Electronic Research Archive

Research article

Generalized estimation equations method for fixed effects panel interval-valued data models

Chi Liu[1], **Ruiqin Tian**[1,*] **and Dengke Xu**[2]

[1] School of Mathematics, Hangzhou Normal University, Hangzhou 311121, China
[2] School of Economics, Hangzhou Dianzi University, Hangzhou 310018, China

* **Correspondence:** Email: ruiqintian@163.com.

Abstract: This paper studied panel interval-valued data models with individual fixed effects, in which the correlation within a group was considered and the group average method was used to eliminate the fixed effects. Then, we applied generalized estimation equations (GEEs) to analyze panel interval-valued data models and gave a computational algorithm to obtain the estimators. Some Monte Carlo simulations and real data analysis showed that, in contrast with the least-squares dummy-variable (LSDV) method, the proposed GEEs method has advantages in forecasting performance.

Keywords: panel interval-valued data; generalized estimation equations; the min-max model; the center and range model; fixed effects

1. Introduction

In the real world, there are a large number of interval-valued data. For example, in the stock market, the K-line chart, as an important decision-making tool, includes the opening price, closing price, maximum, and minimum of the day. In temperature prediction, the minimum and maximum temperatures can be regarded as an interval. On the other hand, due to the complexity of the system, the data obtained is an interval rather than a real point. For example, experts often predict the economic growth rate of the next year in the form of an interval, such as 4–5%. As far as we know, interval-valued data can provide a tool for dealing with large datasets, which were proposed by Diday [1]. It not only contains richer information than point data but also has better interpretability. So, it is meaningful to analyze interval-valued data. Billard and Diday [2] proposed the first algorithm to fit the interval-valued linear regression, which fits the linear regression model to the midpoint of the interval values and obtains the parameters by minimizing the midpoint error. Billard and Diday [3] proposed the min-max method, which defined two models, each corresponding to a response bound. The lower bounds of the response depend on the lower bounds of the regressive variables, and the upper bounds of response depend

on the upper bounds of the regressive variables. Lima Neto et al. [4] introduced the center and range Method (CRM) and proposed two linear models: one for the midpoint and another for the range. Kong and Gao [5] proposed a regularized interval MM estimate (RIMME) for interval-valued regression, which can achieve a good balance between the prediction accuracy and mathematical coherence of the predicted intervals. Xu and Qin [6] proposed a bivariate Bayesian regression model based on CRM with known or unknown covariance matrices. Beyaztas et al. [7] introduced the functional forms of some well-known regression models that take interval-valued data. Interval analysis technologies can also be found in other literature [8–12]. To sum up, it is not difficult to find that the above research on interval-valued data is mainly based on cross-sectional data.

Panel data or longitudinal data is obtained by multiple observations of individuals at different times. Statistical models that combine cross-section and time series real value data are becoming more and more popular in economic research, and panel data sets have certain advantages over traditional pure cross-section or pure time series data sets. Galvao Jr. [13] studied a quantile regression dynamic panel model with fixed effects. Aristodemou [14] studied semiparametric identification in linear index discrete response panel data models with fixed effects. Hamiye Beyaztas and Bandyopadhyay [15] proposed a new, weighted likelihood-based robust estimation procedure for linear panel data models with fixed and random effects. Bai et al. [16] studied generalized least squares (GLS) estimation for linear panel data models. Baltagi and Li [17] focused on prediction in spatial models based on panel data. Elhorst [18] studied the specification and estimation of spatial panel data models. Zhang et al. [19] studied a penalized quantile regression for spatial panel models with fixed effects. In particular, in recent years, a few scholars have begun to study interval panel data models. For example, Ji et al. [20] introduced fixed effects panel data regression models for interval-valued data, and presented three kinds of panel interval-valued data regression models and the estimation of their parameters. Zhang et al. [21] proposed a robust estimation method based on the iterative weighted least squares technique to reduce the impact of outliers on models. Considering the correlation between the center and range or the upper and lower bounds of intervals, Ji et al. [22] proposed the bivariate maximum likelihood (BML) method for estimating unknown parameters of the model. Usually, the observation data of different individuals are independent of each other, but there is correlation between the observation data of the same individual at different times. As far as we know, current research on panel interval-valued data has not considered the correlation within a group of the data.

Generalized estimation equations (GEEs) proposed by Liang and Zeger [23] are commonly used to analyze the longitudinal data with correlation within individuals. Crowder [24] studied the (weak) consistency and inconsistency of the solutions of general estimating equations. Balan and Schiopu-Kratina [25] studied the consistency and asymptotic normality of the GEEs estimation, with the covariate dimension fixed. Wang [26] studied the consistency and asymptotic normality of GEEs estimates when the sample size of the covariate dimension tends to infinity. Wang and Carey [27] showed that the discrepancy between the working correlation structure and the true correlation structure and the estimation method of the correlation coefficient affect the asymptotic relative efficiency. Pan [28] proposed a new model selection criterion that minimizes the expected predictive bias of estimating equations. Wang [29] presented a systematic review on GEEs, covering foundational concepts as well as several recent developments due to practical challenges in real applications. Seaman and Copas [30] described doubly robust (DR) GEEs, and illustrated their use on simulated data. To sum up, previous studies have only applied GEEs to point value data. Therefore, we intend to apply it to panel

interval-valued data in this paper.

In this paper, unlike the independent hypothesis of other studies, we consider the correlation within a group of the panel interval-valued data to obtain better parameter estimates. GEEs are commonly used to analyze the longitudinal data with correlation within individuals. We intend to apply it to panel interval-valued data in this paper. The results of Monte Carlo simulations reveal that our proposed method outperforms the LSDV method based on independent assumptions in terms of estimation accuracy and robustness. When the working correlation matrix is assumed correctly, the GEEs method performs better than the LSDV method. When the working correlation matrix is assumed incorrectly, the performance of the two methods is similar. In addition, the higher the degree of correlation within the group, the better the performance of the GEEs method. The empirical results also demonstrate the superiority of our proposed method.

The rest of this paper are organized as follows: Current regression models for fixed effects panel interval-valued data and the LSDV method are provided in Section 2. The regression method based on the generalized estimation equations is presented in Section 3. In Section 4, some Monte Carlo simulations and real data analysis show that, in contrast with the least-squares dummy-variable (LSDV) method, the proposed GEEs method has advantages in forecasting performance. Finally, Section 5 gives the conclusion and discussion.

2. Fixed effects panel interval-valued data models

In this section, the LSDV method and three existing kinds of fixed effects panel interval-valued data models are introduced: the min-max (Min-Max or MM) model, the center and range (CRM) model, the center (CM) model. These three kinds models hold their own advantages and disadvantages.

For the panel interval-valued dataset $Z = \{(X_{it}, y_{it}) \mid i = 1, 2, ..., N; t = 1, 2, ..., T\}$, let $X_{it} = (x_{it1}, x_{it2}, ..., x_{itp})$, where $x_{itk} = [x_{itk}^l, x_{itk}^u]$, $y_{it} = [y_{it}^l, y_{it}^u]$, $x_{itk}^c = \frac{x_{itk}^l + x_{itk}^u}{2}$, $x_{itk}^r = \frac{x_{itk}^u - x_{itk}^l}{2}$, $y_{it}^c = \frac{y_{it}^l + y_{it}^u}{2}$, $y_{it}^r = \frac{y_{it}^u - y_{it}^l}{2}$, $k = 1, 2, ..., p$. $x_{itk}^l(x_{itk}^u, x_{itk}^c$ or $x_{itk}^r)$ are considered as explanatory variables and $y_{it}^l(y_{it}^u, y_{it}^c$ or $y_{it}^r)$ are considered as response variables. In addition, let $u_i^l(u_i^u)$ be the individual effect and $u_i^c = \frac{u_i^u + u_i^l}{2}$, $u_i^r = \frac{u_i^u - u_i^l}{2}$.

2.1. The min-max model

The min-max model of fixed effects panel interval-valued data regression consists of two different models. Two regression models are established for the upper and lower bounds of the response variable by using the upper and lower bounds of the explanatory variables. The specific form is given as follows:

$$y_{it}^l = u_i^l + \sum_{j=1}^{p} \beta_j^l x_{itj}^l + \epsilon_{it}^l, \tag{2.1}$$

$$y_{it}^u = u_i^u + \sum_{j=1}^{p} \beta_j^u x_{itj}^u + \epsilon_{it}^u, \tag{2.2}$$

where the errors $\epsilon_{it}^l \sim N(0, \sigma^2)$, $\epsilon_{it}^u \sim N(0, \sigma^2)$, $i = 1, 2, ..., N; t = 1, 2..., T$.

Based on the lower bounds of y_{it} and x_{itj} $(i = 1, 2..., N; t = 1, 2, ..., T; j = 1, 2, ..., p)$, the parameters

u_i^l and $\beta^l = (\beta_1^l, \beta_2^l, ..., \beta_p^l)^T$ are estimated by the least-squares dummy-variable (LSDV) method:

$$\hat{\beta}^l = (\hat{\beta}_1^l, \hat{\beta}_2^l, ..., \hat{\beta}_p^l)^T = (A^l)^{-1} B^l,$$

$$\hat{u}_i^l = \bar{y}_i^l - \sum_{j=1}^{p} \hat{\beta}_j^l \bar{x}_{i.j}^l,$$

where $A^l = \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{it}^l - \overline{X}_i^l)(X_{it}^l - \overline{X}_i^l)^T$, $B^l = \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{it}^l - \overline{X}_i^l)(y_{it}^l - \bar{y}_i^l)^T$, $X_{it}^l = (x_{it1}^l, x_{it2}^l, ..., x_{itp}^l)^T$, $\overline{X}_i^l = \frac{1}{T} \sum_{t=1}^{T} X_{it}^l$, $\bar{y}_i^l = \frac{1}{T} \sum_{t=1}^{T} y_{it}^l$, $\bar{x}_{i.j}^l = \frac{1}{T} \sum_{t=1}^{T} x_{itj}^l$.

The estimates of β^u and u_i^u have the following similar conclusion:

$$\hat{\beta}^u = (\hat{\beta}_1^u, \hat{\beta}_2^u, ..., \hat{\beta}_p^u)^T = (A^u)^{-1} B^u,$$

$$\hat{u}_i^u = \bar{y}_i^u - \sum_{j=1}^{p} \hat{\beta}_j^u \bar{x}_{i.j}^u,$$

where $A^u = \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{it}^u - \overline{X}_i^u)(X_{it}^u - \overline{X}_i^u)^T$, $B^u = \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{it}^u - \overline{X}_i^u)(y_{it}^u - \bar{y}_i^u)^T$, $X_{it}^u = (x_{it1}^u, x_{it2}^u, ..., x_{itp}^u)^T$, $\overline{X}_i^u = \frac{1}{T} \sum_{t=1}^{T} X_{it}^u$, $\bar{y}_i^u = \frac{1}{T} \sum_{t=1}^{T} y_{it}^u$, $\bar{x}_{i.j}^u = \frac{1}{T} \sum_{t=1}^{T} x_{itj}^u$.

