10.2478/v10048-011-0007-0

On the Possibilistic Approach to Linear Regression with Rounded or Interval-Censored Data

Michal Černý*, Miroslav Rada

Department of Econometrics, University of Economics, Winston Churchill Square 4, 130 67 Prague, Czech Republic

Consider a linear regression model where some or all of the observations of the dependent variable have been either rounded or interval-censored and only the resulting interval is available. Given a linear estimator $\widehat{\beta}$ of the vector of regression parameters, we consider its possibilistic generalization for the model with rounded/censored data, which is called the OLS-set in the special case $\widehat{\beta}$ = Ordinary Least Squares. We derive a geometric characterization of the set: we show that it is a zonotope in the parameter space. We show that even for models with a small number of regression parameters and a small number of observations, the combinatorial complexity of the polyhedron can be high. We therefore derive simple bounds on the OLS-set. These bounds allow to quantify the worst-case impact of rounding/censoring on the estimator $\widehat{\beta}$. This approach is illustrated by an example. We also observe that the method can be used for quantification of the rounding/censoring effect in advance, before the experiment is made, and hence can provide information on the choice of measurement precision when the experiment is being planned.

Keywords: Linear regression; rounding; inexact data; interval-censored data.

1. 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). Throughout the text we assume X is a fixed matrix of constants.

In this text we deal with the situation when the observations y of the dependent variable cannot be observed directly; instead, only the interval vector $Y = [\underline{Y}, \overline{Y}]$ is known such that the vector of unobservable values y fulfills $y \in Y$.

A typical setup, when only Y instead of the exact values y are available, is the presence of rounding. 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 reserved for representation of the non-integer part. For example, if we store data as integers, then we know only the interval $Y = [\tilde{y} - 0.5, \tilde{y} + 0.5]$ instead of the exact value y, where $\tilde{y}$ is y rounded to the nearest integer.

However, the setting may be understood more generally, for example:

- The data 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.
- 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.

There is an interesting difference between rounded data and interval-censored data.

(a) If the data y have been rounded, then the widths of all intervals $Y_1, \ldots Y_n$ are the same; for example, if we are rounding to integers, then every interval in Y has width 1.

(b) If the intervals Y resulted from censoring, then the intervals $Y_1, \ldots, Y_n$ may be of different widths. In particular, only some portion of the data may have been censored: then, for some $I \subseteq \{1, \ldots, n\}$, the values Y_i with $i \in I$ are crisp (i.e. $\underline{Y}_i = \overline{Y}_i$).

*Corresponding author: cernym@vse.cz

The case (a) can be seen as a special case of (b). The method introduced in the following sections is applicable to the more general case (b).

A variety of methods for estimation of regression parameters in regression involving interval data has been developed; they are studied in statistics [1, 2, 3, 4, 5, 6, 7], where also robust regression methods have been proposed [8, 9], in fuzzy theory [10, 11, 12, 13, 14, 15, 16] as well as in computer science [17], [18], [19]. An algebraic treatment of least squares methods for interval data has been considered in [20] and [21].

There are classical works dealing with rounding of data included in regression analysis [22, 23, 24] as well as modern works on the topic [25, 26, 27, 28].

A majority of the cited papers deals with the basic issue how to derive a "good" crisp estimator of β from data affected by rounding/censoring. Our approach is complementary: our goal is not to derive an estimator of β but rather to describe the set in which a given linear estimator β can be when crisp values of y are replaced by rounded/censored values.

2. The possibilistic approach

Definition 1. *Let Y denote the interval vector $[\underline{Y},\overline{Y}]$. The **OLS-set** associated with Y (and the matrix X, which is assumed to be fixed) is defined as*

$$OLS(Y) = \{\beta \in \mathbb{R}^p : (\exists y \in Y)[X^{\mathrm{T}}X\beta = X^{\mathrm{T}}y]\}.$$

The motivation for Definition 1 is straightforward. Our aim is to use least squares to obtain an estimate of the unknown vector of regression parameters β in the model (1). However, we only know intervals Y that are guaranteed to contain the directly unobservable data y. Then, the set $OLS(Y)$ contains *all possible* values of $\widehat{\beta}$ as y ranges over Y. The set $OLS(Y)$ is a possibilistic version of the notion of the OLS-estimator.

