Towards 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 modelling discord or variance between sources (i.e., inter-source) of information. Increasingly however, research is focusing on uncertainty arising at the level of individual measurements (i.e., within- or intra-source), 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 paper 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 behaviour, advantages, and pitfalls when applied to synthetic and real-world data sets 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, or IRG, which effectively communicates the impact of both position and range of variables within the regression models—offering a leap in their interpretability. Finally, the paper 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—enables 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—Intervals, regression, uncertainty, intra-source, inter-source

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 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 data sets 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 modelled according to set arithmetic [10]–[14]. On the other hand, in the SDA setting, the classical regression model is expanded to IV data where intervals are treated as bi-variate vectors [15]–[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]–[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]–[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], Wáng 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
Fig. 1: (a) Visualization of IV data Set-1 and (b) IRG with respect to range, \( Y^{\text{w}} \approx 2Y^r \) using the PM method [21].

Fig. 2: (a) Visualization of IV data Set-2 and (b) IRG with respect to range, \( Y^{\text{w}} \approx 2Y^r \) using the PM method [21].

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 the paper.

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 cybersecurity, uncertainty around the vulnerability of a system may lead to certainty 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 data sets. Here, in Set-1 (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 vs their horsepower. In contrast, for Set-2 (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 data set—based on its properties—is essential. Ferson et al. in [26], discuss the complexity of analysing real-world IV data in detail, elucidating the value of focusing on key properties for data sets 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 conjunctive 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 paper 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.

\(^1\)Throughout the remainder of this paper, for simplicity, we refer to the range of the interval as representing uncertainty as a general term capturing vagueness, lack of information, etc.
A closed interval $\hat{\pi}$ is characterized by its two endpoints, $a^-$ and $a^+$ with $a^- \leq a^+ \leq a^+$. Let $b_{ij}$ denote the $i$th entry in $\hat{\pi}_j$. Generally, $\hat{\pi}$ is denoted as $\hat{\pi} = [a^-, a^+]$, and its range is given by $a^{\text{w}} = [a^- - a^+]$. Alternatively, the same interval can be represented as $\hat{\pi} = [a^- - a^+, a^+ - a^-]$. In this section, we review linear regression methods for full IV data sets (i.e. both dependent and independent variables are intervals) where intervals are treated as bi-variate vectors in the regression process. We note $Y = (\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_n)$ as an IV regressand with $n$ intervals, where $\hat{\pi}_i = [y_i^- , y_i^+]$, $1 \leq i \leq n$. Further, $\{X_1, X_2, \ldots, X_p\}$ presents $p$ (i.e. 1) IV regressor(s) where each $X_j$ has also $n$ intervals, $X_j = \{x_{ji}^- , x_{ji}^+\}$, $1 \leq i \leq n$, $1 \leq j \leq p$. Thus, there are $p$ pairs of $(Y, X)$ in the data set. In addition, the estimated value of $Y$ is given by $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_n)$. Since about the year 2000, a variety of linear regression approaches for full IV data sets have been put forward. We briefly review them below in chronological order.

A. The Center Method (2000)

Billard and Diday [15] proposed a linear regression model, known as the Center Method (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. Equation (1) defines the CM regression model,

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

where $y_i^-$ and $y_i^+$ are the lower and upper bounds of $\hat{\pi}_i$. Similarly, $x_{ji}^-$ and $x_{ji}^+$ are the the lower and upper bounds of $\hat{\pi}_{ji}$. $\beta_j^c \in \mathbb{R}$ are the regression coefficients for the center estimation, $0 \leq j \leq p$, $p$ is the total number of regressors while $\epsilon_i^-$ and $\epsilon_i^+$ are error terms. Both lower and upper bound models of (1) use the same coefficients, therefore, they can be captured together using (2).

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

where $y_i^+$ and $x_{ji}^+$ are the center of $\hat{\pi}_i$ and $\hat{\pi}_{ji}$ respectively and $\epsilon_i^+ = (\epsilon_i^- + \epsilon_i^+)/2$ is error term. Equation (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 $Y^c = (y_1^c \quad y_2^c \quad \ldots \quad y_n^c)^T$, $\beta^c = (\beta_0^c \quad \beta_1^c \quad \ldots \quad \beta_p^c)^T$, and $X^c = \left(\begin{array}{cccc}
x_{i1}^c & \ldots & x_{ip1}^c \\
x_{i2}^c & \ldots & x_{ip2}^c \\
\vdots & \ddots & \vdots \\
x_{in}^c & \ldots & x_{ipn}^c
\end{array}\right)$.

The coefficients, $\beta^c$ are estimated using the least squares (LS) method [30], defined in (3).

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

---

2) 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://lucidersearch.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.2

The paper 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 interval regression graph (IRG) and discusses adaptations of the LM method [20] for real-world applications, followed by IV data sets (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 data sets. 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 the paper with future works. Table I presents a list of acronyms and notation used in this paper to assist the reader.

---

2) 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 cross-validation [28]—offering scope for future research for IV data.
By using \( \hat{\beta}^c \), (4) estimates the bounds of regressand, \( Y \).

\[
\hat{y}_i^+ = \hat{\beta}_0^+ + \sum_{j=1}^{p} \hat{\beta}_j^+ x_{ji} \\
\hat{y}_i^- = \hat{\beta}_0^- + \sum_{j=1}^{p} \hat{\beta}_j^- x_{ji}
\]

(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.