For a new observation $X_{it} = (x_{it1}, x_{it2}, ..., x_{itp})^T$, with $x_{itj} = [x_{itj}^l, x_{itj}^u]$, the prediction of $\hat{y}_{it} = [\hat{y}_{it}^l, \hat{y}_{it}^u]$ is given by:

$$\hat{y}_{it}^l = \hat{u}_i^l + \sum_{j=1}^{p} \hat{\beta}_j^l x_{itj}^l,$$

$$\hat{y}_{it}^u = \hat{u}_i^u + \sum_{j=1}^{p} \hat{\beta}_j^u x_{itj}^u.$$

Sometimes the min-max model does not guarantee the mathematical coherence of the predicted interval bounds, and then the response variables are defined as follows:

$$\hat{y}_{it} = \left[\hat{y}_{it}^l, \hat{y}_{it}^u\right] = \begin{cases} \left[\frac{\hat{y}_{it}^l + \hat{y}_{it}^u}{2}, \frac{\hat{y}_{it}^l + \hat{y}_{it}^u}{2}\right] & \text{if } \hat{y}_{it}^l > \hat{y}_{it}^u; \\ \left[\hat{y}_{it}^l, \hat{y}_{it}^u\right] & \text{if } \hat{y}_{it}^l \leq \hat{y}_{it}^u. \end{cases}$$

2.2. *The center and range model*

The center and range model of fixed effects panel interval-valued data regression is as follows:

$$y_{it}^c = u_i^c + \sum_{j=1}^{p} \beta_j^c x_{itj}^c + \epsilon_{it}^c, \tag{2.3}$$

$$y_{it}^r = u_i^r + \sum_{j=1}^{p} \beta_j^r x_{itj}^r + \epsilon_{it}^r, \tag{2.4}$$

where the errors $\epsilon_{it}^c \sim N(0, \sigma^2)$, $\epsilon_{it}^r \sim N(0, \sigma^2)$, $i = 1, 2, ..., N; t = 1, 2..., T$. The estimates of β^c and u_i^c given by the LSDV method are:

$$\hat{\beta}^c = (\hat{\beta}_1^c, \hat{\beta}_2^c, ..., \hat{\beta}_p^c)^T = (A^c)^{-1} B^c,$$

$$\hat{u}_i^c = \overline{y}_i^c - \sum_{j=1}^{p} \hat{\beta}_j^c \overline{x}_{i.j}^c,$$

where $A^c = \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{it}^c - \overline{X}_i^c)(X_{it}^c - \overline{X}_i^c)^T$, $B^c = \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{it}^c - \overline{X}_i^c)(y_{it}^c - \overline{y}_i^c)^T$, $X_{it}^c = (x_{it1}^c, x_{it2}^c, ..., x_{itp}^c)^T$, $\overline{X}_i^c = \frac{1}{T} \sum_{t=1}^{T} X_{it}^c$, $\overline{y}_i^c = \frac{1}{T} \sum_{t=1}^{T} y_{it}^c$, $\overline{x}_{i.j}^c = \frac{1}{T} \sum_{t=1}^{T} x_{itj}^c$.

Similarly, we can obtain the estimates of β^r and u_i^r :

$$\hat{\beta}^r = (\hat{\beta}_1^r, \hat{\beta}_2^r, ..., \hat{\beta}_p^r)^T = (A^r)^{-1} B^r,$$

$$\hat{u}_i^r = \overline{y}_i^r - \sum_{j=1}^{p} \hat{\beta}_j^r \overline{x}_{i.j}^r,$$

where $A^r = \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{it}^r - \overline{X}_i^r)(X_{it}^r - \overline{X}_i^r)^T$, $B^r = \sum_{i=1}^{N} \sum_{t=1}^{T} (X_{it}^r - \overline{X}_i^r)(y_{it}^r - \overline{y}_i^r)^T$, $X_{it}^r = (x_{it1}^r, x_{it2}^r, ..., x_{itp}^r)^T$, $\overline{X}_i^r = \frac{1}{T} \sum_{t=1}^{T} X_{it}^r$, $\overline{y}_i^r = \frac{1}{T} \sum_{t=1}^{T} y_{it}^r$, $\overline{x}_{i.j}^r = \frac{1}{T} \sum_{t=1}^{T} x_{itj}^r$.

For a new observation $X_{it} = (x_{it1}, x_{it2}, ..., x_{itp})^T$, with $x_{itj} = [x_{itj}^l, x_{itj}^u]$, the predictions of $\hat{y}_{it}^c$ and $\hat{y}_{it}^r$ are given by:

$$\hat{y}_{it}^c = \hat{u}_i^c + \sum_{j=1}^{p} \hat{\beta}_j^c x_{itj}^c,$$

$$\hat{y}_{it}^r = \hat{u}_i^r + \sum_{j=1}^{p} \hat{\beta}_j^r x_{itj}^r.$$

Then, the predictions of $\hat{y}_{it}^l$ and $\hat{y}_{it}^u$ are as follows:

$$\hat{y}_{it} = \left[\hat{y}_{it}^l, \hat{y}_{it}^u\right] = \begin{cases} \left[\hat{y}_{it}^c - \hat{y}_{it}^r, \hat{y}_{it}^c + \hat{y}_{it}^r\right] & \text{if } \hat{y}_{it}^r \geq 0; \\ \left[\hat{y}_{it}^c, \hat{y}_{it}^c\right] & \text{if } \hat{y}_{it}^r < 0. \end{cases}$$

2.3. The center model

Based on the center value of the panel interval-valued data, the liner regression model with fixed individual effects is as follows:

$$y_{it}^c = u_i^c + \sum_{j=1}^{p} \beta_j^c x_{itj}^c + \epsilon_{it}^c, \tag{2.5}$$

where the errors $\epsilon_{it}^c \sim N(0, \sigma^2)$, $i = 1, 2, ..., N$; $t = 1, 2, ..., T$.

As described earlier, the estimates of β^c and u_i^c are given by the LSDV method:

$$\hat{\beta}^c = (\hat{\beta}_1^c, \hat{\beta}_2^c, ..., \hat{\beta}_p^c)^T = (A^c)^{-1} B^c,$$

$$\hat{u}_i^c = \overline{y}_i^c - \sum_{j=1}^{p} \hat{\beta}_j^c \overline{x}_{i.j}^c.$$

For a new observation $X_{it} = (x_{it1}, x_{it2}, ..., x_{itp})^T$, with $x_{itj} = [x_{itj}^l, x_{itj}^u]$, the prediction of $\hat{y}_{it} = [\hat{y}_{it}^l, \hat{y}_{it}^u]$ is given by:

$$\hat{y}_{it}^l = \hat{u}_i^c + \sum_{j=1}^{p} \hat{\beta}_j^c x_{itj}^l,$$

$$\hat{y}_{it}^u = \hat{u}_i^c + \sum_{j=1}^{p} \hat{\beta}_j^c x_{itj}^u.$$

3. Fixed effects panel interval-valued data models based on the GEEs method

This section introduces generalized estimation equations(GEEs) for fixed effects panel interval-valued regression models that consider the data correlation within individuals.

For panel interval-valued data set $Z = \{(X_{it}, y_{it}) \mid i = 1, 2, ..., N; t = 1, 2, ..., T\}$, $X_{it} = (x_{it1}, x_{it2}, ..., x_{itp})^T$ is assumed as $p \times 1$ interval-valued explanatory variables with $x_{itj} = [x_{itj}^l, x_{itj}^u]$, $y_{it} = [y_{it}^l, y_{it}^u]$ as the observed interval-valued dependent variables, $i = 1, 2, ..., N; t = 1, 2, ..., T; j = 1, 2, ..., p$. u_i is an unobserved effect for individual i and is time invariant. Let $Y_i^a = (y_{i1}^a, y_{i2}^a, ..., y_{iT}^a)^T$, $X_i^a = (X_{i1}^a, X_{i2}^a, ..., X_{iT}^a)^T$, where $a = l, u, c$, or r. Let $Y^a = ((Y_1^a)^T, (Y_2^a)^T, ..., (Y_N^a)^T)^T$, $X^a = ((X_1^a)^T, (X_2^a)^T, ..., (X_N^a)^T)^T$, $u^a = (u_1^a, u_2^a, ..., u_N^a)^T$.

3.1. The GEEs method of the min-max model

The matrix form of the min-max model of fixed effects panel interval-valued data regression is as follows:

$$Y^l = X^l \beta^l + (I_N \otimes \tau_T) u^l + \epsilon^l, \tag{3.1}$$

$$Y^u = X^u \beta^u + (I_N \otimes \tau_T) u^u + \epsilon^u. \tag{3.2}$$

To save space, we only present the estimation method of the model (3.1). Considering the correlation between the observations of the same individual at different times, and the independence between the observations of different individuals, we suppose that $\epsilon^l \sim N(0, \Sigma^l)$, $cov(\epsilon_{ik}^l, \epsilon_{jl}^l) = 0 (i \neq j)$, where $\epsilon^l = (\epsilon_1^l, \epsilon_2^l, ..., \epsilon_N^l)^T$, $\epsilon_i^l = (\epsilon_{i1}^l, \epsilon_{i2}^l, ..., \epsilon_{iT}^l)$,

$$\Sigma^l = \begin{pmatrix} \Sigma_1^l & 0 & \cdots & 0 \\ 0 & \Sigma_2^l & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_N^l \end{pmatrix}, \Sigma_i^l = \begin{pmatrix} var(\epsilon_{i1}^l) & cov(\epsilon_{i1}^l, \epsilon_{i2}^l) & \cdots & cov(\epsilon_{i1}^l, \epsilon_{iT}^l) \\ cov(\epsilon_{i2}^l, \epsilon_{i1}^l) & var(\epsilon_{i2}^l) & \cdots & cov(\epsilon_{i2}^l, \epsilon_{iT}^l) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\epsilon_{iT}^l, \epsilon_{i1}^l) & cov(\epsilon_{iT}^l, \epsilon_{i2}^l) & \cdots & var(\epsilon_{iT}^l) \end{pmatrix},$$

and similarly, $\epsilon^u \sim N(0, \Sigma^u)$.

There is a problem for us to find the estimation of β^l in the model (3.1) because of the fixed effects u^l. Specifically, simultaneously estimating u^l and β^l directly may incur "curse of dimensionality" since u^l is high dimensional when N becomes too large. First, similar to Section 2.1, we use the projection techniques to eliminate fixed effects in the model (3.1). Specifically, let $Q = I_{NT} - P$, $P = I_N \otimes (J_T/T)$, $J_T = \tau_T \tau_T'$, where I_N is an $N \times N$ identity matrix and τ_T is an $N \times 1$ vector of ones. Multiplying by Q on the left for model (3.1), we have

$$\widetilde{Y}^l = \widetilde{X}^l \beta^l + \widetilde{\epsilon}^l, \tag{3.3}$$

where $\widetilde{Y}^l = QY^l$, $\widetilde{X}^l = QX^l$, $\widetilde{\epsilon}^l = Q\epsilon^l$, $Q(I_N \otimes \tau_T) = 0$, and $\widetilde{\epsilon}^l \sim N(0, Q\Sigma^l Q^T)$, $Q =$

$diag(Q_1, Q_2, ..., Q_N), Q_i = I_T - J_T/T,$

$$QΣ^lQ^T = \begin{pmatrix} Q_1Σ_1^lQ_1^T & 0 & \cdots & 0 \\ 0 & Q_2Σ_2^lQ_2^T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Q_NΣ_N^lQ_N^T \end{pmatrix}.$$

Then, the generalized estimating equations (GEEs [23]) can be defined as follows:

$$\sum_{i=1}^{N}(\widetilde{X}_i^l)^T(V_i^l)^{-1}\left(\widetilde{Y}_i^l - (\widetilde{X}_i^l)^T\beta\right) = 0, \tag{3.4}$$

where $\widetilde{Y}_i^l = Q_iY_i^l = (\tilde{y}_{i1}^l, \ldots, \tilde{y}_{iT}^l)^T$, $\widetilde{X}_i^l = Q_iX_i^l$, V_i^l is the covariance matrix of $\widetilde{Y}_i^l$. According to Liang and Zeger ([23]), we simplify the covariance of the ith subject V_i by taking $V_i^l = (A_i^l)^{1/2}R_i(\alpha)(A_i^l)^{1/2}$, where $A_i^l = diag\{var(\widetilde{Y}_{i1}^l), \ldots, var(\widetilde{Y}_{iT}^l)\}$, and $R_i(\alpha)$ is a common working correlation with nuisance parameter α. In general, V_i^l is unknown, and we often use the empirical estimators $\hat{V}_i^l$ for the V_i^l. Specifically,

$$\hat{V}_i^l = (A_i^l)^{1/2}(\hat{\beta}^l)R_i(\hat{\alpha})(A_i^l)^{1/2}(\hat{\beta}^l),$$

where $\hat{\alpha}$ can be obtained by following the method proposed by Liang and Zeger ([23]) in Section 3.4.1. $A_i^l(\hat{\beta}^l)$ is the empirical estimator of A_i based on the current estimator of β^l.

