Toward Handling Uncertainty-At-Source in AI—A Review and Next Steps for Interval Regression

Shaily Kabir, Christian Wagner, Senior Member, IEEE, and Zack Ellerby

Abstract—Most of statistics and AI draw insights through modeling discord or variance between sources (i.e., intersource) of information. Increasingly however, research is focusing on uncertainty arising at the level of individual measurements (i.e., within- or intrasource), such as for a given sensor output or human response. Here, adopting intervals rather than numbers as the fundamental data-type provides an efficient, powerful, yet challenging way forward—offering systematic capture of uncertainty-at-source, increasing informational capacity, and ultimately potential for additional insight. Following progress in the capture of interval-valued data in particular from human participants, conducting machine learning directly upon intervals is a crucial next step. This article focuses on linear regression for interval-valued data as a recent growth area, providing an essential foundation for broader use of intervals in AI. We conduct an in-depth analysis of state-of-the-art methods, elucidating their behavior, advantages, and pitfalls when applied to synthetic and real-world datasets with different properties. Specific emphasis is given to the challenge of preserving mathematical coherence, i.e., models maintain fundamental mathematical properties of intervals. In support of real-world applicability of the regression methods, we introduce and demonstrate a novel visualization approach, the interval regression graph (IRG), which effectively communicates the impact of both position and range of variables within the regression models—offering a leap in their interpretability. Finally, this article provides practical recommendations concerning regression-method choice for interval data and highlights remaining challenges and important next steps for developing AI with the capacity to handle uncertainty-at-source.

Impact Statement—Capturing information as intervals provides a powerful means for handling uncertainty and inherent range in data. This paper focuses on a basic building block of statistics and AI: linear regression. In recent years, regression for intervals has become a topic of interest in AI and has been applied to domains ranging from marketing to cyber-security. It allows direct modelling of relationships not only between variables per-se, but also their associated uncertainty. For example, we can infer not only how a snack’s nutritional benefits impact consumer purchase intention, but also how uncertainty about these benefits impacts purchase intention and its associated uncertainty. Nonetheless, while there are considerable upsides to interval regression and AI, substantial challenges remain. This paper reviews and extends state-of-the-art interval regression methods, presents in-depth experimental results and introduces a novel visualization approach for interval regression—enhances interpretability and facilitates real-world insight. Finally, it provides recommendations for algorithm selection based on key properties of data sets, discusses next steps for interval-valued AI beyond linear regression, and is complemented by an open-source implementation of the algorithms discussed.

Index Terms—Intersource, intervals, inrasource, regression, uncertainty.

NOMENCLATURE

| Term         | Definition                                                                 |
|--------------|----------------------------------------------------------------------------|
| CM           | Center method [15].                                                        |
| MinMax       | MinMax method [16].                                                       |
| CRM          | Center and range method [17].                                              |
| CCRM         | Constrained center and range method [18].                                 |
| CIM          | Complete information method [19].                                         |
| LM           | Linear model [20].                                                        |
| PM           | Parametrized model [21].                                                  |
| LM_c         | Constrained linear model.                                                 |
| LM_w         | Weakly constrained linear model.                                          |
| IV           | Interval-valued.                                                          |
| IRG          | Interval regression graph.                                                |
| σ          | Interval \(\{\bar{\sigma} \subseteq \mathbb{R} : \bar{\sigma} = [a^-, a^+], a^- \leq a^+\}\). |
| a_w         | Range of interval, \(a_w = [a^+ - a^-]\).                                |
| a_c         | Center of interval, \(a_c = (a^+ + a^-)/2\).                             |
| a_r         | Radius of interval \(a_r = \frac{a_w}{2}\).                             |
| Y           | Estimated Y.                                                              |
| X           | Interval-valued regressor.                                                |
| μY          | Mean and standard deviation of range of Y.                                |
| μX          | Mean and standard deviation of range of X.                                 |
| σY          | Mean and standard deviation of center of Y.                                |
| σX          | Mean and standard deviation of center of X.                                |

I. INTRODUCTION

Interval-valued (IV) data have gained increasing attention across a variety of contexts as they offer a systematic way to capture information complete with an intrinsic representation of range or uncertainty in each individual “measurement,” which is not possible using point-values, such as numbers or ranks. Such data may arise due to imprecision and uncertainty in measurement (e.g., sensor data), inherent uncertainty of outcome (e.g., estimations of stock prices [1], climate/weather forecasting [2]), or inherent vagueness or nuance (e.g., in linguistic terms [3]). A growing body of evidence also suggests that subjective judgement of humans (e.g., experts, consumers) may be better represented by intervals [4], [5], where the interval width/size

Manuscript received 25 August 2022; revised 13 October 2022; accepted 20 November 2022. Date of publication 8 January 2024. This work was supported by the UK EPSRC’s Leveraging the Multi-Stakeholder Nature of Cyber Security EP/P011918/1 and Horizon: Trusted Data-Driven Products EP/T022493/1 grants. (Corresponding author: Shaily Kabir.)

The authors are with LUCID, the Lab for Uncertainty in Data and Decision Making, School of Computer Science, University of Nottingham, Nottingham NG7 2RD, U.K. (e-mail: shaily.kabir3@nottingham.ac.uk; christian.wagner@nottingham.ac.uk; zack.ellerby@nottingham.ac.uk).

Digital Object Identifier 10.1109/TAL.2023.3234930

2691-4581 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
captures the response range as a conjunctive set (e.g., the reals between 2 and 4), or degree of uncertainty with respect to an estimate or rating—a disjunctive set (e.g., confidence interval) [6].

Regression analysis for estimating one or more variables is an important task in data analysis and machine learning under the wider umbrella of AI. While regression is most commonly applied to numeric data, researchers have begun to develop methods of applying (linear) regression analysis to IV data [7]. To date, models designed to handle datasets where both independent (regressor) and dependent (regressand) variables are IV have been developed within the frameworks of random sets and symbolic data analysis (SDA) [8], [9]. In the random set framework, intervals are viewed as compact convex sets in \( \mathbb{R} \), and their interaction is modeled according to set arithmetic [10], [11], [12], [13], [14]. On the other hand, in the SDA setting, the classical regression model is expanded to IV data where intervals are treated as bivariate vectors [15], [16], [17], [18], [19], [20], [21], [22].

Current linear regression models for intervals use different reference points, such as the regressors’ center values, their lower and upper bounds, or both their center and range (width), to estimate the IV regressand [21]. Earlier interval regression approaches [15], [16], [17] risk “flipping” the interval bounds, respectively, generating a negative interval range, thus breaking mathematical coherence (i.e., violating the definition of interval). The more recent approaches [18], [19], [20], [21] impose either positivity restrictions on parameters, or design the regression model so as to ensure that the lower bound does not exceed the upper bound. For instance, Neto and Carvalho [18] apply the Lawson and Hanson algorithm (LHA) [23], Wang et al. [19] use Moore’s linear combination [24], and Souza et al. [21] the Box-Cox transformation [25] to guarantee mathematical coherence. Although Sun and Ralescu [20] consider positivity restrictions on parameters in their model setting, they do not provide a method to maintain these restrictions. A detailed review and analysis of these models is provided in Section II and the general issue is discussed throughout this article.

An important aspect is that most current approaches, including the above, restrict parameters based on the assumption that an increase in uncertainty\(^1\) of a regressor (i.e., input) leads to greater uncertainty in the dependent variable or regressand. However, there are plausible cases in the real world where this positive correlation does not hold. For example, in cyber-security, uncertainty around the vulnerability of a system may lead to uncertainty in the need to invest resources into controls—thus mitigating the risk of the system being compromised, that is, a negative correlation. Similarly, in a multivariate case, consider an economics example; greater uncertainty in future product prices and interest rates may cause more certainty with respect to present consumption choices.

For illustration, Figs. 1(a) and 2(a) show a synthetic example of two IV datasets. Here, in Set-1 (see Fig. 1(a)), the position (center) and range (uncertainty) of the regressand, \( Y \) increases with the position (center) and range (e.g. the uncertainty) of regressor, \( X \)—that is, they vary in unison—this reflects the most common case as outlined above, e.g., the price of cars versus their horsepower. In contrast, for Set-2 [see Fig. 2(a)], the range (uncertainty) of the regressand, \( Y \) increases irrespective of the range (uncertainty) of the regressor, \( X \). This reflects a less common scenario, e.g., the wealth of people with respect to their age. The relationships between the variables in each figure are captured through interval regression graphs (IRGs) in Figs. 1(b) and 2(b), respectively. IRGs will be introduced in detail in Section III-B, but for now, note how Fig. 1(b) shows a gentle increase in the range of the dependent variable (\( 2Y^r \)) for an increase in the range of the independent variable (\( 2X^r \)) for Set-1, while Fig. 2(b) shows, as expected, no such change for Set-2. The IRGs also highlight the much less drastic increase in range of Set-1’s dependent variable that is attributed to an increase in the position of the independent variable (\( X^c \)), compared to Set-2, in which variance in regressand range (\( 2Y^r \)) is attributed solely to regressor position (\( X^c \)).

Aspects such as the above highlight the complexity but also the potential of IV data in directly capturing uncertainty and range at-source, making it accessible for reasoning. With intervals encoding additional information by comparison to numbers, the choice of a “universally best” analysis approach, as can often be found for numeric cases, is, however, highly challenging. Instead, the selection of appropriate techniques for a given IV dataset—based on its properties—is essential. Ferson et al. [26] discuss the complexity of analysing real-world IV data in detail, elucidating the value of focusing on key properties for datasets

---

\(^1\)Throughout the rest of this article, for simplicity, we refer to the range of the interval as representing uncertainty as a general term capturing vagueness, lack of information, etc.
to inform the choice of methods, such as whether the intervals exhibit strong overlap (i.e., “puffy” [26]: wider, more uncertain intervals) or very little overlap (i.e., “skinny” [26]: narrow, less uncertain); whether there are outliers; whether intervals are highly scattered, densely placed or even nested within each other. Beyond these properties, more fundamental questions arise from the semantic meaning of the given intervals, as already alluded to above; whether they represent an epistemic, disjunctive set, with the interval describing the range within which a correct answer is contained (this is arguably the most common use of intervals in science, cf. confidence intervals), or indeed an ontic, conjunctive set with a true range, such as the real numbers between 1 and 5, or a time span [5], [27]. For conciseness, we do not discuss the latter further in this manuscript, but will revisit it, and its impact on the analysis of AI and machine learning more generally, in future publications.

At this stage, we also note that of course one can go beyond intervals to model uncertainty, leveraging distributions with a probabilistic or fuzzy interpretation, depending on the dis- or con-junctive nature of the information. Focusing on intervals a priori provides advantages, such as increasingly established pathways to collecting data (e.g., [5]), while also laying foundations for more advanced methods. For example, in fuzzy set theory, $\alpha$-cut decomposition provides a direct translation mechanism between intervals and fuzzy sets. This article makes four principal contributions.

