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

Marco E. G. V. Cattaneo · **Andrea Wiencierz**

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 Interval-censored data · Nonparametric likelihood inference · Robust regression · Least median of squares · Exact algorithm · R package

M. Cattaneo
Department of Statistics, LMU Munich, Ludwigstraße 33, 80539 München, Germany
E-mail: cattaneo@stat.uni-muenchen.de

A. Wiencierz
Department of Statistics, LMU Munich, Ludwigstraße 33, 80539 München, Germany
E-mail: andrea.wiencierz@stat.uni-muenchen.de

This is the unformatted version of: `doi:10.1007/s00180-013-0459-9`

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 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 the residuals' quantiles 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 practically 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 (see Wiencierz and Cattaneo, 2012, for some preliminary ideas). 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). Here, we develop the algorithm for the robust LIR method in full detail and generality. In particular, we do not assume that the data are "in general position", since this assumption (which is usual in the context of least median of squares regression) would be too restrictive for interval-censored data.

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 used to analyze data from the European Social Survey (ESS) 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 $\mathscr{F} := \{f_{a,b} : a,b \in \mathbb{R}\}$, where the functions $f_{a,b} : \mathbb{R} \to \mathbb{R}$ are defined by $f_{a,b}(x) = a + bx$ 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 $f \in \mathscr{F}$. For each $f \in \mathscr{F}$, 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 $f \in \mathscr{F}$, and that there is a unique $f_0 \in \mathscr{F}$ 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 $f \in \mathscr{F}$ and each $q \in \mathbb{R}_{\geq 0}$, let

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

be the closed band of (vertical) width $2q$ 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 $f\in\mathscr{F}$, let $\mathscr{C}_f$ 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, $\mathscr{C}_f$ 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.

If the quantities of interest were precisely observed, $\mathscr{C}_f$ would be the empirical likelihood confidence interval obtained by thresholding the empirical likelihood ratio at level β. The fact that the quantities of interest are only imprecisely observed entails an enlargement of this interval. Therefore, $\mathscr{C}_f$ is asymptotically a (conservative) confidence region of level $F_{\chi^2}(-2\log\beta)$ for the p-quantile of the distribution of the (absolute) residual R_f, where F_{χ^2} is the cumulative distribution function of the chi-square distribution with 1 degree of freedom (see Owen, 2001; Cattaneo and Wiencierz, 2012, for more details).

In order to obtain an explicit formula for the confidence regions $\mathscr{C}_f$, we define

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

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

where $\underline{i}:=\lceil(p-\varepsilon)n\rceil$ and $\overline{i}:=\lfloor(p+\varepsilon)n\rfloor+1$ (i.e., $\underline{i}-1$ is the largest integer smaller than $(p-\varepsilon)n$, while $\overline{i}$ is the smallest integer larger than $(p+\varepsilon)n$). Clearly, the two integers $\underline{k}$ and $\overline{k}$ depend on ε, p, n, and β, and satisfy

$$0\leq\underline{k}\leq\underline{i}-1<(p-\varepsilon)n\leq pn\leq(p+\varepsilon)n<\overline{i}\leq\overline{k}\leq n.$$

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 $\underline{k}=\underline{i}-1$ and $\overline{k}=\overline{i}$.

Now, for each function $f\in\mathscr{F}$ 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 $\underline{r}_{f,i}\in\mathbb{R}_{\geq0}$, while $\overline{r}_{f,i}\in\overline{\mathbb{R}}_{\geq0}$. 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

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

for all $f \in \mathscr{F}$. That is, the likelihood-based confidence region $\mathscr{C}_f \subseteq \mathbb{R}_{\geq 0}$ 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 $\mathscr{C}_f$ is proper (i.e., it contains more than one value), even when β is so large that $\underline{k} = \underline{i} - 1$ and $\overline{k} = \overline{i}$. In this case, $\mathscr{C}_f$ 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 $\underline{i} = n/2$ and $\overline{i} = n/2 + 1$, 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 $f \in \mathscr{F}$, we have an interval estimate $\mathscr{C}_f$ for the p-quantile of the distribution of the (absolute) residual R_f. As in least quantile of squares regression, we would like to select the regression function $f \in \mathscr{F}$ by minimizing the estimate of the residual's p-quantile, but comparing the intervals $\mathscr{C}_f$ gives us only a partial order on $\mathscr{F}$. The linear functions $f \in \mathscr{F}$ 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 $f' \in \mathscr{F}$ such that $\overline{r}_{f',(\overline{k})} < \underline{r}_{f,(\underline{k}+1)}$. In order to simplify the description of the undominated functions, define

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

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

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

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 $\mathscr{U}$ 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 $\mathscr{U}$. As regards

the first part, we will show that in order to determine $\overline{q}_{LRM}$, it suffices to minimize $\overline{r}_{f,(\overline{k})}$ over a finite subset of $\mathscr{F}$. For the second part of the algorithm, we will see that the set of all the intercept-slope pairs corresponding to the undominated regression functions is the union of finitely many polygons. As a by-product, we obtain a representation of the result of the least quantile of squares regression in the case of precise data, without any assumption about the data being "in general position". We will show that the computational complexity of our exact algorithm is $O(n^3 \log n)$. We have implemented this 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 $\mathscr{D}$ 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 $\mathscr{B} := \{0\}$ if there are less than $\overline{k}$ bounded imprecise observations (i.e., if $|\mathscr{D}| < \overline{k}$, where $|\mathscr{D}|$ denotes the cardinality of the set $\mathscr{D}$), and