Thus, we implement the iterative algorithm to solve Eq (3.4). Given the current estimate $\hat{\alpha}$, we use the following iterative procedure for β^l:

$$\hat{\beta}^{l(k+1)} = \hat{\beta}^{l(k)} + \left(\sum_{i=1}^{N}(\widetilde{X}_i^l)^T(\hat{V}_i^l)^{-1}\widetilde{X}_i^l\right)^{-1}\left(\sum_{i=1}^{N}(\widetilde{X}_i^l)^T(\hat{V}_i^l)^{-1}(\widetilde{Y}_i^l - \widetilde{X}_i^l\hat{\beta}^{l(k)})\right), \tag{3.5}$$

where $\hat{V}_i^l = (A_i^l)^{1/2}(\hat{\beta}^{l(k)})R_i(\hat{\alpha})(A_i^l)^{1/2}(\hat{\beta}^{l(k)})$.

Then, the estimates of the fixed effects u^l are given as follows:

$$\hat{u}^l = \left((I_N \otimes \tau_T)^T(I_N \otimes \tau_T)\right)^{-1}(I_N \otimes \tau_T)^T(\overline{Y}^l - \overline{X}^l\hat{\beta}^l), \tag{3.6}$$

where $\overline{Y}^l = PY^l$, $\overline{X}^l = PX^l$.

Using similar methods, we can obtain the estimates of $\hat{\beta}^u$ by iterating the following ralationship to convergence:

$$\hat{\beta}^{u(k+1)} = \hat{\beta}^{u(k)} + \left(\sum_{i=1}^{N}(\widetilde{X}_i^u)^T(\hat{V}_i^u)^{-1}\widetilde{X}_i^u\right)^{-1}\left(\sum_{i=1}^{N}(\widetilde{X}_i^u)^T(\hat{V}_i^u)^{-1}(\widetilde{Y}_i^u - \widetilde{X}_i^u\hat{\beta}^{u(k)})\right), \tag{3.7}$$

where $\hat{V}_i^u = (A_i^u)^{1/2}(\hat{\beta}^{u(k)})R_i(\hat{\alpha})(A_i^u)^{1/2}(\hat{\beta}^{u(k)})$, and the estimates of $\hat{u}^u$ are given by

$$\hat{u}^u = \left((I_N \otimes \tau_T)^T(I_N \otimes \tau_T)\right)^{-1}(I_N \otimes \tau_T)^T(\overline{Y}^u - \overline{X}^u\hat{\beta}^u), \tag{3.8}$$

where $\overline{Y}^u = PY^u$, $\overline{X}^u = PX^u$.

For a new observation $X = [X^l, X^u]$, the prediction $\hat{Y} = [\hat{Y}^l, \hat{Y}^u]$ is given by

$$\hat{Y}^l = X^l\hat{\beta}^l + (I_N \otimes \tau_T)\hat{u}^l,$$

$$\hat{Y}^u = X^u\hat{\beta}^u + (I_N \otimes \tau_T)\hat{u}^u.$$

3.2. The GEEs method of the center and range model

The matrix form of the center and range model of fixed effects panel interval-valued data regression is as follows:

$$Y^c = X^c\beta^c + (I_N \otimes \tau_T)u^c + \epsilon^c, \tag{3.9}$$

$$Y^r = X^r\beta^r + (I_N \otimes \tau_T)u^r + \epsilon^r, \tag{3.10}$$

where $\epsilon^c \sim N(0, \Sigma^c), \epsilon^r \sim N(0, \Sigma^r), \Sigma^c = \frac{\Sigma^l + \Sigma^u}{4} = \Sigma^r$.

Similar to Section 3.1, we only present the estimation method of model (3.9). To eliminate the fixed effects, let $\widetilde{Y}^c = QY^c, \widetilde{X}^c = QX^c, \widetilde{\epsilon}^c = Q\epsilon^c$, and $Q(I_N \otimes \tau_T) = 0$. Multiplying (3.9) by Q on the left gives the following intra-group regression model equation:

$$\widetilde{Y}^c = \widetilde{X}^c\beta^c + \widetilde{\epsilon}^c, \tag{3.11}$$

where $\widetilde{\epsilon}^c \sim N(0, Q\Sigma^c Q^T)$.

Then, we can obtain the estimates of $\hat{\beta}^c$ by iterating the following ralationship to convergence:

$$\hat{\beta}^{c\,(k+1)} = \hat{\beta}^{c\,(k)} + \left(\sum_{i=1}^{N} (\widetilde{X}_i^c)^T (\hat{V}_i^c)^{-1} \widetilde{X}_i^c \right)^{-1} \left(\sum_{i=1}^{N} (\widetilde{X}_i^c)^T (\hat{V}_i^c)^{-1} (\widetilde{Y}_i^c - \widetilde{X}_i^c \hat{\beta}^{c\,(k)}) \right),$$

where $\hat{V}_i^c = (A_i^c)^{1/2}(\hat{\beta}^{c\,(k)})R_i(\hat{\alpha})(A_i^c)^{1/2}(\hat{\beta}^{c\,(k)})$, and the estimates of $\hat{u}^c$ are given as follows:

$$\hat{u}^c = \left((I_N \otimes \tau_T)^T (I_N \otimes \tau_T) \right)^{-1} (I_N \otimes \tau_T)^T (\overline{Y}^c - \overline{X}^c \hat{\beta}^c),$$

where $\overline{Y}^c = PY^c, \overline{X}^c = PX^c$.

Similarly, estimates of β^r can be obtained as follows:

$$\hat{\beta}^{r\,(k+1)} = \hat{\beta}^{r\,(k)} + \left(\sum_{i=1}^{N} (\widetilde{X}_i^r)^T (\hat{V}_i^r)^{-1} \widetilde{X}_i^r \right)^{-1} \left(\sum_{i=1}^{N} (\widetilde{X}_i^r)^T (\hat{V}_i^r)^{-1} (\widetilde{Y}_i^r - \widetilde{X}_i^r \hat{\beta}^{r\,(k)}) \right),$$

where $\hat{V}_i^r = (A_i^r)^{1/2}(\hat{\beta}^{r\,(k)})R_i(\hat{\alpha})(A_i^r)^{1/2}(\hat{\beta}^{r\,(k)})$, and the estimates of $\hat{u}^r$ are given by

$$\hat{u}^r = \left((I_N \otimes \tau_T)^T (I_N \otimes \tau_T) \right)^{-1} (I_N \otimes \tau_T)^T (\overline{Y}^r - \overline{X}^r \hat{\beta}^r),$$

where $\overline{Y}^r = PY^r, \overline{X}^r = PX^r$.

For a new observation $X = [X^l, X^u]$, the prediction $\hat{Y} = [\hat{Y}^l, \hat{Y}^u]$ is given as follows:

$$\hat{Y}^l = \hat{Y}^c - \hat{Y}^r, \hat{Y}^u = \hat{Y}^c + \hat{Y}^r,$$

where

$$\hat{Y}^c = X^c\hat{\beta}^c + (I_N \otimes \tau_T)\hat{u}^c,$$

$$\hat{Y}^r = X^r\hat{\beta}^r + (I_N \otimes \tau_T)\hat{u}^r.$$

3.3. The GEEs method of the center model

The matrix form of the center model of fixed effects panel interval-valued data regression is as follows:

$$Y^c = X^c \beta^c + (I_N \otimes \tau_T) u^c + \epsilon^c, \tag{3.12}$$

where $\epsilon^c \sim N(0, \Sigma^c), \Sigma^c = \frac{\Sigma^l + \Sigma^u}{4}$.

As described earlier, the estimates of β^c and u^c given by GEEs are as follows:

$$\hat{\beta}^{c(k+1)} = \hat{\beta}^{c(k)} + \left(\sum_{i=1}^{N} (\widetilde{X}_i^c)^T (\hat{V}_i^c)^{-1} \widetilde{X}_i^c \right)^{-1} \left(\sum_{i=1}^{N} (\widetilde{X}_i^c)^T (\hat{V}_i^c)^{-1} (\widetilde{Y}_i^c - \widetilde{X}_i^c \hat{\beta}^{c(k)}) \right),$$

where $\hat{V}_i^c = (A_i^c)^{1/2} (\hat{\beta}^{c(k)}) R_i(\hat{\alpha}) (A_i^c)^{1/2} (\hat{\beta}^{c(k)})$, and the estimates of $\hat{u}^c$ are obtained as follows:

$$\hat{u}^c = \left((I_N \otimes \tau_T)^T (I_N \otimes \tau_T) \right)^{-1} (I_N \otimes \tau_T)^T (\overline{Y}^c - \overline{X}^c \hat{\beta}^c),$$

where $\overline{Y}^c = PY^c, \overline{X}^c = PX^c$.

For a new observation $X = [X^l, X^u]$, the prediction $\hat{Y} = [\hat{Y}^l, \hat{Y}^u]$ is given by

$$\hat{Y}^l = X^l \hat{\beta}^c + (I_N \otimes \tau_T) \hat{u}^c,$$

$$\hat{Y}^u = X^u \hat{\beta}^c + (I_N \otimes \tau_T) \hat{u}^c.$$

3.4. Computation algorithm

3.4.1. The working correlation matrix

The following introduces the structures of several commonly used working correlation matrices [23, 31, 32]:

- Independence: For the independence working correlation matrix, it is assumed that all the response variables are mutually independent. $\alpha = 0, R_i(\alpha) = I_T$.
- Exchangeable: Suppose that the correlation between any two observed values of the same individual are the same. The form of the correlation matrix is as follows:

$$R_i(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha & \cdots & \alpha \\ \alpha & 1 & \alpha & \cdots & \alpha \\ \alpha & \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \alpha & \cdots & 1 \end{pmatrix},$$

where α is the unknown correlation coefficient,

$$\hat{\alpha} = \frac{(NT - P) \sum_{i=1}^{N} \sum_{t \neq k}^{T} r_{it} r_{ik}}{(T(T-1)N - p) \sum_{i=1}^{N} \sum_{t=1}^{T} r_{it}^2},$$

where $r_{it} = \widetilde{y}_{it} - \widetilde{X}_{it} \beta$.

