SEVERAL FORMULATIONS OF INTERVAL REGRESSION ANALYSIS

Hisao Ishibuchi and Hideo Tanaka
(College of Engineering, University of Osaka Prefecture,
Sakai, Osaka 591, JAPAN)

ABSTRACT

We propose several mathematical programming formulations of regression analysis using linear interval models. A basic form of regression analysis with a linear interval model can be formulated as a linear programming problem. The LP problem is similar to an existing formulation of fuzzy regression analysis with a linear fuzzy model[1,2]. Since the LP formulation is sensitive to outliers, it cannot cope with a case where given data have outliers. Therefore we extend the LP problem to a combinatorial optimization problem in order to deal with outliers. We also formulate a two-objective combinatorial optimization problem where users of regression analysis can choose the best interval regression model through interaction with computer programs.

1. INTRODUCTION

In regression analysis, a regression model for m input-output data points $(\mathbf{x}_j, y_j)$, $j=1,2,\cdots,m$, is

$$y_j = a_0 + a_1x_{1j} + \cdots + a_nx_{nj} + e_j \quad \text{for } j=1,2,\cdots,m \tag{1}$$

where $\mathbf{x}_j = (x_{1j}, \cdots, x_{nj})$ and e_j is the j-th residual. Unknown coefficients $a_0,a_1,\cdots,a_n$ are determined by the following objective function in the least squares regression.

$$\text{minimize} \quad (e_1^2 + e_2^2 + \cdots + e_m^2). \tag{2}$$

In the absolute errors regression, the unknown coefficients are determined by the following objective function.

$$\text{minimize} \quad (\text{abs}(e_1) + \text{abs}(e_2) + \cdots + \text{abs}(e_m)) \tag{3}$$

where abs(e_j) denotes the absolute value of e_j. It should be noted that the estimated regression model:

$$y^* = a_0^* + a_1^* x_1 + \cdots + a_n^* x_n \tag{4}$$

is a hyperplane in an (n+1)-dimensional space.

The aim of this paper is to propose new regression methods based on a linear interval regression model:

$$Y(\mathbf{x}_j) = A_0 + A_1x_{1j} + A_2x_{2j} + \cdots + A_nx_{nj} \quad \text{for } j=1,2,\cdots,m \tag{5}$$

where A_i is an interval coefficient and $Y(\mathbf{x}_j)$ is the predicted interval corresponding to $\mathbf{x}_j = (x_{1j},x_{2j},\cdots,x_{nj})$. Geometrically, the interval regression model (5) represents a region between two hyperplanes in an (n+1)-dimensional space (for example, see Fig.1 and Fig.2 in Section 2.2).

A linear programming problem is formulated to determine the interval coefficients under the following condition.

$$y_j \in Y(\mathbf{x}_j) \quad \text{for } j=1,2,\cdots,m. \tag{6}$$

This condition requires that the interval regression model (5) should include all the given data points.

Since the LP formulation is sensitive to outliers because of the constraint condition (6), it cannot cope with a case where the given data have outliers. Therefore we extend the LP problem to a combinatorial optimization problem where the constraint condition (6) is relaxed as follows.

$$y_j \in Y(\mathbf{x}_j) \quad \text{for k data points out of the given m data points.} \tag{7}$$

We also formulate a two-objective combinatorial optimization problem in order to determine the linear interval regression model through an interactive manner.

2. FORMULATION OF A LINEAR PROGRAMMING PROBLEM

2.1 Formulation

We first formulate a LP problem to determine the interval coefficients of the interval linear regression model. Our problem is to determine the interval coefficients A_i, $i=0,1,\cdots,n$, using the given input-output data points $(\mathbf{x}_j,y_j)$, $j=1,2,\cdots,m$, under the constraint condition (6).

