Efficient First-Approximation Algorithms for Interval-Valued Regression, with Medical Applications in Mind

María Lizeth Reyna Cruz, Martine Ceberio, Christoph Q. Lauter,
Vladik Kreinovich, and Cecilia Alejandra Márquez Barraza

Abstract In many practical situations – in particular, in many medical problems – it is important to find the coefficients of linear regression based on the empirical data. In many such situations, we only know the upper bound on the absolute value of the measurement error – i.e., in effect, we only know intervals containing the actual values. When we know that the dependence is exactly linear, finding the exact ranges of possible values of the regression coefficients is NP-hard – meaning that, in general (unless P = NP), the exact computation of these ranges is not practically feasible. However, in many practical cases – in particular, in many medical applications – linear regression is only an approximate model, obtained by ignoring quadratic and higher order terms. In such cases, it is reasonable to also ignore quadratic order terms in our estimation of the ranges of regression coefficients. We show that this natural idea enables us to design efficient algorithms for estimating these ranges. Specifically, we present a polynomial-time algorithm.

1 Formulation of the Problem

General problem. In many practical situations, we are interested in a physical quantity y that is difficult to measure directly. Such situations are especially common in medical applications, when we are interested in different characteristics of the biological processes happening inside the human body.

María Lizeth Reyna Cruz, Martine Ceberio, Christoph Q. Lauter, and Vladik Kreinovich
Department of Computer Science, University of Texas at El Paso, 500 W. University
El Paso, Texas 79968, USA, e-mail: mlreynacruz@miners.utep.edu, mceberio@utep.edu,
cqlauter@utep.edu, vladik@utep.edu

Cecilia Alejandra Márquez Barraza
Centro Médico de Especialidades, Ciudad Juárez, Chihuahua, México,
e-mail: cecilia.marquez.b@gmail.com

To measure such a quantity, a natural idea is to find some easier-to-measure quantities $x_1, \ldots, x_n$ whose values largely determine the desired quantity y, i.e., for which, with reasonable accuracy, $y = f(x_1, \ldots, x_n)$ for some function $f(x_1, \ldots, x_n)$.

In some cases – e.g., in situations when the process of interest is described by fundamental physical laws – this function $f(x_1, \ldots, x_n)$ is known. However, in many other situations – in particular, in almost all medical applications – we do not know this function, we need to determine it based on the available data, i.e., based on the situations $k = 1, \ldots, K$ when we know both the values $x_1^{(k)}, \ldots, x_n^{(k)}$ and the corresponding value $y^{(k)}$.

In statistics, the problem is reconstructing the dependence from data is known as *regression*; in computer science, it is known as *machine learning*, as we learn a function $f(x_1, \ldots, x_n)$ from data.

Specific problems in which we are interested. The main problem in which we are interested in early detection of the Developmental Dysplasia of the Hip (DDH) (see, e.g., see, e.g., [8, 18]), a problem that occurs in 1-3% of the newborns [14]. When diagnosed early, DDH can often be treated non-invasively using devices that maintain the hip in a corrective posture, such as a special brace (called *Pavlik harness*) or a cast. However, when left undetected, DDH can lead to progressive joint damage, chronic pain, and impaired mobility, frequently requiring surgery. The main tool for detecting DDH is a special ultrasound screening called the *Graf method*; see, e.g., [3]. To detect DDH, one needs to estimate geometric characteristics of the hip joint from the ultrasound image. For DDH, we have already published some preliminary results [13].

Another example to which we plan to apply our technique is the treatment of relapsed Non-Hodgkin Lymphoma (NHL). NHL is one of the most common blood-related cancers. For NHL relapses, an appropriate early treatment largely increases the survival rate; see, e.g., [1, 4, 15]. There are several treatment pathways that a physician can use to treat NHL relapse. Our objective is to help the physicians to select, for each patient, the pathway that is the most appropriate for this particular patient. For the NHL problem, we are still gathering and preprocessing the data, we do not have results yet.

Possibility of linearization. The desired dependence is often smooth, and it is known that sufficiently smooth dependencies can be expanded into Taylor series. This means that to get a reasonable approximation, we can take the sum of the first few terms of the Taylor expansion; see, e.g., [2, 17]. In many practical situations, already the first – linear – terms provide a good approximation to the desired dependence [2, 17]:

$$y = c_0 + c_1 \cdot x_1 + \ldots + c_n \cdot x_n. \tag{1}$$

Such cases are known as *linear regression*. In such cases, to describe the desired function $f(x_1, \ldots, x_n)$, we need to find the coefficients $c_0, \ldots, c_n$.