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. Equation (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 \( \mathbf{Y}^- = \mathbf{X}^- \mathbf{\beta}^- + \mathbf{\epsilon}^- \), where

\[
\mathbf{Y}^- = (y_1^- \ y_2^- \ldots \ y_n^-)^T, \quad \mathbf{\beta}^- = (\beta_0^- \ \beta_1^- \ldots \ \beta_p^-)^T, \\
\mathbf{\epsilon}^- = (\epsilon_1^- \ \epsilon_2^- \ldots \ \epsilon_n^-)^T, \text{ and } \mathbf{X}^- = \begin{bmatrix} 1 & x_{11}^+ & x_{12}^+ & \ldots & x_{1n}^+ \\ 1 & x_{21}^+ & x_{22}^+ & \ldots & x_{2n}^+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{p1}^+ & x_{p2}^+ & \ldots & x_{pn}^+ \end{bmatrix}.
\]

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

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

(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 \( \mathbf{Y}^+ = \mathbf{X}^+ \mathbf{\beta}^+ + \mathbf{\epsilon}^+ \), where

\[
\mathbf{Y}^+ = (y_1^+ \ y_2^+ \ldots \ y_n^+)^T, \quad \mathbf{\beta}^+ = (\beta_0^+ \ \beta_1^+ \ldots \ \beta_p^+)^T, \\
\mathbf{\epsilon}^+ = (\epsilon_1^+ \ \epsilon_2^+ \ldots \ \epsilon_n^+)^T, \text{ and } \mathbf{X}^+ = \begin{bmatrix} 1 & x_{11}^+ & x_{12}^+ & \ldots & x_{1n}^+ \\ 1 & x_{21}^+ & x_{22}^+ & \ldots & x_{2n}^+ \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{p1}^+ & x_{p2}^+ & \ldots & x_{pn}^+ \end{bmatrix}.
\]

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

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

(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]. Further, its estimation performance can be reduced if there is no dependency between the bounds of regressand and regressors [21].

C. The Center and Range Method (2008)

Neto and Carvalho [17] extended the CM method [15] in 2008 to the Center and Range Method (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). Equation (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 \( x_{ji}^c \) are the centers \( \bar{y}_i \) and \( \bar{x}_{ij} \) 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 \( \beta_j^c \) (see (3)). It then estimates the center values for the regressand, \( Y \) using the estimated \( \hat{\beta}^c \) by (11).

\[
\hat{y}_i^c = \hat{\beta}_0^c + \sum_{j=1}^{p} \hat{\beta}_j^c x_{ji}^c
\]

(11)

Similarly, the interval ranges are modelled by the CRM method using (12),

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

(12)

where \( y_i^r \) and \( x_{ji}^r \) are the radius of \( \bar{y}_i \) and \( \bar{x}_{ij} \) respectively. \( \beta_j^r \) are the coefficients for the range estimations, \( 0 \leq j \leq p \) and \( \epsilon_i^r \) is error term. Again, following the same matrix notation, the coefficients, \( \beta_j^r \) are estimated by (13),

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

(13)

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

\[
\hat{y}_i^r = \hat{\beta}_0^r + \sum_{j=1}^{p} \hat{\beta}_j^r x_{ji}^r 
\]

(14)

Finally, the CRM method estimates the bounds of the regressand from the estimated \( \hat{y}_i^c \) and \( \hat{y}_i^r \) using (15).

\[
\hat{y}_i^- = \hat{y}_i^c - \hat{y}_i^r \quad \text{and} \quad \hat{y}_i^+ = \hat{y}_i^c + \hat{y}_i^r
\]

(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 \( y_i^r \) [21].
D. The Constrained Center and Range Method (2010)

Neto and Carvalho [18] later refined the CRM model [17] (see Section II-C) in 2010 to ensure mathematical coherence. The resulting approach is known as the Constrained Center and Range Method (CCRM), where a positivity restriction is enforced on the coefficients which are estimated with respect to the relationship of the radii of regressand and regressor variables. In other words, the overall estimation process in the CCRM method remains the same as for the CRM model (see (10) to (15)) with positivity constraints on the range coefficients $\beta_j$. The CCRM model uses an iterative algorithm proposed by Lawson and Hanson in [23] to ensure $\beta_j^* \geq 0$. For more detail, we refer the reader to [18]. Finally, the CCRM method estimates the bounds of the regressand using (15).

Neto and Carvalho recommend to apply the CRM model in all cases, only adopting the CCRM method as a suitable strategy when the CRM method fails to maintain mathematical coherence, as the use of the CCRM model can lead to biased estimation outcomes [18]. In particular, the positivity restriction within the CCRM method forces any negative range coefficient to 0 (at the same time updating the remaining range coefficients), in turn leading to potentially biased estimation outcomes and poor model fitness [18].

E. The Complete Information Method (2012)

Wang et al. [19] presented the Complete Information Method (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 hyper-cube (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 hyper-cubes. Further, 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 $X_j$,

$$Y = \beta_0 1_n + \sum_{j=1}^{p} \beta_j X_j + \epsilon_i,$$

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

$$y_i^- = \beta_0 + \sum_{j=1}^{p} \beta_j \left( \tau_j x_{ji}^- + (1 - \tau_j) x_{ji}^+ \right)$$

$$y_i^+ = \beta_0 + \sum_{j=1}^{p} \beta_j \left( (1 - \tau_j) x_{ji}^- + \tau_j x_{ji}^+ \right)$$

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,

$$M\beta = b$$

where,

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

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

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

Here, $(X_a, X_b)$ is the inner product between $X_a$ and $X_b$ ($0 \leq a, b \leq p$ and $X_0 = 1$) which is computed using (19).

$$\langle X_a, X_b \rangle = \begin{cases} \frac{1}{p} \sum_{i=1}^{p} (x_{ai} x_{bi} + x_{ai} x_{bi}^+ + x_{ai} x_{bi}^-), & \text{if } X_a = X_b \\ \frac{1}{p} \sum_{i=1}^{p} (x_{ai} x_{bi} + x_{ai} x_{bi}^+ + x_{ai} x_{bi}^-), & \text{if } X_a \neq X_b \end{cases}$$

From (18), the coefficients $\beta$ can be estimated by the inverse of matrix $M$ as defined in (20).

