LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Marco E. G. V. Cattaneo and Andrea Wiencierz

On the implementation of LIR: the case of simple linear regression with interval data

Technical Report Number 127, 2012
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

On the implementation of LIR: the case of simple linear regression with interval data

Marco E. G. V. Cattaneo, Andrea Wiencierz

Department of Statistics, LMU Munich, Ludwigstraße 33, 80539 München, Germany

Abstract

This paper considers the problem of simple linear regression with interval-censored data. That is, n pairs of intervals are observed instead of the n pairs of precise values for the two variables (dependent and independent). Each of these intervals is closed but possibly unbounded, and contains the corresponding (unobserved) value of the dependent or independent variable. The goal of the regression is to describe the relationship between (the precise values of) these two variables by means of a linear function.

Likelihood-based Imprecise Regression (LIR) is a recently introduced, very general approach to regression for imprecisely observed quantities. The result of a LIR analysis is in general set-valued: it consists of all regression functions that cannot be excluded on the basis of likelihood inference. These regression functions are said to be undominated.

Since the interval data can be unbounded, a robust regression method is necessary. Hence, we consider the robust LIR method based on the minimization of the residuals' quantiles. For this method, we prove that the set of all the intercept-slope pairs corresponding to the undominated regression functions is the union of finitely many polygons. We give an exact algorithm for determining this set (i.e., for determining the set-valued result of the robust LIR analysis), and show that it has worst-case time complexity $O(n^3 \log n)$. We have implemented this exact algorithm as part of the R package `linLIR`.

Keywords: simple linear regression, interval data, likelihood inference, robust regression, exact algorithm, R package

1. Introduction

Likelihood-based Imprecise Regression (LIR) is a recently introduced approach to regression for imprecisely observed quantities (see Cattaneo and Wiencierz, 2012, 2011). In this approach, it is assumed that the available data are coarse in the sense of Heitjan and Rubin (1991). That is, precise values of the quantities of interest exist, but we cannot observe them directly. Instead, we have only imprecise observations: these are subsets of the sample space, which we know to contain the precise values of the quantities of interest.

At the two extremes of the range of possible imprecise observations are the precise observations and the missing data, respectively. We have a precise observation when the imprecise observation contains a single value, which we then know to be the precise value of the quantity of interest (which in this case is thus indirectly observed). At the other extreme we have the missing data, which occur when the imprecise observation is the whole sample space, since in this case we learn nothing about the precise value of the quantity of interest.

Between these two extremes lies the whole range of possible imprecise observations, which can be any subset of the sample space. In particular, it can be argued that continuous quantities are always imprecisely observed, since no measuring device can be infinitely precise. Therefore, regression for imprecisely observed quantities is certainly an important topic in statistics. In fact, various regression methods have been proposed in several special cases (see

Email addresses: `cattaneo@stat.uni-muenchen.de` (Marco E. G. V. Cattaneo), `andrea.wiencierz@stat.uni-muenchen.de` (Andrea Wiencierz)

for example Beaton et al., 1976; Buckley and James, 1979; Dempster and Rubin, 1983; Li and Zhang, 1998; Pötter, 2000; Manski and Tamer, 2002; Marino and Palumbo, 2002; Gioia and Lauro, 2005; Ferson et al., 2007; Chen and Van Keilegom, 2009; Utkin and Coolen, 2011). In contrast to most of these proposals, LIR approaches the problem of regression with imprecisely observed quantities from a very general perspective.

The imprecise observations induce a likelihood function on the joint probability distributions of the random variables and random sets representing the precise values and imprecise observations, respectively. The result of a LIR analysis consists of all regression functions that cannot be excluded on the basis of likelihood inference. Hence, the result of a LIR analysis is in general set-valued (set-valued results are obtained for instance also by Manski and Tamer, 2002; Marino and Palumbo, 2002; Gioia and Lauro, 2005; Vansteelandt et al., 2006; Ferson et al., 2007). The extent of the set-valued result of a LIR analysis reflects the whole uncertainty in the regression problem with imprecisely observed quantities. That is, it encompasses the statistical uncertainty due to the finite sample as well as the indetermination related to the fact that the quantities are only imprecisely observed (these two kinds of uncertainty in the set-valued results are discerned for example also by Manski and Tamer, 2002; Vansteelandt et al., 2006).

In the present paper we consider a robust LIR method, in which quantiles of the residuals are used to compare the possible regression functions (see Cattaneo and Wiencierz, 2012, 2011). This method is closely related to the least median (or more generally, quantile) of squares regression, which is a very robust regression method for precisely observed quantities (see for example Rousseeuw, 1984; Hampel, 1975; Hampel et al., 1986; Rousseeuw and Leroy, 1987; Maronna et al., 2006; Huber and Ronchetti, 2009). Besides being a virtue by itself, the robustness of the regression method is almost necessary when dealing with possibly unbounded imprecise observations, because an unbounded imprecise observation means that the precise value can be arbitrarily far away. In practical applications, an unbounded imprecise observation can usually be replaced by a bounded (but very wide) one: the advantage of robust methods is that they do not depend (much) on the choice of the replacing imprecise observation.

In this paper we focus on the case of simple linear regression with interval data. That is, there are two variables of interest, which are real-valued and interval-censored (i.e., the imprecise observations are possibly unbounded intervals). For this situation, we develop the first exact algorithm to determine the set-valued result of the robust LIR method. The first part of this algorithm is related to the first exact algorithm for least median of squares regression, proposed by Steele and Steiger (1986) (see also Rousseeuw and Leroy, 1987, Chapter 5), which was also the basis of many other developments (see for example Souvaine and Steele, 1987; Edelsbrunner and Souvaine, 1990; Stromberg, 1993; Hawkins, 1993; Carrizosa and Plastria, 1995; Watson, 1998; Bernholt, 2005; Mount et al., 2007).

The paper is organized as follows. In the next section, we briefly present the robust LIR method in the framework of simple linear regression with interval data. Section 3 contains the main results of the paper, expressed as two theorems, whose proofs are in the appendix. These results give us an exact algorithm for the robust LIR method. The computational complexity of the algorithm is then studied in Subsection 3.3. We have implemented the algorithm as part of an R package, which is briefly introduced in Subsection 3.4, and applied to an illustrative example in Section 4. The final section is devoted to conclusions and directions for further research.

2. LIR in the case of simple linear regression with interval data

In the case of simple linear regression, the relation between two real-valued variables, X and Y, shall be described by means of a linear function. Hence, the set of all possible regression functions is $\mathcal{F} := \{f_{a,b} : a, b \in \mathbb{R}\}$, where the functions $f_{a,b} : \mathbb{R} \to \mathbb{R}$ are defined by $f_{a,b}(x) = a + b\,x$ for all $x \in \mathbb{R}$. We consider here the case of imprecisely observed quantities, and in particular of interval data. That is, instead of directly observing the realizations of the variables X and Y, we can only observe the realizations of the extended real-valued variables $\underline{X}$, $\overline{X}$, $\underline{Y}$, and $\overline{Y}$, which are the endpoints of the interval data $[\underline{X}, \overline{X}]$ and $[\underline{Y}, \overline{Y}]$. Throughout the paper, $[\underline{w}, \overline{w}]$ denotes the closed interval consisting of all real numbers w such that $\underline{w} \leq w \leq \overline{w}$. This notation is used for all $\underline{w}, \overline{w} \in \overline{\mathbb{R}}$, so that the interval $[\underline{w}, \overline{w}]$ is empty when $\underline{w} > \overline{w}$, and does not contain its endpoints when these are infinite.

2.1. The probability model

The only assumption about the joint distribution of the six random variables X, Y, $\underline{X}$, $\overline{X}$, $\underline{Y}$, and $\overline{Y}$ is the following:

$$P(\underline{X} \leq X \leq \overline{X} \text{ and } \underline{Y} \leq Y \leq \overline{Y}) \geq 1 - \varepsilon, \tag{1}$$

for some $\varepsilon \in [0, 1/2[$. That is, apart for the choice of ε, the probability model is fully nonparametric: it is only assumed that the (possibly unbounded) rectangle $[\underline{X}, \overline{X}] \times [\underline{Y}, \overline{Y}]$ contains the pair (X, Y) with probability at least $1 - \varepsilon$. In other words, an imprecise observation may not cover the precise data point with probability at most ε. The usual choice of ε is 0 (see for instance Heitjan and Rubin, 1991), but sometimes it can be useful to allow the imprecise data to be incorrect with a positive probability, and $\varepsilon \in]0, 1/2[$ is then an upper bound on this probability. Apart from this assumption, there is no restriction on the set of possible distributions of the precise and imprecise data. In particular, nothing is assumed about the joint distribution of the quantities of interest, X and Y.

The relation between X and Y shall be described by a linear function $f \in \mathcal{F}$. For each $f \in \mathcal{F}$, the quality of the description depends on the marginal distribution of the (absolute) residual

$$R_f := |Y - f(X)|\,.$$

The more this distribution is concentrated near 0, the better is the description of the relation between X and Y. In the robust LIR method that we consider in this paper, the concentration near 0 of the distribution of the residual R_f is evaluated by its median, or more generally by its p-quantile, with $p \in]\varepsilon, 1 - \varepsilon[$. The closer to 0 the p-quantile is, the better f describes the relation between X and Y. In particular, the best description of the relation of interest is a linear function for which the p-quantile of the residual's distribution is minimal.

Assuming for simplicity that the p-quantiles of the distribution of R_f are unique for all $f \in \mathcal{F}$, and that there is a unique $f_0 \in \mathcal{F}$ such that the corresponding p-quantile $q_0 \in \mathbb{R}_{\geq 0}$ is minimal, we can characterize geometrically the best description f_0 as follows. For each $f \in \mathcal{F}$ and each $q \in \mathbb{R}_{\geq 0}$, let

$$\overline{B}_{f,q} := \left\{(x, y) \in \mathbb{R}^2 : |y - f(x)| \leq q\right\}$$

be the closed band of (vertical) width $2\,q$ around the graph of f. Then $\overline{B}_{f_0,q_0}$ is the thinnest band of the form $\overline{B}_{f,q}$ containing (X, Y) with probability at least p. This is in particular the case when Y has for each $x \in \mathbb{R}$ a conditional distribution given $X = x$ that is strictly unimodal and symmetric around $f_0(x)$ (see also Tasche, 2003). That is, in the linear model $Y = a_0 + b_0\,X + E$, the best description in the above sense is $f_0 = f_{a_0,b_0}$, when the conditional distribution of the error term $E \,|\, X = x$ is strictly unimodal and symmetric (around 0) for all $x \in \mathbb{R}$ (e.g., when the error term E is independent of X and normally distributed with mean 0).

