Statistical Methods & Applications
https://doi.org/10.1007/s10260-023-00699-w

ORIGINAL PAPER

Analysis of correlated unit-Lindley data based on estimating equations

Danilo V. Silva[1] · **Hatice Tul Kubra Akdur**[2] · **Gilberto A. Paula**[1]

Accepted: 10 April 2023

Abstract
In this paper we derive estimating equations for modeling unbalanced correlated data sets in which the marginal distributions follow the one parameter unit-Lindley distributions with domain on the interval (0,1). A class of regressions models is proposed for modeling the location parameter and a reweighted iterative process is developed for the joint estimation of the regression coefficients and the correlation structure. Simulation studies are performed to assess the empirical properties of the derived estimators and diagnostic procedures, such as residual analysis and sensitivity studies based on conformal local influence are given. Finally, we analyze the proportion of people in households with inadequate water supply and sewage within federation units of Brazil by the procedures developed in the paper.

Keywords Unit-Lindley distribution · Correlated data · Diagnostic procedures · Estimating equations

1 Introduction

Unit-interval distributions such as beta, simplex and Kumaraswamy distributions are widely known in the statistical literature (see, for instance, Ferrari and Cribari-Neto 2004; Barndorff-Nielsen and Jørgensen 1991; Kumaraswamy 1980) and recent literature includes new distributions for modeling the unit-interval outcomes. For example, Mazucheli et al. (2019) introduced the unit-Lindley distribution along with its regression model as an alternative to the beta distribution using the transformation on the cumulative distribution function of the Lindley distribution (Ghitany et al. 2008) and more recently Altun et al. (2021) presented the log-Bilal distribution using the

Danilo V. Silva, Hatice T. K. Akdur and Gilberto A. Paula have contributed equally to this work.

✉ Gilberto A. Paula
giapaula@ime.usp.br

[1] Department of Statistics, Universidade de São Paulo, São Paulo, Brazil

[2] Department of Statistics, Faculty of Science, Gazi University, Ankara, Turkey

Published online: 16 May 2023

transformation on the Bilal distribution (Abd-Elrahman 2013). Even though Grassia (1977) presented unit-gamma distribution, it was not used as a distribution for the continuous unit-interval outcomes until Mousa et al. (2016) provided its regression model. For independent distributed proportion outcomes, beta, simplex, unit-gamma, unit-Lindley and log-Bilal regression models can be preferred depending on the suitability of the data set. However, repeated measures design, longitudinal clinical trials, cluster sample designs induce multilevel data structures which are not appropriate for analyzing under these regression models.

An example for longitudinal proportional outcomes may be the percent of gas left in the eye from an ophthalmology study, which was analyzed with simplex mixed-effect models (Qiu et al. 2008). Multilevel proportional outcomes such as proportion of diseased tooth sites from a dental study and water quality index are analyzed, respectively, by beta-mixed effect models and quasi-beta longitudinal models by Galvis et al. (2014) and Petterle et al. (2019). Recently, Akdur (2021) has developed unit-Lindley mixed-effect models.

The aim of this paper is to propose an alternative approach for modeling unbalanced correlated unit-Lindley data sets based on estimating equations. From the optimum class of estimating functions proposed by Crowder (1987), we derive an optimum class of estimating equations for modeling correlated data in which the marginal distributions are assumed to follow unit-Lindley distributions on the interval (0,1). The estimating equations and the asymptotic properties of the former estimators are based on the theory developed by Godambe (1997), with the assumption that the within experimental unit correlations follow the same structure of the generalized estimating equations (GEE) proposed by Liang and Zeger (1986). A reweighted iterative process is developed for the parameter estimation and the asymptotic and empirical properties of the derived estimators are discussed. Diagnostic procedures are proposed as well as an application with a real data set is given for illustration.

The paper is organized as follows. In Sect. 2, a brief review on the unit-Lindley distribution on the interval (0,1) is given and a class of regression models is proposed for modeling correlated rates and proportions. An optimum class of estimating functions for the regression coefficients is derived in Sect. 3 as well as a joint iterative process for the parameter estimation is presented with some discussions on the asymptotic properties of the former estimators. Simulation studies to assess the empirical properties of the estimators are performed in Sect. 4 and diagnostic procedures, such residual analysis based on marginal quantile residual and sensitivity studies based on conformal local influence are proposed in Sect. 5. An application with a real data set is presented in Sect. 6 for illustrating the methodology developed in the paper. Section 7 deals with some conclusions and some technical results are presented in Appendices A-C.

2 Unit-Lindley distribution

Denote a random variable y distributed as the unit-Lindley distribution indexed by the parameter $0 < \mu < 1$ for fitting rates and proportions, whose probability density function and the cumulative density function may be written, respectively, as

$$f(y;\mu) = \frac{(1-\mu)^2}{\mu(1-y)^3} \exp\left\{-\frac{y(1-\mu)}{\mu(1-y)}\right\},$$

and

$$F(y;\mu) = 1 - \left(\frac{1-y\mu}{1-y}\right) \exp\left\{-\frac{(1-\mu)y}{(1-y)\mu}\right\},$$

where $0 < y < 1$, with $\mathrm{E}(y) = \mu$. The variance function may be expressed as

$$\mathrm{Var}(y) = \frac{(1-\mu)^2}{\mu}\left[\mathrm{E}_1\left(\frac{1-\mu}{\mu}\right)\exp\left(\frac{1-\mu}{\mu}\right) - \mu\right],$$

whereas $\mathrm{E}_n(x) = \int_1^\infty t^{-n}\exp(-xt)dt$ denotes the exponential integral function, for $n \in \{0, 1, \ldots\}$ and $x \in \mathbb{R}$, with $\mathrm{E}_1(x) = \int_1^\infty t^{-1}\exp(-xt)dt$. Also, the pth quantile takes the form

$$F^{-1}(p;\mu) = \frac{\frac{1}{\mu} + \mathrm{W}_{-1}\left\{\frac{p-1}{\mu\exp(\mu^{-1})}\right\}}{1 + \mathrm{W}_{-1}\left\{\frac{p-1}{\mu\exp(\mu^{-1})}\right\}},$$

where $0 < p < 1$ and $\mathrm{W}_{-1}(a)$ denotes the negative branch of the Lambert-W function, for $a \in [-e^{-1}, 0)$. Various additional properties of unit-Lindley distribution may be found in Mazucheli et al. (2019). We will denote along the paper $y \sim \mathrm{UL}(\mu)$ for a random variable distributed as unit-Lindley. Figure 1 describes the forms of the probability density function and the variance of y.

In the sequel we will present some key results related with the score function for the parameter μ, necessary for writing the estimating functions for modeling correlated data with marginal UL distributions.

The log-likelihood function for μ takes the form

$$\mathrm{L}(\mu) = 2\log\left(\frac{1-\mu}{\mu}\right) - \frac{y(1-\mu)}{\mu(1-y)} - 3\log(1-y),$$

and consequently the score function for μ is given by

$$u = \frac{d\mathrm{L}(\mu)}{d\mu} = \frac{z}{\mu^2} - \frac{1+\mu}{\mu(1-\mu)},$$

where $z = y/(1-y)$ denotes the observed odds. Note that, for μ fixed, z is a monotonic function of y.

Consider satisfied the usual regularity conditions, namely: (1) $\mathrm{E}(u) = 0$ and (2) $\mathrm{E}(-u') = \mathrm{E}(u^2)$, where $u' = du/d\mu$. After some algebraic manipulations (see Appendix A) we obtain

$$\mathrm{Var}(u) = \frac{2-(1-\mu)^2}{\mu^2(1-\mu)^2},\ \mathrm{E}(z) = \frac{\mu(1+\mu)}{1-\mu} \text{ and } \mathrm{Var}(z) = \frac{\mu^2\{2-(1-\mu)^2\}}{(1-\mu)^2}.$$

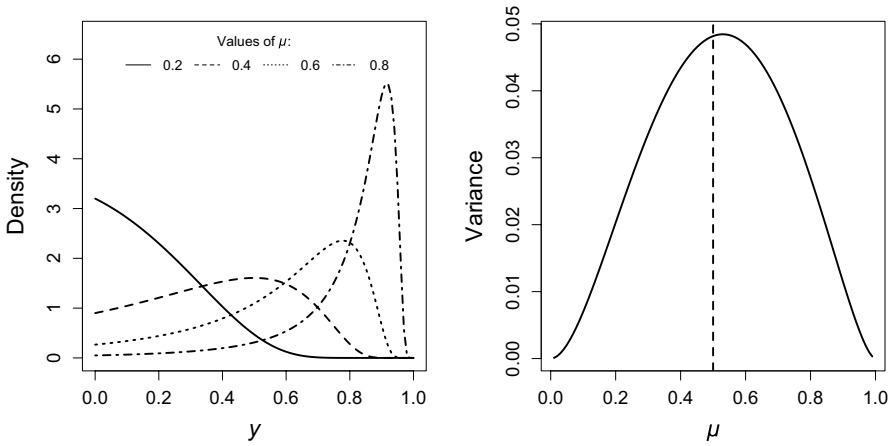

Fig. 1 Forms of the UL distribution. The left panel shows the probability density function for some μ values. The right panel shows the behavior of the variance function

It may be showed that $z \sim \text{Lindley}(\mu^{-1} - 1)$ belongs to the one-parameter exponential family of distributions $f(z;\theta) = \exp\{\theta z - b(\theta) + c(z)\}$ (see Appendix B). So, all the theory developed for generalized linear models (McCullagh and Nelder 1989) may be applied for modeling $\text{E}(z)$. However, our interest in this paper is modeling the parameter μ for appropriate link functions and correlated data.

2.1 UL-GEE models

Let s_i a set of indexes relative to the instants in which the response was observed in the ith experimental unit, with cardinality denoted by $n(s_i)$. Then $\mathbf{y}_i^\top = \{y_{ij} : j \in s_i\}$ is an $n(s_i) \times 1$ vector containing responses (rates or proportions), for $i = 1, \ldots, n$. We will assume that $y_{ij} \sim \text{UL}(\mu_{ij})$ with regression structure $g(\mu_{ij}) = \eta_{ij} = \mathbf{x}_{ij}^\top \boldsymbol{\beta}$, where $g(\cdot)$ denotes a link function with domain on (0, 1) and differentiable, $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp})^\top$ contains values of explanatory variables and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is the regression coefficient vector. The within experimental unit correlation matrix will be represented by the $n(s_i) \times n(s_i)$ matrix $\mathbf{R}(\mathbf{u}_i)$, where $\mathbf{u}_i^\top = \{u_{ij} : j \in s_i\}$ and

$$u_{ij} = \frac{z_{ij}}{\mu_{ij}^2} - \frac{1 + \mu_{ij}}{\mu_{ij}(1 - \mu_{ij})},$$

with $z_{ij} = y_{ij}/(1 - y_{ij})$, for $i = 1, \ldots, n$. For $n(s_i) > 1$, the (j, j')th element of the matrix $\mathbf{R}(\mathbf{u}_i)$ may be denoted as

$$\text{R}_{jj'}(\mathbf{u}_i) = \frac{\text{E}(u_{ij}u_{ij'})}{\sqrt{\text{Var}(u_{ij})}\sqrt{\text{Var}(u_{ij'})}},$$

for $j \neq j'$, whereas $\text{R}_{jj'}(\mathbf{u}_i) = 1$, for $j = j'$, with

$$\mathrm{Var}(u_{ij}) = \frac{2-(1-\mu_{ij})^2}{\mu_{ij}^2(1-\mu_{ij})^2},$$

for $i = 1, \ldots, n$. Since u_{ij} is a linear combination of z_{ij}, then $\mathbf{R}(\mathbf{u}_i)$ agrees with the correlation matrix $\mathbf{R}(\mathbf{z}_i)$, where $\mathbf{z}_i^\top = \{z_{ij} : j \in s_i\}$, and due to the monotonic relationship between z_{ij} and y_{ij}, it is expected a good agreement between the correlation matrices $\mathbf{R}(\mathbf{z}_i)$ and $\mathbf{R}(\mathbf{y}_i)$.

Based on the theory developed by Godambe (1997), we will derive in the next section estimating functions for $\boldsymbol{\beta}$ and a parallel to the Fisher scoring algorithm for estimating $\boldsymbol{\beta}$ by assuming that the matrix $\mathbf{R}(\mathbf{u}_i)$ is replaced by some structured correlation matrix (working correlation matrix) that does not depend on $\boldsymbol{\beta}$. Asymptotic properties of the former estimator for $\boldsymbol{\beta}$ will also be presented. We will named this class as UL-GEE models.

