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

\[ y = X\beta + \varepsilon \]

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

\[ \hat{\beta} = Qy, \]

where \( Q \) is a matrix. In the following text, we shall concentrate on the Ordinary Least Squares (OLS) estimator, which corresponds to the choice \( Q = (X^T X)^{-1} X^T \) 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 \( Q = (X^T \Omega^{-1} X)^{-1} \Omega^{-1} X^T \) in (2), where \( \Omega \) is either known or estimated covariance matrix of \( \varepsilon \). 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.

\[ \tilde{y}^5, \tilde{y}^2, \tilde{y}^1, \text{ and } \tilde{y}^0 \]

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], speech 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 $-\infty < a \leq \pi < \infty$, the interval number $a$ is the closed interval $[a, \pi]$.
(b) Let $X \leq X$ be two $M \times N$ real matrices. The interval matrix $X = [X, \bar{X}]$ is the set
\[
\{X \in \mathbb{R}^{M \times N} : X \leq X \leq \bar{X}\}.
\]
The interval vector $y = [y, \bar{y}]$ is a special case of the interval matrix with one column.

Interval numbers, vectors and matrices are typeset in bold-face.

Arithmetic operations $+$ and $\times$ with interval numbers $a = [a, \pi]$ and $b = [b, \bar{b}]$ are defined in a natural way (see [1]):
\[
\begin{align*}
\alpha + \beta &= [\alpha + \beta, \pi + \bar{b}], \\
\alpha \cdot \beta &= [\min\{\alpha \cdot b, \pi \cdot \pi\}, \max\{\alpha \cdot \bar{b}, \pi \cdot \bar{b}\}].
\end{align*}
\]
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 $(X, y)$ are available instead of the exact values of the data $(X, y)$ such that $X \in X$ and $y \in y$. 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 $\hat{\beta}$. The next definition generalizes the notion of the estimator $\hat{\beta}$ for the case when the crisp values $(X, y)$ in are replaced by intervals $(X, y)$ in (1).

Definition 3. (a) A tuple $(X, y)$, where $X$ is an $n \times p$ interval matrix and $y$ 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 $(X, y)$ is defined as
\[
\text{OLS}(X, y) = \{\beta \in \mathbb{R}^{p} : \exists X \in X, y \in y, X^T X \beta = X^T y\}.
\]
The motivation for the definition is straightforward. Our aim is to use OLS to obtain an estimate of the unknown vector of regression parameters $\beta$ in the model (1). However, observations of both dependent variables $(y)$ and independent variables $(X)$ are interval-censored; i.e., we only know intervals $X$ and $y$ that are guaranteed to contain the directly unobservable data $(X, y)$. Then, the set $\text{OLS}(X, y)$ contains all possible values of OLS-estimates of $\beta$ as $X$ and $y$ range over $X$ and $y$, respectively. We say that $\text{OLS}(X, y)$ is a possibilistic version of the notion of the OLS estimator.

The set $\text{OLS}(X, y)$ 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 estimator may be serious or not. More generally, the user needs a suitable description of the set $\text{OLS}(X, y)$. 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 $\text{OLS}(X, y)$; 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

1. Possibilistic interpretation. If we do not assume any distribution on $X$ or $y$, then the set $\text{OLS}(X, y)$ contains all possible values of $\hat{\beta} = (X^T X)^{-1} X^T y$ as $X$ ranges over $X$ and $y$ ranges over $y$. Then, the boundary of the set $\text{OLS}(X, y)$ 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.

2. Probabilistic interpretation. If $X$ and $y$ are random variables such that the supports of the distributions of $X$ and $y$ are $X$ and $y$, respectively, then the support of the distribution of $(X^T X)^{-1} X^T y$ is $\text{OLS}(X, y)$. Then the set $\text{OLS}(X, y)$ 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 $(X = [X, \bar{X}], y = [y, \bar{y}])$ 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 $X = \bar{X}$;
(ii) interval input – crisp output model is a model with $y = \bar{y}$;
(iii) crisp input – crisp output model is a model with $X = \bar{X}$ and $y = \bar{y}$.

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

If $X$ is crisp, i.e. if $X = \bar{X} = X$, instead of $\text{OLS}(X, y)$ we write $\text{OLS}(X, y)$. (And similarly in the case of $y$ crisp.)
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 \( \text{NP} \) is the class of sets decidable in Turing nondeterministic polynomial time. The class \( \text{co-NP} \) is the class of complements of \( \text{NP} \)-sets, i.e.

\[ \text{co-NP} = \{ A : \text{co-A \in \text{NP}} \}, \]

where \( \text{co-A} \) is the complement of \( A \). The class \( \text{PF} \) 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 \( f \in \text{PF} \) such that

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

The function \( f \) is also called \( \text{reduction} \) of the problem \( A \) to the problem \( B \).

A problem \( C \) is \( \text{NP} \)-complete if \( C \in \text{NP} \) and any problem \( A \in \text{NP} \) is reducible to \( C \). A problem \( C \) is \( \text{co-NP} \)-complete if \( C \in \text{co-NP} \) and any problem \( A \in \text{co-NP} \) is reducible to \( C \).

