On the OLS set in linear regression with interval data

Miroslav Rada   Michal Černý

Department of Econometrics
University of Economics, Prague

SCAN 2012, Novosibirsk
23.9. – 29.9.2012

Linear regression model

Consider the classical statistical instrument for testing and estimating dependencies — linear regression model:

$$y = X\beta + \epsilon$$

- y – observations of dependent var (say *model output*) – $[n \times 1]$
- X – observations of independent vars (say *model input*) – $[n \times p]$
- β – unknown true regression parameters – $[p \times 1]$
- ϵ – disturbances – $[n \times 1]$

Linear regression model

Consider the classical statistical instrument for testing and estimating dependencies — linear regression model:

$$y = X\beta + \epsilon$$

- y – observations of dependent var (say *model output*) – $[n \times 1]$
- X – observations of independent vars (say *model input*) – $[n \times p]$
- β – unknown true regression parameters – $[p \times 1]$
- ϵ – disturbances – $[n \times 1]$

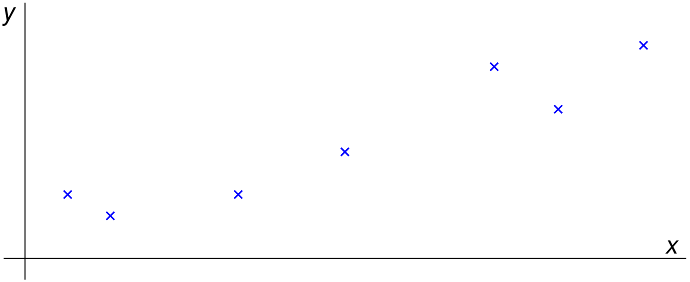
The tuple (X, y) – *data* of the model
Assumption: β can be estimated by a linear estimator, ie. $\hat{\beta} = Qy$.
In case of

- $Q = (X^\mathsf{T} X)^{-1} X^\mathsf{T}$, the estimator is called Ordinary least squares (OLS) and $\hat{\beta} := (X^\mathsf{T} X)^{-1} X^\mathsf{T} y$ is called OLS-solution of classical linear regression model,
- $Q = (X^\mathsf{T} \Omega^{-1} X)^{-1} \Omega^{-1} X^\mathsf{T}$, where Ω is a positive definite matrix, the estimator is called Generalized least squares (GLS).

Linear regression model — example

Consider a model with one input variable and a constant:

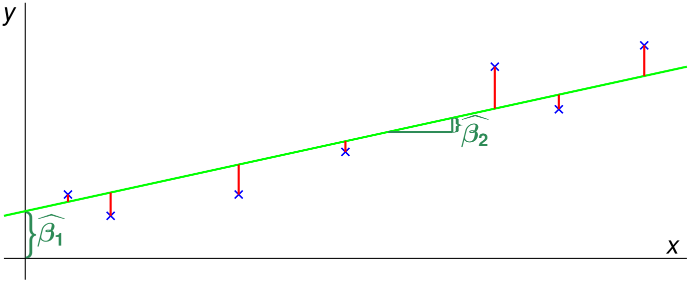
$$X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 10 & 15 & 22 & 26 & 31 \end{pmatrix} \qquad y^{\mathsf{T}} = \begin{pmatrix} 3 & 2 & 3 & 5 & 9 & 7 & 10 \end{pmatrix}$$



Linear regression model — example

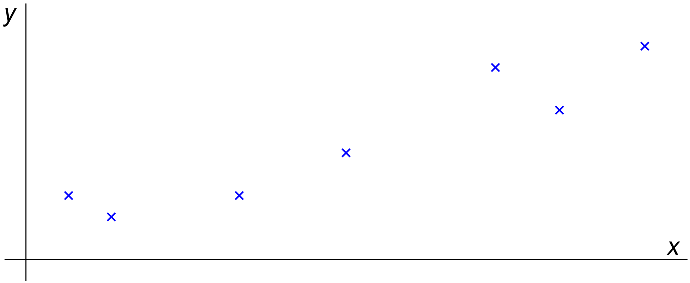
OLS-solution of regression parameters consists of finding such hyper-plane (line), that has the least sum of squares of disturbances:

$$\hat{\beta} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}y$$



Interval data in the model

But what when we allow interval data instead of crisp only?



