mgee2: An R package for marginal analysis of longitudinal ordinal data with misclassified responses and covariates

by Yuliang Xu, Shuo Shuo Liu and Grace Y. Yi

Abstract Marginal methods have been widely used for analyzing longitudinal ordinal data due to their simplicity in model assumptions, robustness in inference results, and easiness in the implementation. However, they are often inapplicable in the presence of measurement errors in the variables. Under the setup of longitudinal studies with ordinal responses and covariates subject to misclassification, Chen et al. (2014) developed marginal methods for misclassification adjustments using the second-order estimating equations and proposed a two-stage estimation approach when the validation subsample is available. Parameter estimation is conducted through the Newton-Raphson algorithm, and the asymptotic distribution of the estimators is established. While the methods of Chen et al. (2014) can successfully correct the misclassification effects, its implementation is not accessible to general users due to the lack of a software package. In this paper, we develop an R package, **mgee2**, to implement the marginal methods proposed by Chen et al. (2014). To evaluate the performance and illustrate the features of the package, we conduct numerical studies.

Introduction

Analysis of longitudinal ordinal data is a common research topic in health science and survey sampling. Typically, Liang and Zeger (1986) introduced the generalized estimating equations (GEE) method that gave consistent estimation with mild assumptions of the joint distribution of the repeated measurements. This method has been used widely in analyzing longitudinal binary and categorical data. The validity of the GEE method hinges on the critical condition that data are precisely observed, which is commonly infeasible and violated in practice (Yi, 2017). Extensive discussions about covariate error (Carroll et al., 2006) and response with binary misclassification (Neuhaus, 1999; Chen et al., 2011; Yi, 2017) have been conducted in the literature. For example, Neuhaus (1999) investigated the bias due to errors in the response. Yi (2008) proposed a simulation–extrapolation (SIMEX) method to handle both dropout and covariate measurement error problems in longitudinal studies. Furthermore, in Yi (2017, Ch5), the impact of covariate measurement error on longitudinal data analysis was investigated, and methods of addressing covariate measurement error effects were described.

To accommodate effects induced from error-prone correlated ordinal responses and ordinal covariates simultaneously, Chen et al. (2014) proposed GEE-based methods for the estimation of both mean and association parameters. The proposed methods are based on formulating unbiased second-order estimating functions and solving the resulting equations using the Newton-Raphson algorithm. The asymptotic distributions for the proposed estimators are established. While the methods of Chen et al. (2014) correct for error effects due to misclassified variables, the methods cannot be used by the analysts without programming the implementation procedures. To expedite the use of the methods for problems in applications, in this paper, we develop an R package, called **mgee2**, to implement the methods of Chen et al. (2014).

Our work offers an R package complement to available R packages for analyzing longitudinal data with misclassified observations. It is relevant to but differs from available R packages about measurement error. For example, the package **SAMBA**, developed by Beesley and Mukherjee (2020), provides resources for fitting logistic regression with misclassified binary outcomes. The R package **misclassGLM** implements inferential procedures for generalized linear models with misclassified covariates proposed by Dlugosz et al. (2017); Zhang and Yi (2019) developed the package **augSIMEX** to implement the method proposed by Yi et al. (2015) for fitting generalized linear models with mixed continuous and discrete covariates subject to mismeasurement.

When the degree of measurement error is very severe, the observed surrogate measurements are virtually useless, and hence the corresponding variables may be alternatively treated as subject to missingness. Regarding the analysis of longitudinal data with missing observations, packages **kml** and **kml3d**, developed by Genolini et al. (2015), describe the implementation procedures of *k*-means for longitudinal clustered data with missing observations. Carey (2015) developed the package **gee** to solve generalized estimation equations with longitudinal data missing completely at random. Xu et al. (2018) developed the package **wgeesel** for using weighted generalized estimating equations approaches to analyze longitudinal clustered data with data missing at random. Xiong and Yi (2019) developed the package **swgee** for analyzing longitudinal data with missingness in the response and

measurement error in covariates. Our package **mgee2** differs from those packages in its ability to simultaneously handle the features of misclassification in correlated ordinal responses and ordinal covariates, which to our best knowledge, is the first software package to address this problem.

The article is organized as follows. Section *Model setup* introduces the notations and estimation procedures proposed by Chen et al. (2014). Section *Package details* describes the usage of the package **mgee2**. Section *Data analysis* illustrates the package by simulation studies and a real dataset. We finally conclude the article in Section *Summary*.

Model setup

We first review the notation and formulations of Chen et al. (2014). For $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$, let Y_{ij} denote an error-prone ordinal response variable for subject i at visit j. Suppose that the response variable Y_{ij} has $(K+1)$ categories, denoted $0, 1, \ldots, K$, and that an error-prone ordinal covariate $\boldsymbol{X}_{ij}$ has $(H+1)$ categories, denoted $0, 1, \ldots, H$. Let $\boldsymbol{X}_{ij} = \left(X_{ij1}, \ldots, X_{ijH}\right)^T$ be the misclassification-prone vector of binary variables such that $X_{ijq} = I(\text{the covariate } \boldsymbol{X}_{ij} \text{ in category } q)$ for $q = 0, 1, \ldots, H$, and let $\boldsymbol{Z}_{ij}$ be the vector of covariates that are free of measurement error, where $I(\cdot)$ is the indicator function. Furthermore, we define $\boldsymbol{X}_i = \left(\boldsymbol{X}_{i1}^T, \ldots, \boldsymbol{X}_{im_i}^T\right)^T$ and $\boldsymbol{Z}_i = \left(\boldsymbol{Z}_{i1}^T, \ldots, \boldsymbol{Z}_{im_i}^T\right)^T$.

Response process

Let

$$\lambda_{ijk} = P\left(\mathsf{Y}_{ij} \geq k | \boldsymbol{X}_i, \boldsymbol{Z}_i\right) \tag{1}$$

be the univariate cumulative probability with $k = 1, \ldots, K$, and adopt the assumption $P\left(\mathsf{Y}_{ij} \geq k | \boldsymbol{X}_i, \boldsymbol{Z}_i\right) = P\left(\mathsf{Y}_{ij} \geq k | \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}\right)$ (Pepe and Anderson, 1994). Consider the proportional odds models

$$\text{logit } \lambda_{ijk} = \boldsymbol{\beta}_{0k} + \boldsymbol{X}_{ij}^T \boldsymbol{\beta}_x + \boldsymbol{Z}_{ij}^T \boldsymbol{\beta}_z,$$

where $\boldsymbol{\beta}_{0k}$, $\boldsymbol{\beta}_x$, and $\boldsymbol{\beta}_z$ are regression parameters, $k = 1, \ldots, K$, $j = 1, \ldots, m_i$, and $i = 1, \ldots, n$. Similar to Williamson et al. (1995), we measure the association between a pair of responses for the same subject at two different visits by the global odds ratio

$$\psi_{i,jk,j'k'} = \frac{P\left(\mathsf{Y}_{ij} \geq k, \mathsf{Y}_{ij'} \geq k' | \boldsymbol{X}_i, \boldsymbol{Z}_i\right) \times P\left(\mathsf{Y}_{ij} < k, \mathsf{Y}_{ij'} < k' | \boldsymbol{X}_i, \boldsymbol{Z}_i\right)}{P\left(\mathsf{Y}_{ij} \geq k, \mathsf{Y}_{ij'} < k' | \boldsymbol{X}_i, \boldsymbol{Z}_i\right) \times P\left(\mathsf{Y}_{ij} < k, \mathsf{Y}_{ij} \geq k' | \boldsymbol{X}_i, \boldsymbol{Z}_i\right)}, \tag{2}$$

where $k, k' = 1, \ldots, K$, and $j \neq j'$. To characterize the dependence of the global odds ratios on covariates, the log-linear models can be expressed as

$$\log \psi_{i,jk,j'k'} = \phi + \phi_k + \phi_{k'} + \phi_{kk'} + \mathbf{u}_{ijj'}^T \boldsymbol{\alpha}_1,$$

where ϕ is the global intercept, ϕ_k and $\phi_{k'}$ correspond to the effect of category k and of category k', respectively, $\phi_{kk'}$ is the interaction effect between categories k and k' with $\phi_{kk'} = \phi_{k'k}$, and $\boldsymbol{\alpha}_1$ is a vector of parameters corresponding to pair-specific covariates, denoted $\mathbf{u}_{ijj'}$. The constraint $\phi_1 = \phi_{1k} = \phi_{k1} = 0$ is set for the model identification for $k = 1, \ldots, K$ (Williamson et al., 1995).