Recall that the most important complexity-theoretic conjecture is that \( P \neq \text{NP} \) 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 \( \text{NP} \)-complete if and only if the problem \( \text{co-A} \) is \( \text{NP} \)-complete;

(b) if \( A \) is \( \text{NP} \)-complete, \( C \in \text{NP} \) and \( A \) is reducible to \( C \), then \( C \) is \( \text{NP} \)-complete;

(c) if \( P \neq \text{NP} \), then for any \( \text{NP} \)-complete set \( C \) it holds \( C \notin P \).

The problems in \( P \) are generally considered to be computationally feasible. The proposition (c) says that, if \( P \neq \text{NP} \), then no \( \text{NP} \)-complete problem is computationally feasible. Indeed, for all \( \text{NP} \)-complete problems we know only exponential time algorithms. The best known example of a \( \text{NP} \)-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 \( \varphi \)—requires time exponential in \( N \). No feasible algorithm for the problem is known and if \( P \neq \text{NP} \) 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 \( X, X, y, \overline{y} \) be rational and denote \( X = [X, X] \) and \( y = [y, \overline{y}] \). Deciding whether the set \( \text{OLS}(X, y) \) is bounded is a \( \text{NP} \)-complete problem.

**Proof.** Let \( X \) be an \( n \times p \) interval matrix. If there is \( X \in X \) with column rank \( < p \), then for any \( y \) the set

\[ \{ \beta : X^T X \beta = X^T y \} \]

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

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

\[ ((X^T X)^{-1})_{ij} = \pm \frac{\det(X^T X)_{ij}}{\det X^T X} \]

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

\[ \{ \det X^T X : X \in X \} \]

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

\[ \left\{ \frac{1}{\det X^T X} : X \in X \right\} \]

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

\[ \{ \pm \det(X^T X)(b, \overline{b}) : X \in X \} \]

is an interval of the form \([d_{ij}, \overline{d}_{ij}]\). Hence we can write

\[ (\beta)_i = \{ ((X^T X)^{-1} X^T y)_i : X \in X, y \in y \} \]

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

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

(a) either to search for descriptions and/or approximations of \( \text{OLS}(X, y) \) 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 \( \text{OLS}(X, y) \) 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 $X, Y$ and $Y$ be rational and denote $X = [X, X]$. Deciding whether the set $OLS(X, y)$ is bounded is a co-$NP$ complete problem.

Proof. The reduction constructed in the proof of Theorem 5 remains valid also if $y$ is crisp.

IV. THE CRISP INPUT – INTERVAL OUTPUT MODEL

The aims of this section are twofold:

- we shall show a geometric characterization of the set $OLS(X, y)$;
- we shall show that though there are natural descriptions of the set $OLS(X, y)$, 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 is quite different.

A. Geometric characterization of the set $OLS(X, y)$

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

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

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

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

where $\text{conv}$ 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

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

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

Instead of $\{s + g_1 + \cdots + g_N\}$ we shall write $\{s \in R^k : g_1 + \cdots + g_N\}$ only.

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

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 $y = [y, y]$ an $n \times 1$ interval vector. Let $\underline{y}$ and $\overline{y}$ denote the $i$-th entry of $y$ and $\overline{y}$, respectively. Then

$$OLS(X, y) = \bigcup_{i = 1}^{n} [\underline{Q_i} - y_i, \overline{Q_i} - y_i],$$  

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

Proof.

$$OLS(X, y) = \{Qy : y \in Y\} = \{Qy : A \in [0, y - y]\}$$  

$$= \{Qy + QA : A \in [0, y - y]\} = \{Qy + QA : A_1 \in [0, y_1 - y_1], A_2 \in [0, y_2 - y_2], \ldots, A_n \in [0, y_n - y_n]\}$$  

$$= \{Qy + Q(A_1 y_1 + A_2 y_2 + \cdots + A_n y_n) : A_1 \in [0, y_1 - y_1], A_2 \in [0, y_2 - y_2], \ldots, A_n \in [0, y_n - y_n]\}$$  

$$= \{Qy + Q(A_1 y_1 + A_2 y_2 + \cdots + A_n y_n) : A_1 \in [0, y_1 - y_1], A_2 \in [0, y_2 - y_2], \ldots, A_n \in [0, y_n - y_n]\}$$  

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 $y$ (called $m$-dimensional cube) such that $Z = \{Qy : y \in Y\}$. 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 $OLS(X, y)$ is an image of $y$ under the mapping determined by the matrix $Q = (X^TX)^{-1}X^T$.

