Proceedings of Machine Learning Research 147:273–279, 2021 ISIPTA 2021

Computing Simple Bounds for Regression Estimates for Linear Regression with Interval-valued Covariates

Georg Schollmeyer GEORG.SCHOLLMEYER@STAT.UNI-MUENCHEN.DE
Department of Statistics, LMU Munich

Abstract

In this paper, we deal with linear regression where the covariates are interval-valued and the dependent variable is precise. Opposed to the case where the dependent variable is interval-valued and the covariates are precise, it is far more difficult to compute the set of all ordinary least squares (OLS) estimates as the precise values of the covariates vary over all possible values, compatible with the given intervals of the covariates. Though the exact solution is difficult to obtain, there are still some simple possibilities to compute bounds for the regression parameters. In this paper we deal with simple linear regression and present three different approaches: The first one uses a simple interval-arithmetic consideration for the equation for the slope parameter. The second approach uses reverse regression to swap the roles of the dependent and the independent variable to make the computation analytically solvable. The obtained solution for the reverse regression then gives an analytical upper bound for the slope parameter of the original regression. The third approach does not directly give bounds for the OLS estimator. Instead, before the actual interval analysis, in a first step, we modify the OLS estimator to another linear estimator which is simply a reasonably weighted convex combination of a number of unbiased estimators, which are themselves based on only two data points of the data set, respectively. It turns out that for the degenerate case of a precise independent variable, this estimator coincides with the OLS estimator. Additionally, the third method does also work if both the independent variable, as well as the dependent variable are interval-valued. Also the case of more than one covariate is manageable. A further nice point is that because of the analytical accessibility of the third estimator, also confidence intervals for the bounds can be established. To compare all three approaches, we conduct a short simulation study.

Keywords: interval regression, measurement error, interval arithmetic, Frisch's true regression, reverse regression, partial identification

1. Introduction

The present paper considers reliable regression analysis under interval data on the independent variable. More concretely, we study the influence of certain covariates x (also called independent variables, explanatory variables, regressors or stimuli) on a response variable y (dependent variable, outcome, response) under the additional difficulty that the covariates cannot be directly observed. Instead, one only observes upper and lower bounds for the unobservable covariates. This is a special case of so-called coarse(ned) data, i.e. data that are not observed in the resolution intended in the subject matter context (see, in particular, [7], for a discussion from a classical viewpoint, and, e.g., [1, Section 7.8] from an imprecise probability perspective). We focus here on simple linear regression of the form

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1,\dots,n \tag{1}$$
$$x_i \in [\underline{x}_i, \overline{x}_i] \quad a.s., \qquad i = 1,\dots,n. \tag{2}$$

Here $(\varepsilon_1,\dots,\varepsilon_n)$ is the vector of error terms which are assumed to be pairwise uncorrelated with expectation 0 and finite variance σ^2. Furthermore, we assume that not all intervals intersect and that y is not constant. (In the first case all three methods would break down and in the second case method 2 would break down.) We make no assumption about the distribution of the unobserved x_i within the observed intervals $[\underline{x}_i, \overline{x}_i]$ beyond (2). This may be seen as somehow unbalanced in comparison to the other assumptions above. However, note that also under much weaker assumptions (for example only assuming that the error variances are uniformly bounded by a constant and only assuming that the pairwise correlation is uniformly bounded below 1) the classical ordinary least squares estimator (OLS), which we will rely on in the sequel, is still an unbiased consistent estimator. This will also translate to the results we obtain here for the case of interval-valued covariates. The x_i's are not observed, one observes only the bounds $\underline{x}_i$ and $\overline{x}_i$, as well as the y_i's. Thus, the linear model is generally only partially identified. A reasonable set-valued estimator for the only partially identified parameters β_0 and β_1 is given by the collection of all OLS estimators where the

covariate values vary over all values, compatible with the observed bounds:[1]

$$OLS = \bigcup \left\{ \operatorname*{argmin}_{\beta} \{||X\beta - y||_2\} \mid X \in [\underline{X}, \overline{X}] \right\}. \quad (3)$$

