An exact algorithm for Likelihood-based Imprecise Regression in the case of simple linear regression with interval data

Andrea Wiencierz and Marco E. G. V. Cattaneo

Abstract Likelihood-based Imprecise Regression (LIR) is a recently introduced approach to regression with imprecise data. Here we consider a robust regression method derived from the general LIR approach and we establish an exact algorithm to determine the set-valued result of the LIR analysis in the special case of simple linear regression with interval data.

Key words: interval data, robust regression, likelihood inference, algorithm

1 Introduction

In [3], Likelihood-based Imprecise Regression (LIR) was introduced as a very general theoretical framework for regression analysis with imprecise data. Within the context of LIR, the term imprecise data refers to imprecisely observed quantities. This means that one is actually interested in analyzing the relation between precise variables, but the available data provide only the partial information that the values each lie in some subset of the observation space. In the general formulation of LIR, the imprecise observations can be arbitrary subsets of the observation space, including as special cases actually precise data (where the subset is a singleton) and missing data (where the subset is the entire observation space).

The aim of a LIR analysis is to identify plausible descriptions of the relation between the unobserved precise quantities on the basis of the imprecise observations. This is achieved by applying a general methodology for likelihood inference with imprecise data to the regression problem with imprecise data as a problem of statistical inference. The mathematical details of the LIR approach are set out in [3].

Department of Statistics, LMU Munich, Ludwigstraße 33, 80539 München, Germany
andrea.wiencierz@stat.uni-muenchen.de · cattaneo@stat.uni-muenchen.de

This is the unformatted version of: doi:10.1007/978-3-642-33042-1_32

In this paper, we deal with the implementation of the robust regression method derived from the general LIR approach in [3]. There, a grid search was proposed as a first implementation, which served to obtain the (approximate) result of the LIR analysis for a quadratic regression problem with interval data. Here, we consider the special case of simple linear regression with interval data and we derive an exact algorithm to determine the set-valued result of the LIR analysis in this particular situation. In the following section, we review the relevant technical details of the robust LIR method, before we establish the exact algorithm in Section 3.

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. Thus, the set of regression functions considered here can be written as $\mathcal{F} = \{f_{a,b} : (a,b) \in \mathbb{R}^2\}$ with $f_{a,b} : \mathbb{R} \to \mathbb{R},\ x \mapsto a + b\,x$. Furthermore, we here focus on the particular case of interval data, where the imprecise data $V_i^* := [\underline{X}_i, \overline{X}_i] \times [\underline{Y}_i, \overline{Y}_i],\ i = 1, \ldots, n$ are (possibly unbounded) rectangles. To keep the notation simple, throughout the paper, we write $[\underline{I}, \overline{I}]$ for the set of all real numbers z such that $\underline{I} \leq z \leq \overline{I}$. This is not the standard notation if $\underline{I} = -\infty$ or $\overline{I} = +\infty$. Figure 1 gives an example of such a data set containing 17 observations with varying amounts of imprecision.

The robust regression method we consider in this paper is based on a fully nonparametric probability model. It is only assumed that the n random objects $(V_i, V_i^*),\ i = 1, \ldots, n$ (where $V_i := (X_i, Y_i)$ are the unobserved precise values) are independent and identically distributed, and that

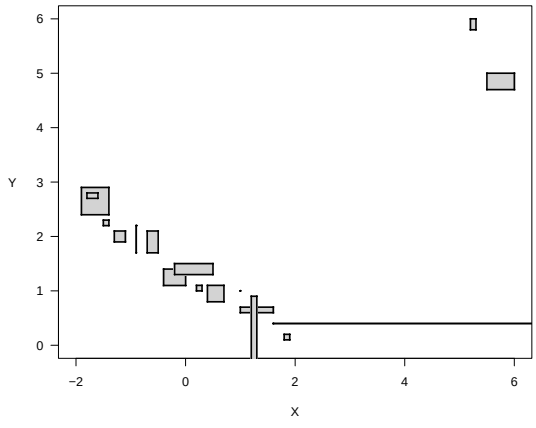

Fig. 1 Example data set containing 17 observations with varying amounts of imprecision: there is one actually precisely observed data point $V_i^* = [1,1] \times [1,1] = \{(1,1)\}$, there are two line segments (one of which is unbounded towards $+\infty$ in the X dimension), and, finally, there are 14 rectangles of different sizes and shapes (one of which is unbounded towards $-\infty$ in the Y dimension).