1) **Review of the state-of-the-art**: In view of the significance and use of intervals as a fundamental data-type in AI, the paper presents a detailed review of all well-known state-of-the-art linear regression approaches based on vector representation of data sets where both independent and dependent variables are IV.

2) **Extensions of current methods**: As part of the review of the methods, this paper puts forward extensions to the Linear Model (LM) [20], enhancing its robustness efficiently for real-world applications. Specifically, we propose a dynamically selective approach to applying restrictions, letting the model’s parameters vary freely unless the mathematical coherence is being compromised.

3) **Practical guidance on algorithm selection and software**: To support the broader adoption of intervals as a fundamental data-type in AI, extant regression approaches and their performance are analysed using synthetic and real world data sets capturing various common and features of intervals, such as ‘puffy’ and ‘skinny’ intervals [26]. The resulting insights are distilled, for the first time, to guide readers in selecting the regression algorithms appropriate to their data. Further, we provide open-source software at http://lucidresearch.org/software, offering, for the first time, direct access to IV regression approaches, with a view to facilitating adoption and replication.

4) **Explainability and visualization**: The explainability of numeric linear regression models is one of their primary assets as an AI technique. This paper introduces a novel visualization approach for IV regression—the Interval Regression Graph (IRG)—which for the first time provides the means to visualize and succinctly communicate the relationship in terms of both position and range between IV regressor and regressand.

The rest of this article is organized as follows. Section II reviews linear regression methods based on vector/matrix representation of intervals in the regression process. Section III introduces the IRG and discusses adaptations of the LM method [20] for real-world applications, followed by IV datasets (synthetic and real-world) and evaluation metrics. Section IV demonstrates the estimation performance of all regression models with respect to both synthetic and real-world datasets. Section V discusses suitability of the existing regression approaches to salient features of IV data, distilling insights into practical guidance for algorithm-selection. Lastly, Section VI concludes this article. Nomenclature presents a list of acronyms and notation used in this article to assist the reader.

II. BACKGROUND AND STATE OF THE ART

A closed interval $\pi$ is characterized by its two endpoints, $a^-$ and $a^+$ with $a^- \leq a \leq a^+$ where $a^-$ and $a^+$ are respectively the lower and upper bounds of $\pi$ [29]. Generally, $\pi$ is denoted as $\pi = [a^-, a^+]$, and its range is given by $a^{\pi} = |a^+ - a^-|$. Alternatively, the same interval can be represented as $\pi = [a^c - a^e, a^c + a^e]$ with its center, $a^c = \frac{(a^+ + a^-)}{2}$ and radius, $a^e = \frac{a^{\pi}}{2}$. In this section, we review linear regression methods for full IV datasets (i.e., where both dependent and independent variables are intervals) where intervals are treated as bi-variate vectors in the regression process. We note $Y = \{\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n\}$ as an IV regressand with $n$ intervals, where $\bar{y}_i = [y_{ij}, y_{ij}^+]$, $1 \leq i \leq n$. Furthermore, $\{X_1, X_2, \ldots, X_p\}$ presents $p \geq 1$ IV regressor(s) where each $X_j$ has also $n$ intervals, $X_j = \{x_{ji}, x_{ji}^+, \ldots, x_{jn}\}$ with $x_{ji} = [x_{ji}^- , x_{ji}^+]$, $1 \leq i \leq n$, $1 \leq j \leq p$. Thus, there are $p$ pairs of $(Y, X)$ in the dataset. In addition, the estimated value of $Y$ is given by $\hat{Y} = \{\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n\}$.

Since about the year 2000, a variety of linear regression approaches for full IV datasets have been put forward. We briefly review them below in chronological order.

A. Center Method (CM) (2000)

Billard and Diday [15] proposed a linear regression model, known as the CM in 2000, to fit the center of both regressand and regressors, and then apply this model to the lower and upper bounds of regressors to separately estimate the upper and lower bounds of the regressand. Eq. (1) defines the CM regression model

$$y_i^- = \beta_0^- + \sum_{j=1}^{p} \beta_j^- x_{ji}^- + \epsilon_i^-$$

$$y_i^+ = \beta_0^+ + \sum_{j=1}^{p} \beta_j^+ x_{ji}^+ + \epsilon_i^+$$

Note that in this paper, we focus on establishing and visualizing the relationship between the variables of a given IV data set. We do not focus on prediction, a different problem setting with implications for example the need to establish model and prediction robustness, e.g., using crossvalidation [28]-offering scope for future research for IV data.
where $y_i^c$ and $y_i^T$ are the lower and upper bounds of $\bar{y}_i$. Similarly, $x_{ji}^+$ and $x_{ji}^-$ are the lower and upper bounds of $\bar{x}_{ji}$. $\beta_j^c \in \mathbb{R}$ are the regression coefficients for the center estimation, $0 \leq j \leq p$, $p$ is the total number of regressors while $\epsilon_i^c$ and $\epsilon_i^+ \epsilon_i^T$ are error terms. Both lower and upper bound models of (1) use the same coefficients, therefore, they can be captured together using the following equation:

$$ y_i^c = \beta_0^c + \sum_{j=1}^{p} \beta_j^c x_{ji}^c + \epsilon_i^c $$ (2)

where $y_i^c$ and $x_{ji}^c$ are the center of $\bar{y}_i$ and $\bar{x}_{ji}$, respectively, and $\epsilon_i^c = (\epsilon_i^- + \epsilon_i^+)/2$ is error term. Eq. (2) can also be expressed for the total set of $n$ observations in the matrix format as $Y^c = X^c \beta^c + \epsilon^c$, where

$$ Y^c = \begin{pmatrix} y_1^c \\ y_2^c \\ \vdots \\ y_n^c \end{pmatrix}, \quad \beta^c = \begin{pmatrix} \beta_0^c \\ \beta_1^c \\ \vdots \\ \beta_p^c \end{pmatrix}^T, \quad \epsilon^c = \begin{pmatrix} \epsilon_1^c \\ \epsilon_2^c \\ \vdots \\ \epsilon_n^c \end{pmatrix}^T, \quad \text{and } X^c = \begin{pmatrix} 1 & x_{11}^c & \cdots & x_{1n}^c \\ 1 & x_{21}^c & \cdots & x_{2n}^c \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1}^c & \cdots & x_{nn}^c \end{pmatrix}. $$

The coefficients, $\beta^c$, are estimated using the least squares (LSs) method [30], defined in the following equation:

$$ \hat{\beta}^c = ((X^c)^T X^c)^{-1} (X^c)^T Y^c. $$ (3)

By using $\hat{\beta}^c$, (4) estimates the bounds of regressor, $Y$

$$ \hat{y}_i^- = \hat{\beta}_0^c + \sum_{j=1}^{p} \hat{\beta}_j^c x_{ji}, $$

$$ \hat{y}_i^+ = \hat{\beta}_0^c + \sum_{j=1}^{p} \hat{\beta}_j^c x_{ji}. $$ (4)

The CM model is simple, as it follows a standard regression approach applied to the center of the intervals and then the resulting model is applied both to upper and lower endpoints. However, by doing so, it risks and often fails to maintain mathematical coherence [17], that is, the estimated lower bound of $\bar{y}_i$ becomes larger than the estimated upper bound.

B. MinMax Method (2002)

Billard and Diday [16] presented the MinMax method in 2002, which directly uses the lower and upper bounds of the regressors to separately estimate the lower and upper bounds of the regressand. Eq. (5) defines the MinMax regression model for the lower and upper bounds of the regressand

$$ y_i^- = \beta_0^- + \sum_{j=1}^{p} \beta_j^- x_{ji} + \epsilon^-i $$

$$ y_i^+ = \beta_0^+ + \sum_{j=1}^{p} \beta_j^+ x_{ji} + \epsilon^+i $$ (5)

where $\beta_j^-$ and $\beta_j^+$ are regression coefficients for the lower and upper bounds, respectively $(\beta_j^-, \beta_j^+ \in \mathbb{R}, 0 \leq j \leq p)$. In matrix notation, the lower bound model of (5) with all $n$ observations is expressed as $Y^- = X^- \beta^- + \epsilon^-$, where

$$ Y^- = \begin{pmatrix} y_1^- \\ y_2^- \\ \cdots \\ y_n^- \end{pmatrix}, \quad \beta^- = \begin{pmatrix} \beta_0^- \\ \beta_1^- \\ \cdots \\ \beta_p^- \end{pmatrix}^T, \quad \epsilon^- = \begin{pmatrix} \epsilon_1^- \\ \epsilon_2^- \\ \cdots \\ \epsilon_n^- \end{pmatrix}^T, \quad \text{and } X^- = \begin{pmatrix} 1 & x_{11}^- & \cdots & x_{1n}^- \\ 1 & x_{21}^- & \cdots & x_{2n}^- \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1}^- & \cdots & x_{nn}^- \end{pmatrix}. $$

Like the CM method, (6) estimates the coefficient $(\beta^-)$ for the lower bound model of (5) using the LS method [30]

$$ \hat{\beta}^- = ((X^-)^T X^-)^{-1} (X^-)^T Y^- $$ (6)

Now, using the estimated $\hat{\beta}^-$, (7) estimates the lower bounds for the regressand, $Y$

$$ \hat{y}_i^- = \hat{\beta}_0^- + \sum_{j=1}^{p} \hat{\beta}_j^- x_{ji}. $$ (7)

Similarly, the upper bound model of (5) is expressed as $Y^+ = X^+ \beta^+ + \epsilon^+$, where

$$ Y^+ = \begin{pmatrix} y_1^+ \\ y_2^+ \\ \cdots \\ y_n^+ \end{pmatrix}, \quad \beta^+ = \begin{pmatrix} \beta_0^+ \\ \beta_1^+ \\ \cdots \\ \beta_p^+ \end{pmatrix}^T, \quad \epsilon^+ = \begin{pmatrix} \epsilon_1^+ \\ \epsilon_2^+ \\ \cdots \\ \epsilon_n^+ \end{pmatrix}^T, \quad \text{and } X^+ = \begin{pmatrix} 1 & x_{11}^+ & \cdots & x_{1n}^+ \\ 1 & x_{21}^+ & \cdots & x_{2n}^+ \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1}^+ & \cdots & x_{nn}^+ \end{pmatrix}. $$

Eq. (8) estimates the coefficients for the upper bound of (5) by the LS method [30]

$$ \hat{\beta}^+ = ((X^+)^T X^+)^{-1} (X^+)^T Y^+. $$ (8)

Now, using $\hat{\beta}^+$, (9) estimates the upper bounds of the regressand, $Y$

$$ \hat{y}_i^+ = \hat{\beta}_0^+ + \sum_{j=1}^{p} \hat{\beta}_j^+ x_{ji}. $$ (9)

In the MinMax method, the use of two separate models to estimate the lower and upper bounds of the regressand improves the model fitness and interpretation compared to the CM method [17]. However, it still does not guarantee mathematical coherence on the estimated bounds [17]. Furthermore, its estimation performance can be reduced if there is no dependency between the bounds of regressand and regressors [21].

