International Journal of Statistics and Probability; Vol. 9, No. 2; March 2020
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Efficient Estimation of Interval-Valued Symbolic Data Regression Model

Chuanhua Wei[1], Nana Zheng[1] & Ke Tian[1]

[1] School of Science, Minzu University of China, Beijing 100081, China

Correspondence: Chuanhua Wei, School of Science, Minzu University of China, Beijing 100081, P.R.China.
E-mail: chweisd@163.com

Received: January 24, 2020 Accepted: February 10, 2020 Online Published: February 20, 2020

doi:10.5539/ijsp.v9n2p1 URL: https://doi.org/10.5539/ijsp.v9n2p1

Abstract

In the last two decades, regression analysis with interval-valued type data has received more and more attention. For the interval-valued symbolic data regression, the Minmax method and the center and range (CR) method are two widely used popular estimating approaches. In this paper, to improve the estimating efficiency of these two estimating methods, seemingly unrelated regression approach has been applied to take account of the dependence information of two regression models of the Minmax method or the CR method. Finally, real data sets are analysed to examine the performance of our proposed procedure.

Keywords: interval-valued data, minmax method, center and range method, seemingly unrelated regression, generalized least squares estimation

1. Introduction

Interval-valued data as a more general class of data type called symbolic data are observed as ranges instead of single values and frequently appear in some fields, such as finance, engineering, and medicine. In the last two decades, regression analysis with interval-valued type data has received more and more attention. Several approaches have been proposed to estimate the regression models with interval-valued data. The center method of Billard and Diday (2000) uses interval midpoints of both response and the associated explanatory variables to build the regression, and apply the fitted model to the lower and upper bounds of the independent variables to generate predictions respectively. In order to use more interval information than that of the center method, Billard and Diday (2002) proposed a MinMax method, which suggests modelling the lower and upper bounds of the intervals of both response and the associated explanatory variables by two linear regression models independently. Using two different models to predict lower and upper bounds can improve the linear fit and give an intuitive response interpretation. To include the information given by both the centre and the range of an interval on a linear regression model to improve the model prediction performance, Lima Neto and De Carvalho (2008) proposed a center and range (CR) method using two independent models: one for the interval midpoints and another for the semi-length of the interval. Other estimating approaches can be found in Xu (2010), Giordani (2015) and Souza *et al.* (2017). Furthermore, some generalized interval data models have been proposed, examples including additive models of Lim (2016), partially linear models of Wei *et al.* (2015).

It is noted that both the MinMax method and CR method analyze the interval data based on two independent linear regression models, and the assumption of independence between two regression models is not always true in practice. If these two regression models have some relationship, then, to solve this problem, how to combine the two regression models is an interesting topic. As we all know, the seemingly unrelated regression (SUR) introduced by Zellner (1962) is an important tool to analyze multiple equations with correlated disturbances. The SUR specification is expressed as a set of linear regressions where the disturbances in the different equations are correlated. Therefore, to take account of the dependence information of two regression models, we propose a seemingly unrelated regression approach based on the MinMax method or the CR method to modelling interval data.

The rest of this paper is organized as follows. We introduce the MinMax and CRM methods in Section 2. The proposed SUR approach is given in Section 3. Real interval-valued data sets are analyzed in Section 4 to illustrate the performance of the proposed approach. Conclusion is presented in Section 5.

2. Minmax and CR Methods for Linear Models With Interval-Valued Data

Let $\mathbf{E} = \{e_1, e_2, \cdots, e_n\}$ be a set of objects that are described by the $p{+}1$ symbolic interval-valued variables $Y, X_1, X_2, \cdots, X_p$. Each example $e_i \in E(i = 1, 2, \cdots, n)$ is represented as an interval quantitative feature vector $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i)$. $\mathbf{x}_i =$

$(x_{i1}, x_{i2}, \cdots, x_{ip})^{\mathrm{T}}$, where $x_{ij} = [a_{ij}, b_{ij}] \in \Im = \{[a, b] : a, b \in \Re, a \leq b\} (j = 1, 2, \cdots, p)$ and $y_i = [y_{Li}, y_{Ui}] \in \Im$ are the ith observed values of X_j and Y , respectively. Let us work with the matrix notation. Denote

$$\mathbf{Y}_L = \begin{bmatrix} y_{L1} \\ y_{L2} \\ \vdots \\ y_{Ln} \end{bmatrix}, \mathbf{Y}_U = \begin{bmatrix} y_{U1} \\ y_{U2} \\ \vdots \\ y_{Un} \end{bmatrix}, \mathbf{X}_L = \begin{bmatrix} 1 & a_{11} & \cdots & a_{1p} \\ 1 & a_{21} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & a_{n1} & \cdots & a_{np} \end{bmatrix}, \mathbf{X}_U = \begin{bmatrix} 1 & b_{11} & \cdots & b_{1p} \\ 1 & b_{21} & \cdots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & b_{n1} & \cdots & b_{np} \end{bmatrix},$$