2.2. The LIR analysis

Let the nonempty (possibly unbounded) rectangles $[\underline{x}_1, \overline{x}_1] \times [\underline{y}_1, \overline{y}_1], \ldots, [\underline{x}_n, \overline{x}_n] \times [\underline{y}_n, \overline{y}_n] \subseteq \mathbb{R}^2$ be n independent realizations of the random set $[\underline{X}, \overline{X}] \times [\underline{Y}, \overline{Y}]$. The LIR analysis consists in using likelihood inference to identify a set of plausible regression functions. The imprecise data induce a (nonparametric) likelihood function on the set of all joint probability distributions (of X, Y, $\underline{X}$, $\overline{X}$, $\underline{Y}$, and $\overline{Y}$) satisfying condition (1). For each $f \in \mathcal{F}$, let C_f be the likelihood-based confidence region with cutoff point β for the p-quantile of the distribution of R_f, where $\beta \in [(\max\{p, 1 - p\} + \varepsilon)^n\,, 1[$. That is, C_f consists of all possible values of the p-quantile of the distribution of R_f, for all probability distributions whose likelihood exceeds β times the maximum of the likelihood function (see Cattaneo and Wiencierz, 2012, for more details).

In order to obtain an explicit formula for the confidence regions C_f, we define

$$\underline{k} := \max\left(\left\{k \in \{1, \ldots, n-1\} : k < (p - \varepsilon)\,n \text{ and } \left(\frac{p - \varepsilon}{k}\right)^k \left(\frac{1 - p + \varepsilon}{n - k}\right)^{n-k} \leq \frac{\beta}{n^n}\right\} \cup \{0\}\right),$$

$$\overline{k} := \min\left(\left\{k \in \{1, \ldots, n-1\} : k > (p + \varepsilon)\,n \text{ and } \left(\frac{p + \varepsilon}{k}\right)^k \left(\frac{1 - p - \varepsilon}{n - k}\right)^{n-k} \leq \frac{\beta}{n^n}\right\} \cup \{n\}\right).$$

Clearly, the two integers $\underline{k}$ and $\overline{k}$ depend on ε, p, n, and β, and satisfy

$$0 \leq \underline{k} < (p - \varepsilon)\,n \leq p\,n \leq (p + \varepsilon)\,n < \overline{k} \leq n.$$

Moreover, when ε, p, and n are fixed, $\underline{k}$ and $\overline{k}$ are an increasing and a decreasing function of β, respectively, and in particular, if β is sufficiently large, then $\underline{k} = \lceil (p - \varepsilon)\,n \rceil - 1$ (i.e., the largest integer smaller than $(p - \varepsilon)\,n$) and $\overline{k} = \lfloor (p + \varepsilon)\,n \rfloor + 1$ (i.e., the smallest integer larger than $(p + \varepsilon)\,n$).

Now, for each function $f \in \mathcal{F}$ and each imprecise observation $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$, we define the lower and upper (absolute) residuals

$$\underline{r}_{f,i} := \min_{(x,y)\in[\underline{x}_i,\overline{x}_i]\times[\underline{y}_i,\overline{y}_i]} |y - f(x)|\,,$$
$$\overline{r}_{f,i} := \sup_{(x,y)\in[\underline{x}_i,\overline{x}_i]\times[\underline{y}_i,\overline{y}_i]} |y - f(x)|\,.$$

Obviously, $\underline{r}_{f,i} \leq \overline{r}_{f,i}$, and $\underline{r}_{f,i} \in \mathbb{R}_{\geq 0}$, while $\overline{r}_{f,i} \in \overline{\mathbb{R}}_{\geq 0}$. In particular, $\overline{r}_{f,i} = +\infty$ if and only if either the linear function f is not constant and the rectangle $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ is unbounded, or f is constant and the interval $[\underline{y}_i, \overline{y}_i]$ is unbounded.

As usual in statistics, $\underline{r}_{f,(i)}$ and $\overline{r}_{f,(i)}$ denote then the ith smallest lower and upper residuals, respectively. That is, $\underline{r}_{f,(1)} \leq \cdots \leq \underline{r}_{f,(n)}$ are the ordered lower residuals and $\overline{r}_{f,(1)} \leq \cdots \leq \overline{r}_{f,(n)}$ are the ordered upper residuals. Then Corollary 2 of Cattaneo and Wiencierz (2012) implies that

$$C_f = [\underline{r}_{f,(\underline{k}+1)}, \overline{r}_{f,(\overline{k})}]$$

for all $f \in \mathcal{F}$. That is, the likelihood-based confidence region $C_f \subseteq \mathbb{R}_{\geq 0}$ is a nonempty closed interval, which is bounded if and only if either f is not constant and there are at least $\overline{k}$ bounded imprecise observations, or f is constant and there are at least $\overline{k}$ imprecise observations $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ such that the interval $[\underline{y}_i, \overline{y}_i]$ is bounded.

It is important to note that in general the interval C_f is proper (i.e., it contains more than one value), even when β is so large that $\underline{k} = \lceil (p - \varepsilon)\, n \rceil - 1$ and $\overline{k} = \lfloor (p + \varepsilon)\, n \rfloor + 1$. In this case, C_f represents the maximum likelihood estimate of the p-quantile of the distribution of R_f, which in general is not a single value because the data are imprecise and quantiles of a distribution are not necessarily unique. For example, if n is even, $\varepsilon = 0$, and $p = {}^1\!/_2$, then $\lceil (p - \varepsilon)\, n \rceil = \lfloor (p + \varepsilon)\, n \rfloor = {}^n\!/_2$, and thus the maximum likelihood estimate of the p-quantile (i.e., the median) of the distribution of R_f is $[\underline{r}_{f,(n/2)}, \overline{r}_{f,(n/2+1)}]$.

Hence, for each linear function $f \in \mathcal{F}$, we have an interval estimate C_f for the p-quantile of the distribution of the (absolute) residual R_f. We would like to select the regression function $f \in \mathcal{F}$ by minimizing this estimate, but comparing the intervals C_f gives us only a partial order on $\mathcal{F}$. The linear functions $f \in \mathcal{F}$ that are minimal according to this partial order are said to be undominated. That is, f is undominated if and only if there is no $f' \in \mathcal{F}$ such that $\overline{r}_{f',(\overline{k})} < \underline{r}_{f,(\underline{k}+1)}$. In order to simplify the description of the undominated functions, define

$$\overline{q}_{LRM} := \inf_{f\in\mathcal{F}} \overline{r}_{f,(\overline{k})}$$

(the name $\overline{q}_{LRM}$ shall be clarified in Subsection 3.1). The set of all undominated regression functions

$$\mathcal{U} := \{f \in \mathcal{F} : \underline{r}_{f,(\underline{k}+1)} \leq \overline{q}_{LRM}\}$$

is the result of the robust LIR method considered in this paper. It represents the whole uncertainty about the linear function that best describes the relation between X and Y, including the statistical uncertainty due to the finite sample as well as the indetermination related to the fact that the quantities are only imprecisely observed.

3. An exact algorithm for LIR

We now present an exact algorithm for determining the result of the robust LIR analysis described in Section 2. That is, an exact algorithm for calculating the set $\mathcal{U}$ of all undominated regression functions, given n nonempty (possibly unbounded) rectangles $[\underline{x}_1, \overline{x}_1] \times [\underline{y}_1, \overline{y}_1], \ldots, [\underline{x}_n, \overline{x}_n] \times [\underline{y}_n, \overline{y}_n] \subseteq \mathbb{R}^2$ and the two integers $\underline{k}$ and $\overline{k}$ with $0 \leq \underline{k} < \overline{k} \leq n$. The algorithm consists of two parts: in the first one, we determine the bound $\overline{q}_{LRM}$, which is then used in the second part to identify the set $\mathcal{U}$. As we will see, the computational complexity of the algorithm is $O(n^3 \log n)$. We have implemented this exact algorithm as part of an R package, which we will briefly introduce at the end of the present section.

3.1. Part 1: Determining the bound $\overline{q}_{LRM}$

Let $\mathcal{D}$ be the set of all $i \in \{1, \ldots, n\}$ such that the rectangle $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ is bounded. Then define $\mathcal{B} := \{0\}$ if there are less than $\overline{k}$ bounded imprecise observations (i.e., if $|\mathcal{D}| < \overline{k}$, where $|\mathcal{D}|$ denotes the cardinality of the set $\mathcal{D}$), and

$$\begin{aligned}\mathcal{B} := &\left\{ \frac{\overline{y}_i - \overline{y}_j}{\underline{x}_i - \underline{x}_j} : (i, j) \in \mathcal{D}^2 \text{ and } \underline{x}_i > \underline{x}_j \text{ and } \overline{y}_i > \overline{y}_j \right\} \cup \left\{ \frac{\underline{y}_i - \underline{y}_j}{\underline{x}_i - \underline{x}_j} : (i, j) \in \mathcal{D}^2 \text{ and } \underline{x}_i > \underline{x}_j \text{ and } \underline{y}_i < \underline{y}_j \right\} \\ &\cup \left\{ \frac{\overline{y}_i - \overline{y}_j}{\overline{x}_i - \overline{x}_j} : (i, j) \in \mathcal{D}^2 \text{ and } \overline{x}_i > \overline{x}_j \text{ and } \overline{y}_i < \overline{y}_j \right\} \cup \left\{ \frac{\underline{y}_i - \underline{y}_j}{\overline{x}_i - \overline{x}_j} : (i, j) \in \mathcal{D}^2 \text{ and } \overline{x}_i > \overline{x}_j \text{ and } \underline{y}_i > \underline{y}_j \right\} \cup \{0\}\end{aligned}$$

otherwise (i.e., if $|\mathcal{D}| \geq \overline{k}$). The central ideas of the first part of the algorithm are that in order to obtain $\overline{q}_{LRM}$ it suffices to consider the linear functions $f_{a,b}$ with slope $b \in \mathcal{B}$, and that for each slope b the intercept $a \in \mathbb{R}$ minimizing $\overline{r}_{f_{a,b},(\overline{k})}$ can be easily calculated, since the problem becomes one-dimensional. These ideas are formalized in the following theorem, but first we need some additional definitions. For each $b \in \mathbb{R}$ and each $i \in \{1, \ldots, n\}$, define

$$\underline{z}_{b,i} = \begin{cases} \underline{y}_i - b\,\underline{x}_i & \text{if } b < 0, \\ \underline{y}_i & \text{if } b = 0, \\ \underline{y}_i - b\,\overline{x}_i & \text{if } b > 0, \end{cases}$$

$$\overline{z}_{b,i} = \begin{cases} \overline{y}_i - b\,\overline{x}_i & \text{if } b < 0, \\ \overline{y}_i & \text{if } b = 0, \\ \overline{y}_i - b\,\underline{x}_i & \text{if } b > 0. \end{cases}$$

For each $b \in \mathbb{R}$ and each $j \in \{1, \ldots, n\}$, as usual, $\underline{z}_{b,(j)}$ and $\overline{z}_{b,(j)}$ denote then the jth smallest value among the $\underline{z}_{b,i}$ and among the $\overline{z}_{b,i}$, respectively. Furthermore, for each $b \in \mathbb{R}$ and each $j \in \{1, \ldots, n - \overline{k} + 1\}$, let $\overline{z}_{b,[j]}$ denote the $\overline{k}$th smallest value among the $\overline{z}_{b,i}$ such that $\underline{z}_{b,i} \geq \underline{z}_{b,(j)}$.