C. Center and Range Method (CRM) (2008)

Neto and Carvalho [17] extended the CM method [15] in 2008 to the CRM, by considering both interval center and half the “range” (i.e., the radius) of the regressor and regressand variables. They build two separate regression models—one for the centers and other for the ranges (or radii). Eq. (10) is used to model the interval centers

$$ y_i^c = \beta_0^c + \sum_{j=1}^{p} \beta_j^c x_{ji} + \epsilon_i^c $$ (10)

where $y_i^c$ and $x_{ji}$ are the centers $\bar{y}_i$ and $\bar{x}_{ji}$, respectively, and $\epsilon_i^c$ is error term. The CRM method follows the same matrix format of
the CM method to represent the center model of (10) and applies the same strategy of the CM method to estimate the coefficients $\beta^T_j$ [see (3)]. It then estimates the center values for the regressand, $Y$, using the estimated $\hat{\beta}^T_j$ by the following equation:

$$\hat{y}^c = \hat{\beta}^T_0 + \sum_{j=1}^{p} \hat{\beta}^T_j x_{ji}.$$  (11)

Similarly, the interval ranges are modeled by the CRM method using the following equation:

$$y_i^c = \beta^T_0 + \sum_{j=1}^{p} \beta^T_j x^T_{ji} + \epsilon^T_i$$  (12)

where $y_i^c$ and $x^T_{ji}$ are the radius of $y_i$ and $x_{ji}$, respectively. $\beta^T_j \in \mathbb{R}$ are the coefficients for the range estimations, $0 \leq j \leq p$ and $\epsilon^T_i$ is error term. Again, following the same matrix notation, the coefficients, $\beta^T_j$ are estimated by the following equation:

$$\hat{\beta}^T = \left((X^T X)^{-1} (X^T Y)\right)^T$$  (13)

and using the estimated $\hat{\beta}^T$, (14) computes the ranges of the regressand, $Y$

$$\hat{y}_i = \hat{\beta}^T_0 + \sum_{j=1}^{p} \hat{\beta}^T_j x^T_{ji}.$$.  (14)

Finally, the CRM method estimates the bounds of the regressand from the estimated $\hat{y}_i^c$ and $\hat{y}_i^c$ using the following equation:

$$\hat{y}_i = \hat{y}_i^c - \delta_i$$ and $\hat{y}_i^c = \hat{y}_i^c + \delta_i$.

(15)

The CRM model provides better estimation when there is a linear dependency between the ranges of regressand and regressors [18], [21]. If this is not the case, it does not ensure mathematical coherence as (14) can produce negative $y_i^c$ [21].

E. Complete Information Method (CIM) (2012)

Wang et al. [19] presented the CIM in 2012 with a focus on considering all the information contained within the intervals. The CIM model considers each interval observation of regressand and regressor variables as a hypercube (if a single valued regressor and associated regressand are considered, each observation effectively reflects a rectangular, as visualized in Fig. 4), and the regression model is built on these hypercubes. Furthermore, it adopts Moore’s linear combination algorithm [24] to ensure the mathematical consistency of the bounds. In this model, $Y$ is presented by a linear combination of $X_j$

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon_i$$  (16)

where $\beta_j$ represents the coefficients with $0 \leq j \leq p$ and $1_n$ is an $n$-dimensional column constant vector of ones. Eq. (17) defines the lower and upper bound estimations of $Y$

$$y_i^- = \beta_0 + \sum_{j=1}^{p} \beta_j (\tau^- x^T_{ji} + (1 - \tau^-)x^+_{ji})$$

$$y_i^+ = \beta_0 + \sum_{j=1}^{p} \beta_j ((1 - \tau^-)x^-_{ji} + \tau^- x^+_{ji})$$  (17)

where

$$\tau_j = \begin{cases} 
0, & \text{if } \beta_j \leq 0 \\
1, & \text{otherwise}.
\end{cases}$$

Using (16), the LS method [30] of (17) generates the following linear system:

$$M \beta = b$$  (18)

where

$$M = \begin{pmatrix} 
\langle 1,1 \rangle & \langle 1,X_1 \rangle & \ldots & \langle 1,X_p \rangle \\
\langle X_1,1 \rangle & \langle X_1,X_1 \rangle & \ldots & \langle X_1,X_p \rangle \\
\vdots & \vdots & \ddots & \vdots \\
\langle X_p,1 \rangle & \langle X_p,X_1 \rangle & \ldots & \langle X_p,X_p \rangle 
\end{pmatrix}$$

$$\beta = \begin{pmatrix} 
\beta_0 \\
\beta_1 \\
\vdots \\
\beta_p \end{pmatrix}$$

and $b = \begin{pmatrix} 
\langle 1,Y \rangle \\
\langle X_1,Y \rangle \\
\vdots \\
\langle X_p,Y \rangle 
\end{pmatrix}$.

Here, $\langle X_a, X_b \rangle$ is the inner product between $X_a$ and $X_b$ ($0 \leq a, b \leq p$ and $X_0 = 1$), which is computed using the following equation:

$$\langle X_a, X_b \rangle = \begin{cases} 
\frac{1}{3} \sum_{i=1}^{n} (x_{ai}^{-2} + x_{ai}^{-+} + x_{ai}^{++}) \text{,} & \text{if } X_a = X_b \\
\frac{1}{4} \sum_{i=1}^{n} (x_{ai}^{-}x_{ai}^{+} + x_{ai}^{-+} + x_{ai}^{++}) \text{,} & \text{if } X_a \neq X_b.
\end{cases}$$  (19)
From (18), the coefficients $\beta$ can be estimated by the inverse of matrix $M$ as defined in the following equation:

$$\hat{\beta} = (M)^{-1}b$$  \hspace{1cm} (20)

Thus, using $\hat{\beta}$ and $\hat{\tau}_j$, the lower and upper bounds of the regressand, $Y$ can be estimated by (17).

The CIM method satisfies mathematical coherence of the estimated bounds of the regressand using $\tau_j$ where its value $(1$ or $0$) depends on the sign $(+$ or $-)$ of $\beta_j$ [19].

F. Linear Model (LM) (2015)

Sun and Ralescu [20] developed the LM in 2015 based on the affine operator in the cone $\mathcal{C} = \{(x, y) \in \mathbb{R}^2 | x \leq y\}$, aiming to maximize model flexibility while preserving its interpretability. This model considers both lower and upper bounds of regressors and their ranges for estimating the bounds of the regressand. In this model, $\bar{y}_i$ is expressed as a linear transformation of $\pi_{ji}$ such that

$$y_i^- = \sum_{j=1}^{p} (\alpha_j x_{ji}^- + \beta_j x_{ji}^+) + \eta + \epsilon_i$$

$$y_i^+ = \sum_{j=1}^{p} ((\alpha_j - \gamma_j)x_{ji}^- + (\beta_j + \gamma_j)x_{ji}^+) + \eta + \theta + \epsilon_i$$  \hspace{1cm} (21)

where the regressor coefficients, $\alpha_j, \beta_j, \eta \in \mathbb{R}$ and $\gamma_j, \theta \geq 0$, $j = 1, 2, \ldots, p$. Eq. (21) can also be written in the following matrix form as $Y = X\beta + \epsilon$, where

$$Y = \begin{pmatrix} Y^- \\ Y^+ \end{pmatrix} \quad \text{with} \quad Y^- = \begin{pmatrix} y_1^- \\ \vdots \\ y_n^- \end{pmatrix} \quad \text{and} \quad Y^+ = \begin{pmatrix} y_1^+ \\ \vdots \\ y_n^+ \end{pmatrix}$$

$$X = \begin{pmatrix} X^- \\ X^+ \end{pmatrix} \quad \text{with} \quad X^+ = \begin{pmatrix} x_{11}^- & x_{11}^+ & \cdots & x_{1p}^- \\ x_{12}^- & x_{12}^+ & \cdots & x_{1p}^+ \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}^- & x_{n1}^+ & \cdots & x_{np}^+ \end{pmatrix}$$

and

$$X^w = \begin{pmatrix} x_{11}^w \\ x_{12}^w \\ \vdots \\ x_{n1}^w \end{pmatrix}$$

and $\theta = \begin{pmatrix} \theta \\ \vdots \\ \theta \end{pmatrix}$, and $\beta = \begin{pmatrix} \beta^a \\ \beta^w \end{pmatrix}$

Again, the LS estimate of the coefficients, $\hat{\beta}$ are computed by the following equation:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$  \hspace{1cm} (22)

and (23) estimates the bounds of the regressand, $Y$

$$\hat{y}_i^- = \sum_{j=1}^{p} (\hat{\alpha}_j x_{ji}^- + \hat{\beta}_j x_{ji}^+) + \hat{\eta}$$

$$\hat{y}_i^+ = \sum_{j=1}^{p} ((\hat{\alpha}_j - \hat{\gamma}_j)x_{ji}^- + (\hat{\beta}_j + \hat{\gamma}_j)x_{ji}^+) + \hat{\eta} + \hat{\theta}$$.  \hspace{1cm} (23)

Even though the authors assume positive constraints on range coefficients (i.e., $\gamma_j \geq 0$ and $\theta \geq 0$) in (21) to maintain mathematical coherence, the actual model does not ensure compliance with these constraints. As a result, (22) can result in negative range coefficients $(\hat{\gamma}_j, \hat{\theta} < 0)$, which may lead to flipped interval bounds $(\hat{y}_i^- > \hat{y}_i^+)$ through (23). The authors do not discuss how to maintain these constraints in practice, though they expect that if any of the estimates of $\hat{\gamma}_j$ and $\hat{\theta}$ turn out to be negative, forcing it to be positive may lead to poor fitness of the LM model [20].

G. Parametrized Model (PM) (2017)

Souza et al. [21] proposed the PM in 2017, which also builds two different models for the regressand bounds. This approach extracts the best reference points from the regressors and uses them to build linear regression models for both lower and upper bounds of the regressand, unlike earlier approaches that use specific reference points on the regressors, such as center, range, and interval bounds.