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

otherwise (i.e., if $|\mathscr{D}| \geq \overline{k}$). 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 $b \in \mathscr{B}$, 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*

$$\overline{q}_{LRM} = +\infty,$$
$$\{f\in\mathscr{F} : \overline{r}_{f,(\overline{k})} = \overline{q}_{LRM}\} = \mathscr{F}.$$

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),*

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

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

where the set on the left-hand side is infinite when the inclusion is strict. However, the inclusion is certainly an equality when the following condition is satisfied: if there is a pair $(i,j)\in\mathscr{D}^2$ *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* $[\underline{y}_j,\overline{y}_j]\subseteq[\underline{y}_i,\overline{y}_i]$*).*

Some further explanations are needed to fully understand the results in Theorem 1. As seen in Subsection 2.2, for each linear function $f\in\mathscr{F}$, 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 $f\in\mathscr{F}$ 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 $f\in\mathscr{F}$ 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 $i\in\{1,\ldots,n\}$ (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 $f\in\mathscr{F}$ 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) 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 $\mathscr{B}$ 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 $b \in \mathscr{B}$, 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 $[\underline{z}_{b,1}, \overline{z}_{b,1}], \ldots, [\underline{z}_{b,n}, \overline{z}_{b,n}]$. This is a finite problem: it suffices to consider the intervals $[\underline{z}_{b,(j)}, \overline{z}_{b,[j]}]$ with $j \in \{1, \ldots, n-\overline{k}+1\}$.

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 $\mathscr{F}$ to a minimization problem on the finite set $\mathscr{B} \times \{1, \ldots, n-\overline{k}+1\}$. This constitutes the first part of our algorithm for the robust LIR analysis.

Besides that, Theorem 1 gives us also an algorithm for finding all LRM regression functions, when the condition at the end of the theorem is satisfied. An explicit formula for the set of all LRM regression functions in the general case (i.e., also when this condition is not satisfied) can be easily obtained, but requires several case distinctions and goes beyond the scope of the present paper. However, a brief comment on the condition at the end of Theorem 1 can be helpful in understanding the theorem. This condition is sufficient (but not necessary) for excluding the cases in which the set of all LRM regression functions is an infinite, proper subset of $\mathscr{F}$. That is, for excluding the situations in which there is a thinnest band of the form $\overline{B}_{f,q}$ containing at least $\overline{k}$ imprecise observations, but the union of these imprecise observations touches each border of the band in only one point. In fact, in such situations these contact points have the same x-coordinate, and can thus be written as $(x, y+\overline{q}_{LRM})$ and $(x, y-\overline{q}_{LRM})$, for some $x, y \in \mathbb{R}$. In this case, there are infinitely many linear functions $f \in \mathscr{F}$ going through the point (x, y) and such that the band $\overline{B}_{f,\overline{q}_{LRM}}$ contains at least $\overline{k}$ imprecise observations. All these functions f are LRM regression functions.

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

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

Theorem 2

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

A linear function $f \in \mathscr{F}$ 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 $[\underline{z}_{b,1}, \overline{z}_{b,1}], \ldots, [\underline{z}_{b,n}, \overline{z}_{b,n}]$. For each $b \in \mathbb{R}$ and each nonempty set $\mathscr{I} \subseteq \{1,\ldots,n\}$, 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 $i \in \mathscr{I}$ if and only if $a \in [\max_{i\in\mathscr{I}} \underline{z}_{b,i} - \overline{q}_{LRM}, \min_{i\in\mathscr{I}} \overline{z}_{b,i} + \overline{q}_{LRM}]$. Therefore,

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

Theorem 2 gives a simpler expression for $\mathscr{U}$, 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

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

of all the intercept-slope pairs corresponding to the undominated regression functions. More precisely, $\mathscr{U}'$ is a subset of the plane $\mathbb{R}^2$ bounded by finitely many line segments and half-lines. However, $\mathscr{U}'$ 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 $\mathscr{U}'$ 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 $i \in \{1,\ldots,n\}$, 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 $j \in \{1,\ldots,n-\underline{k}\}$, 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 $\mathscr{U}'$ is a closed subset of the plane $\mathbb{R}^2$ bounded by finitely many line segments and half-lines. That is, $\mathscr{U}'$ 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 $j \in \{1,\ldots,n-\underline{k}\}$) can be discontinuous at $b = 0$. As a consequence, Theorem 2 implies that $\mathscr{U}'$ is a subset of the plane $\mathbb{R}^2$ bounded by finitely many line segments and half-lines, but $\mathscr{U}'$ is not necessarily closed. However, the two parts $\mathscr{U}' \cap (\mathbb{R} \times \{0\})$ and $\mathscr{U}' \cap (\mathbb{R} \times \mathbb{R}_{\neq 0})$ are relatively closed in $\mathbb{R} \times \{0\}$ and $\mathbb{R} \times \mathbb{R}_{\neq 0}$, respectively.

If the imprecise data are in fact precise (i.e., all interval data are degenerate: $\underline{x}_i = \overline{x}_i$ and $\underline{y}_i = \overline{y}_i$ for all $i \in \{1,\ldots,n\}$) and $\underline{k} = \overline{k} - 1$, then $\mathscr{U}$ is the set of all least quantile of squares regression functions $f \in \mathscr{F}$ minimizing the $\overline{k}$th smallest absolute residual $\underline{r}_{f,(\overline{k})} = \overline{r}_{f,(\overline{k})}$. That is, Theorem 2 gives us in particular an algorithm for calculating the result of the least quantile of squares regression, without any assumption about the data being "in general position".