Need for interval uncertainty. The equation (1) deals with the exact values of the quantities x_i and y. In reality, the only way we can know the values of physical quantities is by measurement, and measurements are never absolutely accurate: the

measurement results $\widetilde{x}_i$ and $\widetilde{y}$ are, in general, somewhat different from the actual (unknown) values x_i and y; see, e.g., [12]. The corresponding differences $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$ and $\Delta y \stackrel{\text{def}}{=} \widetilde{y} - y$ are known as *measurement errors*.

In some practical situations, when the values x_i and y are measured by well-calibrated measuring instruments, we know the probability distributions of all the measurement errors Δx_i and Δy. In such situations, we can use statistical methods to find not only the approximate values of c_i but also the probabilities of different deviations from these approximate values; see, e.g., [16].

However, in many other situations, we do not know the probability distributions for the measurement errors. Such situations are especially common in medicine, when the measurement accuracy depends on the skills of the medical professional who performs the measurements. In such cases, usually, the only information that we have about each measurement error Δx_i or Δy is the upper bound Δ_i or Δ on its absolute value – the upper bound that distinguishes professionals for the trainees who are still learning the corresponding skill: $|\Delta x_i| \le \Delta_i$ and $|\Delta y| \le \Delta$. In this case, after the measurements, the only information that we have about the actual values x_i and y of each quantity is that these values belong to the intervals $[\underline{x}_i, \overline{x}_i] \stackrel{\text{def}}{=} [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$ and $[\underline{y}, \overline{y}] \stackrel{\text{def}}{=} [\widetilde{y} - \Delta, \widetilde{y} + \Delta]$.

So, we arrive at the following two problems.

Resulting practical problems. In both problems, we have the interval-valued results of K measurement cycles; in each cycle, we measure the values of $n+1$ quantities $x_1, \ldots, x_n$, and y.

Problem 1.

- *Available information:* For each $k = 1, \ldots, K$, we know the intervals
$$\left[\underline{x}_1^{(k)}, \overline{x}_1^{(k)}\right], \ldots, \left[\underline{x}_n^{(k)}, \overline{x}_n^{(k)}\right] \text{ and } \left[\underline{y}^{(k)}, \overline{y}^{(k)}\right].$$
- For each $i = 0, \ldots, n$, and for each number c_i, we say that this number is *possible* if there exist values $c_0, \ldots, c_{i-1}, c_{i+1}, \ldots, c_n$, $x_i^{(k)} \in \left[\underline{x}_i^{(k)}, \overline{x}_i^{(k)}\right]$, and $y^{(k)} \in \left[\underline{y}^{(k)}, \overline{y}^{(k)}\right]$, for which
$$y^{(k)} = c_0 + c_1 \cdot x_1^{(k)} + \ldots + c_n \cdot x_n^{(k)}. \tag{2}$$
- *Our objective* is to find, for each $i = 0, 1, \ldots, n$, the set $[\underline{c}_i, \overline{c}_i]$ of all possible values of c_i.

Problem 2.

- *Available information:*
  - For each $k = 1, \ldots, K$, we know the intervals
$$\left[\underline{x}_1^{(k)}, \overline{x}_1^{(k)}\right], \ldots, \left[\underline{x}_n^{(k)}, \overline{x}_n^{(k)}\right] \text{ and } \left[\underline{y}^{(k)}, \overline{y}^{(k)}\right].$$

  - We also know intervals $[\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n]$.

- We say that a real number y is *possible* if there exist values $c_0, \ldots, c_{i-1}, c_i, c_{i+1}, \ldots, c_n$, $x_i^{(k)} \in \left[\underline{x}_i^{(k)}, \overline{x}_i^{(k)}\right]$, $y^{(k)} \in \left[\underline{y}^{(k)}, \overline{y}^{(k)}\right]$, and $x_i \in [\underline{x}_1, \overline{x}_1]$ for which the equalities (1) and (2) are true.
- *Our objective* is to find, the set $\left[\underline{y}, \overline{y}\right]$ of all possible values of y.

Existing methods of interval linear regression and their limitation. There exist algorithms for solving these problems; see, e.g., [5, 7, 9, 10, 11]. However, these algorithms are often very time-consuming – and when limited to reasonable computation time, they often produce *not* the desired intervals themselves, but rather the *enclosures* for the corresponding intervals.

This is not a fault of specific algorithms: it is known that the problem of finding the exact intervals for c_i is, in general, NP-hard; see, e.g., [6].