Let $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_{0k}^T, \boldsymbol{\beta}_x^T, \boldsymbol{\beta}_z^T\right)^T$, $\boldsymbol{\alpha} = \left(\phi, \phi_k, \phi_{kk'}, \boldsymbol{\alpha}_1^T\right)^T$, and $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T\right)^T$. For $k = 1, \ldots, K$, let $\boldsymbol{\mathsf{Y}}_{ij} = \left(\mathsf{Y}_{ij1}, \ldots, \mathsf{Y}_{ijk}\right)^T$ with $\mathsf{Y}_{ijk} = I\left(\mathsf{Y}_{ij} = k\right)$. Define $\boldsymbol{\mathsf{Y}}_i = \left(\boldsymbol{\mathsf{Y}}_{i1}^T, \ldots, \boldsymbol{\mathsf{Y}}_{im_i}^T\right)^T$. For $j < j'$, let $C_{i,jk,j'k'} = \mathsf{Y}_{ijk}\mathsf{Y}_{ij'k'}$, $\boldsymbol{C}_{ijj'} = \left(C_{i,j1,j'1}, C_{i,j1,j'2}, \ldots, C_{i,jK,j'K'}\right)^T$, and $\boldsymbol{C}_i = \left(\boldsymbol{C}_{ijj'}^T, j < j'\right)^T$. The univariate and bivariate marginals, $\boldsymbol{\mu}_i = E\left(\boldsymbol{\mathsf{Y}}_i | \boldsymbol{X}_i, \boldsymbol{Z}_i\right)$ and $\boldsymbol{\xi}_i = E\left(\boldsymbol{C}_i | \boldsymbol{X}_i, \boldsymbol{Z}_i\right)$, can be expressed in terms of the global odds ratios and univariate and bivariate cumulative probabilities; the detailed expressions are given by Chen et al. (2014).

As a result, the estimating functions for the mean and association parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are given by

$$\boldsymbol{U}_{1i}\left(\boldsymbol{\theta}; \boldsymbol{\mathsf{Y}}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i\right) = \boldsymbol{D}_{1i} \boldsymbol{V}_{1i}^{-1}\left(\boldsymbol{\mathsf{Y}}_i - \boldsymbol{\mu}_i\right) \tag{3}$$

and

$$\boldsymbol{U}_{2i}\left(\boldsymbol{\theta}; \boldsymbol{\mathsf{Y}}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i\right) = \boldsymbol{D}_{2i} \boldsymbol{V}_{2i}^{-1}\left(\boldsymbol{C}_i - \boldsymbol{\xi}_i\right), \tag{4}$$

respectively, where $\boldsymbol{D}_{1i} = \partial \boldsymbol{\mu}_i^T / \partial \boldsymbol{\beta}$, $\boldsymbol{D}_{2i} = \partial \boldsymbol{\xi}_i^T / \partial \boldsymbol{\alpha}$, and $\boldsymbol{V}_{1i}$ and $\boldsymbol{V}_{2i}$ are the conditional covariance

matrices for $\mathbf{Y}_i$ and $\boldsymbol{C}_i$, given $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$.

Case 1: Estimation with known misclassification probabilities

If the true measurements of the responses and covariates are available, (3) and (4) can be used for estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. However, in applications, the $\mathbf{Y}_{ij}$ and the $\boldsymbol{X}_{ij}$ may be subject to misclassification. Let $\boldsymbol{S}_{ij}$ and $\boldsymbol{W}_{ij}$ be surrogate measurements of $\mathbf{Y}_{ij}$ and $\boldsymbol{X}_{ij}$, respectively. Let $\tau_{ijkl} = P\left(\boldsymbol{S}_{ij} = l|\mathbf{Y}_{ij} = k, \boldsymbol{Z}_i\right)$ be the conditional probability concerning the response for subject i at visit j where $k, l = 0, \ldots, K$. Let $\pi_{ijqr} = P\left(\boldsymbol{W}_{ij} = r|\boldsymbol{X}_{ij} = q, \boldsymbol{Z}_i\right)$ be the conditional probability concerning the covariate for subject i at visit j where $q, r = 0, \ldots, H$. Consider the generalized logistic models by setting category 0 as the reference:

$$\log\left(\tau_{ijkl}/\tau_{ijk0}\right) = \boldsymbol{L}_{ij}^T\boldsymbol{\gamma}_{kl} \quad \text{for} \quad l = 1, \ldots, K; k = 0, \ldots, K$$

and

$$\log\left(\pi_{ijqr}/\pi_{ijq0}\right) = \boldsymbol{L}_{ij}^{xT}\boldsymbol{\varphi}_{qr} \quad \text{for} \quad r = 1, \ldots, H; q = 0, \ldots, H,$$

where $\boldsymbol{L}_{ij}$ and $\boldsymbol{L}_{ij}^x$ are vectors of variables related to response and covariate misclassification processes, respectively, and $\boldsymbol{\gamma}_{kl}$ and $\boldsymbol{\varphi}_{qr}$ are vectors of regression parameters.

Let $\boldsymbol{\gamma}_k = \left(\boldsymbol{\gamma}_{k1}^T, \ldots, \boldsymbol{\gamma}_{kK}^T\right)^T$ and $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_0^T, \ldots, \boldsymbol{\gamma}_K^T\right)^T$. Let $\boldsymbol{\varphi}_q = \left(\boldsymbol{\varphi}_{q1}^T, \ldots, \boldsymbol{\varphi}_{qH}^T\right)^T$ and $\boldsymbol{\varphi} = \left(\boldsymbol{\varphi}_0^T, \ldots, \boldsymbol{\varphi}_H^T\right)^T$. Let $\boldsymbol{\eta} = \left(\boldsymbol{\gamma}^T, \boldsymbol{\varphi}^T\right)^T$. Define the $K \times K$ matrix $\boldsymbol{R}_{ij} = \left(\boldsymbol{\tau}_{ij1} - \boldsymbol{\tau}_{ij0}, \ldots, \boldsymbol{\tau}_{ijK} - \boldsymbol{\tau}_{ij0}\right)$ and the $H \times H$ matrix $\boldsymbol{G}_{ij} = \left(\boldsymbol{\pi}_{ij1} - \boldsymbol{\pi}_{ij0}, \ldots, \boldsymbol{\pi}_{ijK} - \boldsymbol{\pi}_{ij0}\right)$, where $\boldsymbol{\tau}_{ijk} = \left(\tau_{ijk1}, \ldots, \tau_{ijkK}\right)^T$ and $\boldsymbol{\pi}_{ijk} = \left(\pi_{ijk1}, \ldots, \pi_{ijkK}\right)^T$. Then the unbiased surrogates for $\mathbf{Y}_{ij}$ and $\boldsymbol{X}_{ij}$ are constructed, respectively, by

$$\mathbf{Y}_{ij}^* = \boldsymbol{R}_{ij}^{-1}\left(\boldsymbol{S}_{ij} - \boldsymbol{\tau}_{ij0}\right)$$

and

$$\boldsymbol{X}_{ij}^* = \boldsymbol{G}_{ij}^{-1}\left(\boldsymbol{W}_{ij} - \boldsymbol{\pi}_{ij0}\right),$$

where we write $\mathbf{Y}_{ij}^* = \left(Y_{ij1}^*, \ldots, Y_{ijK}^*\right)^T$, $\boldsymbol{X}_{ij}^* = \left(X_{ij1}^*, \ldots, X_{ijK}^*\right)^T$, and let $\mathbf{Y}_i^* = \left(\mathbf{Y}_{i1}^T, \ldots, \mathbf{Y}_{im_i}^{*T}\right)^T$. Let $\boldsymbol{e}_q$ be an H-dimensional vector whose rth element is an indicator $I(r = q)$ for $q = 1, \ldots, H$ and let $\boldsymbol{e}_0 = 0$.