$P(V_i \in V_i^*) \geq 1-\varepsilon$, for some $\varepsilon \in [0,1]$. If $\varepsilon > 0$, this assumption implies that an imprecise observation may not cover the precise value with probability at most ε. Apart from this assumption, there is no restriction on the set of possible distributions of the data.

The relation between X and Y shall be described by a linear function. Which linear function is a suitable description of the relation when no particular structure of the joint distribution of X and Y is assumed? The basic idea behind the robust LIR method is that a possible description f can be evaluated by the p-quantile, $p \in]0,1[$, of the distribution of the corresponding (absolute) residual $|Y - f(X)|$. The closer to zero the p-quantile is, the better the associated function describes the relation between X and Y. Therefore, the linear function for which the p-quantile of the residual's distribution is minimal can be considered as the best description of the relation of interest. This linear function can be characterized geometrically as the central line of the thinnest band of the form $f \pm q,\ q \geq 0$, that contains (X,Y) with probability at least p.

This idea is very similar to the idea behind the robust regression method of least quantile of squares (or absolute deviations) regression, introduced in [5] as a generalization of the method of least median of squares regression (corresponding to the choice $p = 0.5$). Therefore, the LIR method can be seen as a generalization of these robust regression methods to the setting with imprecise data, where not only the optimal line is estimated, but a whole set of plausible descriptions is idenified.

To see how the robust LIR method works in detail, consider $V_1^* = A_1, \ldots, V_n^* = A_n$ as (nonempty) realizations of the imprecise data. Applying the general methodology for likelihood inference with imprecise data on which the LIR method is based, likelihood-based confidence regions for the p-quantile of the distribution of the precise residuals $R_{f,i} := |Y_i - f(X_i)|$, $i = 1, \ldots, n$, are determined for each considered regression function $f \in \mathcal{F}$. The confidence regions are obtained by cutting the (normalized) profile likelihood function for the p-quantile induced by the imprecise data at some cutoff point $\beta \in]0,1[$. The confidence regions cover the values of the p-quantiles corresponding to all probability distributions that give at least a certain probability to the observations, i.e. whose likelihood exceeds the threshold β.

To obtain the confidence regions, for each $f \in \mathcal{F}$ lower and upper (absolute) residuals are defined as follows

$$\underline{r}_{f,i} = \min_{(x,y)\in A_i} |y - f(x)| \quad \text{and} \quad \overline{r}_{f,i} = \sup_{(x,y)\in A_i} |y - f(x)|, \quad i = 1, \ldots, n.$$

Let $0 =: \underline{r}_{f,(0)} \leq \underline{r}_{f,(1)} \leq \ldots \leq \underline{r}_{f,(n)} \leq \underline{r}_{f,(n+1)} := +\infty$ be the ordered lower residuals and $0 =: \overline{r}_{f,(0)} \leq \overline{r}_{f,(1)} \leq \ldots \leq \overline{r}_{f,(n)} \leq \overline{r}_{f,(n+1)} := +\infty$ be the ordered upper residuals. Furthermore, define $\underline{i} = \max(\lceil (p-\varepsilon)\, n \rceil, 0)$ and $\overline{i} = \min(\lfloor (p+\varepsilon)\, n \rfloor, n) + 1$. According to Corollary 1 of [3] the profile likelihood function for the p-quantile of the distribution of the residuals corresponding to some function $f \in \mathcal{F}$ is a piecewise constant function whose

points of discontinuity are given by $\underline{r}_{f,(0)}, \ldots, \underline{r}_{f,(\underline{i})}, \overline{r}_{f,(\overline{i})}, \ldots, \overline{r}_{f,(n+1)}$. To obtain the confidence region $\mathcal{C}_f$ it thus suffices to identify the $(\underline{k}+1)$-th ordered lower residual and the $\overline{k}$-th ordered upper residual, which correspond to the points where the profile likelihood function jumps above and below the chosen threshold β, provided the condition $(\max\{p, 1-p\} + \varepsilon)^n \leq \beta$ holds. The values of $\underline{k}$ and $\overline{k}$ are determined on the basis of the explicit formula for the profile likelihood function given in [3]. They depend on n, on the choice of p and β, as well as on ε, which is part of the assumed probability model.

