World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:5, No:9, 2011

Computational Aspects of Regression Analysis of Interval Data

Michal Černý

Abstract—We consider linear regression models where both input data (the values of independent variables) and output data (the observations of the dependent variable) are interval-censored. We introduce a possibilistic generalization of the least squares estimator, so called OLS-set for the interval model. This set captures the impact of the loss of information on the OLS estimator caused by interval censoring and provides a tool for quantification of this effect. We study complexity-theoretic properties of the OLS-set. We also deal with restricted versions of the general interval linear regression model, in particular the crisp input – interval output model. We give an argument that natural descriptions of the OLS-set in the crisp input – interval output cannot be computed in polynomial time. Then we derive easily computable approximations for the OLS-set which can be used instead of the exact description. We illustrate the approach by an example.

Keywords—linear regression; interval-censored data; computational complexity

I. Introduction

Consider the linear regression model

$$y = X\beta + \varepsilon \tag{1}$$

where y denotes the vector of observations of the dependent variable, X denotes the design matrix of the regression model, β denotes the vector of unknown regression parameters and ε is the vector of disturbances. We do not make any special assumptions on ε; we just assume that for estimation of β, a linear estimator can be used, i.e. an estimator of the form

$$\widehat{\beta} = Qy, \tag{2}$$

where Q is a matrix. In the following text, we shall concentrate on the Ordinary Least Squares (OLS) estimator, which corresponds to the choice $Q = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ in (2). Nevertheless, the theory is also applicable for other linear estimators, such as the Generalized Least Squares (GLS) estimator, which corresponds to the choice $Q = (X^{\mathrm{T}}\Omega^{-1}X)^{-1}\Omega^{-1}X^{\mathrm{T}}$ in (2), where Ω is either known or estimated covariance matrix of ε. Other examples include estimation methods which, at the beginning, exclude outliers and then apply OLS or GLS. These estimators are often used in robust statistics.

The symbol n stands for the number of observations and the symbol p stands for the number of regression parameters.

The tuple (X, y) is called *input data* for the model (1).

In this text we study computational properties of estimators for the model (1) when the input data cannot be observed directly; instead, only intervals are known such that that values of X and y are guaranteed to be contained in.

M. Černý, Department of Econometrics, University of Economics. Winston Churchill Square 4, 130 67 Prague, Czech Republic. E-mail: cernym@vse.cz.

A variety of methods for estimation of regression parameters in a regression with interval data has been developed; they are studied in statistics ([6], [14], [26], [29], [32], [34], [38], [46]), where also robust regression methods have been proposed ([22], [35]), in fuzzy theory ([15], [19], [20], [43], [44], [45]) as well as in computer science ([9], [21], [24]). An algebraic treatment of least squares methods for interval data has been considered in [4] and [12].

II. Interval data in linear regression models

A. Motivation

Inclusion of interval data in linear regression models is suitable for modeling of a variety of real-world problems. For example:

- The data (X, y) have been interval-censored. This is often the case of medical, epidemiologic or demographic data — only interval-censored data are published while the exact individual values are kept secret.
- Data are rounded. If we store data using data types of restricted precision, then instead of exact values we are only guaranteed that the true value is in an interval of width 2^{-d} where d is the number of bits of the data type for representation of the non-integer part. For example, if we store data as integers, then we know only the interval $[\tilde{y} - 0.5, \tilde{y} + 0.5]$ instead of the exact value y, where $\tilde{y}$ is y rounded to the nearest integer. This application is important in the theory of reliable computing.
- Sometimes, data are intervals by their nature. For instance, financial data have bid-ask spreads.
- Categorial data may be sometimes interpreted as interval data; for example, credit rating grades can be understood as intervals of credit spreads over the risk-free yield curve.

In econometric regression models, it is often the case that varying variables are represented by their average or median values. For example, if the exchange rate for a period of one year should be included in the regression model, usually the average rate of that year is taken. However, it might be more appropriate to regard the exchange rate as an interval inside which the variable changes.

More applications of interval data are found in econometrics [5], information science [8], ergonomics [7], optimization and operational research [10], [27], [30], [42], speeach learning [33] and in pattern recognition [28], [31].

B. Interval numbers, vectors and matrices

If two real matrices X_1, X_2 are of the same dimension, the relation $X_1 \leq X_2$ is understood componentwise.

Definition 1. (a) *If $-\infty < \underline{a} \leq \overline{a} < \infty$, the **interval number** $\boldsymbol{a}$ is the closed interval $[\underline{a}, \overline{a}]$.*
(b) *Let $\underline{X} \leq \overline{X}$ be two $M \times N$ real matrices. The* interval matrix $\boldsymbol{X} = [\underline{X}, \overline{X}]$ *is the set*

$$\{X \in \mathbb{R}^{M \times N} : \underline{X} \leq X \leq \overline{X}\}.$$

The *interval vector* $\boldsymbol{y} = [\underline{y}, \overline{y}]$ is a special case of the interval matrix with one column.

Interval numbers, vectors and matrices are typeset in boldface.

Arithmetic operations $+$ and $\times$ with interval numbers $\boldsymbol{a} = [\underline{a}, \overline{a}]$ and $\boldsymbol{b} = [\underline{b}, \overline{b}]$ are defined in a natural way (see [1]):

$$\begin{aligned} \boldsymbol{a} + \boldsymbol{b} &= [\underline{a} + \underline{b}, \overline{a} + \overline{b}], \\ \boldsymbol{a} \cdot \boldsymbol{b} &= [\min\{\underline{a}\underline{b}, \underline{a}\overline{b}, \overline{a}\underline{b}, \overline{a}\overline{b}\}, \max\{\underline{a}\underline{b}, \underline{a}\overline{b}, \overline{a}\underline{b}, \overline{a}\overline{b}\}]. \end{aligned} \tag{3}$$

From the definition, the following lemma is clear:

Lemma 2. *A finite sequence of sums and products of interval numbers is a bounded set.* □

C. The possibilistic approach to linear regression models with interval data

Assume that only intervals $(\boldsymbol{X}, \boldsymbol{y})$ are available instead of the exact values of the data (X, y) such that $X \in \boldsymbol{X}$ and $y \in \boldsymbol{y}$. Then, of course, we lose some information. The main aim of this text is to quantify the impact of the loss of information caused by interval censoring (or rounding) on the OLS estimator $\widehat{\beta}$. The next definition generalizes of the notion of the estimator $\widehat{\beta}$ for the case when the crisp values (X, y) in are replaced by intervals $(\boldsymbol{X}, \boldsymbol{y})$ in (1).