If $\boldsymbol{\eta}$ is known, then

$$\boldsymbol{U}_{1i}^*(\boldsymbol{\theta}) = \sum_{q_{m_i}=0}^{H} \cdots \sum_{q_1=0}^{H} \left[\boldsymbol{U}_{1i}\left\{\theta; \mathbf{Y}_i^*, \left(\boldsymbol{e}_{q_1}^T, \ldots, \boldsymbol{e}_{q_{i_i}}^T\right)^T, \boldsymbol{Z}_i\right\} \prod_{j=1}^{m_i} X_{ijq_j}^*\right]$$

and

$$\boldsymbol{U}_{2i}^*(\boldsymbol{\theta}) = \sum_{q_{m_i}=0}^{H} \cdots \sum_{q_1=0}^{H} \left[\boldsymbol{U}_{2i}\left\{\theta; \mathbf{Y}_i^*, \left(\boldsymbol{e}_{q_1}^T, \ldots, \boldsymbol{e}_{q_{i_i}}^T\right)^T, \boldsymbol{Z}_i\right\} \prod_{j=1}^{m_i} X_{ijq_j}^*\right]$$

are unbiased estimating functions of $\boldsymbol{\theta}$, as shown in Appendix 2 of Chen et al. (2014). Estimation of $\boldsymbol{\theta}$ can then be obtained by solving

$$\sum_{i=1}^{n} \left\{ \begin{array}{c} \boldsymbol{U}_{1i}^*(\boldsymbol{\theta}) \\ \boldsymbol{U}_{2i}^*(\boldsymbol{\theta}) \end{array} \right\} = \boldsymbol{0} \tag{5}$$

for $\boldsymbol{\theta}$.

Case 2: Estimation with validation data

Case 1 highlights the estimation of $\boldsymbol{\theta}$ when the parameter $\boldsymbol{\eta}$ for the misclassification models is known or specified as a given value. In applications, $\boldsymbol{\eta}$ is unknown and may be estimated from a validation subsample. In this case, we modify the estimation procedure based on (5) and describe a two-stage estimation procedure. First, let $\delta_{ij} = I$(subject i at visit j is included in the validation subsample). Using validation data (i.e., $\delta_{ij} = 1$), we may estimate τ_{ij} and π_{ij}.

Define $\boldsymbol{D}_{\gamma ij} = \partial\boldsymbol{\tau}_{ij}^T/\partial\boldsymbol{\gamma}$ and $\boldsymbol{D}_{\boldsymbol{\varphi} ij} = \partial\boldsymbol{\pi}_{ij}^T/\partial\boldsymbol{\varphi}$, then estimating functions for $\boldsymbol{\gamma}$ and $\boldsymbol{\varphi}$ are given by $\boldsymbol{Q}_{\gamma i}(\boldsymbol{\gamma}) = \sum_{j=1}^{m_i} \boldsymbol{D}_{\gamma ij}\boldsymbol{V}_{\gamma ij}^{-1}\left\{\boldsymbol{S}_{ij} - \boldsymbol{\tau}_{ij}\right\}\delta_{ij}$ and $\boldsymbol{Q}_{\varphi i}(\boldsymbol{\varphi}) = \sum_{j=1}^{m_i} \boldsymbol{D}_{\varphi ij}\boldsymbol{V}_{\varphi ij}^{-1}\left\{\boldsymbol{W}_{ij} - \boldsymbol{\pi}_{ij}\right\}\delta_{ij}$, where $\boldsymbol{V}_{\gamma ij}$ and $\boldsymbol{V}_{\boldsymbol{\varphi} ij}$ are, respectively, the conditional covariance matrix for $\boldsymbol{S}_{ij}$ and $\boldsymbol{W}_{ij}$, given $\mathbf{Y}_{ij}$ and the true covariates.

Let $\tilde{Y}_{ijk} = Y_{ijk}$ if $\delta_{ij} = 1$ and $\tilde{Y}_{ijk} = Y^*_{ijk}$ otherwise, then we write $\tilde{\mathbf{Y}}_{ij} = \left(\tilde{Y}_{ij1}, \ldots, \tilde{Y}_{ijK}\right)^T$. Let $\tilde{X}_{ijq} = X_{ijq}$ if $\delta_{ij} = 1$ and $\tilde{X}_{ijq} = X^*_{ijq}$ otherwise. Then the augmented estimating functions of $\boldsymbol{\theta}$ are given by

$$\tilde{\boldsymbol{U}}_{1i}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \sum_{q_{m_i}=0}^{H} \cdots \sum_{q_1=0}^{H} \left[\boldsymbol{U}_{1i} \left\{ \theta; \tilde{\mathbf{Y}}_i, \left(\boldsymbol{e}_{q_1}^T, \ldots, \boldsymbol{e}_{q_m}^T \right)^T, \boldsymbol{Z}_i \right\} \prod_{j=1}^{m_i} \tilde{X}_{ijq_j} \right] \tag{6}$$

and

$$\tilde{\boldsymbol{U}}_{2i}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \sum_{q_{m_i}=0}^{H} \cdots \sum_{q_1=0}^{H} \left[\boldsymbol{U}_{2i} \left\{ \theta; \tilde{\mathbf{Y}}_i, \left(\boldsymbol{e}_{q_1}^T, \ldots, \boldsymbol{e}_{q_m}^T \right)^T, \boldsymbol{Z}_i \right\} \prod_{j=1}^{m_i} \tilde{X}_{ijq_j} \right]. \tag{7}$$

Consequently, estimation of $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ can be carried out by the two-stage procedure.

Stage 1. Solve $\sum_{i=1}^{n} \begin{Bmatrix} \boldsymbol{Q}_{\gamma i}(\gamma) \\ \boldsymbol{Q}_{\varphi i}(\boldsymbol{\varphi}) \end{Bmatrix} = \mathbf{0}$ for γ and $\boldsymbol{\varphi}$ and write $\hat{\boldsymbol{\eta}} = \left(\hat{\gamma}^T, \hat{\boldsymbol{\varphi}}^T\right)^T$, where $\hat{\gamma}$ and $\hat{\boldsymbol{\varphi}}$ are the estimators for γ and $\boldsymbol{\varphi}$, respectively.

Stage 2. Substitute $\boldsymbol{\eta}$ with $\hat{\boldsymbol{\eta}}$ in (6) and (7) and solve $\sum_{i=1}^{n} \begin{Bmatrix} \tilde{\boldsymbol{U}}_{1i}(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}) \\ \tilde{\boldsymbol{U}}_{2i}(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}) \end{Bmatrix} = \mathbf{0}$ for $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\alpha}}^T\right)^T$ denote the resulting estimator $\boldsymbol{\theta}$.