The set $OLS(Y)$ captures the loss of information caused by rounding/censoring 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. A geometric characterization of that set will be given in the next section.

When $p = 2$ or $p = 3$ then the set $OLS(Y)$ 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 $OLS(Y)$.

The possibilistic approach is essentially algebraic or geometric, not probabilistic: it does not assume any distribution of y on Y. It allows to answer such questions as "*is it true that a given vector b fulfills $b \in OLS(Y)$?*", i.e. *is it true that if the truly observed values y had been available, we could have estimated $\widehat{\beta} = b$?* If b is a bad scenario, then a negative answer allows to rule the scenario out. (See also Section 7.)

The possibilistic approach also allows to derive bounds on the set $OLS(Y)$ giving information about the possible worst-case impact of rounding/censoring on the deviation of the OLS estimator $\widehat{\beta}$ from (say) its central value $\tilde{\beta} := \frac{1}{2}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}(\underline{Y}+\overline{Y})$. This approach is illustrated in Section 6.

Several measures can be introduced to quantify the rounding/censoring effect: the essence is that if the set $OLS(Y)$ is in some sense small, then the rounding/censoring impact on the estimator can be regarded as negligible. Natural measures include the volume of the set $OLS(Y)$ and the radius of the smallest circle circumscribing the set $OLS(Y)$.

However, we can also regard the set $OLS(Y)$ in a probabilistic way.

Probabilistic interpretation of the possibilistic approach. If y is a random vector such that the support of its distribution is Y, then the support of the distribution of $(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$ is $OLS(Y)$. Then the set $OLS(Y)$ can be seen as *100% confidence region* for the OLS estimator. An interesting special case is a regression model with independent disturbances with distributions the supports of which are bounded.

3. Geometry of the set $OLS(Y)$

First we need to review some notions from geometry of convex polyhedra; for further reading see [29].

Definition 2. *The **Minkowski sum** of a set $A \subseteq \mathbb{R}^k$ and a vector $g \in \mathbb{R}^k$ is the set*

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

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

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

where *conv* denotes the convex hull.

Definition 3. *The **zonotope** generated by $g_1,\dots,g_N \in \mathbb{R}^k$ with shift $s \in \mathbb{R}^k$ is the set*

$$\mathscr{Z}(s; g_1,\dots,g_N) = (\cdots((\{s\}\oplus g_1)\oplus g_2)\oplus\cdots\oplus g_N).$$

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

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

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

The main result of this section follows.

Theorem 4. *Let $X \in \mathbb{R}^{n\times p}$ be a matrix of full column rank and $Y = [\underline{Y},\overline{Y}]$ an $n \times 1$ interval vector. Then*

$$OLS(Y) = \mathscr{Z}(Q\underline{Y};\ Q_1(\overline{Y}_1 - \underline{Y}_1),\dots,Q_n(\overline{Y}_n - \underline{Y}_n)),$$