Theorem 1. *If there are less than $\overline{k}$ imprecise observations $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ such that the interval $[\underline{y}_i, \overline{y}_i]$ is bounded, then*

$$\begin{aligned} \overline{q}_{LRM} &= +\infty, \\ \{f \in \mathcal{F} : \overline{r}_{f,(\overline{k})} = \overline{q}_{LRM}\} &= \mathcal{F}. \end{aligned}$$

Otherwise (i.e., when there are at least $\overline{k}$ imprecise observations $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ such that the interval $[\underline{y}_i, \overline{y}_i]$ is bounded),

$$\overline{q}_{LRM} = \tfrac{1}{2} \min_{(b,j) \in \mathcal{B} \times \{1,\ldots,n-\overline{k}+1\}} (\overline{z}_{b,[j]} - \underline{z}_{b,(j)}),$$

$$\{f \in \mathcal{F} : \overline{r}_{f,(\overline{k})} = \overline{q}_{LRM}\} \supseteq \left\{ f_{a',b'} : (b', j') \in \operatorname*{arg\,min}_{(b,j) \in \mathcal{B} \times \{1,\ldots,n-\overline{k}+1\}} (\overline{z}_{b,[j]} - \underline{z}_{b,(j)}) \textit{ and } a' = \tfrac{1}{2} (\underline{z}_{b',(j')} + \overline{z}_{b',[j']}) \right\},$$

where the set on the left-hand side is infinite when the inclusion is strict. However, the inclusion is an equality when the following condition is satisfied: if there is a pair $(i, j) \in \mathcal{D}^2$ such that $\underline{x}_i = \overline{x}_j$ and $\max\{\overline{y}_i, \overline{y}_j\} - \min\{\underline{y}_i, \underline{y}_j\} = 2\,\overline{q}_{LRM}$, then $i \neq j$ and the two intervals $[\underline{y}_i, \overline{y}_i]$ and $[\underline{y}_j, \overline{y}_j]$ are nested (i.e., either $[\underline{y}_i, \overline{y}_i] \subseteq [\underline{y}_j, \overline{y}_j]$, or $[\underline{y}_j, \overline{y}_j] \subseteq [\underline{y}_i, \overline{y}_i]$).

For each linear function $f \in \mathcal{F}$, we have a likelihood-based confidence region $[\underline{r}_{f,(\underline{k}+1)}, \overline{r}_{f,(\overline{k})}]$ for the p-quantile of the residual's distribution. Hence, the functions $f \in \mathcal{F}$ minimizing $\overline{r}_{f,(\overline{k})}$ can be interpreted as the results of a minimax approach to our regression problem: they are called Likelihood-based Region Minimax (LRM) regression functions (see Cattaneo, 2007). For these functions, the upper endpoint of the interval estimate of the p-quantile of the residual's distribution is $\overline{q}_{LRM}$, which explains its name.

Theorem 1 implies in particular that an LRM regression function always exists, though it is not necessarily unique. When it is unique, it is denoted by f_{LRM}. In this case, $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ is the thinnest band of the form $\overline{B}_{f,q}$ containing at least

$\overline{k}$ imprecise observations, for all $f \in \mathcal{F}$ and all $q \in \mathbb{R}_{\geq 0}$. More generally, if there are at least $\overline{k}$ imprecise observations $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ such that the interval $[\underline{y}_i, \overline{y}_i]$ is bounded, then $2\,\overline{q}_{LRM}$ is the (vertical) width of the thinnest bands of the form $\overline{B}_{f,q}$ containing at least $\overline{k}$ imprecise observations (there can be more than one such bands, but only finitely many when the condition at the end of Theorem 1 is satisfied).

If all interval data are degenerate: $\underline{x}_i = \overline{x}_i$ and $\underline{y}_i = \overline{y}_i$ for all $i \in \{1, \dots, n\}$ (i.e., the imprecise data are in fact precise), then the LRM regression functions correspond to the least quantile of squares (or absolute residuals) regression functions $f \in \mathcal{F}$ minimizing the (square of the) $\overline{k}$th smallest absolute residual $\underline{r}_{f,(\overline{k})} = \overline{r}_{f,(\overline{k})}$ (see Rousseeuw and Leroy, 1987). That is, the LRM regression functions can be interpreted as the results of a generalization of the least quantile of squares regression to the case of imprecise data. The first part of our algorithm corresponds to a generalization (to the case of general quantiles and imprecise data) of the first exact algorithm for least median of squares regression, proposed by Steele and Steiger (1986) (see also Rousseeuw and Leroy, 1987, Chapter 5).

The key result behind Theorem 1 is that (when the condition at the end of the theorem is satisfied and there are at least $\overline{k}$ imprecise observations $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ such that the interval $[\underline{y}_i, \overline{y}_i]$ is bounded) if $\overline{B}_{f',q'}$ is one of the thinnest bands of the form $\overline{B}_{f,q}$ containing at least $\overline{k}$ imprecise observations, then the union of these imprecise observations touches one of the two borders of $\overline{B}_{f',q'}$ in at least two different points. This is a simple consequence of general results by Cheney (1982, Chapters 1 and 2), as suggested by Stromberg (1993). From this property it follows that one of the two borders of $\overline{B}_{f',q'}$ (which obviously have the same slope as f') is the line determined by two points on the borders of the imprecise observations contained in $\overline{B}_{f',q'}$. Hence, either the slope of f' is 0, or it is determined by two vertices of a pair of bounded imprecise observations contained in $\overline{B}_{f',q'}$. The set $\mathcal{B}$ consists of all the possible slopes that can be obtained in this way: they are at most $4\binom{n}{2} + 1$. For each possible slope $b \in \mathcal{B}$, finding the thinnest bands of the form $\overline{B}_{f_{a,b},q}$ containing at least $\overline{k}$ imprecise observations (for all $a \in \mathbb{R}$ and all $q \in \mathbb{R}_{\geq 0}$) corresponds to finding the shortest intervals (of the form $[a - q,\ a + q]$) containing at least $\overline{k}$ of the n intervals $[\underline{z}_{b,1}, \overline{z}_{b,1}], \dots, [\underline{z}_{b,n}, \overline{z}_{b,n}]$. This is a finite problem: it suffices to consider the intervals $[\underline{z}_{b,(j)}, \overline{z}_{b,[j]}]$ with $j \in \{1, \dots, n - \overline{k} + 1\}$.

Therefore, Theorem 1 gives us an algorithm for determining the bound $\overline{q}_{LRM}$, by reducing the minimization of $\overline{r}_{f,(\overline{k})}$ on the infinite set $\mathcal{F}$ to a minimization problem on the finite set $\mathcal{B} \times \{1, \dots, n - \overline{k} + 1\}$. Moreover, when there are finitely many LRM regression functions, Theorem 1 gives us an algorithm for finding all of them. An explicit formula for the set of all LRM regression functions in the general case (i.e., also when the condition at the end of the theorem is not satisfied) can be easily obtained, but requires several case distinctions and goes beyond the scope of the present paper.

3.2. Part 2: Identifying the set $\mathcal{U}$

After having determined the bound $\overline{q}_{LRM}$, in the second part of the algorithm we identify the set $\mathcal{U}$ of all undominated regression functions (i.e., the result of the robust LIR analysis described in Section 2).

Theorem 2.

$$\mathcal{U} = \left\{ f_{a,b} : b \in \mathbb{R} \textit{ and } a \in \bigcup_{j=1}^{n-\underline{k}} [\underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM},\ \overline{z}_{b,(j)} + \overline{q}_{LRM}] \right\}.$$

A linear function $f \in \mathcal{F}$ is undominated if and only if $\underline{r}_{f,(\underline{k}+1)} \leq \overline{q}_{LRM}$. That is, if and only if the band $\overline{B}_{f,\overline{q}_{LRM}}$ intersects at least $\underline{k} + 1$ imprecise observations. For each possible slope $b \in \mathbb{R}$, finding all the bands of the form $\overline{B}_{f_{a,b},\overline{q}_{LRM}}$ intersecting at least $\underline{k} + 1$ imprecise observations (for all $a \in \mathbb{R}$) corresponds to finding all the intervals of the form $[a - \overline{q}_{LRM},\ a + \overline{q}_{LRM}]$ intersecting at least $\underline{k} + 1$ of the n intervals $[\underline{z}_{b,1}, \overline{z}_{b,1}], \dots, [\underline{z}_{b,n}, \overline{z}_{b,n}]$. For each $b \in \mathbb{R}$ and each nonempty set $\mathcal{I} \subseteq \{1, \dots, n\}$, the interval $[a - \overline{q}_{LRM},\ a + \overline{q}_{LRM}]$ (with $a \in \mathbb{R}$) intersects all the intervals $[\underline{z}_{b,i}, \overline{z}_{b,i}]$ with $i \in \mathcal{I}$ if and only if $a \in [\max_{i \in \mathcal{I}} \underline{z}_{b,i} - \overline{q}_{LRM},\ \min_{i \in \mathcal{I}} \overline{z}_{b,i} + \overline{q}_{LRM}]$. Therefore,

$$\mathcal{U} = \left\{ f_{a,b} : b \in \mathbb{R} \text{ and } a \in \bigcup_{\mathcal{I} \subseteq \{1,\dots,n\} : |\mathcal{I}| = \underline{k}+1} \left[\max_{i \in \mathcal{I}} \underline{z}_{b,i} - \overline{q}_{LRM},\ \min_{i \in \mathcal{I}} \overline{z}_{b,i} + \overline{q}_{LRM} \right] \right\}.$$

Theorem 2 gives a simpler expression for $\mathcal{U}$, in which the number of intervals in the union is reduced from $\binom{n}{\underline{k}+1}$ to $n - \underline{k}$.