Chen et al. (2014) established the asymptotic distribution of $\hat{\boldsymbol{\theta}}$. Let $\tilde{\boldsymbol{U}}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) = \left\{ \tilde{\boldsymbol{U}}_{1i}^T(\boldsymbol{\theta}, \boldsymbol{\eta}), \tilde{\boldsymbol{U}}_{2i}^T(\boldsymbol{\theta}, \boldsymbol{\eta}) \right\}^T$, $\boldsymbol{Q}_i(\boldsymbol{\eta}) = \left\{ \boldsymbol{Q}_{\gamma i}^T(\gamma), \boldsymbol{Q}_{\varphi i}^T(\boldsymbol{\varphi}) \right\}^T$, $\boldsymbol{\Omega}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) = \tilde{\boldsymbol{U}}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) - E\left\{ \partial \tilde{\boldsymbol{U}}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) / \partial \boldsymbol{\eta}^T \right\} \cdot \left[E\left\{ \partial \boldsymbol{Q}_i(\boldsymbol{\eta}) / \partial \boldsymbol{\eta}^T \right\} \right]^{-1}$, and $\tilde{\boldsymbol{\Gamma}}(\boldsymbol{\theta}, \boldsymbol{\eta}) = E\left\{ \partial \tilde{\boldsymbol{U}}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) / \partial \boldsymbol{\theta}^T \right\}$. Then, under regularity conditions, $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normally distributed with mean $\mathbf{0}$ and covariance matrix $\tilde{\boldsymbol{\Gamma}}^{-1} \tilde{\boldsymbol{\Sigma}} \left(\tilde{\boldsymbol{\Gamma}}^{-1} \right)^T$, where $\tilde{\boldsymbol{\Sigma}} = E\left\{ \boldsymbol{\Omega}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) \boldsymbol{\Omega}_i^T(\boldsymbol{\theta}, \boldsymbol{\eta}) \right\}$.

Package details

We develop an R package, called **mgee2**, to implement the misclassification adjustment method described in the preceding section. This package requires support from the external packages **MASS** (Venables and Ripley, 2002), **Matrix** (Bates and Maechler, 2019), and **ggplot2** (Wickham, 2016). Our **mgee2** package mainly contains two functions, `mgee2k` and `mgee2v`, respectively, implementing cases 1 and 2 described in the previous section. Specifically, `mgee2k` implements the method where the misclassification parameters are given, and `mgee2v` implements the misclassification method for the case where validation data are available to estimate misclassification probabilities. We now describe the details of these two functions.

mgee2k

`mgee2k` implements the misclassification adjustment method outlined in Case 1 of the previous section, where the misclassification parameters are known. In this case, validation data are not required, and only the observed data of the outcome and covariates are needed for the implementation.

The function `mgee2k` requires the data set to be grouped by the individual id, $i = 1, ..., n$, and each individual has m_i rows of data each corresponding to the visit time $j = 1, ..., m_i$. The column name of the individual id is indicated by the argument `id`. The misclassification matrices for the response and covariate variables are recorded by the arguments `gamMat` and `varphiMat`, respectively, which need to be specified by the user.

To call `mgee2k`, we issue the following command,

```
mgee2k(formula, id, data, corstr="exchangeable", misvariable,
            gamMat, varphiMat, maxit=50, tol=1e-3)
```

where the meaning of each argument is described as follows:

- `formula`: a formula object which specifies the relationship between the response and covariates for the observed data.
- `id`: a character object which records individual id in the data.
- `data`: a dataframe or matrix object for the observed data set.
- `corstr`: a character object. The default value is "exchangeable", corresponding to the structure where the association between two paired responses is considered to be a constant. The other option is "log-linear" which indicates the log-linear association between two paired responses.

- misvariable: a character object which names the error-prone covariate W.
- maxit: an integer which specifies the maximum number of iterations. The default is 50.
- tol: a numeric object which indicates the tolerance threshold. The default is 1e-3.
- gamMat: a matrix object which records the misclassification parameter γ for response Y.
- varphiMat: a matrix object which records the misclassification parameter ϕ for covariate X.

The function mgee2k returns a list of components:

- beta: the coefficients in the same order as that specified in the formula for the response and covariates.
- alpha: the coefficients for paired responses global odds ratios. The number of α coefficients corresponds to the paired responses odds ratio structure selected in corstr. When corstr="exchangeable", only one baseline α is fitted.
- variance: the variance-covariance matrix of the estimators of all parameters.
- convergence: a logical variable; TRUE if the model converges.
- iteration: the number of iterations for the estimates of the model parameters to converge.
- call: an unevaluated function call which consists of the named function applied to the given arguments.

mgee2v

The function mgee2v does not require the misclassification parameters to be known, but requires the availability of validation data.

Similar to mgee2k, the function mgee2v needs the data set to be structured by individual id, $i = 1, ..., n$, and visit time, $j = 1, ..., m_i$. The data set should contain the observed response and covariates, S and W. To indicate whether or not a subject is in the validation set, an indicator variable delta should be added in the data set, and we use a column named valid.sample.ind for this purpose. The column name of the error-prone covariate W should also be specified in misvariable.

To call mgee2v, we issue the command,

```
mgee2v(formula, id, data, corstr="exchangeable", misvariable, valid.sample.ind,
               y.mcformula, x.mcformula, maxit=50, tol=1e-3)
```

where the arguments are described as follows:

- formula: a formula object which specifies the relationship between the response and covariates for the observed data.
- id: a character object which records individual id in the data.
- data: a dataframe or matrix object for the observed data set.
- corstr: a character object. The default value is "exchangeable", corresponding to the structure where the association between two paired responses is considered to be a constant. The other option is "log-linear" which indicates the log-linear association between two paired responses.
- misvariable: a character object which names the error-prone covariate W.
- valid.sample.ind: a string object which names the indicator variable delta. When a data point belongs to the validation set, delta = 1; otherwise 0.
- y.mcformula: a string object which indicates the misclassification formula between true response Y and the surrogate response S.
- x.mcformula: a string object which indicates the misclassification formula between true error-prone covariate X and the surrogate W.
- maxit: an integer which specifies the maximum number of iterations. The default is 50.
- tol: a numeric object which indicates the tolerance threshold. The default is 1e-3.

The function mgee2v returns a list of components:

- beta: the coefficients in the same order as that specified in the formula for the response and covariates.
- alpha: the coefficients for paired responses global odds ratios. The number of α coefficients corresponds to the paired responses odds ratio structure selected in corstr. When corstr="exchangeable", only one baseline α is fitted.

- `variance`: the variance-covariance matrix of the estimators of all parameters.
- `convergence`: a logical variable; TRUE if the model converges.
- `iteration`: the number of iterations for the estimates of the model parameters to converge.
- `call`: an unevaluated function call which consists of the named function applied to the given arguments.

ordGEE2

In addition to developing the package **mgee2** to implement the methods of Chen et al. (2014), which accommodate misclassification effects in inferential procedures, we also implement the naive method of ignoring the feature of misclassification and call the resulting function `ordGEE2`. This function can be used together with the preceding described `mgee2k` or `mgee2v` to evaluate the impact of not addressing misclassification effects:

```
ordGEE2(formula, id, data, corstr = "exchangeable", maxit = 50, tol = 0.001)
```

Data analysis

In this section, we conduct numerical studies to demonstrate the usage of our developed R package as well as to show supplementary functions such as summary and plot functions in this package. We first demonstrate all of the external functions in `mgee2` through an example with a simulated data set, known as `obs1`, provided in our package.