2.1 MinMax Method

By Billard and Diday (2002), we consider $X_1, X_2, \cdots, X_p$ related to Y according to the following linear regression relationship

$$y_{Li} = \beta_0^L + \beta_1^L a_{i1} + \beta_2^L a_{i2} + \cdots + \beta_p^L a_{ip} + \varepsilon_{Li}, \tag{1}$$

$$y_{Ui} = \beta_0^U + \beta_1^U b_{i1} + \beta_2^U b_{i2} + \cdots + \beta_p^U b_{ip} + \varepsilon_{Ui}. \tag{2}$$

Model (1) (2) can be written in the matrix form as

$$\mathbf{Y}_m = \mathbf{X}_m \boldsymbol{\beta}_m + \boldsymbol{\varepsilon}_m, \tag{3}$$

where

$$\mathbf{Y}_m = \begin{bmatrix} \mathbf{Y}_L \\ \mathbf{Y}_U \end{bmatrix}, \mathbf{X}_m = \begin{bmatrix} \mathbf{X}_L & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_U \end{bmatrix}, \boldsymbol{\beta}_m = \begin{bmatrix} \boldsymbol{\beta}^L \\ \boldsymbol{\beta}^U \end{bmatrix}, \boldsymbol{\varepsilon}_m = \begin{bmatrix} \boldsymbol{\varepsilon}_L \\ \boldsymbol{\varepsilon}_U \end{bmatrix},$$

and $\boldsymbol{\beta}^L = \left(\beta_0^L, \beta_1^L, \cdots, \beta_p^L\right)^{\mathrm{T}}, \boldsymbol{\beta}^U = \left(\beta_0^U, \beta_1^U, \cdots, \beta_p^U\right)^{\mathrm{T}}, \boldsymbol{\varepsilon}_L = (\varepsilon_{L1}, \cdots, \varepsilon_{Ln})^{\mathrm{T}}, \boldsymbol{\varepsilon}_U = (\varepsilon_{U1}, \cdots, \varepsilon_{Un})^{\mathrm{T}}$.

By applying the least squares approach to model (4), we can get the Minmax estimator of $\boldsymbol{\beta}_{\mathrm{m}}$ as

$$\hat{\boldsymbol{\beta}}^{mm} = \begin{bmatrix} \hat{\boldsymbol{\beta}}^L \\ \hat{\boldsymbol{\beta}}^U \end{bmatrix} = \left[\mathbf{X}_m^{\mathrm{T}} \mathbf{X}_m\right]^{-1} \mathbf{X}_m^{\mathrm{T}} \mathbf{Y}_m = \begin{bmatrix} \left[\mathbf{X}_L^{\mathrm{T}} \mathbf{X}_L\right]^{-1} \mathbf{X}_L^{\mathrm{T}} \mathbf{Y}_L \\ \left[\mathbf{X}_U^{\mathrm{T}} \mathbf{X}_U\right]^{-1} \mathbf{X}_U^{\mathrm{T}} \mathbf{Y}_U \end{bmatrix}. \tag{4}$$

Specifically, the MinMax estimators of $\boldsymbol{\beta}_L$ and $\hat{\boldsymbol{\beta}}_U$ are

$$\hat{\boldsymbol{\beta}}^L = \left[\mathbf{X}_L^{\mathrm{T}} \mathbf{X}_L\right]^{-1} \mathbf{X}_L^{\mathrm{T}} \mathbf{Y}_L, \quad \hat{\boldsymbol{\beta}}^U = \left[\mathbf{X}_U^{\mathrm{T}} \mathbf{X}_U\right]^{-1} \mathbf{X}_U^{\mathrm{T}} \mathbf{Y}_U.$$

2.2 The Centre and Range Method

Let $y_i^c = (y_{Li} + y_{Ui})/2, y_i^r = (y_{Ui} - y_{Li})/2,\ x_{ij}^c = (a_{ij} + b_{ij})/2, x_{ij}^r = (b_{ij} - a_{ij})/2.$ Lima Neto and De Carvalho (2008) considered the following two regression models

$$y_i^c = \beta_0^c + \beta_1^c x_{i1}^c + \beta_2^c x_{i2}^c + \cdots + \beta_p^c x_{ip}^c + \varepsilon_i^c \tag{5}$$

$$y_i^r = \beta_0^r + \beta_1^r x_{i1}^r + \beta_2^r x_{i2}^r + \cdots + \beta_p^r x_{ip}^r + \varepsilon_i^r \tag{6}$$