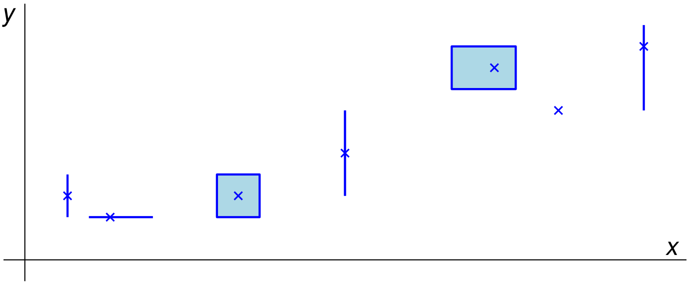
Interval data in the model

But what when we allow interval data instead of crisp only?

$$\boldsymbol{X}^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & [3,6] & [9,11] & 15 & [20,23] & 26 & 31 \end{pmatrix}$$

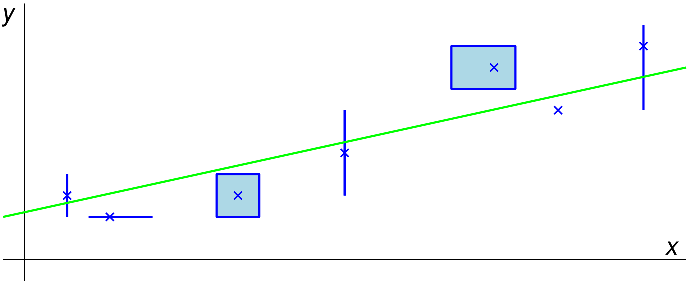
$$\boldsymbol{y}^T = \begin{pmatrix} [2,4] & 2 & [2,4] & 5 & [8,10] & 7 & [7,11] \end{pmatrix}$$

There can be interval observations of output variable, of input variable or of both.



Interval data in the model

One may compute OLS-solution for somehow chosen crisp values from the intervals, for example for the values in previous example.

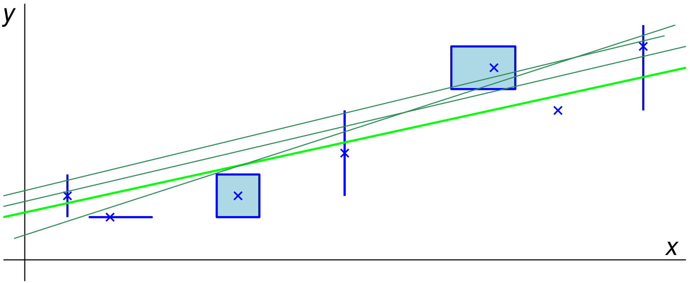


Interval data in the model

One may compute OLS-solution for somehow chosen crisp values from the intervals, for example for the values in previous example.

In our presentation:

- we focus on the set of **all possible** OLS-solutions one can obtain,
- then, we will focus on the special case when model input (X) is crisp (i.e., only the output $\boldsymbol{y}$ is interval).



Interval data in the model

One may compute OLS-solution for somehow chosen crisp values from the intervals, for example for the values in previous example.

In our presentation:

- we focus on the set of **all possible** OLS-solutions one can obtain,
- then, we will focus on the special case when model input (X) is crisp (i.e., only the output $\boldsymbol{y}$ is interval).

More formally, the *interval regression model (IRM)* is the structure

$$\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon,$$

where $\boldsymbol{X}$ is an interval $n \times p$ matrix $[\underline{X}, \overline{X}]$ and $\boldsymbol{y}$ is an interval $n \times 1$ vector $[\underline{y}, \overline{y}]$.

The *OLS-solution of IRM* is the set

$$\{\beta \in \mathbb{R}^p : X^\mathsf{T} X\beta = X^\mathsf{T} y, X \in \boldsymbol{X}, y \in \boldsymbol{y}\}.$$

Applications for IRM

Interval data in regression analysis can arise

- when rounding or representing data as data-types with restricted precision,
- in case of loss of information, for example when categorizing or censoring data, when discretizing continuous data,
- when dealing with unstable data,
  - in case of unstability of physical "constants", e.g. gravity acceleration, though often treated as a constant, slightly changes with position,
  - in case of changes of the observed variable inside a period (the day-closing prices of stocks don't capture their fluctuation during the day),
- in case of expert predictions or forecasts,
- in statistics, e.g. interval predictions of one model can act as input data for another model.