In this method, an interval is considered as a line segment, useful to reach all points within it. For instance, given an interval $\pi$, any point $q \in \pi$ can be computed as $q = a^- (1 - \lambda) + a^+ \lambda$, $0 \leq \lambda \leq 1$. By setting $\lambda$, $\pi$ is turned into a single point. Hence, when $\lambda = 0$, $q = a^-$ (lower bound of $\pi$) and when $\lambda = 1$, $q = a^+$ (upper bound of $\pi$). Similarly, $q = a^0$ (center of $\pi$) when $\lambda = 0.5$. Utilizing this concept, the PM method specifies the linear regression models for the lower and upper bounds of $Y$ in the following equation:

$$y_i^- = \beta_0^0 + \sum_{j=1}^{p} \beta_j^- (1 - \lambda_j) x_{ji}^- + \beta_j^+ \lambda_j x_{ji}^+ + \epsilon_i$$

$$y_i^+ = \beta_0^+ + \sum_{j=1}^{p} \beta_j^+ (1 - \lambda_j) x_{ji}^- + \beta_j^- \lambda_j x_{ji}^+ + \epsilon_i$$  \hspace{1cm} (24)

Eq. (25) simplifies (24) by replacing $\beta_j^- (1 - \lambda_j)$ by $\alpha_j^-$ and $\beta_j^+ \lambda_j$ by $\alpha_j^+$ and $\omega_j^+$ respectively

$$y_i^- = \beta_0^- + \sum_{j=1}^{p} \alpha_j^- x_{ji}^- + \omega_j^+ x_{ji}^+ + \epsilon_i$$

$$y_i^+ = \beta_0^+ + \sum_{j=1}^{p} \alpha_j^+ x_{ji}^- + \omega_j^+ x_{ji}^+ + \epsilon_i$$  \hspace{1cm} (25)
In matrix notation, the lower bound model can be expressed for all \( n \) observations as \( Y^- = X^* \beta^- + \epsilon^- \), where

\[
Y^- = (y^-_1, y^-_2, \ldots, y^-_n)^T, \quad \beta^- = (\beta^-_0, \omega^-_1, \omega^-_2, \ldots, \omega^-_p)^T
\]

\[
\epsilon^- = (\epsilon^-_1, \epsilon^-_2, \ldots, \epsilon^-_n)^T, \quad \text{and } X^* = \begin{pmatrix} 1 & x^-_1 & x^-_2 & \cdots & x^-_p \\ 1 & x^-_{i1} & x^-_{i2} & \cdots & x^-_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x^-_{in} & x^-_{i2} & \cdots & x^-_{ip} \end{pmatrix}.
\]

The LS estimate of the coefficients for the lower bound model, \( \hat{\beta}^- \), is computed by the following equation:

\[
\hat{\beta}^- = ((X^*)^T X^*)^{-1} (X^*)^T Y^-.
\]

The matrix expression follows the same pattern for the upper bound model, \( Y^+ = X^* \beta^+ + \epsilon^+ \), and the LS estimate of the coefficients for the upper bound model, \( \beta^+ \), in defined in the following equation:

\[
\hat{\beta}^+ = ((X^*)^T X^*)^{-1} (X^*)^T Y^+.
\]

Finally, using \( \hat{\beta}^- \) and \( \hat{\beta}^+ \), the lower and upper bounds of \( Y \) are estimated using the following equation:

\[
\hat{y}^- = \hat{\beta}^-_0 + \sum_{j=1}^{p} \hat{\alpha}_j x^-_{ji} + \hat{\omega}_j^- x^+_j,
\]

\[
\hat{y}^+ = \hat{\beta}^+_0 + \sum_{j=1}^{p} \hat{\alpha}_j x^-_{ji} + \hat{\omega}_j^+ x^+_j.
\]

The PM method does not automatically guarantee the mathematical coherence of the bounds [21]. To avoid flipping the interval bounds, the approach estimates the range of \( Y \) using (29) before performing the regression

\[
\hat{Y}^w = X^*((X^*)^T X^*)^{-1} (X^*)^T Y^w.
\]

If all estimated ranges are positive (\( \hat{y}^w \in \hat{Y}^w \)), the model automatically ensures mathematical coherence. However, if at least one of the estimated ranges is negative, it applies the box-cox transformation [25], extended to intervals by the authors [21] [defined in (30)], to transform the regressand so that the desirable coherence is achieved by the PM method

\[
\hat{y}_i^k = \begin{cases} (y^-_i + k_2)^{\gamma_i} - 1, & \text{if } k_1 \neq 0 \\ \log(y^-_i + k_2), & \text{if } k_1 = 0 \end{cases},
\]

where \( \hat{y}_i^k \) is the transformed interval of \( \hat{y}_i \), \( k_1 \) is any real value and \( k_2 \) maintains the following restriction: \( k_1 + k_2 > 0 \).
Here, we first regress the range of the regressand variable \( (y_i^n) \) on the range of regressors \( (x_i^n) \) before performing the estimation using (29) as the estimated range values of the regressand guarantee mathematical coherence of the estimated bounds if they are positive [21]. Therefore, if all estimated range values are positive \( (y_i^n \geq 0, \forall i) \), we compute the LSs estimation of (22) without any parameter restrictions. In contrast, if any of the estimated range values are negative \( (y_i^n < 0, \exists i) \), we enforce positivity restrictions on \( \gamma_j \) and \( \hat{\theta} \) using the LHA [18], [23]. Algorithm 2 summarizes this adapted approach, LMw.

### B. Interval Regression Graph

Visualization of traditional regression models provides a powerful mechanism for interpretability and communication of the relationship between numeric variables. While intervals provide a richer model, enabling the capture of uncertain information, they are positive \( (y_i^n \geq Y_i^\gamma \) on the vertical axis, \( X_i^w \) on the bottom left axis and \( X_i^c \) on the bottom right axis.

Algorithm 3: Interval Regression Graph (IRG) Generation.

**Input:** An IV regression model (IVRM). The IV regressor’s (e.g., from the original data set) minimum range and maximum range are \( range X_{min} \) and \( range X_{max} \), as well as its minimum center and maximum center coordinates are \( center X_{min} \) and \( center X_{max} \).

**Output:** Two IRG plots mapping the regressor to the regressand’s center, \( \bar{Y} \) and range, \( Y^w \).

1. Generate the set \( X^c \) of \( q \) discretizations of the interval \( [center X_{min}, center X_{max}] \).
2. Generate the set \( X^w \) of \( p \) discretizations of the interval \( [center X_{min}, center X_{max}] \).
3. for each discretized \( X^w, 1 \leq i \leq p \) do
4. for each discretized \( X^c, 1 \leq j \leq q \) do
5. Compute \( X^c_{ij} = X^c_j - X^w_i \)
6. Compute \( X^c_{ij} = X^c_j + X^w_i \)
7. Compute \( \hat{Y}^{ij}_{ij}, \hat{Y}^{ij}_{+} \) with \( X^c_{ij}, X^w_{ij} \) using IVRM
8. Compute \( \hat{Y}^{ij}_{ij} = \hat{Y}^{ij}_{ij} + \hat{Y}^{ij}_{+} \)
9. Compute \( \hat{Y}^{ij}_{w} = \hat{Y}^{ij}_{+} - \hat{Y}^{ij}_{ij} \)
10. end for
11. end for
12. Generate \( \hat{Y}^{c} \) IRG plot with \( \hat{Y}^{c} \) on the vertical axis, \( X^w \) on the bottom left axis and \( X^c \) on the bottom right axis.
13. Generate \( \hat{Y}^{w} \) IRG plot with \( \hat{Y}^{w} \) on the vertical axis, \( X^w \) on the bottom left axis and \( X^c \) on the bottom right axis.
14. return IRG plots for \( \hat{Y}^{c} \) and \( \hat{Y}^{w} \)

Note that the IRG generation algorithm is designed to allow for the visualization of both linear and non-linear relationships between variables. In this paper, we focus on linear relationships/regression models only.

**Fig. 3.** Relationship between regressand and regressor with respect to center (position) (a) and range (b) using the PM method for Set-1 (Fig. 1(a)). (a) IRG as to \( \hat{Y}^{c} \). (b) IRG as to \( \hat{Y}^{w} \) (\( \geq 2Y^+ \))
for Set-1 based on the PM regression method. The bottom-left and bottom-right axes show the regressor's center ($X^c$) and range ($X^w \sim 2X^r$), respectively. In Fig. 3(a), the vertical axis denotes the regressand’s estimated center [$Y^c$ (solid plane)], while in Fig. 3(b), it reflects the regressand’s estimated range [$Y^w \sim 2Y^r$ (dotted plane)].

Interpreting these figures, we can see how Fig. 3(a) visualizes that the regressand’s center, $Y^c$ increases with respect to increasing values of both the regressor’s range, $2X^r$ (or $X^w$) and its center, $X^c$. Fig. 3(b) shows how the regressand’s range, $2Y^r$ also increases with respect to both increasing values of the regressor’s range, $2X^r$ and center, $X^c$, and that it does so at a greater rate in each case than does $Y^c$. For compact and complete visualization, IRGs commonly combine the visualizations with respect to both the regressand’s center and range, as shown in Fig. 18(f). We will leverage IRGs throughout the remainder with additional examples.

C. Datasets

We use both synthetic and real-world IV datasets to investigate the performance of linear regression methods. The synthetic datasets are used to explore and visualize the sensitivity of the regression models with respect to different properties of the interval data, whereas the real-world datasets are used to investigate their suitability to real-world scenarios. All datasets are described as follows.

1) Synthetic Datasets: We specifically design a series of synthetic IV datasets to investigate the behavior of the regression models. First, Set-1 and Set-2, shown at the beginning of this article [see Figs. 1(a) and 2(a)], are used to explore the relationship between regressor and regressand with respect to the primary features (i.e., center and range). Second, an additional nine sets of intervals (Set-3 to Set-11), exhibiting different properties are used: we use three sets of skinny intervals (Set-3 to Set-5) with smaller range and three puffy sets with wider range (Set-6 to Set-8) having features of disjointedness, some, and full overlap. We also consider three mixed cases (Set-9 to Set-11), containing puffy and skinny intervals in separated, partially overlapping and nested states. All of these datasets are produced within the range of 0 to 10. They are given in Table VI(a)–(k) in the Appendix including the mean (μ) and standard deviation (σ) of the center and range of intervals. Fig. 4 graphically presents the datasets, Set-3 to Set-11.

2) Real-World Datasets: We focus on three real-world IV datasets with different properties and arising in different application contexts, i.e., medical, consumer food ratings/marketing, and cyber-security, reflecting the broad applicability of IV data in real-world applications.

The first dataset represents medical observations and includes IV systolic and diastolic blood pressure (BP) data of 59 patients of the Hospital Valle del Nalón in Asturias, Spain [11], [14].

The second dataset presents IV consumer ratings of eight different (U.K. market) snack-food products [32]. In this case, 40 consumers rated each product—using “DECSYS” interval survey software [33]—on six different attributes: visual appeal, taste, overall purchase intention (i.e., how likely they were to buy each product).

The third and final dataset captures assessments by cybersecurity experts of the vulnerability of system components in a real-world scenario. Specifically, 38 cybersecurity experts at the UK CESG (Communications-Electronics Security Group) assessed a range of components (referred to as hops) that are commonly encountered during a cyber-attack [34], [35]. They rated the hops on overall difficulty for an attacker to either attack or evade them, as well as rating each on several attributes (seven

3The “skinny” and “puffy” designs have been inspired by Ferson et al. [26].
for attack, and three for evade) that might affect this difficulty. The experts provided their IV ratings on a scale from 0 to 100. For more details of these datasets, see [11], [14], [32], [34], [35].