An example

The simulated data set, called "`obs1`", includes 8 columns and 3000 rows, with each patient having 3 entries of visits. The format of this data set is as follows.

```
> head(obs1)
ID    Y    X treatment visit S W delta
1  1    2    2         1     1 2 2     1
2  1    0    0         1     2 0 0     1
3  1 <NA> <NA>         1     3 1 2     0
4  2 <NA> <NA>         1     1 1 0     0
5  2 <NA> <NA>         1     2 0 1     0
6  2 <NA> <NA>         1     3 0 0     0
> summary(obs1)
ID             Y           X        treatment visit
Min.   :   1.0   0   : 352   0   : 444   0:1500    1:1000
1st Qu.: 250.8   1   : 283   1   : 269   1:1500    2:1000
Median : 500.5   2   : 256   2   : 178             3:1000
Mean   : 500.5   NA's:2109   NA's:2109
3rd Qu.: 750.2
Max.   :1000.0
S        W            delta
0:1181   0:1460   Min.   :0.000
1: 955   1: 944   1st Qu.:0.000
2: 864   2: 596   Median :0.000
Mean   :0.297
3rd Qu.:1.000
Max.   :1.000
```

Here, `Y` and X represent the true outcome and covariate variables, both being ordinal variables, each taking 3 possible values, denoted 0, 1, and 2, whereas S and W are the observed surrogates for `Y` and X, respectively, with a 5% misclassification rate. `delta` is 1 when the subject is in the validation set and 0 otherwise. About 30% of subjects are randomly chosen to be included in the validation set. We include the subscripts i and j to `Y` and X to indicate the measurements for the corresponding variables for subject i at time point j, in considering the proportional odds model indicated by (1),

$$\text{logit}\,\lambda_{ijk} = \beta_{0k} + \beta_{X1}X_{ij1} + \beta_{X2}X_{ij2} + \beta_{Z1}Z_{ij1} + \beta_{Z2}Z_{ij2} + \beta_{Z3}Z_{ij3} \text{ for } k = 1, 2,$$

where λ_{ijk} is defined as for (1); the treatment variable, denoted Z_{ij1}, is an error-free binary variable; we simulated 3 visits for each patient, denoted by dummy variables Z_{ij2} and Z_{ij3}, with the first visit

as a reference level; X_{ij1} and X_{ij2} represent the dummy variables to reflect the three levels of the error-prone covariate, X_{ij}; and $(\beta_{0k}, \beta_{x1}, \beta_{x2}, \beta_{z1}, \beta_{z2}, \beta_{z3})^T$ is the vector of regression coefficients.

In the case `corstr = "exchangeable"`, the association, defined as in (2), between paired responses is assumed to be

$$\log \psi_{i,jk,j'k'} = \phi;$$

while in the case `corstr = "log-linear"`, the association is assumed to be

$$\log \psi_{i,jk,j'k'} = \phi + \phi_2 I(k = 2) + \phi_2 I(k' = 2) + \phi_{22} I(k = 2, k' = 2),$$

where ϕ, ϕ_2, and ϕ_{22} are parameters.

We now apply `mgee2k` and `mgee2v`, in contrast to `ordGEE2`, to fit the data to the models, respectively. The results are displayed as follows. In the summary tables for the R output, we use `"Y>=1"` and `"Y>=2"` to denote the coefficients β_{01} and β_{02}, respectively, and let `"Delta"` correspond to the parameter ϕ in the dependence structure.

mgee2k

To use function `mgee2k`, we need to specify the misclassification matrices beforehand. Here, we set the misclassification matrices the same as used in the simulation process.

```
> data(obs1)
> obs1$visit <- as.factor(obs1$visit)
> obs1$treatment <- as.factor(obs1$treatment)
> obs1$S <- as.factor(obs1$S)
> obs1$W <- as.factor(obs1$W)
> ## set misclassification parameters to be known.
> varphiMat <- gamMat <- log( cbind(0.04/0.95, 0.01/0.95,
+                                   0.95/0.03, 0.02/0.03,
+                                   0.04/0.01, 0.95/0.01) )
> mgee2k.fit = mgee2k(formula = S~W+treatment+visit, id = "ID", data = obs1,
+                  corstr = "exchangeable", misvariable = "W", gamMat = gamMat,
+                  varphiMat = varphiMat)
> summary(mgee2k.fit)
Call:
mgee2k(formula = S ~ W + treatment + visit, id = "ID", data = obs1,
corstr = "exchangeable", misvariable = "W", gamMat = gamMat,
varphiMat = varphiMat)

Summary table of the estimation
            Estimate  Std.Err Z value   Pr(>z)
Y>=1         0.70889  0.08591   8.251 2.22e-16 ***
Y>=2        -0.67521  0.08625  -7.828 4.88e-15 ***
W1           0.58667  0.08719   6.729 1.71e-11 ***
W2           0.94948  0.09745   9.743  < 2e-16 ***
treatment1 -0.70554  0.09114  -7.742 9.77e-15 ***
visit2     -0.24147  0.07735  -3.122   0.0018 **
visit3     -0.62480  0.07571  -8.253 2.22e-16 ***
Delta        1.22606  0.12231  10.024  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

mgee2v

To use `mgee2v`, a column of indicator variable should be specified in `valid.sample.ind`.

```
> data(obs1)
>   obs1$visit <- as.factor(obs1$visit)
>   obs1$treatment <- as.factor(obs1$treatment)
>   obs1$S <- as.factor(obs1$S)
>   obs1$W <- as.factor(obs1$W)
>   mgee2v.fit = mgee2v(formula = S~W+treatment+visit, id = "ID", data = obs1,
+                       y.mcformula = "S~1", x.mcformula = "W~1",
+                       misvariable = "W", valid.sample.ind = "delta",
+                       corstr = "exchangeable")
> summary(mgee2v.fit)
Call:
mgee2v(formula = S ~ W + treatment + visit, id = "ID",
```

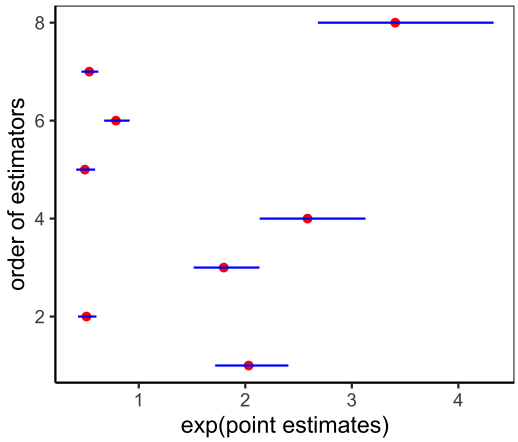

Figure 1: The display of the results in the summary table of applying **mgee2k** method to the example. The vertical axis presents the estimates for the coefficients corresponding to Y>=1, Y>=2, ..., and Delta in the order from the bottom to the top. The horizontal axis shows exp(point estimates) (shown in red dots) for those coefficients indicated by the vertical axis, together with their 95% confidence intervals (shown in blue line segments). The confidence intervals are calculated as $(\exp(C_L), \exp(C_U))$, where (C_L, C_U) is a 95% confidence interval of a coefficient.

