LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Marco E. G. V. Cattaneo and Andrea Wiencierz

Robust regression with imprecise data

Technical Report Number 114, 2011
Department of Statistics
University of Munich

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

Robust regression with imprecise data

Marco E. G. V. Cattaneo, Andrea Wiencierz

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

Abstract

We consider the problem of regression analysis with imprecise data. By imprecise data we mean imprecise observations of precise quantities in the form of sets of values. In this paper, we explore a recently introduced likelihood-based approach to regression with such data. The approach is very general, since it covers all kinds of imprecise data (i.e. not only intervals) and it is not restricted to linear regression. Its result consists of a set of functions, reflecting the entire uncertainty of the regression problem. Here we study in particular a robust special case of the likelihood-based imprecise regression, which can be interpreted as a generalization of the method of least median of squares. Moreover, we apply it to data from a social survey, and compare it with other approaches to regression with imprecise data. It turns out that the likelihood-based approach is the most generally applicable one and is the only approach accounting for multiple sources of uncertainty at the same time.

Keywords: robust regression, imprecise data, nonparametric statistics, likelihood inference, statistical uncertainty, indetermination

1. Introduction

Regression analysis is perhaps the most widely used method of statistical data analysis. The goal of a regression analysis is to learn the relationship between some variables from data. The relationship is typically modeled in expressing one of the quantities (so-called dependent variable) as a function of the other ones (so-called independent variables). Regression analyses are mainly conducted for two major purposes: to predict future observations of the dependent variable or to explain the relationship between the analyzed variables within the population from which the observed sample was drawn. There is a large variety of specific regression methods that can be applied to many different problems of prediction and explanation. Most of these statistical methods are based on the (implicit or explicit) assumption that the data are precise and correct measurements of the quantities of interest. However, this is hardly ever the case. For example, if data are obtained by questionnaires there are many different sources of errors or biases in the data: for instance, for numerical quantities there may be measurement errors due to rounding behavior, and for categorical variables there may be multiple observed categories if respondents cannot decide for one of the categories (see for instance [3]). Some of those errors, like the two examples mentioned, may be accounted for by considering subsets of the observation space, rather than precise values, as data. Similarly, technical measuring instruments usually provide not only a precise measurement but also an assessment of the measurement uncertainty, which translates the measured value into an interval of possible values. The measurement uncertainty should not be ignored in the statistical analysis, therefore, the entire interval of possible values of the measurement could be considered as the observation rather than the precise but possibly incorrect value. Moreover, all continuous variables are generally measured only with a limited precision depending on the number of reported significant digits. That is, we only know that the true value lies in a small interval around the measured value. Thus, data often contain only the information that the (true precise) values of the quantities of interest are in certain subsets of the observation space. We call data that are in this sense imprecise observations (of precise quantities) imprecise data.

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

How can we do regression analysis with imprecisely observed quantities? An intuitive, ad hoc solution to this problem consists in reducing it to one or more regression problems with precise data, to which the usual regression methods can be applied. In particular, we can apply the usual regression methods to all possible precise data sets compatible with the imprecise observations (that is, all precise data sets where each value lies in the corresponding observed set). This idea was used in [1] to study the stability of the regression results obtained without considering the limited precision of the measurements of continuous variables. With this approach, we obtain a set of precise regressions, meaning the set of the (precise) results of the regressions with all possible precise data sets compatible with the imprecise observations. We can consider this set of precise regressions as the imprecise result of the regression with imprecise data: this idea was studied extensively in [16], and for some special cases was already considered in [32, 19, 27]. A closely related idea was studied in more detail in [44].

Another, simpler way of reducing the imprecise data to precise ones (in order to apply the usual regression methods to them) is to represent the imprecise observations by some precise values. In particular, (bounded) interval observations can be represented by their midpoints and lengths. Midpoint regression is very often implicitly or explicitly used with interval data, although it is well known that it can lead to severe error [40, 1, 11]. The literature on regression with interval observations is vast, and it is important to distinguish between authors who interpret them as precise observations of interval-valued quantities, and authors who interpret them as imprecise observations of precise values (as we do in this paper). With the former interpretation we obtain in fact a different problem, and it is difficult in the present framework (based on the latter interpretation) to evaluate the proposed solutions [12, 4, 5, 18, 17, 34, 29, 13, 15, 14, 6, 7]. However, in most of these works the regression with interval data reduces in practice to two regressions with precise data: one for the midpoints of the intervals, and the other for their lengths. When the interval observations are interpreted as in the present paper (that is, as imprecise observations of precise values), the regression problem is complicated by the assumptions about the "coarsening process" connecting the unobserved precise data with the observed interval data. For several special cases, regression methods based on maximum likelihood estimation and its variations have been proposed [11, 30, 28, 20, 10, 43].

In [9] we have introduced a new likelihood-based approach to regression with imprecisely observed data. We call it Likelihood-based Imprecise Regression (LIR): the regression is imprecise in the sense that the result is usually a set of regression functions, instead of a single function. The idea of allowing imprecise results is shared by several other approaches to regression with imprecise data [31, 32, 19, 27, 44, 34, 16, 39, 36]. In our approach the imprecision of the results reflects two kinds of uncertainty: the statistical uncertainty (due to the finite sample) and the indetermination (due to the imprecision of the data). These two kinds of uncertainty in the imprecise results are discerned for example also in [31, 44]. In [9] we have considered in particular the case in which nothing is assumed about the distribution of the (unobserved) precise data or about the "coarsening process". With such extremely weak assumptions, LIR leads to a very robust regression method. In the present paper, we study its most important special case in much more detail, and compare it to alternative methods of regression with imprecise data.

The paper is organized as follows. In Section 2 we formalize the problem of regression with imprecise data and two of its possible solutions. Our robust regression method is presented in Sections 3, 4, and 5: it is first formulated mathematically, then illustrated with a simple example, and finally applied to real data from a social survey. In Section 6 the method is compared to the alternative solutions presented in Section 2, going back to the simple example studied in Section 4. Finally, Section 7 gives some conclusions and outlooks.

2. Regression problem with imprecise data

In regression analysis, we investigate the relationship between a dependent variable Y and one or more independent variables X. We assume that Y is a real number, while X can take values in any set $\mathcal{X}$. In particular, if we have d independent variables, then X is a d-tuple and $\mathcal{X}$ is the Cartesian product of d sets (for instance, $\mathcal{X} = \mathbb{R}^d$). The goal of the regression analysis is to describe the relationship between Y and X by means of some function $f : \mathcal{X} \to \mathbb{R}$, chosen from a particular set $\mathcal{F}$ of possible functions. For example, the regression is called linear when $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{F}$ is the set of all affine functions $f : \mathbb{R}^d \to \mathbb{R}$.

The possible regression functions $f \in \mathcal{F}$ are compared on the basis of some data points $V_1, \ldots, V_n$, where each data point corresponds to a realization of the independent and dependent variables: $V_i = (X_i, Y_i)$. The regression problem consists in trying to identify the function $f \in \mathcal{F}$ minimizing in some sense the absolute residuals $R_{f,i} := |Y_i - f(X_i)|$. For example, the choice of the function minimizing the sum of the absolute residuals (that is, $\arg\min_{f\in\mathcal{F}} \sum_{i=1}^{n} R_{f,i}$)

seems to have been the first solution to the regression problem (see for instance [41, 21]). The most common solution is of course the method of least squares: the choice of the function minimizing the sum of the squares of the (absolute) residuals (that is, $\arg\min_{f\in\mathcal{F}} \sum_{i=1}^{n} R_{f,i}^2$). However, these two methods are not robust: they are very sensitive to outliers (including leverage points). In fact, the breakdown point of these two methods is 0, which means that they cannot even handle a single outlier (see for example [22, 24, 38, 33]).