Thus, if $(\max\{p, 1-p\} + \varepsilon)^n \leq \beta$ is fulfilled, for each function $f \in \mathcal{F}$ the likelihood-based confidence region is the interval $\mathcal{C}_f := [\underline{r}_{f,(\underline{k}+1)}, \overline{r}_{f,(\overline{k})}]$ (see Corollary 2 of [3]). In order to find the best description of the relation between X and Y it is possible to follow a minimax approach and minimize the upper endpoint of the confidence interval over all considered regression functions. When there is a unique $f \in \mathcal{F}$ that minimizes $\sup \mathcal{C}_f$, it is optimal according to the Likelihood-based Region Minimax (LRM) criterion (see [1]) and therefore called f_{LRM}. If we consider the closed bands $\overline{B}_{f,q}$ defined for each function $f \in \mathcal{F}$ and each $q \in [0, +\infty[$ by

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

the function f_{LRM} can be characterized geometrically. The closed band $\overline{B}_{f_{LRM}, \overline{q}_{LRM}}$ (where $\overline{q}_{LRM} := \sup \mathcal{C}_{f_{LRM}}$) is the thinnest band of the form $\overline{B}_{f,q}$ containing at least $\overline{k}$ imprecise data, for all $f \in \mathcal{F}$ and all $q \in [0, +\infty[$. Thus, to determine the function f_{LRM} it suffices to adapt to the case of imprecise data an algorithm for the least quantile of squares regression, as we do in Section 3.1. Figure 2 shows f_{LRM} (solid line) for the LIR analysis of the example data set introduced in Figure 1, as well as the closed band $\overline{B}_{f_{LRM}, \overline{q}_{LRM}}$ of width $2\,\overline{q}_{LRM}$ (dashed lines).

However, f_{LRM} is not regarded as the final result of the LIR analysis. The aim of a LIR analysis is to describe the whole uncertainty about the

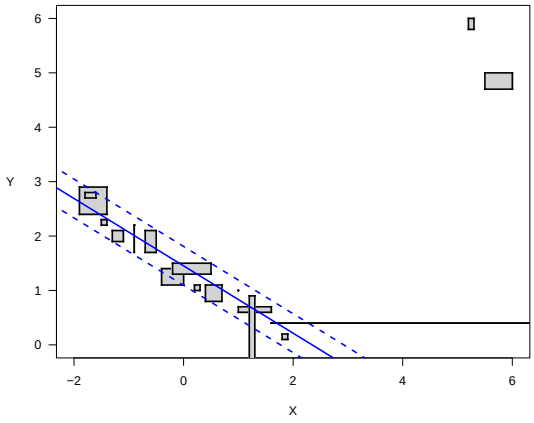

Fig. 2 Function f_{LRM} (solid line) for the LIR analysis of the example data set introduced in Figure 1 with $p = 0.5, \beta = 0.8, \varepsilon = 0$ (implying $\underline{k} = 7$ and $\overline{k} = 10$) and band $\overline{B}_{f_{LRM}, \overline{q}_{LRM}}$ (dashed lines).

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. Therefore, the set of all functions that are plausible in the light of the data is considered as the set-valued result of the LIR analysis, which describes the entire uncertainty involved in the regression problem with imprecise data. A regression function $f \in \mathcal{F}$ is regarded as plausible, if the corresponding confidence interval $\mathcal{C}_f$ is not strictly dominated by another one. Thus, the result of the LIR analysis is the set

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

The undominated functions can be characterized geometrically by the fact that the corresponding closed bands $\overline{B}_{f,\overline{q}_{LRM}}$ (i.e. the bands have width $2\,\overline{q}_{LRM}$) intersect at least $\underline{k}+1$ imprecise data. This characterization is the basis of the second part of the algorithm presented in the next section.

3 An exact algorithm for LIR

As a first implementation of the robust LIR method, we suggested in [3] a grid search over the space of parameters identifying the considered regression functions, while we considered a random search in [2]. Here, we derive an exact algorithm to determine the result of the robust LIR analysis in the case of simple linear regression with interval data. The algorithm consists of two parts: first, we find the optimal function f_{LRM}, which is then used to identify the set of all undominated regression lines. It can be proved that the computational complexity of the algorithm is $O(n^3 \log n)$, i.e. it is of the same order as the complexity of the initial algorithm for least median of squares regression (see [6]).

3.1 Part 1: Finding the LRM line

Analogously to what is shown in [6] for the case with precise data, it is possible to prove that, if the slope b_{LRM} of the function f_{LRM} is different from zero, the band $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ is determined by three imprecise observations V_i^* for which $\overline{r}_{f_{LRM},i} = \overline{q}_{LRM}$. Figure 3 illustrates this fact for the example of Figure 2. From this property follows that b_{LRM} is either zero or given by the slope of the line connecting the corresponding corner points of two of the observations. Thus, in order to identify candidates for b_{LRM} it suffices to consider the four slopes between the corresponding vertices of each pair of (nonidentical) bounded imprecise observations. In this way, we obtain a set of at most $4\binom{n}{2}+1$ candidates for the slope of f_{LRM}.