3 Estimating functions

The well-known generalized estimating equations (GEE) (Liang and Zeger 1986) may be extended for modeling rates and proportions with marginal UL distributions and some dependence among the responses within experimental unit represented by a working correlation matrix. In general, such estimating equations may be derived as a particular case from the optimum estimating function class proposed by Crowder (1987) and defined as

$$\boldsymbol{\Psi}^*(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathrm{E}\left(\frac{\partial \mathbf{u}_i}{\partial \boldsymbol{\beta}^\top}\right)^\top \mathrm{Cov}(\mathbf{u}_i)^{-1}\mathbf{u}_i.$$

After some algebraic manipulation (see Tsuyuguchi et al. (2020), Equation 2) we obtain

$$\boldsymbol{\Psi}^*(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{X}_i^\top \mathbf{D}_i \mathrm{Cov}(\mathbf{u}_i)^{-1}\mathbf{u}_i, \tag{1}$$

where $\mathbf{X}_i$ is an $n(s_i) \times p$ matrix of rows $\mathbf{x}_{ij}^\top$, $\mathbf{D}_i = \mathrm{diag}\{d_{ij} : j \in s_i\}$ with $d_{ij} = -\mathrm{Var}(u_{ij})\{g'(\mu_{ij})\}^{-1}$ and $\mathrm{Cov}(\mathbf{u}_i)$ denotes the variance-covariance matrix of $\mathbf{u}_i$, for $i = 1, \ldots, n$.

Expressing $\mathrm{Cov}(\mathbf{u}_i) = \boldsymbol{\Sigma}_{u_i}^{\frac{1}{2}}\mathbf{R}(\mathbf{u}_i)\boldsymbol{\Sigma}_{u_i}^{\frac{1}{2}}$, where $\boldsymbol{\Sigma}_{u_i} = \mathrm{diag}\{\mathrm{Var}(u_{ij}) : j \in s_i\}$, the idea here is to replace the correlation matrix $\mathbf{R}(\mathbf{u}_i)$ by a working correlation matrix $\mathbf{R}_i(\boldsymbol{\rho})$, that depends only on the correlation vector $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_q)^\top$, which does not depend on $\boldsymbol{\beta}$. Thus, the estimating function (1) assumes the alternative form

$$\boldsymbol{\Psi}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{X}_i^\top \mathbf{D}_i \boldsymbol{\Omega}_i^{-1}\mathbf{u}_i, \tag{2}$$

where $\boldsymbol{\Omega}_i = \boldsymbol{\Sigma}_{u_i}^{\frac{1}{2}}\mathbf{R}_i(\boldsymbol{\rho})\boldsymbol{\Sigma}_{u_i}^{\frac{1}{2}}$, for $i = 1, \ldots, n$.

Since $\mathrm{E}(u_{ij}) = 0$, $\forall ij$, one has a unbiased estimating function, $\mathrm{E}\{\boldsymbol{\Psi}(\boldsymbol{\beta})\} = \mathbf{0}$. From Godambe (1997) the variability matrix of $\boldsymbol{\Psi}(\boldsymbol{\beta})$ is defined as

$\mathbf{V}_{n\Psi}(\boldsymbol{\beta}) = \mathrm{E}\{\boldsymbol{\Psi}(\boldsymbol{\beta})\boldsymbol{\Psi}^{\top}(\boldsymbol{\beta})\}$, and the respective sensitivity matrix of $\boldsymbol{\Psi}(\boldsymbol{\beta})$ is given by $\mathbf{S}_{n\Psi}(\boldsymbol{\theta}) = \mathrm{E}\{\boldsymbol{\Psi}'(\boldsymbol{\beta})\}$. The Godambe information matrix of $\boldsymbol{\beta}$ is a regular estimating function defined as

$$\mathbf{J}_{n\Psi}(\boldsymbol{\beta}) = \mathbf{S}_{n\Psi}(\boldsymbol{\beta})^{\top}\mathbf{V}_{n\Psi}^{-1}(\boldsymbol{\beta})\mathbf{S}_{n\Psi}(\boldsymbol{\theta}),$$

where $\mathbf{V}_{n\Psi}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\mathbf{V}_i(\boldsymbol{\beta})$ with $\mathbf{V}_i(\boldsymbol{\beta}) = \mathbf{X}_i^{\top}\mathbf{W}_i\mathbf{D}_i^{-1}\mathrm{Cov}(\mathbf{u}_i)\mathbf{D}_i^{-1}\mathbf{W}_i\mathbf{X}_i$, $\mathbf{S}_{n\Psi}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\mathbf{S}_i(\boldsymbol{\beta})$ with $\mathbf{S}_i(\boldsymbol{\beta}) = \mathbf{X}_i^{\top}\mathbf{W}_i\mathbf{X}_i$ and $\mathbf{W}_i = \mathbf{D}_i\boldsymbol{\Omega}_i^{-1}\mathbf{D}_i$, for $i = 1, \ldots, n$ (see derivation in Appendix C).

In the next section a reweighted iterative process will be derived for solving $\boldsymbol{\Psi}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ from (2), and some asymptotic properties of the former estimator $\hat{\boldsymbol{\beta}}$ will be presented.

3.1 Iterative process

Similarly to Tsuyuguchi et al. (2020) we will apply the Newton scoring method, that is a parallel to the Fisher scoring method (see, for instance, Jorgensen et al. 1996), for obtaining the estimate $\hat{\boldsymbol{\beta}}$, in which $\boldsymbol{\Psi}'(\boldsymbol{\beta})$ is replaced by its expectation $\mathrm{E}\{\boldsymbol{\Psi}'(\boldsymbol{\beta})\} = \sum_{i=1}^{n}\mathbf{X}_i^{\top}\mathbf{W}_i\mathbf{X}_i$. An advantage of this method is the existence at each step of the iterative process of the inverse $[\mathrm{E}\{\boldsymbol{\Psi}'(\boldsymbol{\beta}^{(m)})\}]^{-1}$, since each $\mathbf{X}_i$ has full column rank. Then, fixing $\boldsymbol{\rho}$, we obtain the following iterative process:

$$\begin{aligned}\boldsymbol{\beta}^{(m+1)} &= \boldsymbol{\beta}^{(m)} - [\mathrm{E}\{\boldsymbol{\Psi}'(\boldsymbol{\beta}^{(m)})\}]^{-1}\boldsymbol{\Psi}(\boldsymbol{\beta}^{(m)}) \\ &= \boldsymbol{\beta}^{(m)} - \left\{\sum_{i=1}^{n}\mathbf{X}_i^{\top}\mathbf{W}_i^{(m)}\mathbf{X}_i\right\}^{-1}\left\{\sum_{i=1}^{n}\mathbf{X}_i^{\top}\mathbf{W}_i^{(m)}(\mathbf{D}_i^{(m)})^{-1}\mathbf{u}_i^{(m)}\right\},\end{aligned} \tag{3}$$

for $m = 0, 1, 2\ldots$. The iterative process (3) may be expressed as the following reweighted iterative process:

$$\boldsymbol{\beta}^{(m+1)} = \left\{\sum_{i=1}^{n}\mathbf{X}_i^{\top}\mathbf{W}_i^{(m)}\mathbf{X}_i\right\}^{-1}\left\{\sum_{i=1}^{n}\mathbf{X}_i^{\top}\mathbf{W}_i^{(m)}\mathbf{t}_i^{(m)}\right\}, \tag{4}$$

for $m = 0, 1, 2\ldots$, where $\mathbf{t}_i = \mathbf{X}_i\boldsymbol{\beta} - \mathbf{D}_i^{-1}\mathbf{u}_i$ is a modified dependent variable, for $i = 1, \ldots, n$. Below we describe moments estimators for $\boldsymbol{\rho}$ by fixing $\boldsymbol{\beta}$, for some usual correlation structures such that $n(s_i) > 1$.

1. I*ndependent*: In this case one has $\mathbf{R}_i(\boldsymbol{\rho}) = \mathbf{I}_{n(s_i)}$, where $\mathbf{I}_{n(s_i)}$ denotes the identity matrix of order $n(s_i)$.
2. *Unstructured*: Here the correlation matrix $\mathbf{R}_i(\boldsymbol{\rho})$ is unstructured and one has $n(s_i)\{n(s_i) - 1\}/2$ parameters to be estimated for each group. Let the set $A_{jj'} = \{i : j, j' \in s_i,\ i = 1, \ldots, n,\ j \neq j'\}$, denoting $\mathbf{R}_i = \{\rho_{ijj'}\}$, the (j, j')-th element of $\mathbf{R}_i$ may be estimated by

$$\hat{\rho}_{jj'} = \frac{1}{n(A_{jj'})} \sum_{i \in A_{jj'}} \frac{\hat{u}_{ij}}{\sqrt{\widehat{\mathrm{Var}}(u_{ij})}} \frac{\hat{u}_{ij'}}{\sqrt{\widehat{\mathrm{Var}}(u_{ij'})}}.$$

3. *Exchangeable*: In this case $\mathbf{R}_i = \mathbf{R}_i(\rho)$, where the (j, j')-th element of $\mathbf{R}_i$ becomes given by $\mathrm{R}_{ijj'} = 1$, for $j = j'$, and $\mathrm{R}_{ijj'} = \rho$, for $j \neq j'$. A consistent estimator for ρ may expressed as

$$\hat{\rho} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{n(s_i)\{n(s_i) - 1\}} \sum_{j \in s_i} \sum_{\substack{j' \in s_i \\ j' \neq j}} \frac{\hat{u}_{ij}}{\sqrt{\widehat{\mathrm{Var}}(u_{ij})}} \frac{\hat{u}_{ij'}}{\sqrt{\widehat{\mathrm{Var}}(u_{ij'})}}.$$

4. *First-order autoregressive*: Here we assume $\mathbf{R}_i = \mathbf{R}_i(\rho)$, where the (j, j')-th element of $\mathbf{R}_i$ becomes given by $\mathrm{R}_{ijj'} = 1$, for $j = j'$, and $\mathrm{R}_{ijj'} = \rho^{|j-j'|}$, for $j \neq j'$. Let the set $A_j = \{i : j, j+1 \in s_i,\ i = 1, \ldots, n\}$ and $B = \bigcap_{i=1}^{n} s_i$, a consistent estimator for ρ may be expressed as

$$\hat{\rho} = \frac{1}{\{n(B) - 1\}} \sum_{j \in B} \frac{1}{n(A_j)} \sum_{i \in A_j} \frac{\hat{u}_{ij}}{\sqrt{\widehat{\mathrm{Var}}(u_{ij})}} \frac{\hat{u}_{i(j+1)}}{\sqrt{\widehat{\mathrm{Var}}(u_{i(j+1)})}}.$$

Thus, denoting $\mathbf{X} = [\mathbf{X}_1^\top, \ldots, \mathbf{X}_n^\top]^\top$ an $N \times p$ matrix of rows $\mathbf{X}_i$ and $g(\mathbf{y}) = [g(\mathbf{y}_1^\top), \ldots, g(\mathbf{y}_n^\top)]^\top$ an $N \times 1$ vector of rows $g(\mathbf{y}_i) = \{g(y_{ij}) : j \in s_i\}^\top$, where $N = \sum_{i=1}^{n} n(s_i)$, we propose the following Algorithm 1 to obtain the parameter estimates:

Algorithm 1 Parameter estimates

Inputs:
  $\mathbf{X}$ full column rank
Initialize:
  $\boldsymbol{\beta}^{(0)} = (\mathbf{X}^\top \mathbf{X})^{-1} g(\mathbf{y})$
repeat
  update $\boldsymbol{\rho}$ from some 1-4 fixed structure
  update $\boldsymbol{\beta}$ from (4)
until the convergence
return $\boldsymbol{\beta}$

3.2 Inference

From Artes and Jorgensen (2000) one has that $\hat{\boldsymbol{\beta}}$, obtained from the iterative process (4), is such as

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} \mathrm{N}_p(\mathbf{0}, \mathbf{J}_\Psi^{-1}(\boldsymbol{\beta})),$$

where $\mathbf{J}_\Psi(\boldsymbol{\beta}) = \lim_{n \to \infty} n^{-1} \mathbf{J}_{n\Psi}(\boldsymbol{\beta})$, with

$$\mathbf{J}_{n\Psi}(\boldsymbol{\beta}) = \left\{\sum_{i=1}^{n}\mathbf{S}_i(\boldsymbol{\beta})\right\}\left\{\sum_{i=1}^{n}\mathbf{V}_i(\boldsymbol{\beta})\right\}^{-1}\left\{\sum_{i=1}^{n}\mathbf{S}_i(\boldsymbol{\beta})\right\}.$$

A consistent estimator of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by

$$\{\widehat{\mathbf{J}}_{\Psi}(\boldsymbol{\beta})\}^{-1} = \left\{\sum_{i=1}^{n}\widehat{\mathbf{S}}_i(\boldsymbol{\beta})\right\}^{-1}\left\{\sum_{i=1}^{n}\mathbf{X}_i^{\top}\widehat{\mathbf{D}}_i\widehat{\boldsymbol{\Omega}}_i^{-1}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i^{\top}\widehat{\boldsymbol{\Omega}}_i^{-1}\widehat{\mathbf{D}}_i\mathbf{X}_i\right\}\left\{\sum_{i=1}^{n}\widehat{\mathbf{S}}_i(\boldsymbol{\beta})\right\}^{-1},$$

with $\widehat{\mathbf{S}}_i(\boldsymbol{\beta}) = \mathbf{X}_i^{\top}\widehat{\mathbf{W}}_i\mathbf{X}_i$, where $\widehat{\mathbf{W}}_i = \widehat{\mathbf{D}}_i\widehat{\boldsymbol{\Omega}}_i^{-1}\widehat{\mathbf{D}}_i$, $\widehat{\boldsymbol{\Omega}}_i = \widehat{\boldsymbol{\Sigma}}_{u_i}^{\frac{1}{2}}\mathbf{R}_i(\hat{\rho})\widehat{\boldsymbol{\Sigma}}_{u_i}^{\frac{1}{2}}$, $\widehat{\mathbf{D}}_i = \text{diag}\{\hat{d}_j : j \in s_i\}$ and $\hat{d}_{ij} = \{g(\hat{\mu}_{ij})\}^{-1}\widehat{\text{Var}}(u_{ij})$, for $i = 1, \ldots, n$.