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 $b \in \mathscr{B}$, we must determine the pair $(\underline{z}_{b,(j)}, \overline{z}_{b,[j]})$ (with $j \in \{1,\ldots,n-\overline{k}+1\}$) 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 $\underline{z}_{b,1},\ldots,\underline{z}_{b,n}$ and $\overline{z}_{b,1},\ldots,\overline{z}_{b,n}$, we sort the two lists, obtaining $\underline{z}_{b,i_1},\ldots,\underline{z}_{b,i_n}$ (with $\underline{z}_{b,i_j} = \underline{z}_{b,(j)}$) and $\overline{z}_{b,(1)},\ldots,\overline{z}_{b,(n)}$. 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 $j \in \{1,\ldots,n-\overline{k}+1\}$ (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 $j \in \{1,\ldots,n-\overline{k}+1\}$, 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 $j \in \{1,\ldots,n-\underline{k}\}$. This can be done in time $O(n \log n)$, since it suffices to calculate and sort the two lists $\underline{z}_{b,1},\ldots,\underline{z}_{b,n}$ and $\overline{z}_{b,1},\ldots,\overline{z}_{b,n}$, 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 $\mathscr{U}'$ 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 $\mathscr{U}'$ 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 $b \mapsto \underline{z}_{b,i} - \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,i} + \overline{q}_{LRM}$ 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 $\mathscr{U}'$ 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 the statistical software environment R (R Development Core Team, 2012). It is part of the package `linLIR` (Wiencierz, 2013), designed for the implementation of 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 robust LIR analysis for two variables out of the data object, as well as 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 LIR results including, e.g., the set $\mathscr{U}'$.

The `linLIR` package provides a ready-to-use first implementation of the robust LIR method for linear regression with interval data, although the current version of the `s.linlir` function is not optimized for computational speed yet. In the following section, we illustrate the implementation by means of an application example.

4 Analysis of ESS data using the `linLIR` package

In recent years, there has been a lively interest in analyzing subjective well-being in various disciplines of the social and behavioral sciences. In this context, one important question is how an increase in income translates to subjective well-being (see, e.g., Deaton, 2012; Clark et al, 2008; Diener and Biswas-Diener, 2002). Empirical studies in this field often use global measures of subjective well-being, which are obtained from a single survey question about the overall satisfaction with life. These global measures are indicators of the state of an individual's well-being, and therefore, it is sensible to use them to analyze subjective well-being (Deaton, 2008), although, of course, they do not capture the entire complexity of the concept of well-being (Huppert et al, 2009). As single-item measures are usually measured on a discrete scale, they can be considered as coarse observations of the latent, continuous variable of interest *degree of subjective well-being*. The coarseness of the discrete values can be represented by intervals, thus, the LIR approach is suitable to analyze this kind of data. Moreover, when investigating the relation between income and subjective well-being, sometimes also the income data are only available as classes,

which represent in fact intervals that form a partition of the associated observation space $\mathbb{R}_{\geq 0}$. Finally, as the relation between income and subjective well-being is usually assumed to be log-linear (see, e.g., Deaton, 2012; Diener and Biswas-Diener, 2002), we can conduct a linear LIR analysis with the logarithm of income as independent variable X and subjective well-being as dependent variable Y to analyze the relation of interest, accounting for the imprecision of the data.

In this section, we analyze data from the fifth round of the ESS (Norwegian Social Science Data Services, 2010) to illustrate the implementation of the linear LIR analysis. The ESS is a biennial multi-country survey established to monitor changing attitudes and behavior of people in Europe. The data collected for the ESS are available free of charge on the ESS website `www.europeansocialsurvey.org`.

Previous empirical studies indicated that the relation between income and subjective well-being on the individual level is not the same in rich countries as in poor countries, and furthermore, that there may be different relations for men than for women (see, e.g., Clark et al, 2005; Diener and Biswas-Diener, 2002). For these reasons, we choose Finland and Bulgaria as representatives for the groups of rich and poor European countries, respectively, and we will analyze only the corresponding subsets of the ESS data set. Furthermore, for each country we will perform separate LIR analyses for the subpopulations of women and men. From the variables included in the ESS data set we retrieve the following ones: *household income* (net per month, in categories corresponding to the decile classes of the income distribution in each country) and *overall satisfaction with life* (on a discrete scale from 0 – *extremely dissatisfied* to 10 – *extremely satisfied*). In a data preprocessing step, the income classes are replaced by the corresponding intervals, then the interval endpoints are divided by the household size and, finally, the logarithmic transformation is made. The data on subjective well-being are changed from discrete values $0, 1, \ldots, 9, 10$ to intervals $[0, 0.5], [0.5, 1.5], \ldots, [8.5, 9.5], [9.5, 10]$. Hence, the independent and dependent precise quantities whose relation is investigated by the linear LIR analysis are the logarithm of monthly net household income per capita in euros and the subjective well-being on a latent, continuous scale from 0 to 10.