Definition 3. (a) *A tuple $(\boldsymbol{X}, \boldsymbol{y})$, where $\boldsymbol{X}$ is an $n \times p$ interval matrix and $\boldsymbol{y}$ is an $n \times 1$ interval vector, is called **an input** (or: **data**) **for an interval regression model**, or just **interval regression model** for short.*
(b) *The **OLS-set** of $(\boldsymbol{X}, \boldsymbol{y})$ is defined as*

$$\begin{aligned} \mathit{OLS}(\boldsymbol{X}, \boldsymbol{y}) = \{\beta \in \mathbb{R}^p : \\ (\exists X \in \boldsymbol{X})(\exists y \in \boldsymbol{y}) X^{\mathrm{T}} X \beta = X^{\mathrm{T}} y\}. \end{aligned}$$

The motivation for the definition is straightforward. Our aim is to use OLS to obtain an estimate of the unknown vector of regression parameters β in the model (1). However, observations of both dependent variables (y) and independent variables (X) are interval-censored; i.e., we only know intervals $\boldsymbol{X}$ and $\boldsymbol{y}$ that are guaranteed to contain the directly unobservable data (X, y). Then, the set $\mathit{OLS}(\boldsymbol{X}, \boldsymbol{y})$ contains *all possible* values of OLS-estimates of β as X and y range over $\boldsymbol{X}$ and $\boldsymbol{y}$, respectively. We say that $\mathit{OLS}(\boldsymbol{X}, \boldsymbol{y})$ is a *possibilistic* version of the notion of the OLS estimator.

The set $\mathit{OLS}(\boldsymbol{X}, \boldsymbol{y})$ captures the loss of information caused by interval censoring (or rounding) of the data included in the regression model. For a user of such a regression model, it is essential to understand whether the set is, in some sense, "large" or "small"; that is, whether the impact of the loss on the OLS esimator may be serious or not. More generally, the user needs a suitable description of the set $\mathit{OLS}(\boldsymbol{X}, \boldsymbol{y})$. When $p = 2$ or $p = 3$, then the set can be visualized in the parameter space using standard numerical methods. However, in higher dimensions visualization is quite complicated. Hence we need methods for a suitable description of the set $\mathit{OLS}(\boldsymbol{X}, \boldsymbol{y})$; in particular, we would like to design computationally feasible methods. In Section 2 we shall show that this task is very hard from the computational point of view.

D. Two interpretations of the possibilistic approach

Possibilistic interpretation. If we do not assume any distribution on $\boldsymbol{X}$ or $\boldsymbol{y}$, then the set $\mathit{OLS}(\boldsymbol{X}, \boldsymbol{y})$ contains all possible values of $\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$ as X ranges over $\boldsymbol{X}$ and y ranges over $\boldsymbol{y}$. Then, the boundary of the set $\mathit{OLS}(\boldsymbol{X}, \boldsymbol{y})$ can be understood as the worst-case impact of interval censoring (or rounding) on the estimator. The possibilistic approach then can be characterized as a tool for analysis of the worst possible case. The worst-case analysis will be illustrated by an example in Section V-C.

Probabilistic interpretation. If X and y are random variables such that the supports of the distributions of X and y are $\boldsymbol{X}$ and $\boldsymbol{y}$, respectively, then the support of the distribution of $(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$ is $\mathit{OLS}(\boldsymbol{X}, \boldsymbol{y})$. Then the set $B(\boldsymbol{X}, \boldsymbol{y})$ can be called as *100% confidence region* for the OLS estimator. An interesting special case is a regression model with independent random errors with distributions the supports of which are bounded.

E. Variants of interval regression models

An interval regression model $(\boldsymbol{X} = [\underline{X}, \overline{X}], \boldsymbol{y} = [\underline{y}, \overline{y}])$ is also called *a general model* or *interval input – interval output model*. Interesting special cases are (see [23]):

(i) *crisp input – interval output model* is a model with $\underline{X} = \overline{X}$;
(ii) *interval input – crisp output model* is a model with $\underline{y} = \overline{y}$;
(iii) *crisp input – crisp output model* is a model with $\underline{X} = \overline{X}$ and $\underline{y} = \overline{y}$.

"Crisp input – crisp output" is just another name for the traditional model (1).

If $\boldsymbol{X}$ is crisp, i.e. if $\underline{X} = \overline{X} =: X$, then instead of $\mathit{OLS}(\boldsymbol{X}, \boldsymbol{y})$ we write $\mathit{OLS}(X, \boldsymbol{y})$. (And similarly in the case of $\boldsymbol{y}$ crisp.)

III. The general model

Our aim is to find a description of the set $\mathit{OLS}(\boldsymbol{X}, \boldsymbol{y})$ given $\boldsymbol{X} = [\underline{X}, \overline{X}]$ and $\boldsymbol{y} = [\underline{y}, \overline{y}]$. Such a description may take several forms — for example, we might try to find a small enclosing ellipse or a small enclosing box (i.e. interval vector). Theorem 5, which will be the main result of this Section, shows that in general we cannot expect to be successful in a computationally feasible way. The point is that any reasonable description of $\mathit{OLS}(\boldsymbol{X}, \boldsymbol{y})$ must allow the user to decide whether the set is bounded or not. Theorem 5 says that there is no polynomial-time method for this question unless $\boldsymbol{P} = \boldsymbol{NP}$.