Let us denote the interval coefficient A_i using its center a_{ci} and radius a_{wi} as

$$A_i = \langle a_{ci},\ a_{wi} \rangle = \{\ a_i\ :\ a_{ci} - a_{wi} \leqq a_i \leqq a_{ci} + a_{wi},\ \ a_i \in R\ \} \tag{8}$$

where R is the set of real numbers and the radius a_{wi} should be non-negative. From interval arithmetic[3,4], the linear interval model (5) can be calculated as follows.

$$Y(\mathbf{x}_j) = \langle y_c(\mathbf{x}_j),\ y_w(\mathbf{x}_j) \rangle \quad \text{for } j=1,2,\cdots,m \tag{9}$$

where

$$y_c(\mathbf{x}_j) = a_{c0} + a_{c1}x_{1j} + \cdots + a_{cn}x_{nj}, \tag{10}$$

$$y_w(\mathbf{x}_j) = a_{w0} + a_{w1}\ \mathrm{abs}(x_{1j}) + \cdots + a_{wn}\ \mathrm{abs}(x_{nj}). \tag{11}$$

Therefore the constraint condition (6) can be represented as follows.

$$y_c(\mathbf{x}_j) - y_w(\mathbf{x}_j) \leqq y_j \leqq y_c(\mathbf{x}_j) + y_w(\mathbf{x}_j) \quad \text{for } j=1,2,\cdots,m. \tag{12}$$

It should be noted that $y_c(\mathbf{x}_j)$ and $y_w(\mathbf{x}_j)$ are linear functions of the center a_{ci} and the radius a_{wi} of the interval coefficient A_i.

In interval regression, we set up the following objective function in order to determine the interval linear regression model.

$$\text{Minimize } (\ y_w(\mathbf{x}_1) + y_w(\mathbf{x}_2) + \cdots + y_w(\mathbf{x}_m)\ ) \tag{13}$$

which means that the sum of the radiuses of the predicted intervals $Y(\mathbf{x}_j)$, $j=1,2,\cdots,m$, is minimized. This objective function is also a linear function of the center a_{ci} and the radius a_{wi} of the interval coefficient A_i. Therefore the following LP problem is formulated to determine the interval linear regression model.

[LP problem]

$$\text{Minimize} \quad z = y_w(\mathbf{x}_1) + y_w(\mathbf{x}_2) + \cdots + y_w(\mathbf{x}_m) \tag{14}$$

$$\text{subject to} \quad y_c(\mathbf{x}_j) - y_w(\mathbf{x}_j) \leqq y_j \leqq y_c(\mathbf{x}_j) + y_w(\mathbf{x}_j) \quad \text{for } j=1,2,\cdots,m \tag{15}$$

$$a_{wi} \geqq 0 \quad \text{for } i=0,1,\cdots,n \tag{16}$$

2.2 Numerical examples

[Example 1] Let us consider the following input-output data points.

Data index	j	1	2	3	4	5	6	7	8
Input value	x_j	2	4	6	8	10	12	14	16
Output value	y_j	14	16	14	18	18	22	18	22

From the LP problem, the following interval regression model is obtained.

$$Y(x_j) = \langle 12,\ 1 \rangle + \langle 0{:}625,\ 0.125 \rangle\ x_j \quad \text{for } j=1,2,\cdots,8 \tag{17}$$

This interval regression model is shown in Fig.1. We can see that all the given data points are included in the interval regression model (17).

[Example 2] Next, let us consider the following input-output data points.

Data index	j	1	2	3	4	5	6	7	8	9	10
Input value	x_j	2	4	6	8	10	12	14	16	2	16
Output value	y_j	14	16	14	18	18	22	18	22	4	32

The ninth and tenth data points, which may be outliers, are added to the data in Example 1. The following interval regression model is obtained from the LP problem.

$$Y(x_j) = \langle 6.548,\ 4.881 \rangle + \langle 1.226,\ 0.060 \rangle\ x_j \quad \text{for } j=1,2,\cdots,10 \tag{18}$$

This model is shown in Fig.2. Since the interval regression model (18) is directly influenced by the two outliers, this model is quite different from (17) in Fig.1.