The resulting data frames contain each four columns: two for each of the analyzed variables, one column for the lower interval endpoint and one for the upper endpoint, which is the required data format for the `linLIR` package. Applying the function `idf.create` to these data frames, we create so-called interval data frame (`idf`) objects, which consist of a list of data frames, each containing the corresponding two columns of interval endpoints of one variable. For these `idf`-objects, the `linLIR` package provides a `summary` method as well as a `plot` method with two options. Figures 1 and 2 show the data plots of the four data sets we will analyze. As the data sets consist of roughly 1 000 observations each, we used the two-dimensional histogram plot by choosing the option `typ="hist"` in the `plot` function. As expected, we notice that the marginal distribution of subjective well-being is concentrated at a higher level in Finland compared to Bulgaria, but there appear to be no big differences between men and women within the countries. Moreover, we can see that there are many observations that are unbounded with respect to X. This is partly caused by the high number of observations in the lowest and highest income classes. In addition to this, there is a significant percentage of completely missing income values

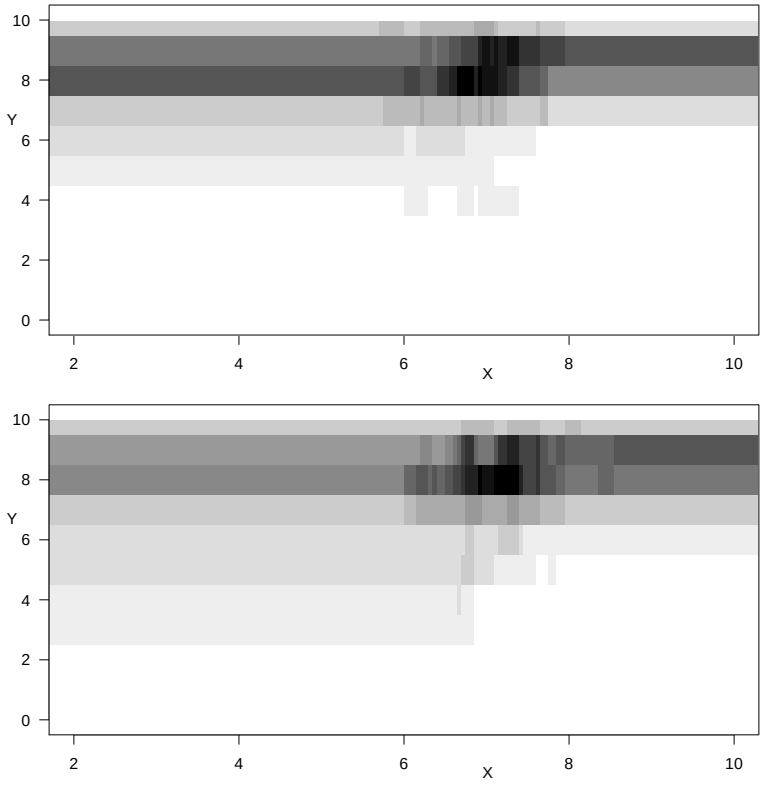

Fig. 1 Histogram plots of the Finnish data sets: women above ($n = 967$), men below ($n = 911$). The darker a rectangle the more observations overlap this rectangle.

(Finland 5–10%, Bulgaria 15–20%), which are represented in the data set as intervals $[\underline{x}_i, \overline{x}_i] = [-\infty, +\infty]$. Given the high degree of data imprecision, we can expect to obtain rather uninformative results from the LIR analyses, reflecting the high uncertainty induced by the interval data. One could argue that using $-\infty$ as lower endpoint of the range of the logarithmic income (instead of using, e.g., zero) entails too much unnecessary data uncertainty. However, the results of the LIR analyses are affected only a little by this, because the LIR method is very robust.

Before conducting the linear LIR analyses, we have to set up the probability model by selecting the only model parameter ε, and furthermore, we need to choose the quantile to be considered and the cutoff point β. For simplicity, we here assume that the imprecise 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 or if we wanted to account for possibly wrong coarsening, a positive ε could be considered in the probability model. This would lead to more imprecise results of the LIR analyses, reflecting the fact that there is additional uncertainty. As the residual's quantile to be minimized we consider the median (i.e., $p = 0.5$) because it can be shown that the robust LIR method yields the most robust results in this case. Finally, we choose $\beta = 0.8$ as cutoff point for the likelihood-based