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

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. The Linear Model (2015)

Sun and Ralescu [20] developed the Linear Model (LM) in 2015 based on the affine operator in the cone $C = \{ (x, y) \in \mathbb{R}^2 | x \leq y \}$, 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 $\bar{X}_{ji}$ such that,

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

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

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

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

$$X = \begin{pmatrix} X^- & 0 \\ X^+ & X^w \end{pmatrix} \text{ with } X^+ = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix},$$

and $X^w = \begin{pmatrix} 1 & x_{1w} & \ldots & x_{1w} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{nw} & \ldots & x_{nw} \end{pmatrix}$, and $\beta = \begin{pmatrix} \beta^+ \\ \beta^w \end{pmatrix}$.
\[ \beta^* = \begin{pmatrix} \alpha_1 \\ \beta_1 \\ \vdots \\ \alpha_p \\ \beta_p \end{pmatrix} \quad \text{and} \quad \beta^w = \begin{pmatrix} \theta \\ \gamma_1 \\ \vdots \\ \gamma_p \end{pmatrix}. \]

Again, the LS estimate of the coefficients, \( \hat{\beta} \), are computed by (22),
\[ \hat{\beta} = (X^T X)^{-1} X^T Y \] (22)
and (23) estimates the bounds of the regressand, \( Y \).
\[ \hat{y}_i = \sum_{j=1}^{p} (\hat{\alpha}_j x_{ji} + \hat{\beta}_j x_{ji}^+) + \hat{\eta}, \]
\[ \hat{y}_i^+ = \sum_{j=1}^{p} (\hat{\alpha}_j - \hat{\gamma}_j) x_{ji} + (\hat{\beta}_j + \hat{\gamma}_j) x_{ji}^+ + \hat{\eta} + \hat{\theta}. \]

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 estimated of \( \hat{\gamma}_j \) and \( \hat{\theta} \) turns out to be negative, forcing it to be positive may lead to poor fitness of the LM model [20].

G. The Parametrized Model (2017)

Souza et al. [21] proposed the Parametrized Model (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, 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 \( q = a^{-1} (1 - \lambda) + a^+ \lambda, 0 \leq \lambda \leq 1 \). By setting \( \lambda, \pi \) is turned into a single point. Hence, when \( \lambda = 0, q = a^{-1} \) (lower bound of \( \pi \)) and when \( \lambda = 1, q = a^+ \) (upper bound of \( \pi \)). Similarly, \( q = a^c \) (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 (24).
\[ y_i = \beta_0^- + \sum_{j=1}^{p} \beta_j^- (1 - \lambda_j) x_{ji} + \beta_j^+ \lambda_j x_{ji}^+ + \epsilon_i, \]
\[ y_i^+ = \beta_0^+ + \sum_{j=1}^{p} \beta_j^+ (1 - \lambda_j) x_{ji}^- + \beta_j^+ \lambda_j x_{ji}^+ + \epsilon_i. \]

Equation (25) simplifies (24) by replacing \( \beta_j^- (1 - \lambda_j) \) by \( \alpha_j \) and \( \beta_j^- \lambda_j \) by \( \omega_j \) for lower bounds, and \( \beta_j^+ (1 - \lambda_j) \) and \( \beta_j^+ \lambda_j \) by \( \alpha_j^+ \) and \( \omega_j^+ \) respectively.
\[ y_i = \beta_0^- + \sum_{j=1}^{p} \alpha_j x_{ji}^- + \omega_j^- x_{ji}^+ + \epsilon_i, \]
\[ y_i^+ = \beta_0^+ + \sum_{j=1}^{p} \alpha_j x_{ji}^- + \omega_j^+ x_{ji}^+ + \epsilon_i. \] (25)

In matrix notation, the lower bound model can be expressed for all \( n \) observations as \( Y^- = X^* \beta^- + \epsilon^- \), where \( Y^- = (y_1^- \ y_2^- \ldots y_n^-)^T, \beta^- = (\beta_0^- \ \alpha_1^- \ \omega_1^- \ldots \ \alpha_p^- \ \omega_p^-)^T, \epsilon^- = (\epsilon_1^- \ \epsilon_2^- \ldots \ \epsilon_n^-)^T, \) and \( X^* = \begin{pmatrix} 1 \ x_{i1}^- \ x_{i2}^- \ldots \ x_{ip}^- \\ 1 \ x_{i1}^- \ x_{i2}^- \ldots \ x_{ip}^- \\ \vdots \ \vdots \ \vdots \ \vdots \\ 1 \ x_{i1}^- \ x_{i2}^- \ldots \ x_{ip}^- \end{pmatrix} \).