Denote

$$\mathbf{Y}^c = \begin{bmatrix} y_1^c \\ y_2^c \\ \vdots \\ y_n^c \end{bmatrix}, \mathbf{Y}^r = \begin{bmatrix} y_1^r \\ y_2^r \\ \vdots \\ y_n^r \end{bmatrix}, \mathbf{X}^c = \begin{bmatrix} 1 & x_{11}^c & \cdots & x_{1p}^c \\ 1 & x_{21}^c & \cdots & x_{2p}^c \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1}^c & \cdots & x_{np}^c \end{bmatrix}, \mathbf{X}^r = \begin{bmatrix} 1 & x_{11}^r & \cdots & x_{1p}^r \\ 1 & x_{21}^r & \cdots & x_{2p}^r \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1}^r & \cdots & x_{np}^r \end{bmatrix},$$

Combing models (5) and (6), we have the following model

$$\mathbf{Y}_{cr} = \mathbf{X}_{cr} \boldsymbol{\beta}_{cr} + \boldsymbol{\varepsilon}_{cr}, \tag{7}$$

where

$$\mathbf{Y}_{cr} = \begin{bmatrix} \mathbf{Y}^c \\ \mathbf{Y}^r \end{bmatrix}, \mathbf{X}_{cr} = \begin{bmatrix} \mathbf{X}^c & \mathbf{0} \\ \mathbf{0} & \mathbf{X}^r \end{bmatrix}, \boldsymbol{\beta}_{cr} = \begin{bmatrix} \boldsymbol{\beta}^c \\ \boldsymbol{\beta}^r \end{bmatrix}, \boldsymbol{\varepsilon}_{cr} = \begin{bmatrix} \boldsymbol{\varepsilon}^c \\ \boldsymbol{\varepsilon}^r \end{bmatrix}.$$

and $\mathbf{Y}^{\mathrm{c}} = \frac{\mathbf{Y}_L + \mathbf{Y}_U}{2}, \mathbf{X}^{\mathrm{c}} = \frac{\mathbf{X}_L + \mathbf{X}_U}{2}, \mathbf{Y}^{\mathrm{r}} = \frac{\mathbf{Y}_L - \mathbf{Y}_U}{2}, \mathbf{X}^{\mathrm{r}} = \frac{\mathbf{X}_L - \mathbf{X}_U}{2}, \boldsymbol{\beta}^c = (\beta_0^c, \beta_1^c, \cdots, \beta_p^c)^{\mathrm{T}}, \boldsymbol{\beta}^r = (\beta_0^r, \beta_1^r, \cdots, \beta_p^r)^{\mathrm{T}}, \boldsymbol{\varepsilon}^c = (\varepsilon_1^c, \varepsilon_2^c, \cdots, \varepsilon_n^c)^{\mathrm{T}}$, and $\boldsymbol{\varepsilon}^r = (\varepsilon_1^r, \varepsilon_2^r, \cdots, \varepsilon_n^r)^{\mathrm{T}}$.

Then, by applying the least squares approach to model (7), we can obtain the CR estimator for $\boldsymbol{\beta}_{cr}$ as

$$\hat{\boldsymbol{\beta}}^{cr} = \left[\begin{array}{c} \hat{\boldsymbol{\beta}}^{c} \\ \hat{\boldsymbol{\beta}}^{r} \end{array} \right] = \left[\mathbf{X}_{cr}^{\mathrm{T}}\mathbf{X}_{cr}\right]^{-1}\mathbf{X}_{cr}^{\mathrm{T}}\mathbf{Y}_{cr} = \left[\begin{array}{c} \left[(\mathbf{X}^{c})^{\mathrm{T}}\mathbf{X}^{c}\right]^{-1}(\mathbf{X}^{c})^{\mathrm{T}}\mathbf{Y}^{c} \\ \left[(\mathbf{X}^{r})^{\mathrm{T}}\mathbf{X}^{r}\right]^{-1}(\mathbf{X}^{r})^{\mathrm{T}}\mathbf{Y}^{r} \end{array} \right]. \tag{8}$$

Specifically, the Centre-Range estimators of $\boldsymbol{\beta}^{c}$ and $\hat{\boldsymbol{\beta}}^{r}$ can be written as

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{c} &= \left[(\mathbf{X}^{c})^{\mathrm{T}}\mathbf{X}^{c}\right]^{-1}(\mathbf{X}^{c})^{\mathrm{T}}\mathbf{y}^{c} = \left[(\mathbf{X}_U+\mathbf{X}_L)^{\mathrm{T}}(\mathbf{X}_U+\mathbf{X}_L)\right]^{-1}(\mathbf{X}_U+\mathbf{X}_L)^{\mathrm{T}}(\mathbf{Y}_U+\mathbf{Y}_L) \\ &= \left(\mathbf{X}_U^{\mathrm{T}}\mathbf{X}_U+\mathbf{X}_L^{\mathrm{T}}\mathbf{X}_L+\mathbf{X}_L^{\mathrm{T}}\mathbf{X}_U+\mathbf{X}_U^{\mathrm{T}}\mathbf{X}_L\right)^{-1}\left(\mathbf{X}_U^{\mathrm{T}}\mathbf{Y}_U+\mathbf{X}_L^{\mathrm{T}}\mathbf{Y}_L+\mathbf{X}_L^{\mathrm{T}}\mathbf{Y}_U+\mathbf{X}_U^{\mathrm{T}}\mathbf{Y}_L\right), \end{aligned}$$