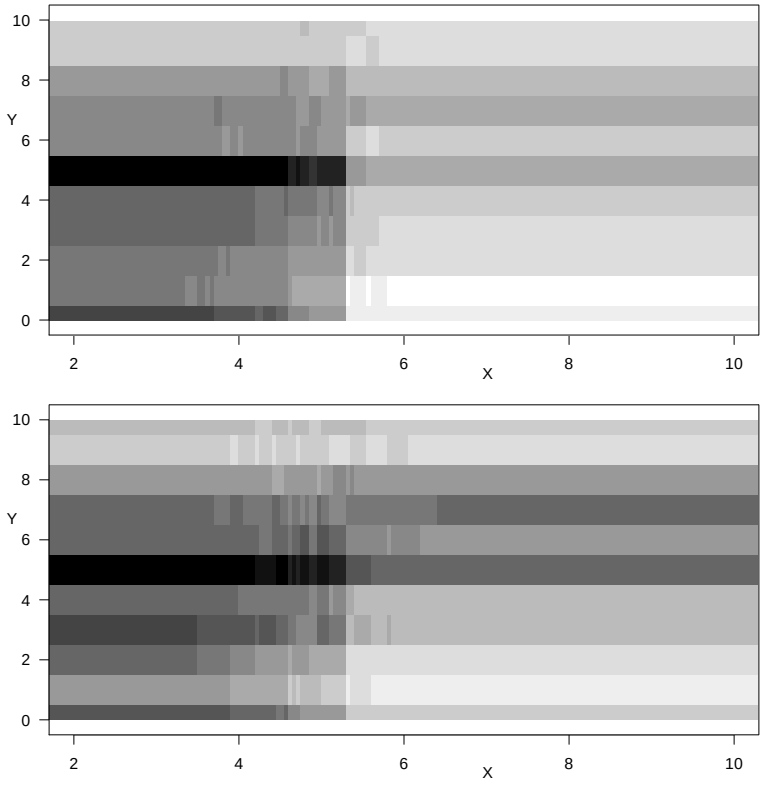

Fig. 2 Histogram plots of the Bulgarian data sets: women above ($n = 1\,370$), men below ($n = 1\,064$). The darker a rectangle the more observations overlap this rectangle.

confidence regions $\mathscr{C}_f$ with $f \in \mathscr{F}$. This choice of β corresponds to an asymptotic confidence level of (approximately) 50% for each $\mathscr{C}_f$ (see Subsection 2.2).

The model parameter ε, the LIR settings p and β, as well as the `idf`-object to be analyzed are handed over to the `s.linlir` function, which determines the set $\mathscr{U}'$ by the introduced algorithm. In the current version of the `linLIR` package, the first part of the `s.linlir` function determines $\overline{q}_{LRM}$. After this, the range $[\underline{b}, \overline{b}]$ of slope values for which there may be undominated functions is identified. This is done by exploiting the representation of the set $\mathscr{U}'$ as the union of finitely many polygons. As described in Subsection 3.3, the possible vertices of the polygons are situated at those values $b \in \mathbb{R}$ at which the graphs of some of the functions $b \mapsto \underline{z}_{b,i} - \overline{q}_{LRM}$ and $b \mapsto \overline{z}_{b,i} + \overline{q}_{LRM}$ cross each other. The set of all these intersection points can be formulated similarly to the set $\mathscr{B}$ in terms of the endpoints of the interval data, and thus, it can easily be determined. Considering these values together with the slopes $b = 0$, the smallest slope minus 100, and the maximum plus 100, ordering them by their size and starting from the smallest value, one can find $\underline{b}$ as the first of these slopes for which the corresponding set $\bigcup_{j=1}^{n-\underline{k}} [\underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM}, \overline{z}_{b,(j)} + \overline{q}_{LRM}]$ is not empty. Analogously, starting from the highest values and descending, the upper endpoint $\overline{b}$ is identified. If $\underline{b}$ corresponds to the smallest or $\overline{b}$ to the highest of the

considered slopes, respectively, the set $\mathscr{U}'$ is unbounded. In this case, in the final part of the `s.linlir` function, the set of undominated functions is approximated only over a coarse grid of slope values ranging at most from -10^9 to 10^9 (if unbounded on both sides). Otherwise, $\mathscr{U}'$ is approximated by determining the corresponding intercept values over a fine grid across the identified range of slope values. As already mentioned at the end of Section 3, the current version of the function `s.linlir` is not optimized for speed. The computations for the present analysis took about 2 to 10 minutes on a standard desktop computer, most of the time is needed for the first part of the algorithm, where $\overline{q}_{LRM}$ is determined.

The `s.linlir` function returns an object of the class "`s.linlir`", a list object whose elements include the ranges of slope and intercept values in $\mathscr{U}'$, a data frame containing the intercept-slope combinations that represent the approximation of the set $\mathscr{U}'$, the bound $\overline{q}_{LRM}$, the analyzed data set, the used LIR settings, $\underline{k}$ and $\overline{k}$, etc. The `linLIR` package provides a `print` method and a `summary` method for these `s.linlir`-objects. To visualize the results, there is furthermore an associated `plot` method with three options, which are to plot only the LRM regression functions (`typ="lrm"`), to plot a random selection of functions out of the set $\mathscr{U}$ (`typ="func"`), or to plot the entire set $\mathscr{U}'$ (`typ="para"`). For Figures 3 and 4 we used the latter plot type with the default option `para.typ="polygon"` to display the results of the conducted linear LIR analyses, the black points indicate the LRM regression functions.

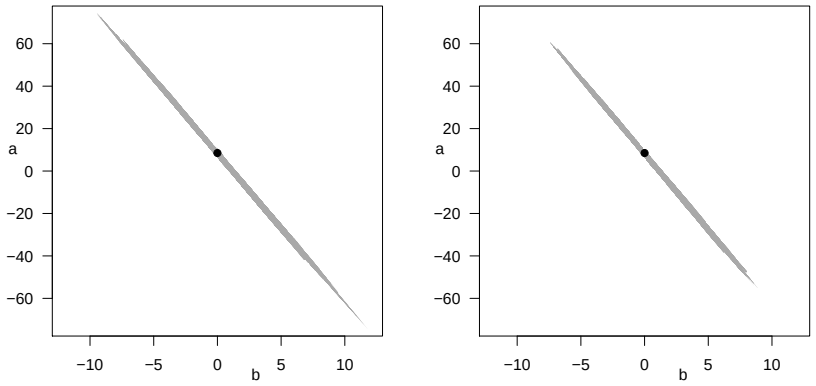

Fig. 3 Sets $\mathscr{U}'$ for Finland: women on the left ($n = 967$), men on the right ($n = 911$).