Hence, Theorem 2 gives us an algorithm for identifying, for each possible slope $b \in \mathbb{R}$, the set of all intercepts $a \in \mathbb{R}$ such that the linear function $f_{a,b}$ is undominated. This suffices for most practical purposes, but Theorem 2 also enables us to precisely describe as union of finitely many (possibly unbounded) polygons the set

$$\mathcal{U}' := \left\{(a, b) \in \mathbb{R}^2 : f_{a,b} \in \mathcal{U}\right\}$$

of all the intercept-slope pairs corresponding to the undominated regression functions. More precisely, $\mathcal{U}'$ is a subset of the plane $\mathbb{R}^2$ bounded by finitely many line segments and half-lines. However, $\mathcal{U}'$ is not necessarily convex nor connected, and if there are imprecise observations $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ such that the interval $[\underline{x}_i, \overline{x}_i]$ is unbounded and $[\underline{y}_i, \overline{y}_i] \neq \mathbb{R}$, then $\mathcal{U}'$ is not even necessarily closed.

Consider first the case with no imprecise observations $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ such that the interval $[\underline{x}_i, \overline{x}_i]$ is unbounded and $[\underline{y}_i, \overline{y}_i] \neq \mathbb{R}$. In this case, for each $i \in \{1, \dots, n\}$, the function $b \mapsto \underline{z}_{b,i}$ on $\mathbb{R}$ is either continuous and piecewise linear, or constant equal $-\infty$, while the function $b \mapsto \overline{z}_{b,i}$ on $\mathbb{R}$ is either continuous and piecewise linear, or constant equal $+\infty$. Therefore, for each $j \in \{1, \dots, n - \underline{k}\}$, the function $b \mapsto \underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM}$ on $\mathbb{R}$ is either continuous and piecewise linear, or constant equal $-\infty$, while the function $b \mapsto \overline{z}_{b,(j)} + \overline{q}_{LRM}$ on $\mathbb{R}$ is either continuous and piecewise linear, or constant equal $+\infty$. Thus, Theorem 2 implies that $\mathcal{U}'$ is a closed subset of the plane $\mathbb{R}^2$ bounded by finitely many line segments and half-lines. That is, $\mathcal{U}'$ is the union of finitely many (possibly unbounded) polygons (see for example Alexandrov, 2005, Subsection 1.1.1).

If $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ is an imprecise observation such that the interval $[\underline{x}_i, \overline{x}_i]$ is unbounded and $[\underline{y}_i, \overline{y}_i] \neq \mathbb{R}$, then at least one of the two functions $b \mapsto \underline{z}_{b,i}$ and $b \mapsto \overline{z}_{b,i}$ on $\mathbb{R}$ has a discontinuity at $b = 0$. Therefore, in this case, the functions $b \mapsto \underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,(j)} + \overline{q}_{LRM}$ on $\mathbb{R}$ (with $j \in \{1, \dots, n - \underline{k}\}$) can be discontinuous at $b = 0$. As a consequence, Theorem 2 implies that $\mathcal{U}'$ is a subset of the plane $\mathbb{R}^2$ bounded by finitely many line segments and half-lines, but $\mathcal{U}'$ is not necessarily closed. However, the two parts $\mathcal{U}' \cap (\mathbb{R} \times \{0\})$ and $\mathcal{U}' \cap (\mathbb{R} \times \mathbb{R}_{\neq 0})$ are relatively closed in $\mathbb{R} \times \{0\}$ and $\mathbb{R} \times \mathbb{R}_{\neq 0}$, respectively.

3.3. Computational complexity

The algorithm consisting of the two parts presented in Subsections 3.1 and 3.2 is the first exact algorithm to determine the result of the robust LIR analysis in the case of simple linear regression with interval data. It has worst-case time complexity $O(n^3 \log n)$, exactly as the first exact algorithm for least median of squares regression (see Steele and Steiger, 1986).

In the first part of the algorithm, described in Subsection 3.1, for each possible slope $b \in \mathcal{B}$, we must determine the pair $(\underline{z}_{b,(j)}, \overline{z}_{b,[j]})$ (with $j \in \{1, \dots, n - \overline{k} + 1\}$) such that the difference $\overline{z}_{b,[j]} - \underline{z}_{b,(j)}$ is minimal. We can do this as follows: after having calculated the values $\underline{z}_{b,1}, \dots, \underline{z}_{b,n}$ and $\overline{z}_{b,1}, \dots, \overline{z}_{b,n}$, we sort the two lists, obtaining $\underline{z}_{b,i_1}, \dots, \underline{z}_{b,i_n}$ (with $\underline{z}_{b,i_j} = \underline{z}_{b,(j)}$) and $\overline{z}_{b,(1)}, \dots, \overline{z}_{b,(n)}$. Then, for each j from 1 to $n - \overline{k} + 1$, we retrieve the pair consisting of the jth entry (i.e., $\underline{z}_{b,i_j}$) in the first list and of the $\overline{k}$th entry in the second one, and after that we remove the value $\overline{z}_{b,i_j}$ from the second list. In this way, the pairs of values that we have retrieved include all the pairs $(\underline{z}_{b,(j)}, \overline{z}_{b,[j]})$ with $j \in \{1, \dots, n - \overline{k} + 1\}$ (and possibly some irrelevant additional pairs with larger differences, if some of the $\underline{z}_{b,i}$ are equal), and we did not have to calculate a new list of $\overline{z}_{b,i}$ for each j in order to determine $\overline{z}_{b,[j]}$.

Hence, for each possible slope, we have to calculate and sort two lists of length n, which can be done in time $O(n \log n)$, and then for each $j \in \{1, \dots, n - \overline{k} + 1\}$, we have to search and remove a value from the second list, which can be done in time $O(\log n)$ using balanced trees (see for example Knuth, 1998, Subsection 6.2.3). Therefore, since there are at most $4\binom{n}{2} + 1$ possible slopes, the worst-case time complexity of the first part of the algorithm is $O(n^3 \log n)$.

In the second part of the algorithm, described in Subsection 3.2, for a given slope $b \in \mathbb{R}$, we must determine the pairs $(\underline{z}_{b,(\underline{k}+j)}, \overline{z}_{b,(j)})$ for all $j \in \{1, \dots, n - \underline{k}\}$. This can be done in time $O(n \log n)$, since it suffices to calculate and sort the two lists $\underline{z}_{b,1}, \dots, \underline{z}_{b,n}$ and $\overline{z}_{b,1}, \dots, \overline{z}_{b,n}$, and then, for each j from 1 to $n - \underline{k}$, retrieve the pair consisting of the $(\underline{k} + j)$th entry in the first list and of the jth entry in the second one.

For example, if we want to graphically represent the set $\mathcal{U}'$ of all the intercept-slope pairs $(a, b) \in \mathbb{R}^2$ corresponding to the undominated regression functions $f_{a,b}$, then it suffices to consider a finite number of possible values for the slope b, resulting in a worst-case time complexity of $O(n \log n)$ for the second part of the algorithm. However, if the goal is to precisely describe the set $\mathcal{U}'$ as union of finitely many (possibly unbounded) polygons, then the (worst-case) number of values $b \in \mathbb{R}$ that must be considered depends on n. In this case, it suffices to consider all values $b \in \mathbb{R}$

such that some of the $2n$ graphs of the functions $b \mapsto \underline{z}_{b,i}$ and $b \mapsto \overline{z}_{b,i}$ cross each other, and five additional values for the slope b. More precisely, these additional values are 0, a positive and a negative value sufficiently near 0 (in order to clarify what happens in the limits $b \downarrow 0$ and $b \uparrow 0$), and finally a positive and a negative value sufficiently far from 0 (in order to clarify what happens in the limits $b \uparrow +\infty$ and $b \downarrow -\infty$). Therefore, the worst-case number of values $b \in \mathbb{R}$ that must be considered is $2\binom{2n}{2} + 5$, and so the worst-case time complexity of the second part of the algorithm is $O(n^3 \log n)$, when the goal is to precisely describe the set $\mathcal{U}'$ as union of finitely many (possibly unbounded) polygons.

Altogether, the worst-case time complexity of the whole algorithm for the robust LIR analysis is thus $O(n^3 \log n)$.

3.4. R package

We have implemented the presented algorithm in R (R Development Core Team, 2012) as part of a package called `linLIR` (Wiencierz, 2012). This R package is created to implement LIR methods for the case of linear regression with interval data. The available version of the `linLIR` package includes a function to create a particular data object for interval-valued observations (`idf.create`), the function `s.linlir` to perform the LIR analysis for two variables out of the data object, and some associated methods for the generic functions `print`, `summary`, and `plot`. Both parts of the algorithm are incorporated in the `s.linlir` function. The corresponding `plot` method provides tools to visualize the results including, e.g., the set $\mathcal{U}'$. Moreover, the R package contains two example data sets including the one analyzed in Section 4. The current version of the R package is not optimized for speed, yet it provides a ready-to-use first implementation of the robust LIR method for linear regression with interval data.

4. Example

In many practical settings data are only available with limited precision. Consider, for example, the situation where it shall be analyzed how particulate matter concentration in the air varies with surface temperature. To investigate this relation, data is collected. The temperature is measured by means of a thermometer at randomly selected points in time during a certain period. From the instructions manual of the thermometer it is known that the measurement accuracy is, e.g., ±0.05°C, which translates the measured values into small intervals of width 0.1°C. Furthermore, there is a nearby measuring station for air pollution where four times a day recent data about particulate matter concentration is published, each time referring to a period of six hours. Among the published data there are the minimum and the maximum concentration measured during the corresponding period. Thus, for each temperature measurement at a particular time, the available information is that the corresponding particulate matter concentration lies in the interval of values measured during the corresponding six-hour period. That is, also this variable is only imprecisely observed.

The data set to be analyzed in the described situation might be similar to the simulated data set shown in Figure 1. This data set consists of 514 interval-valued observations of two variables. The data of the independent variable each have the same amount of imprecision, because the indetermination stems from the measurement accuracy of the thermometer determining the width of the intervals. By contrast, the width of the interval-valued observations of the dependent variable is given by the range of measured values during a fixed period and therefore varies a lot.

In the remainder of this section, we use the simulated data set to illustrate the implementation of the robust LIR method in the `linLIR` package. We here assume that the data are correct in the sense that the observed rectangles contain the correct precise values with probability one, i.e., we assume $\varepsilon = 0$. If we had concerns about the data quality, e.g., if it were likely that there have been some mistakes in recording the data, a positive ε could be considered. This would lead to a more imprecise result of the LIR analysis, reflecting the fact that there is additional uncertainty in the data.

