Analysis of Correlated Data with Error-Prone Response Under Generalized Linear Mixed Models

Grace Y. Yi, Zhijian Chen, and Changbao Wu

Abstract Measurements of variables are often subject to error due to various reasons. Measurement error in covariates has been discussed extensively in the literature, while error in response has received much less attention. In this paper, we consider generalized linear mixed models for clustered data where measurement error is present in response variables. We investigate asymptotic bias induced by nonlinear error in response variables if such error is ignored, and evaluate the performance of an intuitively appealing approach for correction of response error effects. We develop likelihood methods to correct for effects induced from response error. Simulation studies are conducted to evaluate the performance of the proposed methods, and a real data set is analyzed with the proposed methods.

1 Introduction

Generalized linear mixed models (GLMMs) have been broadly used to analyze correlated data, such as clustered/familial data, longitudinal data, and multivariate data. GLMMs provide flexible tools to accommodate normally or non-normally distributed data through various link functions between the response mean and a set of predictors. For longitudinal studies, in which repeated measurements of a response variable are collected on the same subject over time, GLMMs can be used as a convenient analytic tool to account for subject-specific variations [e.g., 5].

Standard statistical analysis with GLMMs is typically developed under the assumption that all variables are precisely observed. However, this assumption is commonly violated in applications. There has been much interest in statistical inference pertaining to error-in-covariates, and a large body of methods have been developed [e.g., 3, 17, 18]. Measurement error in response, however, has received

G.Y. Yi (✉) • C. Wu
Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada N2L 3G1
e-mail: yyi@uwaterloo.ca; cbwu@uwaterloo.ca

Z. Chen
Bank of Nova Scotia, 4 King Street West, Toronto, ON, Canada M5H 1A1
e-mail: zhijian.chen@scotiabank.com

S.E. Ahmed (ed.), *Big and Complex Data Analysis*, Contributions to Statistics,
DOI 10.1007/978-3-319-41573-4_5

much less attention, and this is partially attributed to a misbelief that ignoring response error would still lead to valid inferences. Unfortunately, this is only true in some special cases, e.g., the response variable follows a linear regression model and is subject to additive measurement error. With nonlinear response models or nonlinear error models, inference results can be seriously biased if response error is ignored. Buonaccorsi [1] conducted numerical studies to illustrate induced biases under linear models with nonlinear response measurement error. With binary responses subject to error, several authors, such as Neuhaus [10] and Chen et al. [4], demonstrated that naive analysis ignoring measurement error may lead to incorrect inference results.

Although there is some research on this topic, systematic studies on general clustered/longitudinal data with response error do not seem available. It is the goal of this paper to investigate the asymptotic bias induced by the error in response and to develop valid inference procedures to account for such biases. We formulate the problem under flexible frameworks where GLMMs are used to feature various response processes and nonlinear models are adopted to characterize response measurement error.

Our research is partly motivated by the Framingham Heart Study, a large scale longitudinal study concerning the development of cardiovascular disease. It is well known that certain variables, such as blood pressure, are difficult to measure accurately due to the biological variability and that their values are greatly affected by the change of environment. There has been a large body of work on the analysis of data from the Framingham Heart Study, accounting for measurement error in covariates. For example, Carroll et al. [2] considered binary regression models to relate the probability of developing heart disease to risk factors including error-contaminated systolic blood pressure. Within the framework of longitudinal analysis, the impact of covariate measurement error and missing data on model parameters has been examined. Yi [16] and Yi et al. [19] proposed estimation and inference methods that account for measurement error and missing response observations. Other work can be found in [7, 20], among others. Relative to the extensive analysis of data with covariate error, there is not much work on accounting for measurement error in continuous responses using the data from the Framingham Heart Study.

The remainder of the paper is organized as follows. In Sect. 2, we formulate the response and the measurement error processes. In Sect. 3, we investigate the estimation bias in two analyses: the naive analysis that completely ignores response measurement error, and a partial-adjustment method that fits model to transformed surrogate responses. In Sect. 4, we develop likelihood-based methods to cover two useful situations: measurement error parameters are known, or measurement error parameters are unknown. In Sect. 5, we evaluate the performances of various approaches through simulation studies. In Sect. 6, we illustrate the proposed method using a real data set from the Framingham Heart Study. Discussion and concluding remarks are given in Sect. 7.

2 Model Formulation

2.1 Response Model

Suppose data from a total of n independent clusters are collected. Let Y_{ij} denote the response for the jth subject in cluster i, $i = 1, \ldots, n$, $j = 1, \ldots, m_i$. Let $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ be vectors of covariates for subject j and cluster i, respectively, and write $\mathbf{X}_i = (\mathbf{X}_{i1}^{\mathrm{T}}, \ldots, \mathbf{X}_{im_i}^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{Z}_i = (\mathbf{Z}_{i1}^{\mathrm{T}}, \ldots, \mathbf{Z}_{im_i}^{\mathrm{T}})^{\mathrm{T}}$. Here we use upper case letters and the corresponding lower case letters to denote random variables and their realizations, respectively.

Conditional on random effects $\mathbf{b}_i$ and covariates $\{\mathbf{X}_i, \mathbf{Z}_i\}$, the $Y_{ij} (j = 1, \ldots, m_i)$ are assumed to be conditionally independent and follow a distribution from the exponential family with the probability density or mass function

$$f_{y|x,z,b}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i) = \exp[\{y_{ij}\alpha_{ij} - a_1(\alpha_{ij})\}/a_2(\phi) + a_3(y_{ij}, \phi)], \tag{1}$$

where functions $a_1(\cdot)$, $a_2(\cdot)$, and $a_3(\cdot)$ are user-specified, ϕ is a dispersion parameter, and α_{ij} is the canonical parameter which links the conditional mean, $\mu_{ij}^b = E(Y_{ij}|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i)$, via the identity $\mu_{ij}^b = \partial a_1(\alpha_{ij})/\partial\alpha_{ij}$.

A generalized linear mixed model (GLMM) relates μ_{ij}^b to the covariates and random effects via a regression model

$$g(\mu_{ij}^b) = \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{\mathrm{T}}\mathbf{b}_i, \tag{2}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients for the fixed effects, and $g(\cdot)$ is a link function. Random effects $\mathbf{b}_i$ are assumed to have a distribution, say, $f_b(\mathbf{b}_i; \boldsymbol{\sigma}_{\mathbf{b}})$, with an unknown parameter vector $\boldsymbol{\sigma}_{\mathbf{b}}$. The link function $g(\cdot)$ is monotone and differentiable, and its form can be differently specified for individual applications. For instance, for binary Y_{ij}, $g(\cdot)$ can be chosen as a logit, probit, or complementary log-log link, while for Poisson or Gamma variables Y_{ij}, $g(\cdot)$ is often set as a log link.

A useful class of models belonging to GLMMs is linear mixed models (LMM) where $g(\cdot)$ in (2) is set to be the identity function, leading to

$$Y_{ij} = \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{Z}_{ij}^{\mathrm{T}}\mathbf{b}_i + \epsilon_{ij} \tag{3}$$

where the error term ϵ_{ij} is often assumed to be normally distributed with mean 0 and unknown variance ϕ.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\sigma}_b^{\mathrm{T}}, \phi)^{\mathrm{T}}$ be the vector of model parameters. In the absence of response error, estimation of $\boldsymbol{\theta}$ is based on the likelihood for the observed data:

$$\mathscr{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathscr{L}_i(\boldsymbol{\theta}),$$

where

$$\mathscr{L}_i(\boldsymbol{\theta}) = \int \prod_{j=1}^{m_i} f_{y|x,z,b}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i; \boldsymbol{\theta}) f_b(\mathbf{b}_i; \boldsymbol{\sigma}_{\mathbf{b}}) d\mathbf{b}_i \tag{4}$$

is the marginal likelihood for cluster i, and $f_{y|x,z,b}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i; \boldsymbol{\theta})$ is determined by (1) in combination with (2). Maximizing $\mathscr{L}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ gives the maximum likelihood estimator of $\boldsymbol{\theta}$.

