Regression analysis of interval data

Lev Utkin

Munich, September 2009

Standard regression models

- Suppose that we have two variables Y and X with Y being a dependent variable and X being predictor variable, related to Y according to the relation $Y = f(X)$.
- The simplest case: the linear model $Y = bX + c + \epsilon$. Here b and c are parameters and ϵ is the random errors or the noise having zero mean and the unknown variance σ^2.
- A linear regression model fits a linear function to a set of data points. When the variable X takes n specific values $x_1, ..., x_n$, the variables Y and ϵ take specific values y_i and ϵ_i, respectively, $i = 1, ..., n$, we get

$$y_i = bx_i + c + \epsilon_i, \ \ i = 1, ..., n.$$

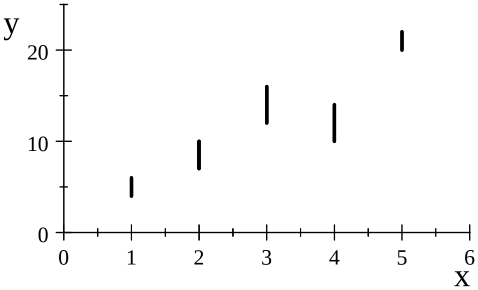
Interval data

- Suppose now that we have interval-valued observations $\mathbf{y}_i = [\underline{y}_i, \overline{y}_i]$ instead of the point-valued ones y_i, $i = 1, ..., n$.
- The simplest way is to randomly take points y_{ij} from $\mathbf{y}_i$, $i = 1, ..., n$, and to construct the j-th standard regression model with the corresponding parameters b_j and c_j. After taking M points, M regression models are constructed. Then the intervals for parameters b and c are determined as $\underline{b} = \min_j b_j$, $\overline{b} = \max_j b_j$ and $\underline{c} = \min_j c_j$, $\overline{c} = \max_j c_j$.
- The main **shortcoming** is that this way provides too wide and often non-informative intervals of the parameters.

The main idea of the proposed approach

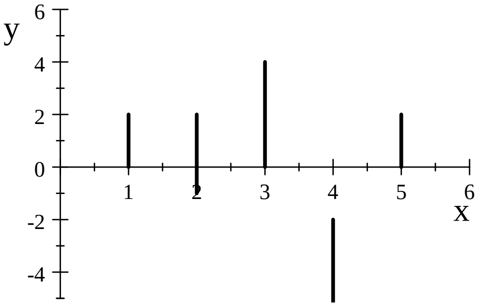
The proposed approach does not use the well-known common way for minimizing the deviation of the observed points from the "optimal" regression function f*. It maximizes the "density" or "overcrowding" of the biased intervals* $\mathbf{y}_i - f(x_i)$*. In other words, the biased intervals have to be maximally overlapped. In this case, the approach is invariant to possible changes of the interval width with changing* X*.*

Overlapping intervals (1)



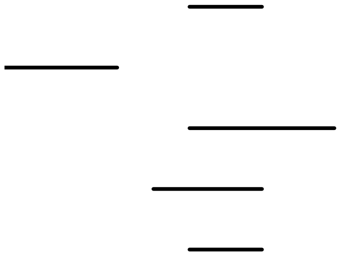
Observed intervals

Overlapping intervals (2)



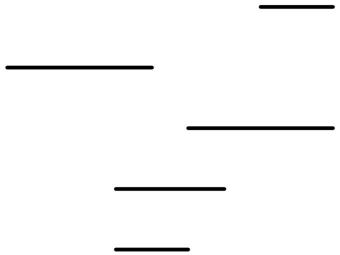
Observed intervals minus $f(x) = 4x$

Overlapping intervals (3)



Observed intervals minus $f(x) = 4x$ in another form

Overlapping intervals (4)



Observed intervals minus $f(x) = 3x$ in another form

The extended imprecise Dirichlet model

Denote

$$\underline{z}_i = \underline{y}_i - bx_i - c, \ \overline{z}_i = \overline{y}_i - bx_i - c.$$