Idea. At first glance, it looks like the NP-hardness result makes a search for more efficient general algorithms hopeless. But this is not necessarily so. NP-hardness has been proven for the case when the dependence is exactly linear. However, in most real-life problems – especially in medical problems – a linear model is simply a good first approximation, obtained by ignoring quadratic and higher order terms. Since we ignore quadratic terms anyway, why not ignore similar quadratic and higher order terms in the constructions of the intervals for c_i and y?

What we do in this paper. In this paper, we show that the above idea indeed leads to an efficient polynomial-time algorithm for both above-described problems related to the (first-approximation) interval-based regression. Specifically, in Section 2, we will describe a feasible algorithm for computing the ranges of the coefficients, and in Section 3, we will describe a feasible algorithm for computing the range of y.

2 Towards a Feasible Algorithm for Computing the Ranges of the Regression Coefficients c_i

Analysis of the problem. For both Problems 1 and 2, we know that the actual (unknown) values c_i, $x_i^{(k)}$, and $y^{(k)}$ satisfy the formula (2):

$$c_0 + \sum_i c_i \cdot x_i^{(k)} = y^{(k)}, \quad k = 1, \ldots, K. \tag{3}$$

Here, we do not know the values $x_i^{(k)}$ and $y^{(k)}$, we only know the intervals $\left[\underline{x}_i^{(k)}, \overline{x}_i^{(k)}\right] = \left[\widetilde{x}_i^{(k)} - \Delta_i^{(k)}, \widetilde{x}_i^{(k)} + \Delta_i^{(k)}\right]$ and $\left[\underline{y}^{(k)}, \overline{y}^{(k)}\right] = \left[\widetilde{y}^{(k)} - \Delta^{(k)}, \widetilde{y}^{(k)} + \Delta^{(k)}\right]$ that contain these values.

As a crude approximation, we can ignore the interval uncertainty, use the midpoints $\widetilde{x}_i^{(k)}$ and $\widetilde{y}^{(k)}$ and apply, e.g., the usual Least Squares Method (see, e.g., [16]) to find the approximate values $\widetilde{c}_i$ of c_i. Let us see how it helps.

Each term $c_i \cdot x_i^{(k)}$ in the formula (3) can be described in the following form:

$$c_i \cdot x_i^{(k)} = c_i \cdot \widetilde{x}_i^{(k)} + c_i \cdot \Delta x_i^{(k)},$$

where we denoted $\Delta x_i^{(k)} \stackrel{\text{def}}{=} x_i^{(k)} - \widetilde{x}_i^{(k)}$. Here, $c_i = \widetilde{c}_i + \Delta c_i$, where we denote $\Delta c_i \stackrel{\text{def}}{=} c_i - \widetilde{c}_i$. Thus:

$$c_i \cdot x_i^{(k)} = c_i \cdot \widetilde{x}_i^{(k)} + \widetilde{c}_i \cdot \Delta x_i^{(k)} + \Delta c_i \cdot \Delta x_i^{(k)}.$$

The last term is quadratic in terms of the uncertainties, and since by using linear regression we already ignores quadratic and higher order terms in the actual dependence, we can safely ignore this second-order term as well, and assume that

$$c_i \cdot x_i^{(k)} = c_i \cdot \widetilde{x}_i^{(k)} + \widetilde{c}_i \cdot \Delta x_i^{(k)}.$$

Substituting this expression into the formula (3), we get the following formulas that are linear in terms of the unknowns c_i, $\Delta x_i^{(k)}$, and $y^{(k)}$:

$$c_0 + \sum_i c_i \cdot \widetilde{x}_i^{(k)} + \sum_i \widetilde{c}_i \cdot \Delta x_i^{(k)} = y^{(k)}, \quad k = 1, \ldots, K. \tag{4}$$

To these equalities, we need to add inequalities describing our uncertainty in $x_i^{(k)}$ and $y^{(k)}$:

$$\widetilde{y}^{(k)} - \Delta^{(k)} \le y^{(k)} \le \widetilde{y}^{(k)} + \Delta^{(k)} \tag{5}$$

and

$$-\Delta_i^{(k)} \le \Delta x_i^{(k)} \le \Delta_i^{(k)}. \tag{6}$$

Now, for each i, we can find the lower bound $\underline{c}_i$ by solving the following problem:

$$\text{minimize } c_i \text{ under the conditions (4)-(6).}$$

Both the objective function c_i and the constraints (4)-(6) are linear with respect to the unknowns c_i, $\Delta x_i^{(k)}$, and $y^{(k)}$, so we can use well-known and efficient linear programming software (see, e.g., [19]) to compute this lower bound.