2.2 Measurement Error Models

When Y_{ij} is subject to measurement error, we observe a value that may differ from the true value; let S_{ij} denote such an observed measurement for Y_{ij}, and we call it a surrogate variable. In this paper we consider the case where Y_{ij} is a continuous variable only. Let $f_{s|y,x,z}(S_{ij}|\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i)$ or $f_{s|y,x,z}(S_{ij}|y_{ij}, \mathbf{x}_{ij}, \mathbf{z}_{ij})$ denote the conditional probability density (or mass) function for S_{ij} given $\{\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i\}$ or $\{Y_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij}\}$, respectively. It is often assumed that

$$f_{s|y,x,z}(s_{ij}|\mathbf{y}_i, \mathbf{x}_i, \mathbf{z}_i) = f_{s|y,x,z}(s_{ij}|y_{ij}, \mathbf{x}_{ij}, \mathbf{z}_{ij}).$$

This assumption says that given the true variables $\{Y_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij}\}$ for each subject j in a cluster i, the observed measurement S_{ij} is independent of variables $\{Y_{ik}, \mathbf{X}_{ik}, \mathbf{Z}_{ik}\}$ of other subjects in the same cluster for $k \neq j$.

Parametric modeling can be invoked to feature the relationship between the true response variable Y_{ij} and its surrogate measurement S_{ij}. One class of useful models are specified as

$$S_{ij} = h(Y_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij}; \boldsymbol{\gamma}_i) + e_{ij}, \tag{5}$$

where the stochastic noise term e_{ij} has mean zero. Another class of models are given by

$$S_{ij} = h(Y_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij}; \boldsymbol{\gamma}_i) \cdot e_{ij}, \tag{6}$$

where the stochastic term e_{ij} has mean 1. These models basically modulate the mean structure of the surrogate variable S_{ij}:

$$E(S_{ij}|\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i) = h(Y_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij}; \boldsymbol{\gamma}_i), \tag{7}$$

where the function form $h(\cdot)$ can be chosen differently to facilitate various applications, and $\boldsymbol{\gamma}_i$ is a vector of error parameters for cluster i. For cases where the measurement error process is homogeneous, e.g., same measuring system is used across clusters, we replace $\boldsymbol{\gamma}_i$ with a common parameter vector $\boldsymbol{\gamma}$.

Specification of $h(\cdot)$ reflects the feature of the measurement error model. For example, if $h(\cdot)$ is set as a linear function, model (5) gives a linear relationship between the response and surrogate measurements:

$$S_{ij} = \gamma_0 + \gamma_1 Y_{ij} + \gamma_2^{\mathrm{T}} \mathbf{X}_{ij} + \gamma_3^{\mathrm{T}} \mathbf{Z}_{ij} + e_{ij},$$

where parameters γ_0, γ_1, γ_2, and γ_3 control the dependence of surrogate measurement S_{ij} on the response and covariate variables; in the instance where both γ_2 and γ_3 are zero vectors, surrogate measurement S_{ij} is not affected by the measurements of the covariates and depends on the true response variable Y_{ij} only. More complex relationships can be delineated by employing nonlinear function forms for $h(\cdot)$. In our following simulation studies and data analysis, linear, exponential, and logarithmic functions are considered for $h(\cdot)$.

We call (5) *additive error models*, and (6) *multiplicative error models* to indicate how noise terms e_{ij} act relative to the mean structure of S_{ij}. Commonly, noise terms e_{ij} are assumed to be independent of each other, of the true responses as well as of the covariates. Let $f(e_{ij}; \boldsymbol{\sigma}_e)$ denote the probability density function of e_{ij}, where $\boldsymbol{\sigma}_e$ is an associated parameter vector. With model (5), the e_{ij} are often assumed to be normally distributed, while for model (6), a log normal or a Gamma distribution may be considered.

3 Asymptotic Bias Analysis

In this section we investigate asymptotic biases caused by response error under the two situations: (1) response error is totally ignored in estimation procedures, and (2) an intuitively compelling correction method is applied to adjust for measurement error in response.

3.1 Naive Analysis of Ignoring Measurement Error

We consider a naive analysis which fits the GLMM (1) to the observed raw data (hereafter referred to as NAI1), i.e., we assume that the S_{ij} are linked with covariates via the same random effects model. Let $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^{*\mathrm{T}}, \boldsymbol{\sigma}_b^{*\mathrm{T}}, \phi^*)^{\mathrm{T}}$ denote the corresponding parameter vector, and the corresponding working likelihood contributed from cluster i is given by

$$\mathscr{L}_i^w(\boldsymbol{\theta}^*) = \int \prod_{j=1}^{m} f_{y|x,z,b}(s_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i^*; \boldsymbol{\theta}^*) f_b(\mathbf{b}_i^*; \boldsymbol{\sigma}_b^*) d\mathbf{b}_i^*.$$

Maximizing $\sum_{i=1}^{n} \log \mathscr{L}_i^w(\boldsymbol{\theta}^*)$ with respect to $\boldsymbol{\theta}^*$ gives an estimator, say $\hat{\boldsymbol{\theta}}^*$, of $\boldsymbol{\theta}^*$.

Adapting the arguments of White [14] it can be shown that under certain regularity conditions, as $n \to \infty$, $\hat{\boldsymbol{\theta}}^*$ converges in probability to a limit that is the solution to a set of estimating equations

$$\mathrm{E}_{\text{true}} \left\{ \sum_{i=1}^{n} \partial \log \mathscr{L}_i^w(\boldsymbol{\theta}^*)/\partial \boldsymbol{\theta}^* \right\} = \mathbf{0}, \tag{8}$$

where the expectation is taken with respect to the true joint distribution of the associated random variables. The evaluation of (8) involves integration over the nonlinear error functions which are often intractable.

To gain insights into the impact of ignoring error in response, we consider a LMM

$$Y_{ij} = \beta_0 + (\beta_1 + b_i)X_{ij} + \epsilon_{ij}, \tag{9}$$

where β_0 and β_1 are regression parameters, the ϵ_{ij} are independent of each other and of other variables, $\epsilon_{ij} \sim N(0, \phi)$ with variance ϕ, and $b_i \sim \text{Normal}(0, \sigma_b^2)$ with variance σ_b^2. We consider the additive error model (5), where the e_{ij} are independent of each other and of other variables, $e_{ij} \sim N(0, \sigma_e^2)$, and the mean error structures are, respectively, specified as one of the following two cases.

Case 1 *Linear measurement error.*

Commonly seen in epidemiologic studies, this structure specifies a linear form for the measurement error

$$h(Y_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij}; \boldsymbol{\gamma}) = \gamma_0 + \gamma_1 Y_{ij},$$

where $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$, γ_0 represents a systematic error of the measuring device at $Y_{ij} = 0$, and γ_1 is a scale factor. It can be easily shown that simple relationship between the true and working parameters is

$$\beta_0^* = \gamma_0 + \gamma_1\beta_0,\ \beta_1^* = \gamma_1\beta_1,\ \sigma_b^{*2} = \gamma_1^2\sigma_b^2,$$

and

$$\phi^* = \gamma_1^2\phi + \sigma_e^2.$$

These results suggest that estimation of fix effect β_1 and variance component σ_b^2 is generally attenuated or inflated by factor γ_1, a factor which governs the difference between the true response Y_{ij} and surrogate measurement S_{ij}. When γ_1 equals 1, even if there is systematic measurement error involved with measuring Y_{ij} (i.e., $\gamma_0 \neq 0$), disregarding error in Y_{ij} does not bias point estimates of fix effect β_1 and variance component σ_b^2, but may reduce estimation precision.

