swgee: An R Package for Analyzing Longitudinal Data with Response Missingness and Covariate Measurement Error

by Juan Xiong and Grace Y. Yi

Abstract Though longitudinal data often contain missing responses and error-prone covariates, relatively little work has been available to simultaneously correct for the effects of response missingness and covariate measurement error on analysis of longitudinal data. Yi (2008) proposed a simulation based marginal method to adjust for the bias induced by measurement error in covariates as well as by missingness in response. The proposed method focuses on modeling the marginal mean and variance structures, and the missing at random mechanism is assumed. Furthermore, the distribution of covariates are left unspecified. These features make the proposed method applicable to a broad settings. In this paper, we develop an R package, called **swgee**, which implements the method proposed by Yi (2008). Moreover, our package includes additional implementation steps which extend the setting considered by Yi (2008). To describe the use of the package and its main features, we report simulation studies and analyses of a data set arising from the Framingham Heart Study.

Introduction

Longitudinal studies are commonly conducted in the health sciences, biochemical, and epidemiology fields; these studies typically collect repeated measurements on the same subject over time. Missing observations and covariate measurement error frequently arise in longitudinal studies and they present considerable challenges in statistical inference about such data (Carroll et al., 2006; Yi, 2008). It has been well documented that ignoring missing responses and covariate measurement error may lead to severely biased results, thus leading to invalid inferences (Fuller, 1987; Carroll et al., 2006).

Regarding longitudinal data with missing responses, there has been extensive methods such as maximum likelihood, multiple imputation, and weighted generalized estimating equations (GEE) method (Little and Rubin, 2002). In terms of methods of handling measurement error in covariate, many methods have been developed for various settings. Comprehensive discussions can be found in Fuller (1987), Gustafson (2003), Carroll et al. (2006), Buonaccorsi (2010) and Yi (2017). However, there has been relatively little work on simultaneously addressing the effects of response missingness and covariate measurement error in longitudinal data analysis, although some work such as Wang et al. (2008), Liu and Wu (2007) and Yi et al. (2012), are available. In particular, Yi (2008) proposed an estimation method based on the marginal model for the response process, which does not require the full specification of the distribution of the response variable but models only the mean and variance structures. Furthermore, a functional method is applied to relax the need of modeling the covariate process. These features make the method of Yi (2008) flexible for many applications.

Relevant to our R package, a set of R packages and statistical software have been available for performing the GEE and weighted GEE analyses for longitudinal data with missing observations. In particular, package **gee** (Carey, 2015) and **yags** (Carey, 2011) perform the GEE analyses under the strong assumption of missing completely at random (MCAR) (Kenward, 1998). Package **wgeesel** (Xu et al., 2018) can perform the multiple model selection based on weighted GEE/GEE. Package **geepack** (Hojsgaard et al., 2016) implements the weighted GEE analyses under the missing at random (MAR) assumption, in which an optional vector of weights can be used in the fitting process but the weight vector has to be externally calculated. In addition, the statistical software SAS/STAT version 13.2 (SAS Institute Inc., 2014) includes an experimental version of the function PROC GEE (Lin and Rodriguez, 2015), which fits weighted GEE models.

Our **swgee** package has several features distinguishing from existing packages. First, **swgee** is designed to analyze longitudinal data with both missing responses and error-prone covariates. To the best of our knowledge, this is the first R package that can simultaneously account for response missingness and covariate measurement error. Secondly, this simulation based marginal method can be applied to a broad range of problems because the associated model assumptions are minimal. **swgee** can be directly applied to handle continuous and binary responses as well as count data with dropouts under the MAR and MCAR mechanisms. Thirdly, observations are weighted inversely proportional to their probability of being observed, with weights calculated internally. Lastly, the **swgee** package employs the simulation extrapolation (SIMEX) algorithm to account for the effect of measurement error in covariates.

The remainder is organized as follows. Section Notation and framework introduces the notation and model setup. In Section Methodology, we describe the method proposed by Yi (2008) and its implementation in R in Section Implementation in R. The developed R package is illustrated with simulation studies and analyses of a data set arising from the Framingham Heart Study in Section Examples. General discussion is included in Section Summary and discussion.

Notation and framework

For $i = 1, \ldots, n$ and $j = 1, \ldots, m$, let Y_{ij} be the response variable for subject i at time point j, let $\mathbf{X}_{ij}$ be the vector of covariates subject to error, and $\mathbf{Z}_{ij}$ be the vector of covariates which are error-free. Write $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{im})'$, $\mathbf{X}_i = (\mathbf{X}'_{i1}, \mathbf{X}'_{i2}, \ldots, \mathbf{X}'_{im})'$, and $\mathbf{Z}_i = (\mathbf{Z}'_{i1}, \mathbf{Z}'_{i2}, \ldots, \mathbf{Z}'_{im})'$.

Response model

For $i = 1, \ldots, n$ and $j = 1, \ldots, m$, let $\mu_{ij} = \mathrm{E}(Y_{ij}|\mathbf{X}_i, \mathbf{Z}_i)$ and $v_{ij} = \mathrm{var}(Y_{ij}|\mathbf{X}_i, \mathbf{Z}_i)$ be the conditional expectation and variance of Y_{ij}, given the covariates $\mathbf{X}_i$ and $\mathbf{Z}_i$, respectively. We model the influence of the covariates on the marginal response mean by means of a regression model:

$$g(\mu_{ij}) = \mathbf{X}'_{ij}\boldsymbol{\beta}_x + \mathbf{Z}'_{ij}\boldsymbol{\beta}_z, \tag{1}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}'_x, \boldsymbol{\beta}'_z)'$ is the vector of regression parameters and $g(\cdot)$ is a specified monotone function. The intercept term, if any, of the model may be included as the first element of $\boldsymbol{\beta}_z$ by including the unit vector as the first column of $\mathbf{Z}_i$.