The method of least squares can also be interpreted as the choice of the function minimizing the mean of the squared residuals (that is, $\arg\min_{f\in\mathcal{F}} {}^1\!/_n \sum_{i=1}^{n} R_{f,i}^2$). The mean is a nonrobust estimator of location: a very robust alternative is the median. If we replace the mean by the median in the previous interpretation of the method of least squares, we obtain a very robust alternative: the so called method of least median of squares (that is, the choice of $\arg\min_{f\in\mathcal{F}} med(R_{f,1}^2, \ldots, R_{f,n}^2)$). More generally, for some $p \in (0, 1)$, we can use the p-quantile instead of the median (which corresponds to the ${}^1\!/_2$-quantile): we obtain the so called method of least quantile of squares. The nearer p is to ${}^1\!/_2$, the more robust is this method. In fact, its breakdown point is $\min\{p, 1 - p\}$: the highest possible breakdown point ${}^1\!/_2$ is reached for $p = {}^1\!/_2$ (that is, for the method of least median of squares [38]). Of course, these two methods involving quantiles of the squared residuals are left essentially unchanged when we replace the squared residuals $R_{f,i}^2$ by the absolute residuals $R_{f,i}$ (the possible differences depend on the exact definition of quantile: that is, on the interpolation scheme used; see for instance [25]).

In this paper, we consider the regression problem when the data are only imprecisely observed, in the following sense: Instead of a data point $V_i \in \mathcal{X} \times \mathbb{R}$, we observe only a subset $V_i^* \subseteq \mathcal{X} \times \mathbb{R}$, whose interpretation is that we only know that $V_i \in V_i^*$. As extreme cases we have the observations of a singleton $V_i^* = \{V_i\}$ and of the whole sample space $V_i^* = \mathcal{X} \times \mathbb{R}$. In the former case the data point V_i is in fact precisely observed, while in the latter case V_i is in fact a missing data point (since we have learned nothing about it by observing V_i^*). Between these two extreme cases, there is of course a large variety of possible kinds of imprecise observations. Of particular importance in applications is the case in which the imprecise observations V_i^* are (multidimensional) intervals: that is, the data are interval-censored (see for instance [26]). By including degenerate and unbounded intervals, this case covers all situations in which each one of the independent and dependent variables are either precisely observed, interval-censored, or missing. In all the examples of the present paper, we shall in fact consider this kind of imprecise data, but our approach can be applied to any kind of imprecise observations V_i^*. Moreover, in our approach we can also allow the imprecise data to be wrong (with a bounded probability): that is, it is sometimes possible that $V_i \notin V_i^*$. In order to precisely formulate this assumption, we need a probabilistic model, which will be introduced in the next section.

A simple example of (artificial) imprecise data is displayed in Figure 1. In this example, $\mathcal{X} = \mathbb{R}$, and the imprecise observations $V_i^* \subseteq \mathbb{R}^2$ are (possibly degenerate) two-dimensional intervals (i.e., rectangles). Hence, we can write $V_i^* = X_i^* \times Y_i^*$ and interpret the intervals X_i^* and Y_i^* as the imprecise observations of X_i and Y_i, respectively. The example data set contains 17 observations with varying amounts of imprecision: there is one actually precisely observed data point $V_i = (1, 1)$ (that is, $X_i^* = Y_i^* = \{1\}$), there are two line segments (which are imprecise observations where either X_i^* or Y_i^* is a singleton), and, finally, there are 14 rectangles of different sizes and shapes. This example data set will be used in Sections 4 and 6 to illustrate our approach and compare it with other approaches to regression with imprecise data.

In general, when the data points V_i are only imprecisely observed, the absolute residuals $R_{f,i}$ are imprecisely observed as well (for each function $f \in \mathcal{F}$), and therefore the above mentioned regression methods cannot be directly applied. As noted in Section 1, an intuitive, ad hoc solution to this problem consists in reducing the imprecise data set to one or more precise data sets, to which the usual regression methods can be applied. We mentioned in particular the set of precise regressions approach and the midpoint regression approach: both ideas are usually considered in connection with the method of least squares. The imprecise result of the former approach consists then of the least squares regression functions corresponding to all possible precise data sets $v_1, \ldots, v_n$ with $v_i \in V_i^*$. When the imprecise observations V_i^* are (multidimensional) bounded intervals, it is possible to define their midpoint $mid(V_i^*)$. In this case, the (precise) result of the latter approach is the least squares regression function corresponding to the precise data set $v_1, \ldots, v_n$ with $v_i = mid(V_i^*)$. It is important to note that the midpoint regression approach cannot be applied when some imprecise observations V_i^* are unbounded (as for example in the case of missing data). In this case, the set of precise regressions approach would also break down, but the problem could be resolved by using for instance the method of least median (or quantile) of squares instead of the method of least squares.

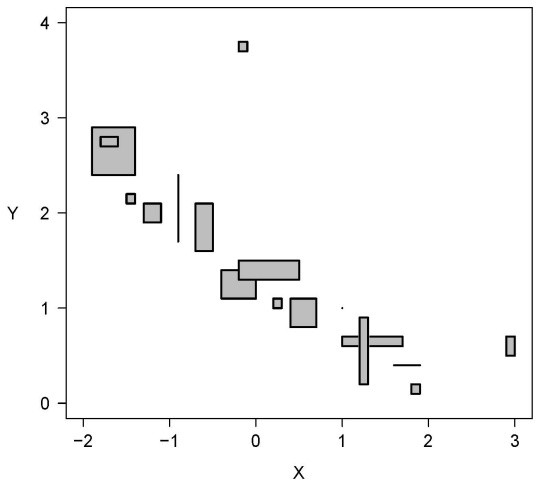

Figure 1: Example data set consisting of 17 imprecise observations $V_i^* \subseteq \mathbb{R}^2$.

3. Likelihood-Based Imprecise Regression (LIR)

LIR is a new approach to regression, directly applicable to the case of imprecise data. The regression is imprecise, in the sense that in general the result is a set of regression functions (that is, a subset of $\mathcal{F}$), whose extent reflects the total amount of uncertainty (about which $f \in \mathcal{F}$ best describes the relationship between the dependent and independent variables). The approach is likelihood-based, which means that it relies on a probabilistic model, and the data are used only through the induced likelihood function. In particular, the robustness of the resulting regression method depends on the strength of the assumptions in the probabilistic model: the weaker the assumptions, the more robust the method. In the present paper we study a very robust regression method, resulting from the following nonparametric probabilistic model.

We assume that the pairs $(V_1, V_1^*), \ldots, (V_n, V_n^*)$ are n independent realizations from a joint probability distribution for the precise (unobserved) data $V_i \in \mathcal{X} \times \mathbb{R}$ and the imprecise (observed) data $V_i^* \subseteq \mathcal{X} \times \mathbb{R}$. The only assumption about this joint distribution is that

$$P(V_i \in V_i^*) \geq 1 - \varepsilon, \tag{1}$$

where $\varepsilon \in [0, 1/2)$ is a fixed bound on the probability of error for the imprecise observations (that is, the probability that $V_i \notin V_i^*$). The usual assumption in alternative approaches to regression with imprecise data is $\varepsilon = 0$: the possibility of allowing the imprecise data to be wrong seems to be new. A positive value of ε means that it is considered possible that some (imprecise) data are incorrectly measured or reported, for instance when the precise data are misclassified. An extreme case of incorrect data are imprecise observations that are empty (i.e., $V_i^* = \emptyset$), for example when the reported right endpoint of an interval is lower that the left one: such data are very problematic when $\varepsilon = 0$, but not when $\varepsilon > 0$. Even if the choice of a particular positive value for ε can be difficult in many applications, it can at least be very interesting to study how the regression results change when ε is increased.