```
data = obs1, corstr = "exchangeable", misvariable = "W",
valid.sample.ind = "delta", y.mcformula = "S~1", x.mcformula = "W~1")

Summary table of the estimation
            Estimate  Std.Err Z value   Pr(>z)
Y>=1         0.64876  0.08851   7.330 2.30e-13 ***
Y>=2        -0.68226  0.08703  -7.839 4.44e-15 ***
W1           0.56507  0.08140   6.942 3.88e-12 ***
W2           0.98411  0.09305  10.577  < 2e-16 ***
treatment1 -0.68153  0.09052  -7.529 5.11e-14 ***
visit2     -0.24694  0.07483  -3.300 0.000966 ***
visit3     -0.60027  0.07335  -8.184 2.22e-16 ***
Delta       1.22862  0.12160  10.103  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

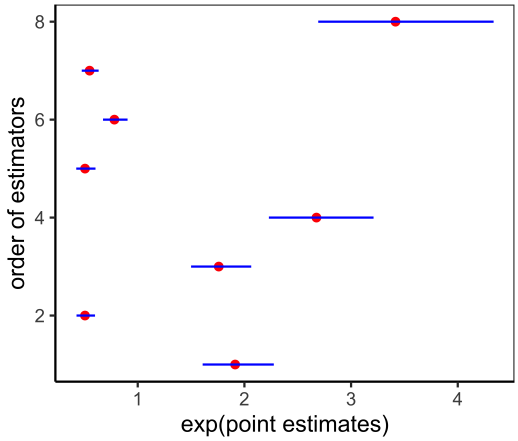

Figure 2: The display of the results in the summary table of applying **mgee2v** method to the example. The vertical axis presents the estimates for the coefficients corresponding to Y>=1, Y>=2, ..., and Delta in the order from the bottom to the top. The horizontal axis shows exp(point estimates) (shown in red dots) for those coefficients indicated by the vertical axis, together with their 95% confidence intervals (shown in blue line segments). The confidence intervals are calculated as $(\exp(C_L), \exp(C_U))$, where (C_L, C_U) is a 95% confidence interval of a coefficient.

ordGEE2

```
> naigee.fit = ordGEE2(formula = S~W+treatment+visit, id = "ID",
+                      data = obs1, corstr = "exchangeable")
> summary(naigee.fit)
Call:
ordGEE2(formula = S ~ W + treatment + visit, id = "ID",
data = obs1, corstr = "exchangeable")

Summary table of the estimation
            Estimate  Std.Err Z value   Pr(>z)
Y>=1         0.73276  0.07990   9.171  < 2e-16 ***
Y>=2        -0.69330  0.08004  -8.662  < 2e-16 ***
W1           0.51237  0.07354   6.967 3.23e-12 ***
W2           0.84890  0.08582   9.892  < 2e-16 ***
treatment1 -0.65954  0.08511  -7.749 9.33e-15 ***
visit2     -0.22766  0.07241  -3.144  0.00167 **
visit3     -0.58407  0.07052  -8.282 2.22e-16 ***
Delta       1.06616  0.09846  10.828  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1
```

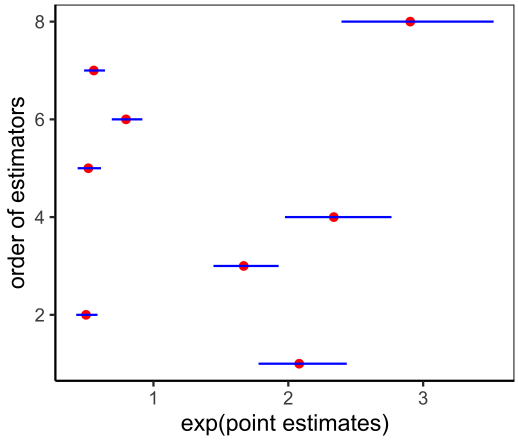

Figure 3: The display of the results in the summary table of applying **ordGEE2** method to the example. The vertical axis presents the estimates for the coefficients corresponding to Y>=1, Y>=2, ..., and Delta in the order from the bottom to the top. The horizontal axis shows exp(point estimates) (shown in red dots) for those coefficients indicated by the vertical axis, together with their 95% confidence intervals (shown in blue line segments). The confidence intervals are calculated as $(\exp(C_L), \exp(C_U))$, where (C_L, C_U) is a 95% confidence interval of a coefficient.

plot_model

We use the function `plot_model` to compare the results obtained from the three functions:

```
> plot_model(naigee.fit)
> plot_model(mgee2.fit)
> plot_model(mgee2v.fit)
```

It is helpful to compare the odds ratios when there are multiple covariates. We use the function `plot_model` to visualize the odds ratios. The estimated odds ratios for this simulated data set across the three methods are displayed in Figure 1, 2, and 3. The red dot gives the odds ratio of each covariate. The horizontal blue line measures the length of each confidence interval. The vertical axes of the graphs indicate the descending order of the covariates. In other words, the red points from the lowest to the highest in the graph represent the first covariate, the second covariate, and so on. It is seen that the three methods yield similar odds ratios.

Simulation studies

To further compare the three methods, a simulation study is conducted. We run 500 simulations where each data set includes 1000 subjects, with three visits for each subjects. `obs1` is one example of the

simulated data. The true values of the coefficients are reported in Table 1:

β_{01}	β_{02}	β_{X1}	β_{X2}	β_{Z1}	β_{Z2}	β_{Z3}
$\log 2$	$\log(1/2)$	$\log 2$	$\log 3$	$\log(1/2)$	$\log(3/4)$	$\log(1/2)$

Table 1: True coefficients.

Table 2 reports the simulation results for the care with a 5% misclassification rate set for both the response and covariate variables, where Bias% records a bias in percentage, EV represents an empirical variance, AMV stands for an average of model-based variance, and CR records a coverage rate of 95% confidence intervals. Simulation results show that the `mgee2` and `mgee2v` perform better than the naive method `ordGEE2`, and they produce reasonable results.

	ordGEE2				mgee2				mgee2v			
	Bias%	EV	AMV	CR	Bias%	EV	AMV	CR	Bias%	EV	AMV	CR
β_{01}	3.119	0.007	0.007	0.942	-0.915	0.008	0.008	0.944	-2.223	0.008	0.008	0.951
β_{02}	3.226	0.007	0.007	0.946	1.562	0.008	0.008	0.940	3.556	0.009	0.008	0.947
β_{x1}	-12.112	0.006	0.006	0.784	1.238	0.008	0.008	0.942	-6.933	0.024	0.014	0.924
β_{x2}	-9.810	0.007	0.008	0.754	1.433	0.009	0.010	0.964	3.393	0.016	0.011	0.941
β_{z1}	-7.032	0.008	0.007	0.922	-0.456	0.009	0.008	0.954	-0.138	0.009	0.008	0.949
β_{z2}	-6.311	0.006	0.005	0.932	0.071	0.006	0.006	0.938	-0.364	0.006	0.006	0.932
β_{z3}	-6.630	0.005	0.005	0.908	0.056	0.006	0.006	0.964	0.143	0.006	0.006	0.962
ϕ	-13.130	0.009	0.010	0.690	0.217	0.014	0.015	0.954	1.257	0.016	0.017	0.956

Table 2: Simulation results with a 5% misclassification rate.

In addition to the preceding simulation with a misclassification rate of 5%, we conducted another simulation with the same parameters except that the misclassification rate is changed to be 20%, and `corstr = "log-linear"`. The results are reported in Table 3, which shows more noticeable differences in implementing the three functions, 'ordGEE2', 'mgee2', and 'mgee2v'.

	ordGEE2				mgee2				mgee2v			
	Bias%	EV	AMV	CR	Bias%	EV	AMV	CR	Bias%	EV	AMV	CR
β_{01}	9.589	0.007	0.007	0.866	0.748	0.015	0.015	0.952	0.872	0.015	0.016	0.966
β_{02}	11.131	0.006	0.007	0.842	-0.210	0.014	0.015	0.958	-0.523	0.015	0.016	0.958
β_{x1}	-48.891	0.005	0.006	0.000	-1.233	0.027	0.029	0.964	-0.874	0.023	0.023	0.940
β_{x2}	-43.506	0.008	0.008	0.000	-0.400	0.026	0.027	0.958	-0.399	0.023	0.023	0.946
β_{z1}	-26.284	0.006	0.006	0.346	0.364	0.011	0.012	0.960	0.084	0.011	0.011	0.940
β_{z2}	-24.870	0.006	0.006	0.832	1.703	0.012	0.011	0.938	1.396	0.010	0.010	0.948
β_{z3}	-26.893	0.006	0.006	0.314	0.228	0.011	0.012	0.954	0.144	0.010	0.011	0.968
ϕ_0	-53.210	0.009	0.009	0.000	1.873	0.078	0.068	0.942	1.034	0.052	0.066	0.976
ϕ_2	-73.139	0.006	0.006	0.042	-0.353	0.052	0.047	0.948	-1.241	0.037	0.049	0.970
ϕ_{22}	-59.955	0.011	0.011	0.000	0.256	0.075	0.075	0.944	0.188	0.053	0.079	0.978

Table 3: Simulation results with a 20% misclassification rate.