- AR(1): Suppose that the correlation depends on the time. The form of the correlation matrix is as follows:

$$R_i(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{T-1} \\ \alpha & 1 & \alpha & \cdots & \alpha^{T-2} \\ \alpha^2 & \alpha & 1 & \cdots & \alpha^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^{T-1} & \alpha^{T-2} & \alpha^{T-3} & \cdots & 1 \end{pmatrix},$$

where

$$\hat{\alpha} = \frac{(NT - p) \sum_{i=1}^{N} \sum_{t=1}^{T-1} r_{it} r_{i,t+1}}{(T(T-1)N - p) \sum_{i=1}^{N} \sum_{t=1}^{T} r_{it}^2}.$$

3.4.2. Algorithm

In this subsection, we present a feasible algorithm as follows. Consider the general form of the model: $Y = X\beta + (I_N \otimes \tau_T)u + \epsilon$.

Step 1: Use the matrix Q to eliminate fixed effects u. Then the initial value $\hat{\beta}^{(0)}$ is obtained from the independent assumption.

Step 2: The estimates $\hat{\alpha}^{(k)}$ and $A_i^{1/2}(\hat{\beta}^{(k)})$ are obtained based on the assumed structure of the correlation matrix as in Section 3.4.1. Then $\hat{V}_i^{(k)} = A_i^{1/2}(\hat{\beta}^{(k)})R_i(\hat{\alpha})A_i^{1/2}(\hat{\beta}^{(k)})$ is obtained. Update $\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \left(\sum_{i=1}^{N} (\widetilde{X}_i)^T (\hat{V}_i^{(k)})^{-1} \widetilde{X}_i \right)^{-1} \left(\sum_{i=1}^{N} (\widetilde{X}_i)^T (\hat{V}_i^{(k)})^{-1} (\widetilde{Y}_i - \widetilde{X}_i \hat{\beta}^{(k)}) \right)$.

Step 3: Iterate Step 2 until convergence, and denote the final estimators as the GEEs estimators. Then $\hat{u} = \left((I_N \otimes \tau_T)^T (I_N \otimes \tau_T) \right)^{-1} (I_N \otimes \tau_T)^T (\overline{Y} - \overline{X}\hat{\beta})$.

Here we can also utilize the existing algorithm to obtain estimates of β, which can be obtained by using the geepack function package in the statistical software R language.

4. Numerical illustrations

This section compares the proposed GEEs method with the LSDV method proposed by Ji et al. [20] through computational experiments demonstration. In computational experiments, two types of panel interval-valued datasets are considered. The first is a dataset of synthetic panel interval-valued data, while the second is the air pollution dataset of the 26 cities in China's Yangtze River Delta.

4.1. Monte Carlo simulations

Without loss of generality, we provide two configuration schemes for synthetic panel interval-valued datasets: one from the min-max model and another derived from the CRM model. In the simulation framework, 20 or 50 individuals are considered, and each individual consists of three attributes of explanatory variables over 14 continuous periods, that is, $N = 20$ or 50 and $T = 14$. Let the first 10 periods of data be the train set, and the rest be the test set. Considering the actual problem scenarios, let the correlation matrix be the first-order autoregressive (AR(1)) correlation setting $R_1(\alpha)$ or exchangeable correlation setting $R_2(\alpha)$.

$$R_1(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{T-1} \\ \alpha & 1 & \alpha & \cdots & \alpha^{T-2} \\ \alpha^2 & \alpha & 1 & \cdots & \alpha^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^{T-1} & \alpha^{T-2} & \alpha^{T-3} & \cdots & 1 \end{pmatrix}, R_2(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha & \cdots & \alpha \\ \alpha & 1 & \alpha & \cdots & \alpha \\ \alpha & \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \alpha & \cdots & 1 \end{pmatrix}.$$

4.1.1. Dataset configuration based on the CRM model

In this section, we consider the following panel interval-valued data-generating processes.

$$y_{it}^c = u_i^c + \sum_{j=1}^{3} \beta_j^c x_{itj}^c + \epsilon_{it}^c,$$

$$y_{it}^r = u_i^r + \sum_{j=1}^{3} \beta_j^r x_{itj}^r + \epsilon_{it}^r,$$

where $\beta^c = (\beta_1^c, \beta_2^c, \beta_3^c) = (1, 3, 2)$, $\beta^r = (\beta_1^r, \beta_2^r, \beta_3^r) = (1, 2, 1)$; $x_{ijt}^c \sim U[5, 10]$, $x_{ijt}^r \sim U[1, 3]$; fixed effects u_i^c and $u_i^r \sim U[0, 0.5]$; $\epsilon_i^c = (\epsilon_{i1}^c, \epsilon_{i2}^c, ..., \epsilon_{iT}^c) \sim N(0, \sigma_c^2 R_i(\alpha))$, $\epsilon_i^r = (\epsilon_{i1}^r, \epsilon_{i2}^r, ..., \epsilon_{iT}^r) \sim N(0, \sigma_r^2 R_i(\alpha))$, where $\sigma_c^2 = \sigma_r^2 = 4, i = 1, 2$; and let the correlation coefficient $\alpha = 0.9, 0.5, 0.3$ to generate data with different degrees of correlation.

4.1.2. Dataset configuration based on the min-max model

In this section, we consider the following panel interval-valued data-generating processes.

$$y_{it}^l = u_i^l + \sum_{j=1}^{3} \beta_j^l x_{itj}^l + \epsilon_{it}^l,$$

$$y_{it}^u = u_i^u + \sum_{j=1}^{3} \beta_j^u x_{itj}^u + \epsilon_{it}^u,$$

where $\beta^u = (\beta_1^u, \beta_2^u, \beta_3^u) = (3, 2, 2)$, $\beta^l = (\beta_1^l, \beta_2^l, \beta_3^l) = (2, 1, 1)$; $x_{ijt}^u \sim U[5, 10]$, $x_{ijt}^l \sim U[1, 3]$; fixed effects u_i^u and $u_i^l \sim U[0, 0.5]$; $\epsilon_i^u = (\epsilon_{i1}^u, \epsilon_{i2}^u, ..., \epsilon_{iT}^u) \sim N(0, \sigma_u^2 R_i(\alpha))$, $\epsilon_i^l = (\epsilon_{i1}^l, \epsilon_{i2}^l, ..., \epsilon_{iT}^l) \sim N(0, \sigma_l^2 R_i(\alpha))$, where $\sigma_u^2 = \sigma_l^2 = 4, i = 1, 2$; and let the correlation coefficient $\alpha = 0.9, 0.5, 0.3$ to generate data with different degrees of correlation.

4.1.3. Implementation and evaluation

Four criteria for assessing the performances of different models are shown as follows:
(1) Mean average absolute error (MAE), mean magnitude of relative error (MMER), and root mean square error (RMSE) defined by Ji et al. [20] are:

$$MAE = \frac{1}{2NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left\{ \left| \hat{y}_{it}^l - y_{it}^l \right| + \left| \hat{y}_{it}^u - y_{it}^u \right| \right\},$$

$$MMER = \frac{1}{2NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left\{ \left| \frac{\hat{y}_{it}^l - y_{it}^l}{\hat{y}_{it}^l} \right| + \left| \frac{\hat{y}_{it}^u - y_{it}^u}{\hat{y}_{it}^u} \right| \right\},$$

$$RMSE = \frac{1}{2} \left\{ \sqrt{\frac{\sum_{i=1}^{N} \sum_{t=1}^{T} \left(\hat{y}_{it}^l - y_{it}^l \right)^2}{NT}} + \sqrt{\frac{\sum_{i=1}^{N} \sum_{t=1}^{T} \left(\hat{y}_{it}^u - y_{it}^u \right)^2}{NT}} \right\}.$$

(2) The rate of different intervals (RI) defined by Hu and He [33] is:

$$RI = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{w(y_{it} \cap \hat{y}_{it})}{w(y_{it} \cup \hat{y}_{it})},$$

where $w(.)$ is the interval width.

Table 1. Comparison for the GEEs and LSDV methods based on the dataset generated by the CRM model and AR(1) correlation matrix, with the average and mean square error, where the working correlation matrix is AR(1).

α	Methods	N	$\hat{\beta}_1^c$	$\hat{\beta}_2^c$	$\hat{\beta}_3^c$	$\hat{\beta}_1^r$	$\hat{\beta}_2^r$	$\hat{\beta}_3^r$
0.3	LSDV	20	0.9992	3.0005	1.9987	1.0045	2.0109	1.0089
			(0.0078)	(0.0072)	(0.0069)	(0.0428)	(0.0445)	(0.0451)
		50	1.0001	3.0034	1.9993	1.0060	2.0044	1.0030
			(0.0027)	(0.0027)	(0.0026)	(0.0161)	(0.0163)	(0.0187)
	GEEs	20	0.9999	3.0001	1.9990	1.0025	2.0090	1.0115
			(0.0068)	(0.0064)	(0.0062)	(0.0381)	(0.0410)	(0.0406)
		50	1.0000	3.0032	1.9991	1.0065	2.0045	1.0026
			(0.0025)	(0.0024)	(0.0024)	(0.0144)	(0.0144)	(0.0165)
0.5	LSDV	20	1.0016	3.0024	2.0032	1.0002	1.9993	0.9986
			(0.0064)	(0.0064)	(0.0062)	(0.0399)	(0.0397)	(0.0403)
		50	1.0013	3.0006	2.0008	1.0090	1.9994	0.9969
			(0.0025)	(0.0026)	(0.0025)	(0.0145)	(0.0152)	(0.0157)
	GEEs	20	1.0004	3.0023	2.0020	1.0021	2.0027	0.9960
			(0.0046)	(0.0046)	(0.0044)	(0.0276)	(0.0285)	(0.0291)
		50	1.0018	3.0002	2.0010	1.0079	1.9984	0.9958
			(0.0017)	(0.0019)	(0.0018)	(0.0106)	(0.0113)	(0.0112)
0.9	LSDV	20	0.9989	3.0018	1.9990	0.9944	1.9984	1.0046
			(0.0027)	(0.0030)	(0.0027)	(0.0188)	(0.0167)	(0.0180)
		50	0.9992	2.9987	2.0007	1.0024	1.9975	0.9991
			(0.0011)	(0.0011)	(0.0011)	(0.0070)	(0.0074)	(0.0074)
	GEEs	20	1.0001	3.0001	1.9996	0.9972	1.9996	1.0059
			(0.0009)	(0.0009)	(0.0009)	(0.0057)	(0.0056)	(0.0051)
		50	0.9996	2.9995	2.0002	1.0012	1.9986	0.9989
			(0.0004)	(0.0004)	(0.0003)	(0.0021)	(0.0023)	(0.0022)

Table 2. Comparison for the GEEs and LSDV methods based on the dataset generated by the min-max model and AR(1) correlation matrix, with the average and mean square error, where the working correlation matrix is AR(1).