Here, X is the design matrix

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

and $\underline{X}$ and $\overline{X}$ are the design matrices corresponding to the observed bounds and

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

is the column vector of the values of the dependent variable. Here, we assume a fixed covariate design, but the results naturally translate to a stochastic covariate design. The set *OLS* can serve as a reasonable set-valued estimate for the true but unidentified parameters in the sense that this set always contains the classical OLS estimate one would obtain, if one would know the unobserved covariates. Furthermore (under certain assumptions), this set-valued estimate converges almost surely to the identification region of the best linear predictor. For more details about this, see for example [14, 2, 3] (compare also, e.g., [9, 12], especially for other forms of identification regions). In the sequel, we are only interested in the slope parameter β_1 of the regression. Generally, computing the exact set *OLS* is computationally very hard for large n. Especially, if one is also interested in inference, one presumably has to estimate bounds for the scale parameter σ^2, which is generally **NP**-hard, see [8], cf., also [5]. Therefore, in this paper, we present methods that give non-sharp upper and lower bounds for the slope parameter β_1. We compare three approaches: The first approach uses simple interval-arithmetic, the second uses reverse regression and the known analytical results for a regression where only the dependent variable is interval-valued. The third approach replaces the classical OLS estimator by another linear estimator and then applies an interval-arithmetic analysis. Astonishingly, for the degenerate case of precisely observed covariates, this estimator coincides with the classical OLS estimator, which makes this approach very attractive. Additionally, the third approach can be modified and generalized in many ways, see the outlook given in Section 5.

1. The interval endpoints have to be always interpreted pointwise.

2. Three Approaches for Approximating the OLS Set

2.1. Approach 1: Simple Interval-arithmetic

For simple linear regression the explicit formula for the OLS slope parameter is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - mean(x))(y_i - mean(y))}{\sum_{i=1}^{n} (x_i - mean(x))^2}. \quad (4)$$

If now x cannot be observed and one only knows $x \in [\underline{x}, \overline{x}]$, then by elementary interval-arithmetic one can still deduce that $\hat{\beta}_1$ lies in the interval $[\underline{\hat{\beta}}_1, \overline{\hat{\beta}}_1]$ given by

$$\overline{\hat{\beta}}_1 := \frac{\sum_{i:y_i > mean(y)} \overline{(x_i - mean(x))}(y_i - mean(y))}{\sum_{i=1}^{n} \underline{(x_i - mean(x))^2}} \quad (5)$$

$$+ \frac{\sum_{i:y_i < mean(y)} \underline{(x_i - mean(x))}(y_i - mean(y))}{\sum_{i=1}^{n} \underline{(x_i - mean(x))^2}} \quad (6)$$

$$\underline{\hat{\beta}}_1 := \frac{\sum_{i:y_i < mean(y)} \overline{(x_i - mean(x))}(y_i - mean(y))}{\sum_{i=1}^{n} \overline{(x_i - mean(x))^2}} \quad (7)$$

$$+ \frac{\sum_{i:y_i > mean(y)} \underline{(x_i - mean(x))}(y_i - mean(y))}{\sum_{i=1}^{n} \overline{(x_i - mean(x))^2}} \quad (8)$$

where

$$\overline{x_i - mean(x)} = \overline{x}_i - mean(\underline{x}) \quad (9)$$

$$\underline{(x_i - mean(x))^2} = min\{(\underline{x}_i - mean(\overline{x}))^2, (mean(\underline{x}) - \overline{x_i})^2\} \quad (10)$$

$$\text{if } [\underline{x}_i, \overline{x}_i] \cap [mean(\underline{x}), mean(\overline{x})] = \emptyset \text{ and} \quad (11)$$

$$\underline{(x_i - mean(x))^2} = 0 \text{ else} \quad (12)$$

$$\overline{(x_i - mean(x))^2} = max\{(\overline{x}_i - mean(\underline{x}))^2, (mean(\overline{x}) - \underline{x_i})^2\}. \quad (13)$$