By maximizing instead of minimizing, we get the upper bound $\overline{c}_i$. So, we arrive at the following feasible algorithm.

Resulting feasible algorithm. Let us recall the problem:

- *We are given:* the values $\widetilde{x}_i^{(k)}$, $\Delta_i^{(k)}$, $\widetilde{y}^{(k)}$, and $\Delta^{(k)}$, where $k = 1, \ldots, K$, and $i = 1, \ldots, n$.
- *We know:* that for some $c_0, \ldots, c_n$ and for some values

$$x_i^{(k)} \in \left[\widetilde{x}_i^{(k)} - \Delta_i^{(k)}, \widetilde{x}_i^{(k)} + \Delta_i^{(k)}\right] \text{ and } y^{(k)} \in \left[\widetilde{y}^{(k)} - \Delta^{(k)}, \widetilde{y}^{(k)} + \Delta^{(k)}\right],$$

  the formula (3) holds.
- *We want to find:* for each i, the range $[\underline{c}_i, \overline{c}_i]$ of possible values of c_i.

Here is the proposed algorithm:

- First, we use the Least Squares Method (see, e.g., [16]) to find the approximate values $\widetilde{c}_i$ of all the coefficients.
- Then, to find $\underline{c}_i$, we solve the following linear programming problem: minimize c_i under the constraints (4)-(6).
- Finally, to find $\overline{c}_i$, we solve the following linear programming problem: maximize c_i under the constraints (4)-(6).

3 Towards a Feasible Algorithm for Computing the Range of the Desired Value y

Analysis of the problem. We want to find the range of the value y, i.e., the set of possible values of y:

$$y = c_0 + c_1 \cdot x_1 + \ldots + c_n \cdot x_n, \tag{7}$$

where the only thing we know about x_i is that $x_i \in [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.

We can use similar techniques to predict y based on the approximately known x_i values. Indeed, each term $c_i \cdot x_i$ can be represented in the following form:

$$c_i \cdot x_i = c_i \cdot \widetilde{x}_i + c_i \cdot \Delta x_i,$$

where we denoted $\Delta x_i \stackrel{\text{def}}{=} x_i - \widetilde{x}_i$. Here, $c_i = \widetilde{c}_i + \Delta c_i$, where we denote $\Delta c_i \stackrel{\text{def}}{=} c_i - \widetilde{c}_i$. Thus:

$$c_i \cdot x_i = c_i \cdot \widetilde{x}_i + \widetilde{c}_i \cdot \Delta x_i + \Delta c_i \cdot \Delta x_i.$$

The last term is quadratic in terms of the uncertainties, and since by using linear regression we already ignores quadratic and higher order terms in the actual dependence, we can safely ignore this second-order term as well, and assume that

$$c_i \cdot x_i = c_i \cdot \widetilde{x}_i + \widetilde{c}_i \cdot \Delta x_i.$$

Substituting this expression into the formula (7), we get the following formula:

$$y = c_0 + \sum_i c_i \cdot \widetilde{x}_i + \sum_i \widetilde{c}_i \cdot \Delta x_i. \tag{8}$$

To this equation, we need to add the formulas (4)-(6), and also the following inequalities:

$$-\Delta_i \le \Delta x_i \le \Delta_i. \tag{9}$$

Both the objective expression y and all the constraints are linear in terms of the unknowns – to which we add Δx_i and y. So, to find the bounds for y, we can also use linear programming – feasible technique for optimizing linear functions under linear constraints. Thus, we arrive at the following feasible algorithm.

Resulting feasible algorithm. Let us recall the problem:

- *We are given:* the values $\widetilde{x}_i^{(k)}$, $\Delta_i^{(k)}$, $\widetilde{y}^{(k)}$, and $\Delta^{(k)}$, where $k = 1, \ldots, K$, and $i = 1, \ldots, n$, and the values $\widetilde{x}_i$ and Δ_i.
- *We know:* that for some $c_0, \ldots, c_n$ and for some values

$$x_i^{(k)} \in \left[\widetilde{x}_i^{(k)} - \Delta_i^{(k)}, \widetilde{x}_i^{(k)} + \Delta_i^{(k)}\right] \text{ and } y^{(k)} \in \left[\widetilde{y}^{(k)} - \Delta^{(k)}, \widetilde{y}^{(k)} + \Delta^{(k)}\right],$$

  the formula (3) holds.
- *We want to find:* the range $\left[\underline{y}, \overline{y}\right]$ of possible values of the expression (7) corresponding to all possible values of $x_i \in [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$ and c_i.