α	Methods	N	$\hat{\beta}_1^u$	$\hat{\beta}_2^u$	$\hat{\beta}_3^u$	$\hat{\beta}_1^l$	$\hat{\beta}_2^l$	$\hat{\beta}_3^l$
0.3	LSDV	20	3.0032	2.0012	1.9997	1.9962	0.9965	1.0160
			(0.0074)	(0.0075)	(0.0069)	(0.0465)	(0.0428)	(0.0394)
		50	3.0004	2.0015	1.9975	1.9853	1.0019	1.0033
			(0.0028)	(0.0026)	(0.0027)	(0.0182)	(0.0180)	(0.0171)
	GEEs	20	3.0032	2.0015	1.9998	1.9981	0.9958	1.0145
			(0.0068)	(0.0067)	(0.0062)	(0.0417)	(0.0377)	(0.0350)
		50	3.0009	2.0016	1.9978	1.9874	1.0020	1.0026
			(0.0025)	(0.0023)	(0.0024)	(0.0162)	(0.0160)	(0.0152)
0.5	LSDV	20	3.0008	2.0004	1.9969	1.9988	0.9992	0.9957
			(0.0067)	(0.0065)	(0.0066)	(0.0421)	(0.0433)	(0.0420)
		50	3.0004	2.0016	2.0017	1.9966	0.9982	0.9935
			(0.0025)	(0.0024)	(0.0024)	(0.0158)	(0.0161)	(0.0164)
	GEEs	20	3.0012	1.9989	1.9979	2.0000	0.9978	0.9947
			(0.0047)	(0.0047)	(0.0047)	(0.0312)	(0.0306)	(0.0292)
		50	3.0003	1.9999	2.0016	1.9993	0.9992	0.9922
			(0.0018)	(0.0018)	(0.0017)	(0.0114)	(0.0109)	(0.0115)
0.9	LSDV	20	2.9989	2.0018	1.9990	1.9944	0.9984	1.0046
			(0.0027)	(0.0030)	(0.0027)	(0.0188)	(0.0167)	(0.0180)
		50	2.9999	1.9990	1.9998	2.0007	1.0014	1.0004
			(0.0011)	(0.0011)	(0.0011)	(0.0072)	(0.0071)	(0.0076)
	GEEs	20	3.0001	2.0001	1.9996	1.9972	0.9996	1.0059
			(0.0009)	(0.0009)	(0.0009)	(0.0057)	(0.0056)	(0.0051)
		50	3.0000	1.9993	1.9996	2.0025	1.0004	1.0003
			(0.0003)	(0.0003)	(0.0004)	(0.0023)	(0.0022)	(0.0022)

Table 3. Comparison for the GEEs and LSDV methods based on the dataset generated by the CRM model and AR(1) correlation matrix, with the average and standard deviation, where the working correlation matrix is AR(1).

N	α	Methods	RMSE			MAE			MMER			RI		
			CM	MM	CRM	CM	MM	CRM	CM	MM	CRM	CM	MM	CRM
20	0.3	LSDV	3.4228	2.2270	2.1540	4.0842	2.5173	2.4348	0.1003	0.0587	0.0568	0.6621	0.7303	0.7405
			(0.2662)	(0.1408)	(0.1364)	(0.3657)	(0.1661)	(0.1607)	(0.0101)	(0.0041)	(0.0040)	(0.0241)	(0.0177)	(0.0170)
		GEEs	3.4220	2.2259	2.1529	4.0830	2.5160	2.4334	0.1003	0.0587	0.0568	0.6622	0.7305	0.7407
			(0.2605)	(0.1407)	(0.1362)	(0.3580)	(0.1658)	(0.1604)	(0.0099)	(0.0041)	(0.0039)	(0.0238)	(0.0177)	(0.0170)
	0.5	LSDV	3.4333	2.2389	2.1658	4.0981	2.5350	2.4517	0.1007	0.0591	0.0572	0.6617	0.7285	0.7390
			(0.2783)	(0.1501)	(0.1483)	(0.3855)	(0.1796)	(0.1755)	(0.0106)	(0.0045)	(0.0044)	(0.0266)	(0.0190)	(0.0183)
		GEEs	3.4276	2.2366	2.1630	4.0902	2.5323	2.4487	0.1004	0.0590	0.0571	0.6620	0.7288	0.7393
			(0.2617)	(0.1503)	(0.1483)	(0.3622)	(0.1798)	(0.1755)	(0.0099)	(0.0045)	(0.0044)	(0.0258)	(0.0190)	(0.0183)
	0.9	LSDV	3.2706	1.7869	1.6946	3.9601	2.0200	1.9133	0.0974	0.0472	0.0447	0.6715	0.7713	0.7850
			(0.2928)	(0.1582)	(0.1587)	(0.4088)	(0.1882)	(0.1899)	(0.0109)	(0.0046)	(0.0046)	(0.0328)	(0.0231)	(0.0223)
		GEEs	3.2692	1.7837	1.6904	3.9589	2.0164	1.9084	0.0973	0.0471	0.0446	0.6715	0.7717	0.7855
			(0.2791)	(0.1580)	(0.1585)	(0.3900)	(0.1881)	(0.1896)	(0.0103)	(0.0046)	(0.0046)	(0.0323)	(0.0230)	(0.0223)
50	0.3	LSDV	3.4200	2.2120	2.1399	4.0807	2.4978	2.4161	0.1002	0.0582	0.0563	0.6621	0.7321	0.7423
			(0.1621)	(0.0886)	(0.0870)	(0.2242)	(0.1051)	(0.1024)	(0.0062)	(0.0026)	(0.0025)	(0.0153)	(0.0112)	(0.0107)
		GEEs	3.4193	2.2116	2.1394	4.0798	2.4973	2.4156	0.1002	0.0582	0.0563	0.6621	0.7321	0.7423
			(0.1584)	(0.0886)	(0.0868)	(0.2188)	(0.1051)	(0.1023)	(0.0060)	(0.0026)	(0.0025)	(0.0151)	(0.0112)	(0.0107)
	0.5	LSDV	3.4248	2.2355	2.1626	4.0833	2.5215	2.4398	0.1002	0.0588	0.0569	0.6621	0.7298	0.7401
			(0.1768)	(0.0980)	(0.0968)	(0.2440)	(0.1156)	(0.1139)	(0.0067)	(0.0028)	(0.0028)	(0.0173)	(0.0122)	(0.0118)
		GEEs	3.4244	2.2343	2.1614	4.0826	2.5202	2.4383	0.1002	0.0588	0.0569	0.6621	0.7299	0.7402
			(0.1698)	(0.0978)	(0.0965)	(0.2337)	(0.1154)	(0.1137)	(0.0064)	(0.0028)	(0.0028)	(0.0170)	(0.0122)	(0.0118)
	0.9	LSDV	3.2857	1.7905	1.6982	3.9696	2.0188	1.9134	0.0976	0.0472	0.0447	0.6702	0.7711	0.7845
			(0.1867)	(0.1035)	(0.1045)	(0.2578)	(0.1217)	(0.1237)	(0.0068)	(0.0030)	(0.0030)	(0.0205)	(0.0144)	(0.0143)
		GEEs	3.2851	1.7893	1.6967	3.9688	2.0173	1.9116	0.0975	0.0471	0.0447	0.6703	0.7712	0.7847
			(0.1767)	(0.1033)	(0.1045)	(0.2428)	(0.1215)	(0.1236)	(0.0064)	(0.0030)	(0.0030)	(0.0200)	(0.0144)	(0.0143)

Table 4. Comparison for the GEEs and LSDV methods based on the dataset generated by the min-max model and AR(1) correlation matrix, with the average and standard deviation, where the working correlation matrix is AR(1).

N	α	Methods	RMSE			MAE			MMER			RI		
			CM	MM	CRM	CM	MM	CRM	CM	MM	CRM	CM	MM	CRM
20	0.3	LSDV	3.3163	1.5252	1.6014	4.1985	1.7246	1.8087	0.2120	0.1262	0.1475	0.8115	0.9248	0.9213
			(0.4332)	(0.0910)	(0.0951)	(0.6468)	(0.1095)	(0.1135)	(0.0255)	(0.0115)	(0.0195)	(0.0288)	(0.0048)	(0.0050)
		GEEs	3.3134	1.5245	1.6009	4.1944	1.7238	1.8082	0.2118	0.1261	0.1475	0.8117	0.9248	0.9213
			(0.4239)	(0.0908)	(0.0950)	(0.6330)	(0.1093)	(0.1133)	(0.0251)	(0.0115)	(0.0194)	(0.0282)	(0.0048)	(0.0050)
	0.5	LSDV	3.3147	1.5380	1.6117	4.1942	1.7413	1.8238	0.2119	0.1282	0.1549	0.8117	0.9241	0.9206
			(0.4115)	(0.1088)	(0.1120)	(0.6125)	(0.1290)	(0.1309)	(0.0252)	(0.0133)	(0.0993)	(0.0272)	(0.0056)	(0.0056)
		GEEs	3.3116	1.5359	1.6101	4.1900	1.7390	1.8221	0.2118	0.1280	0.1528	0.8119	0.9242	0.9207
			(0.3832)	(0.1087)	(0.1119)	(0.5705)	(0.1288)	(0.1307)	(0.0238)	(0.0133)	(0.0538)	(0.0253)	(0.0056)	(0.0056)
	0.9	LSDV	3.2148	1.1983	1.2943	4.1311	1.3541	1.4603	0.2096	0.1045	0.1635	0.8143	0.9402	0.9357
			(0.3274)	(0.1122)	(0.1122)	(0.4881)	(0.1333)	(0.1333)	(0.0228)	(0.0212)	(0.3487)	(0.0207)	(0.0058)	(0.0058)
		GEEs	3.2102	1.1953	1.2921	4.1251	1.3506	1.4579	0.2094	0.1043	0.1633	0.8146	0.9403	0.9358
			(0.2888)	(0.1121)	(0.1121)	(0.4295)	(0.1332)	(0.1333)	(0.0212)	(0.0206)	(0.3760)	(0.0180)	(0.0058)	(0.0058)
50	0.3	LSDV	3.3000	1.5166	1.5935	4.1750	1.7129	1.7983	0.2111	0.1252	0.1468	0.8125	0.9252	0.9217
			(0.2773)	(0.0589)	(0.0616)	(0.4146)	(0.0706)	(0.0739)	(0.0164)	(0.0071)	(0.0120)	(0.0184)	(0.0031)	(0.0032)
		GEEs	3.2982	1.5162	1.5932	4.1724	1.7125	1.7981	0.2110	0.1252	0.1468	0.8126	0.9252	0.9217
			(0.2712)	(0.0589)	(0.0616)	(0.4054)	(0.0706)	(0.0739)	(0.0161)	(0.0071)	(0.0120)	(0.0180)	(0.0031)	(0.0032)
	0.5	LSDV	3.3180	1.5288	1.6052	4.2001	1.7258	1.8114	0.2116	0.1272	0.1683	0.8113	0.9246	0.9211
			(0.2575)	(0.0687)	(0.0711)	(0.3841)	(0.0812)	(0.0842)	(0.0158)	(0.0084)	(0.4595)	(0.0169)	(0.0035)	(0.0036)
		GEEs	3.3161	1.5277	1.6044	4.1974	1.7246	1.8105	0.2115	0.1271	0.1525	0.8114	0.9247	0.9211
			(0.2361)	(0.0686)	(0.0710)	(0.3527)	(0.0811)	(0.0841)	(0.0147)	(0.0084)	(0.0277)	(0.0155)	(0.0035)	(0.0036)
	0.9	LSDV	3.2425	1.2003	1.2956	4.1651	1.3531	1.4593	0.2111	0.1049	0.2063	0.8130	0.9402	0.9357
			(0.2095)	(0.0689)	(0.0696)	(0.3121)	(0.0808)	(0.0809)	(0.0143)	(0.0197)	(1.5871)	(0.0134)	(0.0036)	(0.0036)
		GEEs	3.2377	1.1993	1.2948	4.1582	1.3519	1.4585	0.2108	0.1044	0.1552	0.8133	0.9403	0.9357
			(0.1833)	(0.0690)	(0.0696)	(0.2722)	(0.0808)	(0.0810)	(0.0130)	(0.0140)	(0.1258)	(0.0115)	(0.0036)	(0.0036)

Table 5. Comparison for the GEEs and LSDV methods based on the dataset generated by the CRM model and exchangeable correlation matrix, with the average and mean square error, where the working correlation matrix is AR(1).

