Visualization of Interval Regression for Facilitating Data and Model Insight

Shaily Kabir*†, Christian Wagner (Senior Member, IEEE)*
*Lab for Uncertainty in Data and Decision Making (LUCID),
School of Computer Science, University of Nottingham, Nottingham, UK
†Department of Computer Science and Engineering, University of Dhaka, Bangladesh
Email: {shaily.kabir3, christian.wagner}@nottingham.ac.uk, shailykabir@cse.du.ac.bd

***Abstract*—With growing significance of interval-valued data, interest in artificial intelligence methods tailored to this data type is similarly increasing across a range of application domains. Here, regression, i.e., the modelling of the association between interval-valued variables has been shown to be both challenging and rewarding. Beyond the mathematical challenges, fundamentals, such as the visualization of regression models, are not similarly available for interval-valued data, limiting both accessibility and utility of resulting models. Recently, the Interval Regression Graph (*IRG*) was introduced, providing a powerful visualization tool for interval-valued regression models. In this paper, we demonstrate the *IRG* in a practical data-science application, showing how it can rapidly highlight powerful insights of data. Specifically, we focus on consumer characteristics, analyzing potential relationships between their demographic characteristics and their product purchase intentions. We conclude with a brief outlook on the potential and remaining challenges of leveraging interval-valued data using fuzzy systems and artificial intelligence more broadly.**

***Index Terms*—Intervals, regression, uncertainty, linear**

I. Introduction

Interval-valued (IV) data have gained growing importance as a basic data type as they can capture information entirely with an intrinsic representation of range or uncertainty in each individual 'measurement', which is not possible with point-values, such as numbers or ranks [1]. Such IV data may arise from imprecision and uncertainty in measurement in sensor data, uncertainty of outcome in stock prices, or in vagueness or nuance in linguistic terms [2]–[4].

Regression for IV data is a fundamental step from a statistical and artificial intelligence (AI) point of view, and it is being increasingly applied in domains ranging from marketing to cyber-security, modelling of the relationships between variables and their inherent uncertainty or range [5], [6]. For example, the regression of IV consumer preference data can allow us to infer not only how a snack's nutritional benefits influence purchase intention of consumers, but also how *uncertainty* about these benefits impact is associated with *uncertainty* in purchase intention [7], a crucial insight from a marketing perspective.

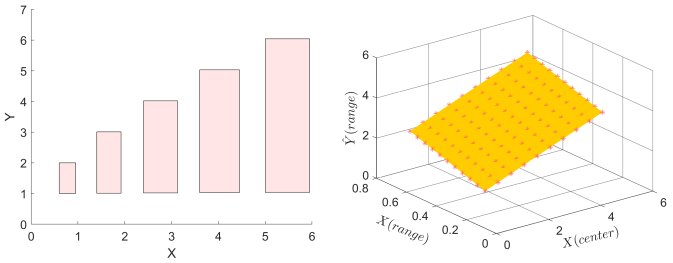
This work was supported in part by the UK EPSRC's Leveraging the Multi-Stakeholder Nature of Cyber Security (EP/P011918/1) and Horizon: Trusted Data-Driven Products (EP/T022493/1) grants.

A number of linear regression approaches have been developed for IV data, using different *reference points* of intervals, such as center values, lower and upper bounds, or center and range (width) as regression variables [3]. While earlier approaches struggled maintaining mathematical coherence in the models, i.e., ensuring lower interval bounds are smaller than upper bounds, [8]–[10], the most recent approaches [11]–[14] adopt refined strategies and algorithms to ensure the coherence of bounds. A detailed review of the state of the art approaches, their behaviour, advantages and pitfalls using both synthetic and real-world data sets having various properties are provided in [15]. Experiments consistently show that among the existing interval linear regression approaches, the Parametrized Model (PM) introduced in [14] produces the best model fit overall, for a variety of IV data sets.

The same paper [15] also introduces the aforementioned novel visualization for IV regression models, drastically improving their interpretability and thus accessibility and utility. These models are referred to as 'interval regression graph' (*IRG*) and succinctly show the complex relationship between IV regressor and regressand in respect to both position (center) and uncertainty (range) within a given IV regression model. To illustrate the *IRG*, we show an example of a synthetic IV data set in Fig. 1(a) where the position and range of the regressand, Y increase with the position and range of regressor, X—that is, they vary in unison. This reflects a common case, e.g., the price of cars vs their horsepower. The relationship between the variables in this figure—as modelled by the regression method—is captured through the *IRG* in Fig. 1(b). The *IRG* will be discussed in more details in Section III.

Going beyond the journal paper, this paper aims to expand on the interpretability of the *IRG*, demonstrating how it can provide rapid insight into the relationship of IV variables via the respective regression models in real-world applications such as marketing and consumer-insight.

Here, extant studies using numeric data suggest that consumers' purchase intention is related to their personal demographic characteristics, such as age, gender, race and income [16], [17]. These studies indicate that consumers desire to buy products are significantly different across age-groups, gender, and income using statistical tests, such as t-tests or ANOVA tests. For example, they highlight the importance of



(a) Set-1 (b) *IRG* for Set-1

Fig. 1: (a) Visualization of IV data Set-1 and (b) *IRG* in respect to $\hat{Y}$(range) using the PM method [14].

ethical consumption or socially responsible consumption with associated low environment footprint in respect to purchase intention [17]. Further, they show that this sustainability focus is more common in older individuals, while environmentally-conscious and health-conscious attitudes were more strongly expressed among females compared to males [17].

With the view to exploring the relevance of uncertainty and range in these variables, and leveraging linear regression, this paper examines such relationships where consumers provide their purchase intention of snack foods using IV data for different attributes of the products. We demonstrate how the *IRG* can support the effective articulation of the relationship between purchase intention and different attributes (e.g., visual appeal, taste, healthiness, ethics) in respect to demographic factors of consumers (e.g., age, gender). We note that while *IRGs* are independent of the underlying regression approach, i.e. can be generated based on all IV regression methods, we only consider the PM method as it consistently produces strong results in terms of model fit among existing methods [14], [15].