The LS estimate of the coefficients for the lower bound model, \( \hat{\beta}^- \), is computed by (26).
\[ \hat{\beta}^- = ((X^*)^T X^*)^{-1}(X^*)^T Y^- \] (26)

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, \( \hat{\beta}^+ \) in defined in (27).
\[ \hat{\beta}^+ = ((X^*)^T X^*)^{-1}(X^*)^T Y^+ \] (27)

Finally, using \( \hat{\beta}^- \) and \( \hat{\beta}^+ \), the lower and upper bounds of \( Y \) are estimated using (28).
\[ \hat{y}_i = \hat{\beta}_0^- + \sum_{j=1}^{p} \hat{\alpha}_j x_{ji}^- + \hat{\omega}_j^- x_{ji}^+, \]
\[ \hat{y}_i^+ = \hat{\beta}_0^+ + \sum_{j=1}^{p} \hat{\alpha}_j x_{ji}^- + \hat{\omega}_j^+ x_{ji}^+. \] (28)

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.
\[ \hat{Y}^w = X^* ((X^*)^T X^*)^{-1} (X^*)^T Y^w \] (29)

If all estimated ranges are positive (\( \hat{y}_i^w \in \hat{Y}^w \)), 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.
\[ \hat{y}_i^k = \begin{cases} \frac{(y_i^{-k} + k)^{k+1} - 1}{k+1}, & \text{if } k_1 \neq 0 \left[ \log (y_i^{-k} + k_2), \log (y_i^{-k} + k_2) \right], \text{if } k_1 = 0. \end{cases} \] (30)
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: \( y_i^{-k} + k_2 > 0 \).

III. Data and Methodology

In Section II, we discussed linear regression models for intervals considering their vector/matrix representation. Table II provides a summary of these models and their features. In this section, we initially put forward an extension to the LM method [20] which ensures mathematical coherence, thus making it more suitable for real-world data. Next, we introduce...
TABLE II: Summary of well-known linear regression models for IV data

<table>
<thead>
<tr>
<th>Methods</th>
<th>Interval reference point(s) used in the regression process</th>
<th>Also applicable to numeric data (interval range (= 0))</th>
<th>Ensure mathematical coherence on interval bounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>CM [15]</td>
<td>Center</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>MinMax [16]</td>
<td>Lower and upper bounds</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>CRM [17]</td>
<td>Center and radius (half of range)</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>CRM [18]</td>
<td>Center and radius (half of range)</td>
<td>No</td>
<td>Yes with positivity restrictions on range coefficients</td>
</tr>
<tr>
<td>CIM [19]</td>
<td>All points within intervals</td>
<td>Yes</td>
<td>Yes with Moore’s linear combination algorithm [24]</td>
</tr>
<tr>
<td>LM [20]</td>
<td>Lower and upper bounds, and range</td>
<td>No</td>
<td>Yes with positivity restrictions as proposed in this paper</td>
</tr>
<tr>
<td>PM [21]</td>
<td>Any point within intervals</td>
<td>Yes</td>
<td>Yes with the interval Box-Cox transformation [25]</td>
</tr>
</tbody>
</table>

A novel approach to intuitively visualizing IV regression—the Interval Regression Graph (IRG). IRGs are designed to facilitate the interpretation and communication of the regression model and thus the relationship between an independent and dependent variable of interest. The latter is a crucial asset of traditional (numeric) regression models, accounting for a substantial part of their utility and popularity. Going beyond the traditional visualization of numeric regression, IRGs communicate the relationships of both position and range (uncertainty) of IV regressor and regressand. Following the introduction of IRGs, we introduce both synthetic and real-world data sets to support the empirical exploration and analysis of the regression approaches in Section II. Finally, we discuss evaluation metrics employed for performance assessment during the remainder of the paper.

A. Extension of the Linear Model [20]

As pointed out in Section II, the LM method [20] considers positivity restrictions on the range coefficients to consistently estimate the regressand bounds. However, it neither guarantees these restrictions through its model setting, nor provides algorithmic solutions to maintaining them. Thus, to make the model suitable for practical real-world deployment, we put forward two alternative extensions to the LM method: (1) universally forcing positive restrictions on parameters as [20] suggested—that is proposing a constrained LM method, and (2) enforcing restrictions only when needed to avoid unnecessary estimation bias [20]—that is, proposing a weakly constrained LM method.

Algorithm 1 A constrained LM method (\(L_{MC}\))

**Input:** Let \(\tilde{Y} = \{\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_n\}\) be the regressand variable’s samples and \(X_1, X_2, ..., X_p\) are the \(p\) (\(\geq 1\)) be the regressor variables.

**Output:** Let \(\hat{Y} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_n\}\) be the estimated \(Y\).

1. Apply the LM method at (22) on \(Y\) and \(X_j\) (\(1 \leq j \leq p\)) to estimate \(\hat{\beta}^*\) and \(\hat{\beta}^w\).
2. if \(\hat{\theta} < 0\) or \(\hat{\gamma}_j < 0\) then \(\hat{\theta}, \hat{\gamma}_j \in \hat{\beta}^w\), then
3. Apply LHA [23] to make \(\hat{\theta}\) and \(\hat{\gamma}_j\) positive
4. end if
5. Estimate \(\hat{Y}^-\) and \(\hat{Y}^+\) using (23)
6. return \(\hat{Y} = [\hat{Y}^-, \hat{Y}^+]\)

B. The Interval Regression Graph (IRG)

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, the crucial interpretation and communication of any insights is complex, yet just as vital as for the numeric case. We introduce a novel 3D visualization approach, the interval regression graph (IRG) which succinctly and clearly captures the change in a regressand’s key features (center and

\[
\begin{align*}
    \text{Algorithm 2 A weakly-constrained LM method (\(L_{MCw}\))} \\
    \text{Input: Let } Y = \{\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_n\} \text{ be the regressand variable’s samples and } X_1, X_2, ..., X_p \text{ are the } p \text{ (} \geq 1\text{) regressor variables.} \\
    \text{Output: Let } \hat{Y} = \{\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_n\} \text{ be the estimated } Y. \\
    1: \text{Estimate the range values of regressand variable, } \hat{Y}^w = \{\hat{y}_1^w, ..., \hat{y}_n^w\} \text{ using (29)} \\
    2: \text{if all } \hat{y}_i^w \geq 0 \text{ then} \\
    3: \text{Apply the LM method at (22) on } Y \text{ and } X_j \text{ (} 1 \leq j \leq p \text{) to generate } \hat{\beta}^* \text{ and } \hat{\beta}^w \\
    4: \text{Estimate } \hat{Y}^- \text{ and } \hat{Y}^+ \text{ using (23)} \\
    5: \text{else} \\
    6: \text{Apply Algorithm 1 to estimate } \hat{Y}^- \text{ and } \hat{Y}^+ \\
    7: \text{end if} \\
    8: \text{return } \hat{Y} = [\hat{Y}^-, \hat{Y}^+] \\
\end{align*}
\]
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\(X_{\text{min}}\) and range\(X_{\text{max}}\), as well as its minimum center and maximum center coordinates are center\(X_{\text{min}}\) and center\(X_{\text{max}}\).