Here is the proposed algorithm:

- First, we use the Least Squares Method (see, e.g., [16]) to find the approximate values $\widetilde{c}_i$ of all the coefficients.
- Then, to find $\underline{y}$, we solve the following linear programming problem: minimize y under the constraints (4)-(6) and (8)-(9).
- Finally, to find $\overline{y}$, we solve the following linear programming problem: maximize y under the constraints (4)-(6) and (8)-(9).

Acknowledgments

This work was supported by the AT&T Fellowship in Information Technology, by the Institute for Risk and Reliability, Leibniz Universitaet Hannover, Germany, by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Focus Program SPP 100+ 2388, Grant Nr. 501624329, by the European Union under the project ROBOPROX (No. CZ.02.01.01/00/22 008/0004590), by the Center of Excellence in Econometrics, Faculty of Economics, Chiang Mai University, Thailand, by the Ho Chi Minh City University of Banking, Vietnam, and by Thang Long University, Hanoi, Vietnam.

References

1. L. de Leval and E. S. Jaffe, "Lymphoma classification", *The Cancer Journal*, 2020, Vol. 26, No. 3, pp. 176–185.
2. R. Feynman, R. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Addison Wesley, Boston, Massachusetts, 2005.
3. R. Graf, *Hip Sonography: Diagnosis and Management of Infant Hip Dysplasia*, Springer, Cham, Switzerland, 2006.
4. Instituto Mexicano del Seguro Social, *Guía de Práctica Clínica, Linfomas No Hodgkin en el Adulto*, México, 2009.
5. L. Jaulin, M. Kiefer, O. Didrit, and E. Walter, *Applied Interval Analysis, with Examples in Parameter and State Estimation, Robust Control, and Robotics*, Springer, London, 2012.
6. V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational Complexity and Feasibility of Data Processing and Interval Computations*, Kluwer, Dordrecht, 1998.

7. B. J. Kubica, *Interval Methods for Solving Nonlinear Constraint Satisfaction, Optimization, and Similar Problems: from Inequalities Systems to Game Solutions*, Springer, Cham, Switzerland, 2019.
8. S. T. Mahan, J. N. Katz, and Y.-J. Kim, "To screen or not to screen? A decision analysis of the utility of screening for developmental dysplasia of the hip," *The Journal of Bone & Joint Surgery (JBJS)*, 2009, Vol. 91, No. 7, pp. 1705–1719.
9. G. Mayer, *Interval Analysis and Automatic Result Verification*, de Gruyter, Berlin, 2017.
10. R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*, SIAM, Philadelphia, 2009.
11. A. Neumaier, *Interval Methods for Systems of Equations*, Cambridge University Press, Cambridge, UK, 1990.
12. S. G. Rabinovich, *Measurement Errors and Uncertainty: Theory and Practice*, Springer Verlag, New York, 2005.
13. M. L. Reyna Cruz, R. Tabares, M. Ceberio, V. Kreinovich, C. Lauter, and C. A. Márquez Barraza, "Machine learning-based screening for pediatric hip dysplasia: Towards a validated approach", *Proceedings of the 58th Asilomar Conference on Signals, Systems, and Computers, Pacific Grove, October 27–30, 2024*, IEEE Press, pp. 1891–1894.
14. A. Sarmiento-Piñeros, S. Muñoz-Medina, and S. Quevedo, "Incidencia de displasia del desarrollo de cadera. Estandarizando la radiografía con un dispositivo anti rotatorio Orthohip (Incidence of hip development displasia. Standardizing the radiography with the Orthohip antirotatory device)", *Revista Colombiana de Ortopedia y Traumatología*, 2022, Vol. 36, No. 3, pp. 140–146.
15. K. R. Shankland, J. O. Armitage, and B. W. Hancock. "Non-Hodgkin Lymphoma", *The Lancet*, 2-12, Vol. 380, No. 9844, pp. 848–857.
16. D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.
17. K. S. Thorne and R. D. Blandford, *Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics*, Princeton University Press, Princeton, New Jersey, 2021.
18. V. V. Upasani, J. D. Bomar, T. H. Matheney, W. N. Sankar, K. Mulpuri, C. T. Price, C. F. Moseley, S. P. Kelley, U. Narayanan, N. M. P. Clarke, et al., "Evaluation of brace treatment for infant hip dislocation in a prospective cohort: defining the success rate and variables associated with failure", *The Journal of Bone & Joint Surgery (JBJS)*, 2016, Vol. 98, No. 14, pp. 1215–1221.
19. R. J. Vanderbei, *Linear Programming: Foundations and Extensions*, Springer, New York, 2014.