This simple interval-arithmetic analysis constitutes our first approach. (Note that there is a plenty of other forms of interval-arithmetic analyses, for example one other possibility would be to make an interval-arithmetic analysis for the representation of $\hat{\beta}$ as

$$\hat{\beta} = (X'X)^{-1}X'y. \tag{14}$$

For analysing the inverse of $(X'X)$ one could use for example the results given in [10].)

2.2. Approach 2: Using Reverse Regression and Analytical Bounds

The second approach uses the idea of reverse regression. For a sample $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, let $\beta_{yx} = \frac{Cov(x,y)}{Var(x)}$ be the slope parameter if we regress y on x and let $\beta_{xy} = \frac{Cov(x,y)}{Var(y)}$ be the slope of the reverse regression, i.e., the slope parameter if we regress x on y. Then it is well known that because of the Cauchy-Schwarz inequality we have

$$|\beta_{yx}| \leq \frac{1}{|\beta_{xy}|}. \tag{15}$$

Thus, if we can establish a lower bound for the slope parameter for the reverse regression, then we can compute an upper bound for the original regression. (For simplicity, we assume here that both slope parameters are non-negative.) Since for the reverse regression the roles of x and y are swapped, we have an interval-type regression problem where only the dependent variable is interval-valued. Concretely, we have $\beta_{yx} = \left[(Y'Y)^{-1}Y'x\right]_{11}$ with precise Y and only x interval-valued. The set of all OLS-parameters where x varies over all possible values is then simply the image of an n dimensional interval under a linear map, thus a zonotope, which is easily enough to describe, see [14] for a detailed analysis. Thus, we can analytically compute the smallest slope parameter for the reverse regression and get an upper bound for the original regression. Note that usually the bound given by the reverse regression is not sharp, it is sharp if the error term is zero and it usually gets more loose if the error term has a higher variance.

2.3. Approach 3: Replacing OLS by Another Linear Estimator

The third approach is based on the observation that it is enough to have two different data points $(x_i, y_i); (x_j, y_j)$ to estimate the slope of the regression in an unbiased way as $\frac{y_j - y_i}{x_j - x_i}$. Since one usually has more than two data points, one can make use of all pairs of data points and use as an estimate a weighted mean of the form:

$$\hat{\beta}_1 := \sum_{j>i} \alpha_{ji} \cdot \frac{y_j - y_i}{x_j - x_i} \tag{16}$$

with weights $\alpha_{ji} \geq 0$ such that $\sum_{j>i} \alpha_{ji} = 1$. Since every term $\frac{y_j - y_i}{x_j - x_i}$, if treated as a random variable, has expectation β_1, the estimator $\hat{\beta}_1$ is in fact an unbiased estimator of β_1. The variance of the estimator is[2]

$$\begin{aligned}
Var(\hat{\beta}_1) &= Var\left(\sum_{j>i} \alpha_{ji} \cdot \frac{y_j - y_i}{x_j - x_i}\right) && (17)\\
&= \sum_{j>i} Var\left(\alpha_{ji} \cdot \frac{y_j - y_i}{x_j - x_i}\right) && (18)\\
&\quad + \sum_{\substack{j>i,l>k,\\(i,j)\neq(k,l)}} Cov\left(\alpha_{ji} \cdot \frac{y_j - y_i}{x_j - x_i}, \alpha_{lk} \cdot \frac{y_l - y_k}{x_l - x_k}\right) && (19)\\
&= \sum_{j>i} \frac{\alpha_{ji}^2}{(x_j - x_i)^2} \cdot Var(y_j - y_i) && (20)\\
&\quad + \sum_{\substack{j>i,l>k,\\(i,j)\neq(k,l),\\|\{i,j,k,l\}|\leq 3}} \frac{\alpha_{ji} \cdot \alpha_{lk}}{(x_j - x_i)(x_l - x_k)} Cov(y_j - y_i, y_l - y_k) && (21)\\
&= \sum_{j>i} \frac{\alpha_{ji}^2}{(x_j - x_i)^2} \cdot 2\sigma^2 && (22)\\
&\quad + \sum_{\substack{j>i,l>k,\\(i,j)\neq(k,l),\\|\{i,j,k,l\}|\leq 3}} \frac{\alpha_{ji} \cdot \alpha_{lk}}{(x_j - x_i)(x_l - x_k)} \cdot \pm_{ijkl}\sigma^2 && (23)
\end{aligned}$$