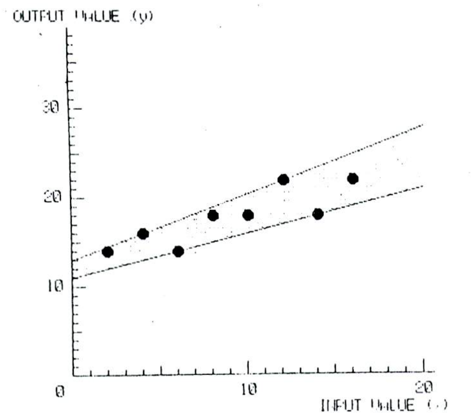

Fig.1 Interval regression model for Example 1 obtained by the LP problem

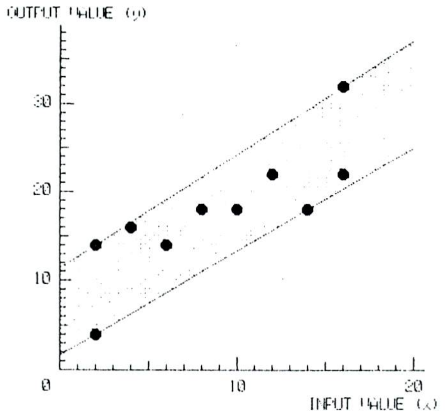

Fig.2 Interval regression model for Example 2 obtained by the LP problem

3. FORMULATION OF A COMBINATORIAL OPTIMIZATION PROBLEM

3.1 Formulation

As is shown in Fig.2, the LP problem (14)-(16) is sensitive to outliers. Therefore we formulate a combinatorial optimization problem to deal with outliers.

First we relax the constraint condition (15) as follows.

$$y_c(\mathbf{x}_j) - y_w(\mathbf{x}_j) \leqq y_j \leqq y_c(\mathbf{x}_j) + y_w(\mathbf{x}_j) \quad \text{for} \quad j \in J \tag{18}$$

where J is an unknown subset of $\{1,2,\cdots,m\}$. In order to require that the constraint condition (18) holds for at least k data points out of the given m data points, we impose the following condition on the subset J.

$$\mathrm{Card}(J) = k \tag{19}$$

where Card(J) represents the cardinality of J and k is an integer such that $k < m$.

The objective function (14) is also modified as follows.

$$\text{Minimize} \quad z = \sum_{j \in J} y_w(\mathbf{x}_j) \tag{20}$$

Therefore the following combinatorial optimization problem is formulated.

[Combinatorial optimization problem]

$$\text{Minimize} \quad z = \sum_{j \in J} y_w(\mathbf{x}_j) \tag{21}$$

$$\text{subject to} \quad y_c(\mathbf{x}_j) - y_w(\mathbf{x}_j) \leqq y_j \leqq y_c(\mathbf{x}_j) + y_w(\mathbf{x}_j) \quad \text{for} \quad j \in J, \tag{22}$$

$$a_{w_i} \geqq 0 \quad \text{for } i=0,1,\cdots,n, \tag{23}$$

$$\mathrm{Card}(J) = k, \tag{24}$$

$$J \subset \{1,2,\cdots,m\}. \tag{25}$$

3.2 Numerical example

Let us consider Example 2 in Section 2.2 again. Since two data points may be outliers, we set k=8. This means that a linear interval model should include at least eight data points out of the given ten data points. From the combinatorial optimization problem (21)-(25), the following linear interval model is obtained.

$$Y(x_j) = \langle 12,\ 1\rangle + \langle 0.625,\ 0.125\rangle\, x_j \quad \text{for} \quad j=1,2,\cdots,10. \tag{26}$$

It should be noted that this model is the same as (17) obtained from the data in Example 1 which have no outliers. In Fig.3, We show this model together with the given data points. We can see that two outliers are not included in this model.

4. FORMULATION OF A TWO-OBJECTIVE PROGRAMMING PROBLEM

4.1 Formulation

Since there are many cases where it is difficult for users of regression analysis to determine an appropriate value of k in the combinatorial optimization problem, let us consider the parameter k as an objective function. Therefore the following two-objective combinatorial optimization problem is formulated.