The sets $\mathscr{U}'$ resulting from the LIR analyses of the data sets of women and men in Finland are displayed in Figure 3. Both sets of parameter values are bounded and have a similar shape, admitting both lines with positive and negative slopes ranging approximately from -9.5 to 12. For the sample of Bulgarian women, the shape of the obtained set $\mathscr{U}'$ is much different, as shown in the left part of Figure 4. In this particular data set, there are 687 observations $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$ such that $\underline{x}_i = -\infty$ and $[\underline{y}_i, \overline{y}_i] \neq \mathbb{R}$. A line with an arbitrarily high slope will always go through these observations at the lower end of the income range as long as the intercept is not too low, and conversely, a line with a negative slope will always intersect these obser-

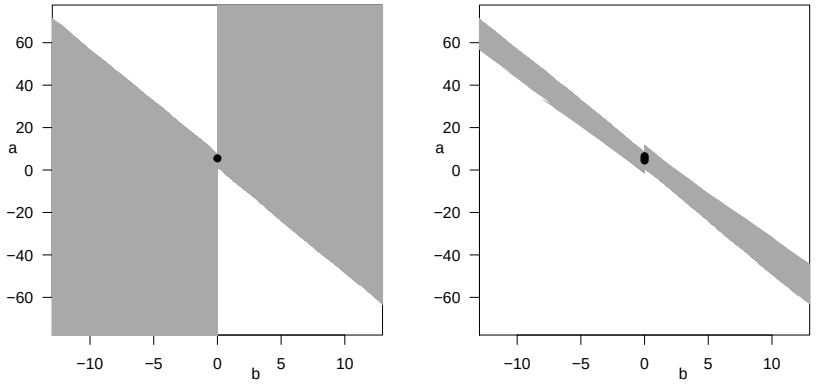

Fig. 4 Sets $\mathscr{U}'$ for Bulgaria: women on the left ($n = 1\,370$), men on the right ($n = 1\,064$).

vations if the intercept is not too high. As here $\underline{k} + 1 = 673$, all lines intersecting these 687 observations are undominated. Therefore, the obtained set of undominated functions is unbounded, reflecting the high degree of imprecision inherent in this data set. Furthermore, we here observe the particular data situation discussed at the end of Subsection 3.2, where the set $\mathscr{U}'$ is not closed. (The borders at $b = 0$ are not included.) In the LIR results for the sample of men in Bulgaria, $\mathscr{U}'$ is not unbounded, but large, which is to some extent due to the almost 20% of missing income values. In the right part of Figure 4, we displayed only the middle section of $\mathscr{U}'$. Interestingly, in this LIR analysis we find three LRM regression lines. These lines can be characterized geometrically by the fact that the closed bands of width $2\overline{q}_{LRM} = 4$ around them completely include at least $\overline{k} = 543$ observations. In the present data set, there are only 500 observations bounded with respect to X, therefore, only the band around a horizontal line can contain at least 543 observations. Hence, each of the three functions has slope 0.

The results of the LIR analyses do not give a clear answer to the question how an increase in income translates to subjective well-being. However, the obtained results are more or less in line with current research in this field, as there is no clear evidence about the direct relationship between these two variables. Some empirical studies in rich countries found only very weak positive effects of income on subjective well-being, while others even suggested a negative effect at the upper end of the income distribution (Diener and Biswas-Diener, 2002). These two possibilities are also admitted by the LIR results for the Finnish data sets, admitting increasing and decreasing functions. In poorer countries, several studies found a strong positive effect, reflecting the fact that in these countries an increase in income is more often used to fulfill basic material needs, clearly improving the individual living standard (Diener and Biswas-Diener, 2002). The LIR result for the sample of Bulgarian men admits more extreme slope and intercept values, while the data of the sample of Bulgarian women are too imprecise to obtain informative results.

5 Conclusion

In the present work, 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 studied 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). In particular, when the data are precise, the algorithm can calculate the (possibly infinite) set of all least median of squares regression functions, without any assumption about the data being "in general position".

We have implemented this exact algorithm as part of the R package `linLIR` (Wiencierz, 2013). In the present paper, we analyzed data of the fifth round of the ESS (Norwegian Social Science Data Services, 2010) to illustrate the implementation of the robust LIR method in the `linLIR` package. The obtained results are in line with current research in the field. In addition to that, 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 (which can also be generalized to imprecise data other than intervals). In particular, the first part of our algorithm can be easily extended to the problem of multiple linear regression by adapting the ideas of Stromberg (1993) to the case of interval data.

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

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

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

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

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

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

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

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

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 $j \in \{k,\ldots,n-1\}$,

$$[\underline{w}_j, \overline{w}_{j-k+1:j}] \cup [\underline{w}_{j+1}, \overline{w}_{j+1-k+1:j+1}] = [\underline{w}_j, \overline{w}_{(j-k+1)}] \cup [\underline{w}_{j+1}, \overline{w}_{j+1-k+1:j+1}]. \tag{2}$$

Since $\overline{w}_{(j-k+1)} \le \overline{w}_{j-k+1:j}$ always holds, (2) 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 $i \in \{j+1,\ldots,n\}$ such that $\overline{w}_i \le \overline{w}_{(j-k+1)}$, but then

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

and thus both unions in (2) are equal to the interval $[\underline{w}_j, \overline{w}_{j+1-k+1:j+1}]$. Therefore, using (2) for each j from k to $n-1$, we obtain

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

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

A.1 Proof of Theorem 1

As noted in Subsection 2.2, for each linear function $f \in \mathscr{F}$, we have $\overline{r}_{f,(\bar{k})} < +\infty$ if and only if either f is not constant and there are at least $\bar{k}$ bounded imprecise observations, or f is constant and there are at least $\bar{k}$ 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 $\bar{k}$ 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 $\overline{r}_{f,(\bar{k})} = +\infty$ for all $f \in \mathscr{F}$, which proves the first part of the theorem. Otherwise, if there are at least $\bar{k}$ 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 $\overline{r}_{f,(\bar{k})} < +\infty$ at least for the constant functions $f \in \mathscr{F}$, which implies $\overline{q}_{LRM} < +\infty$.