Case 2 *Exponential measurement error.*

The second error structure specifies an exponential form for the measurement error

$$h(Y_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij}; \gamma) = \exp(\gamma Y_{ij}),$$

which may be useful to feature transformed response variables that are not measured precisely.

The bias in the naive estimator for fixed effect β_1 does not have an analytic form when the response is subject to nonlinear measurement error. To illustrate the induced bias in estimation of β_1 with response error ignored, we undertake a numerical study. The covariates X_{ij} are independently generated from a normal distribution $N(0, 1)$. We fix the values of β_0 and ϕ at -1 and 0.01, respectively, and consider values of σ_b^2 to be 0.01, 0.25, and 1, respectively. The error parameters are, respectively, specified as $\gamma = 0.5$ and 1, and $\sigma_e^2 = 0.01$, 0.25, and 0.75.

As shown in Fig. 1, the relationship between the naive limit β_1^* and the value of β_1 is nonlinear. For instance, when $\gamma = 0.5$, the naive estimate is attenuated for small values of β_1 but is inflated for large values of β_1. In general, the direction and magnitude of the bias induced by nonlinear response error depend on the function form of $h(\cdot)$ as well as the magnitude of the parameters in the measurement error process.

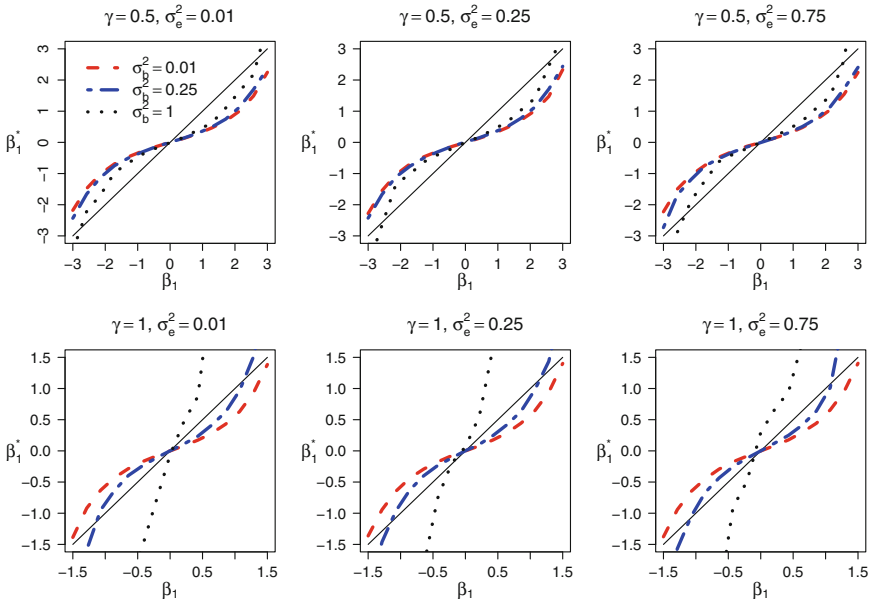

Fig. 1 Bias in β_1^* from the completely naive approach induced by an exponential error model. The *dashed*, *two-dash*, and *dotted lines* correspond to $\sigma_b^2 = 0.01$, 0.25, and 1, respectively

3.2 Analysis of Transformed Data

With the response process modeled by an LMM, Buonaccorsi [1] considered an intuitively tempting method to correct for response error in estimation. The idea is to employ a two-step approach to correct for response error effects. In the first step, keeping the covariates fixed, we use the mean function $h(\cdot)$ of the measurement error model and find its inverse function $h^{-1}(\cdot)$, and then calculate a pseudo-response

$$\tilde{Y}_{ij} = h^{-1}(S_{ij}; \boldsymbol{\gamma}).$$

In the second step, we perform standard statistical analysis with $\tilde{Y}_{ij}$ taken as a response variable. This approach (hereafter referred to as NAI2) is generally preferred over NAI1, as it reduces a certain amount of bias induced by response measurement error. However, this method does not completely remove the biases induced from response error.

To evaluate the performance of using pseudo-response in estimation procedures, we may follow the same spirit of Sect. 3.1 to conduct bias analysis. As it is difficult to obtain analytic results for general models, here we perform empirical studies by employing the same response model (9) and the measurement error model for Case 2 as in Sect. 3.1.

It is seen that as expected, the asymptotic bias, displayed in Fig. 2, is smaller than that from the NAI1 analysis. This confirms that the NAI2 method outperforms

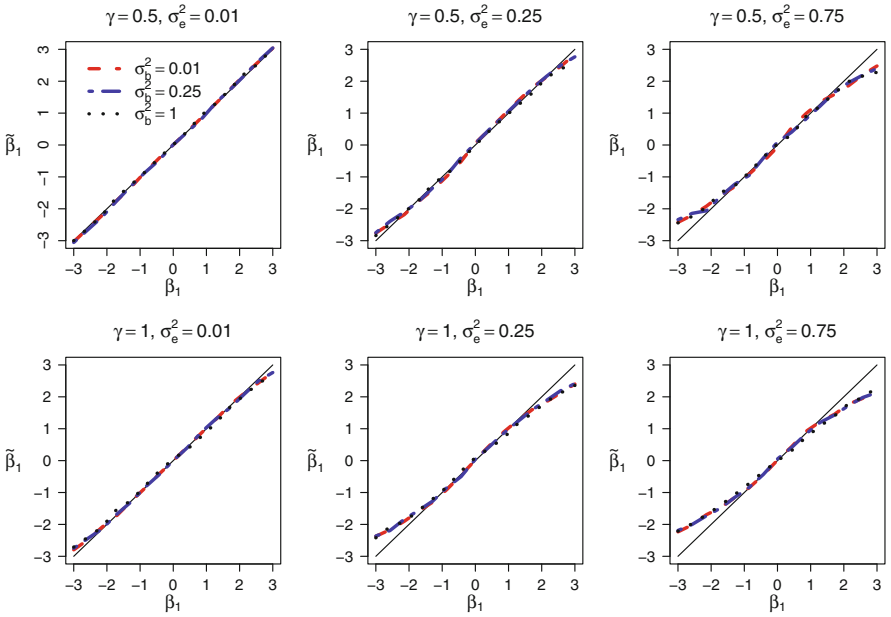

Fig. 2 Bias in $\tilde{\beta}_1$ from NAI2 analyses with response subject to exponential error. The *dashed*, *two-dash*, and *dotted lines* are for $\sigma_b^2 = 0.01$, 0.25, and 1, respectively

the NAI1 method. However, the NAI2 method does not completely remove the bias induced in the response error. The asymptotic bias involved in the NAI2 method is affected by the size of the covariate effect as well as the degree of response error. The asymptotic bias increases as the size of β_1 increases. Furthermore, the values of the error parameters γ and σ_e^2 have significant impact on the bias; the size of the bias tends to increase as σ_e^2 increases.