The paper is organized as follows: Section II describes IV data sets and reviews the best IV linear regression methods where both dependent and independent variables are IV, based on the vector representation of intervals in the regression process. Section III introduces the interval regression graph (*IRG*) and Section IV demonstrates the behaviour of the *IRG* in interpreting the relationship between variables in terms of their position and range for the real-world data set. Lastly, Section V concludes the paper and highlights future work. Table I presents a list of acronyms and notation used in this paper to assist the reader.

II. Background

In this section, all IV data sets used in this paper are introduced. Then, a brief review of well-known linear regression models for IV data is provided followed by a detailed discussion on one of the leading regression models—the Parametrized Model (PM) [14]. Note that this paper focuses on IV regression as it provides a natural underpinning for future extension to more complex data types such as fuzzy sets.

A. IV Data Sets

An interval $\overline{a}$ is defined by its left and right endpoints, a^- and a^+ with $a^- \leq a^+$ [18]. a^- and a^+ are also referred to as the lower and upper bounds of $\overline{a}$. It is generally presented as $\overline{a} = [a^-, a^+]$, however, its alternate representation is $[a^c - \frac{a^w}{2}, a^c + \frac{a^w}{2}]$ where $a^c = \frac{(a^+ + a^-)}{2}$ is its center and $a^w = |a^+ - a^-|$ is its range [18][1]. A set of intervals forms an IV data set. In this paper, all data sets are 'fully' IV, i.e., all variables, both independent and dependent variables are IV.

TABLE I: Acronyms and Notation

CM	Center Method [8]
MinMax	MinMax Method [9]
CRM	Center and Range Method [10]
CCRM	Constrained Center and Range Method [11]
CIM	Complete Information Method [12]
LM	Linear Model [13]
PM	Parametrized Model [14]
IV	Interval-Valued
IRG	Interval Regression Graph
$\overline{a}$	Interval $\{\overline{a} \subseteq \mathbb{R} : \overline{a} = [a^-, a^+],\ a^- \leq a^+\}$
a^w	Range of Interval, $a^w = \lvert\overline{a}\rvert = \lvert a^+ - a^-\rvert$
a^c	Center of Interval, $a^c = \frac{(a^+ + a^-)}{2}$
a^r	Radius of Interval $a^r \simeq \frac{1}{2}a^w$ (Half Range)
$\mathbf{Y}$	IV Regressand
$\hat{\mathbf{Y}}$	Estimated $\mathbf{Y}$
$\mathbf{X}$	IV Regressor(s)

B. Linear Regression Models for IV Data Sets

A number of linear regression approaches have been put forward for IV regressand and regressor(s). We succinctly review key regression approaches in chronological order. In this review, we consider $\mathbf{Y} = \{\overline{y}_1, \overline{y}_2, \ldots, \overline{y}_n\}$ as a regressand with n intervals where $\overline{y}_i = [y_i^-, y_i^+]$, $1 \leq i \leq n$ and $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p\}$ as $p \geq 1$ regressors where each $\mathbf{X}_j$ also has n intervals, $\mathbf{X}_j = \{\overline{x}_{j1}, \overline{x}_{j2}, \ldots, \overline{x}_{jn}\}$ with $\overline{x}_{ji} = [x_{ji}^-, x_{ji}^+]$, $1 \leq j \leq p$, $1 \leq i \leq n$. The estimated regressand is defined as $\hat{\mathbf{Y}} = \{\hat{\overline{y}}_1, \hat{\overline{y}}_2, \ldots, \hat{\overline{y}}_n\}$.

Within the existing interval regression approaches, the Center Method (CM), proposed by Billard and Diday [8] in 2000, is considered as the initial approach to perform regression on interval regressor and regressand. It uses interval center of both regressand and regressor to build the bi-variate vectors and then compute regressor coefficients. These coefficients are later applied with the regressor lower and upper bounds to separately estimate the regressand lower and upper bounds. This approach is simple but faces two major drawbacks. First, it uses the same coefficients to estimate both regressand bounds which often leads to poor estimation and violation of the mathematical coherence of regressand bounds where the estimated lower bound is greater than the estimated upper bound [11]. Second, the resulting regression is often too restrictive as it imposes the centers' behavior on the bounds [14].

To improve regression performance, Billard and Diday [9] later developed the MinMax method in 2002 to directly utilize the lower and upper bounds of the regressors to separately estimate the lower and upper bounds of the regressand. Using two separate models in the MinMax approach to estimate the

[1]For an interval $\overline{a}$, its half of range is often termed as its radius, defined as $a^r \simeq \frac{1}{2}a^w$ [18].

regressand bounds improves the model fitness and interpretation compared to the CM method, however, it does not guarantee the mathematical coherence of regressand bounds [10]. In addition, the model fitness of the MinMax method can be reduced if there is not a clear dependency between the respective bounds of regressand and regressor [14]. In this regard, Neto and Carvalho [10] also developed the Center and Range Method (CRM) in 2008 which considers not only the interval centers but also involves the range of regressor and regressand variables to estimate the regressor coefficients. They build two separate regression models—one for the centers and other for the ranges of variables. They compute the regressor coefficients separately for the center and range estimations and apply them along with the center and range of regressor to estimate those of regressand. The CRM method subsequently uses the estimated regressand center and range to compute its lower and upper bounds. The CRM model provides better estimation than the CM method when there is a linear dependency between the ranges of regressand and regressors [11], [14]. However, this improved fitness can be observed only when there exists range dependency. In addition, it still does not ensure mathematical coherence on bounds [14].

