Least Absolute Deviation Estimators for Interval Regression

Seung Hoe Choi [1] and James J. Buckley [2]

Abstract

In this paper we introduce the least absolute deviation estimators to construct an interval regression model, having interval output and crisp input data. Two numerical examples are presented comparing a performance of the proposed model using the least absolute deviation estimators with the interval regression model based on the least squares method when the data contains interval outliers.

1. Introduction

Regression analysis is a statistical method to estimate any functional relationship that might exist among a dependent variable and one or more independent variables. Applications of regression analysis exist in natural science, technology, economics, education and etc. The goal of regression analysis is to determine the values of parameters for a function that cause the function to best fit a set of data observations. Tanaka and coworker[4-6] introduced initially an interval linear regression model, which has interval parameter and suggested a linear programming method to estimate interval parameters. Inuiguchi et al.[3] and Buckley and Choi[1] used a least squares method to construct the interval regression model. However, the sensitive of the least squares estimators has been well known in the regression analysis. Therefore, we need robust methods, which are less sensitive to some outliers, to estimate interval coefficients in the interval regression model.

This paper deals with the interval regression using the least absolute deviation estimators, and compare the performance of our proposed model with the interval regression model based on the least squares method.

2. Interval Least Absolute Deviation Model

In this paper we consider the following interval regression model:

$$\begin{aligned} Y_i &= Y(\mathbf{x}_i) + E_i \\ &= (f(\mathbf{x}_i, C), g(|\mathbf{x}_i|, W)) + E_i, \end{aligned}$$

where Y_i is an observed interval with the center y_i^c and the width y_i^w, $f$and g are real-valued functions, E_i is an interval error with the center e_i^c and width e_i^w, $C = (c_0, \cdots, c_p)$, $W = (w_0, \cdots, w_p)$ and $|\mathbf{x}_i| = (|x_{i1}|, \cdots, |x_{ip}|)$.

When crisp input data and interval output data are given as

$$\mathbf{x}_\mathrm{i} = (x_{i1}, \cdots, x_{ip}) \quad \text{and} \quad Y_i = (y_i^c, y_i^w),$$

an interval least absolute deviation estimators based on $(\mathbf{x}_i : Y_i)$, denoted by $(\widehat{C}, \widehat{W})$, is defined as the value minimizing the following functions

$$\sum_{i=1}^{n} |y_i^c - f(\mathbf{x}_i, C)|$$

and

$$\sum_{i=1}^{n} |y_i^w - g(|\mathbf{x}_i|, W)|$$

subject to

$$w_j \geq 0 \text{ for each } j.$$

Then, the interval least absolute deviation regression(ILADR) model is as follow:

$$\widehat{Y}_i = (\hat{y}_i^c, \hat{y}_i^w) = \left(f(\mathbf{x}_i, \widehat{C}), g(|\mathbf{x}_i|, \widehat{W}) \right).$$

Now, we consider a method to construct an upper and a lower regression model using the least absolute deviation estimators. For the upper regression model (Y_i^*) which is to satisfy $Y_i \subseteq Y_i^*$ for each i, let $l_{y_i} = y_i^c - y_i^w$, $r_{y_i} = y_i^c + y_i^w$, $l_{\hat{y}_i} = \hat{y}_i^c - \hat{y}_i^w$ and $r_{\hat{y}_i} = \hat{y}_i^c + \hat{y}_i^w$. Then we get a revised width as follow:

$$\hat{y}_i^{w*} = \hat{y}_i^w \text{ if } Y_i \subseteq \widehat{Y}_i$$

and

$$\hat{y}_i^{w*} = \hat{y}_i^w + \max\{|l_{y_i} - l_{\hat{y}_i}|, |r_{y_i} - r_{\hat{y}_i}|\}$$

[1] Department of General Studies, Hankuk Aviation University, Koyang 411, Korea

[2] Department of Mathematics, University of Alabama at Birmingham, Birmingham, AL 35294, USA

when $Y_i \not\subseteq \widehat{Y}_i$.

Thus, we obtain the upper estimation model based on the least absolute deviation estimators, denoted by $\widehat{Y}^*_i = (\hat{y}^{c*}_i, \hat{y}^{w*}_i)$.