4 Inference Methods

The analytic and numerical results in Sect. 3 demonstrate that disregarding response error may yield biased estimation results. To account for the response error effects, in this section we develop valid inference methods for the response model parameter vector $\boldsymbol{\theta}$. Our development accommodates different scenarios pertaining to the knowledge of response measurement error. Let $\boldsymbol{\eta}$ denote the parameter vector associated with a parametric model of the response measurement error process. Estimation of $\boldsymbol{\theta}$ may suffer from nonidentifiability issues in the presence of measurement error in the variables. To circumvent this potential problem, we consider three useful situations: (i) $\boldsymbol{\eta}$ is known, (ii) $\boldsymbol{\eta}$ is unknown but a validation subsample is available, and (iii) $\boldsymbol{\eta}$ is unknown but replicates for the surrogates are available.

The first situation highlights the idea of addressing the difference between the surrogate measurements and the response variables without worrying about model nonidentifiability issues. The second and third scenarios reflect useful practical settings where error model parameter $\boldsymbol{\eta}$ is often unknown, but estimable from additional data sources such as a validation subsample or replicated surrogate measurements. For each of these three situations, we propose strategies for estimating the response model parameters and derive the asymptotic properties of the resulting estimators.

4.1 $\boldsymbol{\eta}$ Is Known

In some applications, the value of $\boldsymbol{\eta}$ is known to be $\boldsymbol{\eta}_0$, say, from a priori study, or specified by the analyst for sensitivity analyses. Inference about $\boldsymbol{\theta}$ is then carried out based on the marginal likelihood of the observed data:

$$\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\eta}_0) = \prod_{i=1}^{n} \mathscr{L}_i(\boldsymbol{\theta}, \boldsymbol{\eta}_0)$$

where

$$\mathscr{L}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) = \int \left\{ \prod_{j=1}^{m_i} \int f_{s|y,x,z}(s_{ij}|y_{ij}, \mathbf{x}_{ij}, \mathbf{z}_{ij}; \boldsymbol{\eta}) \right.$$

$$\left. \times f_{y|x,z,b}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i; \boldsymbol{\theta}) dy_{ij} \right\} f_b(\mathbf{b}_i; \boldsymbol{\sigma}_b) d\mathbf{b}_i,$$

which requires the conditional independence assumption

$$f_{s|y,x,z,b}(s_{ij}|y_{ij}, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i; \boldsymbol{\eta}) = f_{s|y,x,z}(s_{ij}|y_{ij}, \mathbf{x}_{ij}, \mathbf{z}_{ij}; \boldsymbol{\eta}); \tag{10}$$

$f_{s|y,x,z,b}(s_{ij}|y_{ij}, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i; \boldsymbol{\eta})$ and $f_{s|y,x,z}(s_{ij}|y_{ij}, \mathbf{x}_{ij}, \mathbf{z}_{ij}; \boldsymbol{\eta})$ represent the conditional probability density function of S_{ij} given $\{Y_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij}, \mathbf{b}_i\}$ and $\{Y_{ij}, \mathbf{X}_{ij}, \mathbf{Z}_{ij}\}$, respectively.

Maximizing $\sum_{i=1}^{n} \log \mathscr{L}_i(\boldsymbol{\theta}, \boldsymbol{\eta}_0)$ with respect to the parameter $\boldsymbol{\theta}$ gives the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. Let $\mathbf{U}_i(\boldsymbol{\theta}, \boldsymbol{\eta}_0) = \partial \log \mathscr{L}_i(\boldsymbol{\theta}, \boldsymbol{\eta}_0)/\partial \boldsymbol{\theta}$. From standard likelihood theory, under regularity conditions, $\hat{\boldsymbol{\theta}}$ is a consistent estimator for $\boldsymbol{\theta}$. As $n \to \infty$, $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normally distributed with mean $\mathbf{0}$ and variance $\mathscr{I}^{-1}$, where $\mathscr{I} = E\{-\partial \mathbf{U}_i(\boldsymbol{\theta}, \boldsymbol{\eta}_0)/\partial \boldsymbol{\theta}^{\mathrm{T}}\}$. By the Bartlett identity and the Law of Large Numbers, $\mathscr{I}$ can be consistently estimated by $n^{-1} \sum_{i=1}^{n} \mathbf{U}_i(\hat{\boldsymbol{\theta}}, \boldsymbol{\eta}_0) \mathbf{U}_i^{\mathrm{T}}(\hat{\boldsymbol{\theta}}, \boldsymbol{\eta}_0)$.

4.2 $\boldsymbol{\eta}$ Is Estimated from Validation Data

In many applications, $\boldsymbol{\eta}$ is often unknown and must be estimated from additional data sources, such as a validation subsample or replicates of measurements of Y_{ij}. Here we consider the case that a validation subsample is available, and in the next section we discuss the situation with replicated measurements.

Assume that the validation subsample is randomly selected, and let $\delta_{ij} = 1$ if Y_{ij} is available and $\delta_{ij} = 0$ otherwise. Specifically, if $\delta_{ij} = 1$, then measurements $\{y_{ij}, s_{ij}, \mathbf{x}_{ij}, \mathbf{z}_{ij}\}$ are available; when $\delta_{ij} = 0$, measurements $\{s_{ij}, \mathbf{x}_{ij}, \mathbf{z}_{ij}\}$ are available. Let $N_v = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \delta_{ij}$ be the number of the measurements in the validation subsample. The full marginal likelihood of the main data and the validation data contributed from cluster i is given by

$$\mathscr{L}_{Fi}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \int \left[\prod_{j=1}^{m_i} \left\{ f_{s|x,z,b}(s_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i; \boldsymbol{\theta}, \boldsymbol{\eta}) \right\}^{1-\delta_{ij}} \right.$$

$$\left. \times \left\{ f_{s,y|x,z,b}(s_{ij}, y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i; \boldsymbol{\theta}, \boldsymbol{\eta}) \right\}^{\delta_{ij}} \right] f_b(\mathbf{b}_i; \boldsymbol{\sigma}_b) d\mathbf{b}_i, \tag{11}$$

where $f_{s,y|x,z,b}(s_{ij}, y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i; \boldsymbol{\theta}, \boldsymbol{\eta})$ represents the conditional probability density functions of $\{S_{ij}, Y_{ij}\}$, given the covariates $\{\mathbf{x}_{ij}, \mathbf{z}_{ij}\}$ and random effects $\mathbf{b}_i$.

Under the conditional independence assumption (10), we obtain

$$f_{s|x,z,b}(s_{ij}|\mathbf{x}_{ij},\mathbf{z}_{ij},\mathbf{b}_i;\boldsymbol{\theta},\boldsymbol{\eta}) = \int f_{s|y,x,z}(s_{ij}|y_{ij},\mathbf{x}_{ij},\mathbf{z}_{ij};\boldsymbol{\eta}) f_{y|x,z,b}(y_{ij}|\mathbf{x}_{ij},\mathbf{z}_{ij},\mathbf{b}_i;\boldsymbol{\theta}) dy_{ij},$$

and

$$f_{s,y|x,z,b}(s_{ij},y_{ij}|\mathbf{x}_{ij},\mathbf{z}_{ij},\mathbf{b}_i;\boldsymbol{\theta},\boldsymbol{\eta}) = f_{s|y,x,z}(s_{ij}|y_{ij},\mathbf{x}_{ij},\mathbf{z}_{ij};\boldsymbol{\eta}) f_{y|x,z,b}(y_{ij}|\mathbf{x}_{ij},\mathbf{z}_{ij},\mathbf{b}_i;\boldsymbol{\theta}),$$

where $f_{s|y,x,z}(s_{ij}|y_{ij},\mathbf{x}_{ij},\mathbf{z}_{ij};\boldsymbol{\eta})$ is the conditional probability density function determined by the measurement error model such as (5) or (6), and $f_{y|x,z,b}(y_{ij}|\mathbf{x}_{ij},\mathbf{z}_{ij},\mathbf{b}_i;\boldsymbol{\theta})$ is the conditional probability density function specified by the GLMM (1) in combination with (2).