Before conducting the LIR analysis, we have to choose the quantile p to be considered and the cutoff point β. It can be proved that the LIR analysis yields the most robust results when the median of the distribution of the residuals is considered, as for the least quantile of squares regression (see also Rousseeuw and Leroy, 1987, Chapter 3). Therefore, we set $p = 0.5$. Furthermore, we choose $\beta = 0.5$ as cutoff point for the likelihood-based confidence regions C_f with $f \in \mathcal{F}$. The confidence regions C_f are asymptotically (conservative) confidence intervals of level $F_{\chi^2}(-2 \log \beta)$, where F_{χ^2} is the cumulative distribution function of the chi-square distribution with 1 degree of freedom. Thus, the choice of $\beta = 0.5$ implies an asymptotic lower bound for the confidence level of C_f of 76.1% (see Cattaneo and Wiencierz, 2012). The `s.linlir` function of the R package provides also the finite-sample level of the (conservative)

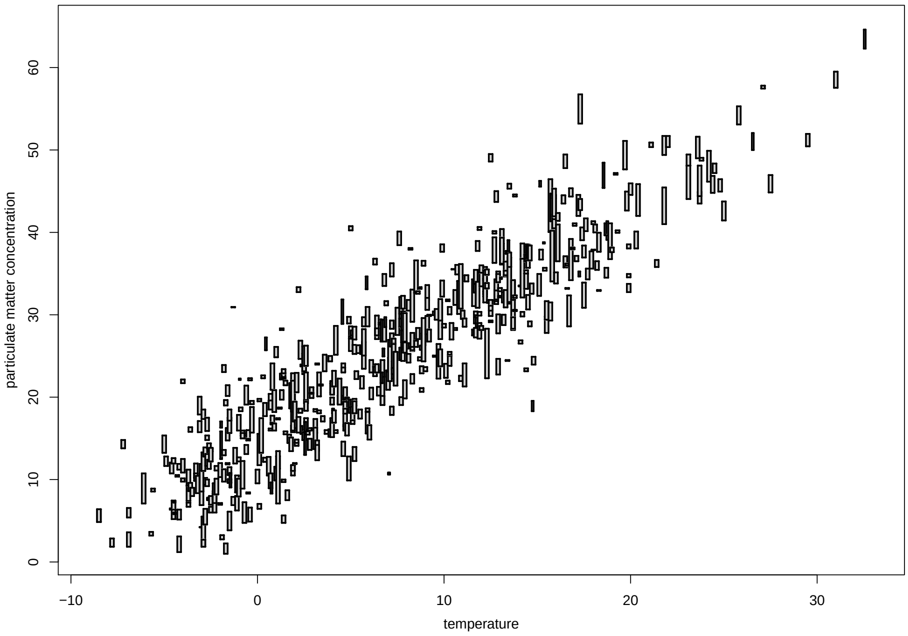

Figure 1: Simulated data set with 514 interval-valued observations of two variables, labeled as temperature and particulate matter concentration.

confidence intervals, which can be derived by simple combinatorial arguments. In the present situation the exact minimum confidence level is 78.28%. A thorough exhibition of the argumentation, however, would go beyond the scope of the present paper.

Using the function `s.linlir` of the R package `linLIR` with the above choices, we obtain the following results:

```
Estimated parameters of the function f.lrm:
intercept of f.lrm:  15.61906
slope of f.lrm:  1.293166

Ranges of parameter values of the undominated functions:
intercept of f in [6.537292,23.53457]
slope of f in [0.6552,2.156]

Number of observations: 514

LIR settings:
p: 0.5    beta: 0.5    epsilon: 0    k.l: 243    k.u: 271
confidence level of each confidence interval: 78.28 %
```

We find that there is a unique function that minimizes the right endpoint of the confidence interval for the median of the residuals' distribution, namely $f_{LRM}(x) = 15.62 + 1.29\,x$. This function corresponds to the line at the center of the thinnest closed band containing at least $\overline{k} = 271$ of the given imprecise observations. The set $\mathcal{U}$ of all undominated regression functions covers lines with intercepts between 6.54 and 23.53 and slopes ranging from 0.66 to 2.16. The closed bands $\overline{B}_{f,\overline{q}_{LRM}}$ of (vertical) width $2\,\overline{q}_{LRM} = 8.18$ around the lines $f \in \mathcal{U}$ each intersect at least $\underline{k} + 1 = 244$ imprecise observations. To visualize the results, we can use the `plot` method associated with the output of the function `s.linlir`. One can choose between plotting a random selection of functions out of the set $\mathcal{U}$, or the entire set $\mathcal{U}'$. Figure 2 shows 500 randomly selected undominated regression lines, which clearly indicate that there is a positive correlation between the investigated variables. The set of all intercept-slope combinations corresponding to the undominated regression lines is displayed in Figure 3, providing a nice illustration of the complex shape of this

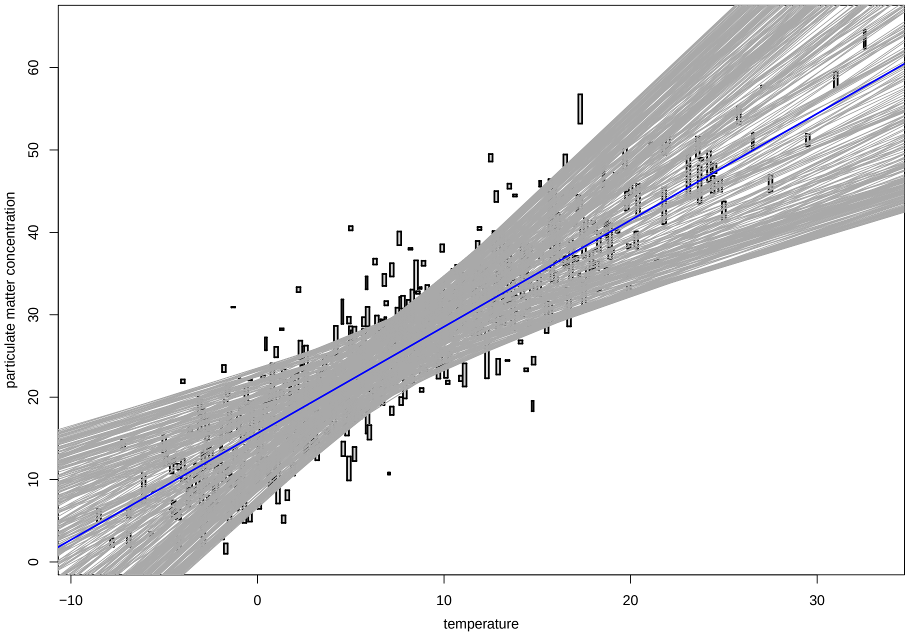

Figure 2: Selection of 500 undominated regression lines out of the set $\mathcal{U}$. The function f_{LRM} is highlighted.

set. In both cases, the line f_{LRM} or the corresponding intercept-slope combination (b_{LRM}, a_{LRM}) is highlighted.

As we already mentioned at the end of Section 3, the current version of the function `s.linlir` has not been optimized for speed yet. The computations for the present analysis took roughly 70 minutes on a desktop computer, most of the time is needed for the first part of the algorithm, where $\overline{q}_{LRM}$ is determined.

5. Conclusion

In this paper, we considered the LIR approach to regression for imprecisely observed quantities (see Cattaneo and Wiencierz, 2012, 2011). The result of a LIR analysis is in general set-valued: it consists of all regression functions that cannot be excluded on the basis of likelihood inference. These regression functions are said to be undominated. In this paper, we considered in particular the robust LIR method based on the residuals' quantiles, in the special case of simple linear regression with interval data. For this situation, we proved that the set of all the intercept-slope pairs corresponding to the undominated regression functions is the union of finitely many polygons, and we gave an exact algorithm for determining this set (i.e., for determining the set-valued result of the robust LIR method).

We have implemented this exact algorithm as part of the R package `linLIR` (Wiencierz, 2012). In the present paper, we illustrated the implementation of the robust LIR method in the `linLIR` package by means of an example. Furthermore, we showed that the algorithm has worst-case time complexity $O(n^3 \log n)$. In fact, the first part of the algorithm is related to the first exact algorithm for least median of squares regression, which has the same (asymptotic) worst-case time complexity (see Steele and Steiger, 1986; Rousseeuw and Leroy, 1987). This algorithm for least median of squares regression was then improved (see for example Souvaine and Steele, 1987; Edelsbrunner and Souvaine, 1990; Carrizosa and Plastria, 1995; Mount et al., 2007) and extended to multiple linear regression (see for instance Stromberg, 1993; Hawkins, 1993; Watson, 1998; Bernholt, 2005). In future work, we intend to do the same with the algorithm for the robust LIR method. Moreover, this algorithm can also be generalized to imprecise data other than intervals.

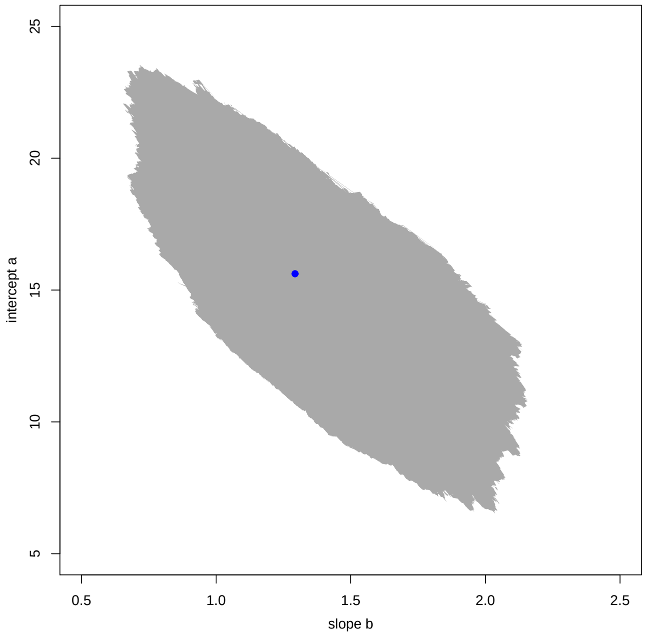

Figure 3: Set $\mathcal{U}'$ of all intercept-slope pairs corresponding to the undominated regression functions. The point (b_{LRM}, a_{LRM}) is highlighted.

Appendix A. Proofs

The following lemma gives us a method for writing the union of all $\binom{n}{k}$ possible intersections of k out of n intervals as the union of $n-k+1$ other intervals. It will be used in the proof of Theorem 2, but can be useful also for other problems, such as constructing an explicit formula for the set of all LRM regression functions in the general case (i.e., also when the condition at the end of Theorem 1 is not satisfied).