To model the variance of Y_{ij}, we consider

$$v_{ij} = h(\mu_{ij}; \phi), \tag{2}$$

where $h(\cdot;\cdot)$ is a given function and ϕ is the dispersion parameter that is known or to be estimated. We treat ϕ as known with emphasis setting on estimation of the $\boldsymbol{\beta}$ parameter. Here we assume that $\mathrm{E}(Y^k_{ij}|\mathbf{X}_i, \mathbf{Z}_i) = \mathrm{E}(Y^k_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij})$ for $k = 1$ and 2, that is, the dependence of the mean μ_{ij} and the variance v_{ij} on the subject-level covariates $\mathbf{X}_i$ and $\mathbf{Z}_i$ is completely reflected by the dependence on the time-specific covariates $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$. This assumption has been widely used in marginal analysis of longitudinal analysis (e. g. , Diggle and Kenward, 1994; Lai and Small, 2007). The necessity of these assumptions was discussed by Yi (2017, Section 5.1.1).

Missing data model

For $i = 1, \ldots, n$ and $j = 1, \ldots, m$, let O_{ij} be 1 if Y_{ij} is observed and 0 otherwise, and let $\mathbf{O}_i = (O_{i1}, O_{i2}, \ldots, O_{im})'$ be the vector of missing data indicators. Dropouts or monotone missing data patterns are considered here. That is, $O_{ij} = 0$ implies $O_{ij'} = 0$ for all $j' > j$. We assume that $O_{i1} = 1$ for every subject i. To reflect the dynamic nature of the observation process over time, we assume an MAR mechanism for the missing process. That is, given the covariates, the missingness probability depends on the observed responses but not unobserved response components (Little and Rubin, 2002). Let $\lambda_{ij} = P(O_{ij} = 1|O_{i,j-1} = 1, \mathbf{X}_i, \mathbf{Z}_i, \mathbf{Y}_i)$ and $\pi_{ij} = P(O_{ij} = 1|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{Y}_i)$, then

$$\pi_{ij} = \prod_{t=2}^{j} \lambda_{it}. \tag{3}$$

Logistic regression models are used to model the dropout process:

$$\mathrm{logit}(\lambda_{ij}) = \mathbf{u}'_{ij}\boldsymbol{\alpha}, \tag{4}$$

for $j = 2, \ldots, m$, where $\mathbf{u}_{ij}$ is the vector consisting of the information of the covariates $\mathbf{X}_i$, $\mathbf{Z}_i$ and the observed responses, and $\boldsymbol{\alpha}$ is the vector of regression parameters. Write $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$ and let $q = \dim(\boldsymbol{\theta})$.

Measurement error model

For $i = 1, \ldots, n$ and $j = 1, \ldots, m$, let $\mathbf{W}_{ij}$ be the observed measurements of the covariates $\mathbf{X}_{ij}$. Covariates $\mathbf{X}_{ij}$ and their observed measurements $\mathbf{W}_{ij}$ are assumed to follow a classical additive measurement error model:

$$\mathbf{W}_{ij} = \mathbf{X}_{ij} + \mathbf{e}_{ij}, \tag{5}$$

where the $\mathbf{e}_{ij}$ are independent of $\mathbf{X}_i$, $\mathbf{Z}_i$ and $\mathbf{Y}_i$. And $\mathbf{e}_{ij}$ follows $\mathrm{N}(\mathbf{0}, \mathbf{\Sigma}_e)$ with the covariance matrix $\mathbf{\Sigma}_e$. This model has been widely used in the context of handling measurement error problems. Yi (2008) assumed that $\mathbf{\Sigma}_e$ is known or can be estimated from replication experiments (e. g. , Carroll et al., 2006; Yi, 2017).

Methodology

Weighted estimation function

The inverse probability weighted generalized estimating equations method is often employed to accommodate the missing data effects (e. g. , Robins et al., 1995; Preisser et al., 2002; Qu et al., 2011) when primary interest lies in the estimation of the marginal mean parameters $\boldsymbol{\beta}$ in the model (1). For $i = 1, \ldots, n$, let M_i be the random dropout time for subject i and m_i be a realization. Define $L_i(\boldsymbol{\alpha}) = (1 - \lambda_{im_i}) \prod_{t=2}^{m_i - 1} \lambda_{it}$, where λ_{it} is determined by model (4). Let $\mathbf{S}_i(\boldsymbol{\alpha}) = \partial logL_i(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$ be the vector of score functions contributed from subject i. Let $\mathbf{D}_i = \partial\boldsymbol{\mu}_i'/\partial\boldsymbol{\beta}$ be the matrix of the derivatives of the mean vector $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im})'$ with respect to $\boldsymbol{\beta}$ and let $\mathbf{\Delta}_i = \mathrm{diag}(I(O_{ij} = 1)/\pi_{ij}, j = 1, 2, \ldots, m)$ be the weighted matrix accommodating missingness, where $I(\cdot)$ is the indicator function. Let $\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{C}_i\mathbf{A}_i^{1/2}$ be the conditional covariance matrix of $\mathbf{Y}_i$, given $\mathbf{X}_i$ and $\mathbf{Z}_i$, where $\mathbf{A}_i = \mathrm{diag}(v_{ij}, j = 1, 2, \ldots, m)$ and $\mathbf{C}_i = [\rho_{i;jk}]$ is the correlation matrix with diagonal elements equal 1 and $\rho_{i;jk}$ being the conditional correlation coefficient of response components Y_{ij} and Y_{ik} for $j \neq k$, given $\mathbf{X}_i$ and $\mathbf{Z}_i$. Define