Before we state and prove the Theorem, we briefly review some definitions from complexity theory.

A. Some complexity-theoretic notions

We sketch basic definitions needed for further reading only; more details can be found in [2], [39].

The class ***P*** is the class of sets decidable in Turing deterministic polynomial time. The class ***NP*** is the class of sets decidable in Turing nondeterministic polynomial time. The class *co-**NP*** is the class of complements of ***NP***-sets, i.e.

$$\textit{co-}\boldsymbol{NP} = \{A : \textit{co-}A \in \boldsymbol{NP}\},$$

where *co-*A is the complement of A. The class ***PF*** is the class of functions computable in Turing deterministic polynomial time.

A set A is also called *problem* A.

A problem A is *reducible* to problem B if there is a function $f \in \boldsymbol{PF}$ such that

$$(\forall x)[x \in A \iff f(x) \in B].$$

The function f is also called *reduction* of the problem A to the problem B.

A problem C is ***NP***-complete if $C \in \boldsymbol{NP}$ and any problem $A \in \boldsymbol{NP}$ is reducible to C. A problem C is *co-**NP***-complete if $C \in$ *co-**NP*** and any problem $A \in$ *co-**NP*** is reducible to C.

Recall that the most important complexity-theoretic conjecture is that $\boldsymbol{P} \neq \boldsymbol{NP}$ which is generally believed to be true. We shall need the following elementary lemma which can be found in any textbook on complexity theory (see [2], [39]).

Lemma 4. (a) *The problem A is **NP**-complete if and only if the problem co-A is co-**NP**-complete;*
(b) *if A is **NP**-complete, $C \in \boldsymbol{NP}$ and A is reducible to C, then C is **NP**-complete;*
(c) *if $\boldsymbol{P} \neq \boldsymbol{NP}$, then for any co-**NP** complete set C it holds $C \notin \boldsymbol{P}$.* □

The problems in ***P*** are generally considered to be computationally feasible. The proposition (c) says that, if $\boldsymbol{P} \neq \boldsymbol{NP}$, then no *co-**NP***-complete problem is computationally feasible. Indeed, for all *co-**NP***-complete problems we know only exponential time algorithms. The best known example of a *co-**NP*** complete problem is the problem to determine whether a given boolean formula $\varphi(x_1, \ldots, x_N)$ is a tautology. Observe that the simplest method for this problem—construction of the truth table of φ—requires time exponential in N. No feasible algorithm for the problem is known and if $\boldsymbol{P} \neq \boldsymbol{NP}$ then none exist.

B. The main result of Section III

Let $(X)_{ij}$ and $(y)_i$ denote the (i,j)-th component of the matrix X and i-th component of the vector y, respectively.

Theorem 5. *Let $\underline{X}, \overline{X}, \underline{y}, \overline{y}$ be rational and denote $\boldsymbol{X} = [\underline{X}, \overline{X}]$ and $\boldsymbol{y} = [\underline{y}, \overline{y}]$. Deciding whether the set OLS$(\boldsymbol{X}, \boldsymbol{y})$ is bounded is a co-**NP** complete problem.*

Proof. Let $\boldsymbol{X}$ be an $n \times p$ interval matrix. If there is $X \in \boldsymbol{X}$ with column rank $< p$, then for any y the set

$$\{\beta : X^{\mathrm{T}} X \beta = X^{\mathrm{T}} y\}$$

is an affine space of dimension at least one, and hence is unbounded.

Assume that for every $X \in \boldsymbol{X}$, the column rank of X is p. Then $(X^{\mathrm{T}}X)^{-1}$ exists for each $X \in \boldsymbol{X}$. By Cramer's Rule, we can write

$$((X^{\mathrm{T}}X)^{-1})_{ij} = \pm \frac{\det(X^{\mathrm{T}}X)^{[i,j]}}{\det X^{\mathrm{T}}X}$$

where $A^{[i,j]}$ results from A by deleting the j-th row and the i-th column. By continuity of $\det(\cdot)$ on the compact set $\boldsymbol{X}$, the set

$$\{\det X^{\mathrm{T}}X : X \in \boldsymbol{X}\}$$

is a closed interval which, by assumption, does not contain zero. It follows that the set

$$\left\{ \frac{1}{\det X^{\mathrm{T}}X} : X \in \boldsymbol{X} \right\}$$

is a closed interval. Let us denote the interval $[\underline{d}, \overline{d}]$. Also the set

$$\{\pm \det(X^{\mathrm{T}}X)^{[i,j]} : X \in \boldsymbol{X}\}$$

is an interval of the form $[\underline{\delta}_{ij}, \overline{\delta}_{ij}]$. Hence we can write

$$\begin{aligned}(\widehat{\beta})_i &= \{((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y)_i : \ X \in \boldsymbol{X}, \ y \in \boldsymbol{y}\} \\ &= \left\{ \sum_{j=1}^{p} ((X^{\mathrm{T}}X)^{-1})_{ij} \cdot \sum_{k=1}^{n} (X)_{kj} \cdot y_k : X \in \boldsymbol{X}, y \in \boldsymbol{y} \right\} \\ &\subseteq \sum_{j=1}^{p} [\underline{d}, \overline{d}] \cdot [\underline{\delta}_{ij}, \overline{\delta}_{ij}] \cdot \sum_{k=1}^{n} [(\underline{X})_{kj}, (\overline{X})_{kj}] \cdot [(\underline{y})_k, (\overline{y})_k]\end{aligned}$$

and the last expression is a finite sequence of sums and products of intervals. By Lemma 2 it follows that it is a bounded set.