with $\pm_{ijkl} \in \{-1, +1\}$ depending on the exact order of i, j, k and l. This is a positive semidefinite quadratic form in the coefficients α_{ji}. Thus, we can minimize the variance of the estimator by solving a quadratic program. Note that the independent variable is treated here as fixed, but the analysis also naturally translates to an analysis of the conditional variance given x within a stochastic covariate design. With this, we have an estimator which is linear in y and we can now use a simple interval-arithmetic analysis to deal with an imprecisely observed x by defining

$$\overline{\hat{\beta}}_1 := \sum_{j>i} \alpha_{ji} \cdot \frac{y_j - y_i}{\underline{x_j - x_i}} \tag{24}$$

$$\underline{\hat{\beta}}_1 := \sum_{j>i} \alpha_{ji} \cdot \frac{y_j - y_i}{\overline{x_j - x_i}} \tag{25}$$

where

$$\underline{x_j - x_i} := min\{x_j - x_i \mid x_j \in [\underline{x}_j, \overline{x}_j], x_i \in [\underline{x}_i, \overline{x}_i]\} \tag{26}$$

$$\overline{x_j - x_i} := max\{x_j - x_i \mid x_j \in [\underline{x}_j, \overline{x}_j], x_i \in [\underline{x}_i, \overline{x}_i]\}. \tag{27}$$

Here, we assume that $[\underline{x}_i, \overline{x}_i] \cap [\underline{x}_j, \overline{x}_j] = \emptyset$. Otherwise we could simply set the corresponding coefficient α_{ji} to zero. Furthermore, ties in the covariates can be handled by removing duplicated values and weighting the sample accordingly beforehand. Analogously to above, also for this imprecise situation, one can calculate the α_{ji}'s such that the variances

2. $|\cdot|$ denotes here the cardinality of a set.

of the estimators are minimal. But note that, opposed to the precise case, for different choices of the coefficients, also the expectation of the estimators vary such that one has here a tradeoff between the narrowness of the interval

$$\left[\mathbb{E}\left(\underline{\hat{\beta}_1}\right), \mathbb{E}\left(\overline{\hat{\beta}_1}\right)\right] \tag{28}$$

and the variability of the interval

$$[\underline{\hat{\beta}_1}, \overline{\hat{\beta}_1}]. \tag{29}$$

This is especially important if one wants to calculate confidence intervals, because the choice of the tradeoff then dependends on the exact used confidence level. For simplicity, in the simulation study of Section 3, we always used that coefficients α_{ji} that minimize the variability of the estimators. A further point worth mentioning is that the above analysis is only valid for the homoscedastic case. For the heteroscedastic case the situation is more subtle, but one can at least do the following: One can simply take the mid-points of the covariate intervals and then estimate with the midpoints for example only the linear trend of the error variability and then apply a weighted regression. Of course, this estimate of the linear trend would then generally be biased. But one can still hope that this biased estimate is better than an estimate of zero implicitly induced by the assumption of homoscedasticity. (Note that also for a biased estimate of the trend, the weighted OLS is still a consistent estimator.)