1. Belief and plausibility functions (Dempster-Shafer theory) are

$$\mathrm{Bel}(A|\mathbf{Z}) = \frac{\sum_{i:\mathbf{z}_i \subseteq A} 1}{n}, \quad \mathrm{Pl}(A|\mathbf{Z}) = \frac{\sum_{i:\mathbf{z}_i \cap A \neq \varnothing} 1}{n}.$$

The extended imprecise Dirichlet model

Denote

$$\underline{z}_i = \underline{y}_i - bx_i - c,\ \overline{z}_i = \overline{y}_i - bx_i - c.$$

1. Belief and plausibility functions (Dempster-Shafer theory) are

$$\mathrm{Bel}(A|\mathbf{Z}) = \frac{\sum_{i:\mathbf{z}_i \subseteq A} 1}{n},\quad \mathrm{Pl}(A|\mathbf{Z}) = \frac{\sum_{i:\mathbf{z}_i \cap A \neq \varnothing} 1}{n}.$$

2. Extended belief and plausibility functions with the cautious parameter s are

$$\underline{P}(A|\mathbf{Z}, s) = \frac{\sum_{i:\mathbf{z}_i \subseteq A} 1}{n+s},\quad \overline{P}(A|\mathbf{Z}, s) = \frac{s + \sum_{i:\mathbf{z}_i \cap A \neq \varnothing} 1}{n+s}.$$

The likelihood function and its maximum

The likelihood function is

$$L(\mathbf{Z}) = \Pr\left\{\underline{z}_1 \leq \epsilon_1 \leq \overline{z}_1, ..., \underline{z}_n \leq \epsilon_n \leq \overline{z}_n\right\}.$$

Proposition

If random variables $\epsilon_1, ..., \epsilon_n$ are independent, then there holds

$$\max_{\mathbf{M}} L(\mathbf{Z}) = \prod_{j=1}^{n} \left\{\overline{P}(\mathbf{z}_j|\mathbf{Z}, s) - \underline{P}(\mathbf{z}_j|\mathbf{Z}, s)\right\}.$$

The imprecise regression model

$$\max_{\mathbf{M}} L(\mathbf{Z}) = \frac{1}{(n+s)^n} \prod_{j=1}^{n} \left\{ s + \sum_{i:\mathbf{z}_i \cap \mathbf{z}_j \neq \varnothing} 1 - \sum_{i:\mathbf{z}_i \subseteq \mathbf{z}_j} 1 \right\}.$$

Parameters b and c are computed by maximizing the above function, i.e. by solving the problem

$$\prod_{j=1}^{n} \left\{ s + \sum_{i:\mathbf{z}_i \cap \mathbf{z}_j \neq \varnothing} 1 - \sum_{i:\mathbf{z}_i \subseteq \mathbf{z}_j} 1 \right\} \to \max_{b,c}.$$

Interesting properties

1. The objective function is invariant with respect to parameter c, i.e. its maximum does not depend on the parameter c.

Interesting properties

1. The objective function is invariant with respect to parameter c, i.e. its maximum does not depend on the parameter c.
2. The parameter b which maximizes the likelihood function is generally interval-valued with lower and upper bounds $\underline{b}$ and $\overline{b}$, respectively, i.e., the largest values of the likelihood function are achieved for $b \in [\underline{b}, \overline{b}]$.

Interesting properties

1. The objective function is invariant with respect to parameter c, i.e. its maximum does not depend on the parameter c.
2. The parameter b which maximizes the likelihood function is generally interval-valued with lower and upper bounds $\underline{b}$ and $\overline{b}$, respectively, i.e., the largest values of the likelihood function are achieved for $b \in [\underline{b}, \overline{b}]$.
3. For linear model, the parameter b does not depend on the caution parameter s.

Possible ways for computing the parameter c

1. The point-valued parameter c is computed as the mean value of $y_i' - b'x_i$, $i = 1, ..., n$. Here y_i' and b' are the middle points of intervals $\mathbf{y}_i$ and $[\underline{b}, \overline{b}]$, respectively.