We have shown that the set $B(\boldsymbol{X}, \boldsymbol{y})$ is unbounded if and only if there is an $X \in \boldsymbol{X}$ such that the column rank of X is $< p$. By [40], the latter problem is ***NP*** complete. We have constructed a reduction from an ***NP***-complete problem to the problem $C :=$ *"is OLS$(\boldsymbol{X}, \boldsymbol{y})$ unbounded?"*. By the statements (a) and (b) of Lemma 4, the problem *co-*$C =$ *"is OLS$(\boldsymbol{X}, \boldsymbol{y})$ bounded?"* is *co-**NP***-complete. □

It follows that if we want to find a computationally feasible description of *OLS*$(\boldsymbol{X}, \boldsymbol{y})$ we must reformulate the problem. We can follow (at least) two ways:

(a) either to search for descriptions and/or approximations of *OLS*$(\boldsymbol{X}, \boldsymbol{y})$ model which are guaranteed to be correct only under additional assumptions,
(b) or to consider special cases of the general model separately.

There is a variety of approaches to (a), see [1], [25], [16], [17], [18], [36], [40] and a comparison study [37].

In the next section we follow the way (b) and study the restriction to the crisp input – interval output model. Observe that this restriction is the only interesting restriction among (i) – (iii) (see Section II-E). In the crisp input – crisp output model, the set *OLS*(X, y) is trivial — it is either a single point or an affine space in the parameter space. And the restriction

to the interval input – crisp output model is ruled out by the following observation.

Corollary 6 (to the proof of Theorem 5)**.** *Let $\underline{X}, \overline{X}$ and y be rational and denote $\boldsymbol{X} = [\underline{X}, \overline{X}]$. Deciding whether the set OLS$(\boldsymbol{X}, y)$ is bounded is a co-**NP** complete problem.*

Proof. The reduction constructed in the proof of Theorem 5 remains valid also if $\boldsymbol{y}$ is crisp. □

IV. The crisp input – interval output model

The aim of this section is twofold:

- we shall show a geometric characterization of the set *OLS*$(X, \boldsymbol{y})$;
- we shall show that though there are natural descriptions of the set *OLS*$(X, \boldsymbol{y})$, they cannot be computed in polynomial time.

Hence, from the computational point of view, the situation is (in some sense) as disappointing as in the general case. However, the reason in quite different.

A. Geometric characterization of the set OLS$(X, \boldsymbol{y})$

First we need to review some notions from geometry of convex polyhedra; for further reading see [47].

Definition 7. *The **Minkowski sum** of a set $A \subseteq \mathbb{R}^k$ and a vector $g \in \mathbb{R}^k$ is the set*

$$A \dotplus g = \{a + \lambda g : \ a \in A, \ \lambda \in [0,1]\}.$$

It is easily seen that for a convex set A, it holds

$$A \dotplus g = \mathit{conv}(A \cup \{a + g : a \in A\}),$$

where *conv* denotes the convex hull.

Definition 8. *The **zonotope** generated by $g_1, \ldots, g_N \in \mathbb{R}^k$ with shift $s \in \mathbb{R}^k$ is the set*

$$\mathcal{Z}(s; g_1, \ldots, g_N) = (\cdots((\{s\} \dotplus g_1) \dotplus g_2) \dotplus \cdots \dotplus g_N).$$

*The vectors $g_1, \ldots, g_N$ are called **generators**.*

Instead of $(\cdots((\{s\} \dotplus g_1) \dotplus g_2) \dotplus \cdots \dotplus g_N)$ we shall write $\{s\} \dotplus g_1 \dotplus g_2 \dotplus \cdots \dotplus g_N$ only.

It is easily seen that a zonotope is a convex polyhedron; see Figure 1.

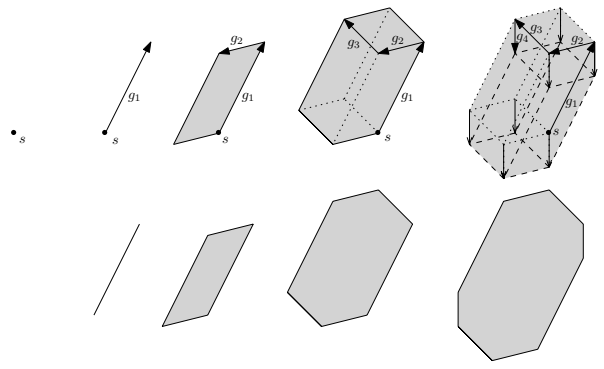

Fig. 1. The evolution of a zonotope $\mathcal{Z}(s; g_1, g_2, g_3, g_4)$.

The main result of this section follows.

Theorem 9. *Let $X \in \mathbb{R}^{n \times p}$ be a matrix of full column rank and $\boldsymbol{y} = [\underline{y}, \overline{y}]$ an $n \times 1$ interval vector. Let $\underline{y}_i$ and $\overline{y}_i$ denote the i-th entry of $\underline{y}$ and $\overline{y}$, respectively. Then*

$$\mathit{OLS}(X, \boldsymbol{y}) = \mathcal{Z}(Q\underline{y};\ Q_1(\overline{y}_1 - \underline{y}_1), \ldots, Q_n(\overline{y}_n - \underline{y}_n)),$$

where $Q = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ and Q_i is the i-th column of Q.

Proof.