2.4. Relation to the OLS Estimator

In this short section, we show that the estimator from approach 3 is closely related to the OLS estimator by stating the following

Theorem 1 *For the case of degenerated interval-valued covariates (i.e., $\underline{x} = \overline{x}$), the bounds given in equations (24) and (25) coincide with the value of the slope parameter of the classical OLS estimator.*

Proof First, without loss of generality let us assume that the values $(x_1, \ldots, x_n)$ are already ordered increasingly. Then the OLS estimator is a linear form in y:

$$\beta_1^{OLS} = \sum_{i=1}^{n} c_i \cdot y_i. \tag{30}$$

with fixed coefficients $c_1, \ldots, c_n$. As can be seen from equation (4) the c_i's are then also ordered increasingly. Furthermore, we have $\sum_{i=1}^{n} c_i = 0$ because the OLS estimator is an unbiased estimator and if we would have $\sum_{i=1}^{n} c_i = C \neq 0$, this would contradict the fact that for example for $y \equiv 1$ we would get $\hat{\beta}_1^{OLS} \equiv C \neq 0$. Now we will show that $\hat{\beta}_1^{OLS}$ can be represented as

$$\sum_{j>i} \underbrace{\frac{\alpha_{ji}}{x_j - x_i}}_{=:d_{ji}} (y_j - y_i) \tag{31}$$

with non-negative α_{ji} and therefore also non-negative d_{ji}. Because the OLS estimator is unbiased, necessarily the α_{ji}'s sum up to 1. With this representation we can then argue that the OLS estimator is one special estimator in the set of estimators considered in approach 3. Since the OLS estimator is the estimator with the lowest variance under all unbiased linear estimators, it coincides with the minimal-variance estimator of approach 3.

Now, to establish the above representation we take $d_{ji} \neq 0$ only for $j = i+1$. Then we have to find values $d_{i+1,i}$ which satisfy

$$\sum_{i=1}^{n} c_i y_i = \sum_{i=1}^{n-1} d_{i+1,i}(y_{i+1} - y_i) \tag{32}$$

for arbitrary y. To guarantee this, we have to ensure that

$$c_n = d_{n,n-1} \tag{33}$$
$$c_{n-1} = d_{n-1,n-2} - d_{n,n-1} \tag{34}$$
$$c_{n-2} = d_{n-2,n-3} - d_{n-1,n-2}. \tag{35}$$
$$\vdots \tag{36}$$

For this it is enough to set

$$d_{n,n-1} = c_n \tag{37}$$
$$d_{n-1,n-2} = c_{n-1} + d_{n,n-1} = c_{n-1} + c_n \tag{38}$$
$$d_{n-2,n-3} = c_{n-2} + d_{n-1,n-2} = c_{n-2} + c_{n-1} + c_n. \tag{39}$$
$$\vdots \tag{40}$$

We now only have to make sure that all d_{ji}'s and therefore all α_{ji}'s are non-negative. But this is clearly the case because the sum of all increasingly ordered $c_1, \ldots, c_n$ is zero and thus taking only a sum $c_k + \ldots + c_n$ will necessarily also give non-negative values. ■

2.5. Confidence Intervals for Approach 3

From the above analysis, for a known dispersion σ^2, one can simply get the variability of the estimators. This can be used for constructing approximate confidence intervals. (Note that the same would also apply for approach 1.) However, one usually does not know the dispersion σ^2 of the unobservable error terms ε_i and this dispersion parameter is in our situation generally only partially identified. Moreover, computing a tight upper bound for $\hat{\sigma^2}$ is NP-hard, see

[8]. However, one can still use non-sharp bounds to get conservative confidence intervals for the estimators of the bounds. One possibility for doing so is to use the results given in [8]. There it is shown that (assuming the intercept and the slope to be non-negative)

$$\hat{\sigma}^2 \leq \min\{||w||_2 \mid \overline{X}\beta - w \leq \underline{y}, -\underline{X}\beta - w \leq -\overline{y}, \beta \geq 0\} \tag{41}$$

The right hand side of this inequality can be computed using linear programming. This result will be used in our short simulation study.