A case study

To illustrate the usage of the developed R package, we analyze a dataset arising from the Framingham Heart Study, obtained from the NIH website (https://biolincc.nhlbi.nih.gov/teaching/). Similar to Chen et al. (2014), we consider those 915 male patients who completed both exams #2 and #3, and age between 31 and 65 at the entry of the study. The response variable, HBP, is a categorical variable indicating the status of the systolic blood pressure (SBP), where HBP=0 if SBP is below 140 mmHg, HBP=1 if SBP is between 140 mmHg and 159 mmHg, and HBP=2 if SBP is larger than 160 mmHg.

We are interested in understanding the relationship between HBP and covariates, including the serum cholesterol level (CHOL), age, and the current smoking status (CURSMOKE), as well as the

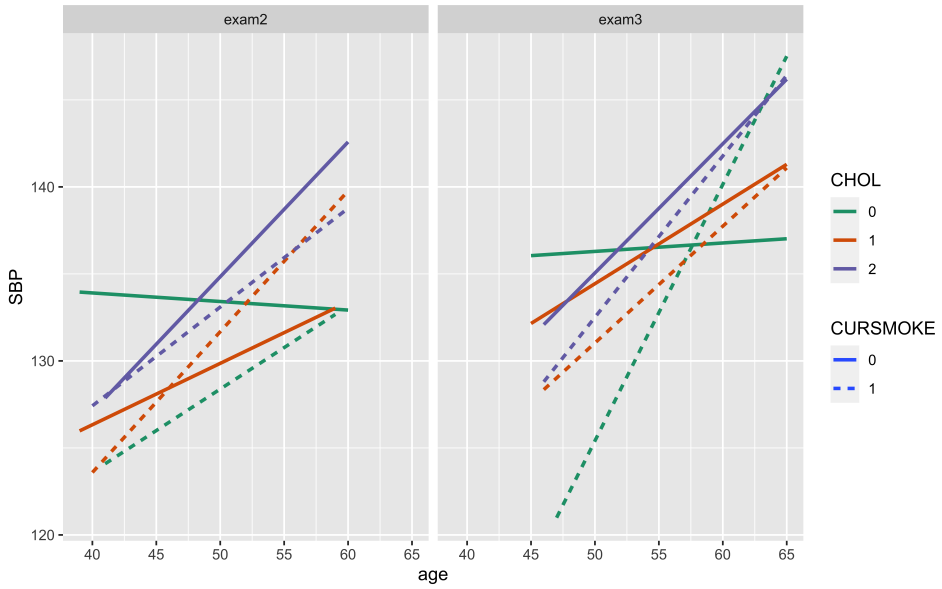

Figure 4: The least squares regression lines by fitting scattered data of SBP against age under different categories stratified by the combination of current smoking status (CURSMOKE) and cholesterol level(CHOL), for patients at different exam times. For example, the dotted red line on the left panel is a linear model fit for SBP against AGE for smokers with level 1 cholesterol at exam 2.

examination status, denoted as "Exam3". CHOL is classified as three categories, with 0, 1, and 2 representing normal (less than 200 mg/dL), borderline high (200-239mg/dL), and hypercholesterolemia (greater than 240 mg/dL), respectively. Exam3 is a dummy variable, with 1 indicating observations for exam 3 and 0 for exam 2.

First, we visualize how SBP may change with age by stratifying the study subjects into different categories according to the exam time, smoking status, or CHOL. To see the trend, we display simple linear regression lines that fit scattered points of SBP against AGE for patients in each category, as shown in Figure 4. Except for the patients with CHOL value 0 and CURSMOKE value 0 at exam 2, there is generally an upward trend of SBP versus age for each category, though the degree varies. While each patient takes 2 exams, the time interval between two exams is different from patient to patient. To reflect this feature of different gap times for the study subjects, in Figure 5 we further display spaghetti plots (Hedeker and Gibbons, 2006) for patients in different categories, where the two endpoints of each black line segment mark SBP and age for the corresponding study subject at exams 2 and 3 in each category, respectively. The blue curve represents the loess smooth curve in each panel to show the trend of SBP against AGE. The loess smooth function is a tool to create smooth lines for scattered plots using polynomial approximations. The code for producing Figures 4 and 5 is included in the help file of data set *heart* in our R package.

Next, we use the proportional odds model to examine how SBP may be quantitatively associated with the covariates. For the ith patient at the jth visit, $X_{ij,CHOL=1}$ and $X_{ij,CHOL=2}$ are binary indicator variables recording whether the patient's cholesterol level is 1 and 2, respectively; $Z_{ij,smoker}$ is a binary variable whether or not the patient is a smoker; $Z_{ij,exam3}$ is a binary variable showing whether or not the patient is taking exam #3; and $Z_{i,age}$ records the age of the ith patient at the entry of the study.

As defined in (1), consider the model

$$\begin{aligned}\text{logit}\,\lambda_{ijk} =& \beta_{0k} + \beta_{X,CHOL=1}X_{ij,CHOL=1} + \beta_{X,CHOL=2}X_{ij,CHOL=2} \\ &+ \beta_{Z,age}Z_{i,age} + \beta_{Z,smoker}Z_{ij,smoker} + \beta_{Z,exam3}Z_{ij,exam3}\end{aligned} \tag{8}$$

for $k = 1, 2$, where $\beta_{0k}, \beta_{X,CHOL=1}, \beta_{X,CHOL=2}, \beta_{Z,age}, \beta_{Z,smoker}$, and $\beta_{Z,exam3}$ are the parameters.

The data set used in our example is included in our package called "`heart`". To demonstrate the usage of the developed package, we perceive that the response HBP level and the covariate cholesterol level are prone to misclassification. Since this example does not have a validation data set, we only analyze the data using the naive method, "`ordGEE2`", and the corrected method with a specified known misclassification rate, "`mgee2k`", where the misclassification rates for both the outcome and the covariate are assumed to be 5%, and the exchangeable dependence structure is considered. The analysis results are shown in Table 4. Overall, the naive method and the corrected method indicate the same significant health factors, yet the magnitude of the coefficient estimates and their standard