The above probabilistic model is completely nonparametric: there is no assumption at all about the distribution of the data points $V_i = (X_i, Y_i)$. In particular, we do not assume that the error $Y_i - f(X_i)$ is independent of X_i (for some theoretically correct function $f \in \mathcal{F}$). Moreover, the only assumption about the "coarsening process" connecting each data point V_i with its imprecise observation V_i^* is the assumption (1). In particular, we do not assume that the data are coarsened at random [23]. It is important to note that an additional assumption that only some imprecise observations are possible (that is, the assumption that $V_i^* \in \mathcal{V}^*$, for some fixed set $\mathcal{V}^*$ of subsets of $\mathcal{X} \times \mathbb{R}$) would not change the regression results, although it would change the probabilistic model.

Assuming the above probabilistic model, it is possible to estimate the quantiles of the distribution of the absolute residuals $R_{f,i}$ on the basis of the imprecise observations $V_1^*, \ldots, V_n^*$ (for each function $f \in \mathcal{F}$). The regression method introduced in [9] basically consists in choosing the regression function $f \in \mathcal{F}$ by minimizing for some $p \in (0, 1)$ the estimate of the p-quantile of the distribution of the absolute residuals $R_{f,i}$. Hence, this method can be interpreted

as a generalization of the method of least quantile of squares (or least quantile of absolute residuals), discussed in Section 2. In the present paper, we consider only the generalization of the method of least median of squares: that is, we set $p = 1/2$, which is the choice of p leading to the most robust regression method.

For each $f \in \mathcal{F}$, let C_f be the likelihood-based confidence interval with cutoff point β for the median of the distribution of the absolute residuals $R_{f,i}$. That is, C_f is the set of all values of the median of $R_{f,i}$ corresponding to probability distributions whose (normalized) likelihood is larger than β. We refer to [9] for a more detailed mathematical description, and to Section 4 for a practical illustration of the method. To obtain the confidence interval C_f, the cutoff point $\beta \in (0, 1)$ must be chosen: this choice can be guided by the following simple connection between β and the confidence level of C_f. When $\varepsilon = 0$, the confidence level is asymptotically at least $F_{\chi^2}(-2 \log \beta)$, where F_{χ^2} is the cumulative distribution function of the chi-square distribution with one degree of freedom (see for example [35]). As shown in [9], the likelihood-based confidence intervals C_f can be easily calculated: the results for the case with $p = 1/2$ can be formulated as follows. Assume that the imprecise observations $V_1^*, \ldots, V_n^*$ are nonempty, and that β and n are sufficiently large to satisfy $\sqrt[n]{\beta} > 1/2 + \varepsilon$ (when these assumptions are not satisfied, the method is still applicable, only the expressions below must be modified). Let $d : (0, 1/2 - \varepsilon) \to (1/2 + \varepsilon, 1)$ be the decreasing bijection defined by

$$d(\gamma) = \left(\frac{1/2 - \varepsilon}{1/2 - \varepsilon - \gamma}\right)^{1/2-\varepsilon-\gamma} \left(\frac{1/2 + \varepsilon}{1/2 + \varepsilon + \gamma}\right)^{1/2+\varepsilon+\gamma}.$$

Hence, $\delta := d^{-1}\left(\sqrt[n]{\beta}\right)$ is well-defined, and it decreases when β or n increase, with $\lim_{\beta\uparrow 1} \delta = \lim_{n\uparrow\infty} \delta = 0$. We can now define the integers

$$\underline{k} := \lfloor (1/2 - \varepsilon - \delta)\, n \rfloor \quad \text{and} \quad \overline{k} := \lceil (1/2 + \varepsilon + \delta)\, n \rceil, \tag{2}$$

which satisfy $0 \le \underline{k} < (1/2 - \varepsilon)\, n \le n/2 \le (1/2 + \varepsilon)\, n < \overline{k} \le n$ and $\underline{k} = n - \overline{k}$ (that is, $\underline{k}$ and $\overline{k}$ are symmetric around $n/2$). Then the confidence interval C_f consists of all $q \in [0, +\infty)$ that are sufficiently large for the closed band (of width $2\,q$ around f)

$$\overline{B}_{f,q} := \{(x, y) \in \mathcal{X} \times \mathbb{R} : |y - f(x)| \le q\}$$

to intersect at least $\underline{k} + 1$ imprecise observations, and sufficiently small for the open band (of width $2\,q$ around f)

$$\underline{B}_{f,q} := \{(x, y) \in \mathcal{X} \times \mathbb{R} : |y - f(x)| < q\}$$

to contain at most $\overline{k} - 1$ imprecise observations. Therefore, the left endpoint $\underline{q}$ of the interval C_f is the largest $q \in [0, +\infty)$ such that $\underline{B}_{f,q}$ intersect at most $\underline{k}$ imprecise data, while the right endpoint $\overline{q}$ is the smallest $q \in [0, +\infty)$ such that $\overline{B}_{f,q}$ contains at least $\overline{k}$ imprecise data. In other words, for each function $f \in \mathcal{F}$ we consider the open and closed bands around f: the widest open band intersecting at most $\underline{k}$ imprecise observations is $\underline{B}_{f,\underline{q}}$, while the thinnest closed band containing at least $\overline{k}$ imprecise observations is $\overline{B}_{f,\overline{q}}$.

In particular, when β is sufficiently large, $\underline{k}$ is the largest integer less than $(1/2 - \varepsilon)\, n$, and $\overline{k}$ is the smallest integer greater than $(1/2 + \varepsilon)\, n$. In this case, for each $f \in \mathcal{F}$, the interval C_f represents the maximum likelihood estimate of the median of the distribution of the absolute residuals $R_{f,i}$ (in the sense that all $q \in C_f$ are maximum likelihood estimates). In general C_f is a proper interval, even when it represents the maximum likelihood estimate, $\varepsilon = 0$, and the data points V_i are in fact precisely observed (i.e., the imprecise observations V_i^* are the singletons $\{V_i\}$). Since in general the maximum likelihood estimate of the median of $R_{f,i}$ is a proper interval, there is no particular reason to prefer it to an interval estimate C_f with a higher confidence level (that is, a lower cutoff point β).

For each $f \in \mathcal{F}$ we thus have an interval estimate C_f of the median of the distribution of the absolute residuals $R_{f,i}$. The goal is to choose the regression function f by minimizing this estimate. In order to obtain a single regression function, we could reduce in some way the interval estimate to a single value, but this would be somewhat unjustified. Instead, the idea of LIR is to consider as the imprecise result of the regression the set of all functions that cannot be excluded on the basis of the likelihood inference. That is, the imprecise result consists of all regression functions $f \in \mathcal{F}$ that are not (strictly) dominated by another function $f' \in \mathcal{F}$, in the sense that there is no $f' \in \mathcal{F}$ such that $q' < q$ for all $q' \in C_{f'}$ and all $q \in C_f$.

The set of all undominated functions (that is, the imprecise result of the regression) has a simple geometrical interpretation. Let $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ be the thinnest band of the form $\overline{B}_{f,q}$ containing at least $\overline{k}$ imprecise observations, for all $f \in \mathcal{F}$ and all $q \in [0, +\infty)$. For simplicity, assume that this band is unique and the interval $C_{f_{LRM}}$ is right-closed. Then

the set of all undominated functions consists of all functions $f \in \mathcal{F}$ such that $\overline{B}_{f,\overline{q}_{LRM}}$ (that is, the closed band of width $2\,\overline{q}_{LRM}$ around f) intersects at least $\underline{k}+1$ imprecise observations. In other words, to determine the set of undominated functions, we first look for the thinnest band containing at least $\overline{k}$ imprecise data, and then consider all bands of the same width intersecting at least $\underline{k}+1$ imprecise data: the "centers" of these bands are the undominated functions. This procedure is exemplified in the next section: see in particular Figure 2.