3. A Short Simulation Study

We now conduct a short simulation study. We take $n = 50$ data points and $\beta_0 = \beta_1 = 10$. We simulate under three different scenarios which are different in the distribution of the covariates. Within every scenario we look at three different situations with a low, a medium and a large dispersion of the error terms, respectively:

Scenario 1: Uniform covariate distribution: $x = (1,\ldots,50)$, dispersion $\sigma \in \{10, 100, 1000\}$. The lower bounds for the covariates are defined as $\underline{x}_i = x_i - 0.4$. The upper bounds are defined as $\overline{x}_i = x_i + 0.4$.

Scenario 2: right skewed covariate distribution: $x = (1^2, 2^2, \ldots, 50^2)$, dispersion $\sigma \in \{1000, 10000, 20000\}$. The lower bounds for the covariates are defined as $\underline{x}_i = (i - 0.4)^2$. The upper bounds are defined as $\overline{x}_i = (i + 0.4)^2$.

Scenario 3: left skewed covariate distribution: $x = (\sqrt{1}, \sqrt{2}, \ldots, \sqrt{50})$, dispersion $\sigma \in \{10, 50, 100\}$. The lower bounds for the covariates are defined as $\underline{x}_i = \sqrt{i - 0.4}$. The upper bounds are defined as $\overline{x}_i = \sqrt{i + 0.4}$.

We always simulated 100 times. Table 1 to table 3 show for every method both the estimated expectation and the estimated standard deviation for the corresponding estimator of the upper bound. (We omitted the results for the lower bounds because approach 2 does not give a lower bound and the results for the lower bounds for the other methods are similar to the results for the upper bounds.) Additionally, for method 3 we computed the upper bound $\hat{\overline{\sigma}}$ for the scale parameter based on [8, Theorem 7.4 (c)], averaged over all 100 simulations. Additionally, to get a feeling for the sharpness of the bounds, we did a constrained optimization and directly maximized the slope parameter for the classical OLS estimator constrained on the condition $x \in [\underline{x}, \overline{x}]$. This non-linear constrained optimization was solved using the algorithm described in [4]. But note that it is not clear if the optimizer did in fact find the global maximum. Thus, the bounds given by the optimization procedure (called reference in the tables) are only lower bounds for the upper bound of the OLS set.

Table 1: Scenario 1: Uniform covariate distribution.

Approach	$\sigma = 10$	$\sigma = 100$	$\sigma = 1000$
1, expectation	11.55	11.57	13.78
1, standard deviation	0.10	0.98	9.51
2, expectation	10.29	15.01	3257.02
2, standard deviation	0.09	1.02	26715.19
3, expectation	10.24	10.18	9.63
3, standard deviation	0.09	0.88	8.77
3, $\hat{\overline{\sigma}}$	13.41	103.00	1003.21
reference	10.24	10.31	10.97
reference, sd	0.09	0.87	8.60

Table 2: Scenario 2: right skewed covariate distribution.

Approach	$\sigma = 10^3$	$\sigma = 10^4$	$\sigma = 2 \cdot 10^4$
1, expectation	11.55	11.69	12.09
1, standard deviation	0.20	1.93	3.80
2, expectation	10.42	28.08	93.38
2, standard deviation	0.17	3.81	48.29
3, expectation	10.54	10.45	10.36
3, standard deviation	0.18	1.79	3.58
3, $\hat{\overline{\sigma}}$	1177.3	10158.9	20172.35
reference	10.24	10.25	10.38
reference, sd	0.17	1.72	3.42

Table 3: Scenario 3: left skewed covariate distribution.