Lemma 1. *If $\underline{w}_1,\dots,\underline{w}_n,\overline{w}_1,\dots,\overline{w}_n \in \overline{\mathbb{R}}$ with $\underline{w}_i \leq \overline{w}_i$ for all $i \in \{1,\dots,n\}$, then for each $k \in \{1,\dots,n\}$,*

$$\bigcup_{\mathcal{I}\subseteq\{1,\dots,n\}\,:\,|\mathcal{I}|=k}\ \bigcap_{i\in\mathcal{I}}[\underline{w}_i,\overline{w}_i] = \bigcup_{j=k}^{n}[\underline{w}_{(j)},\overline{w}_{(j-k+1)}],$$

where for each $j \in \{1,\dots,n\}$, as usual, $\underline{w}_{(j)}$ and $\overline{w}_{(j)}$ denote the jth smallest value among $\underline{w}_1,\dots,\underline{w}_n$ and among $\overline{w}_1,\dots,\overline{w}_n$, respectively.

This lemma can be proved as follows. Assume without loss of generality that $\underline{w}_1 \leq \cdots \leq \underline{w}_n$ (i.e., $\underline{w}_{(j)} = \underline{w}_j$), and for all $j, j' \in \{1,\dots,n\}$ with $j \leq j'$, let $\overline{w}_{j:j'}$ denote the jth smallest value among $\overline{w}_1,\dots,\overline{w}_{j'}$ (hence, in particular, $\overline{w}_{(j)} = \overline{w}_{j:n}$). Then, for each set $\mathcal{I} \subseteq \{1,\dots,n\}$ with cardinality $|\mathcal{I}| = k$,

$$\bigcap_{i\in\mathcal{I}}[\underline{w}_i,\overline{w}_i] = \left[\max_{i\in\mathcal{I}}\underline{w}_i, \min_{i\in\mathcal{I}}\overline{w}_i\right] = \left[\underline{w}_{\max\mathcal{I}}, \min_{i\in\mathcal{I}}\overline{w}_i\right] \subseteq [\underline{w}_{\max\mathcal{I}}, \overline{w}_{\max\mathcal{I}-k+1:\max\mathcal{I}}],$$

and obviously $\max\mathcal{I} \in \{k,\dots,n\}$. Furthermore, for each $j \in \{k,\dots,n\}$, there are at most $j-k$ indices $i \in \{1,\dots,n\}$ such that $\overline{w}_i < \overline{w}_{(j-k+1)}$, and thus there is a set $\mathcal{I}_j \subseteq \{1,\dots,j\}$ with cardinality $|\mathcal{I}_j| = k$ such that $\overline{w}_i \geq \overline{w}_{(j-k+1)}$ for all $i \in \mathcal{I}_j$. Therefore,

$$\bigcup_{j=k}^{n}[\underline{w}_{(j)},\overline{w}_{(j-k+1)}] \subseteq \bigcup_{j=k}^{n}\left[\max_{i\in\mathcal{I}_j}\underline{w}_i, \min_{i\in\mathcal{I}_j}\overline{w}_i\right] = \bigcup_{j=k}^{n}\bigcap_{i\in\mathcal{I}_j}[\underline{w}_i,\overline{w}_i] \subseteq \bigcup_{\mathcal{I}\subseteq\{1,\dots,n\}\,:\,|\mathcal{I}|=k}\ \bigcap_{i\in\mathcal{I}}[\underline{w}_i,\overline{w}_i]$$

$$\subseteq \bigcup_{\mathcal{I}\subseteq\{1,\dots,n\}\,:\,|\mathcal{I}|=k}[\underline{w}_{\max\mathcal{I}}, \overline{w}_{\max\mathcal{I}-k+1:\max\mathcal{I}}] = \bigcup_{j=k}^{n}[\underline{w}_j, \overline{w}_{j-k+1:j}].$$

Hence, in order to complete the proof of the lemma, it suffices to show that the first and last unions of $n-k+1$ intervals in the above expression are equal. To this goal, we first show that for each $j \in \{k, \dots, n-1\}$,

$$[\underline{w}_j, \overline{w}_{j-k+1:j}] \cup [\underline{w}_{j+1}, \overline{w}_{j+1-k+1:j+1}] = [\underline{w}_j, \overline{w}_{(j-k+1)}] \cup [\underline{w}_{j+1}, \overline{w}_{j+1-k+1:j+1}]. \tag{A.1}$$

Since $\overline{w}_{(j-k+1)} \leq \overline{w}_{j-k+1:j}$ always holds, (A.1) could be wrong only if $\overline{w}_{(j-k+1)} < \overline{w}_{j-k+1:j}$, which can be the case only if there is an index $i \in \{j+1, \dots, n\}$ such that $\overline{w}_i \leq \overline{w}_{(j-k+1)}$, but then

$$\underline{w}_j \leq \underline{w}_{j+1} \leq \underline{w}_i \leq \overline{w}_i \leq \overline{w}_{(j-k+1)} < \overline{w}_{j-k+1:j} \leq \overline{w}_{j+1-k+1:j+1},$$

and thus both unions in (A.1) are equal to the interval $[\underline{w}_j, \overline{w}_{j+1-k+1:j+1}]$. Therefore, using (A.1) for each j from k to $n-1$, we obtain

$$\bigcup_{j=k}^{n} [\underline{w}_j, \overline{w}_{j-k+1:j}] = \left(\bigcup_{j=k}^{n-1} [\underline{w}_j, \overline{w}_{(j-k+1)}] \right) \cup [\underline{w}_n, \overline{w}_{n-k+1:n}] = \left(\bigcup_{j=k}^{n-1} [\underline{w}_{(j)}, \overline{w}_{(j-k+1)}] \right) \cup [\underline{w}_{(n)}, \overline{w}_{(n-k+1)}] = \bigcup_{j=k}^{n} [\underline{w}_{(j)}, \overline{w}_{(j-k+1)}].$$

Appendix A.1. Proof of Theorem 1

As noted in Subsection 2.2, for each linear function $f \in \mathcal{F}$, we have $\overline{r}_{f,(\overline{k})} < +\infty$ if and only if either f is not constant and there are at least $\overline{k}$ bounded imprecise observations, or f is constant and there are at least $\overline{k}$ imprecise observations $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ such that the interval $[\underline{y}_i, \overline{y}_i]$ is bounded. Therefore, if there are less than $\overline{k}$ imprecise observations $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ such that the interval $[\underline{y}_i, \overline{y}_i]$ is bounded, then $\overline{r}_{f,(\overline{k})} = +\infty$ for all $f \in \mathcal{F}$, which proves the first part of the theorem. Otherwise, if there are at least $\overline{k}$ imprecise observations $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ such that the interval $[\underline{y}_i, \overline{y}_i]$ is bounded, as we assume from now on, then $\overline{r}_{f,(\overline{k})} < +\infty$ at least for the constant functions $f \in \mathcal{F}$, which implies $\overline{q}_{LRM} < +\infty$.

For each function $f_{a,b} \in \mathcal{F}$ and each imprecise observation $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$,

$$\underline{z}_{b,i} = \inf_{(x,y) \in [\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]} (y - b\,x), \tag{A.2}$$

$$\overline{z}_{b,i} = \sup_{(x,y) \in [\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]} (y - b\,x), \tag{A.3}$$

and therefore

$$\overline{r}_{f_{a,b},i} = \max \left\{ \sup_{(x,y) \in [\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]} (y - a - b\,x), \quad \sup_{(x,y) \in [\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]} (a + b\,x - y) \right\} = \max\{\overline{z}_{b,i} - a,\ a - \underline{z}_{b,i}\}.$$

Hence, the $\overline{k}$th smallest upper residual of $f_{a,b}$ is

$$\overline{r}_{f_{a,b},(\overline{k})} = \min_{\mathcal{I} \subseteq \{1,\dots,n\} : |\mathcal{I}| = \overline{k}} \max_{i \in \mathcal{I}} \max\{\overline{z}_{b,i} - a,\ a - \underline{z}_{b,i}\} = \min_{\mathcal{I} \subseteq \{1,\dots,n\} : |\mathcal{I}| = \overline{k}} \max \left\{ \max_{i \in \mathcal{I}} \overline{z}_{b,i} - a,\ a - \min_{i \in \mathcal{I}} \underline{z}_{b,i} \right\}.$$

Now, for each set $\mathcal{I} \subseteq \{1, \dots, n\}$ with cardinality $|\mathcal{I}| = \overline{k}$, there is a $j \in \{1, \dots, n - \overline{k} + 1\}$ such that $\underline{z}_{b,(j)} = \min_{i \in \mathcal{I}} \underline{z}_{b,i}$, and in this case, since $\underline{z}_{b,i} \geq \underline{z}_{b,(j)}$ for all $i \in \mathcal{I}$, the smallest possible value of $\max_{i \in \mathcal{I}} \overline{z}_{b,i}$ is $\overline{z}_{b,[j]}$. Thus we obtain

$$\overline{r}_{f_{a,b},(\overline{k})} = \min_{j \in \{1,\dots,n-\overline{k}+1\}} \max\{\overline{z}_{b,[j]} - a,\ a - \underline{z}_{b,(j)}\}.$$

Clearly, for each $b \in \mathbb{R}$ and $j \in \{1, \dots, n - \overline{k} + 1\}$ such that the interval $[\underline{z}_{b,(j)}, \overline{z}_{b,[j]}]$ is bounded, the maximum of $\overline{z}_{b,[j]} - a$ and $a - \underline{z}_{b,(j)}$ is uniquely minimized by the interval center $a = 1/2\,(\underline{z}_{b,(j)} + \overline{z}_{b,[j]})$. This implies

$$\overline{q}_{LRM} = \inf_{(a,b) \in \mathbb{R}^2} \overline{r}_{f_{a,b},(\overline{k})} = \tfrac{1}{2} \inf_{(b,j) \in \mathbb{R} \times \{1,\dots,n-\overline{k}+1\}} (\overline{z}_{b,[j]} - \underline{z}_{b,(j)}),$$

$$\{f \in \mathcal{F} : \overline{r}_{f,(\overline{k})} = \overline{q}_{LRM}\} = \left\{ f_{a',b'} : (b', j') \in \operatorname*{arg\,min}_{(b,j) \in \mathbb{R} \times \{1,\dots,n-\overline{k}+1\}} (\overline{z}_{b,[j]} - \underline{z}_{b,(j)}) \text{ and } a' = \tfrac{1}{2}\,(\underline{z}_{b',(j')} + \overline{z}_{b',[j']}) \right\}.$$

Therefore, in order to complete the proof of the theorem, it suffices to show that the set

$$\mathcal{M} := \left\{ b' : (a',b') \in \operatorname*{arg\,min}_{(a,b)\in\mathbb{R}^2} \overline{r}_{f_{a,b},(\overline{k})} \right\} = \left\{ b' : (b',j') \in \operatorname*{arg\,min}_{(b,j)\in\mathbb{R}\times\{1,\dots,n-\overline{k}+1\}} (\overline{z}_{b,[j]} - \underline{z}_{b,(j)}) \right\}$$

intersects $\mathcal{B}$ (i.e., $\mathcal{M} \cap \mathcal{B} \neq \emptyset$), that $\mathcal{M}$ is infinite when $\mathcal{M} \nsubseteq \mathcal{B}$, and that $\mathcal{M} \subseteq \mathcal{B}$ when the condition at the end of the theorem is satisfied.