B. Descriptions of the set $OLS(X, y)$

In order the user can understand how the set $OLS(X, y)$ 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 $OLS(X, y)$ by the shift vector and the set of generators;

(b) description of the zonotope $OLS(X, y)$ by the enumeration of vertices;

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

The description (a) has been given by the Theorem 9.
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 \( X \), \( y \) and \( \pi \). 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 \( F(Z) \leq 2^p \binom{n}{p-1} \), 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

\[
F(Z) \leq 2 \binom{n}{p-1} = \frac{2n(n-1)\cdots(n-p+2)}{(p-1)!} \leq 2n^{p-1} \leq O(n^{p-1})
\]

and

\[
V(Z) \leq 2 \sum_{k=0}^{p-1} \binom{n-1}{k} \leq 2p \max_{k \in \{0, \ldots, p-1\}} \binom{n-1}{k} \leq O(n^{k_{\text{max}}}) = O(n^{p-1}),
\]

where \( k_{\text{max}} \) 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 \( k_{\text{max}} = p - 1 \). In the inequality (\( \ast \)) 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 \( OLS(X, y) \)

A. Interval approximation

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

\[
\sum_{j=1}^{n} \min \{(Q)_{ij}y_j, (Q)_{ij}(\bar{y})_j\} \leq (b)_i \leq \sum_{j=1}^{n} \max \{(Q)_{ij}y_j, (Q)_{ij}(\bar{y})_j\}
\]

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

\[
B = [b, \bar{b}]
\]

is the smallest cube overscribing the set \( OLS(X, y) \).

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 \([b]_i, [\bar{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 \( E \) containing \( \text{OLS}(X, y) \) 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 \( \tilde{E} \) is described by a center point \( s \) and a positive definite matrix \( E \) such that

\[
\tilde{E} = \{ x \in \mathbb{R}^p : (x-s)^T E^{-1} (x-s) \leq 1 \}. 
\]

We do not know a polynomial-time algorithm for construction of the Löwner-John ellipse for the set \( \text{OLS}(X, y) \). 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 \( \tilde{E} = (E, s) \) can be seen as a weaker form:

\[
s = \frac{1}{2} Q (\bar{y} + y), \\
E = Q \cdot \text{diag} \left( \frac{1}{2}((\bar{y})_{1} - (y)_{1})^2, \ldots, \frac{1}{2}((\bar{y})_{n} - (y)_{n})^2 \right) \cdot Q^T, 
\]

where \( Q = (X^T X)^{-1} X^T \) and \( \text{diag}(\xi_1, \ldots, \xi_n) \) 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 \( y \) in \( \mathbb{R}^n \) under the mapping \( v \mapsto Qv \). This proves \( Z \subseteq \tilde{E} \).

\[
\begin{array}{ccccccc}
  i & 1 & 2 & 3 & 4 & 5 & 6 \\
  x_i & -2 & -1 & 0 & 1 & 2 & 3 \\
  \bar{y}_i & 1.5 & -1.5 & -0.5 & 3.5 & 3.5 & 5.5 \\
  y_i & 2 & -1 & 0 & 4 & 4 & 6 \\
  \bar{y}_i & 2.5 & -0.5 & 0.5 & 4.5 & 4.5 & 6.5 \\
  y_i & 7 & 8 & 9 & 10 & 11 \\
  y_i & 4 & 5 & 6 & 7 & 8 \\
  y_i & 8.5 & 6.5 & 10.5 & 10.5 & 9.5 \\
  \bar{y}_i & 9 & 7 & 11 & 11 & 10 \\
  \bar{y}_i & 9.5 & 7.5 & 11.5 & 11.5 & 10.5 \\
\end{array}
\]

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

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

and with (5) we get

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

We can conclude that the interval-censoring effect couldn’t have caused an error higher than \( \pm 0.565 \) \([= \frac{1}{2}(2.69 - 1.56)]\)

in the estimate of \( \beta_1 \) and an error higher than \( \pm 0.14 \) in the estimate of \( \beta_2 \).

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

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure3.png}
\caption{The set (zonotope) \( \text{OLS}(X, y) \) for the regression model in the Example and its approximations \( B \) and \( \tilde{E} \) given by (6) and (7), respectively.}
\end{figure}

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 \([1.65, 1.06] \notin \tilde{E} \).

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 \( \Delta \); i.e. whether the point \( \tilde{\beta} + \Delta \) is in \( \text{OLS}(X, y) \) or not. A vector \( b \) (in particular, a vector \( b \) of the form \( b = \tilde{\beta} + \Delta \)) is called admissible if \( b \in \text{OLS}(X, y) \).

\begin{proposition}
Admissibility can be tested in polynomial time.
\end{proposition}

\begin{proof}
The vector \( b \) is admissible if and only if there is a \( y \in \mathbb{R}^n \) such that

\[
Q y = b \quad \text{and} \quad \bar{y} \leq y \leq \bar{y}
\]

where \( Q = (X^T X)^{-1} X^T \). 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].
\end{proof}

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

Proposition 12, combined with (5), suggests a simple procedure for Monte-Carlo approximation of the volume of the set \( \text{OLS}(X, y) \), which is a natural measure of its size. The procedure just generates a random point \( b \in [\bar{b}, \bar{b}] \) and tests its admissibility. This procedure is interesting in particular in higher dimensions, where the zonotope \( \text{OLS}(X, y) \) 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 \( \text{OLS}(X, y) \) 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 Anocho and Miroslav Rada.

REFERENCES


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.