Approach	$\sigma = 10$	$\sigma = 50$	$\sigma = 100$
1, expectation	11.97	12.84	14.24
1, standard deviation	0.89	4.29	8.52
2, expectation	14.11	133.39	7927.76
2, standard deviation	0.85	110.42	50217.69
3, expectation	10.28	10.06	9.80
3, standard deviation	0.77	3.86	7.72
3, $\hat{\overline{\sigma}}$	10.37	50.25	100.25
reference	10.31	10.57	11.06
reference, sd	0.77	3.78	7.51

4. Discussion of the Results

Now, let us shortly discuss the results.With exception of Scenario 2 and $\sigma = 1000$, both the expectation, as well as the standard deviation of the estimator for the upper bound is always the smallest for approach 3. The performance of the 3 methods are very similar for small dispersions of the

error terms. As expected, for method 2 both the expectation, as well as the standard deviation of the estimator is clearly larger for larger dispersions of the error term. This makes method 2 unattractive. But note that in a situation where the covariates are not only observed in intervals, but where the unobserved precise covariates are additionally prone to measurement error, then, already for the case of precise covariates, the OLS estimator systematically underestimates the true slope. In this situation, the reverse regression still gives an unbiased estimate of a sharp upper bound for the (then only partially identified) true slope parameter. This is referred to as 'Frisch's true regression' in [13], (cf., particularly [6]). This analysis would then naturally translate to the case of interval-valued covariates with additional measurement-error in the unobserved precise covariates. In this case it would thus be reasonable to use method 2 for the upper bound and method 3 for the lower bound of the slope parameter. With respect to inference, as could be expected, the estimates $\hat{\overline{\sigma}}$ tend to overestimate the true unknown scale parameter. Thus, confidence intervals based on these estimates will generally be conservative. Alternatively, one could also use bootstrap methods for inference. If there are more analytical, non-asymptotic exact solutions, not for estimating σ, which is only partially identified, but for the variability of the precise estimators of the bounds, seems to be an open question.

5. Summary and Outlook

In this paper, we have investigated three simple methods to obtain bounds for regression parameters for simple linear regression. Especially method 3 showed that it can often be useful to not directly adopt a known statistical problem from the precise to the interval-valued case, but instead, in a first step, to rethink the original statistical problem, here the problem of linear regression: Something that is (in a certain sense) most convincing (or convenient) in the precise case (here the OLS estimator as the uniformly best unbiased linear estimator under certain assumptions) is not necessarily the most convincing starting point for a generalization for the interval-valued case. Method 3 additionally allows for a plenty of modifications and generalizations:

1) Instead of a weighted mean one can also use a weighted median or another robust measure of location. This will usually make the estimator more robust. Also for the estimation of an upper bound for the scale parameter σ^2 one could adopt robust methods. For example the location free scale estimator analyzed in [11] is based on triples of data points. Also this idea can be adopted straightforwardly to the interval-valued case via interval-arithmetic. Also the variance of the resulting scale estimator can then be minimized by using a weighted median. (If one wants a high breaking point then maybe one should stick to the non-weighted median.)

2) Method 3 is also applicable for the case where both the covariates, as well as the dependent variable are interval-valued. We focused here on the case of a precise dependent variable only because method 2 is not capable of dealing with the situation of an interval-valued dependent variable. Note further that in this situation, under additional measurement error, although method 2 is not applicable, one can still apply method 3 to the original, as well as the reverse regression to get valid lower and upper bounds also in this situation.

3) Also more than one covariate is no problem for method 3. For example for two covariates one would have to look at triples instead of pairs of data points and compute a weighted mean over all triples. Note that also in the case of multiple regression p data points can identify all slope parameters and these parameters are simple linear forms in the y's. Furthermore, the extremal slope parameters are obtained if the precise unobserved covariate values are set to certain extremal points of the intervals. Of course, looking at all $p+1$-tuples of data points would be computationally challenging if one has a high number p of covariates. However, for example one could look not at all $p+1$-tuples of data points, but instead on only that $p+1$ tuples for which the data points have a high enough pairwise distance. Note further that unfortunately, for additional measurement error, the idea of reverse regression seems to not apply any more for more than one covariate. A further open question is if theorem 1 also translates to the case of multiple linear regression. This is an interesting point for further research.