For each set $\mathcal{I} \subseteq \{1,\dots,n\}$ with cardinality $|\mathcal{I}| = \overline{k}$, let $g_{\mathcal{I}}$ be the function $(a,b) \mapsto \max_{i\in\mathcal{I}} \overline{r}_{f_{a,b},i}$ on $\mathbb{R}^2$. Then, for all $a,b \in \mathbb{R}$,

$$\overline{r}_{f_{a,b},(\overline{k})} = \min_{\mathcal{I}\subseteq\{1,\dots,n\}\,:\,|\mathcal{I}|=\overline{k}} g_{\mathcal{I}}(a,b).$$

Let $\mathcal{S}$ be the (nonempty) set of all sets $\mathcal{I} \subseteq \{1,\dots,n\}$ with cardinality $|\mathcal{I}| = \overline{k}$ such that $\inf_{(a,b)\in\mathbb{R}^2} g_{\mathcal{I}}(a,b) = \overline{q}_{LRM}$. Then, defining for each $\mathcal{I} \in \mathcal{S}$,

$$\mathcal{M}_{\mathcal{I}} := \left\{ b' : (a',b') \in \operatorname*{arg\,min}_{(a,b)\in\mathbb{R}^2} g_{\mathcal{I}}(a,b) \right\},$$

we obtain $\mathcal{M} = \bigcup_{\mathcal{I}\in\mathcal{S}} \mathcal{M}_{\mathcal{I}}$. Hence, in order to complete the proof of the theorem, it suffices to show for each $\mathcal{I} \in \mathcal{S}$, that the set $\mathcal{M}_{\mathcal{I}}$ intersects $\mathcal{B}$ (i.e., $\mathcal{M}_{\mathcal{I}} \cap \mathcal{B} \neq \emptyset$), that $\mathcal{M}_{\mathcal{I}}$ is infinite when $\mathcal{M}_{\mathcal{I}} \nsubseteq \mathcal{B}$, and that $\mathcal{M}_{\mathcal{I}} \subseteq \mathcal{B}$ when the condition at the end of the theorem is satisfied.

Let $\mathcal{I} \in \mathcal{S}$, and consider first the case with $\mathcal{I} \nsubseteq \mathcal{D}$. In this case, there is an $i \in \mathcal{I}$ such that the rectangle $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ is unbounded, and since $\overline{q}_{LRM} < +\infty$, there are $a,b \in \mathbb{R}$ such that $\overline{r}_{f_{a,b},i} < +\infty$. As noted in Subsection 2.2, this implies that the interval $[\underline{y}_i, \overline{y}_i]$ is unbounded, and then $\overline{r}_{f_{a,b},i} < +\infty$ if and only if the function $f_{a,b}$ is constant. That is, $g_{\mathcal{I}}(a,b) < +\infty$ if and only if $b = 0$, and therefore $\mathcal{M}_{\mathcal{I}} = \{0\} \subseteq \mathcal{B}$.

Consider now the case with $\mathcal{I} \subseteq \mathcal{D}$ (i.e., the rectangle $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ is bounded for all $i \in \mathcal{I}$), which implies in particular $|\mathcal{D}| \geq \overline{k}$. In this case,

$$g_{\mathcal{I}}(a,b) = \max_{i\in\mathcal{I}} \max_{(x,y)\in\{\underline{x}_i,\overline{x}_i\}\times\{\underline{y}_i,\overline{y}_i\}} |y - a - b\,x|$$

for all $a,b \in \mathbb{R}$, since for a bounded imprecise observation $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$, the upper residual $\overline{r}_{f_{a,b},i}$ is the maximum of the four residuals corresponding to the vertices of the rectangle $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$. The Existence Theorem of Cheney (1982, page 20) implies then that $\arg\min_{(a,b)\in\mathbb{R}^2} g_{\mathcal{I}}(a,b)$ is not empty (i.e., $\mathcal{M}_{\mathcal{I}} \neq \emptyset$). Let thus $(a',b') \in \arg\min_{(a,b)\in\mathbb{R}^2} g_{\mathcal{I}}(a,b)$ (hence, $b' \in \mathcal{M}_{\mathcal{I}}$). From the Characterization Theorem of Cheney (1982, page 35) it follows that there are $(x,y),(x',y') \in \bigcup_{i\in\mathcal{I}} \{\underline{x}_i, \overline{x}_i\} \times \{\underline{y}_i, \overline{y}_i\}$ such that either $x \neq x'$ and both points $(x,y),(x',y')$ lie on the graph of one of the two functions $f_{a'+\overline{q}_{LRM},b'}$ and $f_{a'-\overline{q}_{LRM},b'}$, or $x = x'$ and the point (x,y) lies on the graph of the function $f_{a'+\overline{q}_{LRM},b'}$, while the point (x',y') lies on the graph of the function $f_{a'-\overline{q}_{LRM},b'}$.

All the (bounded) rectangles $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ with $i \in \mathcal{I}$ are contained in the closed band $\overline{B}_{f_{a',b'},\overline{q}_{LRM}}$ of (vertical) width $2\,\overline{q}_{LRM}$ around the graph of the function $f_{a',b'}$, and the points $(x,y),(x',y')$ are vertices of these rectangles lying on the border of the band $\overline{B}_{f_{a',b'},\overline{q}_{LRM}}$. If $x \neq x'$, then (x,y) and (x',y') lie on the same border of $\overline{B}_{f_{a',b'},\overline{q}_{LRM}}$, and thus determine its slope

$$b' = \frac{y - y'}{x - x'}.$$

It can be easily checked that the set $\mathcal{B}$ contains all the slopes that can be obtained in this way by the vertices of the bounded imprecise observations $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$. Therefore, if $x \neq x'$, then $b' \in \mathcal{B}$.

Assume now that $b' \notin \mathcal{B}$. In order to complete the proof of the theorem, it suffices to show that in this case the set $\mathcal{M}_{\mathcal{I}}$ is infinite and intersects $\mathcal{B}$, and that the condition at the end of the theorem cannot be satisfied. The assumption $b' \notin \mathcal{B}$ implies $x = x'$. Hence, the points (x,y) and (x',y') are two vertices of two (bounded) rectangles $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ and $[\underline{x}_j, \overline{x}_j] \times [\underline{y}_j, \overline{y}_j]$ (with $i,j \in \mathcal{I}$), and lie on the upper and on the lower borders of the band $\overline{B}_{f_{a',b'},\overline{q}_{LRM}}$, respectively. If either $x \neq \overline{x}_i$ and $x' \neq \overline{x}_j$, or $x \neq \underline{x}_i$ and $x' \neq \underline{x}_j$, then the intervals $[\underline{x}_i, \overline{x}_i]$ and $[\underline{x}_j, \overline{x}_j]$ are proper (i.e., they contain more than one value) and extend on the same side of $x = x'$, but this would imply $b' = 0 \in \mathcal{B}$, because the two rectangles $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ and $[\underline{x}_j, \overline{x}_j] \times [\underline{y}_j, \overline{y}_j]$ must be contained in the band $\overline{B}_{f_{a',b'},\overline{q}_{LRM}}$. Therefore, $\underline{x}_i = \overline{x}_j$ or $\overline{x}_i = \underline{x}_j$, and $\max\{\overline{y}_i, \overline{y}_j\} - \min\{\underline{y}_i, \underline{y}_j\} = y - y' = 2\,\overline{q}_{LRM}$. That is, one of the two pairs $(i,j),(j,i) \in \mathcal{I}^2 \subseteq \mathcal{D}^2$ satisfies the premise of the condition at the end of the theorem. Now, if $[\underline{y}_i, \overline{y}_i] \subseteq [\underline{y}_j, \overline{y}_j]$, then the interval $[\underline{x}_j, \overline{x}_j]$ must be degenerate (i.e., $\underline{x}_j = \overline{x}_j$), because otherwise we would have $b' = 0 \in \mathcal{B}$, since the rectangle $[\underline{x}_j, \overline{x}_j] \times [\underline{y}_j, \overline{y}_j]$ must be contained in the band $\overline{B}_{f_{a',b'},\overline{q}_{LRM}}$. Analogously, if $[\underline{y}_j, \overline{y}_j] \subseteq [\underline{y}_i, \overline{y}_i]$, then $\underline{x}_i = \overline{x}_i$. Hence, if the two intervals $[\underline{y}_i, \overline{y}_i]$ and $[\underline{y}_j, \overline{y}_j]$ are nested, then one of the two pairs $(i, i), (j, j) \in \mathcal{D}^2$ satisfies the premise of the condition at the end of the theorem. So this condition is contradicted by at least one of the four pairs $(i, j), (j, i), (i, i), (j, j) \in \mathcal{D}^2$.