Let

$$\mathscr{L}_{\theta i}(\boldsymbol{\theta},\boldsymbol{\eta}) = \int \left[\prod_{j=1}^{m_i} \left\{ f_{s|x,z,b}(s_{ij}|\mathbf{x}_{ij},\mathbf{z}_{ij},\mathbf{b}_i;\boldsymbol{\theta},\boldsymbol{\eta}) \right\}^{1-\delta_{ij}} \times \left\{ f_{y|x,z,b}(y_{ij}|\mathbf{x}_{ij},\mathbf{z}_{ij},\mathbf{b}_i;\boldsymbol{\theta}) \right\}^{\delta_{ij}} \right] f_b(\mathbf{b}_i;\boldsymbol{\sigma}_b) d\mathbf{b}_i,$$

and

$$\mathscr{L}_{\eta i}(\boldsymbol{\eta}) = \prod_{j=1}^{m_i} \left\{ f_{s|y,x,z}(s_{ij}|y_{ij},\mathbf{x}_{ij},\mathbf{z}_{ij};\boldsymbol{\eta}) \right\}^{\delta_{ij}},$$

then $\mathscr{L}_{Fi}(\boldsymbol{\theta},\boldsymbol{\eta}) = \mathscr{L}_{\theta i}(\boldsymbol{\theta},\boldsymbol{\eta})\, \mathscr{L}_{\eta i}(\boldsymbol{\eta})$.

Inference about $\{\boldsymbol{\theta},\boldsymbol{\eta}\}$ can, in principle, be conducted by maximizing $\prod_{i=1}^{n} \mathscr{L}_{Fi}(\boldsymbol{\theta},\boldsymbol{\eta})$, or $\sum_{i=1}^{n} \log \mathscr{L}_{Fi}(\boldsymbol{\theta},\boldsymbol{\eta})$, with respect to $\{\boldsymbol{\theta},\boldsymbol{\eta}\}$. When the dimension of $\{\boldsymbol{\theta},\boldsymbol{\eta}\}$ is large, direct maximization of $\sum_{i=1}^{n} \log \mathscr{L}_{Fi}(\boldsymbol{\theta},\boldsymbol{\eta})$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ can be computationally demanding. We propose to use a two-stage estimation procedure as an alternative to the joint maximization procedure.

Let $\mathbf{U}_i^*(\boldsymbol{\theta},\boldsymbol{\eta}) = \partial \log \mathscr{L}_{\theta i}(\boldsymbol{\theta},\boldsymbol{\eta})/\partial\boldsymbol{\theta}$ and $\mathbf{Q}_i(\boldsymbol{\eta}) = \partial \log \mathscr{L}_{\eta i}(\boldsymbol{\eta})/\partial\boldsymbol{\eta}$. In the first stage, estimator for $\boldsymbol{\eta}$, denoted by $\hat{\boldsymbol{\eta}}$, is obtained by solving

$$\sum_{i=1}^{n} \mathbf{Q}_i(\boldsymbol{\eta}) = \mathbf{0}.$$

In the second stage, replace $\boldsymbol{\eta}$ with $\hat{\boldsymbol{\eta}}$ and solve

$$\sum_{i=1}^{n} \mathbf{U}_i^*(\boldsymbol{\theta},\hat{\boldsymbol{\eta}}) = \mathbf{0} \tag{12}$$

for $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}_p$ denote the solution to (12).

Assume that the size of the validation sample is increasing with the sample size n on the same scale, i.e., as $n \to \infty$ and $N_v/n \to \rho$ for a positive constant ρ. Then under regularity conditions, $\sqrt{n}(\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta})$ is asymptotically normally distributed with mean 0 and variance given by

$$\begin{aligned}\boldsymbol{\Sigma}^* = &\left[-E\{\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\theta}^{\mathrm{T}}\}\right]^{-1} + \left[E\{\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\theta}^{\mathrm{T}}\}\right]^{-1} E\{\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\} \\ &\times \left[E\{\partial \mathbf{Q}_i(\boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\}\right]^{-1} \left[E\{\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\}\right]^{\mathrm{T}} \left[E\{\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\theta}^{\mathrm{T}}\}\right]^{-1}.\end{aligned}$$

The proof is outlined in the Appendix. An estimate of $\boldsymbol{\Sigma}^*$ can be obtained by replacing $E\{\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\theta}^{\mathrm{T}}\}$, $E\{\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\}$, and $E\{\partial \mathbf{Q}_i(\boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\}$ with their empirical counterparts $n^{-1}\sum_{i=1}^n \partial \mathbf{U}_i^*(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\eta}})/\partial \boldsymbol{\theta}^{\mathrm{T}}$, $n^{-1}\sum_{i=1}^n \partial \mathbf{U}_i^*(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\eta}})/\partial \boldsymbol{\eta}^{\mathrm{T}}$, and $n^{-1}\sum_{i=1}^n \partial \mathbf{Q}_i(\hat{\boldsymbol{\eta}})/\partial \boldsymbol{\eta}^{\mathrm{T}}$, respectively.

4.3 Inference with Replicates

In this section we discuss inferential procedures for the setting with replicates of the surrogate measurements for Y_{ij}. Suppose the response variable for each subject in a cluster is measured repeatedly, and let S_{ijr} denote the rth observed measurement for subject j in cluster i, $r = 1, \ldots, d_{ij}$, where the replicate number d_{ij} can vary from subject to subject. For $r \neq r'$, S_{ijr} and $S_{ijr'}$ are assumed to be conditionally independent, given $\{\mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i, \mathbf{b}_i\}$. The marginal likelihood contributed from cluster i is given by

$$\begin{aligned}\mathscr{L}_{Ri}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \int f_b(\mathbf{b}_i; \boldsymbol{\sigma}_b) \prod_{j=1}^{m_i} \Bigg\{ \int f_{y|x,z,b}(y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i; \boldsymbol{\theta}) \\ \times \prod_{r=1}^{d_{ij}} f_{s|y,x,z,b}(s_{ijr}|y_{ij}, \mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i; \boldsymbol{\eta}) dy_{ij} \Bigg\} d\mathbf{b}_i.\end{aligned}$$

Unlike the two-stage estimation procedure for the case with validation data, estimation for $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ generally cannot be separated from each other, because information on the underlying true responses and the measurement process is mixed together. A joint estimation procedure for $\{\boldsymbol{\theta}, \boldsymbol{\eta}\}$ by maximizing $\prod_{i=1}^n \mathscr{L}_{Ri}(\boldsymbol{\theta}, \boldsymbol{\eta})$ is particularly required.

Specifically, let

$$\mathscr{U}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) = \partial \log \mathscr{L}_{Ri}(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\theta}, \quad \text{and} \quad \mathscr{Q}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) = \partial \log \mathscr{L}_{Ri}(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\eta}$$

be the score functions. Define

$$\Psi_{Ri}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \begin{pmatrix} \mathscr{Q}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) \\ \mathscr{U}_i(\boldsymbol{\theta}, \boldsymbol{\eta}) \end{pmatrix}.$$

The maximum likelihood estimators for $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ is obtained by solving

$$\sum_{i=1}^{n} \Psi_{Ri}(\boldsymbol{\theta}, \boldsymbol{\eta}) = \mathbf{0};$$

we let $(\hat{\boldsymbol{\theta}}_R, \hat{\boldsymbol{\eta}}_R)$ denote the solution.

