New approach to interval linear regression

Milan Hladík [1] Michal Černý [2]

[1] Faculty of Mathematics and Physics,
Charles University in Prague, Czech Republic

[2] Faculty of Computer Science and Statistics,
University of Economics, Prague, Czech Republic

MEC EurOPT 2010, İzmir, Turkey
June 23 – 26

Linear regression

Model

$$y_j = x_{j,1}a_1 + x_{j,2}a_2 + \cdots + x_{j,n}a_n = X_{j*}a, \quad j = 1, \ldots, p,$$

or,

$$Xa,$$

where $X \in \mathbb{R}^{p \times n}$ is an input matrix, and $y \in \mathbb{R}^p$ is an output vector.

Problem

Find a the best approximation to $y = Xa$.

Methods

- L_2-norm estimate (least squares method);
- L_1-norm estimate (least abscissae method);
- L_∞-norm estimate;
- ...

Interval linear regression

Model (crisp input – crisp output)

$$y_j = x_{j,1}\mathbf{a}_1 + x_{j,2}\mathbf{a}_2 + \cdots + x_{j,n}\mathbf{a}_n = X_{j*}\mathbf{a}, \quad j = 1, \ldots, p,$$

or,

$$X\mathbf{a},$$

where $\mathbf{a}_i = [a_i - c_i, a_i + c_i]$, $i = 1, \ldots, n$, are intervals.

Problem

Find as narrow as possible interval vector $\mathbf{a}$ such that $y \subseteq X\mathbf{a}$, i.e.,

$$\forall j = 1, \ldots, p \ \exists a' \in \mathbf{a} : \ y_j = X_{j*}a'.$$

Interval linear regression

Applications

- Ergonomics (Chang et al., 1996);
- Market sales forecasting (Heshmaty & Kandel, 1985);
- System identification (Kaneyoshi et al., 1990);
- Speech learning systems (Liu, 2009).

Methods

(1) Linear programming formulation (Lee & Tanaka, 1999, Tanaka, 1987, Tanaka & Watada, 1988);

(2) Quadratic programming formulation (Tanaka & Lee, 1998);

(3) Support vector machines (Hao, 2009, Huang & Kao, 2009).

Pros/Cons

(1) simple, but some parameters crisp,

(2)-(3) more complex, but small improvement.

New approach

Preliminary

Find an interval vector in the form $\mathbf{a} = [a - \delta c_\Delta, a + \delta c_\Delta]$, where

- a is an initial real-valued estimate to $y = Xa$;
- $c_\Delta \geq 0$ is given (usually $c_\Delta = |a|$ or $c_\Delta = 1$);
- $\delta \geq 0$ is a tolerance quotient in demand.

Theorem

If there is $j \in \{1, \ldots, p\}$ such that $|X|_{j*} c_\Delta = 0$ and $y_j \neq X_{j*}a$ then there exists no allowable δ. Otherwise let

$$\delta^* := \max_{j:|X|_{j*}c_\Delta > 0} \frac{|y_j - X_{j*}a|}{|X|_{j*}c_\Delta},$$

where $\max \emptyset = 0$ *by definition. Then δ^* is the minimal tolerance quotient.*

New approach

Properties

- Very cheap to calculate δ^*.
- The widths of interval in $\mathbf{a} = [a - \delta c_\Delta, a + \delta c_\Delta]$ are proportional to c_Δ and minimal in some sense.
- Easy to interpret δ^*.
- δ^* can be used to measure a fitness of a to the model.
- But: δ^* highly depends on the initial estimate a.

Examples: house price model

Example (House price model, Lee & Tanaka, 1999)

$$y = \mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \mathbf{a}_3 x_3 + \mathbf{a}_4 x_4.$$

j	y	x_1	x_2	x_3	x_4
1	606	1	1	38.09	36.43
2	710	1	1	62.10	26.50
3	808	1	1	63.76	44.71
4	826	1	1	74.52	38.09
5	865	1	1	75.38	41.10
6	852	1	2	52.99	26.49
7	917	1	2	62.93	26.49
8	1031	1	2	72.04	33.12
9	1092	1	2	76.12	43.06
10	1203	1	2	90.26	42.64
11	1394	1	3	85.70	31.33
12	1420	1	3	95.27	27.64
13	1601	1	3	105.98	27.64
14	1632	1	3	79.25	66.81
15	1699	1	3	120.5	32.25

x_1 ... absolute term,
x_2 ... quality of material,
x_3 ... the area of the first floor (m^2),
x_4 ... the area of the second floor (m^2),
y ... the sale price (10,000 JPY).

Examples: house price model

Example (cont.)

Solution by a linear programming-based method (Tanaka & Lee, 1998)

$$y = [0,\ 0] + [208,\ 283]x_2 + [5.85,\ 5.85]x_3 + [4.79,\ 4.79]x_4.$$

The quadratic programming-based method (Tanaka & Lee, 1998):

$$y = [-8.81,\ 8.81] + [209,\ 259]x_2 + [6.14,\ 6.14]x_3 + [4.41,\ 5.39]x_4.$$

Our method:

- $a := (X^T X)^{-1} X^T y = (-239.28, 264.84, 7.47, 6.76)^T$,
- $c_\Delta := |a|$,
- calculate $\delta^* = 0.0482$.