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

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

$$\overline{z}_{b,i} = \sup_{(x,y)\in[\underline{x}_i,\overline{x}_i]\times[\underline{y}_i,\overline{y}_i]} (y - bx), \tag{4}$$

and therefore

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

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

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

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

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

Clearly, for each $b\in\mathbb{R}$ and $j\in\{1,\ldots,n-\overline{k}+1\}$ 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

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

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

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

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

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

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

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

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

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

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

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

Let $\mathscr{I}\in\mathscr{S}$, and consider first the case with $\mathscr{I}\nsubseteq\mathscr{D}$. In this case, there is an $i\in\mathscr{I}$ 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, $g_{\mathscr{I}}(a,b)<+\infty$ if and only if $b=0$, and therefore $\mathscr{M}_{\mathscr{I}}=\{0\}\subseteq\mathscr{B}$.

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

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

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 $\operatorname{arg\,min}_{(a,b)\in\mathbb{R}^2} g_{\mathscr{I}}(a,b)$ is not empty (i.e., $\mathscr{M}_{\mathscr{I}}\neq\varnothing$). Let thus $(a',b')\in\operatorname{arg\,min}_{(a,b)\in\mathbb{R}^2} g_{\mathscr{I}}(a,b)$ (hence, $b'\in\mathscr{M}_{\mathscr{I}}$). From the Characterization Theorem of Cheney (1982, page 35) it follows that there are $(x,y),(x',y')\in\bigcup_{i\in\mathscr{I}}\{\underline{x}_i,\overline{x}_i\}\times\{\underline{y}_i,\overline{y}_i\}$ 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 $i \in \mathscr{I}$ 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 $\mathscr{B}$ 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 $b' \in \mathscr{B}$.

Assume now that $b' \notin \mathscr{B}$. In order to complete the proof of the theorem, it suffices to show that in this case the set $\mathscr{M}_{\mathscr{I}}$ is infinite and intersects $\mathscr{B}$, and that the condition at the end of the theorem cannot be satisfied. The assumption $b' \notin \mathscr{B}$ 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 $i,j \in \mathscr{I}$), 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 $b' = 0 \in \mathscr{B}$, 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 $(i,j),(j,i) \in \mathscr{I}^2 \subseteq \mathscr{D}^2$ 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 $b' = 0 \in \mathscr{B}$, 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 $(i,i),(j,j) \in \mathscr{D}^2$ 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 $(i,j),(j,i),(i,i),(j,j) \in \mathscr{D}^2$.

In order to complete the proof of the theorem, it remains to show that the set $\mathscr{M}_{\mathscr{I}}$ is infinite and intersects $\mathscr{B}$. We have that $b \in \mathscr{M}_{\mathscr{I}}$ 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 $i \in \mathscr{I}$. 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 $a = a_b := {}^1\!/_2\,(y'+y) - bx$ (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

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

where the second case is implied by the fact that $\mathscr{B}_{x'',y''}$ is not empty (since $b' \in \mathscr{M}_{\mathscr{I}} \subseteq \mathscr{B}_{x'',y''}$), while in the other two cases the endpoints of $\mathscr{B}_{x'',y''}$ 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,

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

is a (nonempty) closed interval, which is either $\mathbb{R}$ or it is bounded. When $\mathscr{M}_{\mathscr{I}} = \mathbb{R}$, obviously it is infinite and intersects $\mathscr{B}$. Otherwise, $\mathscr{M}_{\mathscr{I}}$ is a bounded interval whose endpoints are elements of $\mathscr{B}$, 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 $\mathscr{M}_{\mathscr{I}}$ intersects $\mathscr{B}$ and is infinite, since $b' \notin \mathscr{B}$ is an interior point of the interval $\mathscr{M}_{\mathscr{I}}$.

A.2 Proof of Theorem 2

For each function $f_{a,b} \in \mathscr{F}$ and each imprecise observation $[\underline{x}_i, \overline{x}_i] \times [\underline{y}_i, \overline{y}_i]$, using (3) and (4), we obtain that $\underline{r}_{f_{a,b},i} \leq \overline{q}_{LRM}$ if and only if the set

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

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 $\mathscr{I} \subseteq \{1,\dots,n\}$ such that $|\mathscr{I}| = \underline{k}+1$ and $a \in [\underline{z}_{b,i} - \overline{q}_{LRM}, \overline{z}_{b,i} + \overline{q}_{LRM}]$ for all $i \in \mathscr{I}$. 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

$$\begin{aligned}\bigcup_{\mathscr{I} \subseteq \{1,\dots,n\} : |\mathscr{I}| = \underline{k}+1} \bigcap_{i \in \mathscr{I}} [\underline{z}_{b,i} - \overline{q}_{LRM}, \overline{z}_{b,i} + \overline{q}_{LRM}] &= \bigcup_{j=\underline{k}+1}^{n} [\underline{z}_{b,(j)} - \overline{q}_{LRM}, \overline{z}_{b,(j-\underline{k})} + \overline{q}_{LRM}] \\ &= \bigcup_{j=1}^{n-\underline{k}} [\underline{z}_{b,(\underline{k}+j)} - \overline{q}_{LRM}, \overline{z}_{b,(j)} + \overline{q}_{LRM}].\end{aligned}$$