For assessing the hypothesis testing $\text{H}_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{m}$ against $\text{H}_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{m}$, where $\mathbf{C}$ is a $r \times p$ matrix of row rank r ($r \leq p$), one may apply a Wald-type test whose respective statistic is given by $\xi_W = (\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{m})^{\top}[\mathbf{C}\{\widehat{\mathbf{J}}_{\Psi}(\boldsymbol{\beta})\}^{-1}\mathbf{C}^{\top}]^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{m})$. For large sample and under usual regularity conditions it follows that $\xi_W \sim \chi_r^2$, where χ_r^2 denotes the chi-squared distribution with r degrees of freedom.

4 Simulation study

In order to assess the large sample behavior of the estimators derived from the iterative process described in Sect. 3.1, we will present in this section a simulation study based on the following UL-GEE model:

1. $y_{ij}|x_{ij} \sim \text{UL}(\mu_{ij})$,
2. $\Phi^{-1}(\mu_{ij}) = \beta_0 + \beta_1 x_{ij}$,
3. $\mathbf{R}_i = \mathbf{R}_i(\rho)$,

where $\Phi(\cdot)$ denotes the cumulative density function of the standard normal distribution, the correlation matrix $\mathbf{R}_i(\rho)$ follows exchangeable and first-order autoregressive structures among the elements of $\mathbf{z}_i^{\top} = \{z_{ij} : j \in s_i\}$, respectively. The correlation coefficient ρ and x_{ij}'s are fixed values generated from a uniform distribution in the range [0, 1], for $s_i = \{1, \ldots, s\}$ and $i = 1, \ldots, n$. The values assigned for the parameters are $\beta_0 = -3$ and $\beta_1 = 6$, $\rho = -0.1, 0.3, 0.7$, $n = 10, 50, 500$ and $s = 3, 5, 10$, whereas the bias (in absolute value) and the mean squared error (MSE) were calculated for each scenario considered. Under each simulation scheme, we uniformly covering all the parametric space and not just a specific region. We also consider for each scenario a negative correlation to illustrate a case in which the mixed model cannot handle.

The bias and MSE for the parameter θ were calculated, respectively, as $|\bar{\hat{\theta}} - \theta_0|$ and $R^{-1}\sum_{r=1}^{R}(\hat{\theta}^{(r)} - \theta_0)^2$, where $\bar{\hat{\theta}} = R^{-1}\sum_{r=1}^{R}\hat{\theta}^{(r)}$ with $\hat{\theta}^{(r)}$ being the estimate of θ from the rth replicate and θ_0 denotes the true parameter value. A total of $R = 5000$

replicates was considered for $\theta = \beta_0, \beta_1, \rho$ (see results in Tables 1, 2, 3, 4). Denote $\mathrm{N}_s(\mathbf{0}, \mathbf{R}_i)$ the s-variate normal distribution of mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{R}_i$, $\Phi(\mathbf{z}_i^*) = \{\Phi(z_{i1}^*), \ldots, \Phi(z_{is}^*))\}^\top$ with $\Phi(\cdot)$ denoting the cumulative density function of N(0, 1) and $F^{-1}(\cdot;\boldsymbol{\mu}_i) = \{F^{-1}(\cdot;\mu_{i1}), \ldots, F^{-1}(\cdot;\mu_{is})\}^\top$ with $F(\cdot;\mu_{ij})$ being the cumulative density function of the UL(μ_{ij}) distribution. We will apply the Algorithm 2 to simulate s correlated values from UL(μ_{i1}), … , UL(μ_{is}) marginal distributions through Gaussian copulas (see, for instance, Wicklin 2013) and fitted the UL-GEE model with the specified correlation structure being correct or incorrect to verify the behavior of the coefficient estimators.

We implemented the simulation study in the software R (R Core Team 2022) and add-on the packages: `lamW` (Adler 2022), `expint` (Goulet 2022), `pbapply` (Solymos 2023), `mvnfast` (Fasiolo 2023), `Pracma` (Borchers 2022), `gamlss.dist` (Stasinopoulos 2022), `gamlss` (Stasinopoulos 2023), `Matrix` (Maechler 2022), `RcppEigen` (Eddelbuettel 2022) and `Rcpp` (Eddelbuettel 2023).

Algorithm 2 Simulate correlated UL

1: **Inputs:**
   $\mathbf{R}_i$ positive-definite
2: **Initialize:**
   $\mathbf{z}_i^* =$ random vector from $\mathrm{N}_s(\mathbf{0}, \mathbf{R}_i)$
3: **return** $F^{-1}\{\Phi(\mathbf{z}_i^*);\ \boldsymbol{\mu}_i\}$

From Tables 1 and 3 that describe the results from the scenarios in which the data are generated and fitted under the same correlation structure, we may notice that the bias and MSE of $\hat{\beta}_0$ and $\hat{\beta}_1$ decrease as s and n increase, with indicative of consistency for both estimators. However, for $\hat{\rho}$ we may observe that the moment estimator is biased for the correlation coefficient. These results are expected since the data are generated for correlated UL's observations, whereas the moment estimator is calculated for correlated z_{ij}'s observations. Due to the monotonic relationship between y_{ij} and u_{ij}'s it is expected a good agreement between the true correlation value and the estimated one. But, small differences may still appear even for large sample, as we may observe for $n = 500$, where at the convergence the values for $\hat{\rho}$ were approximately -0.09, 0.27 and 0.67, respectively, whereas the correlation values considered in the data generation were -0.1, 0.3 and 0.7, respectively.

Tables 2 and 4 describe the results of the simulation study under misspecification of the correlation structure. We may notice a similar behavior for the bias and MSE of $\hat{\beta}_0$ and $\hat{\beta}_1$ with the ones observed in Tables 1 and 3. However, for $\hat{\rho}$, in general there is indication of inconsistency, particularly when the data are generated under the AR(1) structure and are fitted under the exchangeable structure.

Table 1 Average estimates $\bar{\hat{\beta}}_0$, $\bar{\hat{\beta}}_1$ and $\bar{\hat{\rho}}$, biases (in absolute value) and mean squared errors (MSEs) of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\rho}$ from a simulation study in which the data are generated from multivariate unit-Lindley distribution with AR(1) correlation structure and fitted under UL-GEE model with the same correlation structure

ρ	s	$\bar{\hat{\beta}}_0$	$\hat{\beta}_0$		$\bar{\hat{\beta}}_1$	$\hat{\beta}_1$		$\bar{\hat{\rho}}$	$\hat{\rho}$	
			Bias	MSE		Bias	MSE		Bias	MSE
		AR(1) with $n = 500$								
−0.1	10	−3.0000	0.0000	0.0001	5.9999	0.0001	0.0002	–0.0849	0.0151	0.0004
	5	−3.0004	0.0004	0.0002	6.0006	0.0006	0.0005	–0.0849	0.0151	0.0007
	3	−3.0004	0.0004	0.0003	6.0005	0.0005	0.0008	–0.0842	0.0158	0.0012
0.3	10	−3.0004	0.0004	0.0001	6.0003	0.0003	0.0002	0.2697	0.0303	0.0013
	5	−3.0005	0.0005	0.0002	6.0001	0.0001	0.0004	0.2692	0.0308	0.0018
	3	−3.0004	0.0004	0.0004	6.0000	0.0000	0.0007	0.2693	0.0307	0.0025
0.7	10	−3.0008	0.0008	0.0001	6.0004	0.0004	0.0001	0.6691	0.0309	0.0028
	5	−3.0010	0.0010	0.0003	6.0004	0.0004	0.0002	0.6699	0.0301	0.0040
	3	−3.0010	0.0010	0.0004	6.0005	0.0005	0.0004	0.6673	0.0327	0.0051
		AR(1) with $n = 50$								
−0.1	10	−3.0005	0.0005	0.0010	6.0001	0.0001	0.0022	–0.0861	0.0139	0.0022
	5	−3.0015	0.0015	0.0021	6.0004	0.0004	0.0048	–0.0884	0.0116	0.0047
	3	−3.0020	0.0020	0.0036	6.0006	0.0006	0.0083	–0.0879	0.0121	0.0093
0.3	10	−3.0006	0.0006	0.0012	6.0000	0.0000	0.0021	0.2651	0.0349	0.0052
	5	−3.0033	0.0033	0.0025	6.0022	0.0022	0.0044	0.2611	0.0389	0.0100
	3	−3.0033	0.0033	0.0041	6.0007	0.0007	0.0076	0.2553	0.0447	0.0168
0.7	10	−3.0037	0.0037	0.0016	6.0022	0.0022	0.0012	0.6368	0.0632	0.0151
	5	−3.0050	0.0050	0.0027	6.0030	0.0030	0.0024	0.6107	0.0893	0.0232
	3	−3.0053	0.0053	0.0039	6.0031	0.0031	0.0044	0.5891	0.1109	0.0304
		AR(1) with $n = 10$								

Table 1 (Continued)

ρ	s	$\bar{\hat{\beta}}_0$	$\hat{\beta}_0$		$\bar{\hat{\beta}}_1$	$\hat{\beta}_1$		$\bar{\hat{\rho}}$	$\hat{\rho}$	
			Bias	MSE		Bias	MSE		Bias	MSE
−0.1	10	−3.0043	0.0043	0.0053	6.0027	0.0027	0.0119	–0.0910	0.0090	0.0099
	5	−3.0068	0.0068	0.0111	6.0013	0.0013	0.0254	–0.0932	0.0068	0.0210
	3	−3.0106	0.0106	0.0207	6.0026	0.0026	0.0479	–0.1029	0.0029	0.0375
0.3	10	−3.0056	0.0056	0.0064	6.0027	0.0027	0.0111	0.2457	0.0543	0.0184
	5	−3.0110	0.0110	0.0134	6.0020	0.0020	0.0237	0.2101	0.0899	0.0352
	3	−3.0187	0.0187	0.0227	6.0082	0.0082	0.0439	0.1670	0.1330	0.0592
0.7	10	−3.0132	0.0132	0.0082	6.0070	0.0070	0.0061	0.5207	0.1793	0.0541
	5	−3.0172	0.0172	0.0143	6.0076	0.0076	0.0135	0.4545	0.2455	0.0891
	3	−3.0255	0.0255	0.0227	6.0110	0.0110	0.0275	0.4041	0.2959	0.1236

Table 2 Average estimates $\bar{\hat{\beta}}_0$, $\bar{\hat{\beta}}_1$ and $\bar{\hat{\rho}}$, biases (in absolute value) and mean squared errors (MSEs) of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\rho}$ from a simulation study in which the data are generated from multivariate unit-Lindley distribution with AR(1) correlation structure and fitted under UL-GEE model with exchangeable correlation structure