All of the regression models discussed so far face one common drawback—they do not guarantee the mathematical coherence of regressand bounds—one of the fundamental properties of intervals (left endpoint $\leq$ right endpoint). To maintain such coherence, Neto and Carvalho [11] later adapted the CRM model [10] by enforcing positivity restrictions on the coefficients which are estimated in respect to the relationship of the range of regressand and regressor variables. The adapted method is known as the Constrained Center and Range Method (CCRM), where the overall process of estimating the regressand bounds remains the same as for the CRM model with positivity constraints on the range coefficients. To enforce the constraints, the CCRM method applies an iterative algorithm proposed by Lawson and Hanson in [19]. While the CCRM model guarantees mathematical coherence, it can lead to biased estimation outcomes [11]. In this regard, 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 such coherence [11]. In particular, the positivity restriction within the CCRM method forces any negative range coefficient to 0 and updates the remaining range coefficients, in turn leading to potentially biased estimation outcomes and poor model fitness [11].

To reduce bias in the estimation process, Wang et al. [12] proposed the Complete Information Method (CIM) in 2012 which considers all internal points of intervals in the estimation process. It basically models each interval observation of regressand and regressor variables as a hyper-cube and builds the regression model on these hyper-cubes. It adopts Moore's linear combination algorithm [20] through an indicator function to ensure the consistency of bounds, where an indicator attached to a coefficient turns into 0 whenever the coefficient is negative, otherwise it is set to 1. This positivity indication helps keeping the mathematical coherence by the CIM method but at the price of potentially poor model fit [14], [15].

To maximize model performance while preserving its flexibility and interpretability, Sun and Ralescu [13] developed the Linear Model (LM) in 2015 based on the affine operator in the cone $\mathcal{C} = \{(x, y) \in \mathbb{R}^2 | x \leq y\}$. The LM method considers both lower and upper bounds of regressors and their ranges for estimating the bounds of the regressand. Here, the IV regressand is considered as a linear transformation of the IV regressor. This approach also imposes positivity constraints on the range coefficients to ensure the coherence of interval bounds. Even though the authors assume positivity constraints on range coefficients, the actual model setting does not ensure compliance with these constraints. As a result, it can result in negative range coefficients—which may lead to flipped interval bounds. The authors do not discuss how to maintain these constraints in practice, though they expect that if any estimated range coefficient turns out to be negative, forcing it to be positive may lead to poor fitness of the LM model. In this regard, the LM method has been extended by enforcing positive restrictions on range coefficients only when needed to avoid unnecessary estimation bias and made it suitable for practical real-world deployment [15].

From the above discussion, it is clear that imposing positivity restriction on coefficients in the regression approach to ensure the mathematical coherence can lead to poorer regression performance. In the next section, we describe one of the most recent interval regression approaches—Parametrized Model [14]—which maintains coherence, adjusting the model as needed-only and delivers overall superior model fit in comparison to the other state-of-the-art approaches [6].

C. The Parametrized Model for IV Data

Souza et al. [14] developed the Parametrized Model (PM) in 2017 which also uses two different models for the regressand bounds. Instead of using specific interval points, such as center, range, interval bounds, the PM method automatically extracts the best reference points from the regressors and uses them to build regression models for both lower and upper bounds of the regressand. Here, an interval is considered as a line segment. For instance, given an interval $\overline{a}$, any point $q \in \overline{a}$ can be computed as $q = a^-(1-\lambda) + a^+\lambda$, $0 \leq \lambda \leq 1$. By setting λ, $\overline{a}$ is turned into a single point. Hence, when $\lambda = 0$, $q = a^-$ (lower bound of $\overline{a}$) and when $\lambda = 1$, $q = a^+$ (upper bound of $\overline{a}$). Similarly, $q = a^c$ (center of $\overline{a}$) when $\lambda = 0.5$. Utilizing this concept, the PM method specifies the linear regression models for the lower and upper bounds of Y in (1).

$$
\begin{aligned}
y_i^- &= \beta_0^- + \sum_{j=1}^{p} \beta_j^-(1-\lambda_j)x_{ji}^- + \beta_j^-\lambda_j x_{ji}^+ + \epsilon_i^-, \\
y_i^+ &= \beta_0^+ + \sum_{j=1}^{p} \beta_j^+(1-\lambda_j)x_{ji}^- + \beta_j^+\lambda_j x_{ji}^+ + \epsilon_i^+.
\end{aligned} \tag{1}
$$

Equation (2) simplifies (1) by replacing $\beta_j^-(1-\lambda_j)$ by α_j^- and $\beta_j^-\lambda_j$ by ω_j^- for lower bounds, and $\beta_j^+(1-\lambda_j)$ and $\beta_j^+\lambda_j$

by α_j^+ and ω_j^+ respectively.

$$\begin{aligned} y_i^- &= \beta_0^- + \sum_{j=1}^{p} \alpha_j^- x_{ji}^- + \omega_j^- x_{ji}^+ + \epsilon_i^- \\ y_i^+ &= \beta_0^+ + \sum_{j=1}^{p} \alpha_j^+ x_{ji}^- + \omega_j^+ x_{ji}^+ + \epsilon_i^+. \end{aligned} \tag{2}$$

In matrix notation, the lower bound model can be expressed for all n observations as $\mathbf{Y}^- = \mathbf{X}^* \boldsymbol{\beta}^- + \boldsymbol{\epsilon}^-$, where

$$\mathbf{Y}^- = (y_1^- \quad y_2^- \quad \dots \quad y_n^-)^T, \quad \boldsymbol{\beta}^- = (\beta_0^- \quad \alpha_1^- \quad \omega_1^- \quad \dots \quad \alpha_p^- \quad \omega_p^-)^T,$$

$$\boldsymbol{\epsilon}^- = (\epsilon_1^- \quad \epsilon_2^- \quad \dots \quad \epsilon_n^-)^T, \text{ and } \mathbf{X}^* = \begin{pmatrix} 1 & x_{11}^- & x_{11}^+ & \dots & x_{p1}^+ \\ 1 & x_{12}^- & x_{12}^+ & \dots & x_{p2}^+ \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{1n}^- & x_{1n}^+ & \dots & x_{pn}^+ \end{pmatrix}.$$