In order to complete the proof of the theorem, it remains to show that the set $\mathcal{M}_{\mathcal{I}}$ is infinite and intersects $\mathcal{B}$. We have that $b \in \mathcal{M}_{\mathcal{I}}$ if and only if there is an $a \in \mathbb{R}$ such that the closed band $\overline{B}_{f_{a,b},\overline{q}_{LRM}}$ of (vertical) width $2\,\overline{q}_{LRM}$ around the graph of the function $f_{a,b}$ contains the $4\,\overline{k}$ vertices of the rectangles $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ with $i \in \mathcal{I}$. For each $b \in \mathbb{R}$, since the two vertices $(x, y), (x', y')$ satisfy $x = x'$ and $y - y' = 2\,\overline{q}_{LRM}$, the band $\overline{B}_{f_{a,b},\overline{q}_{LRM}}$ can contain the $4\,\overline{k}$ vertices only if $a = a_b := {}^1\!/\!{}_2\,(y' + y) - b\,x$ (i.e., only if the midpoint of (x, y) and (x', y') is contained in the graph of the linear function $f_{a,b}$). Now, for each vertex (x'', y''), the set of all $b \in \mathbb{R}$ such that the band $\overline{B}_{f_{a_b,b},\overline{q}_{LRM}}$ contains (x'', y'') is the closed interval

$$\mathcal{B}_{x'',y''} = \begin{cases} \left[\dfrac{y' - y''}{x' - x''}, \dfrac{y - y''}{x - x''}\right] & \text{if } x'' < x = x', \\ \mathbb{R} & \text{if } x'' = x = x', \\ \left[\dfrac{y'' - y}{x'' - x}, \dfrac{y'' - y'}{x'' - x'}\right] & \text{if } x'' > x = x', \end{cases}$$

where the second case is implied by the fact that $\mathcal{B}_{x'',y''}$ is not empty (since $b' \in \mathcal{M}_{\mathcal{I}} \subseteq \mathcal{B}_{x'',y''}$), while in the other two cases the endpoints of $\mathcal{B}_{x'',y''}$ are the slopes b determined by the pairs of points $(x, y), (x'', y'')$ or $(x', y'), (x'', y'')$ lying on the same border of $\overline{B}_{f_{a_b,b},\overline{q}_{LRM}}$. Therefore,

$$\mathcal{M}_{\mathcal{I}} = \bigcap_{i \in \mathcal{I}} \; \bigcap_{(x'',y'') \in \{\underline{x}_i, \overline{x}_i\} \times \{\underline{y}_i, \overline{y}_i\}} \mathcal{B}_{x'',y''}$$

is a (nonempty) closed interval, which is either $\mathbb{R}$ or it is bounded. When $\mathcal{M}_{\mathcal{I}} = \mathbb{R}$, obviously it is infinite and intersects $\mathcal{B}$. Otherwise, $\mathcal{M}_{\mathcal{I}}$ is a bounded interval whose endpoints are elements of $\mathcal{B}$, since they are slopes b determined by a pair of vertices lying on the same border of $\overline{B}_{f_{a_b,b},\overline{q}_{LRM}}$. Hence, also in this case $\mathcal{M}_{\mathcal{I}}$ intersects $\mathcal{B}$ and is infinite, since $b' \notin \mathcal{B}$ is an interior point of the interval $\mathcal{M}_{\mathcal{I}}$.

Appendix A.2. Proof of Theorem 2

For each function $f_{a,b} \in \mathcal{F}$ and each imprecise observation $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$, using (A.2) and (A.3), we obtain that $\underline{r}_{f_{a,b},i} \leq \overline{q}_{LRM}$ if and only if the set

$$\{y - f_{a,b}(x) : (x, y) \in [\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]\} = [\underline{z}_{b,i} - a,\, \overline{z}_{b,i} - a]$$

intersects the interval $[-\overline{q}_{LRM}, \overline{q}_{LRM}]$. That is, $\underline{r}_{f_{a,b},i} \leq \overline{q}_{LRM}$ if and only if $a \in [\underline{z}_{b,i} - \overline{q}_{LRM},\, \overline{z}_{b,i} + \overline{q}_{LRM}]$. Hence, $\underline{r}_{f_{a,b},(\underline{k}+1)} \leq \overline{q}_{LRM}$ if and only if there is a set $\mathcal{I} \subseteq \{1, \ldots, n\}$ such that $|\mathcal{I}| = \underline{k} + 1$ and $a \in [\underline{z}_{b,i} - \overline{q}_{LRM},\, \overline{z}_{b,i} + \overline{q}_{LRM}]$ for all $i \in \mathcal{I}$. That is, using Lemma 1 with $k = \underline{k} + 1$, we obtain that $\underline{r}_{f_{a,b},(\underline{k}+1)} \leq \overline{q}_{LRM}$ if and only if a lies in the set

$$\bigcup_{\mathcal{I} \subseteq \{1,\ldots,n\} : |\mathcal{I}| = \underline{k}+1} \; \bigcap_{i \in \mathcal{I}} [\underline{z}_{b,i} - \overline{q}_{LRM},\, \overline{z}_{b,i} + \overline{q}_{LRM}] = \bigcup_{j=\underline{k}+1}^{n} [\underline{z}_{b,(j)} - \overline{q}_{LRM},\, \overline{z}_{b,(j-\underline{k})} + \overline{q}_{LRM}] = \bigcup_{j=1}^{n-\underline{k}} [\underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM},\, \overline{z}_{b,(j)} + \overline{q}_{LRM}].$$

Therefore,

$$\mathcal{U} = \{f_{a,b} \in \mathcal{F} : \underline{r}_{f_{a,b},(\underline{k}+1)} \leq \overline{q}_{LRM}\} = \left\{ f_{a,b} : b \in \mathbb{R} \text{ and } a \in \bigcup_{j=1}^{n-\underline{k}} [\underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM},\, \overline{z}_{b,(j)} + \overline{q}_{LRM}] \right\}.$$

References

Alexandrov, A.D., 2005. Convex Polyhedra. Springer.

Beaton, A.E., Rubin, D.B., Barone, J.L., 1976. The acceptability of regression solutions: Another look at computational accuracy. J. Am. Stat. Assoc. 71, 158–168.

Bernholt, T., 2005. Computing the least median of squares estimator in time $O(n^d)$, in: Gervasi, O., Gavrilova, M.L., Kumar, V., Laganà, A., Lee, H.P., Mun, Y., Taniar, D., Tan, C.J.K. (Eds.), Computational Science and Its Applications — ICCSA 2005, Springer. pp. 697–706.

Buckley, J., James, I., 1979. Linear regression with censored data. Biometrika 66, 429–436.

Carrizosa, E., Plastria, F., 1995. The determination of a “least quantile of squares regression line” for all quantiles. Comput. Stat. Data Anal. 20, 467–479.

Cattaneo, M., 2007. Statistical Decisions Based Directly on the Likelihood Function. Ph.D. thesis. ETH Zurich.

Cattaneo, M., Wiencierz, A., 2011. Robust regression with imprecise data. Technical Report 114. Department of Statistics, LMU Munich.

Cattaneo, M., Wiencierz, A., 2012. Likelihood-based Imprecise Regression. Int. J. Approx. Reasoning, in press. A preliminary version is available as Technical Report 116, Department of Statistics, LMU Munich.

Chen, S.X., Van Keilegom, I., 2009. A review on empirical likelihood methods for regression. Test 18, 415–447.

Cheney, E., 1982. Introduction to Approximation Theory. Chelsea Publishing. 2nd edition.

Dempster, A.P., Rubin, D.B., 1983. Rounding error in regression: The appropriateness of Sheppard’s corrections. J. R. Stat. Soc., Ser. B 45, 51–59.

Edelsbrunner, H., Souvaine, D.L., 1990. Computing least median of squares regression lines and guided topological sweep. J. Am. Stat. Assoc. 85, 115–119.

Ferson, S., Kreinovich, V., Hajagos, J., Oberkampf, W., Ginzburg, L., 2007. Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty. Technical Report SAND2007-0939. Sandia National Laboratories.

Gioia, F., Lauro, C.N., 2005. Basic statistical methods for interval data. Ital. J. Appl. Stat. 17, 75–104.

Hampel, F.R., 1975. Beyond location parameters: Robust concepts and methods. Bull. Int. Stat. Inst. 46, 375–382.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust Statistics: The Approach Based on Influence Functions. Wiley.

Hawkins, D.M., 1993. The feasible set algorithm for least median of squares regression. Comput. Stat. Data Anal. 16, 81–101.

Heitjan, D.F., Rubin, D.B., 1991. Ignorability and coarse data. Ann. Stat. 19, 2244–2253.

Huber, P.J., Ronchetti, E.M., 2009. Robust Statistics. Wiley. 2nd edition.

Knuth, D.E., 1998. The Art of Computer Programming. Volume 3: Sorting and Searching. Addison-Wesley. 2nd edition.

Li, G., Zhang, C.H., 1998. Linear regression with interval censored data. Ann. Stat. 26, 1306–1327.

Manski, C.F., Tamer, E., 2002. Inference on regressions with interval data on a regressor or outcome. Econometrica 70, 519–546.

Marino, M., Palumbo, F., 2002. Interval arithmetic for the evaluation of imprecise data effects in least squares linear regression. Ital. J. Appl. Stat. 14, 277–291.

Maronna, R.A., Martin, D.R., Yohai, V.J., 2006. Robust Statistics: Theory and Methods. Wiley.

Mount, D.M., Netanyahu, N.S., Romanik, K., Silverman, R., Wu, A.Y., 2007. A practical approximation algorithm for the LMS line estimator. Comput. Stat. Data Anal. 51, 2461–2486.

Pötter, U., 2000. A multivariate Buckley-James estimator, in: Kollo, T., Tiit, E.M., Srivastava, M. (Eds.), Multivariate Statistics. VSP, pp. 117–131.

R Development Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Used R version 2.15.0.

Rousseeuw, P.J., 1984. Least median of squares regression. J. Am. Stat. Assoc. 79, 871–880.

Rousseeuw, P.J., Leroy, A.M., 1987. Robust Regression and Outlier Detection. Wiley.

Souvaine, D.L., Steele, J., 1987. Time- and space-efficient algorithms for least median of squares regression. J. Am. Stat. Assoc. 82, 794–801.

Steele, J., Steiger, W., 1986. Algorithms and complexity for least median of squares regression. Discrete Appl. Math. 14, 93–100.

Stromberg, A.J., 1993. Computing the exact least median of squares estimate and stability diagnostics in multiple linear regression. SIAM J. Sci. Comput. 14, 1289–1299.

Tasche, D., 2003. Unbiasedness in least quantile regression, in: Dutter, R., Filzmoser, P., Gather, U., Rousseeuw, P.J. (Eds.), Developments in Robust Statistics, Physica-Verlag. pp. 377–386.

Utkin, L.V., Coolen, F.P.A., 2011. Interval-valued regression and classification models in the framework of machine learning, in: Coolen, F., de Cooman, G., Fetz, T., Oberguggenberger, M. (Eds.), ISIPTA ’11, Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications, SIPTA. pp. 371–380.

Vansteelandt, S., Goetghebeur, E., Kenward, M.G., Molenberghs, G., 2006. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. Stat. Sin. 16, 953–979.

Watson, G.A., 1998. On computing the least quantile of squares estimate. SIAM J. Sci. Comput. 19, 1125–1138.

Wiencierz, A., 2012. linLIR: linear Likelihood-based Imprecise Regression. R package version 1.0-1.