Under suitable regularity conditions, $n^{1/2} \begin{pmatrix} \hat{\boldsymbol{\theta}}_R - \boldsymbol{\theta} \\ \hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta} \end{pmatrix}$ is asymptotically normally distributed with mean $\mathbf{0}$ and covariance matrix $[E\{\Psi_{Ri}(\boldsymbol{\theta}, \boldsymbol{\eta})\Psi_{Ri}^{\mathrm{T}}(\boldsymbol{\theta}, \boldsymbol{\eta})\}]^{-1}$.

4.4 Numerical Approximation

To implement the proposed methods, numerical approximations are often needed because integrals involved in the likelihood formulations do not have analytic forms in general. With low dimensional integrals, Gaussian–Hermite quadratures may be invoked to handle integrals without a closed form. For example, the integral with an integrand of form $\exp(-u^2)f(u)$ is approximated by a sum

$$\int_{-\infty}^{\infty} \exp(-u^2) f(u) du \approx \sum_{k=1}^{K} w_k f(t_k),$$

where $f(\cdot)$ is a given function, K is the number of selected points, and t_k and w_k are the value and the weight of the kth designated point, respectively. The approximation accuracy relies on the order of the quadrature approximations. We found in our simulation that a quadrature approximation with order 5 performs adequately well for a single integral; as the number of random components increases, more quadrature points are required in order to obtain a good approximation. When $f(\cdot)$ is a symmetric or nearly symmetric function, the approximation is generally good, even when the number of quadrature points is chosen to be small.

Computation quickly becomes infeasible as the number of nested random components grows [9]. The convergence of an optimization procedure can be very slow if the dimension of the random components is high. One approach to deal with such integrals is to linearize the model with respect to the random effects, e.g., using a first-order population-averaged approximation to the marginal distribution by expanding the conditional distribution about the average random effect [12]. Alternatively, Laplace's approximation can be useful to obtain an approximate likelihood function with a closed form [12, 15]. The basic form of linearization using Laplace's approximation is a second-order Taylor series expansion of the integrand $f(\mathbf{u})$ and is given by $\int_{\mathbb{R}^d} f(\mathbf{u}) d\mathbf{u} \approx (2\pi)^{d/2} f(\mathbf{u}_0) \left| -\partial^2 \log f(\mathbf{u}_0) / \partial \mathbf{u} \partial \mathbf{u}^{\mathrm{T}} \right|^{-1/2}$, where d is the dimension of $\mathbf{u}$, and $\mathbf{u}_0$ is the mode of $f(\mathbf{u})$, i.e., the solution to $\partial \log f(\mathbf{u}) / \partial \mathbf{u} = \mathbf{0}$. To construct Laplace's approximation, the first two derivatives of $\log f(\mathbf{u})$ are basically required.

5 Simulation Studies

We conduct simulation studies to assess the performance of the proposed likelihood-based methods. We consider the setting with $n = 100$ and $m_i = 5$ for $i = 1, \ldots, n$. The covariates X_{ij} are simulated from the standard normal distribution, and random effects b_i are generated from a normal distribution with mean 0 and variance $\sigma_b^2 = 0.04$. The response measurements are generated from the model

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + b_i X_{ij} + \epsilon_{ij},$$

where $\epsilon_{ij} \sim N(0, \phi)$, and the parameter values are set as $\beta_0 = -1$, $\beta_1 = \log(0.5)$, and $\phi = 0.04$.

We consider two models for the measurement error process. That is, surrogate measurements S_{ij} are simulated from one of the two measurement error models:

(M1). $S_{ij} = \exp(\gamma Y_{ij}) + e_{ij}$,
(M2). $S_{ij} = \gamma_0 + \gamma_1 Y_{ij} + e_{ij}$,

where e_{ij} is independent of $\mathbf{Y}_i$ and $\mathbf{X}_i$, and follows a normal distribution with mean 0 and variance $\sigma_e^2 = 0.04$. For error model (M1), the error parameters are specified as $\gamma = 0.5$. For error model (M2), the parameters are specified as $\gamma_0 = 0.5$ and $\gamma_1 = 0.5$.

Let $\boldsymbol{\eta}$ denote the vector of associated parameters for the measurement error model. Specifically, in error model (M1), $\boldsymbol{\eta} = (\gamma, \sigma_e^2)^{\mathrm{T}}$; while in error model (M2), $\boldsymbol{\eta} = (\gamma_0, \gamma_1, \sigma_e^2)^{\mathrm{T}}$. We evaluate the proposed methods under two scenarios regarding the knowledge of $\boldsymbol{\eta}$: (i) $\boldsymbol{\eta}$ is treated as known, and (ii) $\boldsymbol{\eta}$ is estimated from internal validation data. For scenario (ii), we obtain a validation subsample by randomly selecting one subject from each cluster. We use Gaussian quadrature of order 15 in the numerical approximation for the likelihood-based approaches. Two thousand simulations are run for each parameter configuration.

We conduct three analyses for each simulated data set: the two naive approaches described in Sects. 3.1 and 3.2 and the proposed methods described in Sect. 4. We report the simulation results based on four measures: relative bias in percent (Bias%), sample standard deviation of the estimates (SD), average of model-based standard errors (ASE), and coverage probability of the 95 % confidence interval (CP%).

Table 1 reports the results for the exponential measurement error model (M1). As expected, the NAI1 approach produces very biased (attenuated) estimates of the fixed-effect parameter β_1, and the coverage rates of the 95 % confidence interval are close to 0. The NAI2 approach, which analyzes transformed surrogate responses, produces slightly better estimates of β_1. The magnitude of the relative bias, although smaller than that from NAI1, is still substantial. In contrast, the proposed likelihood approaches give consistent estimates for β_1 in both scenarios, and the coverage rates of its 95 % confidence intervals are close to the nominal value.

Table 2 reports the results for the linear measurement error model (M2). Again the estimates for β_1 from the NAI1 approach are biased, and the values are scaled

Table 1 Simulation results for cases with measurement error model (M1) (2000 simulations)

	NAI1[a]				NAI2[b]				Proposed[c]			
	Bias%	SD	ASE	CP%	Bias%	SD	ASE	CP%	Bias%	SD	ASE	CP%
	Scenario (i): $\boldsymbol{\eta}$ is known											
β_0	−164.5	0.011	0.010	< 0.1	16.7	0.051	0.046	5.0	−0.56	0.037	0.035	94.2
β_1	−67.7	0.013	0.014	< 0.1	16.0	0.064	0.060	52.9	−0.83	0.040	0.039	94.9
σ_b^2	−80.3	0.003	0.184	100.0	235.8	0.090	0.253	94.3	−2.13	0.016	0.018	93.8
ϕ	17.6	0.003	0.035	100.0	2439.1	0.234	0.035	< 0.1	3.65	0.022	0.025	96.0
	Scenario (ii): $\boldsymbol{\eta}$ is estimated from internal validation data											
β_0	–	–	–	–	13.7	0.072	0.042	19.8	−0.04	0.040	0.042	94.8
β_1	–	–	–	–	13.1	0.067	0.054	61.0	0.32	0.053	0.049	94.6
σ_b^2	–	–	–	–	183.4	0.079	0.242	96.4	4.57	0.024	0.028	95.7
ϕ	–	–	–	–	1975.9	0.220	0.035	< 0.1	−3.14	0.017	0.014	93.3

[a]NAI1: naive LMM analysis of observed data ignoring measurement error.
[b]NAI2: naive LMM analysis of the constructed pseudo-response data.
[c]Proposed: the proposed likelihood method that accounts for measurement error.