For the lower regression model (Y_{i*}) which is to satisfy $Y_{i*} \subseteq Y_i$ for each i, we consider the least absolute deviation estimators with constrained conditions, denoted by $(\widetilde{C}, \widetilde{W})$, as follow:

$$\sum_{i=1}^{n} |y^c_i - f(\mathbf{x}_i, C)| = \min!$$

subject to

$$|f(\mathbf{x}_i, C) - y^c_i| \leq y^w_i$$

and

$$\sum_{i=1}^{n} |y^w_i - g(|\mathbf{x}_i|, W)| = \min!$$

subject to

$$w_j \geq 0 \text{ for each } j.$$

In order to estimate the lower regression model, let $\tilde{y}^c_i = f(\mathbf{x}_i, \widetilde{C})$ and $\tilde{y}^w_i = g(|\mathbf{x}_i|, \widetilde{W})$. Then we use a new width of the predicted interval as follow:

$$\hat{y}^w_{i*} = \tilde{y}^w_i \text{ if } \widetilde{Y}_i \subseteq Y_i$$

and

$$\hat{y}^w_{i*} = \min\{\tilde{y}^w_i, |\tilde{y}^c_i - r_{y_i}|, |\tilde{y}^c_i - l_{y_i}|\}$$

when $\widetilde{Y}_{i*} \not\subseteq Y_i$.

Thus, from the above width we obtain the lower estimation model, denoted by $\widehat{Y}_{i*} = (\hat{y}^c_{i*}, \hat{y}^w_{i*})$ where $\hat{y}^c_{i*} = f(\mathbf{x}_i, \widetilde{C})$.

3. Numerical Examples

This section defines an interval outlier for the interval regression model, and suggests a measure to investigate a performance of the proposed model and the interval regression model using the least squares method. For this, let $\overline{Y}_i = (\bar{y}^c_i, \bar{y}^w_i)$ be a predicted interval of the observed interval Y_i and k a positive number. Also, let $q^c_1(q^c_3)$ be the first(third) quantiles of the set $\{r^c_i : r^c_i = |\bar{y}^c_i - y^c_i|, i = 1, \cdots, n\}$ and $q^w_1(q^w_3)$ the first(third) quartiles of the set $\{r^w_i : r^w_i = |\bar{y}^w_i - y^w_i|, i = 1, \cdots, n\}$, respectively.

A C-type interval outlier $Y^o_i = (y^{co}_i, y^{wo}_i)$ is defined as

$$r^{co}_i < q^c_1 - k(q^c_3 - q^c_1)$$

or

$$r^{co}_i > q_3 + k(q^c_3 - q^c_1)$$

where $r^{co}_i = |\bar{y}^{co}_i - y^c_i|$. A W-type interval outlier $Y^o_i = (y^{co}_i, y^{wo}_i)$ satisfies

$$r^{wo}_i < q^w_1 - k(q^w_3 - q^w_1)$$

or

$$r^{wo}_i > q_3 + k(q^w_3 - q^w_1)$$

where $r^{wo}_i = |\bar{y}^{wo}_i - y^w_i|$. Specially, when $k = 1.5$ and $k = 3$, we will called the outlier as mild outlier and extreme outlier, respectively. Buckley and Choi[1] introduced the length of the symmetric difference between intervals to compare interval regression models. The length of the symmetric difference for two intervals A_i and A_j, denoted by $l(A_i \triangle A_j)$, is equal to

$$\begin{cases} |l_i - l_j| + |r_i - r_j| & \text{if } A_i \cap A_j \neq \varnothing \\ (r_i - l_i) + (r_j - l_j) & \text{if } A_i \cap A_j = \varnothing, \end{cases}$$

where $l_i(l_j)$ and $r_i(r_j)$ are the left and the right endpoints of the interval $A_i(A_j)$. Now, we consider the following measure as a criterion to investigate the performance of an estimated interval regression model $\widehat{Y}_i$

$$M_Y = \sum_{i=1}^{n} M_{\widehat{Y}_i}$$

where

$$M_{\widehat{Y}_i} = \begin{cases} \frac{m(Y_i \triangle \hat{Y}_i)}{l(Y_i)}, & \text{if} \quad l(Y_i) > 1 \\ m(Y_i \triangle \hat{Y}_i), & \text{if} \quad l(Y_i) \leq 1, \end{cases}$$

$l(Y_i)$ denotes a length of the interval Y_i, and $m(Y_i \triangle \hat{Y}_i) = l(Y_i \triangle \hat{Y}_i) + \min\{|s_i - \hat{s}_j| : s_i \in Y_i, \hat{s}_i \in \hat{A}_i\}$, and . On the other hand, they suggested a measure M^*_Y to consider the performance of the estimated upper regression model $\widehat{Y}^*_i$ and lower regression model $\widehat{Y}_{i*}$ simultaneously as follow:

$$M^*_Y = \sum_{i=1}^{n} M^*_{\widehat{Y}_i},$$

where

$$M^*_{\hat{Y}_i} = \begin{cases} \left(\frac{\hat{y}^{w*}_i - \hat{y}^{w}_{i*}}{y^w_i}\right) & \text{if} \quad i \in A \\ (\hat{y}^{w*}_i - \hat{y}^{w}_{i*}) & \text{if} \quad i \in A^C. \end{cases}$$

and A^C denotes the complement set of $A = \{i | l(Y_i) > 1\}$.