$$\begin{aligned}
&\mathit{OLS}(X, \boldsymbol{y}) \\
&= \{Qy : y \in \boldsymbol{y}\} \\
&= \{Q\underline{y} + Q\Lambda : \Lambda \in [0, \overline{y} - \underline{y}]\} \\
&= \{Q\underline{y} + Q\Lambda : \Lambda_1 \in [0, \overline{y}_1 - \underline{y}_1],\ \Lambda_2 \in [0, \overline{y}_2 - \underline{y}_2],\ \ldots, \\
&\qquad \Lambda_n \in [0, \overline{y}_n - \underline{y}_n]\} \\
&= \left\{Q\underline{y} + Q\begin{pmatrix}\Lambda_1 \\ 0 \\ \vdots \\ 0\end{pmatrix} + Q\begin{pmatrix}0 \\ \Lambda_2 \\ \vdots \\ 0\end{pmatrix} + \cdots + Q\begin{pmatrix}0 \\ 0 \\ \vdots \\ \Lambda_n\end{pmatrix} : \right. \\
&\qquad \Lambda_1 \in [0, \overline{y}_1 - \underline{y}_1],\ \Lambda_2 \in [0, \overline{y}_2 - \underline{y}_2], \ldots, \\
&\qquad \left. \Lambda_n \in [0, \overline{y}_n - \underline{y}_n]\right\} \\
&= \{Q\underline{y} + Q_1\Lambda_1 + Q_2\Lambda_2 + \cdots + Q_n\Lambda_n : \\
&\qquad \Lambda_1 \in [0, \overline{y}_1 - \underline{y}_1],\ \Lambda_2 \in [0, \overline{y}_2 - \underline{y}_2],\ \ldots, \\
&\qquad \Lambda_n \in [0, \overline{y}_n - \underline{y}_n]\} \\
&= \{Q\underline{y} + Q_1(\overline{y}_1 - \underline{y}_1)\lambda_1 + Q_2(\overline{y}_2 - \underline{y}_2)\lambda_2 + \cdots \\
&\qquad + Q_n(\overline{y}_n - \underline{y}_n)\lambda_n : \\
&\qquad \lambda_1 \in [0,1],\ \lambda_2 \in [0,1],\ \ldots,\ \lambda_n \in [0,1]\} \\
&= \{Q\underline{y}\} \dotplus Q_1(\overline{y}_1 - \underline{y}_1) \dotplus Q_2(\overline{y}_2 - \underline{y}_2) \dotplus \cdots \\
&\qquad \dotplus Q_n(\overline{y}_n - \underline{y}_n). \quad \square
\end{aligned}$$

There is a nice geometric characterization of zonotopes. Namely, a set $Z \subseteq \mathbb{R}^k$ is a zonotope if and only if *there exists a number m, a matrix $Q \in \mathbb{R}^{k \times m}$ and an interval m-dimensional vector $\boldsymbol{y}$ (called m-dimensional cube) such that $Z = \{Qy : y \in \boldsymbol{y}\}$*. The interesting case is $m > k$. In that case we can say that zonotopes are images of "high-dimensional" cubes in "low-dimensional" spaces under linear mappings, see Figure 2. In our setting, the set *OLS*$(X, \boldsymbol{y})$ is an image of $\boldsymbol{y}$ under the mapping determined by the matrix $Q = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$.

B. Descriptions of the set OLS$(X, \boldsymbol{y})$

In order the user can understand how the set *OLS*$(X, \boldsymbol{y})$ looks like, she/he can use any standard description applicable for convex polyhedra. In particular, three descriptions come to mind:

(a) description of the zonotope *OLS*$(X, \boldsymbol{y})$ by the shift vector and the set of generators;
(b) description of the zonotope *OLS*$(X, \boldsymbol{y})$ by the enumeration of vertices;
(c) description of the zonotope *OLS*$(X, \boldsymbol{y})$ by the enumeration of facets, i.e. in terms of a p-column matrix A and a vector c such that *OLS*$(X, \boldsymbol{y}) = \{\beta \in \mathbb{R}^p : A\beta \leq c\}$.

The description (a) has been given by the Theorem 9.

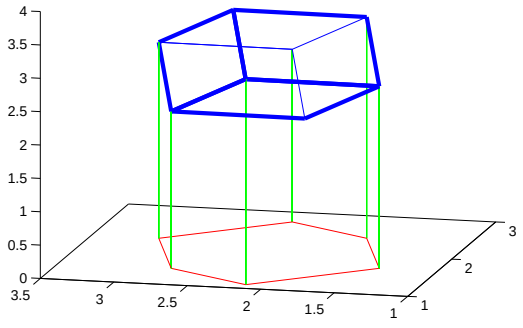

Fig. 2. A zonotope as an image of a higher-dimensional cube.

C. A negative complexity result for the descriptions (b) and (c)

It is an interesting question whether there are efficient algorithms which can construct the enumerations (b) and (c) given X, $\underline{y}$ and $\overline{y}$. We give an argument that the answer is negative. The answer follows from the simple fact that zonotopes can have too many vertices and facets.

Theorem 10 ([47]). *For a zonotope $Z \subseteq \mathbb{R}^p$ with n generators it holds $V(Z) \leq 2\sum_{k=0}^{p-1}\binom{n-1}{k}$ and $F(Z) \leq 2\binom{n}{p-1}$, where $V(Z)$ is the number of vertices and $F(Z)$ is the number of facets of Z. In general the bounds cannot be improved.* □

The numbers $V(Z)$ and $F(Z)$ cannot be bounded by a polynomial in n and p; hence, the functions enumerating vertices and facets are not in ***PF*** for the simple reason that their output cannot be bounded by a polynomial in the size of the input.

D. A positive complexity result for the descriptions (b) and (c)

However, Theorem 10 has an interesting corollary if we treat the number p as a fixed constant (i.e. if we restrict ourselves to a class of regression models with a fixed number of regression parameters).

Corollary 11. *If p is fixed then $V(Z) \leq O(n^{p-1})$ and $F(Z) \leq O(n^{p-1})$.*

Proof. We have

$$\begin{aligned} F(Z) &\leq 2\binom{n}{p-1} \\ &= \frac{2n(n-1)\cdots(n-p+2)}{(p-1)!} \\ &\leq 2n^{p-1} \\ &\leq O(n^{p-1}) \end{aligned} \tag{4}$$

and

$$\begin{aligned} V(Z) &\leq 2\sum_{k=0}^{p-1}\binom{n-1}{k} \\ &\leq 2p \cdot \max_{k\in\{0,\ldots,p-1\}}\binom{n-1}{k} \\ &\overset{(\star)}{\leq} O(n^{k_{\max}}) \\ &= O(n^{p-1}), \end{aligned}$$

where $k_{\max}$ is the $k \in \{0,\ldots,p-1\}$ for which the maximum is attained. By well-known properties of binomial coefficients, for n large enough it holds $k_{\max} = p-1$. In the inequality $(\star)$ we used a similar estimate as in (4). □

In the literature on computational geometry, several algorithms for enumeration of vertices and facets of a zonotope given by the set of generators are known. Moreover, there are methods with computation time which is bounded by a polynomial in the size of input and size of output. In Corollary 11 we have shown that if p is fixed then the size of the output is polynomially bounded in the size of the input. Hence, if p is fixed then these methods work in polynomial time.