Possible ways for computing the parameter c

1. The point-valued parameter c is computed as the mean value of $y_i' - b' x_i$, $i = 1, ..., n$. Here y_i' and b' are the middle points of intervals $\mathbf{y}_i$ and $[\underline{b}, \overline{b}]$, respectively.
2. The interval-valued parameter:

$$\underline{c} = n^{-1} \sum_{i=1}^{n} \left(\underline{y}_i - \overline{b} x_i \right), \ \overline{c} = n^{-1} \sum_{i=1}^{n} \left(\overline{y}_i - \underline{b} x_i \right).$$

Possible ways for computing the parameter c

1. The point-valued parameter c is computed as the mean value of $y_i' - b' x_i$, $i = 1, ..., n$. Here y_i' and b' are the middle points of intervals $\mathbf{y}_i$ and $[\underline{b}, \overline{b}]$, respectively.
2. The interval-valued parameter:

$$\underline{c} = n^{-1} \sum_{i=1}^{n} \left(\underline{y}_i - \overline{b} x_i \right), \ \overline{c} = n^{-1} \sum_{i=1}^{n} \left(\overline{y}_i - \underline{b} x_i \right).$$

3. Extended cautious mean values of c:

$$\underline{\mathbb{E}}_s X = (n+s)^{-1} \left(s \cdot \Omega_* + \sum_{i=1}^{n} \left(\underline{y}_i - \overline{b} x_i \right) \right),$$

$$\overline{\mathbb{E}}_s X = (n+s)^{-1} \left(s \cdot \Omega^* + \sum_{i=1}^{n} \left(\overline{y}_i - \underline{b} x_i \right) \right).$$

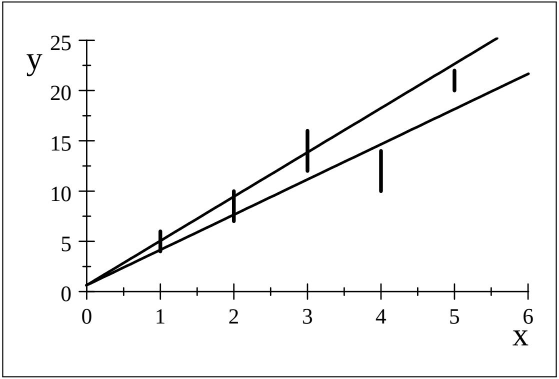
A numerical example (1)

5 pairs $(x_i, [\underline{y}_i, \overline{y}_i])$, $i = 1, ..., 5$.

i	x_i	$\underline{y}_i$	$\overline{y}_i$
1	1	4	6
2	2	7	10
3	3	12	16
4	4	10	18
5	5	20	22

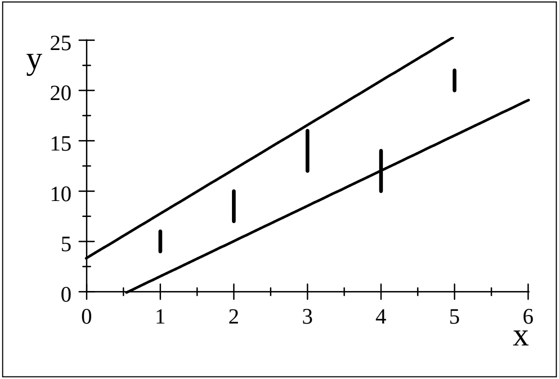
Solution 1: $\underline{b} = 3.5$, $\overline{b} = 4.4$ and $c = 0.65$.
Solution 2: $\underline{b} = 3.5$, $\overline{b} = 4.4$ and $\underline{c} = -1.97$, $\overline{c} = 3.36$.

A numerical example (2)



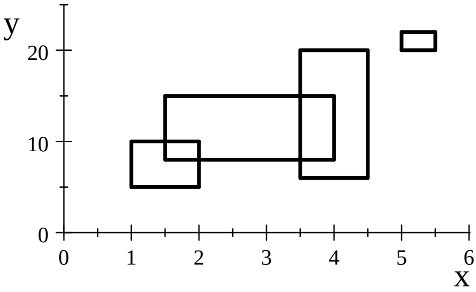
The relationship of intervals and linear functions by the first method for computing c

A numerical example (3)



The relationship of intervals and linear functions by the second method for computing c

A more general case



Observed values (x, y), $x \in [\underline{x}, \overline{x}]$, $y \in [\underline{y}, \overline{y}]$.

Questions

?