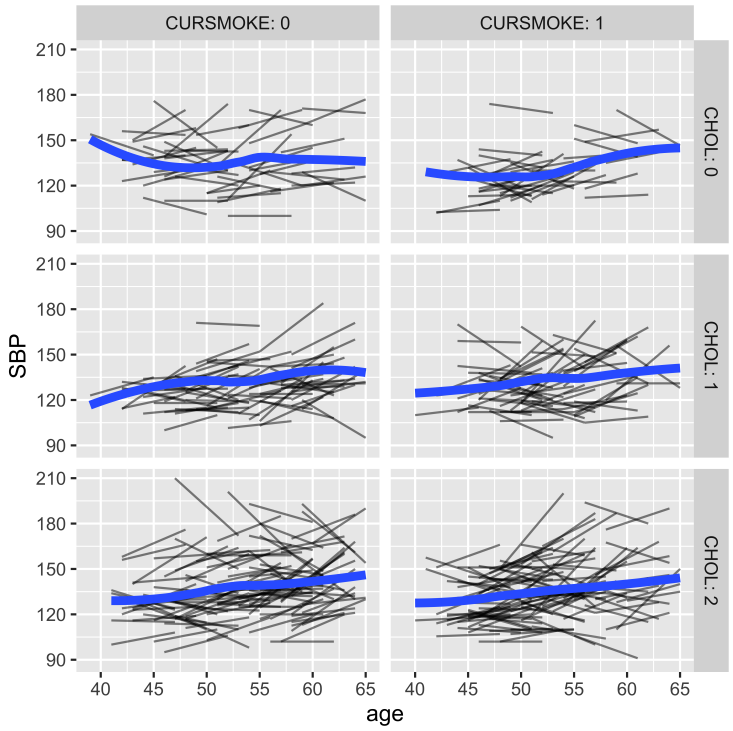

Figure 5: The spaghetti plots of SBP at exams 2 and 3 for the patients classified into different groups by different values of current smoking status and cholesterol level, where the varying lengths of the black line segments reflect the fact that the gap time between exams 2 and 3 differ from patient to patient. The blue curve in each panel is fitted using the loess function.

errors are different. Higher cholesterol levels and older ages appear to be positively correlated with high blood pressure.

Summary

Analysis of longitudinal ordinal data is important for research in health science, epidemiological studies, and social science. Marginal analysis using generalized estimating equations has been extensively employed in applications. However, such a strategy is challenged by the presence of mismeasurement of variables. To address this challenge, Chen et al. (2014) developed estimation methods for analyzing correlated ordinal responses and ordinal covariates, which are subject to misclassification.

To allow analysts to apply the useful methods of Chen et al. (2014) without doing individual codes, we develop an R package **mgee2** to implement the methods for general use. Our package provides three methods for estimation, including the two methods of corrections for misclassification effects, as opposed to the naive method, which disregards the feature of mismeasurement in variables. The package can be used for modeling longitudinal ordinal data with misclassified response and covariates. It provides consistent estimation results by directly inputting the data under required assumptions.

Acknowledgements

The authors thank the review team for the helpful comments on the initial version. Yi's research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC). Yi is Canada Research Chair in Data Science (Tier 1). Her research was undertaken, in part, thanks to funding from the Canada Research Chairs program.

	ordGEE2			mgee2k		
	Est.	SD	p-vlue	Est.	SD	p-vlue
β_{01}	-4.291	0.635	<0.001	-4.943	0.737	<0.001
β_{02}	-5.623	0.638	<0.001	-6.195	0.740	<0.001
$\beta_{x,CHOL=1}$	0.068	0.133	0.608	0.117	0.180	0.515
$\beta_{x,CHOL=2}$	0.352	0.140	0.012	0.474	0.186	0.011
$\beta_{z,age}$	0.063	0.012	<0.001	0.071	0.014	<0.001
$\beta_{z,smoker}$	-0.044	0.105	0.673	-0.042	0.118	0.722
$\beta_{z,exam3}$	0.145	0.097	0.133	0.182	0.110	0.097
ϕ	2.301	0.207	<0.001	2.301	0.318	<0.001

Table 4: A case study of a data subset arising from the Framingham Heart Study, with an assumed misclassification rate at 5%.

Bibliography

D. Bates and M. Maechler. *Matrix: Sparse and Dense Matrix Classes and Methods*, 2019. URL https://CRAN.R-project.org/package=Matrix. R package version 1.2-17. [p435]

L. J. Beesley and B. Mukherjee. Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification. *Biometrics*, 2020. [p432]

V. J. Carey. *gee: Generalized Estimation Equation Solver*, 2015. URL https://CRAN.R-project.org/package=gee. R package version 4.13-19. [p432]

R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models*. Chapman and Hall/CRC, 2006. URL https://doi.org/10.1201/9781420010138. [p432]

Z. Chen, G. Y. Yi, and C. Wu. Marginal methods for correlated binary data with misclassified responses. *Biometrika*, 98(3):647–662, 2011. URL http://www.jstor.org/stable/23076137. [p432]

Z. Chen, G. Y. Yi, and C. Wu. Marginal analysis of longitudinal ordinal data with misclassification in both response and covariates. *Biometrical Journal*, 56(1):69–85, 2014. URL https://doi.org/10.1002/bimj.201200195. [p432, 433, 434, 435, 437, 441, 443]

S. Dlugosz, E. Mammen, and R. A. Wilke. Generalized partially linear regression with misclassified data and an application to labour market transitions. *Computational Statistics & Data Analysis*, 110: 145–159, 2017. [p432]

C. Genolini, X. Alacoque, M. Sentenac, and C. Arnaud. kmlandkml3d: R packages to cluster longitudinal data. *Journal of Statistical Software*, 65(4), 2015. URL https://doi.org/10.18637/jss.v065.i04. [p432]

D. Hedeker and R. D. Gibbons. *Longitudinal Data Analysis*. John Wiley & Sons, 2006. [p442]

K.-Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73 (1):13–22, 1986. URL https://doi.org/10.1093/biomet/73.1.13. [p432]

J. M. Neuhaus. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika*, 86(4):843–855, 1999. URL http://www.jstor.org/stable/2673589. [p432]

M. S. Pepe and G. L. Anderson. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics - Simulation and Computation*, 23(4):939–951, 1994. URL https://doi.org/10.1080/03610919408813210. [p433]

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. [p435]

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. URL https://ggplot2.tidyverse.org. [p435]

J. M. Williamson, K. Kim, and S. R. Lipsitz. Analyzing bivariate ordinal data using a global odds ratio. *Journal of the American Statistical Association*, 90(432):1432–1437, 1995. URL https://doi.org/10.1080/01621459.1995.10476649. [p433]

J. Xiong and G. Y. Yi. swgee: An r package for analyzing longitudinal data with response missingness and covariate measurement error. *The R Journal*, 11(1):416–426, 2019. [p432]

C. Xu, Z. Li, and M. Wang. *wgeesel: Weighted Generalized Estimating Equations and Model Selection*, 2018. URL https://CRAN.R-project.org/package=wgeesel. R package version 1.5. [p432]

G. Y. Yi. A simulation-based marginal method for longitudinal data with dropout and mismeasured covariates. *Biostatistics*, 9(3):501–512, 2008. URL https://doi.org/10.1093/biostatistics/kxm054. [p432]

G. Y. Yi. *Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application*. Springer, New York, 2017. [p432]

G. Y. Yi, Y. Ma, D. Spiegelman, and R. J. Carroll. Functional and structural methods with mixed measurement error and misclassification in covariates. *Journal of the American Statistical Association*, 110(510):681–696, 2015. [p432]

Q. Zhang and G. Y. Yi. R package for analysis of data with mixed measurement error and misclassification in covariates: augsimex. *Journal of Statistical Computation and Simulation*, 89(12):2293–2315, 2019. [p432]

Yuliang Xu
School of Public Health
University of Michigan
500 S. State Street, Ann Arbor, Michigan 48109 USA
yuliangx@umich.edu

Shuo Shuo Liu
Department of Statistics
Pennsylvania State University
201 Old Main, University Park, Pennsylvania 16802 USA
ORCiD: 0000-0001-8396-4515
shuoshuo.liu@psu.edu

Grace Y. Yi
Department of Statistical and Actuarial Sciences
Department of Computer Science
University of Western Ontario
1151 Richmond Street, London, Ontario, Canada N6A 3K7
gyi5@uwo.ca