It means that all entries of a perturb within 4.82% tolerance:

$$y = [-251,\ -228] + [252,\ 278]x_2 + [7.11,\ 7.83]x_3 + [6.43,\ 7.08]x_4.$$

Examples: outliers

Example (Outliers, Ishibuchi & Tanaka, 1990)

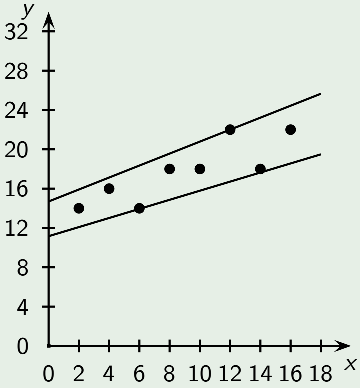

Figure: Basic interval regression model without outliers.

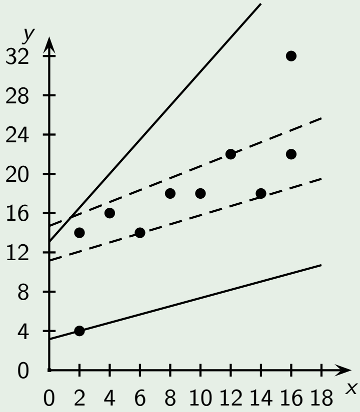

Figure: Basic interval regression model with outliers.

Examples: outliers

Example (Outliers, Ishibuchi & Tanaka, 1990)

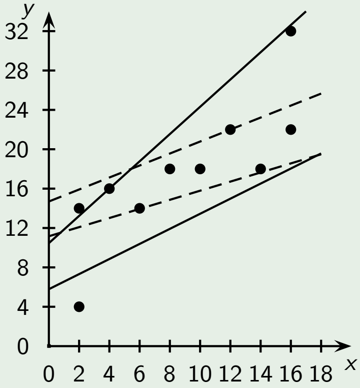

Figure: Improved interval regression model; a is obtained by least squares.

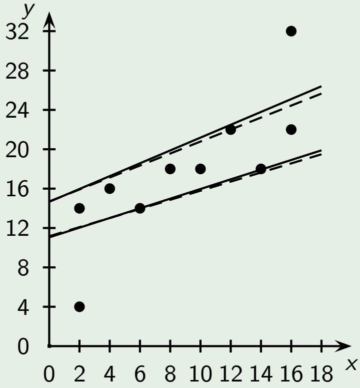

Figure: Improved interval regression model; a is obtained by L_1 regression.

Examples: risk management

Example (Price development of oil and kerosene)

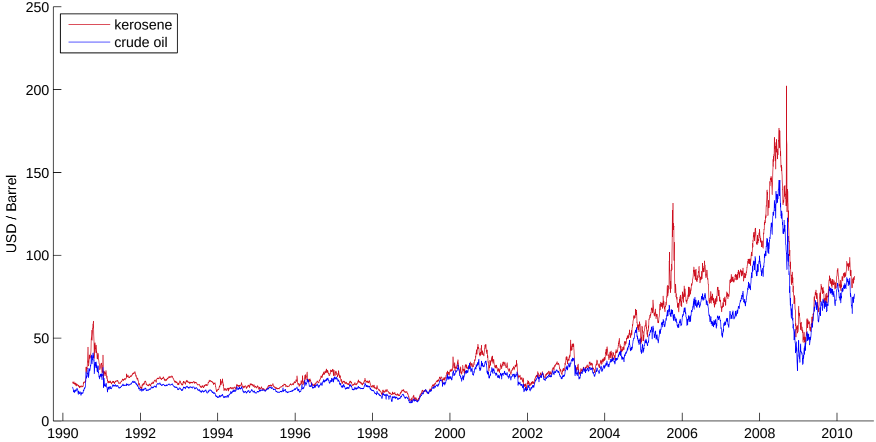

Figure: Evolution of prices of kerosene and crude oil (WTI) in dollars per barrel.

Examples: risk management

Example (cont.)

- An airline company may hedge expected future purchases of kerosene by crude oil futures.
- A proper hedge ratio must be determined to decrease the market risk.

Assume that price of kerosene y is a linear function of the price of oil x:

$$y = a_1 + a_2 x,$$

A least square estimate of a is $a = (-1.319, 1.225)^T$ and the model reads

$$y = -1.319 + 1.225x.$$

- Put $c_\Delta := |a|$ and calculate $\delta^* = 0.7204$.
- Hence the hedge ratio a_2 is quite unstable;
- Removing 10 of the worst outliers we decrease δ^* to 0.5053;
- Removing 100 of the worst outliers we decrease δ^* to 0.1758.

Conclusion and future work

Conclusion

- The proposed method is more flexible;
- The widths of the resulting interval parameters are constructed proportionally;
- The tolerance quotient is easily interpreted by a user;
- It can used as a fitness measure of the model;
- Outliers are easy to detect and handle.

Future work

- Theoretical properties of the quotient as a fitness measure.
- Extension of the method to crisp input – interval output models.
- Extension of the method to interval input – interval output models.