ρ	s	$\bar{\hat{\beta}}_0$	$\hat{\beta}_0$		$\bar{\hat{\beta}}_1$	$\hat{\beta}_1$		$\bar{\hat{\rho}}$	$\hat{\rho}$	
			Bias	MSE		Bias	MSE		Bias	MSE
		Exchangeable with $n = 500$								
–0.1	10	−3.0000	0.0000	0.0001	5.9999	0.0001	0.0002	–0.0158	0.0842	0.0071
	5	−3.0004	0.0004	0.0002	6.0006	0.0006	0.0005	–0.0316	0.0684	0.0048
	3	−3.0005	0.0005	0.0003	6.0006	0.0006	0.0008	–0.0534	0.0466	0.0027
0.3	10	−3.0004	0.0004	0.0001	6.0002	0.0002	0.0002	0.0726	0.2274	0.0518
	5	−3.0005	0.0005	0.0002	6.0001	0.0001	0.0005	0.1364	0.1636	0.0273
	3	−3.0004	0.0004	0.0004	6.0000	0.0000	0.0007	0.2051	0.0949	0.0103
0.7	10	−3.0005	0.0005	0.0002	6.0005	0.0005	0.0002	0.3238	0.3762	0.1424
	5	−3.0005	0.0005	0.0003	6.0001	0.0001	0.0003	0.4868	0.2132	0.0475
	3	−3.0008	0.0008	0.0004	6.0004	0.0004	0.0004	0.5948	0.1052	0.0144
		Exchangeable with $n = 50$								
–0.1	10	−3.0005	0.0005	0.0010	6.0001	0.0001	0.0023	–0.0173	0.0827	0.0072
	5	−3.0015	0.0015	0.0021	6.0002	0.0002	0.0049	–0.0345	0.0655	0.0058
	3	−3.0022	0.0022	0.0036	6.0008	0.0008	0.0083	–0.0583	0.0417	0.0073
0.3	10	−3.0004	0.0004	0.0013	5.9998	0.0002	0.0023	0.0686	0.2314	0.0547
	5	−3.0031	0.0031	0.0026	6.0022	0.0022	0.0046	0.1295	0.1705	0.0337
	3	−3.0030	0.0030	0.0041	6.0005	0.0005	0.0078	0.1922	0.1078	0.0230
0.7	10	−3.0050	0.0050	0.0019	6.0027	0.0027	0.0019	0.3032	0.3968	0.1639
	5	−3.0059	0.0059	0.0030	6.0033	0.0033	0.0032	0.4391	0.2609	0.0803
	3	−3.0061	0.0061	0.0041	6.0039	0.0039	0.0049	0.5230	0.1770	0.0481
		Exchangeable with $n = 10$								

Table 2 (Continued)

ρ	s	$\bar{\hat{\beta}}_0$	$\hat{\beta}_0$		$\bar{\hat{\beta}}_1$	$\hat{\beta}_1$		$\bar{\hat{\rho}}$	$\hat{\rho}$	
			Bias	MSE		Bias	MSE		Bias	MSE
–0.1	10	−3.0042	0.0042	0.0054	6.0027	0.0027	0.0124	–0.0232	0.0768	0.0073
	5	−3.0068	0.0068	0.0113	6.0010	0.0010	0.0260	–0.0450	0.0550	0.0095
	3	−3.0110	0.0110	0.0208	6.0030	0.0030	0.0486	–0.0775	0.0225	0.0230
0.3	10	−3.0061	0.0061	0.0066	6.0034	0.0034	0.0120	0.0544	0.2456	0.0655
	5	−3.0108	0.0108	0.0138	6.0014	0.0014	0.0252	0.0881	0.2119	0.0599
	3	−3.0185	0.0185	0.0231	6.0080	0.0080	0.0454	0.1139	0.1861	0.0680
0.7	10	−3.0183	0.0183	0.0100	6.0098	0.0098	0.0095	0.2255	0.4745	0.2405
	5	−3.0197	0.0197	0.0156	6.0082	0.0082	0.0167	0.3083	0.3917	0.1797
	3	−3.0256	0.0256	0.0236	6.0109	0.0109	0.0293	0.3472	0.3528	0.1584

Table 3 Average estimates $\bar{\hat{\beta}}_0$, $\bar{\hat{\beta}}_1$ and $\bar{\hat{\rho}}$, biases (in absolute value) and mean squared errors (MSEs) of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\rho}$ from a simulation study in which the data are generated from multivariate unit-Lindley distribution with exchangeable correlation structure and fitted under UL-GEE model with the same correlation structure

ρ	s	$\bar{\hat{\beta}}_0$	$\hat{\beta}_0$		$\bar{\hat{\beta}}_1$	$\hat{\beta}_1$		$\bar{\hat{\rho}}$	$\hat{\rho}$	
			Bias	MSE		Bias	MSE		Bias	MSE
		Exchangeable with $n = 500$								
–0.1	10	−3.0000	0.0000	0.0001	6.0000	0.0000	0.0002	–0.0846	0.0154	0.0002
	5	−3.0001	0.0001	0.0002	6.0000	0.0000	0.0004	–0.0850	0.0150	0.0003
	3	−3.0002	0.0002	0.0003	5.9999	0.0001	0.0008	–0.0851	0.0149	0.0007
0.3	10	−3.0006	0.0006	0.0002	6.0004	0.0004	0.0002	0.2692	0.0308	0.0016
	5	−3.0002	0.0002	0.0003	5.9999	0.0001	0.0004	0.2691	0.0309	0.0019
	3	−3.0002	0.0002	0.0004	6.0000	0.0000	0.0007	0.2682	0.0318	0.0025
0.7	10	−3.0010	0.0010	0.0002	6.0003	0.0003	0.0001	0.6695	0.0305	0.0048
	5	−3.0012	0.0012	0.0003	6.0004	0.0004	0.0002	0.6662	0.0338	0.0049
	3	−3.0009	0.0009	0.0004	6.0004	0.0004	0.0004	0.6656	0.0344	0.0051
		Exchangeable with $n = 50$								
–0.1	10	−3.0007	0.0007	0.0007	6.0004	0.0004	0.0020	–0.0847	0.0153	0.0003
	5	−3.0009	0.0009	0.0019	6.0002	0.0002	0.0046	–0.0859	0.0141	0.0012
	3	−3.0017	0.0017	0.0034	5.9999	0.0001	0.0078	–0.0880	0.0120	0.0051
0.3	10	−3.0015	0.0015	0.0018	6.0004	0.0004	0.0020	0.2635	0.0365	0.0073
	5	−3.0036	0.0036	0.0029	6.0016	0.0016	0.0042	0.2573	0.0427	0.0104
	3	−3.0041	0.0041	0.0042	6.0016	0.0016	0.0075	0.2574	0.0426	0.0158
0.7	10	−3.0066	0.0066	0.0022	6.0034	0.0034	0.0012	0.6006	0.0994	0.0261
	5	−3.0053	0.0053	0.0029	6.0027	0.0027	0.0022	0.5890	0.1110	0.0284
	3	−3.0062	0.0062	0.0040	6.0030	0.0030	0.0042	0.5815	0.1185	0.0311
		Exchangeable with $n = 10$								

Table 3 (Continued)

ρ	s	$\bar{\hat{\beta}}_0$	$\hat{\beta}_0$		$\bar{\hat{\beta}}_1$	$\hat{\beta}_1$		$\bar{\hat{\rho}}$	$\hat{\rho}$	
			Bias	MSE		Bias	MSE		Bias	MSE
–0.1	10	−3.0031	0.0031	0.0039	5.9991	0.0009	0.0110	–0.0817	0.0183	0.0006
	5	−3.0061	0.0061	0.0107	6.0013	0.0013	0.0259	–0.0925	0.0075	0.0043
	3	−3.0130	0.0130	0.0204	6.0039	0.0039	0.0489	–0.1017	0.0017	0.0216
0.3	10	−3.0086	0.0086	0.0094	6.0038	0.0038	0.0103	0.2153	0.0847	0.0242
	5	−3.0106	0.0106	0.0151	6.0022	0.0022	0.0231	0.1897	0.1103	0.0342
	3	−3.0161	0.0161	0.0234	6.0030	0.0030	0.0432	0.1651	0.1349	0.0524
0.7	10	−3.0242	0.0242	0.0124	6.0119	0.0119	0.0061	0.4230	0.2770	0.1057
	5	−3.0263	0.0263	0.0173	6.0129	0.0129	0.0128	0.4147	0.2853	0.1121
	3	−3.0272	0.0272	0.0229	6.0120	0.0120	0.0246	0.3897	0.3103	0.1309

Table 4 Average estimates $\bar{\hat{\beta}}_0$, $\bar{\hat{\beta}}_1$ and $\bar{\hat{\rho}}$, biases (in absolute value) and mean squared errors (MSEs) of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\rho}$ from a simulation study in which the data are generated from multivariate unit-Lindley distribution with exchangeable correlation structure and fitted under UL-GEE model with AR(1) correlation structure

ρ	s	$\bar{\hat{\beta}}_0$	$\hat{\beta}_0$		$\bar{\hat{\beta}}_1$	$\hat{\beta}_1$		$\bar{\hat{\rho}}$	$\hat{\rho}$	
			Bias	MSE		Bias	MSE		Bias	MSE
		AR(1) with $n = 500$								
–0.1	10	−3.0001	0.0001	0.0001	6.0001	0.0001	0.0002	–0.0848	0.0152	0.0004
	5	−3.0000	0.0000	0.0002	5.9999	0.0001	0.0005	–0.0853	0.0147	0.0006
	3	−3.0002	0.0002	0.0004	5.9999	0.0001	0.0008	–0.0849	0.0151	0.0010
0.3	10	−3.0006	0.0006	0.0002	6.0004	0.0004	0.0002	0.2694	0.0306	0.0017
	5	−3.0001	0.0001	0.0003	5.9998	0.0002	0.0004	0.2689	0.0311	0.0022
	3	−3.0003	0.0003	0.0004	6.0000	0.0000	0.0007	0.2682	0.0318	0.0028
0.7	10	−3.0008	0.0008	0.0002	6.0003	0.0003	0.0001	0.6684	0.0316	0.0045
	5	−3.0011	0.0011	0.0003	6.0003	0.0003	0.0003	0.6663	0.0337	0.0050
	3	−3.0009	0.0009	0.0004	6.0004	0.0004	0.0005	0.6656	0.0344	0.0055
		AR(1) with $n = 50$								
–0.1	10	−3.0010	0.0010	0.0008	6.0009	0.0009	0.0023	–0.0851	0.0149	0.0021
	5	−3.0007	0.0007	0.0019	5.9998	0.0002	0.0047	–0.0860	0.0140	0.0041
	3	−3.0017	0.0017	0.0034	6.0001	0.0001	0.0079	–0.0874	0.0126	0.0083
0.3	10	−3.0006	0.0006	0.0019	6.0000	0.0000	0.0022	0.2613	0.0387	0.0084
	5	−3.0034	0.0034	0.0030	6.0014	0.0014	0.0046	0.2573	0.0427	0.0131
	3	−3.0042	0.0042	0.0043	6.0017	0.0017	0.0078	0.2587	0.0413	0.0192
0.7	10	−3.0078	0.0078	0.0026	6.0034	0.0034	0.0015	0.6021	0.0979	0.0255
	5	−3.0055	0.0055	0.0032	6.0026	0.0026	0.0026	0.5899	0.1101	0.0290
	3	−3.0059	0.0059	0.0042	6.0028	0.0028	0.0047	0.5807	0.1193	0.0327
		AR(1) with $n = 10$								

Table 4 (Continued)

ρ	s	$\bar{\hat{\beta}}_0$	$\hat{\beta}_0$		$\bar{\hat{\beta}}_1$	$\hat{\beta}_1$		$\bar{\hat{\rho}}$	$\hat{\rho}$	
			Bias	MSE		Bias	MSE		Bias	MSE
–0.1	10	−3.0053	0.0053	0.0042	6.0024	0.0024	0.0123	–0.0824	0.0176	0.0089
	5	−3.0069	0.0069	0.0106	6.0022	0.0022	0.0254	–0.0921	0.0079	0.0185
	3	−3.0132	0.0132	0.0203	6.0041	0.0041	0.0484	–0.1018	0.0018	0.0355
0.3	10	−3.0082	0.0082	0.0099	6.0034	0.0034	0.0114	0.2130	0.0870	0.0296
	5	−3.0106	0.0106	0.0155	6.0026	0.0026	0.0245	0.1875	0.1125	0.0436
	3	−3.0162	0.0162	0.0239	6.0032	0.0032	0.0447	0.1630	0.1370	0.0639
0.7	10	−3.0281	0.0281	0.0144	6.0134	0.0134	0.0077	0.4247	0.2753	0.1066
	5	−3.0261	0.0261	0.0185	6.0118	0.0118	0.0151	0.4115	0.2885	0.1164
	3	−3.0271	0.0271	0.0238	6.0118	0.0118	0.0269	0.3866	0.3134	0.1363

5 Diagnostic procedures

Model checking consists in a set of various diagnostic procedures to assess the assumptions made for the model, as well as to detect the existence of outlying observations and the sensitivity of the parameter estimates under perturbations made in the model or data. In the context of estimating equations there is a vast literature, but concentrated on procedures developed for generalized estimating equations (see, for instance, Preisser and Qaqish 1996; Venezuela et al. 2011; Hardin and Hilbe 2012 and Manghi et al. 2019). An extension of such procedures for Godambe estimating equations has been performed recently for Birnbaum-Saunders-GEE models (Tsuyuguchi et al. 2020). Thus, based on this work, we will derive in this section some diagnostic procedures for UL-GEE models.