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

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

*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.

range) for given changes in a regressor’s key features (center and range). The IRG visualizes the correlation between a regressor’s and regressand’s position (or range), based on a given regression model [31]. Note that in this paper, for conciseness, we only introduce and focus on IRGs generated for visualizing the relationship between one regressor and one regressand. Algorithm 3 presents the pseudocode for constructing IRGs for a regressand’s center and range with respect to a regressor’s center and range—for a given regression model.

To illustrate the IRG and its utility, consider the IV Set-1 (Fig. 1(a)), presented in the introductory section. Figures 3(a) and (b) individually show the two different aspects of the IRG 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 \((X^w \simeq 2X^c)\) 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 \((Y^w \simeq 2Y^c)\) (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, \(2X^c\) (or \(X^w\)) and its center, \(X^c\). Fig. 3(b) shows how the regressand’s range, \(2Y^c\) also increases with respect to both increasing values of the regressor’s range, \(2X^c\) 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. Data Sets

We use both synthetic and real-world IV data sets to investigate the performance of linear regression methods. The synthetic data sets 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 data sets are used to investigate their suitability to real-world scenarios. All data sets are described below.

1) Synthetic Data Sets: We specifically design a series of synthetic IV data sets to investigate the behavior of the regression models. First, Set-1 and Set-2, shown at the beginning of the paper (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 data sets are produced within the range of 0 to 10. They are given in Table VII (a) to (k) in the Appendix including the mean (\(\mu\)) and standard deviation (\(\sigma\)) of the center and range of intervals. Figure 4 graphically presents the data sets, Set-3 to Set-11.

2) Real-world Data Sets: We focus on three real-world IV data sets 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 data set represents medical observations and includes interval-valued systolic and diastolic blood pressure (BP) data of 59 patients of the Hospital Valle del Nalón in Asturias, Spain [11], [14].

The second data set presents IV consumer ratings of eight different (UK market) snack-food products [32]. In this case,
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 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 data sets, 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 paper, 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), $Y = \{\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n\}$ and its estimated value(s), $\hat{Y} = \{\bar{\hat{y}}_1, \bar{\hat{y}}_2, \ldots, \bar{\hat{y}}_n\}$.

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

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

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

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,

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

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

where $0 \leq MAE^-, MAE^+ \leq \mathbb{R}^+$. 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:

$$MMRE = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \left\{ \left| \frac{\bar{y}_i^- - \bar{\hat{y}}_i^-}{\bar{y}_i^-} \right| + \left| \frac{\bar{y}_i^+ - \bar{\hat{y}}_i^+}{\bar{y}_i^+} \right| \right\},$$

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

IV. RESULTS

We explore the performance of existing regression approaches together with two proposed variants of the LM method (LM$_c$ and LM$_p$) for synthetic and real-world IV data sets (see Section III-C). Their performance is assessed using different metrics described in Section III-D. All these regression methods are implemented in Matlab and the full source code will be available at http://lucidresearch.org/software with the publication of the paper.
A. Performance Analysis with Synthetic Data Sets

In Section III-B, we briefly introduced the IRGs using IV Set-1 (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 (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 regressor 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 III and IV 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 approach which does not guarantee mathematical coherence, and the PM—the most recent approach, which does. IRGs for all methods are shown in the appendix (Figs. 18 and 19). We independently present the relationship of the center and range of the test regressor on the center (subfigures (a)) and the range (subfigures (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 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 data set. The rest of the methods also do not perform according to the expectation (see Appendix, Fig. 18).

Further, 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 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 data sets (Set-3—Set-11), each designed to reflect a key
property regularly encountered in IV data sets. This will
highlight how different data sets—with different features—are
differentiable for different regression methods, as later
results will expand upon in more detail. For data Set-3 to
highlight how different data sets—with different features—are
property regularly encountered in IV data sets. This will

bold Bold Best

TABLE V: Performance of different regression models for IV synthetic data sets (Set-3 to Set-11)

<table>
<thead>
<tr>
<th>Methods</th>
<th>RMSE</th>
<th>MAE</th>
<th>MMRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CM</td>
<td>1.215</td>
<td>0.185</td>
<td></td>
</tr>
<tr>
<td>MaxMin</td>
<td>1.204</td>
<td>0.185</td>
<td></td>
</tr>
<tr>
<td>CRM</td>
<td>1.185</td>
<td>0.215</td>
<td></td>
</tr>
<tr>
<td>CCRM</td>
<td>1.195</td>
<td>0.215</td>
<td></td>
</tr>
<tr>
<td>CIM</td>
<td>1.215</td>
<td>0.215</td>
<td></td>
</tr>
<tr>
<td>PM</td>
<td>0.343</td>
<td>0.059</td>
<td></td>
</tr>
<tr>
<td>LM</td>
<td>0.343</td>
<td>0.059</td>
<td></td>
</tr>
</tbody>
</table>

Bold = Best

Set-3 (All disjoint skinny intervals)

Set-4 (Some overlapping skinny intervals)

Set-5 (All overlapping skinny intervals)

Set-6 (All disjoint puffy intervals)

Set-7 (Some overlapping puffy intervals)

Set-8 (All overlapping puffy intervals)

Set-9 (Disjoint mixed intervals)

Set-10 (Some overlapping mixed intervals)

Set-11 (Nested/all overlapping mixed intervals)

to three metrics and the LM<sub>w</sub> 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.