Therefore,

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

References

Alexandrov AD (2005) Convex Polyhedra. Springer, Berlin

Beaton AE, Rubin DB, Barone JL (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 ML, Kumar V, Laganà A, Lee HP, Mun Y, Taniar D, Tan CJK (eds) Computational Science and Its Applications — ICCSA 2005, Springer, Berlin, 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. PhD thesis, ETH Zurich, DOI 10.3929/ethz-a-005463829

Cattaneo M, Wiencierz A (2011) Robust regression with imprecise data. Tech. Rep. 114, Department of Statistics, LMU Munich

Cattaneo M, Wiencierz A (2012) Likelihood-based Imprecise Regression. Int J Approx Reasoning 53:1137–1154

Chen SX, Van Keilegom I (2009) A review on empirical likelihood methods for regression. Test 18:415–447

Cheney E (1982) Introduction to Approximation Theory, 2nd edn. AMS Chelsea Publishing, Providence

Clark A, Etilé F, Postel-Vinay F, Senik C, Van der Straeten K (2005) Heterogeneity in reported well-being: Evidence from twelve European countries. Econ J 115:C118–C132

Clark A, Frijters P, Shields MA (2008) Relative income, happiness, and utility: An explanation for the Easterlin paradox and other puzzles. J Econ Lit 46:95–144

Deaton A (2008) Income, health, and well-being around the world: Evidence from the Gallup World Poll. J Econ Perspect 22:53–72

Deaton A (2012) The financial crisis and the well-being of Americans. Oxf Econ Pap 64:1–26

Dempster AP, Rubin DB (1983) Rounding error in regression: The appropriateness of Sheppard's corrections. J R Stat Soc, Ser B 45:51–59

Diener E, Biswas-Diener R (2002) Will money increase subjective well-being? Soc Indicators Res 57:119–169

Edelsbrunner H, Souvaine DL (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. Tech. Rep. SAND2007-0939, Sandia National Laboratories

Gioia F, Lauro CN (2005) Basic statistical methods for interval data. Ital J Appl Stat 17:75–104

Hampel FR (1975) Beyond location parameters: Robust concepts and methods. Bull Int Stat Inst 46:375–382

Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust Statistics: The Approach Based on Influence Functions. Wiley, New York

Hawkins DM (1993) The feasible set algorithm for least median of squares regression. Comput Stat Data Anal 16:81–101

Heitjan DF, Rubin DB (1991) Ignorability and coarse data. Ann Stat 19:2244–2253

Huber PJ, Ronchetti EM (2009) Robust Statistics, 2nd edn. Wiley, New York

Huppert FA, Marks N, Clark A, Siegrist J, Stutzer A, Vittersø J, Wahrendorf M (2009) Measuring well-being across Europe: Description of the ESS well-being module and preliminary findings. Soc Indicators Res 91:301–315

Knuth DE (1998) The Art of Computer Programming. Volume 3: Sorting and Searching, 2nd edn. Addison-Wesley, Boston

Li G, Zhang CH (1998) Linear regression with interval censored data. Ann Stat 26:1306–1327

Manski CF, 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 RA, Martin DR, Yohai VJ (2006) Robust Statistics: Theory and Methods. Wiley, New York

Mount DM, Netanyahu NS, Romanik K, Silverman R, Wu AY (2007) A practical approximation algorithm for the LMS line estimator. Comput Stat Data Anal 51:2461–2486

Norwegian Social Science Data Services (2010) ESS Round 5: European Social Survey Round 5 Data. Data file edn. 3.0

Owen AB (2001) Empirical Likelihood. Chapman & Hall/CRC, Boca Raton

Pötter U (2000) A multivariate Buckley-James estimator. In: Kollo T, Tiit EM, Srivastava M (eds) Multivariate Statistics, VSP, Utrecht, pp 117–131

R Development Core Team (2012) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, used R version 2.15.2

Rousseeuw PJ (1984) Least median of squares regression. J Am Stat Assoc 79:871–880

Rousseeuw PJ, Leroy AM (1987) Robust Regression and Outlier Detection. Wiley, New York

Souvaine DL, 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 AJ (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 PJ (eds) Developments in Robust Statistics, Springer, Berlin, pp 377–386

Utkin LV, Coolen FPA (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, SIPTA, Manno, pp 371–380

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

Watson GA (1998) On computing the least quantile of squares estimate. SIAM J Sci Comput 19:1125–1138

Wiencierz A (2013) linLIR: linear Likelihood-based Imprecise Regression. R package version 1.1-1

Wiencierz A, Cattaneo M (2012) An exact algorithm for Likelihood-based Imprecise Regression in the case of simple linear regression with interval data. In: Kruse R, Berthold MR, Moewes C, Gil MÁ, Grzegorzewski P, Hryniewicz O (eds) Synergies of Soft Computing and Statistics for Intelligent Data Analysis, Springer, Berlin, pp 293–301