D. Evaluation Metrics

Evaluation poses one of the key challenges for IV data analysis and AI more broadly. While “quality of fit” for numeric data is straightforward, the same is not the case for intervals. Comparing the latter requires application-led assumptions (on the nature of the intervals and whether for example interval size is more important than position), which in turn feed into similarity measures or arithmetic means of comparison. We do not delve further into this challenge in this article, instead following three fairly standard metrics to evaluate the performance of the linear regression models. Each of the evaluation metrics is defined with respect to the regressand (samples), $Y = \{y_1, y_2, \ldots, y_n\}$ and its estimated value(s), $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$.

1) **Root mean squared error (RMSE)** estimates error by considering the average deviation of the estimated values from the actual ones [18]. The RMSE for the lower and upper bounds of $Y$ is defined as

$$\text{RMSE}^- = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i^+ - \hat{y}_i^-)^2}$$

$$\text{RMSE}^+ = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i^- - \hat{y}_i^+)^2}$$

where a smaller RMSE value indicates a better fitted model.

2) **Mean absolute error (MAE)** calculates the average of absolute difference between the actual and estimated values [36]. For the upper and lower bounds of $Y$, the MAE is defined as

$$\text{MAE}^- = \frac{1}{n} \sum_{i=1}^{n} |y_i^+ - \hat{y}_i^-|$$

$$\text{MAE}^+ = \frac{1}{n} \sum_{i=1}^{n} |y_i^- - \hat{y}_i^+|$$

where $0 \leq \text{MAE}^-, \text{MAE}^+ \leq \mathbb{R}^+$. Again, a lower value of MAE implies a better fitted model.

3) **Mean magnitude of relative error (MMRE)** measures the mean of dispersion of the actual and estimated values with respect to the actual values [37]. It is calculated as follows:

$$\text{MMRE} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \left( \frac{|y_i^+ - \hat{y}_i^-|}{y_i^-} + \frac{|y_i^- - \hat{y}_i^+|}{y_i^+} \right)$$

where $0 \leq \text{MMRE} \leq \mathbb{R}^+$. In this case, the lower the MMRE, the better the model.

A. Performance Analysis With Synthetic Datasets

In Section III-B, we briefly introduced the IRGs using IV Set-1 [see Fig. 1(a)]. We expand on this example here, reporting performance of all regression models covered above with this set, as well as for Set-2 [see Fig. 2(a)].

As discussed earlier, Set-1 shows the more commonly considered and intuitive scenario where uncertainty and position in regressor and regressand vary in unison. On the other hand, Set-2 presents a less commonly considered scenario, where the uncertainty of the regressand increases irrespective of the uncertainty of the regressor. Said differently, for Set-2, uncertainty and position of the regressand vary quasi exclusively with respect to the position of the regressor variable.

Tables II and III present the fitness results for all regression approaches for Sets 1 and 2, respectively. These show that the PM method provides the best fitness in each case, though the MinMax method provides almost equal fitness for Set-2 (they do differ with respect to MAE$^+$). For conciseness, we focus further discussion of the regression results on two of the regression approaches: the CRM—representing an earlier and more basic discussion of the regression results on two of the regression approaches: the CRM and PM—the most recent approach, which does. IRGs for all methods are shown in the Appendix (see Figs. 18 and 19). We independently present the relationship of the center and range of the test regressor on the center [see (a)] and the range [see (b)] of the regressand, for the CRM and PM methods in Figs. 3, 5, 6, and 7. Note that, Figs. 3(b) and 6(b) are repeated for clarity from Figs. 1(b) and 2(b) in Section I. The IRGs in Figs. 3(a) and (b) show that, according to the PM method, both changes in the estimated center and range of Set-1’s regressand are associated with the changes in the center and range of the regressor—in-line with our expectation. On the other hand, the IRGs in Figs. 5(a) and (b) show that for the CRM method, the estimated center of Set-1’s regressand is influenced only by the center of the regressor, and the estimated range of Set-1 is influenced only by the range of the regressor. Thus, the CRM method is not in line with our expectation for the Set-1 dataset. The rest of the
methods also do not perform according to the expectation (see Appendix, Fig. 18).

Furthermore, the IRGs in Figs. 6(a) and (b) show how, according to the PM model, both center and range of the regressand of Set-2 vary solely with respect to the center of the regressor—in line with expectation. The IRG in Fig. 7(a) visualizes how, using the CRM method, the estimated center of Set-2’s regressand is similarly positively influenced by the changes in the center, but not the range of the regressor—as expected. However, in contrast to the PM model outputs, the CRM-derived IRG for the same data, shown in Fig. 7(b), indicates no strong relationship between regressand range and either facet of the regressor. If any relationship is present at all, then this is a negative effect of the range of the regressor on that of the regressand—these results do not fit those expected when viewing the data. Fig. 19 (in the Appendix) visualizes that the MinMax method shows a similar performance to the PM method. The remaining methods perform less well.

Next, we will briefly discuss equivalent analysis and model visualizations, using the IRGs, for an additional nine synthetic datasets (Set-3—Set-11), each designed to reflect a key property regularly encountered in IV datasets. This will highlight how different datasets— with different features—are differently suitable for different regression methods, as later results will expand upon in more detail. For data Set-3 to Set-11, all shown in Fig. 4, we briefly summarize individual performance results across methods below before discussing performance more broadly. Table IV(a)—(i) presents the fitness of regression methods for each of these datasets.

1) **Set-3 (All disjoint skinny intervals):** The PM and LM methods produce the best fit for all disjoint skinny intervals. The LM method results in equally good fit well with respect to two metrics (RMSE and MMRE).

2) **Set-4 (Some overlapping skinny intervals):** Again, the PM and LM methods produce the best fit. The LM method also gives the same fitness with respect to three metrics (RMSE, MAE, and MMRE). The fitness of other methods improves with some overlapping of intervals.

3) **Set-5 (All overlapping skinny intervals):** The PM and MinMax methods come out as the best approach. Among others, CRM, CCRM, LMc, and LMw methods produce quite similar results with respect to some of the evaluation metrics.

4) **Set-6 (All disjoint puffy intervals):** The PM method gives the best fit, though the LM and LM methods also produce quite similar results for puffy intervals.

5) **Set-7 (Some overlapping puffy intervals):** The PM and LM methods emerge as the best approaches with respect to different metrics, though the MinMax, CRM, CCRM, and LM methods also produce similarly good fit in terms of MMRE.

6) **Set-8 (All overlapping puffy intervals):** For all overlapping puffy intervals, the PM method gives the best fit with respect to three metrics and the LM method is for two evaluation metrics. Except CM and CIM methods, growing overlap among intervals appears to lower estimation errors by the CRM, CCRM, and MinMax methods.

7) **Set-9 (Disjoint mixed intervals):** The LM, LM, and PM methods perform best. All other methods result in relatively low fitness. In particular, the CM and CIM methods achieve the worst results.

8) **Set-10 (Some overlapping mixed intervals):** The PM, LM, and LM methods result in similar model fit for almost all metrics. The CRM and CCRM methods give better fitness results with respect to one metric (MMRE).

9) **Set-11 (Nested/all overlapping mixed intervals):** The LM, LM, and PM methods give the best fit. The CRM and CCRM methods also closely fit the data as to MMRE. Overall, the PM, and LM (and often LM) methods perform best and result in reliably good—and almost identical—fitness for disjoint skinny or puffy intervals. With growing overlap among intervals (skinny or puffy), model fitness by the CRM,
TABLE IV
PERFORMANCE OF DIFFERENT REGRESSION MODELS FOR IV SYNTHETIC DATASETS (SET-3 TO SET-11)

| Method      | RMSE* | RMSE− | MAE* | MAE− | MMRE |
|-------------|-------|-------|------|------|------|
| CM          | 1.215 | 1.181 | 0.672 | 1.035 | 0.185 |
| MinMax      | 1.204 | 1.169 | 1.004 | 1.002 | 0.185 |
| CRM         | 1.185 | 1.149 | 0.666 | 1.021 | 0.185 |
| CCRM        | 1.195 | 1.179 | 0.065 | 1.011 | 0.185 |
| CMM         | 1.215 | 1.181 | 0.672 | 1.035 | 0.185 |
| LM          | 0.535 | 0.293 | 0.920 | 0.312 | 0.509 |
| LM−         | 0.343 | 0.392 | 0.555 | 0.312 | 0.059 |

| Method      | RMSE* | RMSE− | MAE* | MAE− | MMRE |
|-------------|-------|-------|------|------|------|
| CM          | 0.656 | 0.601 | 0.592 | 0.407 | 0.074 |
| MinMax      | 0.631 | 0.574 | 0.483 | 0.432 | 0.066 |
| CRM         | 0.616 | 0.591 | 0.475 | 0.451 | 0.069 |
| CCRM        | 0.626 | 0.581 | 0.482 | 0.443 | 0.069 |
| CMM         | 0.656 | 0.601 | 0.592 | 0.407 | 0.074 |
| LM          | 0.263 | 0.192 | 0.249 | 0.181 | 0.032 |
| LM−         | 0.263 | 0.192 | 0.249 | 0.181 | 0.032 |

| Method      | RMSE* | RMSE− | MAE* | MAE− | MMRE |
|-------------|-------|-------|------|------|------|
| CM          | 0.370 | 0.362 | 0.341 | 0.341 | 0.052 |
| MinMax      | 0.141 | 0.115 | 0.120 | 0.107 | 0.017 |
| CRM         | 0.143 | 0.120 | 0.125 | 0.108 | 0.014 |
| CCRM        | 0.143 | 0.121 | 0.128 | 0.111 | 0.018 |
| CMM         | 0.204 | 0.177 | 0.159 | 0.149 | 0.024 |
| LM          | 0.139 | 0.115 | 0.128 | 0.107 | 0.018 |
| LM−         | 0.143 | 0.120 | 0.128 | 0.109 | 0.018 |

**Bold = Best**

C CRM, and MinMax methods generally improves. Nevertheless, they consistently fail to reach the performance levels of the PM and LM methods. For the mixed sets with puffy and skinny intervals, where these are disjoint or exhibit some overlap, both PM and both variants of the LM method result in similarly good (best) fit. Others, particularly the CRM and CCRM approaches, give nearly identical results for some metrics, for all overlapping, respectively, nested, mixed cases. In all cases, the estimates by the CIM and CM methods deviate substantially from the actual values and hence they come out as poor performers.