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

Set-10 (Some overlapping mixed intervals): The PM, LM<sub>c</sub>
and LM<sub>wt</sub> 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).

Set-11 (Nested/all overlapping mixed intervals): The LM<sub>o</sub>,
LM<sub>wt</sub>, 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<sub>o</sub> (and often LM<sub>c</sub>) methods per-
form best and result in reliably good—and almost identical—
fitness for disjoint skinny or puffy intervals. With growing
overlap among intervals (skiny or puffy), model fitness by the
CRM, CCRM 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

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
carried out additional experiments going beyond the scope of this paper—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 [38]. 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) [38]. We will review this aspect with real-world data in future.

B. Real-world Data Sets

In this section, three real-world IV data sets from the application contexts described in Section III-C, are used to explore and evaluate the behaviour 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., \( \text{RMSE}, \text{MAE}, \text{MMRE} \)) as before.

In addition, the larger degrees of freedom—compared to the synthetic data—of these real world data sets, permit us to empirically determine a measure of the variability associated with each performance value, utilizing a non-parametric resampling approach [39]. We therefore report bootstrapped (percentile) 95% confidence intervals [39][40] 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 non-significance of differences between the individual regression approaches’ error measures in question. As these were calculated on the same resampled data sets, the measures on each instance were non-independent, and so it is possible for variability evident in the CIs to represent common inter-simulation variance between measures, while for example one was a reliably better performer than another.

1) Systolic and Diastolic BP Data Set: For this data set, we regress the IV diastolic BP (regressand) on those of the systolic BP (regressor) to estimate the bounds of diastolic BP. Figure 8 presents this data set and the associated performance of the regression methods.

Figure 8(b) shows that the PM, \( \text{LM}_{\text{IRG}} \), and \( \text{LM}_{\text{c}} \) methods produce comparable performance in estimating the regressand bounds. With respect to some of the evaluation metrics (\( \text{RMSE}^+ \) and \( \text{RMSE}^- \), and \( \text{MAE}^+ \)), the PM produces the best estimation, whereas the \( \text{LM}_{\text{IRG}} \) and \( \text{LM}_{\text{c}} \) methods perform better for \( \text{MAE}^- \); all these three methods perform equally best for \( \text{MMRE} \). Fit for the CRM, CCRM and MinMax methods for these data is also better according to \( \text{MMRE} \). In this particular case, the CM and CMM 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 \( \text{IRGs} \) using the PM method. Figures 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 Data Set: With this data set, 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. 10 and 12). The estimation results in Figs. 10(b) and 12(b) show that the PM, \( \text{LM}_{\text{IRG}} \), \( \text{LM}_{\text{c}} \), 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 data set. This is especially true for the PM method for the data set representing taste and overall purchase intention—demonstrating how in cases where transformation is required to maintain mathematical coherence, this can have a large
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.

3) Cyber-security Data Set: Section III-C provided an overview of this data set. 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 paper. 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 VI summarises 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 interval-valued attribute ratings, and by comparison with experts’ actual overall ratings. For concise- ness, we present two evade hops – number 4 and number 22, shown in Figs. 14 and 15 respectively.

Figure 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 LMw method produces the joint best fit on mae+. The LMc method also performs very close to the LMw method. Again, the CM method provides poor fit to the data. As for the Evade-22 data set, Fig. 15 (b) displays similar results for Evade-22. Particularly, Fig. 15 reveals the PM method as the best approach in terms of most metrics. Further, the LMw, LMc, 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 data sets in Appendix (see Table VIII and IX). Note here again that bootstrap confidence intervals indicate substantially higher variability in model fit than for the first data set. However, in contrast to the second data set—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 data set, 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. 12: Graphical presentation of taste and overall purchase intention [32] and performance of different regression models as to evaluation metrics using this set.

(a) with respect to \( \hat{Y}^c \)  
(b) with respect to \( \hat{Y}^w \) (\( \approx 2\hat{Y}^w \))

Fig. 13: IRGs showing the relationship between taste and overall purchase intention [32] in terms of center and range for the PM method.

V. DISCUSSION ON SUITABILITY OF EXISTING REGRESSION APPROACHES

This section discusses the suitability of different regression techniques with respect to interval data sets 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 LM\(_w\) (and often LM\(_c\)) 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. In all cases, both MinMax and CRM require checking as to whether they maintain consistency of interval bounds. The real-world data experiments largely substantiate the synthetic experiments in terms of relative model performance.

Results reveal that the LM\(_w\) 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\(_w\) 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 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 data set. Generally, both CM and CIM methods yield poor model-fit.

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

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 favour the PM method as consistently producing strong model fit.
Fig. 14: Graphical presentation of Evade-4 data set [34] and performance of different regression models as to evaluation metrics using this data set.

Fig. 15: Graphical presentation of Evade-22 data set [34] and performance of different regression models as to evaluation metrics for this data set.

Fig. 16: A flowchart presenting recommendations on linear regression methods with respect to different IV data sets.

VI. CONCLUSIONS

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 visualisation in data analysis.