α	Methods	N	$\hat{\beta}_1^c$	$\hat{\beta}_2^c$	$\hat{\beta}_3^c$	$\hat{\beta}_1^r$	$\hat{\beta}_2^r$	$\hat{\beta}_3^r$
0.3	LSDV	20	0.9976	2.9994	2.0028	0.9989	1.9941	0.9976
			(0.0058)	(0.0050)	(0.0054)	(0.0344)	(0.0299)	(0.0360)
		50	1.0028	3.0011	1.9999	1.0058	2.0037	1.0013
			(0.0023)	(0.0022)	(0.0021)	(0.0127)	(0.0137)	(0.0133)
	GEEs	20	0.9975	2.9991	2.0030	0.9983	1.9943	0.9968
			(0.0058)	(0.0052)	(0.0055)	(0.0354)	(0.0305)	(0.0365)
		50	1.0029	3.0013	2.0000	1.0053	2.0039	1.0014
			(0.0023)	(0.0022)	(0.0021)	(0.0128)	(0.0137)	(0.0135)
0.5	LSDV	20	1.0030	2.9975	1.9990	1.0054	1.9942	1.0067
			(0.0039)	(0.0040)	(0.0038)	(0.0237)	(0.0223)	(0.0246)
		50	1.0014	2.9979	2.0001	1.0030	2.0004	0.9961
			(0.0015)	(0.0013)	(0.0015)	(0.009)	(0.0095)	(0.0094)
	GEEs	20	1.0033	2.9974	1.9991	1.0048	1.9932	1.0073
			(0.0040)	(0.0041)	(0.0038)	(0.0244)	(0.0223)	(0.0251)
		50	1.0014	2.9980	1.9999	1.0027	2.0004	0.9963
			(0.0015)	(0.0014)	(0.0015)	(0.0091)	(0.0095)	(0.0095)
0.9	LSDV	20	1.0004	2.9998	2.0014	0.9979	1.9990	0.9999
			(0.0007)	(0.0008)	(0.0007)	(0.0045)	(0.0045)	(0.0049)
		50	1.0004	2.9999	1.9997	1.0008	1.9964	1.0001
			(0.0003)	(0.0003)	(0.0003)	(0.0019)	(0.0019)	(0.0020)
	GEEs	20	1.0004	2.9996	2.0014	0.9978	1.9988	0.9996
			(0.0007)	(0.0008)	(0.0007)	(0.0046)	(0.0045)	(0.0049)
		50	1.0003	3.0000	1.9998	1.0008	1.9963	1.0001
			(0.0003)	(0.0003)	(0.0003)	(0.0019)	(0.0019)	(0.0020)

Table 6. Comparison for the GEEs and LSDV methods based on the dataset generated by the min-max model and exchangeable correlation matrix, with the average and mean square error, where the working correlation matrix is AR(1).
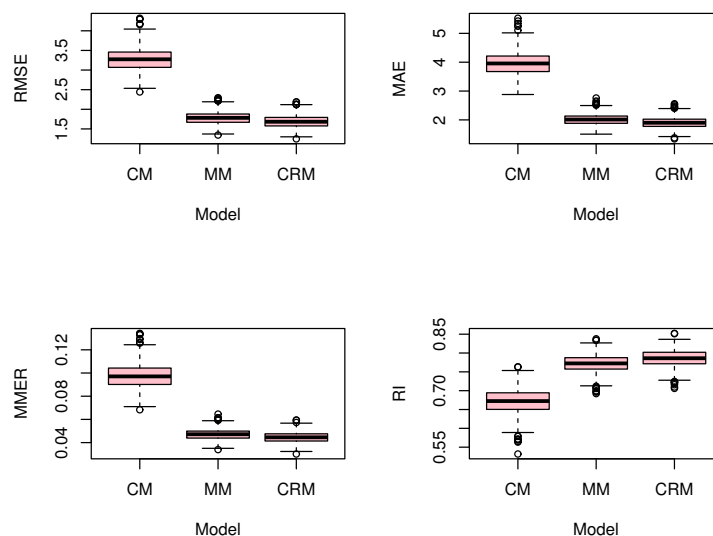
α	Methods	N	$\hat{\beta}_1^u$	$\hat{\beta}_2^u$	$\hat{\beta}_3^u$	$\hat{\beta}_1^l$	$\hat{\beta}_2^l$	$\hat{\beta}_3^l$
0.3	LSDV	20	2.9976 (0.0058)	1.9994 (0.0050)	2.0028 (0.0054)	1.9989 (0.0344)	0.9941 (0.0299)	0.9976 (0.0360)
		50	3.0028 (0.0023)	2.0011 (0.0022)	1.9999 (0.0021)	2.0058 (0.0127)	1.0037 (0.0137)	1.0013 (0.0133)
	GEEs	20	2.9975 (0.0058)	1.9991 (0.0052)	2.0030 (0.0055)	1.9983 (0.0354)	0.9943 (0.0305)	0.9968 (0.0365)
		50	3.0029 (0.0023)	2.0013 (0.0022)	2.0000 (0.0021)	2.0053 (0.0128)	1.0039 (0.0137)	1.0014 (0.0135)
0.5	LSDV	20	3.0014 (0.0039)	2.0021 (0.0036)	1.9991 (0.0037)	1.9915 (0.0237)	1.0069 (0.0234)	0.9988 (0.0221)
		50	3.0014 (0.0015)	1.9979 (0.0013)	2.0001 (0.0015)	2.0030 (0.0090)	1.0004 (0.0095)	0.9961 (0.0094)
	GEEs	20	3.0014 (0.004)	2.0021 (0.0037)	1.9995 (0.0038)	1.9922 (0.0242)	1.0070 (0.0238)	0.9992 (0.0223)
		50	3.0014 (0.0015)	1.998 (0.0014)	1.9999 (0.0015)	2.0027 (0.0091)	1.0004 (0.0095)	0.9963 (0.0095)
0.9	LSDV	20	2.9997 (0.0007)	1.9984 (0.0008)	1.9998 (0.0007)	1.9985 (0.0045)	1.0015 (0.0049)	1.0021 (0.0044)
		50	3.0001 (0.0003)	1.9992 (0.0003)	2.0000 (0.0003)	2.0008 (0.0018)	1.0011 (0.0019)	1.0015 (0.0020)
	GEEs	20	2.9998 (0.0007)	1.9983 (0.0008)	1.9998 (0.0007)	1.9990 (0.0046)	1.0016 (0.0049)	1.0019 (0.0044)
		50	3.0001 (0.0003)	1.9993 (0.0003)	2.0000 (0.0003)	2.0007 (0.0018)	1.0010 (0.0019)	1.0016 (0.0020)

Table 7. Comparison for the GEEs and LSDV methods based on the dataset generated by the CRM model and exchangeable correlation matrix, with the average and standard deviation, where the working correlation matrix is AR(1).

N	α	Methods	RMSE			MAE			MMER			RI		
			CM	MM	CRM	CM	MM	CRM	CM	MM	CRM	CM	MM	CRM
20	0.3	LSDV	3.3003 (0.2606)	1.8590 (0.1071)	1.7703 (0.1015)	3.9815 (0.3689)	2.1006 (0.1285)	1.9990 (0.1220)	0.0979 (0.0100)	0.0491 (0.0033)	0.0467 (0.0031)	0.6698 (0.0265)	0.7662 (0.0161)	0.7786 (0.0150)
		GEEs	3.3002 (0.2613)	1.8594 (0.1073)	1.7708 (0.1016)	3.9813 (0.3701)	2.1010 (0.1286)	1.9995 (0.1221)	0.0979 (0.0101)	0.0491 (0.0033)	0.0467 (0.0031)	0.6698 (0.0265)	0.7662 (0.0162)	0.7786 (0.0150)
	0.5	LSDV	3.2303 (0.2741)	1.5992 (0.0927)	1.4939 (0.0853)	3.9284 (0.3860)	1.8089 (0.1120)	1.6887 (0.1032)	0.0964 (0.0103)	0.0422 (0.0028)	0.0394 (0.0026)	0.6738 (0.0293)	0.7920 (0.0160)	0.8068 (0.0144)
		GEEs	3.2296 (0.2760)	1.5996 (0.0926)	1.4944 (0.0853)	3.9275 (0.3886)	1.8094 (0.1119)	1.6894 (0.1033)	0.0964 (0.0104)	0.0422 (0.0028)	0.0394 (0.0026)	0.6739 (0.0295)	0.7920 (0.0160)	0.8067 (0.0144)
	0.9	LSDV	3.0836 (0.3005)	0.8751 (0.0506)	0.6681 (0.0391)	3.8407 (0.4224)	0.9933 (0.0631)	0.7553 (0.0469)	0.0943 (0.0108)	0.0232 (0.0016)	0.0177 (0.0012)	0.6806 (0.0354)	0.8738 (0.0134)	0.9041 (0.0099)
		GEEs	3.0839 (0.3005)	0.8753 (0.0506)	0.6682 (0.0390)	3.8411 (0.4225)	0.9935 (0.0631)	0.7555 (0.0468)	0.0943 (0.0108)	0.0232 (0.0016)	0.0177 (0.0012)	0.6806 (0.0354)	0.8738 (0.0134)	0.9041 (0.0099)
50	0.3	LSDV	3.3036 (0.1634)	1.8475 (0.0651)	1.7593 (0.0631)	3.9812 (0.2322)	2.0856 (0.0791)	1.9865 (0.0766)	0.0977 (0.0062)	0.0487 (0.0020)	0.0464 (0.0019)	0.6694 (0.0163)	0.7672 (0.0101)	0.7794 (0.0096)
		GEEs	3.3028 (0.1638)	1.8475 (0.0651)	1.7594 (0.0631)	3.9801 (0.2329)	2.0856 (0.0791)	1.9865 (0.0766)	0.0977 (0.0063)	0.0487 (0.0020)	0.0464 (0.0019)	0.6695 (0.0163)	0.7672 (0.0101)	0.7794 (0.0096)
	0.5	LSDV	3.2364 (0.1703)	1.5898 (0.0562)	1.4867 (0.0517)	3.9337 (0.2421)	1.7966 (0.0677)	1.6800 (0.0623)	0.0965 (0.0064)	0.0420 (0.0017)	0.0392 (0.0016)	0.6729 (0.0182)	0.7928 (0.0096)	0.8072 (0.0088)
		GEEs	3.2363 (0.1705)	1.5899 (0.0562)	1.4868 (0.0517)	3.9335 (0.2424)	1.7967 (0.0677)	1.6801 (0.0624)	0.0965 (0.0064)	0.0420 (0.0017)	0.0392 (0.0016)	0.6729 (0.0182)	0.7928 (0.0096)	0.8072 (0.0088)
	0.9	LSDV	3.0932 (0.1826)	0.8716 (0.0317)	0.6654 (0.0237)	3.8466 (0.2606)	0.9873 (0.0384)	0.7511 (0.0284)	0.0947 (0.0066)	0.0231 (0.0010)	0.0176 (0.0007)	0.6800 (0.0215)	0.8747 (0.0081)	0.9049 (0.0058)
		GEEs	3.0932 (0.1830)	0.8716 (0.0317)	0.6654 (0.0236)	3.8466 (0.2612)	0.9874 (0.0384)	0.7511 (0.0284)	0.0947 (0.0066)	0.0231 (0.0010)	0.0176 (0.0007)	0.6800 (0.0216)	0.8747 (0.0081)	0.9049 (0.0058)