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

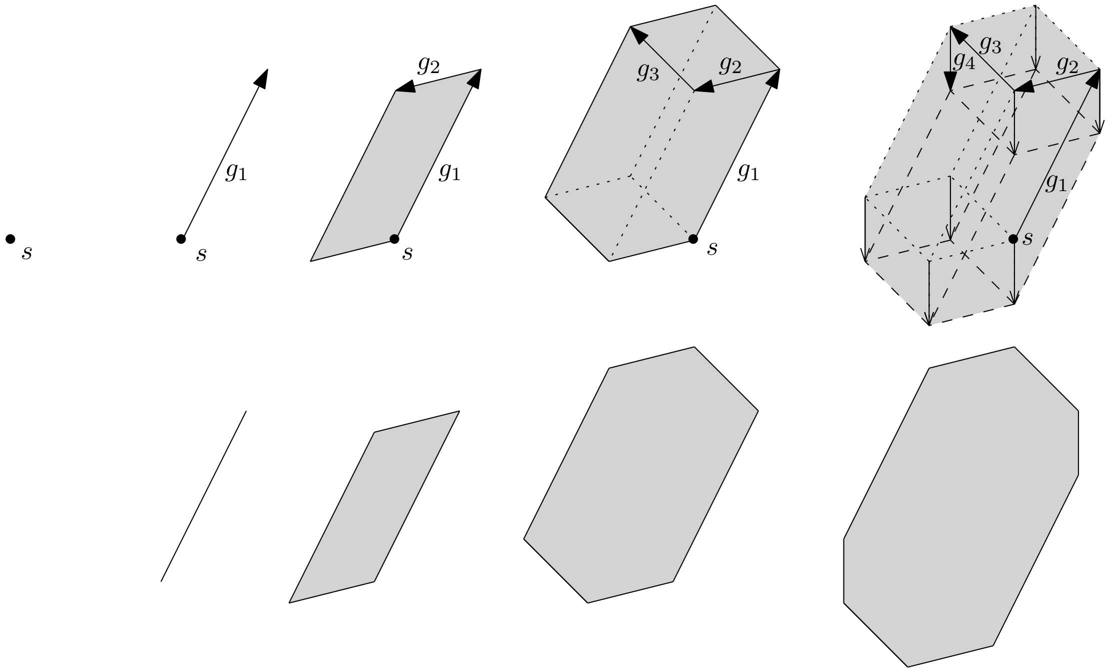

Fig. 1: The evolution of a zonotope $\mathscr{Z}(s;g_1,g_2,g_3,g_4)$.

Proof.

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

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 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 $OLS(Y)$ is an image of Y under the mapping determined by the matrix $Q = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$.

Hence, we have found that the set $OLS(Y)$ is a convex polyhedron in the space of regression parameters. Moreover, from the Figure 1 it is clear that the set $OLS(Y)$ is center-symmetric and the center point is $\tilde{\beta} = \frac{1}{2}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}(\underline{Y}+\overline{Y})$.

4. Complexity of the polyhedron $OLS(Y)$

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

(b) description of the zonotope $OLS(Y)$ by the enumeration of vertices;

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

The description (a) has been given by Theorem 4.

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.

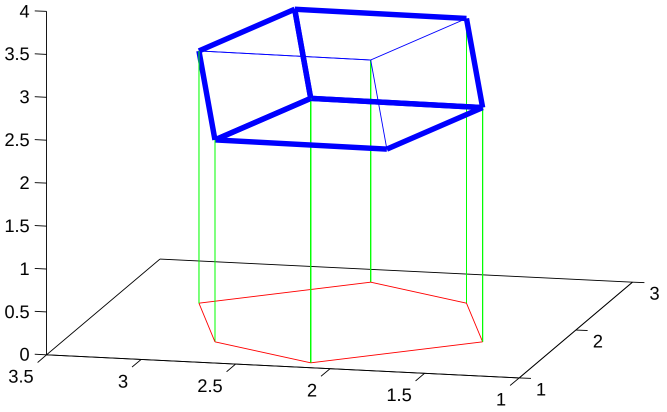

Fig. 2: A zonotope as an image of a higher-dimensional cube.

Theorem 5 ([29]). *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 computable in polynomial time.

However, a short look at Theorem 5 shows that we can also derive a positive result. 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), then we have:

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

Proof. We have

$$\begin{aligned} F(Z) &\leq 2\binom{n}{p-1} \\ &= \frac{2n(n-1)\cdots(n-p+2)}{(p-1)!} \\ &\leq 2n^{p-1} \\ &\leq O(n^{p-1}) \end{aligned} \tag{3}$$

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 (3). □

The Corollary shows that if p is fixed, then the set $OLS(Y)$ cannot have more than a polynomially bounded number of vertices and facets. Now a question arises whether the enumerations of them can be computed in polynomial time.