$$\mathbf{U}_i(\boldsymbol{\theta}) = \mathbf{D}_i\mathbf{V}_i^{-1}\mathbf{\Delta}_i(\mathbf{Y}_i - \boldsymbol{\mu}_i)$$

and

$$\mathbf{H}_i(\boldsymbol{\theta}) = (\mathbf{U}_i'(\boldsymbol{\theta}), \mathbf{S}_i'(\boldsymbol{\alpha}))'. \tag{6}$$

In the absence of measurement error, that is, covariates $\mathbf{X}_{ij}$ are precisely observed, we have $\mathrm{E}[\mathbf{H}_i(\boldsymbol{\theta})] = \mathbf{0}$. Hence, $\mathbf{H}(\boldsymbol{\theta}) = \sum_{i=1}^n \mathbf{H}_i(\boldsymbol{\theta})$ are unbiased estimation functions for $\boldsymbol{\theta}$ (e. g. , Yi, 2017, Chapter 1). Under regularity conditions, the consistent estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be obtained by solving

$$\mathbf{H}(\boldsymbol{\theta}) = \mathbf{0}, \tag{7}$$

where the weight matrix $\mathbf{\Delta}_i$ is used to adjust for the contributions of subject i with his/her missingness probabilities incorporated. Specifically, the probability π_{ij} is determined by (3) in conjunction with (4). Correlation matrix $\mathbf{C}_i$ can be replaced by the moment estimate, or alternatively, a working independence matrix $\mathbf{A}_i$ may be used to replace $\mathbf{V}_i$ (Liang and Zeger, 1986). A detail discussion can be found in Yi (2017, Chapter 4).

SIMEX approach

When measurement error is present in covariates $\mathbf{X}_{ij}$, $\mathbf{H}(\boldsymbol{\theta})$ is no longer unbiased if naively replacing $\mathbf{X}_{ij}$ with its observed measurement $\mathbf{W}_{ij}$. Yi (2008) developed a simulation-extrapolation (SIMEX) method to adjust for the bias induced by using $\mathbf{W}_{ij}$, as well as the missingness effects in the response variables. This method originates from the SIMEX method by Cook and Stefanski (1994) who considered cross-sectional data with measurement error alone. The basic idea of the SIMEX method is to first add additional variability to the observed measurement $\mathbf{W}_{ij}$, then establish the trend how different degrees of measurement error may induce bias in estimation of the model parameters, and finally extrapolate this trend to the case of no measurement error.

Now, we describe the SIMEX method developed by Yi (2008). Let B be a given positive integer and $\mathbf{\Lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_M\}$ be a sequence of nonnegative numbers taken from $[0, \lambda_M]$ with $\lambda_1 = 0$.

- Step 1: Simulation For $i = 1, \ldots, n$ and $j = 1, \ldots, m$, generate $\mathbf{e}_{ijb} \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}_e)$ for $b = 1, 2, \ldots, B$. For a given $\lambda \in \mathbf{\Lambda}$, set
$$\mathbf{W}_{ij}(b, \lambda) = \mathbf{W}_{ij} + \sqrt{\lambda}\mathbf{e}_{ijb}.$$
- Step 2: Estimation For given λ and b, we obtain an estimate $\widehat{\boldsymbol{\theta}}(b, \lambda)$ by solving equation (7) with $\mathbf{X}_{ij}$ replaced by $\mathbf{W}_{ij}(b, \lambda)$. Let $\widehat{\mathbf{\Gamma}}(b, \lambda) = \sum_{i=1}^n [\partial\mathbf{H}_i'(\boldsymbol{\theta}; b, \lambda)/\partial\boldsymbol{\theta}]|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}(b,\lambda)}$ and $\widehat{\mathbf{\Sigma}}(b, \lambda) = \sum_{i=1}^n [\mathbf{H}_i(\boldsymbol{\theta}; b, \lambda)\mathbf{H}_i'(\boldsymbol{\theta}; b, \lambda)]|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}(b,\lambda)}$, then the covariance matrix of $\widehat{\boldsymbol{\theta}}(b, \lambda)$ is estimated by:
$$\widehat{\mathbf{\Omega}}(b, \lambda) = n \cdot \left\{ [\widehat{\mathbf{\Gamma}}(b, \lambda)]^{-1} \cdot \widehat{\mathbf{\Sigma}}(b, \lambda) \cdot [\widehat{\mathbf{\Gamma}}(b, \lambda)]^{-1'} \right\}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}(b,\lambda)}.$$