We shall not describe the methods here; we recommend the papers [3] and [11].

V. Approximations of the set $OLS(X, \boldsymbol{y})$

A. Interval approximation

By basic properties of interval arithmetic (3), it is easily seen that for every i and every $b \in OLS(X, \boldsymbol{y})$ it holds

$$\overbrace{\sum_{j=1}^{n}\min\{(Q)_{ij}(\underline{y})_j, (Q)_{ij}(\overline{y})_j\}}^{=:(\underline{b})_i} \leq (b)_i \leq \underbrace{\sum_{j=1}^{n}\max\{(Q)_{ij}(\underline{y})_j, (Q)_{ij}(\overline{y})_j\}}_{=:(\overline{b})_i} \tag{5}$$

where $Q = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. Moreover, the cube

$$B = [\underline{b}, \overline{b}] \tag{6}$$

is the smallest cube overscribing the set $OLS(X, \boldsymbol{y})$.

The bound B can be easily computed in polynomial time. Moreover, it allows us to quantify the effect of interval censoring on each regression parameter separately. Often it is the case that we are interested in estimation of a single regression parameter or a subset of regression parameters; then, if the interval $[(\underline{b})_i, (\overline{b})_i]$ is narrow, this fact can be interpreted as *the interval-censoring effect is insignificant for estimation of the i-th parameter*.

B. Ellipsoidal approximation

The smallest ellipse $\mathcal{E}$ containing $OLS(X, \boldsymbol{y})$ is called *the Löwner-John ellipse*. Combinatorially complex polyhedra are often approximated with ellipses: an ellipse is a convex set which is quite flexible to approximate the shape of the polyhedron and it is sufficiently simple to be described. An ellipse $\mathcal{E}$ is described by a center point s and a positive definite matrix E such that

$$\mathcal{E} = \{x \in \mathbb{R}^p : (x-s)^{\mathrm{T}} E^{-1} (x-s) \leq 1\}.$$

We do not know a polynomial-time algorithm for construction of the Löwner-John ellipse for the set $OLS(X, \boldsymbol{y})$. It is an intriguing research problem; however, we expect a hardness result on this computational problem rather than a polynomial-time algorithm. (More on algorithms for finding ellipses overscribing polyhedra is found in [13].)

The following ellipse $\mathcal{E} = (E, s)$ can be seen as a weaker form:

$$\begin{aligned} s &= \tfrac{1}{2} Q(\overline{y} + \underline{y}), \\ E &= Q \cdot diag\left(\tfrac{n}{4}((\overline{y})_1 - (\underline{y})_1)^2, \ldots, \tfrac{n}{4}((\overline{y})_n - (\underline{y})_n)^2\right) \cdot Q^{\mathrm{T}}, \end{aligned} \tag{7}$$

where $Q = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ and $diag(\xi_1, \ldots, \xi_n)$ denotes the diagonal matrix with diagonal entries $\xi_1, \ldots, \xi_n$. This is the ellipse which is the image of the smallest ellipse overscribing $\boldsymbol{y}$ in $\mathbb{R}^n$ under the mapping $v \mapsto Qv$. This proves $Z \subseteq \mathcal{E}$.

C. Example

Consider the regression model

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \tag{8}$$

with $n = 11$ observations collected in the following table. Only interval-censored values are available to us:

$$(\boldsymbol{y})_i = [(\underline{y})_i, (\overline{y})_i] = [(\tilde{y})_i - \tfrac{1}{2}, (\tilde{y})_i + \tfrac{1}{2}], \quad i = 1, \ldots, 11$$

where $\tilde{y}$ denotes the center of $\boldsymbol{y}$.

i	1	2	3	4	5	6
x_i	−2	−1	0	1	2	3
$\underline{y}_i$	1.5	−1.5	−0.5	3.5	3.5	5.5
$\tilde{y}_i$	2	−1	0	4	4	6
$\overline{y}_i$	2.5	−0.5	0.5	4.5	4.5	6.5
i	7	8	9	10	11	
x_i	4	5	6	7	8	
$\underline{y}_i$	8.5	6.5	10.5	10.5	9.5	
$\tilde{y}_i$	9	7	11	11	10	
$\overline{y}_i$	9.5	7.5	11.5	11.5	10.5	

Using the central estimator $\tilde{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\tilde{y}$ we get

$$\tilde{\beta}_1 = 2.12, \quad \tilde{\beta}_2 = 1.2$$

and with (5) we get

$$[(\underline{b})_1, (\overline{b})_1] = [1.56, 2.69], \quad [(\underline{b})_2, (\overline{b})_2] = [1.06, 1.34].$$

We can conclude that the interval-censoring effect couldn't have caused an error higher than ± 0.565 $[= \frac{1}{2}(2.69 - 1.56)]$ in the estimate of β_1 and an error higher than ± 0.14 in the estimate of β_2.

The set (zonotope) $OLS(X, \boldsymbol{y})$, together with the enclosure B given by (6) and the ellipse (7), is plotted in Figure 3.

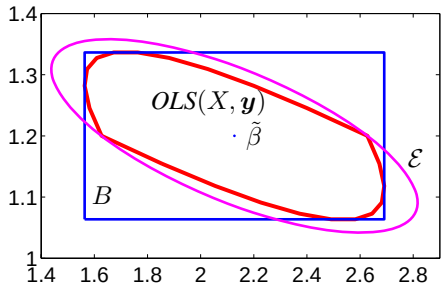

Fig. 3. The set (zonotope) $OLS(X, \boldsymbol{y})$ for the regression model in the Example and its approximations B and $\mathcal{E}$ given by (6) and (7), respectively.