The "center" f_{LRM} of the thinnest band containing at least $\overline{k}$ imprecise data is the regression function obtained by minimizing the right endpoint $\overline{q}$ of the interval estimate C_f (that is, $\arg\min_{f\in\mathcal{F}} \sup C_f$). Hence, f_{LRM} is a minimax solution to the problem described by the likelihood-based confidence intervals C_f: in fact, LRM means "Likelihood-based Region Minimax" [8]. When the data points V_i are precisely observed (i.e., the imprecise observations V_i^* are the singletons $\{V_i\}$), the absolute residuals $R_{f,i}$ are precisely observed as well (for each function $f \in \mathcal{F}$), and from the above considerations it follows that $\overline{q}$ is simply the $\overline{k}$-th ordered absolute residual $R_{f,(\overline{k})}$. Therefore, in this case the regression function f_{LRM} corresponds to the (precise) result of the method of least quantile of squares, when the $\overline{k}/n$-quantile is considered (and the inverse of the empirical distribution function is used as definition of sample quantiles: see for instance [25]). In particular, if $\varepsilon = 0$ and β is sufficiently large for C_f to represent the maximum likelihood estimate, then f_{LRM} corresponds to the result of the method of least median of squares. Hence, in general f_{LRM} can be seen as the result of a generalization of the method of least median (or quantile) of squares to the case of imprecise data. This is of great practical importance, because we can calculate f_{LRM} by adapting to the case of imprecise data the algorithms for the method of least median (or quantile) of squares (see for example [2, 38, 45]), but this goes beyond the scope of the present paper.

The extent of the imprecise result of the regression (that is, of the set of all undominated functions) reflects the total amount of uncertainty about which function $f \in \mathcal{F}$ best describes the relationship between the dependent and independent variables. This total uncertainty consists of two different kinds of uncertainty. On the one hand we have the statistical uncertainty, which is due to the fact that we have only n (imprecise) observations: it decreases when n increases. On the other hand we have the indetermination, which is due to the fact that the n observations are imprecise: this kind of uncertainty is unavoidable under our very weak assumption (1). In the above geometrical interpretation of the imprecise result of the regression, the statistical uncertainty is reflected by the spread between $(\underline{k}+1)/n$ and $\overline{k}/n$, while the indetermination is reflected by the difference between "containing" and "intersecting" the imprecise observations. When the observations are in fact precise, the result of the regression is in general still imprecise, but its extent reflects only the statistical uncertainty. By contrast, with the usual choices of $\varepsilon = 0$ and maximum likelihood estimation (that is, β sufficiently large), the extent of the imprecise result reflects only the indetermination. In this case, very many imprecise data always give an imprecise result, while very few precise data always give a precise result. This is unsatisfactory for a statistical analysis, because in general very few precise data contain less information than very many imprecise data: this topic will be further discussed in Section 6.

4. Illustration of the regression method

In this section, some features of the presented LIR method are studied and illustrated. For this purpose, we consider a simple linear regression problem with the example data set of Section 2. As the majority of the data indicate a (decreasing) linear relationship, we consider the set $\mathcal{F} = \{f_{a,b} : \ f_{a,b}(X) = a + b\,X,\ (a,b) \in \mathbb{R}^2\}$ as the set of possible regression functions. Note, however, that within the LIR framework, it is possible to consider arbitrary functions to describe the relationship between the analyzed variables, i.e. LIR is not restricted to linear regression.

Then, for some choice of β and ε, our regression method consists in finding the function f_{LRM} minimizing the right endpoint of the interval estimate of the median of the absolute residuals and then identifying all regression functions that are not (strictly) dominated (in the above described sense) by f_{LRM}. For the present linear regression problem, we have implemented this search as a random search over the parameter space of (a, b), namely $\mathbb{R}^2$. All computations and graphs are made within the statistical software environment R [37].

At first, we determine the integers $\underline{k}$ and $\overline{k}$, on the basis of which the endpoints of the confidence intervals for the median of the absolute residuals can be computed. Since the absolute residuals are imprecise in the case of imprecise data, for each observation and each function $f \in \mathcal{F}$, the corresponding infimum and supremum of the absolute residuals are considered. Then, for any function $f \in \mathcal{F}$, the left endpoint q of the confidence interval C_f is given by the $(\underline{k}+1)$-th of the ordered infima of the residuals, while the right endpoint $\overline{q}$ is given by the $\overline{k}$-th ordered supremum. This way, we determine the confidence intervals for a large set of possible regression functions randomly

chosen from $\mathcal{F}$ and identify the one whose confidence interval has the smallest right endpoint. Finally, we determine the set of undominated regression functions in comparing the left endpoints of the other confidence intervals with this smallest right endpoint $\overline{q}_{LRM}$. Figure 2 shows the results obtained in the regression analysis of the example data set, for the choice $\beta = 0.5$ and $\varepsilon = 0$. In the following, we examine how different choices of β and ε affect the regression's result.

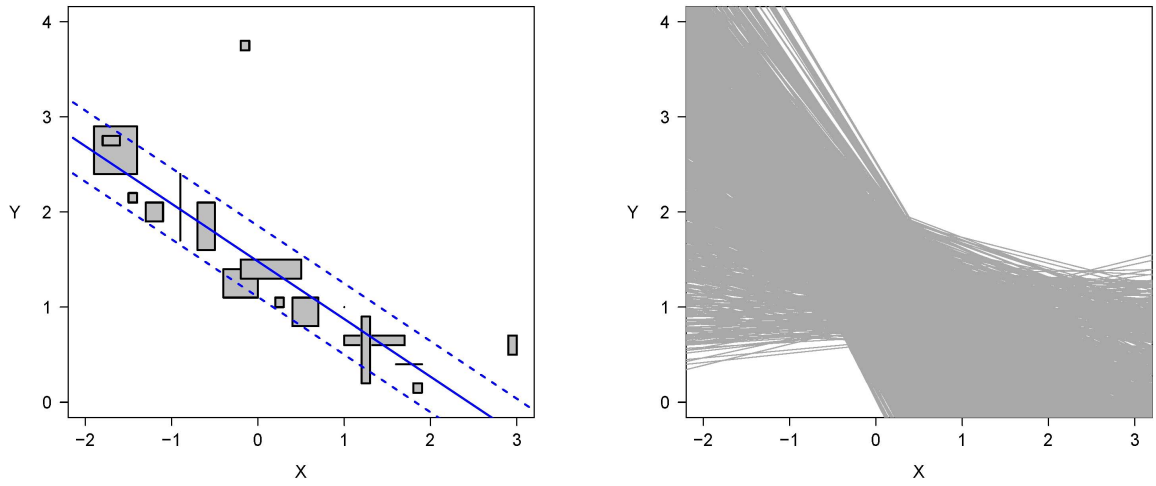

Figure 2: Function f_{LRM} (left, solid line) with closed band $\overline{B}_{f_{LRM},\overline{q}_{LRM}}$ (left, dashed lines) and set of undominated regression functions (right), for $\beta = 0.5$ and $\varepsilon = 0$.

Different choices of $\beta \in (0, 1)$ imply different confidence levels of the interval estimates C_f. For example, when $\varepsilon = 0$, a likelihood-based confidence interval with cutoff point $\beta = 0.5$ would be a conservative 76% confidence interval for the median of the (precise) residuals, given the imprecise data. It corresponds to the set of median values that would not be rejected by a corresponding likelihood ratio test at level 14%. Therefore, β determines how much of the statistical uncertainty is accounted for in the result of the LIR analysis. A high confidence level of the interval estimates of the median of the absolute residuals requires a small choice of β and vice versa. Thus, the higher β, the lower the confidence level and consequently the narrower the set of undominated regression functions. In Figure 3 different results of LIR analyses with other choices of β are displayed. For a low cutoff point such as $\beta = 0.15$, the regression's result is very imprecise, admitting different directions of the relationship between the analyzed variables. In contrast to that, a high cutoff point such as $\beta = 0.8$ leads to a less imprecise result containing practically only decreasing lines. Thus, there is a trade off between confidence in the result and inferential strength of the result. In a practical setting, the two characteristics have to be balanced in light of the analyzed question and the purpose of the analysis in order to choose an appropriate value for β.