Table 8. Comparison for the GEEs and LSDV methods based on the dataset generated by the min-max model and exchangeable correlation matrix, with the average and standard deviation, where the working correlation matrix is AR(1).

| N	α	Methods	RMSE			MAE			MMER			RI		
			CM	MM	CRM	CM	MM	CRM	CM	MM	CRM	CM	MM	CRM
20	0.3	LSDV	3.2503	1.2518	1.3449	4.1679	1.4152	1.5191	0.2110	0.1047	0.1536	0.8128	0.9376	0.9332
			(0.3808)	(0.0718)	(0.0775)	(0.5679)	(0.0866)	(0.0934)	(0.0233)	(0.0093)	(0.6604)	(0.0250)	(0.0039)	(0.0042)
		GEEs	3.2488	1.2521	1.3451	4.1655	1.4155	1.5194	0.2109	0.1048	0.1339	0.8129	0.9376	0.9332
			(0.3873)	(0.0718)	(0.0776)	(0.5774)	(0.0867)	(0.0935)	(0.0237)	(0.0093)	(0.1223)	(0.0254)	(0.0039)	(0.0043)
	0.5	LSDV	3.1975	1.0590	1.1665	4.1315	1.1990	1.3174	0.2100	0.0899	0.1329	0.8144	0.9468	0.9417
			(0.3736)	(0.0587)	(0.0666)	(0.5604)	(0.0711)	(0.0789)	(0.0239)	(0.0085)	(0.1922)	(0.0245)	(0.0032)	(0.0036)
		GEEs	3.1965	1.0592	1.1668	4.1298	1.1992	1.3176	0.2100	0.0899	0.1359	0.8145	0.9468	0.9417
			(0.3752)	(0.0588)	(0.0667)	(0.5628)	(0.0711)	(0.0789)	(0.0239)	(0.0084)	(0.2484)	(0.0246)	(0.0032)	(0.0036)
	0.9	LSDV	3.1294	0.4733	0.6814	4.1257	0.5347	0.7585	0.2106	0.0432	0.1035	0.8147	0.9758	0.9658
			(0.2887)	(0.0263)	(0.0400)	(0.4282)	(0.0319)	(0.0468)	(0.0206)	(0.0403)	(0.1589)	(0.0179)	(0.0015)	(0.0022)
		GEEs	3.1294	0.4734	0.6815	4.1257	0.5348	0.7586	0.2106	0.0436	0.4464	0.8147	0.9758	0.9658
			(0.2905)	(0.0263)	(0.0401)	(0.4308)	(0.0319)	(0.0468)	(0.0207)	(0.0525)	(10.0362)	(0.0180)	(0.0015)	(0.0022)
50	0.3	LSDV	3.2533	1.2431	1.3354	4.1718	1.4038	1.5071	0.2117	0.1040	0.1330	0.8127	0.9381	0.9338
			(0.2455)	(0.0455)	(0.0482)	(0.3687)	(0.0550)	(0.0581)	(0.0151)	(0.0061)	(0.0511)	(0.0162)	(0.0025)	(0.0026)
		GEEs	3.2526	1.2432	1.3355	4.1708	1.4039	1.5071	0.2117	0.1040	0.1309	0.8127	0.9381	0.9338
			(0.2478)	(0.0455)	(0.0482)	(0.3720)	(0.0550)	(0.0581)	(0.0152)	(0.0061)	(0.0257)	(0.0164)	(0.0025)	(0.0026)
	0.5	LSDV	3.2242	1.0513	1.1589	4.1695	1.1874	1.3074	0.2119	0.0894	0.1247	0.8129	0.9473	0.9422
			(0.2226)	(0.0365)	(0.0410)	(0.3352)	(0.0444)	(0.0494)	(0.0140)	(0.0064)	(0.0629)	(0.0148)	(0.0021)	(0.0023)
		GEEs	3.2235	1.0513	1.1590	4.1684	1.1875	1.3075	0.2119	0.0894	0.1271	0.8129	0.9473	0.9422
			(0.2237)	(0.0366)	(0.0411)	(0.3366)	(0.0444)	(0.0494)	(0.0141)	(0.0062)	(0.1286)	(0.0148)	(0.0021)	(0.0023)
	0.9	LSDV	3.1533	0.4700	0.6778	4.1561	0.5310	0.7543	0.2122	0.0502	0.1021	0.8136	0.9760	0.9660
			(0.1739)	(0.0165)	(0.0238)	(0.2577)	(0.0202)	(0.0275)	(0.0123)	(0.2603)	(0.1045)	(0.0108)	(0.0010)	(0.0013)
		GEEs	3.1534	0.4700	0.6778	4.1562	0.5310	0.7543	0.2122	0.0424	0.1438	0.8136	0.9760	0.9660
			(0.1744)	(0.0165)	(0.0238)	(0.2587)	(0.0202)	(0.0276)	(0.0123)	(0.0173)	(0.9041)	(0.0109)	(0.0010)	(0.0013)

Figure 1. The box plots of four evaluation indices from the GEEs and LSDV methods with the dataset generated by the CRM model and AR(1) correlation matrix, and the working correlation matrix specified correctly with $N = 20$, $\alpha = 0.9$.

Tables 1–4 show the fitting results when the working matrix is specified correctly, while Tables 5–8 show the fitting results when the working matrix is specified incorrectly. Figures 1–4 display the box plots of four evaluation indices calculated from the GEEs method in different models with the dataset generated based on the CRM model or the min-max model, when the working matrices are specified
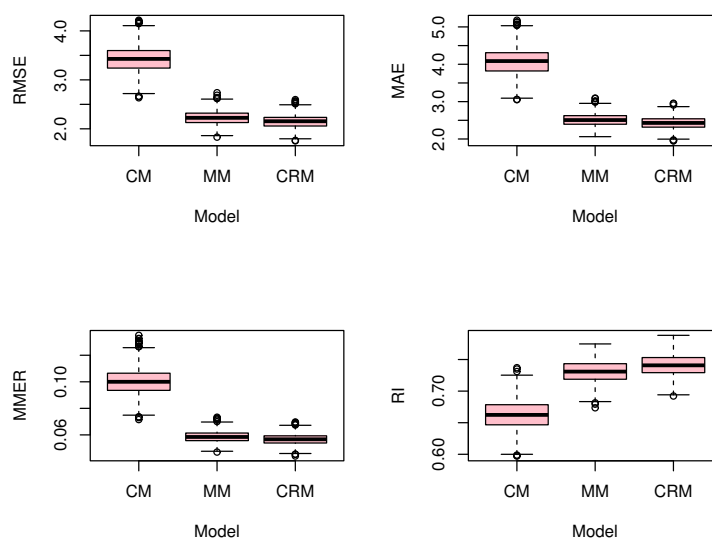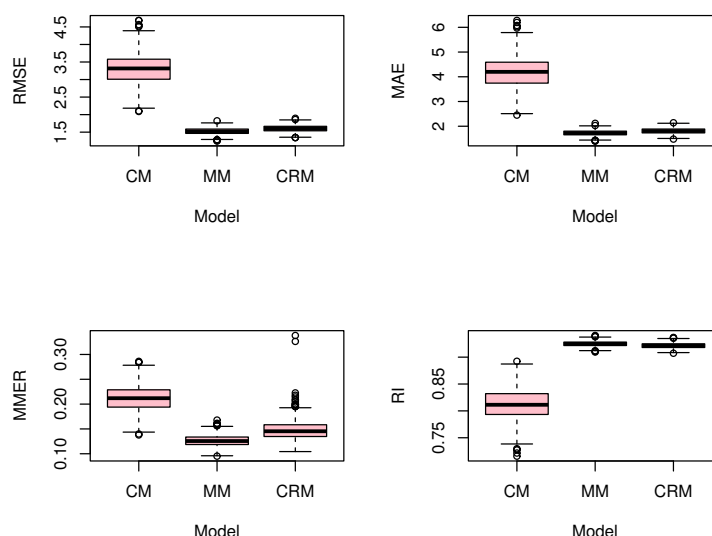
Figure 2. The box plots of four evaluation indices from the GEEs and LSDV methods with the dataset generated by the CRM model and AR(1) correlation matrix, and the working correlation matrix specified correctly with $N = 20$, $\alpha = 0.3$.
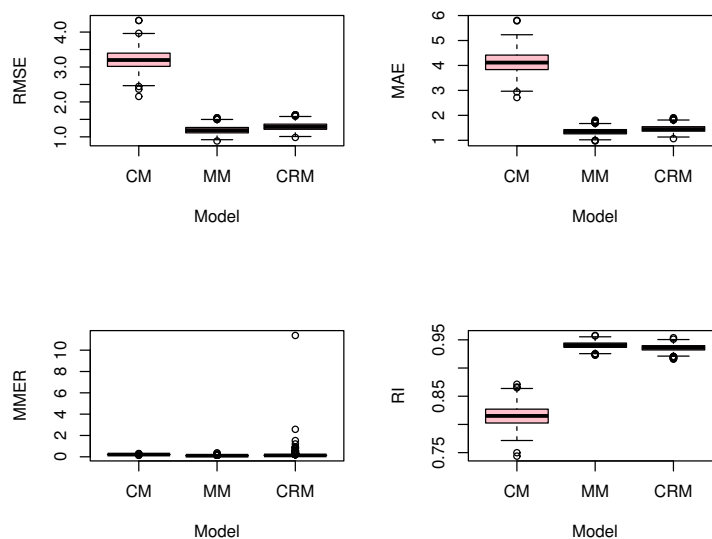
Figure 3. The box plots of four evaluation indices from the GEEs and LSDV methods with the dataset generated by the min-max model and AR(1) correlation matrix, and the working correlation matrix specified correctly with $N = 20$, $\alpha = 0.9$.

Figure 4. The box plots of four evaluation indices from the GEEs and LSDV methods with the dataset generated by the min-max model and AR(1) correlation matrix, and the working correlation matrix specified correctly with $N = 20$, $\alpha = 0.3$.

correctly. The results of the box plots are similar, when the working matrices are specified incorrectly.

- Tables 1 and 2 list the fitting results of β, and we have the following findings: (i) Regardless of the data generated by which model, the GEEs method performs better than the LSDV method. (ii) As α or N increases, the mean square error of GEEs estimation decreases. It means that the higher the degree of the correlation within the group, the better the performance of the GEEs. This also verifies the consistency of the estimate.
- Tables 3 and 4 show that (i) the GEEs method performs better than the LSDV method, regardless of the degree of correlation within the group of the data and the sample size, according to the four indicators. (ii) The CRM model performs best when data is generated based on centers and the radius. Similarly, the min-max model performs best when data is generated based on minimum and maximum values.
- From Tables 5–8, we can see that the fitting result of the GEEs method is close to that of the LSDV method when the working correlation matrix is incorrectly specified.
- Figures 1 and 2 display the box plots of four evaluation indices calculated from the GEEs method in different models with the dataset generated based on the CRM model, when the working matrices are specified correctly with $N = 20$, $\alpha = 0.9, 0.3$. We can see that the CRM model performs better than the other models.
- Figures 3 and 4 display the box plots of four evaluation indices calculated from the GEEs method in different models with the dataset generated based on the min-max model, when the working matrices are specified correctly with $N = 20$, $\alpha = 0.9, 0.3$. We can see that the min-max model performs better than the other models.