Let $\widehat{\theta}_r(b,\lambda)$ be the rth component of $\widehat{\boldsymbol{\theta}}(b,\lambda)$ and let $\widehat{\Omega}_r(b,\lambda)$ be the rth diagonal element of $\widehat{\boldsymbol{\Omega}}(b,\lambda)$ for $r = 1,2,\ldots,q$. We then calculate the average of those estimates over b for each λ:

$$\widehat{\theta}_r(\lambda) = B^{-1}\sum_{b=1}^{B}\widehat{\theta}_r(b,\lambda);$$

$$\widehat{\Omega}_r(\lambda) = B^{-1}\sum_{b=1}^{B}\widehat{\Omega}_r(b,\lambda);$$

$$\widehat{S}_r(\lambda) = (B-1)^{-1}\sum_{b=1}^{B}(\widehat{\theta}_r(b,\lambda) - \widehat{\theta}_r(\lambda))^2;$$

$$\widehat{\tau}_r(\lambda) = \widehat{\Omega}_r(\lambda) - \widehat{S}_r(\lambda).$$

- Step 3: Extrapolation For $r = 1,2,\ldots,q$, fit a regression model to each of the sequences $\{(\lambda,\widehat{\theta}_r(\lambda)) : \lambda \in \boldsymbol{\Lambda}\}$ and $\{(\lambda,\widehat{\tau}_r(\lambda)) : \lambda \in \boldsymbol{\Lambda}\}$, respectively, and extrapolate it to $\lambda = -1$, let $\widehat{\theta}_r(-1)$ and $\widehat{\tau}_r(-1)$ denote the resulting predicted values. Then, $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1,\widehat{\theta}_2,\ldots,\widehat{\theta}_q)'$ is the SIMEX estimator of $\boldsymbol{\theta}$ and $\sqrt{\widehat{\tau}_r}$ is the associated standard error for the estimator $\widehat{\theta}_r$ for $r = 1,2,\ldots,q$.

The SIMEX approach is very appealing because of its simplicity of implementation and no requirement of modeling the true covariates $\mathbf{X}_i$. However, to use this method, several aspects need to be considered. As suggested by Carroll et al. (2006), the specification of Λ is not unique; a typical choice of grid Λ is the equal cut points of interval $[0,2]$ with $M = 5$ or 9. Choosing $B = 100$ or 200 is often sufficient for many applications. The quadratic regression function is commonly used for Step 3 to yield reasonable results. (e. g. , He et al., 2012).

Finally, we extend the method by Yi (2008) to accommodating the case where the covariance matrix $\boldsymbol{\Sigma}_e$ for model (5) is unknown but repeated surrogate measurements of $\mathbf{X}_{ij}$ are available. Let $\mathbf{W}_{ijk}$ denote the repeated surrogate measurements of $\mathbf{X}_{ij}$ for $i = 1,\ldots,n$; $j = 1,\ldots,m$; and $k = 1,\ldots,K$. The surrogate measurements $\mathbf{W}_{ijk}$ and the true covariate $\mathbf{X}_{ij}$ are linked by the model

$$\mathbf{W}_{ijk} = \mathbf{X}_{ij} + \mathbf{e}_{ijk}, \tag{8}$$

where the $\mathbf{e}_{ijk}$ are independent of $\mathbf{X}_i$, $\mathbf{Z}_i$ and $\mathbf{Y}_i$, and $\mathbf{e}_{ijk}$ follows $\mathrm{N}(\mathbf{0},\boldsymbol{\Sigma}_e)$ with the covariance matrix $\boldsymbol{\Sigma}_e$. We now adapt the arguments of Devanarayan and Stefanski (2002) to modify the simulation step of the preceding SIMEX method. For a given b and $\lambda \in \boldsymbol{\Lambda}$, set

$$\mathbf{W}_{ij}(b,\lambda) = \overline{\mathbf{W}}_{ij} + \sqrt{\lambda/K}\sum_{k=1}^{K}c_{ijk}(b)\mathbf{W}_{ijk}, \tag{9}$$

where $\overline{\mathbf{W}}_{ij} = K^{-1}\sum_{k=1}^{K}\mathbf{W}_{ijk}$ and $\mathbf{c}_{ij}(b) = (c_{ij1}(b),\ldots,c_{ijk}(b))'$ is a normalized contrast satisfying $\sum_{k=1}^{K}c_{ijk} = 0$ and $\sum_{k=1}^{K}c_{ijk}^2 = 1$.

A simple way to generate a contrast $\mathbf{c}_{ij}(b)$ can be done by independently generating K variates, $d_{ijk}(b)$, from $\mathrm{N}(0,1)$ for $k = 1,\ldots,K$ and a given b. Let $\overline{d}_{ij}(b) = K^{-1}\sum_{k=1}^{K}d_{ijk}(b)$. Then $c_{ijk}(b)$ is set as

$$c_{ijk}(b,\lambda) = \frac{d_{ijk}(b) - \overline{d}_{ij}(b)}{\sqrt{\sum_{k=1}^{K}\{d_{ijk}(b) - \overline{d}_{ij}(b)\}^2}}.$$

Once $\mathbf{W}_{ij}(b,\lambda)$ of (9) is available, we repeat Steps 2 and 3 to obtain the SIMEX estimator and the associated standard error.

Implementation in R

We implement the SIMEX procedure described in Section Methodology in R and develop the package, called **swgee**. Our package **swgee** takes the advantage of existing R packages **geepack** (Hojsgaard et al., 2016) and **mvtnorm** (Genz and Bretz, 2009; Genz et al., 2018). Specifically, the function swgee produces the estimates for elements of the parameter vector $\boldsymbol{\beta}$, which are of primary interest, the associated standard errors, and P-values.