5.1 Residual analysis

In order to assess the assumptions made for the UL-GEE model, particularly on the marginal distributions with the proposed correlation structure, and to detect the presence of outlying observations, we will consider the marginal quantile residual (Dunn and Smyth 1996) defined as

$$r_{q_{ij}} = \Phi^{-1}\{F(y_{ij};\hat{\mu}_{ij})\},$$

where

$$F(y_{ij};\hat{\mu}_{ij}) = 1 - \left(\frac{1-\hat{y}_{ij}\mu_{ij}}{1-y_{ij}}\right)\exp\left\{-\frac{(1-\hat{\mu}_{ij})y_{ij}}{(1-y_{ij})\hat{\mu}_{ij}}\right\},$$

denotes the cumulative density function of $y_{ij} \sim \text{UL}(\mu_{ij})$ evaluated at $\hat{\mu}_{ij} = g^{-1}(\hat{\eta}_{ij})$ and $\Phi(\cdot)$ is the cumulative density function of the standard normal distribution, for $i = 1, \ldots, n$. Under the hypothesis of independence between u_{ij} and $u_{ij'}$, for $j \neq j'$, one has that $r_{q_{ij}}$ are asymptotically N(0, 1). However, in practice, one may have correlated observations within experimental unit, so some empirical confidence band should be added into the normal probability plot with the marginal quantile residual $r_{q_{ij}}$. Thus, departures of the former residuals from the empirical confidence band may indicate that the assumption of UL marginal distribution with the proposed working correlation matrix is not suitable to fit the data. In addition, the graph may reveal outlying observations. We may apply, for generating the empirical band, the same algorithm proposed in the previous section for the simulation studies.

Alternatively, one may randomly select n residuals, namely $r^*_{q_1}, \ldots, r^*_{q_n}$, with $r^*_{q_i}$ being randomly selected from the residual set $\{r_{q_{ij}} : j \in s_i\}$ of the ith experimental unit, $i = 1, \ldots, n$. Thus, one has that $r^*_{q_1}, \ldots, r^*_{q_n}$ are asymptotically independent N(0, 1). With this set of residuals we may perform various residual graphs, such as the quantile residual against the fitted value, the normal probability plot and the worm plot as in GAMLSS (see, for instance, Stasinopoulos et al. 2017). Since there are $\prod_{i=1}^{n} n(s_i)$ possible sets of residuals, we may display the graphs of m different sets.

5.2 Sensitivity studies

The idea of sensitivity studies is to assess the influence of observations on the parameter estimates under perturbations made in the model or data. The main challenge in this kind of study is to detect observations that have disproportional influence on the parameter estimates, particularly with inferential change. However, such observations may be masked, requiring a careful analysis of the influence graphs. There are various procedures developed for regression models, such as the traditional case deletion (see, for instance, Cook and Weisberg 1982), local influence (Cook 1986), conformal local influence (Poon and Poon 1999) and forward search (Atkinson and Riani 2000), among others. In the context of estimating equations (Hardin and Hilbe 2012) present a review for GEE, whereas in (Tsuyuguchi et al. 2020) one has some extensions for models out of exponential family. Based on the last work we will derive in this section the conformal local curvature for UL-GEE models. Measures based on dropping observations will be applied only in the confirmatory analysis of the highlighted observations by the conformal local influence.

The log-likelihood function, usually applied in models in which the full likelihood is known, is replaced in estimating equations by the fit function $\mathcal{F}(\boldsymbol{\beta})$ (Cadigan and Farrell 2002), that is assumed twice differentiable in $\boldsymbol{\beta}$ with unique interior parameter estimate and defined as

$$\boldsymbol{\Psi}(\hat{\boldsymbol{\beta}}) = \left\{ \frac{\partial \mathcal{F}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right\}\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \mathbf{0}.$$

Then, an appropriate influence measure is given by $\mathrm{FD}_{\omega} = 2\{\mathcal{F}(\hat{\boldsymbol{\beta}}) - \mathcal{F}(\hat{\boldsymbol{\beta}}_{\omega})\}$, with $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_N)^{\top}$ denoting the perturbation vector and $\hat{\boldsymbol{\beta}}_{\omega}$ is the solution of the perturbed estimating equations $\boldsymbol{\Psi}(\hat{\boldsymbol{\beta}}_{\omega}|\boldsymbol{\omega}) = \mathbf{0}$. The no perturbation vector $\boldsymbol{\omega}_0$ is defined as $\boldsymbol{\Psi}(\boldsymbol{\beta}_{\omega}|\boldsymbol{\omega}_0) = \boldsymbol{\Psi}(\boldsymbol{\beta})$.

Poon and Poon (1999) derived the conformal normal curvature in the unitary direction $\boldsymbol{\ell}$ as

$$\mathrm{B}_{\ell}(\boldsymbol{\beta}) = |\boldsymbol{\ell}^{\top}\mathbf{B}\boldsymbol{\ell}|/\sqrt{tr(\mathbf{B}^2)},$$

where $0 \leq \mathrm{B}_{\ell}(\boldsymbol{\beta}) \leq 1$, $\mathbf{B} = \boldsymbol{\Delta}^{\top}\{\ddot{\mathcal{F}}(\boldsymbol{\beta})\}^{-1}\boldsymbol{\Delta}$ is a symmetric non-negative definite matrix with $\boldsymbol{\Delta} = \partial\boldsymbol{\Psi}(\boldsymbol{\beta}|\boldsymbol{\omega})/\partial\boldsymbol{\omega}^{\top}$ being evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, $\boldsymbol{\rho} = \hat{\boldsymbol{\rho}}$ and $\boldsymbol{\omega} = \boldsymbol{\omega}_0$. Influence graphs based on aggregate measures of the non null eigenvalues and the corresponding eigenvectors of the matrix $\mathbf{B}$ were proposed by Poon and Poon (1999). In particular we will consider the aggregate measure B_{ij}, that corresponds to the conformal normal curvature evaluated in the direction $\boldsymbol{\ell}_{ij}$ of the (i, j)th observation, where $\boldsymbol{\ell}_{ij}$ denotes an $N \times 1$ vector of zeros with one in the (i, j)th position. Lee and Xu (2004) suggest highlight possible influential observations such that $\mathrm{B}_{ij} > \bar{\mathrm{B}} + \mathrm{SD(B)}c^*$, where $\bar{\mathrm{B}}$ and SD(B) denote, respectively, the mean and the standard deviation of $\{\mathrm{B}_{ij}, j \in s_i; i = 1, \ldots, n\}$ with c^* being selected appropriately.

To assess the effect of the highlighted observations under the adopted perturbation scheme, we will apply the MRC (Maximum Relative Chance) proposed by Lee et al. (2006) and expressed as

$$\text{MRC} = \max_{1\le k\le p}\left|\frac{\hat{\beta}_k - \hat{\beta}_k^0}{\hat{\beta}_k}\right|,$$

where $\hat{\beta}_k^0$ denotes the estimate of β_k after dropping the pointed out observations. The criterion is to compare the MRC with the ones obtained from a set of random samples from the not highlighted observations. In the sequel we will derive the matrix $\mathbf{\Delta}$ for two usual perturbation schemes.

In order to assist the choice of a suitable correlation structure, we will extend the Quasi-likelihood Independence Criterion (QIC) (see, for example, Hardin and Hilbe 2012) for the UL-GEE class. The respective measure may be expressed as

$$\text{QIC} = -2\sum_{i=1}^{n}\sum_{j\in s_i}\log\{f(y_{ij};\hat{\mu}_{ij})\} + 2\text{tr}\left[\{\widehat{\mathbf{J}}_\Psi(\boldsymbol{\beta})\}^{-1}\left\{\sum_{i=1}^{n}\hat{\mathbf{S}}_{il}(\boldsymbol{\beta})\right\}\right],$$

wherein $\hat{\mu}_{ij}$ is the estimate for a specific correlation matrix $\mathbf{R}_i(\rho)$ and $\mathbf{S}_{il}(\boldsymbol{\beta})$ denotes the matrix $\mathbf{S}_i(\boldsymbol{\beta})$ evaluate under the independent correlation structure. The criterion is to select the correlation structure such that QIC is minimized.

5.3 Case-weight perturbation scheme

Under the case-weight perturbation scheme the estimating function for $\boldsymbol{\beta}$ is expressed as

$$\mathbf{\Psi}(\boldsymbol{\beta}|\boldsymbol{\omega}) = \sum_{i=1}^{n}\mathbf{X}_i^\top\mathbf{W}_i\mathbf{D}_i^{-1}\text{diag}(\boldsymbol{\omega}_i)\mathbf{u}_i, \tag{5}$$

where $\boldsymbol{\omega}_i^\top = \{\omega_{ij} : j \in s_i\}$ denotes the perturbations applied in the elements of the ith experimental unit, $0 \le \omega_{ij} \le 1$, for $i = 1, \ldots, n$, whereas $\boldsymbol{\omega} = (\boldsymbol{\omega}_1^\top, \ldots, \boldsymbol{\omega}_n^\top)^\top$. The no perturbation vector $\boldsymbol{\omega}_0$ is a $N \times 1$ vector formed by 1's. One may re-express the estimating function (5) in the matrix form

$$\mathbf{\Psi}(\boldsymbol{\beta}|\boldsymbol{\omega}) = \mathbf{X}^\top\mathbf{W}\mathbf{D}^{-1}\text{diag}(\boldsymbol{\omega})\mathbf{u},$$

where $\mathbf{W} = \text{blockdiag}\{\mathbf{W}_1, \ldots, \mathbf{W}_n\}$, $\mathbf{D} = \text{blockdiag}\{\mathbf{D}_1, \ldots, \mathbf{D}_n\}$, $\mathbf{X} = (\mathbf{X}_1^\top, \ldots, \mathbf{X}_n^\top)^\top$ and $\mathbf{u} = (\mathbf{u}_1^\top, \ldots, \mathbf{u}_n^\top)^\top$. Consequently, we obtain

$$\mathbf{\Delta} = \left.\frac{\partial\mathbf{\Psi}(\boldsymbol{\beta}|\boldsymbol{\omega})}{\partial\boldsymbol{\omega}^\top}\right|_{(\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\rho=\hat{\rho},\boldsymbol{\omega}=\boldsymbol{\omega}_0)} = \mathbf{X}^\top\widehat{\mathbf{W}}\widehat{\mathbf{D}}^{-1}\text{diag}(\hat{\mathbf{u}}).$$

Thus, the $\mathbf{B}$ matrix is approximated by $\mathbf{\Delta}^\top[\text{E}\{\ddot{\mathcal{F}}(\hat{\boldsymbol{\beta}})\}]^{-1}\mathbf{\Delta}$ which is expressed as $\text{diag}(\hat{\mathbf{u}})\widehat{\mathbf{D}}^{-1}\widehat{\mathbf{W}}\mathbf{X}\{\mathbf{X}^\top\widehat{\mathbf{W}}\mathbf{X}\}^{-1}\mathbf{X}^\top\widehat{\mathbf{W}}\widehat{\mathbf{D}}^{-1}\text{diag}(\hat{\mathbf{u}})$.