The LS estimate of the coefficients for the lower bound model, $\boldsymbol{\beta}^-$ is computed by (3).

$$\hat{\boldsymbol{\beta}}^- = ((\mathbf{X}^*)^T \mathbf{X}^*)^{-1} (\mathbf{X}^*)^T \mathbf{Y}^- \tag{3}$$

The matrix expression follows the same pattern for the upper bound model, $\mathbf{Y}^+ = \mathbf{X}^* \boldsymbol{\beta}^+ + \boldsymbol{\epsilon}^+$, and the LS estimate of the coefficients for the upper bound model, $\boldsymbol{\beta}^+$ in defined in (4).

$$\hat{\boldsymbol{\beta}}^+ = ((\mathbf{X}^*)^T \mathbf{X}^*)^{-1} (\mathbf{X}^*)^T \mathbf{Y}^+ \tag{4}$$

Finally, using $\hat{\boldsymbol{\beta}}^-$ and $\hat{\boldsymbol{\beta}}^+$, the lower and upper bounds of $\mathbf{Y}$ are estimated using (5).

$$\begin{aligned} \hat{y}_i^- &= \hat{\beta}_0^- + \sum_{j=1}^{p} \hat{\alpha}_j^- x_{ji}^- + \hat{\omega}_j^- x_{ji}^+ \\ \hat{y}_i^+ &= \hat{\beta}_0^+ + \sum_{j=1}^{p} \hat{\alpha}_j^+ x_{ji}^- + \hat{\omega}_j^+ x_{ji}^+. \end{aligned} \tag{5}$$

The PM method does not automatically guarantee the mathematical coherence of the bounds [14]. To avoid flipping the interval bounds, the approach estimates the range of Y using (6) before performing the regression.

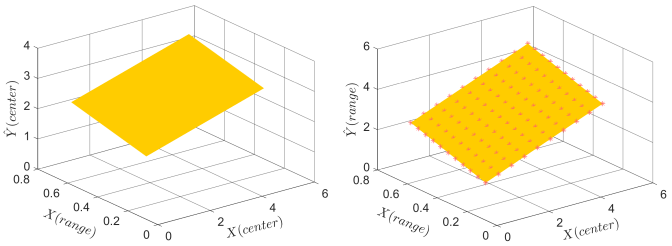
$$\hat{\mathbf{Y}}^w = \mathbf{X}^* ((\mathbf{X}^*)^T \mathbf{X}^*)^{-1} (\mathbf{X}^*)^T \mathbf{Y}^w \tag{6}$$

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 [21], extended to intervals by the authors [14], to transform the regressand so that the desirable coherence is achieved by the PM method. Equation (7) defines the extended Box-Cox transformation for the interval $\overline{y}_i = [y_i^-, y_i^+]$.

$$\overline{y}_i^k = \begin{cases} \left[\frac{(y_i^- + k_2)^{k_1} - 1}{k_1}, \frac{(y_i^+ + k_2)^{k_1} - 1}{k_1} \right], & \text{if } k_1 \neq 0. \\ \left[log(y_i^- + k_2), log(y_i^+ + k_2) \right], & \text{if } k_1 = 0. \end{cases} \tag{7}$$

where k_1 is any real value and k_2 is under the following restriction: $y_i^- + k_2 > 0$.

In the next section, we present the recently introduced visualization approach for IV regression[1]—the interval regression graph (*IRG*)—which visualizes the relationship in terms of both center and range between IV regressor and regressand.



(a) *IRG* as to $\hat{Y}(center)$ (b) *IRG* as to $\hat{Y}(range)$

Fig. 2: Relationship between regressand and regressor in respect to center (position) (a) and range (b) using the PM method for Set-1 (Fig. 1(a)).

III. The Interval Regression Graph (IRG) for IV Regression

Visualization of regression provides a powerful way to interpret and communicate the relationship between variables. Intervals are complex compared to the numeric data, and similarly, the interpretation and communication of any insights from intervals and/or associated regression can be complex. To facilitate and enhance the interpretability of interval regression, a powerful and novel 3D visualization approach—the interval regression graph (*IRG*)—is introduced [15] which succinctly visualizes the relationship between an independent and dependent variable of interest. In other words, *IRG*s capture the change in a regressand's key features (center and range) for given changes in a regressor's key features (center and range)–for a given regression model. Note that while it is tempting to think about such a visualisation as a comparatively simple 2D representation using two regression lines, e.g. one for the upper and one for the lower endpoints, this is not possible as there are not sufficient degrees of freedom in a 2D visualization to represent both center and range of both regressor and regressand.

In this paper, we focus on regression and *IRG*s between individual variables. We will consider the multi-variate case in future publications. Algorithm 1 presents the pseudocode for generating the *IRG*s for regressand center and range in respect to a regressor's center and range–for a given regression model.

To illustrate the *IRG* and its use, consider the IV Set-1 (Fig. 1(a)), presented in the introductory section. Figures 2(a) and (b) separately present the two different aspects of the *IRG* for Set-1 based on the PM regression method. The bottom-left and bottom-right axes always show the range ($X(range)$) and regressor's center ($X(center)$) respectively. In Fig. 2(a), the vertical axis denotes the regressand's estimated center ($\hat{Y}(center)$), while in Fig. 2(b), it reflects the regressand's estimated range ($\hat{Y}(range)$) .