Ideas for studying IRM

- Replacement of (X, y) by $(\boldsymbol{X}, \boldsymbol{y})$ brings some uncertainty or loss of information
- Our aim is to quantify such loss of information
- The OLS-solution of IRM describes all possible estimates of classical model as (X, y) ranges over $(\boldsymbol{X}, \boldsymbol{y})$.
- OLS-solution of IRM can be viewed as "implicit representation" of the brought uncertainty, should be studied to analyse whether the uncertainty is "significant" or "serious".

Goal:
We want to find out how the OLS-solution of IRM looks like, to find some descriptive characteristics of it.

Negative result — studying IRM is "hard"

The definition

$$\{\beta \in \mathbb{R}^p : X^\mathsf{T} X\beta = X^\mathsf{T} y, X \in \boldsymbol{X}, y \in \boldsymbol{y}\}.$$

of the OLS-solution of IRM doesn't testify how it looks like. So, one may be familiar with constructing interval enclosure for that set. Unfortunately, it is not easy task, as follows from the following theorem:

Theorem

*It is **NP**-hard to decide whether the OLS-solution of IRM is a bounded set.*

Hence, the construction of interval enclosure, neither tighter nor less tight, is very hard problem in general.

Furthermore, the OLS-solution of IRM need not be a convex set.

Positive result — special cases can be handled easier

In the rest of the presentation, we will focus on the special case of IRM: the crisp input-interval output model.

Definition

Let $X \in \mathbb{R}^{n \times p}$, $\boldsymbol{y} \in \mathbb{IR}^n$ and let Q be defined as $Q := (X^\mathsf{T} X)^{-1} X^\mathsf{T}$.
The tuple $(X, \boldsymbol{y})$ denotes the (data of) **interval-output** (linear regression) **model** (henceforth shortly **IOM**).
The set $\{\beta \in \mathbb{R}^p : \beta = Qy; \underline{y} \leq y \leq \overline{y}; \}$ is called *OLS-solution of IOM*.

The OLS-solution of IOM is clearly bounded and convex, and thus computationaly easier to handle.

Positive result — special cases can be handled easier

In the rest of the presentation, we will focus on the special case of IRM: the crisp input-interval output model.

Definition

Let $X \in \mathbb{R}^{n \times p}$, $\boldsymbol{y} \in \mathbb{IR}^n$ and let Q be defined as $Q := (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}$.
The tuple $(X, \boldsymbol{y})$ denotes the (data of) **interval-output** (linear regression) **model** (henceforth shortly **IOM**).
The set $\{\beta \in \mathbb{R}^p : \beta = Qy; \underline{y} \leq y \leq \overline{y}; \}$ is called *OLS-solution of IOM*.

The OLS-solution of IOM is clearly bounded and convex, and thus computationaly easier to handle.
The interval data $\boldsymbol{y}$ describes a box in dimension n.

Positive result — special cases can be handled easier

In the rest of the presentation, we will focus on the special case of IRM: the crisp input-interval output model.

Definition

Let $X \in \mathbb{R}^{n \times p}$, $\boldsymbol{y} \in \mathbb{IR}^n$ and let Q be defined as $Q := (X^\mathsf{T}X)^{-1}X^\mathsf{T}$.
The tuple $(X, \boldsymbol{y})$ denotes the (data of) **interval-output** (linear regression) **model** (henceforth shortly **IOM**).
The set $\{\beta \in \mathbb{R}^p : \beta = Qy; \underline{y} \leq y \leq \overline{y}; \}$ is called *OLS-solution of IOM*.

The OLS-solution of IOM is clearly bounded and convex, and thus computationaly easier to handle.
The interval data $\boldsymbol{y}$ describes a box in dimension n.
The box is then linearly projected to parameter space $\mathbb{R}^p$.
Observe that the image must be bounded and convex. In fact, it's a zonotope, well-known type of polytope.

Possible characteristics of OLS-solution of IOM

Knowing that the OLS-solution of IOM is a polytope, we can describe it using characteristics common for polytopes, and possibly use known algorithms for obtaining such characteristics. The characteristics are:

1. interval enclosure – extremal values for individual regression parameters,
2. ellipsoidial approximations – replacement of a combinatorially complex polytope by a simple set, an ellipsoid (sometimes referred to as “rounding of polytopes”),
3. volume – natural measure of uncertainty brought to the model by replacement crisp data by interval data,
4. list of vertices – extremal values for all parameters together and
5. list of facets.