The answer is positive. 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 the size of output; see [30] and [31]. In Corollary 1 we have shown that if p is fixed then the size of the output is polynomially bounded in the size of the input. Hence:

Corollary 2. *Let p be fixed. If the vectors $\underline{Y}, \overline{Y}$ are rational and the matrix X is rational and has full column rank, then:*

(a) the list of vertices of the polyhedron $OLS(Y)$ can be computed in time bounded by a polynomial in n;

(b) a matrix A and a vector c such that

$$OLS(Y) = \{b \in \mathbb{R}^p : Ab \leq c\}$$

can be computed in time bounded by a polynomial in n.

5. Approximations of the polyhedron $OLS(Y)$

By Corollary 2, the descriptions of the set $OLS(Y)$ in terms of the lists of vertices and facets can be constructed in polynomial time when p is fixed. However, these descriptions need not be user-friendly: if, say, $p = 4$ and $n = 100$ then the enumeration of vertices and facets can fill up a thick book!

In this section we derive two simple approximations that can be useful in practice.

Interval approximation. It is easily seen that for every i and every $b \in OLS(Y)$ it holds

$$\overbrace{\sum_{j=1}^{n} \min\{Q_{ij}\underline{Y}_j, Q_{ij}\overline{Y}_j\}}^{=:\underline{b}_i} \leq b_i$$

$$\leq \underbrace{\sum_{j=1}^{n} \max\{Q_{ij}\underline{y}_j, Q_{ij}\overline{y}_j\}}_{=:\overline{b}_i} \qquad (4)$$

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

$$B = [\underline{b}, \overline{b}] \qquad (5)$$

is the smallest cube enclosing the polyhedron $OLS(Y)$.

The bound B can be easily computed in polynomial time.

The bound B 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 rounding/censoring effect is insignificant for estimation of the i-th parameter*.

Ellipsoidal approximation. The smallest ellipse $\mathscr{E}$ containing $OLS(X,y)$ 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 $\mathscr{E}$ is described by a center point s and a positive definite matrix E such that

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

We do not know a polynomial-time algorithm for construction of the Löwner-John ellipse for the set $OLS(Y)$. 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 circumscribing polyhedra is found in [32].)

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

$$\begin{aligned} s &= \tfrac{1}{2}Q(\overline{Y} + \underline{Y}), \\ E &= Q \cdot diag\left(\tfrac{n}{4}(\overline{Y}_1 - \underline{Y}_1)^2, \ldots, \tfrac{n}{4}(\overline{Y}_n - \underline{Y}_n)^2\right) \cdot Q^{\mathrm{T}}, \end{aligned} \qquad (6)$$

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 circumscribing Y in $\mathbb{R}^n$ under the mapping $\upsilon \mapsto Q\upsilon$. This proves $Z \subseteq \mathscr{E}$.

The ellipse $\mathscr{E}$ can be computed in polynomial time.

6. Example

Consider the regression model

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \qquad (7)$$

with $n = 11$ observations collected in the following table. Only integer-rounded values $\tilde{y}_1, \ldots \tilde{y}_{11}$ are available to us; thus, for all $i = 1, \ldots, 11$,

$$Y_i = [\underline{Y}_i, \overline{Y}_i] = [\tilde{y}_i - \tfrac{1}{2}, \tilde{y}_i + \tfrac{1}{2}].$$

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

Using the central estimator

$$\tilde{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\tilde{y} \qquad (8)$$

we get

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

and using (4) we get

$$[\underline{b}_1, \overline{b}_1] = [1.56, 2.69], \quad [\underline{b}_2, \overline{b}_2] = [1.06, 1.34]. \qquad (9)$$

The rounding 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 zonotope Z, together with the cube $[\underline{b}, \overline{b}]$ and the ellipse (6), is plotted in Figure 3.

Though the approximations 1 and 2 are quite trivial, their combination gives some nontrivial information. The interval $[\underline{b}, \overline{b}]$ contains the point $[1.56, 1.06]$; hence, the enclosure (9) does not rule out the case that *both regression parameters could be 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 $[1.65, 1.06] \notin \mathscr{E}$.