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{r} &= \left[(\mathbf{X}^{r})^{\mathrm{T}}\mathbf{X}^{r}\right]^{-1}(\mathbf{X}^{r})^{\mathrm{T}}\mathbf{y}^{r} = \left[(\mathbf{X}_U-\mathbf{X}_L)^{\mathrm{T}}(\mathbf{X}_U-\mathbf{X}_L)\right]^{-1}(\mathbf{X}_U-\mathbf{X}_L)^{\mathrm{T}}(\mathbf{Y}_U-\mathbf{Y}_L) \\ &= \left(\mathbf{X}_U^{\mathrm{T}}\mathbf{X}_U+\mathbf{X}_L^{\mathrm{T}}\mathbf{X}_L-\mathbf{X}_L^{\mathrm{T}}\mathbf{X}_U-\mathbf{X}_U^{\mathrm{T}}\mathbf{X}_L\right)^{-1}\left(\mathbf{X}_U^{\mathrm{T}}\mathbf{Y}_U+\mathbf{X}_L^{\mathrm{T}}\mathbf{Y}_L-\mathbf{X}_L^{\mathrm{T}}\mathbf{Y}_U-\mathbf{X}_U^{\mathrm{T}}\mathbf{Y}_L\right). \end{aligned}$$

3. The Efficient SUR-Based Estimation

For both the Minmax and CR method, we can improve the estimating efficiency by considering the dependence between their two models. We propose the SUR-based Minmax estimator by applying the SUR method to models (1) and (2) directly. Similarly, we propose the SUR-based CR estimator by applying the SUR method to models (5) and (6) directly. It is noted that models (5) and (6) are derived by models (1) and (2), then we construct a two-step SUR-based CR estimator based on the relation between models (1)-(2) and models (5)-(6).

3.1 The Efficient SUR-Based Minmax Estimator

For models (1)-(2), we assume that

$$\mathrm{E}(\varepsilon_{Li}) = 0, \mathrm{D}(\varepsilon_{Li}) = \sigma_{LL}, \mathrm{E}(\varepsilon_{Ui}) = 0, \mathrm{D}(\varepsilon_{Ui}) = \sigma_{UU}, \mathrm{Cov}(\varepsilon_{Li}, \varepsilon_{Lj}) = 0, \mathrm{Cov}(\varepsilon_{Ui}, \varepsilon_{Uj}) = 0, i \neq j.$$

Then, we have

$$\mathrm{Cov}(\boldsymbol{\varepsilon}_L) = E\boldsymbol{\varepsilon}_L\boldsymbol{\varepsilon}_L^{\mathrm{T}} = \sigma_{LL}\mathbf{I}_n, \mathrm{Cov}(\boldsymbol{\varepsilon}_U) = E\boldsymbol{\varepsilon}_U\boldsymbol{\varepsilon}_U^{\mathrm{T}} = \sigma_{UU}\mathbf{I}_n, \tag{9}$$

Different to Billard and Diday (2002) and Lima Neto and De Carvalho (2008), for the models (1)(2), we consider the following assumptions of their disturbances

$$E(\varepsilon_{Li}\varepsilon_{Uj}) = \begin{cases} \sigma_{LU}, & i = j, \\ 0, & \text{otherwise,} \end{cases}$$

for $1 \leq i, j \leq n$. Then, we have

$$\mathrm{Cov}(\boldsymbol{\varepsilon}_L, \boldsymbol{\varepsilon}_U) = E\boldsymbol{\varepsilon}_L\boldsymbol{\varepsilon}_U^{\mathrm{T}} = \sigma_{LU}\mathbf{I}_n, \tag{10}$$

with $\mathbf{I}_n$ is the identity matrix of order n. Therefore, the $2n \times 1$ disturbance vector $\boldsymbol{\varepsilon}_{\mathrm{m}} = (\boldsymbol{\varepsilon}_L^{\mathrm{T}}, \boldsymbol{\varepsilon}_U^{\mathrm{T}})^{\mathrm{T}}$ has the following variance-covariance matrix

$$\boldsymbol{\Omega}_{\mathrm{m}} = E(\boldsymbol{\varepsilon}_m\boldsymbol{\varepsilon}_m^{\mathrm{T}}) = \boldsymbol{\Sigma}_{LU} \otimes \mathbf{I}_n, \tag{11}$$

with

$$\boldsymbol{\Sigma}_{LU} = \left[\begin{array}{cc} \sigma_{LL} & \sigma_{LU} \\ \sigma_{LU} & \sigma_{UU} \end{array} \right].$$