As well as the empirical analyzes reported above, we carried out additional experiments going beyond the scope of this article—exploring the impact of differences in mean interval size/range, differences in the standard deviation of the mean size/range, and the variation in interval centers (i.e., dispersion), on model fitness or estimation errors. Results suggest that estimation errors tend to grow for all regression methods with respect to both higher standard deviation of ranges and more distant placement of intervals (centers). In this particular case, the CM and CIM methods result in the worst fit.

In addition, the larger degrees of freedom—compared to the synthetic data—of these real world datasets, permit us to empirically determine a measure of the variability associated with each performance value, utilizing a nonparametric resampling approach. We, therefore, report bootstrapped (percentile) 95% confidence intervals, associated with each error measure, for each method and all three real-world cases. Note here that although many of the confidence intervals overlap, this does not necessarily imply nonsignificance of differences between the individual regression approaches’ error measures in question. As these were calculated on the same resampled datasets, the measures on each instance were nonindependent, and so it is possible for variability evident in the CIs to represent common intersimulation variance between measures, while, for example, one was a reliably better performer than another.

**B. Real-World Datasets**

In this section, three real-world IV datasets from the application contexts described in Section III-C, are used to explore and evaluate the behavior of different regression methods with respect to the mixed set of features of the data (in the sense of the previous section) present in these sets. The fitness of each method is compared using the same evaluation metrics (i.e., RMSE, MAE, MMRE) as before.

**Systolic and Diastolic BP Dataset:** For this dataset, we regress the IV diastolic BP (regressand) on those of the systolic BP (regressor) to estimate the bounds of diastolic BP. Fig. 8 presents this dataset and the associated performance of the regression methods.

Fig. 8(b) shows that the PM, LMw, and LM methods produce comparable performance in estimating the regressand bounds. With respect to some of the evaluation metrics (RMSE+, and MAE+), the PM produces the best estimation, whereas the LMw and LM methods perform better for MAE−; all these three methods perform equally best for MMRE. Fit for the CRM, CCRM, and MinMax methods for these data is also better according to MMRE. In this particular case, the CM and CIM methods result in the worst fit.
Along with evaluating model fitness, we also explore the relationship between the uncertainty (and position) of the systolic BP and that of the diastolic BP in Fig. 9. Note that for space saving, we only present here the IRGs using the PM method. Figs. 9(a) and (b) show that the center of systolic BP is positively correlated with the center of the diastolic BP and the range of systolic BP exhibits the same trend, i.e., increasing rapidly with the increasing range of the diastolic BP. These figures also highlight that the center of systolic BP decreases, at a lesser rate, as the range in diastolic BP increases, and vice versa, i.e., the range of systolic BP decreases as the center of diastolic BP increases.

2) Food Snacks Purchase Intention Dataset: With this dataset, we first remove outliers and then separately regress the overall purchase intention (regressand) on each of the six attributes of snack-bars (i.e., visual appeal, value for money, healthiness, taste, branding, and ethics). Due to page limitations, we present regression results only with respect to two attributes—ethics and taste (see Figs. 11 and 12). The estimation results in Figs. 11(b) and 12(b) show that the PM, LM_a, LM_e, and MinMax methods produce the best fit with respect to one or more indices. However, bootstrap confidence intervals indicate substantially greater proximity and variability in these estimates than for the previous BP dataset. This is especially true for the PM method for the dataset representing taste and overall purchase intention—demonstrating how in cases where transformation is required to maintain mathematical coherence, this can have a large detrimental impact on model fit, as discussed earlier.

Beyond the evaluation of model fit, we also use the IRGs to explore how the selected attributes of snack-bars influence overall purchase intention, as discussed in [32]. We construct IRGs to show the impact of two attributes—ethics and taste on those of the overall purchase intention. Again, to save space, we only show the IRGs corresponding to the PM method. Fig. 10 shows how the position and uncertainty of ethics affect those of overall purchase intention. The center of overall purchase intention, \( \hat{Y}_c \), increases with the center of ethics, \( \hat{X}_e \), as well as its range, \( 2\hat{X}_e \). Similarly, uncertainty around overall purchase intention, \( 2\hat{Y}_r \), is positively influenced to some extent by both position, \( \hat{X}_e \) and associated uncertainty of the ethics, \( 2\hat{X}_e \), though much more strongly by the latter. The IRGs (shown in Fig. 13) relating to taste also suggest implications for both position and width on each aspect of overall purchase intention. However, comparing the IRGs for taste and ethics, we observe that the position and uncertainty of two attributes tend to have distinctive impacts on overall purchase intention. For example, increasing uncertainty in ethics (together with increasing center values) increases position of overall purchase intention, whereas this tends to decline for rising uncertainty in taste (while also increasing for higher center values).

These results, which are uniquely captured by the IV data, are made visible via the IRGs. We emphasise that the regression analysis using only two variables as shown here, while a useful illustration, is limited and a broader analysis would be expected to extract real-world insights. We point the interested reader to [32], where a more detailed analysis—but based on a discrete representation of the data, is conducted.
We explore how well the regression methods estimate the IV overall difficulty ratings for selected evade components, based upon the three IV attribute ratings, and by comparison with experts’ actual overall ratings. For conciseness, we present two evade hops—number 4 and number 22, shown in Figs. 14 and 15, respectively. Fig. 14(b) shows that the PM method appears to be the best performer for Evade-4 with respect to four evaluation metrics (RMSE, RMSE+, MAE, and MMRE) whereas the \( \text{LM}_w \) method produces the joint best fit on MAE-. The \( \text{LM}_c \) method also performs very close to the \( \text{LM}_w \) method. Again, the \( \text{CM} \) method provides poor fit to the data. As for the Evade-4 dataset, Fig. 15 displays similar results for Evade-22. Particularly, Fig. 15 reveals the PM method as the best approach in terms of most metrics. Furthermore, the \( \text{LM}_w, \text{LM}_c, \text{CRM}, \) and CCRM methods come out as suitable measures as to either RMSE+ or MAE+. We provide all estimated coefficients for both synthetic and real-world datasets in the Appendix (see Tables VII and

3) Cyber-Security Dataset: Section III-C provided an overview of this dataset. Here, we only consider the evade hops as with three attributes these provide a concise basis for presenting the results within the constraints of this article. For each evade hop, 38 experts gave ratings on following three attributes: complexity, availability of information, and maturity. They were also asked to provide an overall rating on how difficult it would be overall for an attacker to evade a given hop. These ratings were given on a scale from 0 to 100. Table V summarizes the attributes for the evade hops.

We explore how well the regression methods estimate the IV overall difficulty ratings for selected evade components, based upon the three IV attribute ratings, and by comparison with experts’ actual overall ratings. For conciseness, we present two evade hops—number 4 and number 22, shown in Figs. 14 and 15, respectively. Fig. 14(b) shows that the PM method appears to be the best performer for Evade-4 with respect to four evaluation metrics (RMSE, RMSE+, MAE, and MMRE) whereas the \( \text{LM}_w \) method produces the joint best fit on MAE-. The \( \text{LM}_c \) method also performs very close to the \( \text{LM}_w \) method. Again, the \( \text{CM} \) method provides poor fit to the data. As for the Evade-4 dataset, Fig. 15 displays similar results for Evade-22. Particularly, Fig. 15 reveals the PM method as the best approach in terms of most metrics. Furthermore, the \( \text{LM}_w, \text{LM}_c, \text{CRM}, \) and CCRM methods come out as suitable measures as to either RMSE+ or MAE+. We provide all estimated coefficients for both synthetic and real-world datasets in the Appendix (see Tables VII and

| Attributes for Evade Hops and Associated Questions |
|-----------------------------------------------|
| **Complexity** | How complex is the job of providing this kind of defense? |
| **Availability of information** | How likely is that there will be publicly available information that could help with evading defense? |
| **Maturity** | How mature is this type of technology? |
| **Overall rating** | Overall, how difficult would it be for an attacker to do this? |

Fig. 15 reveals the PM method as the best approach in terms of most metrics. Furthermore, the \( \text{LM}_w, \text{LM}_c, \text{CRM}, \) and CCRM methods come out as suitable measures as to either RMSE+ or MAE+. We provide all estimated coefficients for both synthetic and real-world datasets in the Appendix (see Tables VII and

We explore how well the regression methods estimate the IV overall difficulty ratings for selected evade components, based upon the three IV attribute ratings, and by comparison with experts’ actual overall ratings. For conciseness, we present two evade hops—number 4 and number 22, shown in Figs. 14 and 15, respectively. Fig. 14(b) shows that the PM method appears to be the best performer for Evade-4 with respect to four evaluation metrics (RMSE, RMSE+, MAE, and MMRE) whereas the \( \text{LM}_w \) method produces the joint best fit on MAE-. The \( \text{LM}_c \) method also performs very close to the \( \text{LM}_w \) method. Again, the \( \text{CM} \) method provides poor fit to the data. As for the Evade-4 dataset, Fig. 15 displays similar results for Evade-22. Particularly, Fig. 15 reveals the PM method as the best approach in terms of most metrics. Furthermore, the \( \text{LM}_w, \text{LM}_c, \text{CRM}, \) and CCRM methods come out as suitable measures as to either RMSE+ or MAE+. We provide all estimated coefficients for both synthetic and real-world datasets in the Appendix (see Tables VII and
Fig. 14. Graphical presentation of Evade-4 dataset [34] and performance of different regression models as to evaluation metrics using this dataset. (a) Evade-4. (b) Model performance.

VIII. Note here again that bootstrap confidence intervals indicate substantially higher variability in model fit than for the first dataset. However, in contrast to the second dataset—for which the PM method often showed substantially higher upper confidence bounds than the other methods—the PM method’s upper confidence bounds are now often the lowest, or nearly lowest. This may indicate relatively few cases in which transformation was required to maintain mathematical coherence for the PM method with this dataset, avoiding the associated model fitness penalties.

Across the experiments, confidence bounds for the PM and the LM methods are overall fairly similar. For the BP data, Fig. 8, errors in model fit are fairly small in size due to densely placed intervals. Similarly, for the cyber-security datasets, Figs. 14 and 15, the PM and the LM methods show similar variability in confidence bounds, but the errors are larger as the intervals are relatively scattered, invariably impacting the feasibility of an LM fit. Similarly, for the food snacks data with respect to ethics, Fig. 11, the methods show similar variability in CI bounds, but bigger errors in terms of size—similar to the cyber-security dataset.

The only exception where the PM method shows wider variability in CI bounds than other approaches is for the taste versus overall purchase intention dataset, Fig. 12. This higher variability in CI bounds may be due to greater incidence of very wide and scattered intervals within the dataset, resulting in increased application of the box-cox transformation and poorer fit across a larger subset of samples/simulations.

Fig. 20 in the Appendix shows the total number of times across the 10000 simulations the regression approaches (CCRM, PM, LMc, and LMo) apply the positivity restrictions/regressand bound transformations to maintain coherence for the real-world datasets. Fig. 20 reveals that for the taste versus overall purchase intention dataset, for a significant proportion (22.85% of total 10000 simulations), positivity restrictions/ transformation of bounds are applied by the PM and LMo methods, compared to other sets. Exactly why the PM method is more strongly impacted than the LMo method for this case requires further research, which we will address in a future publication.