Remark. Observe that in the Example, the width of the interval $[\underline{b}_1, \overline{b}_1]$ in (9) for the intercept β_1 in the model (7) is greater than one, while all of the intervals $[\underline{Y}_i, \overline{Y}_i]$ are of width 1. Hence it is not true that the maximal intercept β_1 is achieved in the case $y = \overline{Y}$ and the minimal intercept is achieved in the case $y = \underline{Y}$ (as these two cases produce intercepts the difference of which is 1). Indeed, $(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y^* = (2.69, 1.12)^{\mathrm{T}}$ and $(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y^{**} = (1.56, 1.28)^{\mathrm{T}}$ with

$$y^* = (\overline{Y}_1, \overline{Y}_2, \overline{Y}_3, \overline{Y}_4, \overline{Y}_5, \overline{Y}_6, \overline{Y}_7, \overline{Y}_8, \overline{Y}_9, \underline{Y}_{10}, \underline{Y}_{11})^{\mathrm{T}}$$

and

$$y^{**} = (\underline{Y}_1, \underline{Y}_2, \underline{Y}_3, \underline{Y}_4, \underline{Y}_5, \underline{Y}_6, \underline{Y}_7, \underline{Y}_8, \underline{Y}_9, \overline{Y}_{10}, \overline{Y}_{11})^{\mathrm{T}}.$$

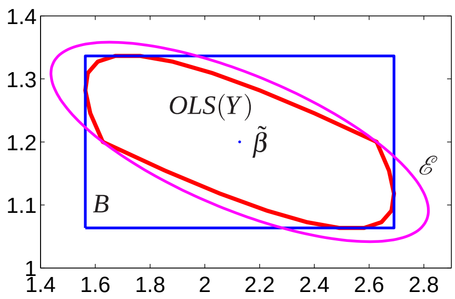

Fig. 3: The zonotope Z for the regression model in the Example and its approximations B and $\mathscr{E}$ given by (5) and (6), respectively.

7. Admissibility; volume of $OLS(Y)$

As motivated by the 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 $\tilde{\beta}+\Delta$ is in $OLS(Y)$ or not. A vector b (in particular, a vector b of the form $b=\tilde{\beta}+\Delta$) is called *admissible* if $b \in OLS(Y)$.

Proposition 6. *Admissibility can be tested in polynomial time.*

Proof. The vector b is admissible if and only if there is a y such that $Qy=b$ and $\underline{y} \le y \le \overline{y}$, where $Q=(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. Hence, deciding admissibility amounts to deciding the feasibility of a system of linear (in)equalities, which is essentially a linear programming problem. □

The Proposition, combined with (4), suggests a procedure for Monte-Carlo approximation of the volume of $OLS(Y)$, which is a natural measure of its size: just generate a random point $b \in [\underline{b},\overline{b}]$ and test its admissibility. This procedure is interesting in particular in higher dimensions, where the polyhedron $OLS(Y)$ cannot be easily visualized.

Though the volume of $OLS(Y)$ can be computed exactly, no polynomial-time algorithm (in n,p) is known; hence, the Monte Carlo approximation is a reasonable choice.

8. Another example

In this example we show that the underlying theory can be used as a simple proof technique. Consider the model of location

$$y_i=\beta+\varepsilon_i, \quad i=1,\dots,n, \tag{10}$$

with rounded observations $Y_i=[\underline{Y}_i,\overline{Y}_i]$. The parameter space is one-dimensional in this case; now $OLS(Y)$ is a one-dimensional interval which coincides with (4). Thus,

$$OLS(Y)=[\underline{b},\overline{b}]=\left[\frac{1}{n}\sum_{i=1}^{n}\underline{Y}_i,\frac{1}{n}\sum_{i=1}^{n}\overline{Y}_i\right].$$