Table 2 Simulation results for cases with linear measurement error model (M2) (2000 simulations)

	NAI1[a]				NAI2[b]				Proposed[c]			
	Bias%	SD	ASE	CP%	Bias%	SD	ASE	CP%	Bias%	SD	ASE	CP%
	Scenario (i): $\boldsymbol{\eta}$ is known											
β_0	−100.0	0.010	0.010	< 0.1	−0.04	0.021	0.021	95.2	−0.04	0.021	0.022	95.8
β_1	−50.0	0.015	0.015	< 0.1	0.08	0.030	0.030	94.3	0.10	0.030	0.030	94.8
σ_b^2	−75.0	0.003	0.162	100.0	−0.01	0.012	0.162	100.0	−3.28	0.012	0.013	94.9
ϕ	25.0	0.003	0.035	100.0	400.12	0.014	0.035	< 0.1	−0.96	0.014	0.014	95.1
	Scenario (ii): $\boldsymbol{\eta}$ is estimated from internal validation data											
β_0	–	–	–	–	−0.13	0.037	0.021	75.3	−0.11	0.031	0.030	94.3
β_1	–	–	–	–	0.44	0.043	0.030	84.0	0.22	0.039	0.041	95.6
σ_b^2	–	–	–	–	0.96	0.013	0.161	100.0	−2.72	0.017	0.017	94.8
ϕ	–	–	–	–	406.53	0.025	0.035	< 0.1	−2.02	0.018	0.022	95.7

[a]NAI1: naive LMM analysis of observed data ignoring measurement error.
[b]NAI2: naive LMM analysis of the constructed pseudo-response data.
[c]Proposed: the proposed likelihood method that accounts for measurement error.

approximately by a factor of γ_1, which is in agreement with the analytical result shown in Sect. 3. The NAI2 approach yields good estimates for β_0, β_1, and σ_b^2 with small finite sample biases. The NAI2 estimates for ϕ, however, are very biased, resulting in coverage rates of corresponding confidence intervals far from the nominal value of 95 %. In contrast, the proposed likelihood-based approach gives consistent estimators for the fixed-effect and variance component, and the associated standard errors are similar to the empirical standard deviations. As a result, the coverage rates of the 95 % confidence intervals are close to the nominal value.

6 Application

We illustrate our proposed methods by analyzing the data from the Framingham Heart Study. The data set includes exams #2 and #3 for $n = 1615$ male subjects aged 31–65 [3]. Two systolic blood pressure (SBP) readings were taken during each exam. One of the clinical interests is to understand the relationship between SBP and potential risk factors such as baseline smoking status and age [6, 8, 11]. The risk factors, however, may not have linear effects on SBP directly.

Preliminary exploration shows that SBP measurements are skewed, and using a square-root transformation to $(T_{ij} - 50)$ is reasonably satisfactory for obtaining a symmetric data distribution, where T_{ij} represents the true SBP measurement for subject i at time j, where $j = 1$ corresponds to exam #2, and $j = 2$ for exam #3, and $i = 1, \ldots, n$. We now let Y_{ij} denote such a transformed variable, i.e., $Y_{ij} = \sqrt{T_{ij} - 50}$. We assume that the Y_{ij} follow the model

$$Y_{ij} = \beta_0 + \beta_{\text{age}} X_{ij1} + \beta_{\text{smoke}} X_{ij2} + \beta_{\text{exam}} X_{ij3} + b_i + \epsilon_{ij}, \quad j = 1, 2, \ i = 1, \ldots, n,$$

where X_{ij1} is the baseline age of subject i at exam #2, X_{ij2} is the indicator variable for baseline smoking status of subject i at exam #1, X_{ij3} is 1 if $j = 2$ and 0 otherwise, and b_i and ϵ_{ij} are assumed to be independently and normally distributed with means 0 and variances given by σ_b^2 and ϕ, respectively.

Because a subject's SBP changes over time, the two individual SBP readings at each exam are regarded as replicated surrogates. Several measurement error models for SBP reading have been proposed by different researchers [2, 7, 13]. Let T^*_{ijr} be the rth observed SBP reading for subject i at time j, $i = 1, \ldots, n, j = 1, 2, r = 1, 2$. We consider an error model $\log(T^*_{ijr} - 50) = \log(T_{ij} - 50) + e_{ijr}$, suggested by Wang et al. [13], where the e_{ijr} are assumed to be independent of each other and of other variables, and are normally distributed with mean 0 and variance σ_e^2. Let S_{ijr} denote $\log(T^*_{ijr} - 50)$, then the measurement error model is equivalently given by

$$S_{ijr} = 2 \log(Y_{ij}) + e_{ijr}.$$

Table 3 Analysis of data from the Framingham Heart Study

	NAI1[a]			NAI2[b]			Proposed[c]		
	Est.	SE	p-value	Est.	SE	p-value	Est.	SE	p-value
β_0	4.117	0.030	< 0.001	7.727	0.140	< 0.001	7.729	0.156	< 0.001
β_{age}	0.006	0.001	< 0.001	0.029	0.003	< 0.001	0.027	0.003	< 0.001
β_{smoke}	-0.027	0.012	0.031	-0.122	0.057	0.032	-0.120	0.061	0.048
β_{exam}	-0.020	0.004	< 0.001	-0.086	0.018	< 0.001	-0.087	0.017	< 0.001
σ_b^2	0.036	0.021	0.083	0.782	0.020	< 0.001	0.754	0.040	< 0.001
ϕ	0.013	0.018	0.474	0.248	0.018	< 0.001	0.120	0.007	< 0.001

[a]NAI1: naive LMM analysis of observed data ignoring measurement error.
[b]NAI2: naive LMM analysis of the constructed pseudo-response data.
[c]Proposed: the proposed likelihood method that accounts for measurement error.

Table 3 reports results from analyses using the proposed method and the two naive approaches. The estimated regression coefficients β_{age}, β_{smoke}, and β_{exam} from the proposed method are 0.027, -0.120, and -0.087, respectively. At the 5 % significance level, age is significantly associated with increasing blood pressure. The negative coefficient for smoking status may suggest an effect of smoking on decreasing blood pressure. As expected, the results from the NAI2 approach are similar to those from the proposed method due to the small value of the measurement error variance. The NAI1 estimates, however, are not comparable to those from the NAI2 and the proposed method, possibly in part due to a different scale of the response variable.

7 Discussion

In this paper, we exploit analysis of response-error-contaminated clustered data within the framework of generalized linear mixed models. Although in some situations ignoring error in response does not alter point estimates of regression parameters, ignoring error in response does affect inference results for general circumstances. Error in response can produce seriously biased results.

In this paper we perform asymptotic bias analysis to assess the impact of ignoring error in response. We investigate the performance of a partial-error-correction method that was intuitively used in the literature [1]. To fully account for error effects, we develop valid inferential procedures for various practical settings which pertain to the information on response error. Simulation studies demonstrate satisfactory performance of the proposed methods under various settings.