Our R function swgee requires the input data set to be sorted by subject i and visit time j for $i = 1,\ldots,n$ and $j = 1,\ldots,m$. If a subject is missing at a certain time, the corresponding measurements should be recorded as NAs. As long as the user provides the missing data model (4), the function swgee can internally generate the missing data indicators O_{ij} for $i = 1,\ldots,n$ and $j = 1,\ldots,m$, and then apply the user specified model (4) to fit the data. The missingness probabilities π_{ij} are calculated by (3) and then used to construct the weight matrix $\mathbf{\Delta}_i$ for the estimating equation (6). The estimate of the missing data model (4) parameter $\boldsymbol{\alpha}$ can also be retrieved from the function swgee output.

The form of calling function swgee is given by

```
swgee(formula, data, id, family, corstr, missingmodel, SIMEXvariable,
    SIMEX.err, repeated = FALSE, repind = NULL, B, lambda)
```

where the arguments are described as follows:

- `formula`: This argument specifies the model to be fitted, with the variables coming with data. See the documentation of geeglm and its `formula` for details.
- `data`: This is the same as the data argument in the R function geeglm, which specifies the data frame showing how variables occur in the formula, along with the id variable.
- `id`: This is the vector which identifies the labels of subjects. i.e., the id for subject i is i, using the notation of Section Response model, where $i = 1, 2, \ldots, n$. Data are arranged so that observations for the same subject are listed in consecutive rows in order of time, and consequently, the id for a subject would repeat the same number of times as the observation times.
- `family`: This argument describes the error distribution together with the link function for model (1). See the documentation of geeglm and its argument `family` for details.
- `corstr`: This is a character string specifying the correlation structure. See the documentation of geeglm and its argument `corstr` for details.
- `missingmodel`: This argument specifies the formula to be fitted for the missing data model (4). See the documentation of geeglm and its `formula` for details.
- `SIMEXvariable`: This is the vector of characters containing the names of the covariates which are subject to measurement error.
- `SIMEX.err`: This argument specifies the covariance matrix of measurement errors in the measurement error model (5).
- `repeated`: This is the indicator whether measurement error model is given by (5) or by (8). The default value FALSE corresponding to model (5).
- `repind`: This is the index of the repeated surrogate measurements $\mathbf{W}_{ijk}$ for each covariate $\mathbf{X}_{ij}$. It has an R list form. If repeated = TRUE, repind must be specified.
- `B`: This argument sets the number of simulated samples for the simulation step. The default is set to be 50.
- `lambda`: This is the vector $\{\lambda_1, \lambda_2, \ldots, \lambda_M\}$ we describe in Step 1 of Section SIMEX approach. Its values need to be specified by the user.

Examples

An example data set

To illustrate the usage of the developed R package **swgee**, we apply the package to a subset of GWA13 (Genetic Analysis Workshops) data arising from the Framingham Heart Study. The data set consists of measurements of 100 patients from a series of exams with 5 assessments for each individual. Measurements such as height, weight, age, systolic blood pressure (SBP) and cholesterol level (CHOL) are collected at each assessment, and 14% patients dropped out of the study. The original data were analyzed by Yi (2008). It is of interest to study how an individual's obesity may change with age (Z_{ij}) and how it is associated with SBP (X_{ij1}) and CHOL (X_{ij2}), where $i = 1, \ldots, 100$, and $j = 1, \ldots, 5$. The response Y_i is the indicator of obesity status of subject i as in Yi (2008); SBP is rescaled as $\log(\text{SBP} - 50)$ as in Carroll et al. (2006); and CHOL is standardized. The response and the covariates are postulated by the logistic regression model:

$$\text{logit}\,\mu_{ij} = \beta_0 + \beta_{x1} X_{ij1} + \beta_{x2} X_{ij2} + \beta_z Z_{ij},$$

where β_0, β_{x1}, β_{x2} and β_z are regression coefficients of interest. We assume that errors in both risk factors X_{ij1} and X_{ij2} can be represented by model (5). The missing data process is characterized by the logistic regression model:

$$\text{logit}\lambda_{ij} = \alpha_1 + \alpha_2 Y_{i,j-1} + \alpha_3 X_{i,j-1,1} + \alpha_4 X_{i,j-1,2} + \alpha_{5c} z_{i,j-1},$$

for $j = 2, \ldots, 5$.

We now apply the developed R package **swgee**, which can be downloaded from CRAN and then loaded in R:

```
R> library("swgee")
```

Next, load the data that are properly organized with the variable names specified. In the example here, the data set, named as bmidata, is included by issuing

```
R> data("BMI")
R> bmidata <- BMI
```

We are concerned how measurement error in SBP and CHOL impacts estimation of parameter $\boldsymbol{\beta} = (\beta_0, \beta_{x1}, \beta_{x2}, \beta_z)'$. For illustrative purposes, we use setting with $B = 100$, $\lambda_M = 2$ and $M = 5$. In this example, we assume that parameters in $\boldsymbol{\Sigma}_e = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$ with $\sigma_{12} = \sigma_{21}$ are known. This is a typical case when conducting sensitivity analysis. Here we set $\sigma_1 = \sigma_2 = 0.5$ and $\sigma_{12} = \sigma_{21} = 0$ as an example.