The central estimator (8) takes the form

$$\tilde{\beta}=\frac{1}{2n}\sum_{i=1}^{n}(\overline{Y}_i+\underline{Y}_i)$$

in the model (10). For any estimator $\widehat{\beta}$, define the error function

$$\eta(\widehat{\beta})=\begin{cases}\max\{\overline{b}-\widehat{\beta},\widehat{\beta}-\underline{b}\} & \text{if } \widehat{\beta}\in OLS(Y),\\ \infty & \text{if } \widehat{\beta}\notin OLS(Y).\end{cases}$$

Now $\tilde{\beta}$, being the central estimator, minimizes $\eta(\widehat{\beta})$. Hence, in this sense it is optimal. This is a justification of the intuitive fact that taking centers (i.e. the rounded values) is the best we can do.

9. Conclusion

It is interesting to observe that while the location of the polyhedron $OLS(Y)$ in the parameter space depends on both $\underline{Y}$ and $\overline{Y}$, its size and shape depends only on $\overline{Y}-\underline{Y}$ (assuming the matrix X fixed), i.e. on the widths of the intervals $Y_1,\dots,Y_n$. Therefore, the bounds on the worst-case error introduced in Section 5 (say, the numbers $\overline{b}_i-\underline{b}_i$ in (4) or the length of the longest semiaxis of the ellipse (6)) depend only on the widths of the intervals $Y_1,\dots,Y_n$, which are often known or may be chosen in advance, for example by the choice of precision of measurement or precision of data storage. It follows that the impact of rounding/censoring on the OLS estimator of regression parameters can be analyzed in advance, before the measurement of y is performed. The analysis of the shape and size of the set $OLS(Y)$ then can give useful information on the choice of precision in an experiment being planned.

10. Acknowledgements

The work of both authors was supported by Project No. F4/18/2011 of Internal Grant Agency of University of Economics, Prague, Czech Republic. Thanks to anonymous referees for fruitful comments.

References

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

[2] Jun-peng, G., Wen-hua, L. (2008). 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.

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

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

[5] Moral-Arce, I., Rodríguez-Póo, J. M., Sperlich, S. (2011). Low dimensional semiparametric estimation in a censored regression model. *Journal of Multivariate Analysis* 102 (1), 118–129.

[6] Pan, W., Chappell, R. (1998). 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), 33–50.

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

[8] Inuiguchi, M., Fujita, H., Tanino, T. (2002). 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.

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

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

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

[12] Hladík, M., Černý, M. (2010). New approach to interval linear regression. In: Kasımbeyli, R., 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, 2010, 167–171.

[13] Tanaka, H., Lee, H. (1997). 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.

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

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

[16] Černý, M., Rada, M. (2010). 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, 276–282.

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

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

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

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

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

[22] Sheppard, W. (1898). On the calculation of the most probable values of frequency constants for data arranged according to equidistant divisions of a scale. *Proceedings of the London Mathematical Society* 29, 353–380.

[23] Kendall, M. G. (1938). The conditions under which Sheppard's corrections are valid. *Journal of the Royal Statistical Society* 101, 592–605.

[24] Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics* 3, 1–21.

[25] Schneeweiss, H., Komlos, J. (2008). Probabilistic rounding and Sheppard's correction. *Technical report* 45, Department of Statistics, University of Munich. Available at: http://epub.ub.uni-muenchen.de/8661/1/tr045.pdf.

[26] Di Nardo, E. (2010). A new approach to Sheppard's corrections. *Mathematical Methods of Statistics*, 19 (2), 151-162.

[27] Wimmer, G., Witkovský, V. (2002). Proper rounding of the measurement results under the assumption of uniform distribution. *Measurement Science Review* 2 (1), 1–7.

[28] Wimmer, G., Witkovský, V., Duby, T. (2000). Proper rounding of the measurement results under normality assumptions. *Measurement Science and Technology* 11, 1659–1665.

[29] Ziegler, G. (2004). *Lectures on polytopes*, Germany: Springer.

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

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

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

Received January 24, 2011.
Accepted April 27, 2011.