5.4 Response perturbation scheme

In this case is usual to perform the following perturbation in each observed response

$$y_{\omega_{ij}} = y_{ij} + \omega_{ij}\sigma_{ij},$$

where $\omega_{ij} \in \mathbb{R}$, so that $0 < y_{ij} < 1$ and σ_{ij} denotes the standard deviation of y_{ij}. Under this perturbation scheme, the estimating function for $\boldsymbol{\beta}$ may be expressed in the matrix form

$$\boldsymbol{\Psi}(\boldsymbol{\beta}|\boldsymbol{\omega}) = \mathbf{X}^\top \mathbf{W}\mathbf{D}^{-1}\text{diag}(\boldsymbol{\omega})\mathbf{u}_\omega,$$

where $\boldsymbol{\omega} = (\boldsymbol{\omega}_1^\top, \ldots, \boldsymbol{\omega}_n^\top)^\top$, $\boldsymbol{\omega}_i^\top = \{\omega_{ij} : j \in s_i\}$, $\mathbf{u}_\omega = (\mathbf{u}_{\omega_1}^\top, \ldots, \mathbf{u}_{\omega_n}^\top)^\top$, $\mathbf{u}_{\omega_i}^\top = \{u_{\omega_{ij}} : j \in s_i\}$ and

$$u_{\omega_{ij}} = \frac{z_{\omega_{ij}}}{\mu_{ij}^2} - \frac{(1+\mu_{ij})}{\mu_{ij}(1-\mu_{ij})},$$

with $z_{\omega_{ij}} = y_{\omega_{ij}}/(1-y_{\omega_{ij}})$. It may be showed that $\partial u_{\omega_{ij}}/\partial\omega_{ij} = f_{ij}$, where $f_{ij} = \sigma_{ij}/\{\mu_{ij}^2(1-y_{ij})^2\}$, for $i = 1, \ldots, n$.

Then, we obtain

$$\boldsymbol{\Delta} = \left.\frac{\partial\boldsymbol{\Psi}(\boldsymbol{\beta}|\boldsymbol{\omega})}{\partial\boldsymbol{\omega}^\top}\right|_{(\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\rho=\hat{\rho},\boldsymbol{\omega}=\boldsymbol{\omega}_0)} = \mathbf{X}^\top\widehat{\mathbf{W}}\widehat{\mathbf{D}}^{-1}\text{diag}(\hat{\mathbf{f}}),$$

where $\mathbf{f} = (\mathbf{f}_1^\top, \ldots, \mathbf{f}_n^\top)^\top$ with $\mathbf{f}_i^\top = \{f_{ij} : j \in s_i\}$, for $i = 1, \ldots, n$. Thus, the $\mathbf{B}$ matrix is approximated by $\boldsymbol{\Delta}^\top[\text{E}\{\ddot{\mathcal{F}}(\hat{\boldsymbol{\beta}})\}]^{-1}\boldsymbol{\Delta}$ which is expressed as $\text{diag}(\hat{\mathbf{f}})\widehat{\mathbf{D}}^{-1}\widehat{\mathbf{W}}\mathbf{X}\{\mathbf{X}^\top\widehat{\mathbf{W}}\mathbf{X}\}^{-1}\mathbf{X}^\top\widehat{\mathbf{W}}\widehat{\mathbf{D}}^{-1}\text{diag}(\hat{\mathbf{f}})$.

6 Application

As illustration of the UL-GEE models proposed in this paper, we will analyze a data set from the Brazilian census of 1991, 2000 and 2010 for the 27 federation units. The data were extracted from the Atlas of Brazil Human Development database, available at http://www.atlasbrasil.org.br/consulta. Particularly, the relationship between the proportion of people in households with inadequate water supply and sewage (`psewage`) and the Gini coefficient (`Gini`). That is, to assess in the 27 federation units how the social inequality affects over time the evolution of the population with inadequate water supply and sewage. The data set is presented in Table 5. Figure 2 describes the boxplot, for each year (`time`), of the logit(`psewage`), its scatter plot (with tendency) versus `Gini` and the empirical histogram of `psewage` versus the theoretical for UL distribution. We may notice from these graphs a good agreement of the UL distribution with the data, and that logit(`psewage`) does not change much over time, but increases for each year as the Gini coefficient increases, with indication of interaction between `time` and `Gini`.

Table 5 Proportion of people in households with inadequate water supply and sewage (`psewage`) and the Gini coefficient (`Gini`) for the 27 Brazilian federation units from the census of 1991, 2000 and 2010

Federation unit	Census					
	1991		2000		2010	
	Gini	psewage	Gini	psewage	Gini	psewage
AC	0.63	0.1491	0.64	0.3730	0.63	0.2809
AL	0.63	0.2929	0.68	0.1813	0.63	0.1307
AP	0.58	0.1296	0.62	0.2898	0.60	0.2151
AM	0.62	0.2085	0.67	0.2933	0.65	0.2098
BA	0.67	0.2571	0.66	0.1322	0.62	0.0935
CE	0.66	0.3589	0.67	0.1533	0.61	0.1099
DF	0.62	0.0406	0.63	0.0314	0.63	0.0072
ES	0.60	0.0409	0.60	0.0354	0.56	0.0099
GO	0.59	0.0204	0.60	0.0670	0.55	0.0338
MA	0.60	0.3278	0.65	0.2600	0.62	0.2399
MT	0.60	0.0117	0.62	0.0932	0.56	0.0393
MS	0.60	0.0340	0.62	0.1582	0.55	0.0516
MG	0.61	0.0423	0.61	0.0433	0.56	0.0184
PA	0.64	0.2980	0.63	0.1453	0.61	0.1175
PB	0.60	0.0091	0.60	0.0354	0.53	0.0097
PR	0.62	0.1607	0.65	0.3960	0.62	0.2905
PE	0.65	0.2319	0.66	0.1486	0.62	0.1083
PI	0.64	0.3551	0.65	0.0731	0.61	0.0815
RJ	0.61	0.0125	0.60	0.0262	0.59	0.0167
RN	0.63	0.2888	0.64	0.2031	0.60	0.0940
RS	0.59	0.0170	0.58	0.0297	0.54	0.0089
RO	0.62	0.0381	0.60	0.2446	0.56	0.1243
RR	0.63	0.0217	0.61	0.2095	0.63	0.1244
SC	0.55	0.0071	0.56	0.0216	0.49	0.0078
SP	0.55	0.0055	0.58	0.0087	0.56	0.0060
SE	0.63	0.1619	0.65	0.1498	0.62	0.1102
TO	0.63	0.0970	0.65	0.1929	0.60	0.0807

Based on these descriptive analysis, we propose the following UL-GEE model:

1. $\texttt{psewage}_{ij} | \texttt{Gini}_{ij} \sim \text{UL}(\mu_{ij})$
2. $$\log\left\{\frac{\mu_{ij}}{1-\mu_{ij}}\right\} = \begin{cases} \alpha_1 + \beta_1 \texttt{Gini}_{ij} & \text{for } 1991 \\ \alpha_2 + \beta_2 \texttt{Gini}_{ij} + \tau_2 \texttt{Gini}_{ij}^2 & \text{for } 2000 \\ \alpha_3 + \beta_3 \texttt{Gini}_{ij} & \text{for } 2010 \end{cases}$$
3. $\mathbf{R}_i = \mathbf{R}_i(\rho)$,

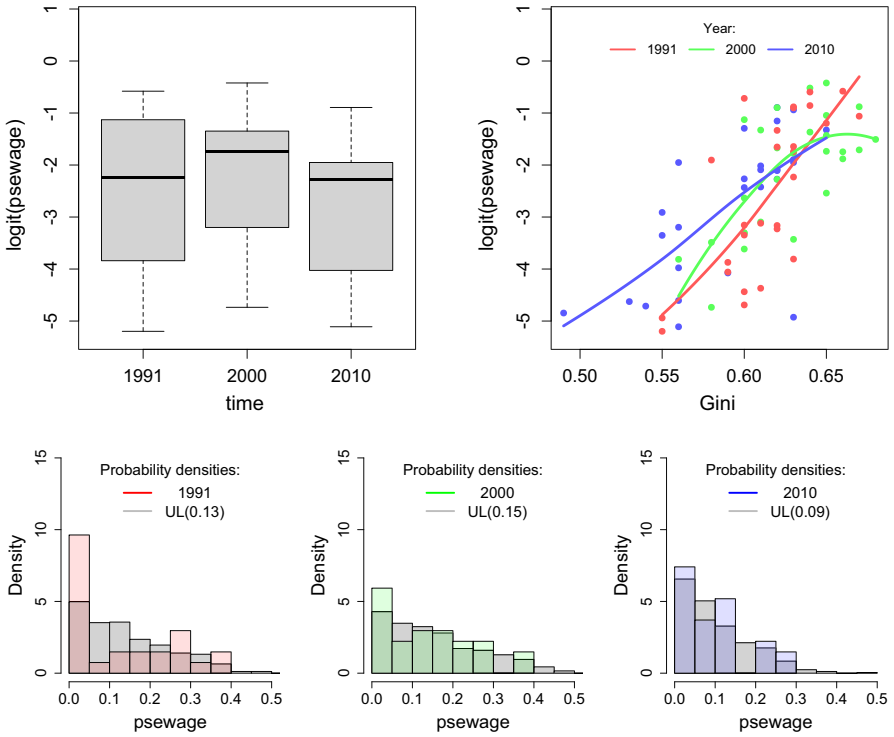

Fig. 2 The left upper panel describes the boxplot of the logit(`psewage`), whereas the right upper panel presents the scatter plot between logit(`psewage`) and `Gini` with smooth curve and the bottom panel shows the empirical histogram of `psewage` and the theoretical for UL distribution for each year and the 27 Brazilian federation units

Table 6 Parameter estimates and the respective approximate standard errors from the UL-GEE model with exchangeable correlation structure fitted to explain the proportion of people in households with inadequate water supply and sewage in the 27 Brazilian federation units in the years of 1991, 2000 and 2010 given the Gini coefficient

Parameter	Estimate	Std. Error	z-value	P-value
α_1	–18.41	3.59	–5.13	< 0.0001
α_2	–103.98	25.09	–4.14	< 0.0001
α_3	–11.69	2.00	–5.86	< 0.0001
β_1	26.50	5.61	4.72	< 0.0001
β_2	316.01	79.63	3.97	0.0001
β_3	15.57	3.30	4.71	< 0.0001
τ_2	–243.83	63.14	–3.86	0.0001
ρ	0.45			

where μ_{ij} denotes the expected proportion of people in households with inadequate water supply and sewage of the ith federation unit in the jth year, for $i = 1, \ldots, 27$ and $j = 1, 2, 3$, whereas $\mathbf{R}_i(\rho)$ is the correlation structure. Based on diagnostic analysis and simplicity we choose the exchangeable correlation structure with the 2nd smallest QIC value of 5928.91. The smallest value was obtained for the unstructured

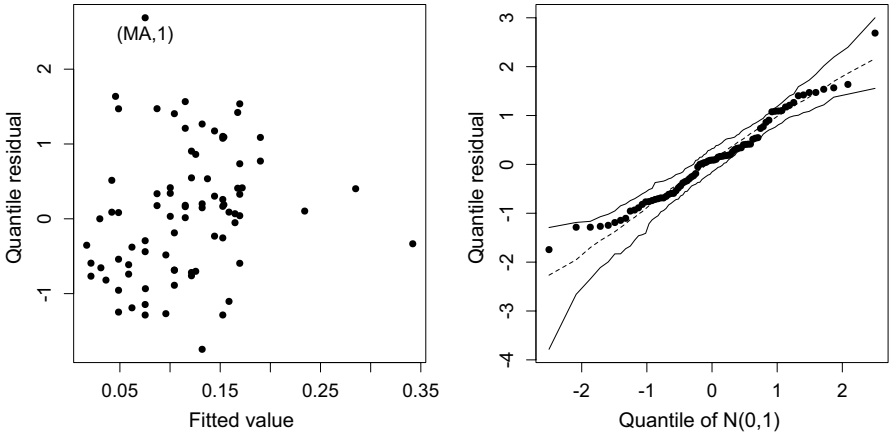

Fig. 3 Scatter plot between the quantile residual and the fitted value (left panel) and the normal probability plot of the quantile residual added by a 99% confidence band (right panel) from the UL-GEE model with exchangeable correlation structure fitted to explain the proportion of people in households with inadequate water supply and sewage in the 27 Brazilian federation units in the years of 1991, 2000 and 2010 given the Gini coefficient

correlation structure, QIC = 4877.56, but this structure spends three correlation estimates whereas exchangeable spends just one. The parameter estimates with their approximate standard errors from the selected model are presented in Table 6.