In the following examples we compare the performance of the ILADR model and the interval least squares regression(ILSR) model for the data with the interval outliers.

Example 3.1 Table 3.1 shows the numerical value used by Tanaka et al.[7] and Chen[2] with modifications to introduce abnormal value in the fuzzy regression.

Input	Output Y_i		Errors $M_{\hat{Y}_i}$		Errors $M^*_{\hat{Y}_i}$	
x_i	y^c_i	y^w_i	ILSR	ILADR	ILSR	ILADR
1	8	1.8	0.93	1.28	5.687	6.79
2	6.4	2.2	0.81	0.55	5.605	4.09
3	9.5	2.6	0.2	0.12	1.437	0.62
4	13.5	2.6	0.63	0.81	4.791	5.79
5	13	2.4	0.29	0.13	2.201	0.89
6	15.2	2.3	0.15	0	1.233	0.02
7	17	2.2	0.18	0.05	1.465	0.29
8	19.3	**4.8**	0.48	0.52	4.64	5.02
9	20.1	1.9	0.52	0.42	3.529	2.49
10	24.3	2	0.69	0.75	4.599	5.19
Total error			4.88	4.63	35.187	31.19

Table 3.1 : Data and Error in Example 3.1

The eighth observation in Table 3.1 is a W-type extreme outlier. The result in Table 3.1 explains that the ILADR model is better than the ILSR model in this example with the interval outlier.

Example 3.2 The underlying model considered in this example is given in the following function:

$$Y_i = (\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2}, \beta_0) + (e^c_i, e^w_i),$$

where $\beta_0 = 3.3, x_{i2} = x^2_{i1}$ and the coefficients $(\alpha_0, \alpha_1, \alpha_2) = (2.5, 1.4, 0.9)$.

Input	Output Y_i		Errors $M_{\hat{Y}_i}$		Errors $M^*_{\hat{Y}_i}$	
x_i	y^c_i	y^w_i	ILSR	ILADR	ILSR	ILADR
1	3.3	6	0.27	0.25	3.71	3.21
1.5	5.125	6.1	0.35	0.33	4.47	4.09
2	7.7	5.1	0.45	0.42	4.56	5.04
2.5	17.625	5.1	0.89	0.92	9.22	5.45
3	19.6	5.9	0.51	0.55	6.28	7.17
3.5	19.325	6.1	0.18	0.14	2.14	1.84
4	22	5.6	0.48	0.43	5.18	5.04
4.5	30.425	5.3	0.21	0.28	2.3	3.55
5	33.8	5.7	0.11	0.04	1.18	0.96
5.5	41.325	5.7	0.28	0.37	2.9	3.85
6	43.7	4.4	0.42	0.27	5.55	6.06
6.5	46.725	5.8	0.86	0.73	10.78	4.92
7	56.5	4.6	0.39	0.2	5.7	6.01
7.5	**72.225**	5.4	1.3	1.48	12.38	8.44
8	71.4	4.4	0.26	0.25	5.76	4.31
Total error			6.96	6.66	82.11	69.93

Table 3.2 : Data and Error in Example 3.2

A random sample of size 15 is generated as follows:

The input x_{i1} satisfies $x_{11} = 1$ and $x_{(i+1)1} = x_{i1} + 0.5$.

The center e^c_i is distributed under Cauchy distribution with local zero and scale two.

The width e^w_i is uniformly distributed within the interval $[1, 3]$.

The fourteenth observation in Table 3.2 is a C-type mild interval outlier. Table 3.2 shows that the ILADR model is better than the ILSR model for this example.

4. Conclusion

In this paper we introduced the interval regression model using the least absolute deviation estimators. This paper has shown that the proposed model is better than the interval regression model based on the least squares method when there are interval outliers.

Reference

(1) Buckley J. J. and Choi S. H. (2006) Interval Regression using Symmetric Difference, under review.

(2) Chen Y. (2001) Outlier detection and confidence interval modification in fuzzy regression. Fuzzy Sets and Systems **119**: 259-272

(3) Inuiguchi M., Fujita H., Tanino T. (2001) Interval linear regression analysis bases on Minkowski difference. Proc. International Conference on Information Systems. Analysis and Synthesis **7**: 112-117

(4) Tanaka H., Hayashi I., Watada J. (1987) Interval regression analysis. 3rd Fuzzy System Symposium: 9-12

(5) Tanaka H., Hayashi I., Watada J. (1989) Possibilistic linear regression analysis for fuzzy data. Eur J Oper Res **40**: 389-396

(6) Tanaka H., Lee H. (1998) Interval regression analysis by quadratic programming approach. IEEE Transactions on Fuzzy Systems **6**: 473-481

(7) Tanaka H, Uejima S, Asai K (1982) Linear regression analysis with fuzzy model, IEEE Trans. Systems. Man Cybernet **12**: 903-907