Acknowledgments

The author would like to thank the four anonymous reviewers for their very helpful comments and the LMU Mentoring program, supporting young researchers for providing financial support.

References

[1] Thomas Augustin, Gero Walter, and Frank P. A. Coolen. *Statistical inference*, chapter 7, pages 135–189. John Wiley & Sons, Ltd, 2014. ISBN 9781118763117. doi: https://doi.org/10.1002/9781118763117.ch7.

[2] Arie Beresteanu and Francesca Molinari. Asymptotic properties for a class of partially identified models. *Econometrica*, 76(4):763–814, 2008. doi: https://doi.org/10.1111/j.1468-0262.2008.00859.x.

[3] Arie Beresteanu, Ilya Molchanov, and Francesca Molinari. Sharp identification regions in models with convex moment predictions. *Econometrica*, 79 (6):1785–1821, 2011. doi: https://doi.org/10.3982/ECTA8680.

[4] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5):1190–1208, 1995.

[5] Scott Ferson, Lev Ginzburg, Vladik Kreinovich, Luc Longpré, and Monica Aviles. Computing variance for interval data is np-hard. *SIGACT News*, 33(2): 108–118, June 2002. ISSN 0163-5700. doi: 10.1145/564585.564604.

[6] Ragnar Frisch. *Statistical confluence analysis by means of complete regression systems*, volume 5. Universitetets Økonomiske Instituut, 1934. URL https://www.sv.uio.no/econ/om/tall-og-fakta/nobelprisvinnere/ragnar-frisch/News/Confluence%20Analysis_2%5B1%5D.pdf.

[7] Daniel F. Heitjan and Donald B. Rubin. Ignorability and coarse data. *The Annals of Statistics*, 19(4):2244–2253, 1991. ISSN 00905364. URL http://www.jstor.org/stable/2241929.

[8] M. Hladík and M. Černý. Two optimization problems in linear regression with interval data. *Optimization*, 66(3):331–349, 2017. doi: 10.1080/02331934.2016.1274988.

[9] Maria Ponomareva and Elie Tamer. Misspecification in moment inequality models: back to moment equalities? *The Econometrics Journal*, 14(2):186–203, 06 2011. ISSN 1368-4221. doi: 10.1111/j.1368-423X.2010.00332.x.

[10] Jiri Rohn and Raena Farhadsefat. Inverse interval matrix: a survey. *The Electronic Journal of Linear Algebra*, 22, 2011. doi: https://doi.org/10.13001/1081-3810.1468.

[11] Peter J. Rousseeuw and Mia Hubert. Regression-free and robust estimation of scale for bivariate data. *Computational Statistics & Data Analysis*, 21(1):67–85, 1996. ISSN 0167-9473. doi: https://doi.org/10.1016/0167-9473(95)00004-6.

[12] Georg Schollmeyer and Thomas Augustin. Statistical modeling under partial identification: Distinguishing three types of identification regions in regression analysis with interval data. *International Journal of Approximate Reasoning*, 56:224–248, 2015. ISSN 0888-613X. doi: https://doi.org/10.1016/j.ijar.2014.07.003. Eighth International Symposium on Imprecise Probability: Theories and Applications (ISIPTA 2013).

[13] Elie Tamer. Partial identification in econometrics. *Annual Review of Economics*, 2(1):167–195, 2010. doi: 10.1146/annurev.economics.050708.143401.

[14] Michal Černý and Miroslav Rada. On the possibilistic approach to linear regression with rounded or interval-censored data. *Measurement Science Review*, 11(2): 34–40, 01 Jan. 2011. doi: https://doi.org/10.2478/v10048-011-0007-0.