By applying the generalized least squares estimation method for linear regression model (3) with (9), we can obtain the seeming unrelated regression-based Minmax estimator of $\boldsymbol{\beta}_{\mathrm{m}}$ as

$$\hat{\boldsymbol{\beta}}_{SUR}^{mm} = \left[\begin{array}{c} \hat{\boldsymbol{\beta}}_{SUR}^{L} \\ \hat{\boldsymbol{\beta}}_{SUR}^{U} \end{array} \right] = \left[\mathbf{X}_m^{\mathrm{T}}\boldsymbol{\Omega}_m^{-1}\mathbf{X}_m\right]^{-1}\mathbf{X}_m^{\mathrm{T}}\boldsymbol{\Omega}_m^{-1}\mathbf{Y}_m. \tag{12}$$

However, the covariance matrix $\boldsymbol{\Omega}_m$ is unknown, then $\hat{\boldsymbol{\beta}}_{SUR}^{mm}$ is infeasible. To solve this problem, we can replace the unknown elements $\boldsymbol{\Omega}_{\mathrm{m}}$ by their estimators respectively. Define

$$\hat{\sigma}_{LL} = \frac{(\mathbf{Y}_L - \mathbf{X}_L\hat{\boldsymbol{\beta}}^{L})^{\mathrm{T}}(\mathbf{Y}_L - \mathbf{X}_L\hat{\boldsymbol{\beta}}^{L})}{n-p}, \hat{\sigma}_{UU} = \frac{(\mathbf{Y}_U - \mathbf{X}_U\hat{\boldsymbol{\beta}}^{U})^{\mathrm{T}}(\mathbf{Y}_U - \mathbf{X}_U\hat{\boldsymbol{\beta}}^{U})}{n-p},$$

and

$$\hat{\sigma}_{LU} = \frac{(\mathbf{Y}_L - \mathbf{X}_L\hat{\boldsymbol{\beta}}^L)^{\mathrm{T}}(\mathbf{Y}_U - \mathbf{X}_U\hat{\boldsymbol{\beta}}^U)}{n},$$

where $\hat{\boldsymbol{\beta}}^L$ and $\hat{\boldsymbol{\beta}}^U$ are the Minmax estimators which were defined in equation (4). Then we can define the feasible SUR-Minmax estimator for $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}}^{mm}_{FSUR} = \begin{bmatrix} \hat{\boldsymbol{\beta}}^L_{FSUR} \\ \hat{\boldsymbol{\beta}}^U_{FSUR} \end{bmatrix} = \left[\mathbf{X}_m^{\mathrm{T}}\hat{\boldsymbol{\Omega}}_m^{-1}\mathbf{X}_m\right]^{-1}\mathbf{X}_m^{\mathrm{T}}\hat{\boldsymbol{\Omega}}_m^{-1}\mathbf{Y}_m. \tag{13}$$

with $\hat{\boldsymbol{\Omega}}_m = \hat{\boldsymbol{\Sigma}}_{LU} \otimes \mathbf{I}_n$,

$$\hat{\boldsymbol{\Sigma}}_{LU} = \begin{bmatrix} \hat{\sigma}_{LL} & \hat{\sigma}_{LU} \\ \hat{\sigma}_{LU} & \hat{\sigma}_{UU} \end{bmatrix}.$$

By the classic theory of linear regression model, we know that the generalized least-squares estimator $\hat{\boldsymbol{\beta}}^{mm}_{FSUR}$ is efficient than the ordinary least-squares estimator $\hat{\boldsymbol{\beta}}^{mm}$. If $\sigma_{LU} = 0$, they are equal. In practice, We can first to test whether $\sigma_{LU} = 0$ is or not. we can use the Lagrange multiplier test method of Breusch and Pagan (1980) to this testing problem.

3.2 The Efficient SUR-Based CR Estimator

Similarly, for the linear regression models (5) and (6), the feasible SUR-based CR estimator for $\boldsymbol{\beta}^{cr}$ can be defined as

$$\hat{\boldsymbol{\beta}}^{cr}_{FSUR} = \begin{bmatrix} \hat{\boldsymbol{\beta}}^c_{FSUR} \\ \hat{\boldsymbol{\beta}}^r_{FSUR} \end{bmatrix} = \left[\mathbf{X}_{cr}^{\mathrm{T}}\hat{\boldsymbol{\Omega}}_{cr}^{-1}\mathbf{X}_m\right]^{-1}\mathbf{X}_{cr}^{\mathrm{T}}\hat{\boldsymbol{\Omega}}_{cr}^{-1}\mathbf{Y}_{cr}. \tag{14}$$