Interpreting these figures, we can see how Fig. 2(a) visualizes that the regressand's center, $\hat{Y}(center)$ increases in respect to the increasing values of both the regressor's range, $X(range)$ and center, $X(center)$. Fig. 2(b) shows how the regressand's range, $\hat{Y}(range)$ also increases in respect to both increasing values of the regressor's range, $X(range)$ and

Algorithm 1 Interval Regression Graph (*IRG*) Generation

Input: An IV regression model. We use the PM method [14] here. The IV regressor's (e.g., from the original data set) minimum range and maximum range are $rangeX_{min}$ and $rangeX_{max}$, as well as its minimum center and maximum center coordinates are $centerX_{min}$ and $centerX_{max}$.

Output: Two *IRG* plots mapping the regressor to the regressand's center, $\hat{Y}(center)$ and range, $\hat{Y}(range)$.

1: Generate the set **X(range)** of p discretizations of the interval $[rangeX_{min}, rangeX_{max}]$
2: Generate the set **X(center)** of q discretizations of the interval $[centerX_{min}, centerX_{max}]$
3: **for** each discretized $X(range)_i, 1 \leq i \leq p$ **do**
4: **for** each discretized $X(center)_j, 1 \leq j \leq q$ **do**
5: Compute $X(left)_{ij} = X(center)_j - \frac{X(range)_i}{2}$
6: Compute $X(right)_{ij} = X(center)_j + \frac{X(range)_i}{2}$
7: Compute $\hat{Y}(left)_{ij}$, $\hat{Y}(right)_{ij}$ with $X(left)_{ij}$, $X(right)_{ij}$ using the regression model
8: Compute $\hat{Y}(center)_{ij} = \frac{\hat{Y}(left)_{ij} + \hat{Y}(right)_{ij}}{2}$
9: Compute $\hat{Y}(range)_{ij} = \hat{Y}(right)_{ij} - \hat{Y}(left)_{ij}$
10: **end for**
11: **end for**
12: Generate $\hat{Y}(center)$ 3D *IRG* plot with $\hat{\mathbf{Y}}(\mathbf{center})$ on the vertical axis, **X(range)** on the bottom left axis and **X(center)** on the bottom right axis.
13: Generate $\hat{Y}(range)$ 3D *IRG* plot with $\hat{\mathbf{Y}}(\mathbf{range})$ on the vertical axis, **X(range)** on the bottom left axis and **X(center)** on the bottom right axis.
14: **return** *IRG* plots for $\hat{Y}(center)$ and $\hat{Y}(range)$.

center, $X(center)$, and that it does so at a greater rate in each case than does $\hat{Y}(center)$.

IV. Demonstration

In this section, we demonstrate the use of *IRG*s to visualize and interpret the relationship between IV regressand and regressor in a given data set. We use a real-world data set on IV consumer ratings of eight (UK market) snack-food products [7]. In this set, 40 consumers rated each product—using the 'DECSYS' interval open-source survey software [22]—based on different attributes, such as, their nutritional value, healthiness, their branding, ethics, price, and taste, as well as their overall purchase intention (*OPI*) for the given products. 62% of participating consumers were female, and the rest were male with varying ages between 18 to 55. All responses were collected on a scale from 0 to 100. Table II presents the survey questions given to the consumers.

In this paper, we explore whether and how *IRG*s can serve to capture and visually communicate the inherent relationship between consumers' *OPI* and their demographic characteristics, such as *age*, *gender*—for IV data, similar to how the traditional 'regression line plots' articulate such relationships for discrete data.

TABLE II: Food Snacks Purchase Intention Data Set: Product Attributes and Associated Survey Question [7].

Attribute	Survey Question
Visual Appeal	How much do you like the look of this product?
Value for Money	How happy would you be to pay x for this product? (where x is the retail price per item for the product in question)
Healthiness	How much can this product contribute to a healthy diet?
Taste	How much do you like the taste of this product?
Branding	How much does the product brand appeal to you?
Ethics	How ethical is this product?
Overall Purchase Intention	Overall, how likely are you to buy this product?

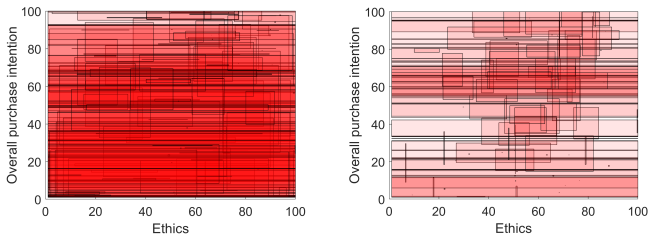
As mentioned in the introduction, studies using numeric variables suggest that consumers desire to buy are linked to their personal demographic characteristics. In particular, *OPI* varies across age groups and gender. Further, younger people tend to value more on taste and visual appeal than older individuals. Similarly, ethical standards and health consciousness appear to have a stronger impact on the *OPI* of females than males. Throughout this section, we will explore whether these or similar insights are found for IV data, and whether additional insights can be identified based on the richer nature of IV data[2].

A. Ethical Standards and Gender

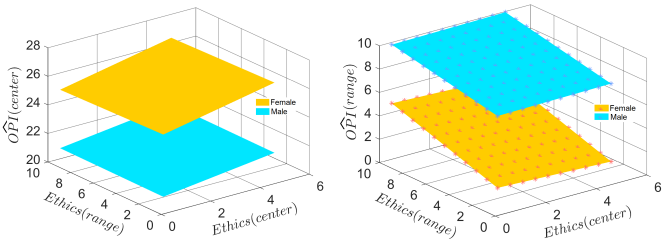
We first explore potential differences in *OPI* in respect to different levels of ethical standards for males and females. We separately regressed *OPI* (regressand) on *ethics* (regressor) for both females and males with the PM method. Figures 3(a) and (b) present the data sets for female and male consumers respectively. Figures 3(c) and (d) present the *IRG*s capturing the relationship of *ethics* on the *OPI* in respect to their center and range (uncertainty) for both groups.