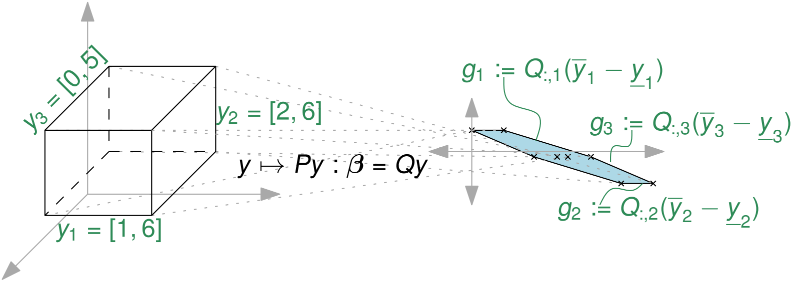
Furthermore, the OLS-solution is a zonotope, a polytope with special properties, that can be utilized for developing more efficient algorithms than algorithms for general polytopes.

Zonotope as an image of a hypercube

Definition

Zonotope is an image of high-dimensional box in a lower (or equal) dimension under a linear projection $y \mapsto Py$.

In fact, we use $P := Q = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}$.



$$X = \begin{pmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 5 \end{pmatrix} \qquad \boldsymbol{y} = \begin{pmatrix} [1,6] \\ [2,6] \\ [0,5] \end{pmatrix} \qquad Q = \begin{pmatrix} -13/12 & 1/4 & -5/12 \\ -1/4 & 0 & 1/4 \end{pmatrix}$$

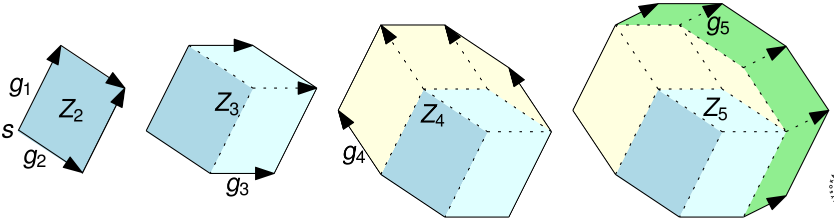
Zonotope as a Minkowski sum

Minkowski sum of set $A \subset \mathbb{R}^p$ and vector $b \in \mathbb{R}^p$ is the operation $A \dot{+} b$ defined as $A \dot{+} b := \{a + \alpha b : a \in A, 0 \le \alpha \le 1\}$.

Lemma

A) Every zonotope $Z \subset \mathbb{R}^p$ can be expressed as a Minkowski sum of a shift $s \in \mathbb{R}^p$ and a set of vectors (called generators) $g_1, \ldots, g_n \in \mathbb{R}^p$.

B) Given an IOM with data $(X, \boldsymbol{y})$, the OLS-solution for that IOM is a zonotope with shift $s := Q(\underline{y}_1 + \cdots + \underline{y}_n)$ and generators $g_1 := Q_{:,1}(\overline{y}_1 - \underline{y}_1), \ldots, g_n := Q_{:,n}(\overline{y}_n - \underline{y}_n)$.



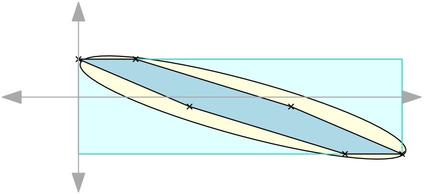
Interval enclosure

From now on we will use terms *OLS-solution of an IOM* and *zonotope* in the same sense.

Standard representation of a zonotope will be the $(n+1)$-tuple $(s, g_1, \ldots, g_n)$ and will be called *generator description*.

The interval enclosure of a zonotope can be constructed easily for example by evaluation of $(X^\mathrm{T} X)^{-1} X^\mathrm{T} \boldsymbol{y}$ using the interval arithmetic. Unfortunately, such enclosure can be very redundant if the zonotope is "noodle-like" in some direction.

Hence, it is reasonable to seek for better enclosures, such as ellipsoidal approximations.