with $\hat{\boldsymbol{\Omega}}_{cr} = \hat{\boldsymbol{\Sigma}}_{cr} \otimes \mathbf{I}_n$, $\quad \hat{\boldsymbol{\Sigma}}_{cr} = \begin{bmatrix} \hat{\sigma}_{cc} & \hat{\sigma}_{cr} \\ \hat{\sigma}_{cr} & \hat{\sigma}_{rr} \end{bmatrix}, \hat{\sigma}_{cr} = \frac{(\mathbf{Y}^c - \mathbf{X}^c\hat{\boldsymbol{\beta}}^c)^{\mathrm{T}}(\mathbf{Y}^r - \mathbf{X}^r\hat{\boldsymbol{\beta}}^r)}{n}$, and

$$\hat{\sigma}_{cc} = \frac{(\mathbf{Y}^c - \mathbf{X}^c\hat{\boldsymbol{\beta}}^c)^{\mathrm{T}}(\mathbf{Y}^c - \mathbf{X}^c\hat{\boldsymbol{\beta}}^c)}{n-p}, \hat{\sigma}_{rr} = \frac{(\mathbf{Y}^r - \mathbf{X}^r\hat{\boldsymbol{\beta}}^r)^{\mathrm{T}}(\mathbf{Y}^r - \mathbf{X}_U\hat{\boldsymbol{\beta}}^r)}{n-p},$$

where $\hat{\boldsymbol{\beta}}^c$ and $\hat{\boldsymbol{\beta}}^r$ are the CR estimators which were defined in equation (8).

3.3 The Two-Step SUR-Based CR Estimation

The direct SUR-based CR estimatior (13) is obtained on the assumption that model (5)-(6) are data generating models. However, it is not ture. Model (5)-(6) are generated from model (1)-(2). Then, the variance and covariance of $\boldsymbol{\varepsilon}^c$ and $\boldsymbol{\varepsilon}^r$ can be computed by the assumptions (9) and (10). We can show that

$$\begin{aligned} \boldsymbol{\Phi}_c = \mathrm{Cov}(\boldsymbol{\varepsilon}_c) &= \mathrm{Cov}\left(\tfrac{\boldsymbol{\varepsilon}_U + \boldsymbol{\varepsilon}_L}{2}\right) = \tfrac{1}{4}\mathrm{E}\left[(\boldsymbol{\varepsilon}_U + \boldsymbol{\varepsilon}_L)(\boldsymbol{\varepsilon}_U + \boldsymbol{\varepsilon}_L)^{\mathrm{T}}\right] \\ &= \tfrac{1}{4}\mathrm{E}(\boldsymbol{\varepsilon}_U\boldsymbol{\varepsilon}_U^{\mathrm{T}} + \boldsymbol{\varepsilon}_L\boldsymbol{\varepsilon}_L^{\mathrm{T}} + 2\boldsymbol{\varepsilon}_U\boldsymbol{\varepsilon}_L^{\mathrm{T}}) \\ &= \tfrac{\sigma_{UU}+\sigma_{LL}+2\sigma_{UL}}{4}\mathrm{I}_n, \end{aligned}$$

$$\boldsymbol{\Phi}_r = \mathrm{Cov}(\boldsymbol{\varepsilon}_r) = \mathrm{Cov}\left(\frac{\boldsymbol{\varepsilon}_U - \boldsymbol{\varepsilon}_L}{2}\right) = \frac{\sigma_{UU} + \sigma_{LL} - 2\sigma_{UL}}{4}\mathrm{I}_n,$$

$$\begin{aligned} \boldsymbol{\Phi}_{cr} = \mathrm{Cov}(\boldsymbol{\varepsilon}_c, \boldsymbol{\varepsilon}_r) &= \tfrac{1}{4}\mathrm{E}\left[(\boldsymbol{\varepsilon}_U + \boldsymbol{\varepsilon}_L)(\boldsymbol{\varepsilon}_U - \boldsymbol{\varepsilon}_L)^{\mathrm{T}}\right] \\ &= \tfrac{1}{4}\mathrm{E}(\boldsymbol{\varepsilon}_U\boldsymbol{\varepsilon}_U^{\mathrm{T}} - \boldsymbol{\varepsilon}_L\boldsymbol{\varepsilon}_L^{\mathrm{T}}) \\ &= \tfrac{\sigma_{UU}-\sigma_{LL}}{4}\mathrm{I}_n. \end{aligned}$$

Therefore, the $2n \times 1$ disturbance vector $\boldsymbol{\varepsilon}_{cr} = (\boldsymbol{\varepsilon}_c^{\mathrm{T}}, \boldsymbol{\varepsilon}_r^{\mathrm{T}})^{\mathrm{T}}$ has the following variance-covariance matrix

$$\boldsymbol{\Lambda}_{cr} = \mathrm{Cov}(\boldsymbol{\varepsilon}_{cr}) = \begin{bmatrix} \boldsymbol{\Phi}_c & \boldsymbol{\Phi}_{cr} \\ \boldsymbol{\Phi}_{cr} & \boldsymbol{\Phi}_r \end{bmatrix}.$$