Figure 3 (left) describes the scatter plot between the quantile residual and the fitted value with the observation (MA,1) (federation unit MA and year 1991) pointed

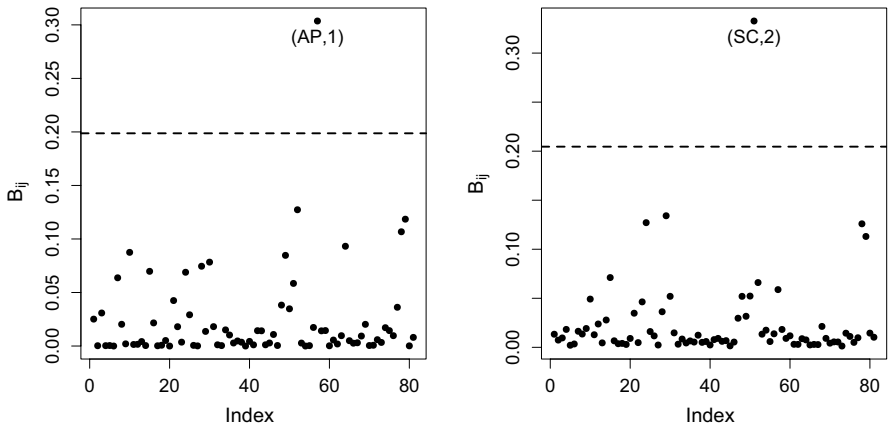

Fig. 4 Index plots of the conformal normal influence measure B_{ij} under the case-weight perturbation scheme (left panel) and under the response perturbation scheme (right panel) from the UL-GEE model with exchangeable correlation structure fitted to explain the proportion of people in households with inadequate water supply and sewage in the 27 Brazilian federation units in the years of 1991, 2000 and 2010 given the Gini coefficient

out as possible outlier, whereas in Fig. 3 (right) one has the normal probability plot of the quantile residual with an empirical confidence band of 99%. We may notice from both graphs indication that the proposed model is not unsuitable. In addition, Fig. 7 presents the worm plots of $m = 12$ sets of randomized quantile residuals taken from each federation unit, confirming the adequacy of the proposed model.

In Fig. 4 one has the sensitivity analysis based on the index plots of the conformal normal curvature B_{ij}, under the case-weight (left) and response (right) perturbation schemes with benchmark $c^* = 4$, and two observations,namely (AP,1)(federation unit AP and year 1991) and (SC,2)(federation unit SC and year 2000), are highlighted. From Fig. 5 we may notice that the three pointed out observations by the diagnostic graphs are in general extreme with respect to the observations in the same year.

In order to confirm the impact of the three highlighted observations, we compare their MRC value as well as the correlation coefficient estimate by dropping these observations with the respective values from a random sample of 15 sets of 3 observations each taken from the no highlighted group of observations. As we may see from Table 7 the MRC value of the highlighted group is 4 times the largest MRC value of the no highlighted group and the correlation coefficient estimate seems to be inflated by the highlighted observations.

Since the inference is based on $n = 27$ experimental units, we perform a simulation study to assess the empirical distribution of the coefficient estimates from the selected model based on Monte Carlo simulations of the selected model. The normal probability plots are described in Fig. 8 for the standardized estimates and we may notice a very good agreement of all estimates with the standard normal distribution.

Finally, from Fig. 6 one has the confidence bands of 95% for the proportion of people in households with inadequate water supply and sewage given the Gini coefficient for 1991, 2000 and 2010 (see, for instance, Piegorsch and Casella 1988). Similar behavior one may observe for the years 1991 and 2010, for which the proportion

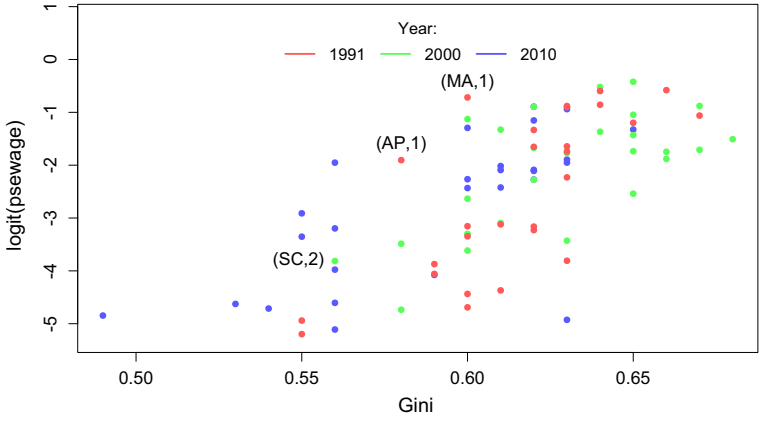

Fig. 5 Scatter plot between logit(`psewage`) and `Gini` for each year and the 27 Brazilian federation units with the identification of the highlighted observations in the residual and sensitivity graphs

Table 7 Comparison of the MRC values and the correlation coefficient estimates between the highlighted observations and 15 sets of random samples of size 3 taken from the set of no highlighted group

Sample of observations	MRC	$\hat{\rho}$
1	0.01	0.42
2	0.12	0.55
3	0.19	0.38
4	0.10	0.42
5	0.05	0.42
6	0.05	0.43
7	0.08	0.44
8	0.04	0.45
9	0.08	0.44
10	0.09	0.38
11	0.01	0.51
12	0.13	0.41
13	0.06	0.41
14	0.02	0.49
15	0.04	0.43
Highlighted	0.76	0.27

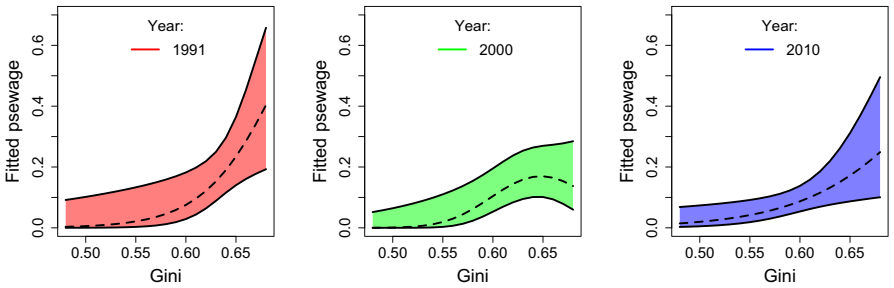

Fig. 6 Confidence bands of 95% for the proportion of people in households with inadequate water supply and sewage given the Gini coefficient for the years 1991, 2000 and 2010

increases as Gini increases, but with larger variability for large Gini values. However, for the year of 2000 we may observe the same tendency with a stability for large Gini values. This last effect may be due some federation units with large Gini in 1991 and important reduction of the proportion in 2000.

7 Concluding remarks

In this paper we derive estimating equations for analyzing correlated rates and proportions in the interval (0, 1) by assuming that the marginal distributions are unit-Lindley (Mazucheli et al. 2019). The within experimental unit dependence is estimated in the same sense of generalized estimating equations (Liang and Zeger 1986)

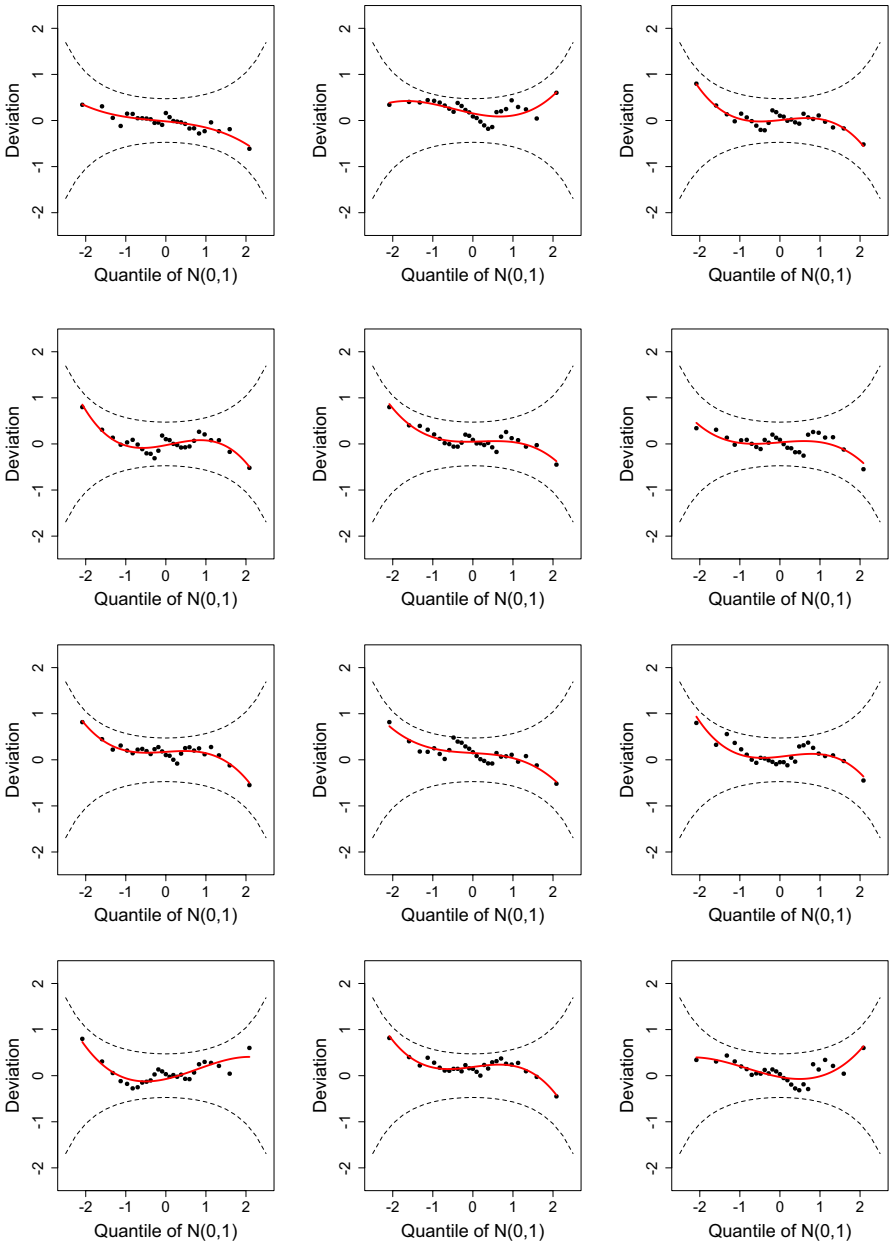

Fig. 7 Worm plots of $m = 12$ sets of randomized quantile residuals taken from each unit federation from the UL-GEE model with exchangeable correlation structure fitted to explain the proportion of people in households with inadequate water supply and sewage in the 27 Brazilian federation units in the years of 1991, 2000 and 2010 given the Gini coefficient

and the inference is based on the Godambe approach (Godambe 1997). Various results are derived in the paper, such as a reweighted iterative process for estimating the regression coefficients jointly with the within experimental unit correlation structure, residual analysis based on the marginal quantile residuals and sensitivity studies based on the conformal normal curvature. In particular, we propose a novel randomized quantile residual to assess the adequacy of the marginal distributions and correlation structure. Unlike from other works on estimating equations, the results derived in this paper may be applied for unbalanced studies.

From the simulation studies one has indication of consistency for the regression coefficient estimates for all the scenarios considered, even under

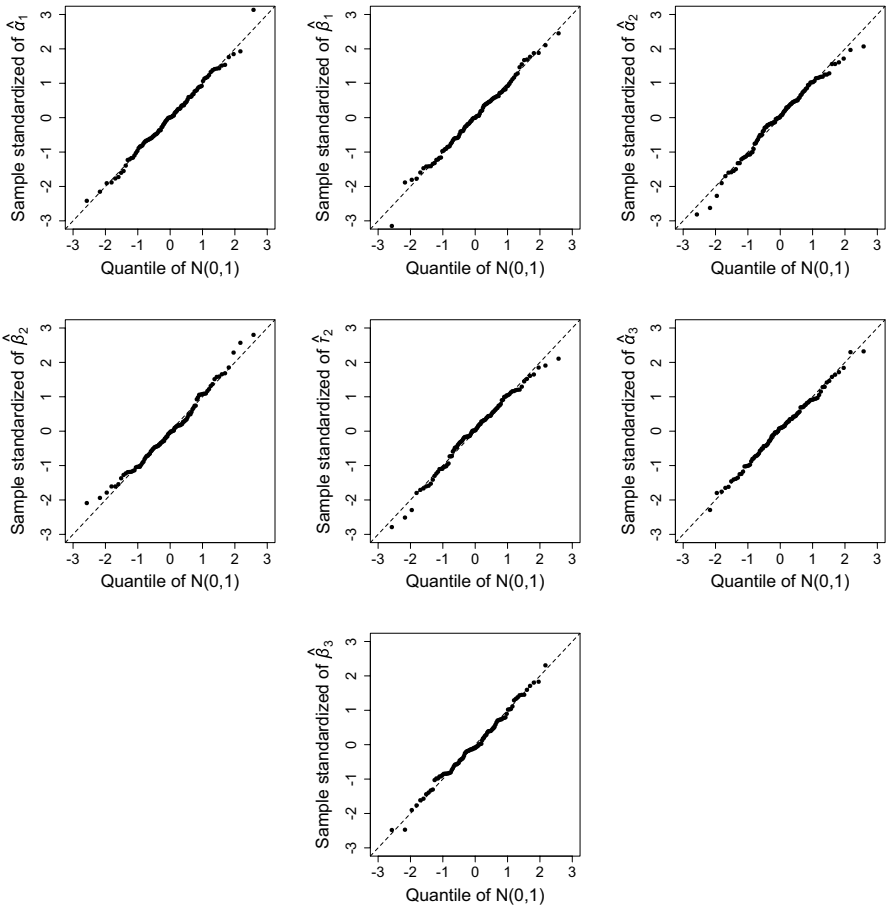

Fig. 8 Normal probability plots for the empirical distribution of the standardized parameter estimates from the UL-GEE model with exchangeable correlation structure fitted to explain the proportion of people in households with inadequate water supply and sewage in the 27 Brazilian federation units in the years of 1991, 2000 and 2010 given the Gini coefficient

misspecification of the correlation structure. An application to a real data set is presented in which the proportion of people in households with inadequate water supply and sewage is related with the Gini coefficient in the Brazilian federation units for a longitudinal study based on the Brazilian census of 1991, 2000 and 2010. Possible extensions of this work are the addition of additive components as described, for example, in Ibacache et al. (2013) and the joint modeling of the correlation structure (see, for instance, Yan and Fine 2004). The R scripts for fitting and diagnostic of UL-GEE models, and to display the simulation and application outputs are available, respectively, in the addresses https://github.com/silva-danilo/ulgee and https://github.com/silva-danilo/ulgee_sup.