The naive GEE approach without considering missingness and measurement error effects in covariates gives the output:

```
R> output1 <- gee(bbmi~sbp+chol+age, id=id, data=bmidata,
+       family=binomial(link="logit"), corstr="independence")

R> summary(output1)

 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Logit
 Variance to Mean Relation: Binomial
 Correlation Structure:     Independent

Call:
gee(formula = bbmi ~ sbp + chol + age, id = id, data = bmidata,
    family = binomial(link = "logit"), corstr = "independence")

Summary of Residuals:
        Min          1Q      Median          3Q         Max
-0.26533967 -0.11385369 -0.08572483 -0.06279540  0.95475735

Coefficients:
               Estimate Naive S.E.    Naive z Robust S.E.   Robust z
(Intercept) -5.43746374 1.42090827 -3.8267521  1.64320527 -3.3090593
sbp          0.59071183 0.30643396  1.9276970  0.24338420  2.4270755
chol         0.11109496 0.13654324  0.8136247  0.23086218  0.4812177
age          0.01297337 0.01339946  0.9682008  0.01814546  0.7149652

Estimated Scale Parameter:  1.017131
Number of Iterations:  1
Working Correlation
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    0    0    0    0
[2,]    0    1    0    0    0
[3,]    0    0    1    0    0
[4,]    0    0    0    1    0
[5,]    0    0    0    0    1
```

To adjust for possible effects of missingness as well as measurement error in variables SBP and CHOL, we call the developed function swgee for the analysis:

```
R> set.seed(1000)
R> sigma <- diag(rep(0.25, 2))
R> output2 <- swgee(bbmi~sbp+chol+age, data=bmidata, id=id,
+      family=binomial(link="logit"), corstr="independence",
+      missingmodel=O~bbmi+sbp+chol+age, SIMEXvariable=c("sbp","chol"),
+      SIMEX.err=sigma, repeated=FALSE, B=100, lambda=seq(0, 2, 0.5))
```

```
> summary(output2)
Call: beta
              Estimate    StdErr t.value   p.value
(Intercept) -8.004577  2.060967 -3.8839 0.0001028 ***
sbp          1.196363  0.356868  3.3524 0.0008011 ***
chol         0.099984  0.264180  0.3785 0.7050810
age          0.012718  0.017201  0.7394 0.4596520
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Call: alpha
        Estimate    StdErr t.value  p.value
alpha1  9.019084  3.086533  2.9221 0.003477 **
alpha2 -0.786135  0.656843 -1.1968 0.231370
alpha3 -0.568740  0.732885 -0.7760 0.437732
alpha4 -0.128941  0.247757 -0.5204 0.602761
alpha5 -0.064257  0.025982 -2.4731 0.013395 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1
```

The function swgee can store individual estimated coefficients in the simulation step, and this enables us to show the extrapolation curve through the developed R function `plot.swgee`. The `plot.swgee` function plots the extrapolation of the estimate of each covariate effect with the quadratic extrapolants. Figure 1 displays the graph for the variable SBP in the example for which the quadratic extrapolation function is applied from the following command:

```
R> plot(output2,"sbp")
```

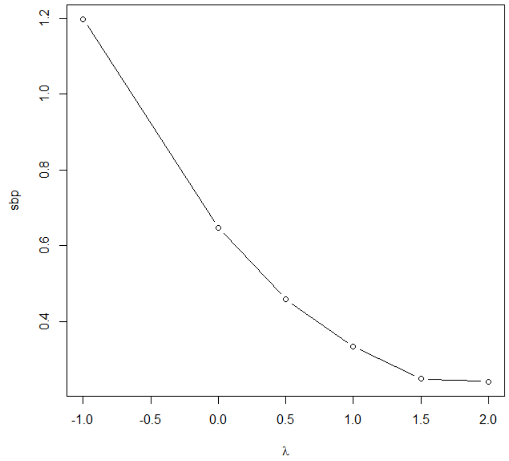

Figure 1: Display of the SIMEX estimate for the example: the dot is the SIMEX estimate obtained from the quadratic extrapolation.

Simulation studies

In this section, we conduct simulation studies to investigate the impact of ignoring covariate measurement error and response missingness on estimation, where the implementation is carried out using the usual GEE method. Furthermore, we assess the performance of the swgee method which accommodates the effects induces from error-prone covariates and missing responses. We set $n = 200$ and $m = 3$, and generate 500 simulations for each parameter configuration. Consider the logistic regression model

$$\text{logit}(\mu_{ij}) = \beta_0 + \beta_{x1}x_{ij1} + \beta_{x2}x_{ij2} + \beta_z z_{ij}, \tag{10}$$

where $\beta_0 = 0$, $\beta_{x1} = \log(1.5)$, $\beta_{x2} = \log(1.5)$, $\beta_z = \log(0.75)$ and z_{ij} is generated independently from $\text{Bin}(1, 0.5)$ to represent a balanced design. The true covariate $\mathbf{X}_{ij} = (x_{ij1}, x_{ij2})'$ is generated