Fig. 15. Graphical presentation of Evade-22 dataset [34] and performance of different regression models as to evaluation metrics for this dataset. (a) Evade-22. (b) Model performance.

V. DISCUSSION ON SUITABILITY OF EXISTING REGRESSION APPROACHES

This section discusses the suitability of different regression techniques with respect to interval datasets with different properties (e.g., skinny, puffy, or mixed), and based upon the results of the analyses conducted above. The experiments consistently show that the PM and LMo (and often LMc) generally produce the best model fit—with reliably good performance for skinny/puffy or mixed sets of intervals, when these are disjoint or overlapping. With increasing overlap among intervals, estimation errors also decrease for the CRM, CCRM, and MinMax methods, but not generally so far as to equal the performance of either PM or LM methods, compared to other sets. Exactly why the PM method is more strongly impacted than the LMo method for this case requires further research, which we will address in a future publication.
Results reveal that the LM_e method produces better fitness than its permanently constrained sibling (LM_c) in most cases, as the weaker constraining approach lets model parameters vary freely unless mathematical coherence is being compromised. Both approaches generate the same outcome when the coherence of the bounds does break down and LM_e effectively reverts to LM_c. Similarly, the PM approach may produce much poorer fit in cases where it requires the box-cox transformation \([21, 25]\) of the regressand to restore coherence of the bounds. We provide such an example in the Appendix (see Fig. 17), and we also hypothesise that this is the cause of the high upper 95% bootstrap confidence bounds in relation to this model’s performance on our second real-world dataset. Generally, both CM and CIM methods yield poor model-fit.

Having said that, the PM and LM models (particularly its weakly-constrained form, LM_e) were overall the most widely suitable regression approaches. Nevertheless, as we discuss throughout, the better performance can be found for datasets following specific properties. To help researchers choose the most suitable method for their data, Fig. 16 provides a flowchart to summarize the findings presented. The figure provides a recommendation on suitability of regression approaches as to different features of the intervals in given datasets.

In addition to comparing model fitness among regression approaches, we have also visualized the relationship between IV variables with respect to their key features (i.e., center and range) by introducing the IRG. The IRG provides a powerful visual tool, akin to the “regression line” of traditional numeric regression, offering rapid, intuitive insight into the data. While the IRG can be generated for all regression models, irrespective of the regression method, we generally favor the PM method as consistently producing strong model fit.

Lastly, as for computational efficiency, the LM_e generally requires more execution time than LM_c, as it generically restricts parameters. Similarly, the PM approach involves extra execution time when it applies the box-cox transformation to restore consistency, in other words, its execution time is dependent on the dataset. Overall, in this article, we do not explore execution time in more detail, instead focusing on the performance as this is the more common driver of data analysis choices in the domains considered.

VI. Conclusion

Inferential statistics and in particular (linear) regression represent a key building block of AI. For numeric variables, linear regression does not only afford a model for estimation, it provides powerful insight into the relationships between variables. Indeed, the regression-line or graph is perhaps the most widely understood visualization in data analysis.

While well understood for the numeric case, linear regression for “fully” IV datasets, i.e., datasets where both regressor and regressand are IV, remain challenging and an active research topic. At the same time, the potential for insights from such datasets is extremely promising, not only for regression, but for AI in general.

The complexity of IV data in terms of how intervals are distributed, how they vary in size, how they overlap—or not, makes it notoriously challenging to establish universal superiority of a single technique. Acknowledging this, in this article, we have carried out an in-depth review and analysis of linear regression models, exploring their behavior with respect to systematically designed synthetic and real-world, fully IV datasets, with a view to support researchers and practitioners in the choice of techniques suitable to their problem and data.

Building on existing literature, we have articulated key features of IV data, such as whether intervals are puffy, skinny, scattered, densely placed, or overlapping. For all methods, we have investigated how they do, or do not, maintain mathematical coherence, and the impact of ensuring coherence on model-fitness. As part of this, we also extended current methods, proposing two variants of the LM method [20] to enhance its general suitability for real-world applications by ensuring mathematical coherence of bounds.

Going beyond the review of the state of the art and refinement of regression approaches, this article discusses the potential of IV data in regression and AI more broadly, in particular, highlighting the value of mapping not only variable position (as in the numeric case), but also uncertainty/range—for example, to infer how uncertainty about a regressor, such as the ethical origin of a snack-food product’s ingredients, impacts uncertainty—and/or position of a regressand, such as a consumer’s purchase-intention.

Acknowledging the powerful aspect of visualization in terms of the popularity of traditional linear regression, we have introduced the IRG, facilitating the interpretation of IV regression models, and by extension, IV data. IRGs succinctly visualize and articulate the relationship between IV regressor and regressand as to both position (center) and uncertainty (range). As IRGs use all three visualizable dimensions, they are limited to the visualization of the relationship between two variables at a time. Naturally, this can be applied to multivariate cases by presenting the relationship between two variables while keeping other variables constant. An alternative to this approach could be to leverage techniques, such as principal component analysis (PCA) for intervals [41], and visualizing the relationships between resulting components. However, here, the intuitive interpretation of said intervals will be complex, potentially limiting one of the primary assets of leveraging
intervals in real-world settings as discussed in this article. We will explore these aspects in future work.

Finally, as alluded to above, we have leveraged the empirical results to provide recommendations on which regression models are best suited to which type of IV datasets, based on the data’s features, such as whether intervals strongly overlap, are narrow or wide, etc. This is supported by complementary insights, such as generally better model prediction observed for IV sets with narrow range or condensed interval placement. Furthermore, in this article, we have provided open-source software which—for the first time—offers access to the suite of IV regression approaches for facilitating adoption and replication.

In summary, analyzes have shown that the PM \([21]\) and the LM \([20]\) methods (with its weaker set of restrictions, i.e., \(L_{M(w)}\)) are the most suitable regression approaches in many cases including where intervals are skinny, puffy, or a mix thereof. The constrained variant of the LM model \([20]\), \(L_{Mc}\) also performs equally well in many cases. However, when intervals are densely placed or some overlapping/mixed, other existing methods, such as CRM \([17]\), CCRM \([18]\), and MinMax \([16]\) may turn into suitable estimation approaches, offering both relatively simple implementation and efficient computation. In such instances, it is desirable to verify if the CRM \([17]\) and MinMax \([16]\) methods maintain consistency of the bounds. Certainly, maintaining this coherence automatically, such as by the PM, LM, and CCRM methods, tends to require more execution time. The latter may be relevant as the number of variables increases.

With increasing popularity of IV data, being able to analyze and derive the important insights available from these data using AI becomes an increasingly important topic. The substantial body of work advancing IV regression over the last ten years, and advances, such as those put forward in this article, provide the foundation for leveraging intervals more generally in AI. Substantial work remains, from further refining our capacity for communicating resulting insights to advancing our modeling techniques, such as to nonlinear approaches.

Beyond advancing techniques, in the future, we aim to explore additional real-world cases to demonstrate and advance the understanding of the value of using intervals, rather than “crisp”, numeric data. Specifically, we will revisit the nature of the information captured by intervals, whether disjunctive (such as in confidence intervals), or conjunctive (true ranges, such as the real numbers between 2 and 6) and explore how different analysis techniques can, and whether they should, deal with such data differently.

Appendix

### Table VI

| Synthetic IV Datasets Used for Sensitivity Analysis (Set-1 to Set-11) |
|-----------------------------|-----------------------------|
| (a) Set-1 | (b) Set-2 | (c) Set-3 | (d) Set-4 |
| \(x^1\) | \(x^2\) | \(x^1\) | \(x^2\) | \(x^1\) | \(x^2\) | \(x^1\) | \(x^2\) |
| \(\mu_1\) | \(\mu_2\) | \(\mu_1\) | \(\mu_2\) | \(\mu_1\) | \(\mu_2\) | \(\mu_1\) | \(\mu_2\) |
| 1.5 | 3.5 | 1.5 | 3.5 | 1.5 | 3.5 | 1.5 | 3.5 |
| 2.0 | 4.0 | 2.0 | 4.0 | 2.0 | 4.0 | 2.0 | 4.0 |
| 2.5 | 5.0 | 2.5 | 5.0 | 2.5 | 5.0 | 2.5 | 5.0 |
| 3.0 | 6.0 | 3.0 | 6.0 | 3.0 | 6.0 | 3.0 | 6.0 |
| 3.5 | 7.0 | 3.5 | 7.0 | 3.5 | 7.0 | 3.5 | 7.0 |
| 4.0 | 8.0 | 4.0 | 8.0 | 4.0 | 8.0 | 4.0 | 8.0 |

Fig. 17. Comparatively low fit of the PM approach. (a) Set-12. (b) 2-D visualization. (c) Model performance.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
### TABLE VII

| Methods   | Set-1                  | Synthetic datasets | Set-4                  | Set-5                  | Set-6                  |
|-----------|------------------------|--------------------|------------------------|------------------------|------------------------|
| CM        | $\beta_0 = 0.397, \beta_1 = 0.43$ | $\beta_0 = 0.316, \beta_1 = 0.46$ | $\beta_0 = 4.608, \beta_1 = 0.26$ | $\beta_0 = 5.637, \beta_1 = 0.172$ | $\beta_0 = 10.54, \beta_1 = 0.691$ |
| MinMax    | $\beta_0 = 0.897, \beta_1 = 0.003$ | $\beta_0 = 0.597, \beta_1 = 0.003$ | $\beta_0 = 4.56, \beta_1 = 0.24$ | $\beta_0 = 5.505, \beta_1 = 0.164$ | $\beta_0 = 11.35, \beta_1 = 0.90$ |
| CRM       | $\beta_0 = 0.397, \beta_1 = 0.43$ | $\beta_0 = 0.316, \beta_1 = 0.46$ | $\beta_0 = 4.608, \beta_1 = 0.26$ | $\beta_0 = 5.637, \beta_1 = 0.172$ | $\beta_0 = 10.54, \beta_1 = 0.691$ |
| CCRM      | $\beta_0 = 0.397, \beta_1 = 0.43$ | $\beta_0 = 0.316, \beta_1 = 0.46$ | $\beta_0 = 4.608, \beta_1 = 0.26$ | $\beta_0 = 5.637, \beta_1 = 0.172$ | $\beta_0 = 10.54, \beta_1 = 0.691$ |
| CIM       | $\beta_0 = 0.397, \beta_1 = 0.43$ | $\beta_0 = 0.316, \beta_1 = 0.46$ | $\beta_0 = 4.608, \beta_1 = 0.26$ | $\beta_0 = 5.637, \beta_1 = 0.172$ | $\beta_0 = 10.54, \beta_1 = 0.691$ |

| PM        | $\beta_0 = 0.397, \beta_1 = 0.003$ | $\beta_0 = 0.597, \beta_1 = 0.003$ | $\beta_0 = 4.56, \beta_1 = 0.24$ | $\beta_0 = 5.505, \beta_1 = 0.164$ | $\beta_0 = 11.35, \beta_1 = 0.90$ |
| LIM       | $\beta_0 = 0.397, \beta_1 = 0.003$ | $\beta_0 = 0.597, \beta_1 = 0.003$ | $\beta_0 = 4.56, \beta_1 = 0.24$ | $\beta_0 = 5.505, \beta_1 = 0.164$ | $\beta_0 = 11.35, \beta_1 = 0.90$ |
| LIMs      | $\beta_0 = 0.397, \beta_1 = 0.003$ | $\beta_0 = 0.597, \beta_1 = 0.003$ | $\beta_0 = 4.56, \beta_1 = 0.24$ | $\beta_0 = 5.505, \beta_1 = 0.164$ | $\beta_0 = 11.35, \beta_1 = 0.90$ |