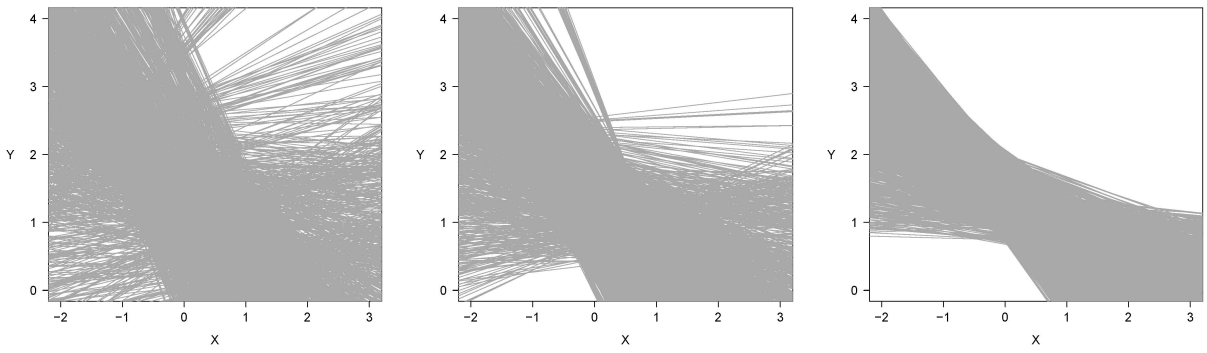

Figure 3: Sets of undominated regression functions for $\varepsilon = 0$ and three different choices of β, namely $\beta = 0.15$ (left), $\beta = 0.26$ (middle), and $\beta = 0.8$ (right), corresponding to the confidence levels 95%, 90%, and 50%, respectively.

In addition to the cutoff point for the likelihood also ε has to be chosen. In the probability model underlying the presented LIR method, ε is the upper bound to the probability that an imprecise observation does not contain the correct precise value. Usually it is assumed that $\varepsilon = 0$, but there might be situations in which the analyst has concerns about the correctness of the imprecise data. For example, as mentioned in the introduction, there are various sources for biases in survey data which should be accounted for in the analysis of such data, at least with a small probability. Hence, the consideration of an $\varepsilon > 0$ means to account for some more uncertainty about the data in addition to the indetermination issuing from the coarseness of the data. It follows directly from definition (2) that increasing ε has the same effect on the width of the confidence intervals as decreasing β, since in both cases, $\underline{k}$ decreases and $\overline{k}$ increases. Thus, the worse the assumed data quality, the more imprecise the result. If we set $\beta = 0.5$ in our example and assume $\varepsilon = 0.01$, we obtain the same result as for $\beta = 0.26$ assuming $\varepsilon = 0$, shown in Figure 3. However, the interpretation is different: Whereas in the case of increasing β for a fixed ε the amount of statistical uncertainty reflected by the result is reduced, in the case of increasing ε for a fixed β the assumptions of the underlying nonparametric probability model are weakened.

Both aspects of the uncertainty of a statistical analysis of imprecise data — statistical uncertainty and indetermination — are crucial and should be reflected in the result. Within the LIR framework, both parts of the uncertainty are expressed in the same way, that is, they determine the extent of the imprecise result of the regression analysis. So far, we have seen how different choices of the confidence level and different assumptions about the correctness of the (imprecise) data are reflected in the regression's result. In order to illustrate how varying degrees of imprecision of the data are represented in the result of a LIR analysis, we consider an application example in the following section.

5. Application to social survey data

One interesting question in the social sciences is how the individual income evolves with age. Yet, analyzing this question on the basis of data is a particular challenge. Personal income is a very sensitive information that respondents in a social survey often give only imprecisely, incorrectly, or do not give at all. Furthermore, it is a common practice in surveys to gather data about continuous variables in a discretized way, i.e. with predefined answer categories that form a partition of the possible range of the variables. Finally, also for privacy reasons, personal data might be available in categories only. Thus, the available data are often imprecise, which makes the presented LIR method perfectly suitable to investigate this question.

Here, we analyze data from the "Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) — German General Social Survey" of 2008, provided by GESIS — Leibniz Institute for the Social Sciences, to investigate the above question. In this survey, 3 469 persons have been interviewed. The data set contains the age of the respondents as a precise variable (in years) as well as categorized into six age classes. In both cases, there are 12 missing values. For the income data there are also two variables: a precise variable with the net personal income (on average per month in euros) and a discrete variable with 22 possible income classes. Here, 1 063 respondents did not give their precise income, 381 of which however gave their income class, while 682 values are completely missing. (Details on the data set can be found in [42].)

Thanks to the multiple information, the ALLBUS data set allows us to consider situations with different degrees of imprecision of the data. First, we analyze a data set composed of the most precise available information: for *income* there are the available 2 406 precise values together with the 381 income classes for which no precise value is available, and *age* is given by intervals $X_i^* = [age, age + 1)$, because the "precise" age of a person (in years) actually contains this imprecise information. We then consider the more imprecise case where the income data is only available in income classes, while the age data is again almost precise. Finally, we analyze the most imprecise data set where both, *age* and *income*, are categorized. The missing values are always replaced by the entire observation space as imprecise observation, i.e., $X_i^* = [18, 100)$ and $Y_i^* = [0, +\infty)$, respectively.

For each of the three data situations, we conduct a LIR analysis. As the relationship between *income* and *age* is usually modeled by a quadratic function in *age*, we consider the following set of possible regression functions $\mathcal{F} = \{f_{a,b_1,b_2} : f_{a,b_1,b_2}(X) = a + b_1 X + b_2 X^2,\ (a, b_1, b_2) \in \mathbb{R}^3\}$. Here, we conduct the regression analyses by means of a grid search over the parameter space $\mathbb{R}^3$. We set $\beta = 0.15$, aiming at very reliable results, and $\varepsilon = 0$, assuming that the data are correct. Although the latter assumption is highly questionable as regards the variable *income*, we leave these concerns aside here, since we are mainly interested in the impact of the indetermination issuing from the coarseness of the observations.

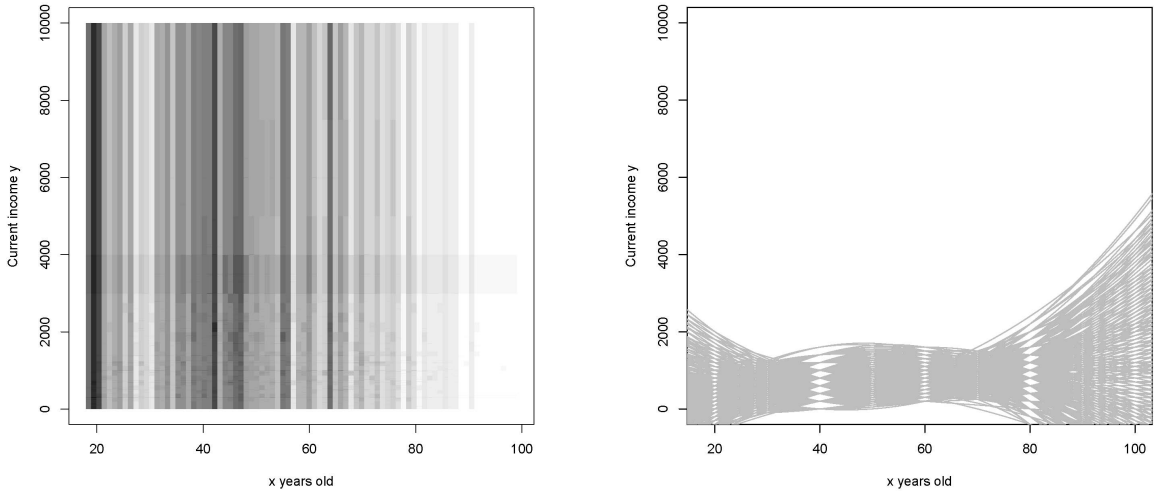

Figure 4: Regression analysis of the ALLBUS data set with *age* in one-year intervals and *income* partly in classes.