Appendix

Let $\Psi_i(\boldsymbol{\theta}, \boldsymbol{\eta}) = \begin{pmatrix} \mathbf{Q}_i(\boldsymbol{\eta}) \\ \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta}) \end{pmatrix}$. Because $(\hat{\boldsymbol{\theta}}_p, \hat{\boldsymbol{\eta}})$ is a solution to $\Psi_i(\boldsymbol{\theta}, \boldsymbol{\eta}) = 0$, by first-order Taylor series approximation, we have

$$n^{1/2} \begin{pmatrix} \hat{\boldsymbol{\eta}} - \boldsymbol{\eta} \\ \hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta} \end{pmatrix} = - \begin{pmatrix} E\left\{\partial \mathbf{Q}_i(\boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\right\} & 0 \\ E\left\{\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\right\} & E\left\{\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\theta}^{\mathrm{T}}\right\} \end{pmatrix}^{-1} \times n^{-1/2} \sum_{i=1}^{n} \Psi_i(\boldsymbol{\theta}, \boldsymbol{\eta}) + o_p(1).$$

It follows that $n^{1/2}(\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta})$ equals

$$-n^{-1/2} \left[E\left\{\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\theta}^{\mathrm{T}}\right\}\right]^{-1} \left\{ \sum_{i=1}^{n} \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta}) - E\left\{\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\right\} \times \left[E\left\{\partial \mathbf{Q}_i(\boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\right\}\right]^{-1} \sum_{i=1}^{n} \mathbf{Q}_i(\boldsymbol{\eta}) \right\} + o_p(1) = -n^{-1/2} \Gamma^{-1}(\boldsymbol{\theta}, \boldsymbol{\eta}) \sum_{i=1}^{n} \Omega_i(\boldsymbol{\theta}, \boldsymbol{\eta}) + o_p(1),$$

where $\Omega_i(\boldsymbol{\theta}, \boldsymbol{\eta}) = \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta}) - E\{\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\}[E\{\partial \mathbf{Q}_i(\boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\}]^{-1} \mathbf{Q}_i(\boldsymbol{\eta})$, and $\Gamma(\boldsymbol{\theta}, \boldsymbol{\eta}) = E\{\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\theta}^{\mathrm{T}}\}$.

Applying the Central Limit Theorem, we can show that $n^{1/2}(\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta})$ is asymptotically normally distributed with mean 0 and asymptotic covariance matrix given by $\Gamma^{-1} \Sigma (\Gamma^{-1})^{\mathrm{T}}$, where $\Sigma = E\{\Omega_i(\boldsymbol{\theta}, \boldsymbol{\eta}) \Omega_i^{\mathrm{T}}(\boldsymbol{\theta}, \boldsymbol{\eta})\}$. But under suitable regularity conditions and correct model specification, $E\{\mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta}) \mathbf{U}_i^{*\mathrm{T}}(\boldsymbol{\theta}, \boldsymbol{\eta})\} = E\{-\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\theta}^{\mathrm{T}}\}$, $E\{\mathbf{Q}_i(\boldsymbol{\eta}) \mathbf{Q}_i^{\mathrm{T}}(\boldsymbol{\eta})\} = E\{-\partial \mathbf{Q}_i(\boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\}$, and $E\{\mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta}) \mathbf{Q}_i^{\mathrm{T}}(\boldsymbol{\eta})\} = E\{-\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\}$. Thus,

$$\Sigma = E\{-\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\theta}^{\mathrm{T}}\} + E\{\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\} \left[E\{\partial \mathbf{Q}_i(\boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\}\right]^{-1} \times \left[E\{\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\}\right]^{\mathrm{T}}.$$

Therefore, the asymptotic covariance matrix for $n^{1/2}(\hat{\boldsymbol{\theta}}_p - \boldsymbol{\theta})$ is

$$\Sigma^* = \left[E\{-\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\theta}^{\mathrm{T}}\}\right]^{-1} + \left[E\{-\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\theta}^{\mathrm{T}}\}\right]^{-1} E\{\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\} \times \left[E\{\partial \mathbf{Q}_i(\boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\}\right]^{-1} \left[E\{\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\eta}^{\mathrm{T}}\}\right]^{\mathrm{T}} \left[E\{-\partial \mathbf{U}_i^*(\boldsymbol{\theta}, \boldsymbol{\eta})/\partial \boldsymbol{\theta}^{\mathrm{T}}\}\right]^{-1}.$$

Acknowledgements This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (G. Y. Yi and C. Wu).

References

1. Buonaccorsi, J.P.: Measurement error in the response in the general linear model. J. Am. Stat. Assoc. **91**, 633–642 (1996)
2. Carroll, R.J., Spiegelman, C.H., Gordon, K.K., Bailey, K.K., Abbott, R.D.: On errors-in-variables for binary regression models. Biometrika **71**, 19–25 (1984)
3. Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M.: Measurement error in nonlinear models: a modern perspective, 2nd edn. Chapman and Hall/CRC, London (2006)
4. Chen, Z., Yi, G.Y., Wu, C.: Marginal methods for correlated binary data with misclassified responses. Biometrika **98**, 647–662 (2011)
5. Diggle, P.J., Heagerty, P., Liang, K.Y., Zeger, S.L.: Analysis of Longitudinal Data, 2nd edn. Oxford University Press, New York (2002)
6. Ferrara, L.A., Guida, L., Iannuzzi, R., Celentano, A., Lionello, F.: Serum cholesterol affects blood pressure regulation. J. Hum. Hypertens. **16**, 337–343 (2002)
7. Hall, P., Ma, Y.Y.: Semiparametric estimators of functional measurement error models with unknown error. J. R. Stat. Soc. Ser. B **69**, 429–446 (2007)
8. Jaquet, F., Goldstein, I.B., Shapiro, D.: Effects of age and gender on ambulatory blood pressure and heart rate. J. Hum. Hypertens. **12**, 253–257 (1998)
9. McCulloch, C.E., Searle, S.R.: Generalized, Linear, and Mixed Models. Wiley, New York (2001)
10. Neuhaus, J.M.: Bias and efficiency loss due to misclassified responses in binary regression. Biometrika **86**, 843–855 (1996)
11. Primatesta, P., Falaschetti, E., Gupta, S., Marmot, M.G., Poulter, N.R.: Association between smoking and blood pressure - evidence from the health survey for England. Hypertension **37**, 187–193 (2001)
12. Vonesh, E.F.: A note on the use of Laplace's approximation for nonlinear mixed-effects models. Biometrika **83**, 447–452 (1996)
13. Wang, N., Lin, X., Gutierrez, R.G., Carroll, R.J.: Bias analysis and SIMEX approach in generalized linear mixed measurement error models. J. Am. Stat. Assoc. **93**, 249–261 (1998)
14. White, H.: Maximum likelihood estimation of misspecified models. Econometrica **50**, 1–25 (1982)
15. Wolfinger, R.: Laplace's approximation for nonlinear mixed models. Biometrika **80**, 791–795 (1993)
16. Yi, G.Y.: A simulation-based marginal method for longitudinal data with dropout and mismeasured covariates. Biostatistics **9**, 501–512 (2008)
17. Yi, G.Y.: Measurement error in life history data. Int. J. Stat. Sci. **9**, 177–197 (2009)
18. Yi, G.Y., Cook, R.J. Errors in the measurement of covariates. In: The Encyclopedia of Biostatistics, 2nd edn., vol. 3, pp. 1741–1748. Wiley, New York (2005)
19. Yi, G.Y., Liu, W., Wu, L.: Simultaneous inference and bias analysis for longitudinal data with covariate measurement error and missing responses. Biometrics 67, 67–75 (2011)
20. Zucker, D.M.: A pseudo-partial likelihood method for semiparametric survival regression with covariate errors. J. Am. Stat. Assoc. **100**, 1264–1277 (2005)