First, the *IRG*s in Fig. 3(c) reveal that the center/position of *OPI* varies solely in respect to the center/position of *ethical standards* for both males and females. It also shows that higher *ethical values* lead to higher *OPI* in both cases. However, ethical standards seem more important to female than male customers in the sample overall. Perhaps more interestingly, and uniquely 'visible' for IV data, the *IRG* in Fig. 3(d) shows that the range/uncertainty of *OPI* varies quasi uniquely in respect to the range/uncertainty of *ethics*, with a little impact of its center, i.e. uncertainty on ethical standards is directly related to the uncertainty in *OPI*. Males are overall more uncertain, while the positive relationship is slightly stronger for females consumers.

[2]Note that the sample size is small, i.e., it is sufficient for the aims of this paper, i.e., demonstration of *IRG*s, but we emphasise that no general insights into consumer behavior should be inferred from this paper. We are consciously not exploring a further statistical analysis of the correlations here for the same reason. At the same time, the intriguing insights identified in this paper in our view serve to underline the value of using IV data, including as part of a much larger, balanced sample to further our broader understanding of consumers.

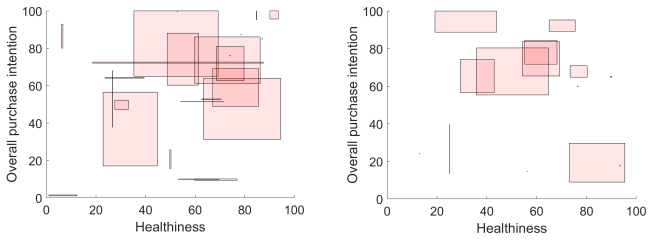


(a) *Ethics* and *OPI* for female consumers (b) *Ethics* and *OPI* for male consumers

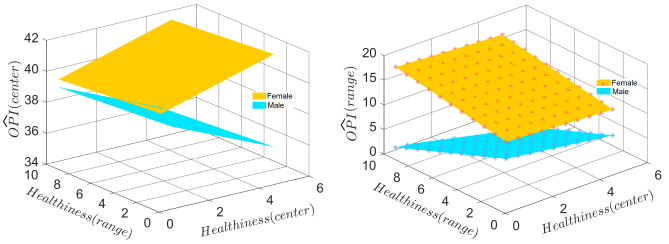


(c) *IRG* in respect to *OPI*(center) (d) *IRG* in respect to *OPI*(range)

Fig. 3: Ethics and Gender. *OPI=Overall purchase intention*.



(a) *Healthiness* and *OPI* for female consumers (b) *Healthiness* and *OPI* for male consumers



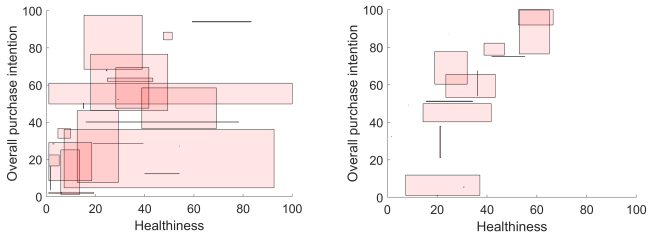
(c) *IRG* in respect to *OPI*(center) (d) *IRG* in respect to *OPI*(range)

Fig. 4: Health Consciousness and Gender (for health-focused, branded snack bar). *OPI=Overall purchase intention*.

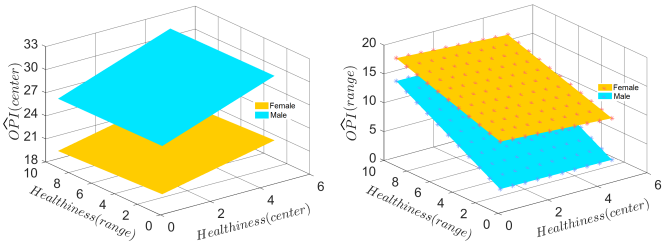
B. Health Consciousness and Gender

This section inspects if female consumers' health consciousness differs from that of male consumers in respect to *OPI*. We split all products into two categories considering their nutritional value: one category termed as 'health-focused, branded snack bar', i.e., products with higher nutritional value and the other category as 'value snack bar', i.e., generally cheaper products with lower nutritional value. For the purposes of this paper, we selected two products, one from each category and explored their perception by female and male groups. We separately performed regression of *OPI* (regressand) in respect to *visual appeal* (regressor) for both sets with the PM method. Figures 4(a) and (b) present the data sets for female and male consumers respectively for the 'health-focused, branded snack bar'. Similarly, Figs. 5(a) and (b) show the data sets for female and male consumers respectively for the 'value snack bar'. Figures 4(c) and (d) (also Figs. 5(c) and (d)) present the *IRG*s capturing the relationship of *healthiness* on the *OPI* in respect to their center and range for the female and male consumers.

From the *IRG*s of the center of *OPI* in Figs. 4(c) and 5(c), it is visible that female consumers give more importance to the healthiness and nutritional value of the products in their intention to purchase them compared to the male consumers and their purchasing intention increases with the increase of healthiness of the product. On the other hand, the *IRG*s of the range of *OPI* in Figs. 4(d) and 5(d), show that female consumers are getting more uncertain in their intention to purchase a product when they are not certain about the healthiness and nutritional value of that product as compared to male consumers.



(a) *Healthiness* and *OPI* for female consumers (b) *Healthiness* and *OPI* for male consumers



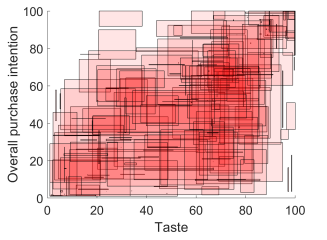
(c) *IRG* in respect to *OPI*(center) (d) *IRG* in respect to *OPI*(range)

Fig. 5: Health Consciousness and Gender (for value snack bar). *OPI=Overall purchase intention*.