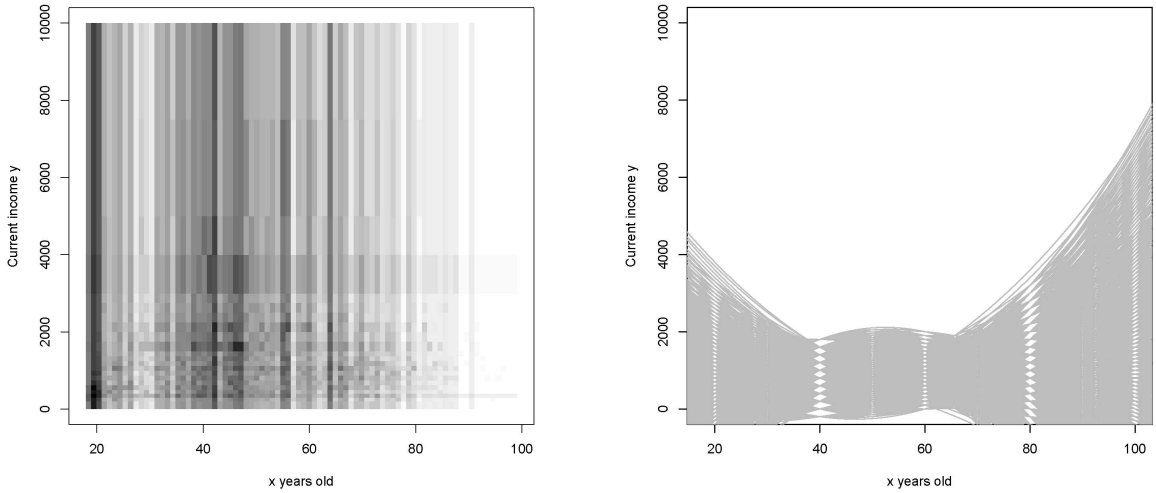

Figure 5: Regression analysis of the ALLBUS data set with *age* in one-year intervals and *income* in classes.

The results of the regressions are displayed in Figures 4, 5, and 6. Each of these shows a two-dimensional histogram of the considered data set (left) and the corresponding set of undominated functions obtained from the LIR analysis (right). All results are fairly imprecise admitting all kinds of shapes for the *age-income* profile. For example, a decreasing line or a concave parabolic curve are both plausible descriptions of the relationship between *income* and *age*. From a social scientist's viewpoint this result might be unsatisfactory even though it is almost sure that the true curve is included in the sets of undominated regression functions due to the low choice of the cutoff point here. However, the imprecise result also reflects the indetermination induced by the data. The regression analysis based on the data set with the most precise information clearly yields the smallest set of undominated functions (Figure 4), whereas the extent of the result for the most imprecise data set is visibly the largest (Figure 6). Thus, the extent of the imprecise result of a LIR analysis is strongly influenced by the degree of imprecision of the data.

Besides serving as an illustration for the effect of the coarseness of the data on the regression's result, this application example further reveals an advantage of LIR over other approaches to regression with interval data. In the ALLBUS data set, the upper income class is $[7\,500, +\infty)$. If a regression method based on the interval midpoints shall be applied to analyze the *age-income* profile, an upper endpoint for this income class has to be fixed in order to determine the midpoint of this interval. The choice of the upper endpoint may have a strong impact on the result, for

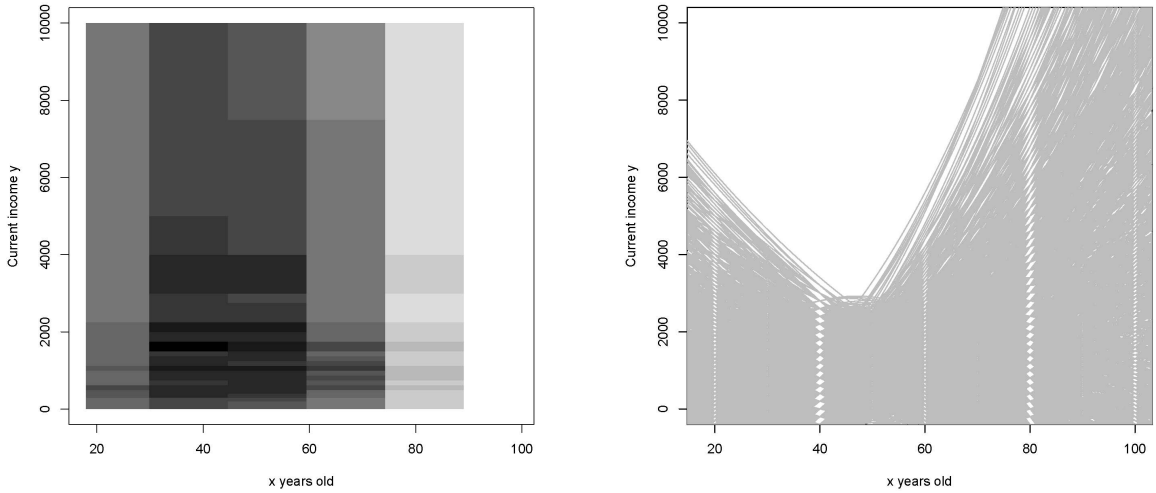

Figure 6: Regression analysis of the ALLBUS data set with both, *age* and *income*, in classes.

example, if least squares estimation is used. For the ALLBUS data set, the effect of choosing different bounds for the upper income class in a least squares regression based on the interval midpoints was studied in [9]. In contrast to that, in the LIR framework it is not necessary to fix an upper endpoint of the income range, i.e. the results are independent of the upper limit, as long as there are at least $\overline{k}$ bounded data. If there are missing values in the data set, the application of a midpoint regression with an (arbitrarily) chosen upper limit is even more questionable, because then the midpoint of the (arbitrarily) bounded range is assigned to the missing values. This is especially problematic in the application example above, where missing income values are much more likely to be values at the margins of the range, that is, particularly high or low incomes. In the LIR approach, the missing values are replaced by the (possibly unbounded) entire observation space, and thus this bias towards the center is avoided.

6. Comparison of LIR with other approaches to regression with imprecise data

Two other types of regression methods for imprecise data were mentioned in Sections 1 and 2: regression methods based (among other things) on interval midpoints and those considering the set of all precise regressions compatible with the imprecise data. In this section, these two approaches are compared with LIR. In the beginning, the general case of regression with imprecise data is discussed, before the special case of precise observations is considered.

In the previous sections, two drawbacks of the midpoint approach to regression with imprecise data were already pointed out. First, reducing interval data to their midpoints and then applying precise regression methods to the midpoint data ignores the uncertainty induced by the coarseness of the observations and generally leads to biased results [40, 1, 11]. The second drawback is that it is impossible to handle unbounded intervals. Furthermore, this approach is restricted to interval data, whereas the LIR framework allows the imprecise observations to be any possible subset of the observation space. Finally, it is important to mention that a midpoint regression yields a precise result, i.e. a point estimate of the regression function, which does not unveil the statistical uncertainty behind it.