4.2. *Real data analysis*

As everyone knows, $PM_{2.5}$ is the main pollutant affecting human health and atmospheric environment quality. With the continuous development of social economy and the acceleration of the industrialization process, the problem of air pollution is also increasing. Studies have shown that $O_3, SO_2, NO_2, CO, PM_{10}$, five air pollutants, have a significant impact on $PM_{2.5}$ concentrations. Exploring the relationship between these air pollutants and $PM_{2.5}$ can provide help for effective prevention and control of air pollution. So, this section selects the 26 cities in China's Yangtze River Delta as the research object, and analyzes the interval-valued data of air pollution for them. The daily data of six air quality indices of the Yangtze River Delta urban agglomeration during the 59 days from January 1 to February 28, 2023, can be obtained from the China Air Quality Online Monitoring platform (https://www.aqistudy.cn/historydata/). Through inspection, there are no outliers or missing values in the data. Then we consider the maximum and minimum values of the daily data of each index as the interval-valued dataset.

The observed data of $PM_{2.5}$ are taken as response variables, and the other five indices are taken as independent variables. The data from the first 54 days are used as the training set, and the remaining 5 days are used as the test set. Substituting the data into the models is shown in Sections 2 and 3. Finally, we demonstrate the fitting performance of the GEEs method and the LSDV method when the working correlation matrix is AR(1). The predictive performance of the two methods are listed in Table 9, and the regression results of the GEEs method are shown in Table 10.

- Table 9 shows that both methods have good fitting effects on actual data in different models. Moreover, we can see that the GEEs method had better fitting performance than the LSDV method in different models. Taking into account the four performance indices, the min-max model performs better than other models, which may be due to the type of experimental data, which is also consistent with the conclusions in the Monte Carlo experiments.
- Table 10 shows that there is a positive correlation between PM_{10}, NO_2, CO, and $PM_{2.5}$, while there is a negative correlation between SO_2, O_3, and $PM_{2.5}$. Especially, CO has the greatest impact on $PM_{2.5}$.

Table 9. Comparison of GEEs and LSDV methods in each model for dealing with the panel interval-valued air pollutant dataset.

Models	Methods	RMSE	MAE	MMER	RI
CM	LSDV	7.20	7.24	0.47	0.69
	GEEs	7.09	7.14	0.27	0.69
Min-Max	LSDV	6.19	6.25	0.19	0.72
	GEEs	5.87	6.00	0.19	0.73
CRM	LSDV	6.13	6.32	0.80	0.73
	GEEs	6.06	6.25	0.71	0.74

Table 10. The estimated regression coefficients β and their robust standard deviation estimated by the GEEs method in different models, with their significance tests shown in parentheses ('***'-0.001,'**'-0.01,'*'-0.05,'.'-0.1).

Models	$\hat{\beta}$	$X_1(PM_{10})$	$X_2(SO_2)$	$X_3(NO_2)$	$X_4(CO)$	$X_5(O_3)$
CM	$\hat{\beta}_c$	0.5705***	-0.5830***	0.0592**	22.7856***	-0.0693***
		(0.0138)	(0.1450)	(0.0182)	(2.5756)	(0.0182)
CRM	$\hat{\beta}_c$	0.5705***	-0.5830***	0.0592**	22.7856***	-0.0693***
		(0.0138)	(0.1450)	(0.0182)	(2.5756)	(0.0182)
	$\hat{\beta}_r$	0.6979***	-0.2790***	0.0033	11.9899***	0.0129
		(0.0161)	(0.0778)	(0.0224)	(2.0010)	(0.0145)
Min-Max	$\hat{\beta}_l$	0.4252***	-0.0409	0.2601***	31.0804***	-0.0228*
		(0.0227)	(0.3231)	(0.0313)	(3.5584)	(0.0096)
	$\hat{\beta}_u$	0.6516***	-0.4963***	0.0295.	16.9246***	-0.0514***
		(0.0116)	(0.1088)	(0.0173)	(2.1394)	(0.0130)

5. Conclusions and discussion

In this paper, we proposed an innovative approach which was to apply GEEs to fixed effects panel interval-valued data models, in which the correlation within a group is considered. Monte Carlo simulation experiments test the feasibility of this method and our proposed method was applied to $PM2.5$ forecasting. Experiments show that the proposed method has a better performance than the LSDV method proposed before.

However, the estimation method we proposed has some limitations. For example, the model can be extended to the nonlinear regression model to capture complex nonlinear relationships between variables. In addition, we did not take into account the correlation between the bounds of the panel interval-valued data. For future studies, we will introduce variance modeling or joint modeling to obtain better estimators.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. E. Diday, The symbolic approach in clustering and related methods of data analysis, classification and related methods of data analysis, in *Proceedings of the first Conference of the Federation of the classification societies. North Holland*, 1988.

2. L. Billard, E. Diday, Regression analysis for interval-valued data, in *Data Analysis, Classification, and Related Methods*, Springer, (2000), 369–374.

3. L. Billard, E. Diday, Symbolic regression analysis, in *Classification, Clustering, and Data Analysis: Recent Advances and Applications*, Springer, (2002), 281–288.

4. E. d. A. L. Neto, F. D. A. De Carvalho, Centre and range method for fitting a linear regression model to symbolic interval data, *Comput. Stat. Data Anal.*, **52** (2008), 1500–1515. https://doi.org/10.1016/j.csda.2007.04.014

5. L. Kong, X. Gao, A regularized MM estimate for interval-valued regression, *Expert Syst. Appl.*, **238** (2024), 122044. https://doi.org/10.1016/j.eswa.2023.122044

6. M. Xu, Z. Qin, A bivariate bayesian method for interval-valued regression models, *Knowl.-Based Syst.*, **235** (2022), 107396. https://doi.org/10.1016/j.knosys.2021.107396

7. U. Beyaztas, H. L. Shang, A. S. G. Abdel-Salam, Functional linear models for interval-valued data, *Commun. Stat.-Simul. Comput.*, **51** (2022), 3513–3532. https://doi.org/10.1080/03610918.2020.1714662

8. Q. Zhao, H. Wang, S. Wang, Robust regression for interval-valued data based on midpoints and log-ranges, *Adv. Data Anal. Classif.*, **17** (2023), 583–621. https://doi.org/10.1007/s11634-022-00518-2

9. L. C. Lin, H. L. Chien, S. Lee, Symbolic interval-valued data analysis for time series based on auto-interval-regressive models, *Stat. Method. Appl.*, **30** (2021), 295–315. https://doi.org/10.1007/s10260-020-00525-7

10. J. Zhang, M. Liu, M. Dong, Variational bayesian inference for interval regression with an asymmetric laplace distribution, *Neurocomputing*, **323** (2019), 214–230. https://doi.org/10.1016/j.neucom.2018.09.083

11. C. Yang, Interval riccati equation-based and non-probabilistic dynamic reliability-constrained multi-objective optimal vibration control with multi-source uncertainties, *J. Sound Vibr.*, **595** (2025), 118742. https://doi.org/10.1016/j.jsv.2024.118742

12. C. Yang, Y. Liu, H. Gao, Reliability-constrained uncertain spacecraft sliding mode attitude tracking control with interval parameters, *IEEE Trans. Aerosp. Electron. Syst.*, **61** (2025), 1–14. https://doi.org/10.1109/TAES.2025.3529798

13. A. F. Galvao Jr, Quantile regression for dynamic panel data with fixed effects, *J. Econom.*, **164** (2011), 142–157. https://doi.org/10.1016/j.jeconom.2011.02.016

14. E. Aristodemou, Semiparametric identification in panel data discrete response models, *J. Econom.*, **220** (2021), 253–271. https://doi.org/10.1016/j.jeconom.2020.04.002

15. B. H. Beyaztas, S. Bandyopadhyay, Robust estimation for linear panel data models, *Stat. Med.*, **39** (2020), 4421–4438. https://doi.org/10.1002/sim.8732

16. J. Bai, S. H. Choi, Y. Liao, Feasible generalized least squares for panel data with cross-sectional and serial correlations, *Empir. Econ.*, **60** (2021), 309–326. https://doi.org/10.1007/s00181-020-01977-2

17. B. H. Baltagi, D. Li, Prediction in the panel data model with spatial correlation, in *Advances in Spatial Econometrics: Methodology, Tools and Applications*, Springer, (2004), 283–295.

18. J. P. Elhorst, Spatial panel data models, in *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*, Springer, (2014), 37–93.

19. Y. Zhang, J. Jiang, Y. Feng, Penalized quantile regression for spatial panel data with fixed effects, *Commun. Stat.-Theory Methods*, **52** (2023), 1287–1299. https://doi.org/10.1080/03610926.2021.1934028

20. A. Ji, J. Zhang, X. He, Y. Zhang, Fixed effects panel interval-valued data models and applications, *Knowl.-Based Syst.*, **237** (2022), 107798. https://doi.org/10.1016/j.knosys.2021.107798

21. J. Zhang, Q. Li, B. Wei, A. Ji, Robust estimation method for panel interval-valued data model with fixed effects, *J. Stat. Comput. Simul.*, **94** (2024), 3230–3251. https://doi.org/10.1080/00949655.2024.2381491

22. A. Ji, J. Zhang, Y. Cao, Bivariate maximum likelihood method for fixed effects panel interval-valued data models, *Comput. Econ.*, **64** (2024), 1–28. https://doi.org/10.1007/s10614-024-10737-8

23. K. Y. Liang, S. L. Zeger, Longitudinal data analysis using generalized linear models, *Biometrika*, **73** (1986), 13–22. https://doi.org/10.1093/biomet/73.1.13

24. M. Crowder, On consistency and inconsistency of estimating equations, *Econom. Theory*, **2** (1986), 305–330. https://doi.org/10.1017/S0266466600011646

25. R. M. Balan, I. S. Kratina, Asymptotic results with generalized estimating equations for longitudinal data, *Ann. Stat.*, **33** (2005), 522–541. https://doi.org/10.1214/009053604000001255

26. L. Wang, Gee analysis of clustered binary data with diverging number of covariates, *Ann. Stat.*, **39** (2011), 389–417. https://doi.org/10.1214/10-AOS846

27. Y. G. Wang, V. Carey, Working correlation structure misspecification, estimation and covariate design: implications for generalised estimating equations performance, *Biometrika*, **90** (2003), 29–41. https://doi.org/10.1093/biomet/90.1.29

28. W. Pan, Model selection in estimating equations, *Biometrics*, **57** (2001), 529–534. https://doi.org/10.1111/j.0006-341X.2001.00529.x

29. M. Wang, Generalized estimating equations in longitudinal data analysis: a review and recent developments, *Adv. Stat.*, **2014** (2014), 303728. https://doi.org/10.1155/2014/303728

30. S. Seaman, A. Copas, Doubly robust generalized estimating equations for longitudinal data, *Stat. Med.*, **28** (2009), 937–955. https://doi.org/10.1002/sim.3520

31. N. R. Chaganty, An alternative approach to the analysis of longitudinal data via generalized estimating equations, *J. Stat. Plan. Infer.*, **63** (1997), 39–54. https://doi.org/10.1016/S0378-3758(96)00203-0

32. N. R. Chaganty, J. Shults, On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter, *J. Stat. Plan. Infer.*, **76** (1999), 145–161. https://doi.org/10.1016/S0378-3758(98)00180-3

33. C. Hu, L. T. He, An application of interval methods to stock market forecasting, *Reliab. Comput.*, **13** (2007), 423–434. https://doi.org/10.1007/s11155-007-9039-4