[Two-objective problem]

$$\text{Minimize} \quad z = \sum_{j\in J} y_w(\mathbf{x}_j) \qquad \text{and} \qquad \text{maximize } k \tag{27}$$

subject to (22)-(25).

4.2 Numerical example

Let us consider Example 2 in Section 2.2 again. Nondominated solutions of the two-objective problem can be obtained by solving the combinatorial optimization problem (21)-(25) with $k=1,2,\cdots,10$. We show ten nondominated solutions in Fig.4. Users of regression analysis can interactively choose a linear interval model using Fig.4. In Choosing process, we can employ the following heuristics: a large value of z suggests that the corresponding interval regression model may be influenced by outliers. The nondominated solutions corresponding to k=10 and k=9 are not good solutions since the corresponding values of z are too large in Fig.4. Therefore we choose the following linear interval model corresponding to k=8.

$$Y(x_j) = \langle 12,\ 1\rangle + \langle 0.625,\ 0.125\rangle\, x_j \qquad \text{for } j=1,2,\cdots,10. \tag{28}$$

If we are not satisfied with this model, we can choose another model using Fig.4. For example, the interval regression model corresponding to k=7 is as follows.

$$Y(x_j) = \langle 12.5,\ 1.5\rangle + \langle 0.5,\ 0\rangle\, x_j \qquad \text{for } j=1,2,\cdots,10. \tag{29}$$

This model is similar to (28) since both models include no outliers.

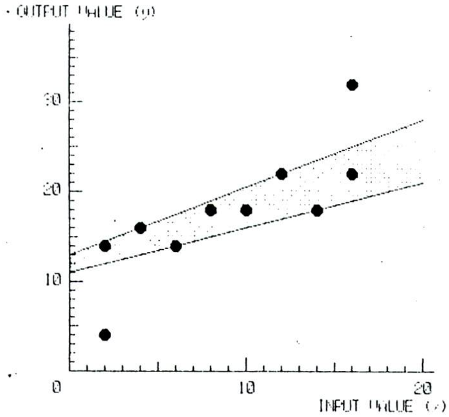

Fig.3 Interval regression model obtained by the combinatorial optimization problem

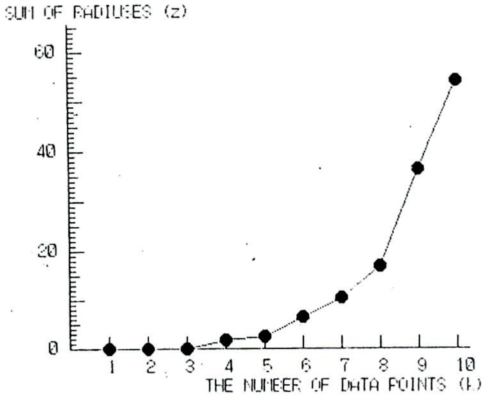

Fig.4 Nondominated solutions of the two-objective problem

5. CONCLUSION

We propose several formulations of regression analysis with linear interval regression models. Our main contribution is to propose a new idea for dealing with the data which may have some outliers. Since the interval regression model discussed in this paper can be considered as one version of fuzzy regression models[1,2], our method can be applied to fuzzy regression analysis.

REFERENCES

1. Tanaka,H., Uejima,S. and Asai,K.: Linear Regression Analysis with Fuzzy Model, IEEE Trans. Syst. Man. Cybern., Vol.SMC-12, No.12, pp.903-907 (1982)

2. Tanaka,H.: Fuzzy Data Analysis by Possibilistic Linear Model, Fuzzy Sets and Systems, Vol.24, No.3, pp.363-375 (1987)

3. Moore,R.E.: Methods and Applications of Interval Analysis, SIAM, Philadelphia (1979)

4. Alefeld,G and Herzberger,J: Introduction to Interval Computations, Academic Press, New York (1983)