Appendix A

Suppose that $u \sim \text{UL}(\mu)$, $0 < \mu < 1$. As showed in Sect. 2 the score function for μ may be expressed as

$$u = \frac{d\text{L}(\mu)}{d\mu} = \frac{z}{\mu^2} - \frac{1+\mu}{\mu(1-\mu)},$$

where $z = y/(1-y)$. Given that $\text{E}(u) = 0$ implies that

$$\text{E}(z) = \frac{\mu(1+\mu)}{1-\mu}.$$

The first derivative of u equals

$$u' = -\frac{2}{(1-\mu)^2} + \frac{1}{\mu^2} - \frac{2z}{\mu^3},$$

and using the previous results, we obtain

$$\text{E}(u') = -\frac{2}{(1-\mu)^2} + \frac{1}{\mu^2} - \frac{2(1+\mu)}{\mu^2(1-\mu)} = \frac{(1-\mu)^2 - 2}{\mu^2(1-\mu)^2}.$$

Therefore, given that $\text{E}(-u') = \text{E}(u^2)$ implies that

$$\text{Var}(u) = \text{E}(-u') = \frac{2-(1-\mu)^2}{\mu^2(1-\mu)^2},$$

and

$$\text{Var}(z) = \mu^4 \text{Var}(u) = \frac{\mu^2\{2-(1-\mu)^2\}}{(1-\mu)^2}.$$

Appendix B

Let $z = y/(1-y)$ be the sample odds, where $y \sim \text{UL}(\mu)$, $0 < \mu < 1$. From Mazucheli et al. (2019) the probability density function of z may be expressed in the form

$$f(z;\mu) = \frac{(1+z)(1-\mu)^2}{\mu}\exp\left\{-\frac{z(1-\mu)}{\mu}\right\},$$

where $z > 0$. We may re-express this pdf in the one-parametric exponential family of distributions

$$f(z;\theta) = \exp\{\theta z - b(\theta) + c(z)\},$$

with $\theta = 1-\mu^{-1}$, $b(\theta) = \log(1-\theta) - 2\log(-\theta)$ and $c(z) = \log(1+z)$ (see, for instance, McCullagh and Nelder 1989). Then, it follows that $\text{E}(z) = b'(\theta)$ and $\text{Var}(z) = b''(\theta)$, where

$$b'(\theta) = -\left(\frac{2}{\theta} + \frac{1}{1-\theta}\right) = \frac{\mu(1+\mu)}{1-\mu},$$

and

$$b''(\theta) = \frac{2}{\theta^2} - \frac{1}{(1-\theta)^2} = \frac{\mu^2\{2-(1-\mu)^2\}}{(1-\mu)^2}.$$

These results agree with the ones obtained in Sect. 2 by considering satisfied the regularity conditions for the score function.

Appendix C

The sensitivity and the variability matrices of $\boldsymbol{\Psi}(\boldsymbol{\beta})$ are, respectively, given by $\mathbf{S}_{n\Psi}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\mathbf{S}_i(\boldsymbol{\beta})$ and $\mathbf{V}_{n\Psi}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\mathbf{V}_i(\boldsymbol{\beta})$, where

$$\mathbf{S}_i(\boldsymbol{\beta}) = \text{E}\{\boldsymbol{\Psi}_i'(\boldsymbol{\beta})\} = \mathbf{X}_i^\top\mathbf{D}_i\boldsymbol{\Omega}_i^{-1}\text{E}\left(\frac{\partial\mathbf{u}_i}{\partial\boldsymbol{\beta}^\top}\right) = \mathbf{X}_i^\top\mathbf{W}_i\mathbf{X}_i,$$

with $\mathbf{W}_i = \mathbf{D}_i\boldsymbol{\Omega}_i^{-1}\mathbf{D}_i$ and $\boldsymbol{\Omega}_i = \boldsymbol{\Sigma}_{u_i}^{\frac{1}{2}}\mathbf{R}_i(\boldsymbol{\rho})\boldsymbol{\Sigma}_{u_i}^{\frac{1}{2}}$, and

$$\mathbf{V}_i(\boldsymbol{\beta}) = \text{E}\left\{\boldsymbol{\Psi}_i(\boldsymbol{\beta})\boldsymbol{\Psi}_i^\top(\boldsymbol{\beta})\right\} = \mathbf{X}_i^\top\mathbf{W}_i\mathbf{D}_i^{-1}\text{Cov}(\mathbf{u}_i)\mathbf{D}_i^{-1}\mathbf{W}_i\mathbf{X}_i.$$

Acknowledgments The authors are grateful to the reviewers for their helpful comments and suggestions. This study was partially supported by CAPES and CNPq, Brazil.

Declarations

Conflict of interest There is no conflict of interest.

References

Abd-Elrahman AM (2013) Utilizing ordered statistics in lifetime distributions production: a new lifetime distribution and applications. J Prob Stat Sci 11:153–164

Adler A (2022) lamW: Lambert-W Function. R package version 2.1.1. https://cran.r-project.org/package=lamW

Akdur HTK (2021) Unit-lindley mixed-effect model for proportion data. J Appl Stat 48:2389–2405

Altun E, El-Morshedy M, Eliwa MS (2021) A new regression model for bounded response variable: an alternative to the beta and unit-lindley regression models. PLoS One 16:1–15

Artes R, Jorgensen B (2000) Longitudinal data estimating equations for dispersion model. Scand J Stat 27:321–334

Atkinson A, Riani M (2000) Robust diagnostic regression analysis. Springer, New York

Barndorff-Nielsen OE, Jørgensen B (1991) Some parametric models on the simplex. J Multivar Anal 39:109–116

Borchers HW (2022) pracma: practical numerical math functions. R package version 2.4.2. https://cran.r-project.org/package=pracma

Cadigan NG, Farrell PJ (2002) Generalized local influence with applications to fish stock cohort analysis. J Appl Stat 51:469–483

Cook RD (1986) Assessment of local influence. J R Stat Soc B 48:133–169

Cook RD, Weisberg S (1982) Residuals and influence in regression. Chapman and Hall/CRC, London

Crowder M (1987) On linear and quadratic estimating functions. Biometrika 74:591–597

Dunn PK, Smyth GK (1996) Randomized quantile residuals. J Comput Graph Stat 5:236–244

Eddelbuettel D (2022) RcppEigen: 'Rcpp' Integration for the 'Eigen' templated linear algebra library. R package version 0.3.3.9.3. https://cran.r-project.org/package=RcppEigen

Eddelbuettel D (2023) Rcpp: Seamless R and C++ Integration. R package version 1.0.10. https://cran.r-project.org/package=Rcpp

Fasiolo M (2023) mvnfast: fast multivariate normal and student's t methods. R package version 0.2.8. https://cran.r-project.org/package=mvnfast

Ferrari SLP, Cribari-Neto F (2004) Beta regression for modelling rates and proportions. J Appl Stat 31:799–815

Galvis DM, Bandyopadhyay D, Lachos VH (2014) Augmented mixed beta regression models for periodontal proportion data. Stat Med 33:3759–3771

Ghitany ME, Atieh B, Nadarajah S (2008) Lindley distribution and its application. Math Comput Simul 78:493–506

Godambe VP (1997) Estimating functions: a synthesis of least squares and maximum likelihood methods. In: Basawa IV , Godambe VP, Taylor RL (eds) Selected proceedings of the symposium on estimating functions. Institute of Mathematical Statistics, California, pp 5–15

Goulet V (2022) expint: exponential integral and incomplete gamma function. R package version 0.1-8. https://cran.r-project.org/package=expint

Grassia A (1977) On a family of distributions with argument between 0 and 1 obtained by transformation of the gamma distribution and derived compound distributions. Aust J Stat 19:108–114

Hardin JW, Hilbe JM (2012) Generalized estimating equations, 2nd edn. Chapman and Hall/CRC, New York

Ibacache P, Paula GA, Cysneiros FJ (2013) Semiparametric additive models under symmetric distributions. TEST 22:103–121

Jorgensen B, Lundbye-Christensen S, Song PX-K, Sun L (1996) State-space models for multivariate longitudinal data of mixed types. Can J Stat 24:385–402

Kumaraswamy P (1980) A generalized probability density function for double-bounded random processes. J Hydrol 46:79–88

Lee SY, Lu B, Song XY (2006) Assessing local influence for nonlinear structural equation models with ignorable mussing data. Comput Stat Data Anal 50:1356–1377

Lee SY, Xu L (2004) Influence analyses of nonlinear mixed-effects models. Comput Stat Data Anal 45:321–341

Liang KY, Zeger SL (1986) Longitudinal analysis using generalized linear models. Biometrika 73:13–22

Maechler M (2022) Matrix: sparse and dense matrix classes and methods. R package version 1.5-3. https://cran.r-project.org/package=Matrix

Manghi RF, Cysneiros FJA, Paula GA (2019) Generalized additive partial linear models for analyzing correlated data. Comput Stat Data Anal 129:49–60
Mazucheli J, Menezes AFB, Chakraborty S (2019) On the one parameter unit-lindley distribution and its associated regression model for proportion data. J Appl Stat 46:700–714
McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman and Hall/CRC, London
Mousa AM, El-Sheikh AA, Abdel-Fattah MA (2016) A gamma regression for bounded continuous variables. Adv Appl Stat 49:305–326
Petterle RR, Bonat WH, Scarpin CT (2019) Quasi-beta longitudinal regression model applied to water quality index data. J Agric Biol Environ Stat 24:346–368
Piegorsch WW, Casella G (1988) Confidence bands for logistic regression with restricted predictor variables. Biometrics 44:739–750
Poon W, Poon YS (1999) Conformal normal curvature and assessment of local influence. J R Stat Soc B 61:51–61
Preisser JS, Qaqish BF (1996) Deletion diagnostics for generalised estimating equations. Biometrika 83:551–562
Qiu Z, Song PX-K, Tan M (2008) Simplex mixed-effects models for longitudinal proportional data. Scand J Stat 35:577–596
R Core Team (2022) R: a language and environment for statistical computing. R foundation for statistical computing. https://www.R-project.org
Solymos P (2023) pbapply: adding progress bar to '*apply' functions. R package version 1.7-0. https://cran.r-project.org/package=pbapply
Stasinopoulos M (2022) gamlss.dist: distributions for generalized additive models for location scale and shape. R package version 6.0-5. https://cran.r-project.org/package=gamlss.dist
Stasinopoulos M (2023) gamlss: generalised additive models for location scale and shape. R package version 5.4-12. https://cran.r-project.org/package=gamlss
Stasinopoulos MD, Righy RA, Gillian ZA, Voudouris V, de Bastiani F (2017) flexible regression and smoothing using GAMLSS in R. Chapman and Hall/CRC, Florida
Tsuyuguchi A, Paula GA, Barros M (2020) Analysis of correlated birnbaum–saunders data based on estimating equations. TEST 29:661–681
Venezuela MK, Sandoval MC, Botter DA (2011) Local influence in estimating equations. Comput Stat Data Anal 55:1867–1883
Wicklin R (2013) Simulating data with SAS. SAS Institute, North Carolina
Yan J, Fine J (2004) Estimating equations for association structures. Stat Med 23:859–874