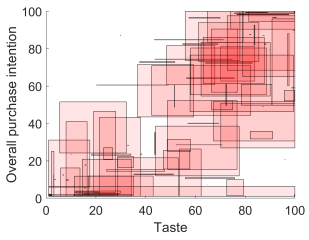
C. Taste and Age Groups

This section explores differences between younger and older consumers in the impact of *taste* on *OPI*. We divided the consumers into two groups—'younger' with age less than or equal to 25 and 'older' with age above 25. This split in the sample was driven purely by the distribution of age, i.e. generating two groups of comparable size. Again, if an actual market research study was conducted, a representative sample with age groups driven for example by target consumer groups would be more meaningful.

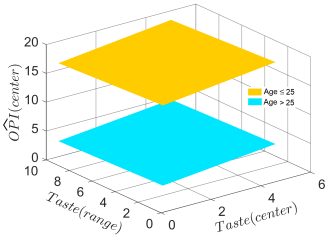
We separately regressed *OPI* on *taste* for both sets with the PM method. Figures 6(a) and (b) present the data sets for 'younger' and 'older' consumers respectively. Figures 6(c) and (d) present the *IRG*s capturing the relationship of *taste* on the



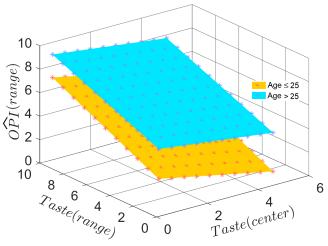
(a) *Taste* and *OPI* for young consumers (age $\leq$ 25)



(b) *Taste* and *OPI* for old consumers (age $>$ 25)

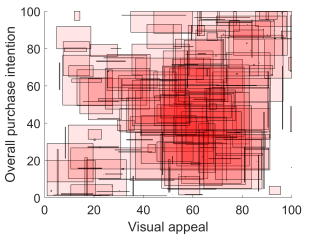


(c) *IRG* in respect to *OPI*(center)

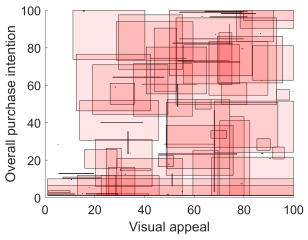


(d) *IRG* in respect to *OPI*(range)

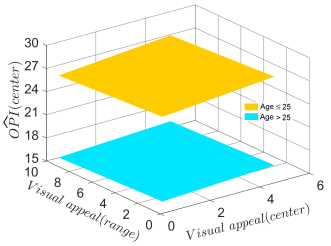
Fig. 6: Taste and Age groups. *OPI=Overall purchase intention*.



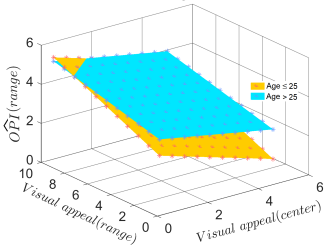
(a) *Visual appeal* and *OPI* for young consumers (age $\leq$ 25)



(b) *Visual appeal* and *OPI* for old consumers (age $>$ 25)



(c) *IRG* in respect to *OPI*(center)



(d) *IRG* in respect to *OPI*(range)

Fig. 7: Visual appeal and Age groups. *OPI=Overall purchase intention*.

OPI in respect to their center and range (uncertainty) for each of 'younger' and 'older' sets. The *IRG* in Fig. 6(c) shows that in this case also the center of *OPT* varies in respect to the center of *taste* for both sets. It also shows that the 'younger' consumers give more importance to the *taste* of products in respect to their *OPI*. The *IRG* in Fig. 6(d) show that the uncertainty in *OPI* increases in respect to the increase in both center and range of *taste*. Interestingly, a higher decline in the uncertainty of *OPI* is observed in particular for the 'younger' consumers for higher center value of *taste*.

D. Visual Appeal and Age Groups

We explore whether younger individuals value visual appeal differently to older consumers in respect to *OPI*, using the same partition for 'younger' and 'older' as in the previous section. We separately regressed *OPI* (regressand) on *visual appeal* (regressor) for both sets with the PM method. Figures 7(a) and (b) present the data sets for 'younger' and 'older' consumers respectively. Figures 7(c) and (d) present the *IRG*s capturing the relationship of *visual appeal* on the *OPI* in respect to their center and range (uncertainty) for the 'younger' and 'older' data sets.

The *IRG* in Fig. 7(c) shows that the center of *OPI* varies a small amount in respect to the center of *visual appeal* for both sets, increasing slightly with improved visual appeal. It also highlights that the *OPI* overall is substantially higher for the 'younger' consumers. The *IRG* in Fig. 7(d) shows that the uncertainty in *OPI* increases in respect to the increasing range/uncertainty on *visual appeal* and decreases in respect to the increasing center/position of the same. Again, a higher decline in the uncertainty of *OPI* is seen for the 'younger' consumers for the higher center value of *visual appeal*.

V. Conclusions

Recognizing the importance of visualization of regression results, this paper presents a series of illustrations of how and where interval-valued (IV) data, combined with recently introduced IV regression models—featured with a novel visual tool—Interval Regression Graphs (*IRG*s)—offer rapid and otherwise inaccessible insight into data, and as in this case, consumer behavior.

Through a series of experiments, we demonstrate how *IRG*s as a novel visualization approach can clearly communicate the intrinsic relationship between the interval-valued (IV) variables in respect to their position (center) and uncertainty (range). We stress that the actual regression examples shown are for illustration-only. They are based on comparatively small samples, and should not be taken as generalisable insight on how purchase intention of different consumer groups varies in respect to different product attributes (note the limitations set out in the footnote of Section IV)[3].