Then, the two-step efficient SUR-based CR estimators are

$$\hat{\boldsymbol{\beta}}^{\mathrm{cr}}_{TSUR} = \begin{bmatrix} \tilde{\boldsymbol{\beta}}^c_{TSUR} \\ \tilde{\boldsymbol{\beta}}^r_{TSUR} \end{bmatrix} = \left[\mathbf{X}_{cr}^{\mathrm{T}}\boldsymbol{\Lambda}_{cr}^{-1}\mathbf{X}_{cr}\right]^{-1}\mathbf{X}_{cr}^{\mathrm{T}}\boldsymbol{\Lambda}_{cr}^{-1}\mathbf{Y}_{cr}. \tag{15}$$

Table 1. Mushroom interval-valued data set

Species	Y	X	Z	Species	Y	X	Z
1	[3,8]	[4,9]	[0.5,2.5]	13	[3.5,8]	[4,10]	[1,2]
2	[6,21]	[4,14]	[1,3.5]	14	[7,14]	[8,14]	[1.5,2.5]
3	[4,8]	[5,11]	[1,2]	15	[8,20]	[9,19]	[3,5]
4	[6,7]	[4,7]	[3,4.5]	16	[2.5,4]	[2.5,4.5]	[0.4,0.7]
5	[5,12]	[2,5]	[1.5,2.5]	17	[7,19]	[8,15]	[2,3.5]
6	[5,15]	[4,10]	[2,4]	18	[5,15]	[6,15]	[2.5,3.5]
7	[4,11]	[3,7]	[0.4,1]	19	[8,12]	[6,12]	[1.5,2]
8	[5,10]	[3,6]	[1,2]	20	[2,6]	[3,7]	[0.4,0.8]
9	[2.5,4]	[3,5]	[0.4,0.7]	21	[6,12]	[6,12]	[1.5,2]
10	[2.5,6]	[1.5,3.5]	[1,1.5]	22	[6,12]	[6,16]	[1,2]
11	[1.5,2.5]	[3,6]	[0.25,0.35]	23	[5,17]	[4,14]	[1,3.5]
12	[4,15]	[4,15]	[1.5,2.5]				

Table 2. Performance of the methods

Method	RMSE$_\mathrm{L}$	RMSE$_\mathrm{U}$	r_L^2	r_U^2
Minmax	1.131198	3.241086	0.6241114	0.6146621
SUR-Minmax	1.132775	3.169804	0.6308301	0.6275454
CR	1.479281	3.003173	0.5105622	0.6631535
SUR-CR	1.351316	2.898771	0.5473502	0.6813018
Two-step SUR-CR	1.348811	2.852429	0.5540649	0.7014462

The corresponding feasible estimator is

$$\hat{\boldsymbol{\beta}}^{cr}_{FTSUR} = \begin{bmatrix} \tilde{\boldsymbol{\beta}}^{c}_{FTSUR} \\ \tilde{\boldsymbol{\beta}}^{r}_{FTSUR} \end{bmatrix} = \left[\mathbf{X}^{\mathrm{T}}_{cr}\hat{\boldsymbol{\Lambda}}^{-1}_{cr}\mathbf{X}_{cr}\right]^{-1}\mathbf{X}^{\mathrm{T}}_{cr}\hat{\boldsymbol{\Lambda}}^{-1}_{cr}\mathbf{Y}_{cr}, \tag{16}$$

where $\hat{\boldsymbol{\Lambda}}$ is defined as $\boldsymbol{\Lambda}$ by replacing unknown parameters $\sigma_{UU}, \sigma_{UU}, \sigma_{UL}$ by their estimators, respectively.

It is noted that if $\sigma_{UL} = 0$, model (1) and (2) are independence, we have $\hat{\boldsymbol{\beta}}^{mm}_{SUR} = \hat{\boldsymbol{\beta}}^{mm}$. If $\sigma_{UU} \neq \sigma_{LL}$, model (5) and (6) are not independence as $\boldsymbol{\Phi}_{cr} = \frac{\sigma_{UU}-\sigma_{LL}}{4}\mathrm{I}_n \neq \mathbf{0}_n$, then $\tilde{\boldsymbol{\beta}}^{cr}_{TSUR} \neq \hat{\boldsymbol{\beta}}^{cr}$. If $\sigma_{UU} = \sigma_{LL}$, then $\boldsymbol{\Phi}_{cr} = \frac{\sigma_{UU}-\sigma_{LL}}{4}\mathrm{I}_n = \mathbf{0}_n$, we have $\hat{\boldsymbol{\beta}}^{cr}_{TSUR} = \hat{\boldsymbol{\beta}}^{cr}$.