While well understood for the numeric case, linear regression for ‘fully’ IV data sets, i.e. data sets where both regressor and regressand are IV, remain challenging and an active research topic. At the same time, the potential for insights from such data sets 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 paper, we have carried out an in-depth review and analysis of linear regression models, exploring their behaviour with respect to systematically designed synthetic and real-world, fully IV data sets, 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, the paper 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 Interval Regression Graph (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 paper. 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 data sets, 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. Further, in this paper, 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, analyses have shown that the PM [21] and the LM [20] methods (with its weaker set of restrictions, i.e., LM_m) 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], LM_m, 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 analyse 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 paper, 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 modelling techniques, such as to non-linear 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 disjointive (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.

ACKNOWLEDGMENT

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.

REFERENCES


### APPENDIX

#### TABLE VII: Synthetic IV data sets used for sensitivity analysis (Set-1 to Set-11)

<table>
<thead>
<tr>
<th>Entry</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\gamma$</th>
<th>$\delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Set-1</td>
<td>1</td>
<td>0.6</td>
<td>0.59</td>
<td>1.0</td>
</tr>
<tr>
<td>(b) Set-2</td>
<td>1</td>
<td>0.6</td>
<td>0.65</td>
<td>1.0</td>
</tr>
<tr>
<td>(c) Set-3</td>
<td>1</td>
<td>0.6</td>
<td>0.71</td>
<td>1.0</td>
</tr>
<tr>
<td>(d) Set-4</td>
<td>1</td>
<td>0.6</td>
<td>0.76</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Example 1: Relatively low fit of the PM approach when the Box-Cox transformation is applied.

(a) Set-12 (b) 2D visualization (c) Model Performance

Fig. 17: Comparatively low fit of the PM approach.
### TABLE VIII: Estimated coefficients of different regression methods for synthetic data sets, Set-1 to set-11

<table>
<thead>
<tr>
<th>Methods</th>
<th>Set-1</th>
<th>Set-2</th>
<th>Synthetic data sets</th>
<th>Set-3</th>
<th>Set-4</th>
<th>Set-5</th>
<th>Set-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>CM</td>
<td>$\beta_0 = 1.256$, $\beta_1 = 0.43$</td>
<td>$\beta_0 = 1.316$, $\beta_1 = 0.46$</td>
<td>$\beta_0 = 4.608$, $\beta_1 = 0.261$</td>
<td>$\beta_0 = 5.637$, $\beta_1 = 0.172$</td>
<td>$\beta_0 = 10.54$, $\beta_1 = 0.091$</td>
<td>$\beta_0 = 5.866$, $\beta_1 = 0.017$</td>
<td></td>
</tr>
<tr>
<td>MinMax</td>
<td>$\beta_0 = 0.997$, $\beta_1 = 0.000$</td>
<td>$\beta_0 = 0.999$, $\beta_1 = 0.008$</td>
<td>$\beta_0 = 4.366$, $\beta_1 = 0.24$</td>
<td>$\beta_0 = 3.505$, $\beta_1 = 0.164$</td>
<td>$\beta_0 = 11.35$, $\beta_1 = 0.090$</td>
<td>$\beta_0 = 4.778$, $\beta_1 = 0.061$</td>
<td></td>
</tr>
<tr>
<td>CRM</td>
<td>$\beta_0 = 1.256$, $\beta_1 = 0.43$</td>
<td>$\beta_0 = 1.316$, $\beta_1 = 0.46$</td>
<td>$\beta_0 = 4.608$, $\beta_1 = 0.261$</td>
<td>$\beta_0 = 5.638$, $\beta_1 = 0.172$</td>
<td>$\beta_0 = 10.54$, $\beta_1 = 0.091$</td>
<td>$\beta_0 = 5.867$, $\beta_1 = 0.017$</td>
<td></td>
</tr>
<tr>
<td>CCRM</td>
<td>$\beta_0 = 1.256$, $\beta_1 = 0.43$</td>
<td>$\beta_0 = 1.316$, $\beta_1 = 0.46$</td>
<td>$\beta_0 = 4.608$, $\beta_1 = 0.261$</td>
<td>$\beta_0 = 5.638$, $\beta_1 = 0.172$</td>
<td>$\beta_0 = 10.54$, $\beta_1 = 0.091$</td>
<td>$\beta_0 = 5.866$, $\beta_1 = 0.017$</td>
<td></td>
</tr>
<tr>
<td>PM</td>
<td>$\alpha = 0.987$, $\omega = -0.025$</td>
<td>$\alpha = 0.999$, $\omega = 0.008$</td>
<td>$\alpha = 1.683$, $\omega = 15.185$</td>
<td>$\alpha = 2.295$, $\omega = -7.692$</td>
<td>$\alpha = 11.66$, $\omega = -0.60$</td>
<td>$\alpha = 7.26$, $\omega = 2.069$</td>
<td></td>
</tr>
<tr>
<td>CRM</td>
<td>$\alpha = 0.987$, $\omega = -0.025$</td>
<td>$\alpha = 0.999$, $\omega = 0.008$</td>
<td>$\alpha = 1.683$, $\omega = 15.185$</td>
<td>$\alpha = 2.295$, $\omega = -7.692$</td>
<td>$\alpha = 11.66$, $\omega = -0.60$</td>
<td>$\alpha = 7.26$, $\omega = 2.069$</td>
<td></td>
</tr>
<tr>
<td>CCRM</td>
<td>$\alpha = 0.987$, $\omega = -0.025$</td>
<td>$\alpha = 0.999$, $\omega = 0.008$</td>
<td>$\alpha = 1.683$, $\omega = 15.185$</td>
<td>$\alpha = 2.295$, $\omega = -7.692$</td>
<td>$\alpha = 11.66$, $\omega = -0.60$</td>
<td>$\alpha = 7.26$, $\omega = 2.069$</td>
<td></td>
</tr>
<tr>
<td>LM</td>
<td>$\eta = 1.003$</td>
<td>$\alpha = 2.045$</td>
<td>$\eta = -1.003$</td>
<td>$\eta = -1.003$</td>
<td>$\eta = -1.003$</td>
<td>$\eta = -1.003$</td>
<td></td>
</tr>
<tr>
<td>LMM</td>
<td>$\eta = 1.003$, $\alpha = 2.045$</td>
<td>$\eta = -1.003$, $\alpha = 2.045$</td>
<td>$\eta = 1.003$, $\alpha = 2.045$</td>
<td>$\eta = 1.003$, $\alpha = 2.045$</td>
<td>$\eta = 1.003$, $\alpha = 2.045$</td>
<td>$\eta = 1.003$, $\alpha = 2.045$</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE IX: Estimated coefficients of different regression methods for real-world IV data sets