For example, we discuss how the *IRG*s show how younger individuals value taste and visual appeal more than older people in buying snack foods, and how female consumers emphasise ethical standards and health more than male consumers. Crucially, we demonstrate how the *IRG*s capture insights uniquely identifiable through IV data, such as that female consumers' uncertainty in respect to their purchase intention grows with growing uncertainty in ethical standards

[3]We note that for this very reason, we purposely do not provide statistical summary information on the quality of fit of the individual models. First, the generation of visualizations of IV regression models is the focus of this paper. Second, the actual notion of traditional measures such as R-squared which used in some cases [23], is actually not trivial for IV data. For example, should *quality of fit* focus on how well the model approximates interval size, or, interval position, or a given combination of both? Depending on the conjunctive or disjunctive nature of the intervals, the answer to this question will drive the relevant *quality of fit*, such as 'R-squared' measure.

or healthiness/nutritional aspects (e.g., calorie, sugar intake) of snack foods.

To emphasise, the individual insights from the experiments, the examples highlight the powerful capacity for IV data to effectively and efficiently provide insights which are not similarly accessible for numeric data. In other words, for comparable effort and cost [4], IV data can provide deeper insight in applications ranging from marketing, to medicine and management, all the way to cyber-security.

In turn, this underlines both the potential, and need for more research in the modelling and reasoning with these data using statistical and computational intelligence techniques. In future work, we will explore more complex cases, with larger IV data sets and multiple regressors—as part of real-world deployments. Further, we are actively working on developing novel approaches to deriving models such as fuzzy sets from IV data and developing the appropriate inference techniques which provide the capacity to both identify and communicate the rich insights in these data to decision makers.

REFERENCES

[1] 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 Systems*, vol. 23, no. 2, pp. 248–269, 2014.

[2] Y. Ren, Y.-H. Liu, J. Rong, and R. Dew, "Clustering interval-valued data using an overlapped interval divergence," in *Proc. 8th Australasian Data Mining Conf.* Melbourne, Australia: Australian Computer Society, Inc., 2009, pp. 35–42.

[3] A. Blanco-Fernández, N. Corral, and G. González-Rodríguez, "Estimation of a flexible simple linear model for interval data based on set arithmetic," *Computational Statistics & Data Analysis*, vol. 55, no. 9, pp. 2568–2578, 2011.

[4] Z. Ellerby, C. Wagner, and S. B. Broomell, "Capturing richer information: On establishing the validity of an interval-valued survey response mode," *Behavior Research Methods*, pp. 1–23, 2021.

[5] P. Hao and J. Guo, "Constrained center and range joint model for interval-valued symbolic data regression," *Computational Statistics & Data Analysis*, vol. 116, pp. 106–138, 2017.

[6] S. Kabir and C. Wagner, "Interval-valued regression-sensitivity to data set features," in *Proc. IEEE Int. Conf. Fuzzy Systems*, 2021, pp. 1–7.

[7] 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 Systems*, 2020, pp. 1–8.

[8] L. Billard and E. Diday, "Regression analysis for interval-valued data," in *Data Analysis, Classification, and Related Methods*. Springer, 2000, pp. 369–374.

[9] ——, "Symbolic regression analysis," in *Classification, Clustering, and Data Analysis*. Springer, 2002, pp. 281–288.

[10] 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," *Computational Statistics & Data Analysis*, vol. 52, no. 3, pp. 1500–1515, 2008.

[11] ——, "Constrained linear regression models for symbolic interval-valued variables," *Computational Statistics & Data Analysis*, vol. 54, no. 2, pp. 333–347, 2010.

[12] H. Wang, R. Guan, and J. Wu, "Linear regression of interval-valued data based on complete information in hypercubes," *J. Systems Science and Systems Engineering*, vol. 21, no. 4, pp. 422–442, 2012.

[13] Y. Sun and D. Ralescu, "A linear model for interval-valued data," *arXiv preprint arXiv:1506.03541*, 2015.

[14] L. C. Souza, R. M. Souza, G. J. Amaral, and T. M. Silva Filho, "A parametrized approach for linear regression of interval data," *Knowledge-Based Systems*, vol. 131, pp. 149–159, 2017.

[15] S. Kabir, C. Wagner, and Z. Ellerby, "Towards handling uncertainty-at-source in ai - a review and next steps for interval regression," *IEEE Transaction on Artificial Intelligence (under review)*, 2022.

[16] L. Wang, P. P. Wong, and E. A. Narayanan, "The demographic impact of consumer green purchase intention toward green hotel selection in china," *Tourism and Hospitality Research*, vol. 20, no. 2, pp. 210–222, 2020.

[17] C. D'Souza, M. Taghian, and R. Khosla, "Examination of environmental beliefs and its impact on the influence of price, quality and demographic characteristics with respect to green purchase intention," *J. targeting, measurement & analysis for marketing*, vol. 15, no. 2, pp. 69–78, 2007.

[18] G. González-Rodríguez, Á. Blanco, N. Corral, and A. Colubi, "Least squares estimation of linear regression models for convex compact random sets," *Advances in Data Analysis and Classification*, vol. 1, no. 1, pp. 67–81, 2007.

[19] C. L. Lawson and R. J. Hanson, *Solving least squares problems*. Prentice-Hall, Newyork, 1974.

[20] R. E. Moore, *Interval analysis*. Prentice-Hall Englewood Cliffs, 1966, vol. 4.

[21] G. E. Box and D. R. Cox, "An analysis of transformations," *J. the Royal Statistical Society: Series B (Methodological)*, vol. 26, no. 2, pp. 211–243, 1964.

[22] Z. Ellerby, J. McCulloch, J. Young, and C. Wagner, "Decsys–discrete and ellipse-based response capture system," in *Proc. IEEE Int. Conf. Fuzzy Systems*, 2019, pp. 1–6.

[23] R. Boukezzoula, S. Galichet, and A. Bisserier, "A midpoint–radius approach to regression with interval data," *Int. J. Approximate Reasoning*, vol. 52, no. 9, pp. 1257–1271, 2011.