The set of precise regressions approach suggests a pragmatic way to take the imprecision of the data into account, but also ignores the statistical uncertainty of the regression problem. Two applications of this approach to the example data set of Section 2 are displayed in Figure 7. The results are approximated by the precise regression estimates of 100 000 randomly drawn precise data sets compatible with the imprecise data. The plot on the left shows the set of least squares regression lines, whereas the one on the right shows the result of a robust variant of this approach using the least median of squares estimation. Both resulting sets are much narrower than all of the LIR results shown in Section 4. This is due to the fact that the LIR results additionally reflect the statistical uncertainty, which is relatively large here, since the example data set contains only 17 observations. In contrast to that, the set of precise regression lines is in fact a set-valued point estimate, therefore, it does not capture the entire information about the relationship between X and Y that is provided by the observations. In the present example, most of the (imprecise) observations

indicate a decreasing linear relationship, however, there are also two observations further away and the sample size is fairly small. Therefore, we actually know only very little about the relationship of interest and inferences should be drawn carefully, accounting for both sources of uncertainty.

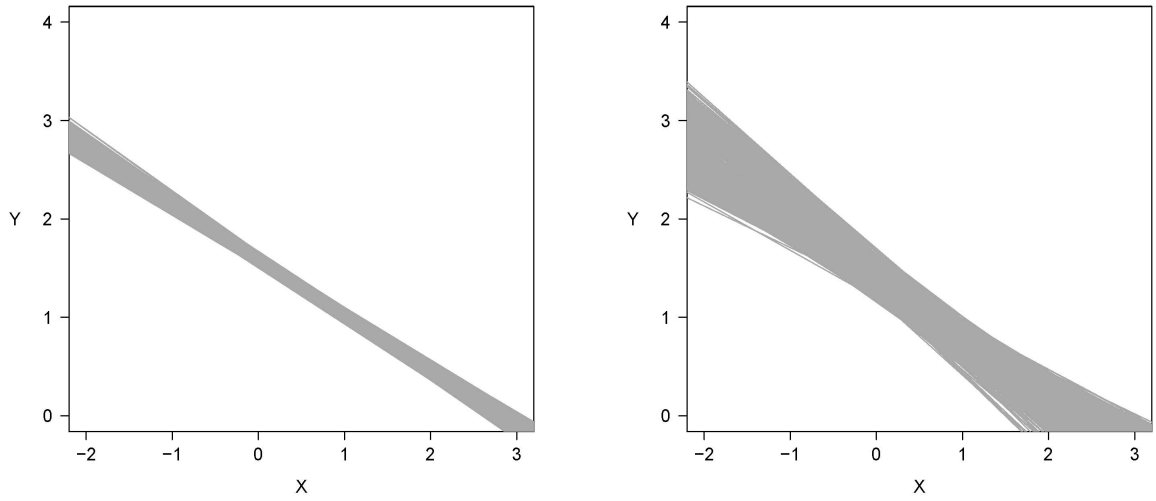

Figure 7: Sets of precise linear regressions compatible with the example data set: least squares estimation (left), and least median of squares estimation (right).

When the data are in fact precise, the set of precise regressions approach always yields a precise result, expressing that there is no uncertainty about the exact data values. In this special case, the simple midpoint regression approach and the set of precise regressions approach lead to the same precise result, if they are based on the same regression method. For example, if they are based on least squares estimation, both methods coincide with the least squares regression. A LIR analysis, by contrast, generally yields an imprecise result, which, in the special case of precise data, reflects only the statistical uncertainty. To illustrate the differences between the approaches in this case, we use the midpoints of the example data set as a precise data set and apply the different regression methods to this data set. In Figure 8, the results of the least squares estimation (left) and of the least median of squares estimation (right) are compared with the set of undominated regression functions obtained by the LIR analysis with $\beta = 0.5$ and $\varepsilon = 0$.

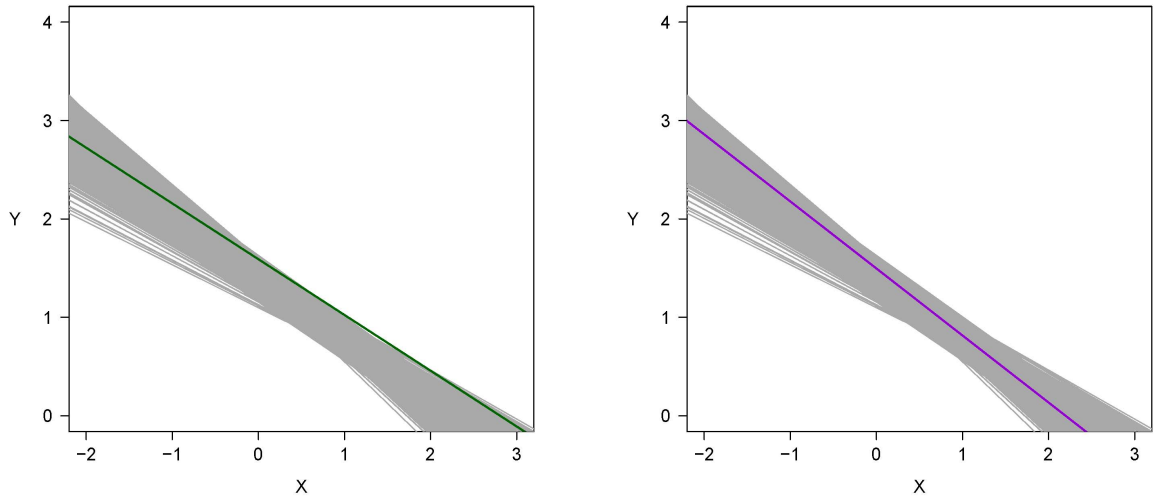

Figure 8: Results of applying the LIR method with $\beta = 0.5$ and $\varepsilon = 0$ (grey lines) versus least squares estimation (left, dark line) and least median of squares estimation (right, dark line) in the case of precise data (consisting of the midpoints of the imprecise example data set).

In standard statistics (implying an ideal precise data situation) it is agreed that obtaining reliable inferences from a point estimate requires an assessment of the statistical uncertainty about this estimate. In this context, such an assessment is often provided in the form of standard errors for the corresponding parameter estimates or (simultaneous) confidence bands for the estimated regression function. Within the frameworks of the other two considered approaches to regression with imprecise data, however, no evaluation of the statistical uncertainty is provided. Neither does the result reflect the statistical uncertainty nor is there a recommendation of how to assess the statistical uncertainty, which means that an estimated regression line based on three observations could be interpreted in the same way as a regression estimate based on 3 000 observations. In contrast to that, in the LIR approach, both aspects of the uncertainty of the regression problem are directly incorporated in the regression method and expressed by the extent of the resulting set of undominated functions, as illustrated in Sections 4 and 5.

One way to account also for the statistical uncertainty in the sets of precise regressions approach could be to consider as its result the set of all functions included in the union of all (simultaneous) confidence bands derived from the precise regressions instead of simply the set of all point estimates. A similar idea was studied in [44], but a thorough analysis of it goes beyond the scope of the present paper.

7. Conclusion

In this paper, we studied a special case of the new LIR approach to regression with imprecise data introduced in [9]. We presented the general framework of LIR and summarized the theoretical results for the special case where we assume only (1) and generalize the least median of squares regression method [38]. Furthermore, some features of the presented LIR method were studied in detail and illustrated with the help of two regression problems: one based on artificial data and the other on real data from a social survey. Finally we compared this method with two other approaches to regression with imprecise data. It turns out that LIR is the most generally applicable approach and is the only one accounting for different sources of uncertainty at the same time.

The result of a LIR analysis is imprecise: it consists of all regression functions that cannot be excluded on the basis of the likelihood inference. Both aspects of the uncertainty, indetermination and statistical uncertainty, affect the extent of the regression's result. That is, the more imprecise the data or the less observations, the larger the resulting set of regression functions. Moreover, the LIR method yields very robust results, because the underlying probability model is completely nonparametric and the regression method is based on sample quantiles of the residuals.