<table>
<thead>
<tr>
<th>Methods</th>
<th>Systolic-Diastolic Blood Pressure data set [14]</th>
<th>Cyber-Security Vulnerability Assessment Data Set [34]</th>
</tr>
</thead>
<tbody>
<tr>
<td>CM</td>
<td>$\beta_0 = 1.886$, $\beta_1 = 0.438$</td>
<td>$\beta_0 = 25.979$, $\beta_1 = 0.029$</td>
</tr>
<tr>
<td>MinMax</td>
<td>$\beta_0 = 0.769$, $\beta_1 = 0.591$</td>
<td>$\beta_0 = -19.068$, $\beta_1 = 0.235$</td>
</tr>
<tr>
<td>CRM</td>
<td>$\beta_0 = 1.886$, $\beta_1 = 0.438$</td>
<td>$\beta_0 = 25.979$, $\beta_1 = 0.029$</td>
</tr>
<tr>
<td>CCRM</td>
<td>$\beta_0 = 2.056$, $\beta_1 = 0.438$</td>
<td>$\beta_0 = 25.979$, $\beta_1 = 0.029$</td>
</tr>
<tr>
<td>PM</td>
<td>$\alpha = 0.720$, $\alpha = 0.591$</td>
<td>$\alpha = 0.720$, $\alpha = 0.591$</td>
</tr>
<tr>
<td>CRM</td>
<td>$\alpha = 0.720$, $\alpha = 0.591$</td>
<td>$\alpha = 0.720$, $\alpha = 0.591$</td>
</tr>
<tr>
<td>CCRM</td>
<td>$\alpha = 0.720$, $\alpha = 0.591$</td>
<td>$\alpha = 0.720$, $\alpha = 0.591$</td>
</tr>
</tbody>
</table>

### Food Snacks Purchase Intention Data Set [35]

<table>
<thead>
<tr>
<th>Methods</th>
<th>Ethics and Overall purchase intention</th>
<th>Taste and Overall purchase intention</th>
</tr>
</thead>
<tbody>
<tr>
<td>CM</td>
<td>$\beta_0 = 23.9$, $\beta_1 = 0.497$</td>
<td>$\beta_0 = 25.395$, $\beta_1 = 0.462$</td>
</tr>
<tr>
<td>MinMax</td>
<td>$\beta_0 = 32.164$, $\beta_1 = 0.328$, $\beta_2 = 18.896$, $\beta_3 = 0.756$</td>
<td>$\beta_0 = 15.877$, $\beta_1 = 0.511$, $\beta_2 = 36.964$, $\beta_3 = 0.342$</td>
</tr>
<tr>
<td>CRM</td>
<td>$\beta_0 = 23.9$, $\beta_1 = 0.497$, $\beta_2 = 3.165$, $\beta_3 = 0.12$</td>
<td>$\beta_0 = 25.395$, $\beta_1 = 0.462$, $\beta_2 = 5.989$, $\beta_3 = 0.274$</td>
</tr>
<tr>
<td>CCRM</td>
<td>$\beta_0 = 23.9$, $\beta_1 = 0.497$, $\beta_2 = 3.165$, $\beta_3 = 0.12$</td>
<td>$\beta_0 = 25.395$, $\beta_1 = 0.462$, $\beta_2 = 5.989$, $\beta_3 = 0.274$</td>
</tr>
<tr>
<td>PM</td>
<td>$\beta_0 = 15.265$, $\beta_1 = 0.955$, $\beta_2 = 0.461$</td>
<td>$\beta_0 = 17.245$, $\beta_1 = 0.656$, $\beta_2 = 0.092$</td>
</tr>
<tr>
<td>CRM</td>
<td>$\beta_0 = 15.265$, $\beta_1 = 0.955$, $\beta_2 = 0.461$</td>
<td>$\beta_0 = 17.245$, $\beta_1 = 0.656$, $\beta_2 = 0.092$</td>
</tr>
<tr>
<td>LMM</td>
<td>$\eta = 13.761$, $\alpha = 0.062$, $\gamma = 0.478$</td>
<td>$\eta = 21.869$, $\alpha = 0.597$, $\gamma = 0.131$</td>
</tr>
<tr>
<td>LMM</td>
<td>$\eta = 13.761$, $\alpha = 0.062$, $\gamma = 0.478$</td>
<td>$\eta = 21.869$, $\alpha = 0.597$, $\gamma = 0.131$</td>
</tr>
</tbody>
</table>

© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.
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 (Fig. 1(a)).

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 (Fig. 2(a)).

Fig. 20: Number of times in 10000 simulations the regression approaches apply positivity restriction/transformation regressand bounds for real-world cases. BP=blood pressure data set, E&OPI=ethics vs overall purchase intention data set, T&OPI=taste vs overall purchase intention data set.