4. Real Data Analysis

In this section, we shall analyse a real data to illustrate the finite sample properties of the proposed procedures.

Mushroom data set consists of a set of 23 species described by 3 interval variables, where the response variable Y is the stipe thickness, and the covariates X is the stipe length, Z is the pileus cap width. These mushroom species are members of the genus Agaricies. The specific variables and their values are extracted from the Fungi of California Species. The data set given in Table 1 was obtained from Billard and Diday (2006).

The performance assessment of the above estimating approaches will be based on the following measures: the lower boundary root mean-square error (RMSE$_\mathrm{L}$) and the upper boundary root mean-square error (RMSE$_\mathrm{U}$), the square of the lower bound correlation coefficient (r_L^2) and the square of the upper bound correlation coefficient (r_U^2) These measures, calculated from the observed values $[\mathrm{y_{Li}}, \mathrm{y_{Ui}}]$ and their corresponding leave-one-out cross-validation predicted values based on Minmax, SUR-Minmax, CR, SUR-CR and Two step SUR-CR estimating methods. They are defined by

$$\mathrm{RMSE_L} = \sqrt{\frac{\sum_{i=1}^{n}(y_{Li} - \hat{y}_{Li})^2}{n}}, \mathrm{RMSE_U} = \sqrt{\frac{\sum_{i=1}^{n}(y_{Ui} - \hat{y}_{Ui})^2}{n}}, \mathrm{r}_\mathrm{L}^2 = \rho^2(Y_L, \hat{Y}_L), \mathrm{r}_\mathrm{U}^2 = \rho^2(Y_U, \hat{Y}_U).$$

The results can be found in Table 2. Thus, we conclude that the SUR-MinMax method outperforms the MinMax method, the two-step SUR-CR method outperforms the SUR-CR method while the SUR-CR method outperforms the CR method.

5. Conclusion

This paper applied the SUR method to improve the efficiency of Minmax method and CR method. Real data sets are analysed to examine the performance of our proposed methods and the results are satisfactory.

For the Minmax method and CR method, we have some comments. In the simulation studies of Lima Neto and De Carvalho (2008) and other literatures, the interval data sets are often constructed in two different ways, the first with values of the centre and range of the intervals simulated independently, the second with values of the interval mid-points and ranges related according to a linear relationship. However, these assumptions are usually not satisfied in real data analysis.

Conflicts of Interest

The author declares no conflict of interest.

References

Billard, L., & Diday, E. (2000). Regression analysis for interval-valued data. In Kiers, H. A. L., Rasson, J. P. (Eds.), *Data Analysis, Classification, and Related Methods, Proceedings IFCS2000, Namur* (pp.369-374). Heidelberg: Springer Verlag. https://doi.org/10.1007/978-3-642-59789-35.8

Billard, L., & Diday, E. (2002). *Symbolic regression analysis.* In Classification, Clustering and Data Analysis, Proceedings of the Eighenth Conference of the International Federation of Classification Societies (IFCS02), Springer, Poland, 281-288. https://doi.org/10.1007/978-3-642-56181-83.1

Billard, L., & Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. John Wiley, New York. https://doi.org/10.1002/9780470090183

Breusch, T. S., & Pagan, A. R. (1980). The Lagrange Multiplier Test and Its Applications to Model Specification in Econometrics. *Review of Economic Studies, 47,* 239-253. https://doi.org/10.2307/2297111

Giordani, P. (2015). Lasso-constrained regression analysis for interval-valued data. *Adv Data Anal Classif, 9*, 5-19. https://doi.org/10.1007/s11634-014-0164-8

Lim, C. (2016). Interval-valued data regression using nonparametric additive models. *Journal of the Korean Statistical Society, 45,* 358-370. https://doi.org/10.1016/j.jkss.2015.12.003

Lima Neto, E. A., & De Carvalho, F. A. T. (2008). Centre and range method to fitting a linear regression model on symbolic interval data. *Computational Statistics and Data Analysis, 52*(3), 1500-1515. https://doi.org/10.1016/j.csda.2007.04.014

Souza, L. C., & Souza, R. M. C. R., & Amaral, G. J. A., & Filho, T. M. S. (2017). A Parametrized Approach for Linear Regression of Interval Data. *Knowledge-Based Systems, 131*, 149-159. https://doi.org/10.1016/j.knosys.2017.06.012

Wei, Y., & Wang, S. S., & Wang, H. W. (2015). Interval-valued data regression using partial linear model. *Journal of Statistical Computation and Simulation, 87,* 3175-3194. https://doi.org/10.1080/00949655.2017.1360298

Xu, W. (2010). Symbolic Data Analysis: Interval-Valued Data Regression (Ph.D. thesis). University of Georgia.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association, 57*, 348-368. https://doi.org/10.1080/01621459.1962.10480664

Copyrights