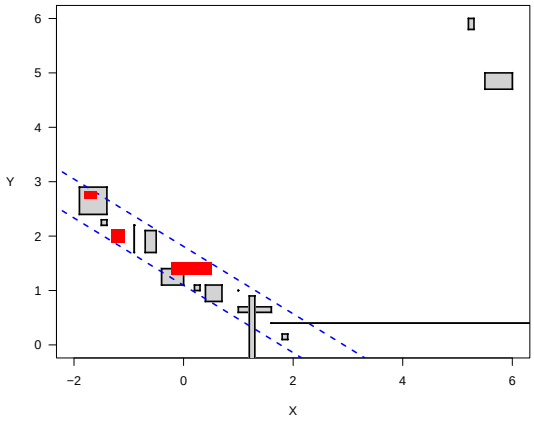

Fig. 3 Band $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ (dashed lines) for the LIR analysis considered in Figure 2. The three imprecise data determining $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ in this case are highlighted.

For a given slope b, it is easy to determine the intercept a for which the width of the resulting closed band around $f_{a,b}$ (containing at least $\overline{k}$ imprecise data) is minimal (over all linear functions with slope b). Consider the transformed data $Z_i^* := [\underline{Z}_i, \overline{Z}_i],\ i = 1, \ldots, n$ obtained as

$$\underline{Z}_i = \begin{cases} \underline{Y}_i - b\,\overline{X}_i\,, & b > 0 \\ \underline{Y}_i - b\,\underline{X}_i\,, & b \leq 0 \end{cases} \quad \text{and} \quad \overline{Z}_i = \begin{cases} \overline{Y}_i - b\,\underline{X}_i\,, & b > 0 \\ \overline{Y}_i - b\,\overline{X}_i\,, & b \leq 0 \end{cases}.$$

Then finding the thinnest band containing (at least) $\overline{k}$ of the imprecise data V_i^* corresponds to finding the shortest interval containing (at least) $\overline{k}$ of the transformed imprecise data Z_i^*. Since the bands $\overline{B}_{f,q}$ are symmetric around f, the optimal intercept for a fixed candidate slope is given by the center of the shortest interval containing (at least) $\overline{k}$ of the transformed imprecise data Z_i^*. It can be proved that this shortest interval is one of the $n-\overline{k}+1$ intervals going from the j-th ordered lower endpoint $\underline{Z}_{(j)}$ to the $\overline{k}$-th of those ordered upper endpoints whose corresponding lower endpoints are not smaller than $\underline{Z}_{(j)}$, for $j = 1, \ldots, n - \overline{k} + 1$. The interval with the shortest length provides the optimal intercept by its midpoint and the corresponding bandwidth by its length.

In this way, we obtain for each of the candidate slopes the associated optimal intercept and the resulting upper endpoint of the confidence interval, which corresponds to half of the width of the associated closed band. The function f_{LRM} is then given by the function with the minimal upper endpoint.

3.2 Part 2: Identifying the set of all undominated lines

Once f_{LRM} and the associated $\overline{q}_{LRM}$ are known, the actual result of the LIR analysis is determined, which is the set of all regression lines that are

not strictly dominated by f_{LRM}. For each $b \in \mathbb{R}$ there is a (possibly empty) set $\mathcal{A}_b$ consisting of all intercept values a such that the function $f_{a,b}$ is not strictly dominated by f_{LRM}.

To determine $\mathcal{A}_b$, we make use of the fact that the closed band of width $2\,\overline{q}_{LRM}$ around an undominated regression line intersects at least $\underline{k}+1$ imprecise data. Consider again the transformed data Z_i^*, then finding the centers of all bands of width $2\,\overline{q}_{LRM}$ that intersect (at least) $\underline{k}+1$ of the imprecise data V_i^* reduces to finding the centers of all intervals (of length $2\,\overline{q}_{LRM}$) that intersect (at least) $\underline{k}+1$ of the transformed imprecise data Z_i^*. Thus, for each b we look for the values a such that the intervals $[a-\overline{q}_{LRM}, a+\overline{q}_{LRM}]$ intersect at least $\underline{k}+1$ of the $Z_i^*,\ i=1,\ldots,n$. For each subset of $\underline{k}+1$ transformed imprecise data, $Z_{i_1}^*,\ldots,Z_{i_{\underline{k}+1}}^*$, the set of undominated interval centers is the interval

$$\left[\max_{i\in\{i_1,\ldots,i_{\underline{k}+1}\}} \underline{Z}_i - \overline{q}_{LRM},\ \min_{i\in\{i_1,\ldots,i_{\underline{k}+1}\}} \overline{Z}_i + \overline{q}_{LRM}\right].$$

If the lower interval endpoint exceeds the upper one, the set of undominated interval centers associated with the considered subset of imprecise data is empty. This means that there is no interval of length $2\,\overline{q}_{LRM}$ intersecting all of the considered imprecise data.