σ_1	σ_2	Method	β_{x1}			β_{x2}			β_z		
			Bias	SE	CR	Bias	SE	CR	Bias	SE	CR
0.25	0.25	gee	-0.0310	0.1228	92.6	-0.0158	0.1246	92.6	0.0063	0.2121	94.6
0.25	0.25	swgee	-0.0062	0.1420	95.0	0.0104	0.1425	95.2	0.0036	0.2354	95.6
0.25	0.50	gee	-0.0019	0.1212	95.4	-0.0997	0.1156	83.4	0.0082	0.2110	94.2
0.25	0.50	swgee	-0.0003	0.1415	95.0	-0.0087	0.1543	93.0	0.0035	0.2361	95.6
0.25	0.75	gee	0.0328	0.1189	95.4	-0.1841	0.1022	51.0	0.0101	0.2100	94.0
0.25	0.75	swgee	0.0205	0.1407	95.8	-0.0660	0.1562	86.4	0.0046	0.2359	95.6
0.50	0.25	gee	-0.1156	0.1114	78.2	0.0139	0.1236	94.2	0.0078	0.2113	94.6
0.50	0.25	swgee	-0.0282	0.1520	93.2	0.0177	0.1431	95.4	0.0031	0.2362	95.2
0.50	0.50	gee	-0.0948	0.1114	81.8	-0.0780	0.1161	85.6	0.0102	0.2099	94.2
0.50	0.50	swgee	-0.0228	0.1510	93.8	-0.0022	0.1542	93.6	0.0030	0.2370	95.4
0.50	0.75	gee	-0.0629	0.1103	87.8	-0.1727	0.1036	55.6	0.0125	0.2088	94.2
0.50	0.75	swgee	-0.0052	0.1499	94.8	-0.0608	0.1570	87.2	0.0042	0.2369	95.2
0.75	0.25	gee	-0.1991	0.0966	45.6	0.0484	0.1216	94.2	0.0092	0.2107	94.6
0.75	0.25	swgee	-0.0870	0.1508	86.4	0.0395	0.1430	93.6	0.0034	0.2366	95.2
0.75	0.50	gee	-0.1889	0.0976	50.0	-0.0458	0.1154	89.8	0.0121	0.2091	94.0
0.75	0.50	swgee	-0.0831	0.1509	87.8	0.0165	0.1539	94.0	0.0034	0.2375	95.4
0.75	0.75	gee	-0.1636	0.0974	58.8	-0.1468	0.1039	66.4	0.0147	0.2077	94.2
0.75	0.75	swgee	-0.0678	0.1505	90.0	-0.0442	0.1574	88.8	0.0046	0.2374	95.2

Table 1: Simulation Results

from the normal distribution $\mathrm{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$, where $\boldsymbol{\mu}_x = (0.5, 0.5)'$ and $\boldsymbol{\Sigma}_x = \begin{pmatrix} \sigma_{x1}^2 & \rho_x\sigma_{x1}\sigma_{x2} \\ \rho_x\sigma_{x1}\sigma_{x2} & \sigma_{x2}^2 \end{pmatrix}$ with $\sigma_{x1} = \sigma_{x2} = 1$. The surrogate value $\mathbf{W}_{ij} = (W_{ij1}, W_{ij2})'$ is generated from $\mathrm{N}(\mathbf{X}_{ij}, \boldsymbol{\Sigma}_e)$ with $\boldsymbol{\Sigma}_e = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$. ρ and ρ_x are set to 0.50 to represent moderate correlations. To feature minor, moderate and severe degrees of measurement error, we consider $\sigma_1, \sigma_2 = 0.25, 0.50$ or 0.75. The missing data indicator is generated from model (4), where $\alpha_0 = \alpha_1 = 0.5$, $\alpha_2 = \alpha_3 = 0.1$, and $\alpha_z = 0.2$. In implementing the swgee method, we choose $B = 100$, $\lambda_M = 2$, $M = 5$, and a quadratic regression for each extrapolation step.

In Table 1, we report on the results of the biases of the estimates (Bias), the empirical standard error (SE), and the coverage rate (CR in percent) for 95% confidence intervals. When measurement error is minor, (i.e. $\sigma_1 = \sigma_2 = 0.25$), both gee and swgee provide reasonable results with fairly small finite sample biases and coverage rates close to the nominal level 95%. When there is moderate or substantial measurement error in covariates $\mathbf{X}_{ij}$, the performance of the gee method deteriorates remarkably in estimation of error-prone covariate effects, leading to considerably biased estimates for β_{x1} and β_{x2}. The corresponding coverage rates for 95% confidence intervals can be quite low. In contrast, the swgee method remarkably improve the performance, providing a lot smaller biases and much higher coverage rates. The estimates for β_z are not subject to much impact of measurement error, which is partially attributed by that the precisely observed covariates z_{ij} are generated independently of error-prone covairates $\mathbf{X}_{ij}$ under the current simulation study.

In summary, ignoring measurement error may lead to substantially biased results. Properly addressing covariate measurement error in estimation procedures is necessary. The proposed swgee method performs reasonably well under various configurations. As expected, its performance may become less satisfactory when measurement error becomes substantial. However, the swgee method does significantly improve the performance of the gee analysis.

Summary and discussion

Missing observations and covariate measurement error commonly arise in longitudinal data. However, there has been relatively little work on simultaneously accounting for the effects of response missingness and covariate measurement error on estimation of response model parameters for longitudinal data. Yi (2008) described a simulation based marginal method to adjust for the biases induced by both missingness and covariate measurement error. The proposed method does not require the full specification of the distribution of the response vector but only requires modeling its mean and covariance structure. In addition, the distribution of covariates is left unspecified, which is desirable for many practical problems. These features make the proposed method flexible.

Here we not only develop the R package **swgee** to implement the method by Yi (2008), but also include an extended setting in the package. Our aim is to provide analysts an accessible tool for the analysis of longitudinal data with missing responses and error-prone covariates. Our illustrations show that the developed package has the advantages of simplicity and versatility.

Acknowledgments

Juan Xiong was supported by the Natural Science Foundation of SZU (grant no.2017094). Grace Y. Yi was supported by the Natural Sciences and Engineering Research Council of Canada. The authors thanks Boston University and the National Heart, Lung, and Blood Institute (NHLBI) for providing the data set from the Framingham Heart Study (No. N01-HC-25195) in the illustration. The Framingham Heart Study is conducted and supported by the NHLBI in collaboration with Boston University. This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI.