As regards the computational aspect, algorithms for the precise regression method of least median of squares can be adapted to be used for the LIR method. This will be one topic of future work, together with a more thorough evaluation of the confidence level of the imprecise results of LIR analyses.

References

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

[2] T. Bernholt, Computing the least median of squares estimator in time $O(n^d)$, in: Computational Science and Its Applications — ICCSA 2005, Springer, 2005, pp. 697–706.

[3] P.P. Biemer, L.E. Lyberg, Introduction to Survey Quality, Wiley, 2003.

[4] L. Billard, E. Diday, Regression analysis for interval-valued data, in: H.A.L. Kiers, J.P. Rasson, P.J.F. Groenen, M. Schader (Eds.), Data Analysis, Classification, and Related Methods, Springer, 2000, pp. 369–374.

[5] L. Billard, E. Diday, From the statistics of data to the statistics of knowledge: symbolic data analysis, J. Am. Stat. Assoc. 98 (2003) 470–487.

[6] A. Blanco-Fernández, N. Corral, G. González-Rodríguez, Estimation of a flexible simple linear model for interval data based on set arithmetic, Comput. Stat. Data Anal. 55 (2011) 2568–2578.

[7] A. Blanco-Fernández, N. Corral, G. González-Rodríguez, A. Palacio, On some confidence regions to estimate a linear regression model for interval data, in: C. Borgelt, G. González-Rodríguez, W. Trutschnig, M.A. Lubiano, M.A. Gil, P. Grzegorzewski, O. Hryniewicz (Eds.), Combining Soft Computing and Statistical Methods in Data Analysis, Springer, 2010, pp. 33–40.

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

[9] M. Cattaneo, A. Wiencierz, Regression with imprecise data: A robust approach, in: F. Coolen, G. de Cooman, T. Fetz, M. Oberguggenberger (Eds.), ISIPTA '11, Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications, SIPTA, 2011, pp. 119–128.

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

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

[12] P. Diamond, Least squares fitting of compact set-valued data, J. Math. Anal. Appl. 147 (1990) 351–362.

[13] M.A.O. Domingues, R.M.C.R. de Souza, F.J.A. Cysneiros, A robust method for linear regression of symbolic interval data, Pattern Recognit. Lett. 31 (2010) 1991–1996.
[14] M.B. Ferraro, A. Colubi, P. Giordani, A linearity test for a simple regression model with *LR* fuzzy response, in: C. Borgelt, G. González-Rodríguez, W. Trutschnig, M.A. Lubiano, M.A. Gil, P. Grzegorzewski, O. Hryniewicz (Eds.), Combining Soft Computing and Statistical Methods in Data Analysis, Springer, 2010, pp. 263–271.
[15] M.B. Ferraro, R. Coppi, G. González-Rodríguez, A. Colubi, A linear regression model for imprecise response, Int. J. Approx. Reasoning 51 (2010) 759–770.
[16] S. Ferson, V. Kreinovich, J. Hajagos, W. Oberkampf, L. Ginzburg, Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty, Technical Report SAND2007-0939, Sandia National Laboratories, 2007.
[17] M.A. Gil, G. González-Rodríguez, A. Colubi, M. Montenegro, Testing linear independence in linear models with interval-valued data, Comput. Stat. Data Anal. 51 (2007) 3002–3015.
[18] M.A. Gil, M.A. Lubiano, M. Montenegro, M.T. López, Least squares fitting of an affine function and strength of association for interval-valued data, Metrika 56 (2002) 97–111.
[19] F. Gioia, C.N. Lauro, Basic statistical methods for interval data, Ital. J. Appl. Stat. 17 (2005) 75–104.
[20] G. Gómez, A. Espinal, S.W. Lagakos, Inference for a linear regression model with an interval-censored covariate, Stat. Med. 22 (2003) 409–425.
[21] A. Hald, Galileo's statistical analysis of astronomical observations, Int. Statist. Rev. 54 (1986) 211–220.
[22] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel, Robust Statistics: The Approach Based on Influence Functions, Wiley, 1986.
[23] D.F. Heitjan, D.B. Rubin, Ignorability and coarse data, Ann. Stat. 19 (1991) 2244–2253.
[24] P.J. Huber, E.M. Ronchetti, Robust Statistics, Wiley, 2nd edition, 2009.
[25] R.J. Hyndman, Y. Fan, Sample quantiles in statistical packages, Am. Stat. 50 (1996) 361–365.
[26] J.P. Klein, M.L. Moeschberger, Survival Analysis: Techniques for Censored and Truncated Data, Springer, 2nd edition, 2003.
[27] C. Lauro, F. Gioia, Dependence and interdependence analysis for interval-valued variables, in: V. Batagelj, H.H. Bock, A. Ferligoj, A. Žiberna (Eds.), Data Science and Classification, Springer, 2006, pp. 171–183.
[28] G. Li, C.H. Zhang, Linear regression with interval censored data, Ann. Stat. 26 (1998) 1306–1327.
[29] E.A. Lima Neto, F.A.T. de Carvalho, Centre and range method for fitting a linear regression model to symbolic interval data, Comput. Stat. Data Anal. 52 (2008) 1500–1515.
[30] J.K. Lindsey, A study of interval censoring in parametric regression models, Lifetime Data Anal. 4 (1998) 329–354.
[31] C.F. Manski, E. Tamer, Inference on regressions with interval data on a regressor or outcome, Econometrica 70 (2002) 519–546.
[32] M. Marino, F. Palumbo, Interval arithmetic for the evaluation of imprecise data effects in least squares linear regression, Ital. J. Appl. Stat. 14 (2002) 277–291.
[33] R.A. Maronna, D.R. Martin, V.J. Yohai, Robust Statistics: Theory and Methods, Wiley, 2006.
[34] W. Näther, Regression with fuzzy random data, Comput. Stat. Data Anal. 51 (2006) 235–252.
[35] A.B. Owen, Empirical Likelihood, Chapman & Hall/CRC, 2001.
[36] H. Prade, M. Serrurier, Why imprecise regression: A discussion, in: C. Borgelt, G. González-Rodríguez, W. Trutschnig, M.A. Lubiano, M.A. Gil, P. Grzegorzewski, O. Hryniewicz (Eds.), Combining Soft Computing and Statistical Methods in Data Analysis, Springer, 2010, pp. 527–535.
[37] R Development Core Team, R: A Language and Environment for Statistical Computing, 2011. Used R versions: 2.13.0 and 2.13.2.
[38] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, Wiley, 1987.
[39] M. Serrurier, H. Prade, A general framework for imprecise regression, in: FUZZ-IEEE 2007, Proceedings of the 2007 IEEE International Conference on Fuzzy Systems, IEEE Press, 2007, pp. 1597–1602.
[40] W.F. Sheppard, On the calculation of the most probable values of frequency-constants for data arranged according to equidistant divisions of a scale, Lond. M. S. Proc. 29 (1898) 353–380.
[41] S.M. Stigler, The History of Statistics: The Measurement of Uncertainty before 1900, Belknap Press of Harvard University Press, 1986.
[42] M. Terwey, S. Baltzer, ALLBUS Datenhandbuch 2008, GESIS, 2009.
[43] L.V. Utkin, F.P.A. Coolen, Interval-valued regression and classification models in the framework of machine learning, in: F. Coolen, G. de Cooman, T. Fetz, M. Oberguggenberger (Eds.), ISIPTA '11, Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications, SIPTA, 2011, pp. 371–380.
[44] S. Vansteelandt, E. Goetghebeur, M.G. Kenward, G. Molenberghs, Ignorance and uncertainty regions as inferential tools in a sensitivity analysis, Stat. Sin. 16 (2006) 953–979.
[45] G.A. Watson, On computing the least quantile of squares estimate, SIAM J. Sci. Comput. 19 (1998) 1125–1138.