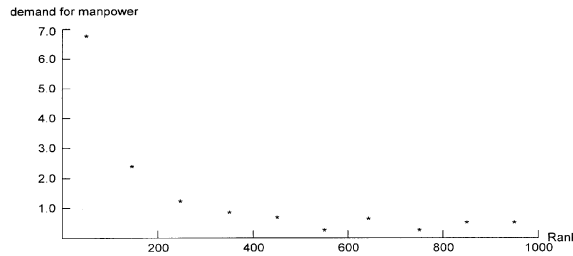


Fig. 5. The relationship between the IE employees and rank of the firms.

Table 2
Center values and spread widths for the fuzzy parameters

Fuzzy parameters	$\ln(\beta_1)$	β_2
Center	4.713379	-0.605585
Width	0.0	3.207107

parameter H is adopted as value 0.5. A generalized reduced gradient algorithm [8], which has been found [7] to be an efficient method, is applied to solve the non-linear programming problem presented in Section 3. Finally, the central and spread ranges of the parameters $\ln(\beta_1)$ are 4.713379 and 0.0, of β_2 are -0.605585 and 3.207107, respectively. All of these are described in Table 2. We then substitute the central value of the above parameters into the original exponential model and integrate the area defined by the regression line, the y -axis and the x -axis. The estimated total manpower demand for profession of IE&M is identified as 3162 persons, which is obtained by calculating the following integration: $\int_X [e^{4.713379 - 0.605585 \ln(X)} - 1] dX$.

The regression line approximately intersects the x -axis when the rank reaches 1228, indicating that the manpower demand for IE&M employees in the firms with ranks larger than 1228 are negligible.

5. Conclusion

In this research, a generalized fuzzy linear function is discussed. Considering a fuzzy linear function as a model for the fuzzy structure of a system, a fuzzy linear regression function analysis is formulated. From this point of view, the vagueness of the output data is caused not only by the indefiniteness of the model parameters, but also by the vagueness of the input data. This is in distinction to the conventional regression model which always assumes that the deviation of the dependent variables from their estimated value to their observed values is caused only by the measurement error and regards the independent variables as being controllable. During the implementation of this approach, such as in the above manpower forecasting example, no assumption such as the normality of the distribution of error terms in the conventional regression analysis are made. Consequently, the robustness of the generalized fuzzy linear model is apparent.

References

- [1] Business Consulting Center, Medium and Small Business Annual Financial Report, Taipei, Taiwan, ROC, 1984.
- [2] D. Dubois, H. Prade, Operations on fuzzy numbers, *Internat. J. Systems Sci.* 9 (1978) 613–626.
- [3] B. Heshmaty, A. Kandel, Fuzzy linear regression and its applications to forecasting in uncertain environment, *Fuzzy Sets and Systems* 15 (1985) 159–191.
- [4] C. Kao, H.T. Lee, An integration model for manpower forecasting, *J. Forecasting* 15 (1996) 543–548.
- [5] K.J. Kim, H. Moskowitz, M. Koksalan, Fuzzy versus statistical linear regression, *European J. Oper. Res.* 92 (1996) 417–434.
- [6] K.J. Kim, H.R. Chen, A comparison of fuzzy and nonparametric linear regression, *Comput. Oper. Res.* 24 (1997) 505–519.
- [7] H.T. Lee, A study of generalized reduced gradient method for solving nonlinear programming problems, Thesis, National Cheng Kung University, Taiwan, ROC, 1986.
- [8] D.G. Luenberger, *Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1984.
- [9] H. Moskowitz, K.J. Kim, On assessing the H value in fuzzy linear regression, *Fuzzy Sets and Systems* 58 (1993) 303–327.
- [10] H. Tanaka, S. Uejima, K. Asai, Fuzzy linear regression model, *IEEE Trans. Systems Man Cybernet.* 10 (1980) 2933–2938.
- [11] H. Tanaka, S. Uejima, K. Asai, Linear regression analysis with fuzzy model, *IEEE Trans. Systems Man Cybernet.* 12 (1982) 903–907.
- [12] L.A. Zadeh, The concept of a linguistic variable and its application to approximate reasoning, Part 1, 2 and 3, *Inform. Sci.* 8 (1975) 199–249, 301–357; 9 (1976) 43–80.
- [13] H.J. Zimmermann, *Fuzzy Set Theory and its Application*, Kluwer Academic Publishers, Reading, 1991.