Though the approximations 1 and 2 are quite trivial, their combination gives some nontrivial information. The enclosure B contains the point $[1.56, 1.06]$; hence, the approximation B does not rule out the case that *both regression parameters could have been affected by the maximal possible error* $[-0.565, -0.14]$ *in the negative direction simultaneously.* However, this case is ruled out by the fact that $[1.65, 1.06] \notin \mathcal{E}$.

D. Testing admissibility

As motivated by the previous Example, it is natural to ask whether it could have happened that all regression parameters had been affected by a simultaneous error Δ; i.e. whether the point $\tilde{\beta} + \Delta$ is in $OLS(X, \boldsymbol{y})$ or not. A vector b (in particular, a vector b of the form $b = \tilde{\beta} + \Delta$) is called *admissible* if $b \in OLS(X, \boldsymbol{y})$.

Proposition 12. *Admissibility can be tested in polynomial time.*

Proof. The vector b is admissible if and only if there is a $y \in \mathbb{R}^n$ such that

$$Qy = b \quad \text{and} \quad \underline{y} \leq y \leq \overline{y}$$

where $Q = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. Hence, deciding admissibility amounts to deciding feasibility of a system of linear (in)equalities, which is essentially a linear programming problem. Linear programming is solvable in polynomial time, see [41]. □

E. Monte Carlo estimation of volume of the set $OLS(X, \boldsymbol{y})$

Proposition 12, combined with (5), suggests a simple procedure for Monte-Carlo approximation of the volume of the set $OLS(X, \boldsymbol{y})$, which is a natural measure of its size. The procedure just generates a random point $b \in [\underline{b}, \overline{b}]$ and tests its admissibility. This procedure is interesting in particular in higher dimensions, where the zonotope $OLS(X, \boldsymbol{y})$ cannot be easily visualized.

Using the Monte Carlo approximation of volume is a reasonable choice: no polynomial-time algorithm (in n, p) for exact computation of volume of the set $OLS(X, \boldsymbol{y})$ is known.

ACKNOWLEDGEMENT

The work was supported by Project No. F4/18/2011 of Internal Grant Agency of University of Economics, Prague, Czech Republic and by Project No. MSM6138439910 of Ministry of Education, Youth and Sports of the Czech Republic.

Thanks to Professor Jaromír Antoch and Miroslav Rada.

REFERENCES

[1] G. Alefeld and J. Herzberger, *Introduction to interval computations,* Computer Science and Applied Mathematics, New York, USA: Academic Press, 1983.

[2] S. Arora and B. Barak, *Computational complexity: A modern approach,* Cambridge, Great Britain: Cambridge University Press, 2009.

[3] D. Avis and K. Fukuda, *Reverse search for enumeration,* Discrete Applied Mathematics 65, 1996, 21–46.

[4] A. H. Bentbib, *Solving the full rank interval least squares problem,* Applied Numerical Mathematics 41 (2), 2002, 283–294.

[5] M. Černý and M. Hladík, *The regression tolerance quotient in data analysis,* in: M. Houda and J. Friebelová (eds.), *Procceding of Mathematical Methods in Economics 2010,* Czech Republic: University of South Bohemia, 2010, 98–104.

[6] M. Černý and M. Rada, *A note on linear regression with interval data and linear programming,* in: *Quantitative methods in economics: Multiple Criteria Decision Making XV,* Slovakia: Kluwer, Iura Edition, 2010, 276–282.

[7] P.-T. Chang, E. S. Lee and S. A. Konz, *Applying fuzzy linear regression to VDT legibility,* Fuzzy Sets and Systems 80 (2), 1996, 197–204.

[8] C. Chuang, *Extended support vector interval regression networks for interval input-output data,* Information Science 178 (3), 2008, 871–891.

[9] J. P. Dunyak and D. Wunsch, *Fuzzy regression by fuzzy number neural networks,* Fuzzy Sets and Systems 112 (3), 2000, 371–380.

[10] T. Entani and M. Inuiguchi, *Group decisions in interval AHP based on interval regression analysis,* in: V.-N. Huynh et al. (eds.), *Integrated uncertainty management and applications,* Advances in Soft Computing, vol. 68, Germany: Springer, 2010, 269–280.

[11] J.-A. Ferrez, K. Fukuda and T. Liebling, *Solving the fixed rank convex quadratic maximization in binary variables by a parallel zonotope construction algorithm,* European Journal of Operational Research 166, 2005, 35–50.

[12] D. M. Gay, *Interval least squares—a diagnostic tool,* in R. E. Moore (ed.), *Reliability in computing, the role of interval methods in scientific computing,* Perspectives in Computing, vol. 19, Boston, USA: Academic Press, 1988, 183–205.

[13] M. Grötschel, L. Lovász and A. Schrijver, *Geometric algorithms and combinatorial optimization,* Germany: Springer, 1993.

[14] P. Guo and H. Tanaka, *Dual models for possibilistic regression analysis,* Computational Statistics & Data Analysis 51 (1), 2006, 253–266.

[15] B. Hesmaty and A. Kandel, *Fuzzy linear regression and its applications to forecasting in uncertain environment,* Fuzzy Sets and Systems 15, 1985, 159–191.

[16] M. Hladík, *Description of symmetric and skew-symmetric solution set,* SIAM Journal on Matrix Analysis and Applications 30 (2), 2008, 509–521.

[17] M. Hladík, *Solution set characterization of linear interval systems with a specific dependence structure,* Reliable Computing 13 (4), 2007, 361–374.

[18] M. Hladík, *Solution sets of complex linear interval systems of equations,* Reliable Computing 14, 2010, 78–87.

[19] M. Hladík and M. Černý, *Interval regression by tolerance analysis approach,* Submitted in Fuzzy Sets and Systems, Preprint: KAM-DIMATIA Series 963, 2010.