Employing this idea, we can prove that for each b the set $\mathcal{A}_b$ can be obtained as the union of the intervals $[\underline{Z}_{(\underline{k}+j)} - \overline{q}_{LRM}, \overline{Z}_{(j)} + \overline{q}_{LRM}],\ j=1,\ldots,n-\underline{k}$, where $\underline{Z}_{(i)}$ and $\overline{Z}_{(i)}$ are the i-th ordered lower and upper endpoints of the imprecise data, respectively. Finally, the whole set of parameters (a,b) identifying the undominated functions is given by the union of the sets $\mathcal{A}_b \times \{b\}$ over all $b \in \mathbb{R}$. It can be shown that this set is polygonal, but it is not necessarily convex nor connected. Figure 4 shows the complex shape of this set in our example and in Figure 5 the corresponding regression functions are plotted.

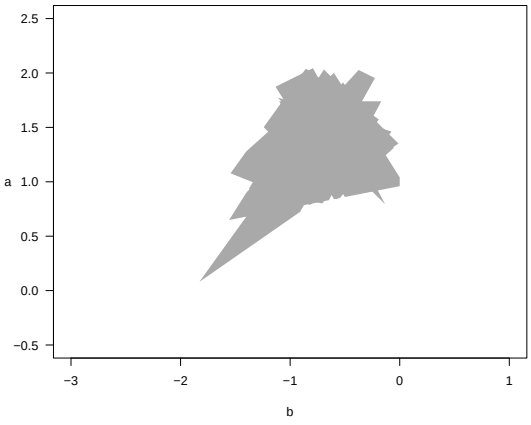

Fig. 4 Set of parameters corresponding to the set of undominated regression lines for the LIR analysis considered in Figure 2.

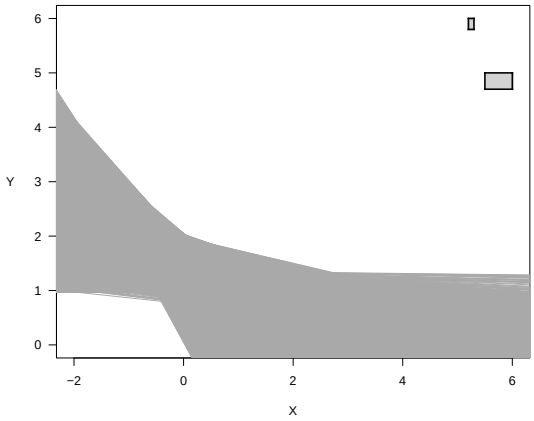

Fig. 5 Set of undominated regression functions for the LIR analysis considered in Figure 2.

4 Conclusions

We presented an algorithm to determine the set-valued result of a robust LIR analysis in the case of simple linear regression with interval data. The algorithm is directly derived from the geometrical properties of the LIR results and it is exact. The proofs will be given in an extended version of this paper. The presented algorithm can be seen as a generalization of an algorithm developed for the least median of squares regression (see [5, 6]), from which it inherits the computational complexity $O(n^3 \log n)$. The algorithm can be further generalized to multiple regression and to other kinds of imprecise data.

So far, we have implemented the algorithm as a general function using the statistical software environment R (see [4]). In future work, we intend to set up an R package for linear regression with LIR.

References

1. Cattaneo M (2007) Statistical Decisions Based Directly on the Likelihood Function. PhD Thesis, ETH Zurich
2. Cattaneo M, Wiencierz A (2011) Robust regression with imprecise data. Technical Report 114, Department of Statistics, LMU Munich
3. Cattaneo M, Wiencierz A (2012) Likelihood-based Imprecise Regression. Accepted for publication in Int J Approx Reason. A preliminary version of the paper is available as Technical Report 116, Department of Statistics, LMU Munich
4. R Development Core Team (2012) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing
5. Rousseeuw PJ, Leroy AM (1987) Robust Regression and Outlier Detection. Wiley
6. Steele JM, Steiger WL (1986) Algorithms and complexity for least median of squares regression. Discret Appl Math 14:93–100