Ellipsoidal approximation in general

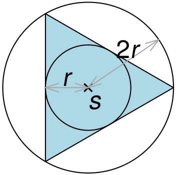
Let $E \in \mathbb{R}^{p \times p}$ be a positive definite matrix and $s \in \mathbb{R}^p$ be a center point. The symbol $\mathcal{E}(E, s)$ stands for the ellipsoid

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

Goffin's algorithm is an algorithm (based on the Shallow Cut Ellipsoid Method) which, for every fixed $\varepsilon > 0$, finds in poly-time an ellipsoidal approximation of a given convex polytope P, represented as an inequality $Ax \leq b$, such that

$$\mathcal{E}(p^{-2}E, s) \subseteq P \subseteq \mathcal{E}(E(1+\varepsilon), s).$$

Observe that approximation is up to the tolerance ε the best possible: the regular 2-simplex serves as example:



where $E = \begin{pmatrix} 4r^2 & 0 \\ 0 & 4r^2 \end{pmatrix}$

Ellipsoidal approximation of a zonotope

We adapted Goffin's algorithm for zonotopes given by generator descriptions, achieving the same tightness of approximation. Hence, we can state the following theorem:

Theorem

Let $\varepsilon > 0$ be fixed. Given a zonotope Z represented by generator description, there exists a polynomial time algorithm that finds a matrix E and shift s such that

$$\mathcal{E}(p^{-2}E, s) \subseteq Z \subseteq \mathcal{E}(E(1+\varepsilon), s).$$

Remark: For a centrally symmetric set Z, Jordan's theorem assures existence of approximation in form $\mathcal{E}(p^{-1}E, s) \subseteq Z \subseteq \mathcal{E}(E(1+\varepsilon), s)$, which theoretically allows better tightness of approximation than we achieved. On the other hand, we couldn't process a crucial step of the algorithm, testing whether polytope contains a ball, in polynomial time, hence we lost the factor p again, achieving the same tightness as the original Goffin's algorithm.

Volume computation

- volume of a zonotope is a natural measure of uncertainty brought to a regression model by interval data
- $\#P$-hard problem (Dyer et. al. (1998))
- given a generator description of a zonotope, volume computation consists in computing $\binom{n}{p}$ determinants

$$\mathrm{vol}(Z) = \sum_{1 \leq i_1 < \cdots < i_p \leq n} \left| \det \left(g_{i_1}, \ldots, g_{i_p} \right) \right|$$

- formula is based on the fact that zonotope can be decomposed into parallelotopes
- we proposed an algorithm called *RRR* (details are omitted) which computes exact volume in time $\mathcal{O}(\binom{m-1}{d-1}(md + d^3))$, doing some computations simultaneously
- there is randomized polynomial algorithm by Dyer et. al. (1998): given relative error bound and a probability for attaining this bound, it computes the volume up to the given bound with the given probability

Vertex and facet enumeration

- problem is in the possible size of output (f_0 – number of vertices, f_{p-1} – number of facets):

$$f_0 \leq 2\sum_{i=0}^{p-1}\binom{n-1}{i}, \qquad f_{p-1} \leq 2\binom{n}{p-1}.$$

  Moreover, these bounds are attained for some zonotopes (which can also be OLS-solutions of IOMs).
- Hence, the number of vertices and facets of a zonotope may be superpolynomial in dimension and number of generators.
- For such problems, algorithms with following properties may be useful:
  - compactness – space complexity polynomial in the input size
  - output-polynomiality – time complexity polynomial in the output size

Vertex and facet enumeration

- Our *RRR* algorithm can be (besides the volume computation) used for enumeration of facets and vertices with minimal added effort.

  However, it is neither compact nor output-polynomial.
- For vertex enumeration, the idea behind *RRR* can be modified to obtain compact and output-polynomial algorithm.

  Another algorithm with such propertites (Fukuda, Avis (1993)) is known, as well as an asymptotically optimal (noncompact) algorithm by Edelsbrunner and O'Rourke (1986).
- For facet enumeration, there is (noncompact) output-polynomial algorithm by Seymour (1994).

Summary

- We dealt with linear regression model with interval data and discussed the properties of the set of all OLS-solutions.
- We showed that for general model with interval input and output "everything is computationally hard".
- For the special case of crisp input-interval output the set of all OLS-solutions is a zonotope — a polytope with special structure.
- For zonotope, we can compute interval enclosure, ellipsoidal approximation and volume approximation in polynomial time.
- Exact volume computation, vertex enumeration and facet enumeration can't be accomplished in polynomial time, although there exists "efficient" algorithms for these problems.

Thank you for attention!