Information Sciences 244 (2013) 26–47

Contents lists available at SciVerse ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

On the possibilistic approach to linear regression models involving uncertain, indeterminate or interval data

Michal Černý [a,*], Jaromír Antoch [b], Milan Hladík [a,c]

[a] Department of Econometrics, Faculty of Computer Science and Statistics, University of Economics, Prague, Winston Churchill Square 4, 13067 Prague, Czech Republic
[b] Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Prague, Sokolovská 83, 18675 Prague, Czech Republic
[c] Department of Applied Mathematics, Faculty of Mathematics and Physics, Charles University, Prague, Malostranské náměstí 25, 11000 Prague, Czech Republic

ARTICLE INFO

Article history:
Received 14 September 2011
Received in revised form 29 March 2013
Accepted 28 April 2013
Available online 6 May 2013

Keywords:
Interval data
Uncertain data
Possibilistic regression
Computational complexity

ABSTRACT

We consider linear regression models where both input data (the observations of independent variables) and output data (the observations of the dependent variable) are affected by loss of information caused by uncertainty, indeterminacy, rounding or censoring. Instead of real-valued (crisp) data, only intervals are available. We study a possibilistic generalization of the least squares estimator, so called OLS-set for the interval model. Investigation of the OLS-set allows us to quantify whether the replacement of real-valued (crisp) data by interval values can have a significant impact on our knowledge of the value of the OLS estimator. We show that in the general case, very elementary questions about properties of the OLS-set are computationally intractable (assuming $P \neq NP$). We also focus on restricted versions of the general interval linear regression model to the crisp input case. Taking the advantage of the fact that in the crisp input – interval output model the OLS-set is a zonotope, we design both exact and approximate methods for its description. We also discuss special cases of the regression model, e.g. a model with repeated observations.

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. For the purposes of this paper, we do not need to make any special assumptions on probabilistic properties of ε. We just assume that for estimation of β a linear estimator can be used, i.e. an estimator of the form

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

where Q is a matrix. In particular we shall concentrate on the Ordinary Least Squares (OLS) estimator, which corresponds to the choice $Q = (X^TX)^{-1}X^T$ in (2). (As it is well-known, this estimator is a "good" estimator e.g. when the disturbances are independent, identically distributed, with zero mean and finite variance.) Nevertheless, the theory is also applicable for other linear

* Corresponding author.
E-mail addresses: cernym@vse.cz (M. Černý), antoch@karlin.mff.cuni.cz (J. Antoch), milan.hladik@vse.cz, hladik@kam.mff.cuni.cz (M. Hladík).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.04.035

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 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 analysis of contaminated data.

Throughout the paper, the symbol n stands for the number of observations and the symbol p stands for the number of regression parameters, as it is usual in statistics.

We shall treat X and y as constants representing observed values of the independent variables and the dependent variable, respectively. Then the tuple (X,y) is called *data* for the regression model (1).

1.1. Interval data in the linear regression model

We shall study the situation when the data (X,y) cannot be observed directly. Instead of y_i and X_{ij}, only intervals of the form $[\underline{y}_i, \overline{y}_i]$ and $[\underline{X}_{ij}, \overline{X}_{ij}]$ are available, where it is guaranteed that for all $i \in \{1,\dots,n\}$ and $j \in \{1,\dots,p\}$,

$$y_i \in [\underline{y}_i, \overline{y}_i] \quad \text{and} \quad X_{ij} \in [\underline{X}_{ij}, \overline{X}_{ij}],$$

where y_i denotes the ith element of y and X_{ij} denotes the (i,j)th element of X.

The replacement of real-valued (crisp) data by intervals is henceforth referred to as "censoring". In some literature, this process is also called "trimming", "uncertaintification" or "intervalization".

1.2. Motivation

Inclusion of interval data in linear regression models is suitable for modeling 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 (i.e., $d = 0$), 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.
- The data are uncertain or unstable. For that reason it might be inappropriate to describe them in terms of fixed values (X,y) only.
- Categorical 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 quantities are represented by their average or median values. For example, if the exchange rate for a period of 1 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.
- Sometimes we use interval predictions as data in regression models. For example, consider a predictor of future inflation (an econometric model or a panel of experts, say), which is assumed to form inflation expectations. The predictions are interval. Then, another model—such as consumption model or capital expenditure model—uses the predicted inflation expectations as a regressor. Thus, the model has to be able to work with an interval regressor.

More applications of interval data in econometrics are found in [7]. Applications in information sciences can be found in [11]; see also applications in ergonomics [10], optimization and operational research [15,37,42,71], speech learning [45] and in pattern recognition [39,43].

A variety of methods for estimation of regression parameters in a regression with interval data has been developed; they are studied in statistics [8,22,36,41,44,49,55,76], where also robust regression methods have been proposed [32,50], in fuzzy theory [24,29,30,72–74] as well as in computer science [12,31,34]. An algebraic treatment of least squares methods for interval data has been considered in [5,18].

1.3. Crisp and interval numbers, vectors and matrices

We need to distinguish between real-valued data and interval data. In the context of this distinction, real-valued (or: numeric) data are called *crisp data*. Then, a *crisp number* is just a real number. Similarly, we say that a matrix/vector is *crisp* when we want to emphasize that all elements of the matrix/vector are real numbers (and not intervals). In general, the term "crisp" can be also understood as "non-interval".

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