### TABLE VIII

| Methods   | Synthetic datasets | Real-world datasets |
|-----------|--------------------|---------------------|
| CM        | $\beta_0 = 0.397, \beta_1 = 0.43$ | $\beta_0 = 0.316, \beta_1 = 0.46$ |
| MinMax    | $\beta_0 = 0.397, \beta_1 = 0.43$ | $\beta_0 = 0.316, \beta_1 = 0.46$ |
| CRM       | $\beta_0 = 0.397, \beta_1 = 0.43$ | $\beta_0 = 0.316, \beta_1 = 0.46$ |
| CCRM      | $\beta_0 = 0.397, \beta_1 = 0.43$ | $\beta_0 = 0.316, \beta_1 = 0.46$ |
| CIM       | $\beta_0 = 0.397, \beta_1 = 0.43$ | $\beta_0 = 0.316, \beta_1 = 0.46$ |

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Fig. 18. IRGs showing the relationship between regressor and regressand in terms of their center and range using different linear regression models for Set-1 [see Fig. 1(a)]. (a) CM method. (b) MinMax method. (c) CRM method. (d) CCRM method. (e) CIM method. (f) PM method. (g) LM\textsubscript{c} method. (h) LM\textsubscript{w} method.

Fig. 19. IRGs showing the relationship between regressor and regressand in terms of their center and range using different linear regression models for Set-2 [see Fig. 2(a)]. (a) CM method. (b) MinMax method. (c) CRM method. (d) CCRM method. (e) CIM method. (f) PM method. (g) LM\textsubscript{c} Method. (h) LM\textsubscript{w} Method.

Fig. 20. Number of times in 10000 simulations the regression approaches apply positivity restriction/transformation regressand bounds for real-world cases. BP=blood pressure dataset, \( \text{E&OPI}=\) ethics versus overall purchase intention dataset, \( \text{T&OPI}=\) taste versus overall purchase intention dataset.

REFERENCES

[1] L.-C. Lin, H.-L. Chien, and S. Lee, "Symbolic interval-valued data analysis for time series based on auto-interval-regressive models," *Statist. Methods Appl.*, vol. 30, pp. 1–21, 2020.
[2] Y. Sun and D. Ralescu, “A normal hierarchical model and minimum contrast estimation for random intervals,” *Ann. Inst. Statist. Math.*, vol. 67, no. 2, pp. 313–333, 2015.
[3] F. Liu and J. M. Mendel, “Encoding words into interval type-2 fuzzy sets using an interval approach,” *IEEE Trans. Fuzzy Syst.*, vol. 16, no. 6, pp. 1503–1521, Dec. 2008.
[4] C. Wagner, S. Miller, J. M. Garibaldi, D. T. Anderson, and T. C. Havens, “From interval-valued data to general type-2 fuzzy sets,” *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 2, pp. 248–269, Apr. 2015.
[5] Z. Ellerby, C. Wagner, and S. B. Broomell, “Capturing richer information: On establishing the validity of an interval-valued survey response mode,” *Behav. Res. Methods*, vol. 54, pp. 1–23, 2021.
[6] S. Kabir, C. Wagner, T. C. Havens, and D. T. Anderson, “A similarity measure based on bidirectional subsethood for intervals,” *IEEE Trans. Fuzzy Syst.*, vol. 28, no. 11, pp. 2890–2904, Nov. 2020.
[7] P. Brito, “Modelling and analysing interval data,” in *Advances in Data Analysis*. Berlin, Germany: Springer, 2007, pp. 197–208.
[8] C.-C. Yeh, “Tree-based regression for interval-valued data,” Master of Statistics, Utah State University, Athens, OH, USA, 2017.
[9] G. González-Rivera and W. Lin, “Constrained regression for interval-valued data,” J. Bus. Econ. Statist., vol. 31, no. 4, pp. 473–490, 2013.

[10] G. González-Rodríguez, Á. Blanco, N. Corral, and A. Colubi, “Least squares estimation of linear regression models for convex compact random sets,” Adv. Data Anal. Classification, vol. 1, no. 1, pp. 67–81, 2007.

[11] A. Blanco-Fernández, N. Corral, and G. González-Rodríguez, “Estimation of a flexible simple linear model for interval data based on set arithmetic,” Comput. Statist. Data Anal., vol. 55, no. 9, pp. 2568–2578, 2011.

[12] R. Boukezzoula, S. Galichet, and A. Bisserier, “A midpoint–radius approach to regression with interval data,” Int. J. Approx. Reasoning, vol. 52, no. 9, pp. 1257–1271, 2011.

[13] A. Blanco-Fernández, A. Colubi, and M. García-BáRzana, “A set arithmetic-based linear regression model for modelling interval-valued responses through real-valued variables,” Inf. Sci., vol. 247, pp. 109–122, 2013.

[14] M. García-BáRzana, A. B. Ramos-Guajardo, A. Colubi, and E. J. Kontoghiorghes, “Multiple linear regression models for random intervals: A set arithmetic approach,” Comput. Statist., vol. 35, no. 2, pp. 755–773, 2020.

[15] L. Billard and E. Diday, “Regression Analysis for Interval-Valued Data,” in Data Analysis, Classification, and Related Methods. Berlin, Germany: Springer, 2000, pp. 369–374.

[16] L. Billard and E. Diday, “Symbolic Regression Analysis,” in Classification, Clustering, and Data Analysis. Berlin, Germany: Springer, 2002, pp. 281–288.

[17] E. D. A. L. Neto and F. D. A. de Carvalho, “Centre and range method for fitting a linear regression model to symbolic interval data,” Comput. Statist. Data Anal., vol. 52, no. 3, pp. 1500–1515, 2008.

[18] E. A. L. Neto and F. D. A. T. Carvalho, “Constrained linear regression models for symbolic interval-valued variables,” Comput. Statist. Data Anal., vol. 54, no. 2, pp. 333–347, 2010.

[19] H. Wang, R. Guan, and J. Wu, “Linear regression of interval-valued data based on complete information in hypercubes,” J. Syst. Sci. Syst. Eng., vol. 21, no. 4, pp. 422–442, 2012.

[20] Y. Sun and D. Ralescu, “A linear model for interval-valued data,” 2015, arXiv:1506.03541.

[21] L. C. Souza, R. M. Souza, G. J. Amaral, and T. M. Silva Filho, “A parametrized approach for linear regression of interval data,” Knowl.-Based Syst., vol. 131, pp. 149–159, 2017.

[22] P. Hao and J. Guo, “Constrained center and range joint model for interval-valued symbolic data regression,” Comput. Statist. Data Anal., vol. 116, pp. 106–138, 2017.

[23] C. L. Lawson and R. J. Hanson, Solving Least Squares Problems. New York, NY, USA: Prentice-Hall, 1974.

[24] R. E. Moore, Interval Analysis, vol. 4. Englewood Cliffs, NJ, USA: Prentice-Hall, 1966.

[25] G. E. Box and D. R. Cox, “An analysis of transformations,” J. Roy. Statist. Soc.: Ser. B. (Methodological), vol. 26, no. 2, pp. 211–243, 1964.

[26] S. Ferson, V. Kreinovich, J. Hajagos, W. Oberkampf, and L. Ginzburg, “Experimental uncertainty estimation and statistics for data having interval uncertainty,” Sandia National Laboratories, Albuquerque, NM, USA, Tech. Rep. SAND2007-0939, 2007.

[27] I. Couso and D. Dubois, “Statistical reasoning with set-valued information: Ontic vs. epistemic views,” Int. J. Approx. Reasoning, vol. 55, no. 7, pp. 1502–1518, 2014.

[28] R. R. Picard and R. D. Cook, “Cross-validation of regression models,” J. Amer. Statist. Assoc., vol. 79, no. 387, pp. 575–583, 1984.

[29] G. Beliakov and S. James, “A penalty-based aggregation operator for non-convex intervals,” Knowl.-Based Syst., vol. 70, pp. 335–344, 2014.

[30] N. R. Draper and H. Smith, Applied Regression Analysis, vol. 326. Hoboken, NJ, USA: Wiley, 1998.

[31] Z. Ellerby, O. Miles, J. McCulloch, and C. Wagner, “Insights from interval-valued ratings of consumer products—A DECSYS appraisal,” in Proc. IEEE Int. Conf. Fuzzy Syst., 2021, pp. 1–7.

[32] Z. Ellerby, J. McCulloch, J. Young, and C. Wagner, “DECSYS–discrete and ellipse-based response capture system,” in Proc. IEEE Int. Conf. Fuzzy Syst., 2019, pp. 1–6.

[33] Z. Ellerby, J. McCulloch, M. Wilson, and C. Wagner, “Exploring how component factors and their uncertainty affect judgements of risk in cybersecurity,” in Proc. Int. Conf. Crit. Inf. Infrastructures Secur., 2019, pp. 31–42.

[34] D. Içen and H. Demirhan, “Error measures for fuzzy linear regression: Monte Carlo simulation approach,” Appl. Soft Comput., vol. 46, pp. 104–114, 2016.

[35] R. A. Fagundes, R. M. de Souza, and Y. M. Soares, “Quantile regression of interval-valued data,” in Proc. 23rd Int. Conf. Pattern Recognit., 2016, pp. 2586–2591.

[36] S. Kabir and C. Wagner, “Interval-valued regression-sensitivity to data set features,” in Proc. IEEE Int. Conf. Fuzzy Syst., 2021, pp. 1–7.

[37] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. Boca Raton, FL, USA: CRC Press, 1994.

[38] T. C. Hesterberg, “What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum,” Amer. Statistician, vol. 69, no. 4, pp. 371–386, 2015.

[39] F. Gioia and C. N. Lauro, “Principal component analysis on interval data,” Comput. Statist., vol. 21, no. 2, pp. 343–363, 2006.