Conflict of Interest: None declared.

Bibliography

J. P. Buonaccorsi. *Measurement Error: Models, Methods, and Applications*. Chapman & Hall/CRC, Boca Raton, Florida, 2010. [p1]

V. J. Carey. **yags***: Yet Another GEE Solve*, 2011. R package version 6.1-13. [p1]

V. J. Carey. **gee***: Generalized Estimation Equation Solver*, 2015. R package version 4.13-19. [p1]

R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall/CRC, Boca Raton, Florida, 2nd edition, 2006. [p1, 3, 4, 5]

J. R. Cook and L. A. Stefanski. Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328, 1994. URL `https://doi.org/10.1080/01621459.1994.10476871`. [p3]

V. Devanarayan and L. A. Stefanski. Empirical simulation extrapolation for measurement error models with replicate measurements. *Statistics and Probability Letters*, 59(3):219–225, 2002. URL `https://doi.org/10.1016/S0167-7152(02)00098-6`. [p4]

P. J. Diggle and M. G. Kenward. Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistics*, 43(1):49–93, 1994. URL `https://doi.org/10.2307/2986113`. [p2]

W. A. Fuller. *Measurement Error Models*. John Wiley & Sons, New York, 1987. [p1]

A. Genz and F. Bretz. *Computation of Multivariate Normal and t Probabilities*. Springer-Verlag, New York, 2009. [p4]

A. Genz, F. Bretz, T. Miwa, X. Mi, and T. Hothorn. **mvtnorm***: Multivariate Normal and t Distributions*, 2018. R package version 1.0-7. [p4]

P. Gustafson. *Measurement Error and Misclassification in Statistics and Epidemiology*. Chapman & Hall/CRC, Boca Raton, Florida, 2003. [p1]

W. He, J. Xiong, and G. Y. Yi. Simex R package for accelerated failure time models with covariate measurement error. *Journal of Statistical Software*, 46(1):1–14, 2012. URL `https://doi.org/10.18637/jss.v046.c01`. [p4]

S. Hojsgaard, U. Halekoh, and J. Yan. **geepack***: Generalized Estimating Equation Package*, 2016. R package version 1.2-1. [p1, 4]

M. G. Kenward. Selection models for repeated measurements with non-random dropout: An illustration of sensitivity. *Statistics in Medicine*, 17(23):2723–2732, 1998. URL `https://doi.org/10.1002/(SICI)1097-0258(19981215)17:23<2723::AID-SIM38>3.0.CO;2-5`. [p1]

T. L. Lai and D. S. Small. Marginal regression analysis of longitudinal data with time-dependent covariates: A generalized method-of-moments approach. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 69(1):79–99, 2007. URL `https://doi.org/10.1111/j.1467-9868.2007.00578.x`. [p2]

K. Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73 (1):13–22, 1986. URL `https://doi.org/10.2307/2336267`. [p3]

G. Lin and R. N. Rodriguez. Weighted methods for analyzing missing data with the gee procedure. *Paper SAS166-2015*, pages 1–8, 2015. [p1]

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, New Jersey, 2nd edition, 2002. [p1, 2]

W. Liu and L. Wu. Simultaneous inference for semiparametric nonlinear mixed-effects models with covariate measurement errors and missing responses. *Biometrics*, 63(2):342–350, 2007. URL https://doi.org/10.1111/j.1541-0420.2006.00687.x. [p1]

J. S. Preisser, K. K. Lohman, and P. J. Rathouz. Performance of weighted estimating equations for longitudinal binary data with drop-outs missing at random. *Statistics in Medicine*, 21(20):3035–3054, 2002. URL https://doi.org/10.1002/sim.1241. [p3]

A. Qu, G. Y. Yi, P. X. K. Song, and P. Wang. Assessing the validity of weighted generalized estimating equations. *Biometrika*, 98(1):215–224, 2011. URL https://doi.org/10.1093/biomet/asq078. [p3]

J. M. Robins, A. Rotnitzky, and L. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429): 106–121, 1995. URL https://doi.org/10.1080/01621459.1995.10476493. [p3]

SAS Institute Inc. *SAS/STAT Software, Version 13.2*. Cary, NC, 2014. URL http://www.sas.com/. [p1]

C. Y. Wang, Y. Huang, E. C. Chao, and M. K. Jeffcoat. Expected estimating equations for missing data, measurement error, and misclassification, with application to longitudinal nonignorable missing data. *Biometrics*, 64(1):85–95, 2008. URL https://doi.org/10.1111/j.1541-0420.2007.00839.x. [p1]

C. Xu, Z. Li, and M. Wang. **wgeesel***: Weighted Generalized Estimating Equations and Model Selection*, 2018. R package version 1.5. [p1]

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. [p1, 2, 3, 4, 5, 8]

G. Y. Yi. *Statistical Analysis with Measurement Error or Misclassification*. Springer-Verlag, New York, 2017. [p1, 2, 3]

G. Y. Yi, Y. Ma, and R. J. Carroll. A functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error. *Biometrika*, 99(1):151–165, 2012. URL https://doi.org/10.1093/biomet/asr076. [p1]

Juan Xiong
Department of Preventive Medicine
School of Medicine
Shenzhen University
3688 Nanhai Avenue, Shenzhen, China 518060
jxiong@szu.edu.cn

Grace Y. Yi
Department of Statistics and Actuarial Science
University of Waterloo
200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1
yyi@uwaterloo.ca