[20] M. Hladík and M. Černý, *New approach to interval linear regression,* in: R. Kasımbeyli et al. (eds.), *24th Mini-EURO conference on continuous optimization and information-based technologies in the financial sector MEC EurOPT 2010, Selected papers,* Vilnius, Lithuania: Technika, 2010, 167–171.

[21] C.-H. Huang and H.-Y. Kao, *Interval regression analysis with soft-margin reduced support vector machine,* Lecture Notes in Computer Science 5579, Germany: Springer, 2009, 826–835.

[22] M. Inuiguchi, H. Fujita and T. Tanino, *Robust interval regression analysis based on Minkowski difference,* in: *SICE 2002, proceedings of the 41st SICE Annual Conference,* vol. 4, Osaka, Japan, 2002, 2346–2351.

[23] H. Ishibuchi and H. Tanaka, *Several formulations of interval regression analysis,* in: *Proceedings of Sino-Japan joint meeting on fuzzy sets and systems,* Beijing, China, 1990, B2-2, 1–4.

[24] H. Ishibuchi, H. Tanaka and H. Okada, *An architecture of neural networks with interval weights and its application to fuzzy regression analysis,* Fuzzy Sets and Systems 57 (1), 1993, 27–39.

[25] C. Jansson, *Calculation of exact bounds for the solution set of linear interval systems,* Linear Algebra and its Applications 251, 1997, 321–340.

[26] G. Jun-peng and L. Wen-hua, *Regression analysis of interval data based on error theory,* in: *Proceedings of 2008 IEEE International Conference on Networking, Sensing and Control, ICNSC,* Sanya, China, 2008, 552–555.

[27] M. Kaneyoshi, H. Tanaka, M. Kamei and H. Furuta, *New system identification technique using fuzzy regression analysis,* in: *Proceedings of the First International Symposium on Uncertainty Modeling and Analysis,* Baltimore, USA, 1990, 528–533.

[28] H. Kashima, K. Yamasaki, A. Inokuchi and H. Saigo, *Regression with interval output values,* in: *19th International Conference on Pattern Recognition ICPR 2008,* Tampa, USA, 2008, 1–4.

[29] H. Lee and H. Tanaka, *Fuzzy regression analysis by quadratic programming reflecting central tendency,* Behaviormetrika 25 (1), 1998, 65–80.

[30] H. Lee and H. Tanaka, *Upper and lower approximation models in interval regression using regression quantile techniques,* Europeran Journal of Operational Research 116 (3), 1999, 653–666.

[31] B. Li, C. Li, J. Si and G. Abousleman, *Interval least-squares filtering with applications to robust video target tracking,* in: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing — Proceedings,* Las Vegas, USA: IEEE Signal Processing Society, 2008, 3397–3400.

[32] E. de A. Lima Neto, F. de A. T. de Carvalho, *Constrained linear regression models for symbolic interval-valued variables,* Computational Statistics & Data Analysis 54 (2), 2010, 333–347.

[33] P. Liu, *Study on a speech learning approach based on interval support vector regression,* in: *Proceedings of 4th International Conference on Computer Science & Education,* Nanning, China, 2009, 1009–1012.

[34] I. Moral-Arce, J. M. Rodríguez-Póo and S. Sperlich, *Low dimensional semiparametric estimation in a censored regression model,* Journal of Multivariate Analysis 102 (1), 118–129.

[35] E. Nasrabadi and S. Hashemi, *Robust fuzzy regression analysis using neural networks,* International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 16 (4), 2008, 579–598.

[36] A. Neumaier, *Interval methods for systems of equations,* Cambridge, Great Britain: Cambridge University Press, 1990.

[37] S. Ning and R. B. Kearfott, *A comparison of some methods for solving linear interval equations,* SIAM Journal on Numerical Analysis 34 (4), 1997, 1289–1305.

[38] W. Pan and R. Chappell, *Computation of the NPMLE of distribution functions for interval censored and truncated data with applications to the Cox model,* Computational Statistics & Data Analysis 28 (1), 1998, 33–50.

[39] C. Papadimitriou, *Computational complexity,* Addison-Wesley Longman, 1995.

[40] J. Rohn, *A handbook of results on interval linear problems,* Prague, Czech Republic: Czech Academy of Sciences, 2005; available at: http://uivtx.cs.cas.cz/∼rohn/handbook/handbook.zip.

[41] A. Schrijver, *Theory of linear and integer programming,* USA: Wiley, 2000.

[42] K. Sugihara, H. Ishii and H. Tanaka, *Interval priorities in AHP by interval regression analysis,* Europeran Journal of Operational Research 158 (3), 2004, 745–754.

[43] H. Tanaka and H. Lee, *Fuzzy linear regression combining central tendency and possibilistic properties,* in: *Proceedings of the Sixth IEEE International Conference on Fuzzy Systems,* vol. 1, Barcelona, Spain, 1997, 63–68.

[44] H. Tanaka and H. Lee, *Interval regression analysis by quadratic programming approach,* IEEE Transactions on Fuzzy Systems 6 (4), 1998, 473–481.

[45] H. Tanaka and J. Watada, *Possibilistic linear systems and their application to the linear regression model,* Fuzzy Sets and Systems 27 (3), 1988, 275–289.

[46] X. Zhang and J. Sun, *Regression analysis of clustered interval-censored failure time data with informative cluster size,* Computational Statistics & Data Analysis 54 (7), 2010, 1817–1823.

[47] G. Ziegler, *Lectures on polytopes,* Germany: Springer, 2004.

Michal Černý (born 1979) is a researcher and lecturer affiliated at the Department of Econometrics, University of Economics, Prague, Czech Republic and the Department of Theoretical Computer Science, Czech Technical University, Prague, Czech Republic. He works in data analysis, optimization and computational complexity. He has published several results on algorithmic problems in statistics, e.g. on experimental design and analysis of interval data.