Published in final edited form as:

Can J Stat. 2015 December ; 43(4): 498–518. doi:10.1002/cjs.11268.

Variable Selection and Inference Procedures for Marginal Analysis of Longitudinal Data with Missing Observations and Covariate Measurement Error

Grace Y. Yi,

Department of Statistics and Actuarial Science University of Waterloo, Waterloo, Ontario, Canada
N2L 3G1

Xianming Tan, and

The Methodology Center Pennsylvania State University University Park, PA16802

Runze Li

Department of Statistics and The Methodology Center, Pennsylvania State University, University Park, PA16802-2111

Summary

In contrast to extensive attention on model selection for univariate data, research on model selection for longitudinal data remains largely unexplored. This is particularly the case when data are subject to missingness and measurement error. To address this important problem, we propose marginal methods that simultaneously carry out model selection and estimation for longitudinal data with missing responses and error-prone covariates. Our method have several appealing features: the applicability is broad because the methods are developed for a unified framework with marginal generalized linear models; model assumptions are minimal in that no full distribution is required for the response process and the distribution of the mismeasured covariates is left unspecified; and the implementation is straightforward. To justify the proposed methods, we provide both theoretical properties and numerical assessments.

Keywords

Longitudinal data; Marginal analysis; Measurement error; Missing data; Model selection; Simulation-extrapolation

1 Introduction

Longitudinal studies have proven to be useful in studying changes of response over time, and have been widely conducted in applications. A variety of modeling approaches have been developed for analysis of longitudinal data. It is often of primary interest to characterize how the mean response is related to the covariates as this relationship features the average response at population level (e.g., Lin and Carroll 2006). To facilitate modeling the dependence of the mean response on covariates, marginal approaches, such as generalized estimating equations (GEE), are often used (e.g., Liang and Zeger 1986, Lin and Carroll 2001). However, the use of existing methods is impeded by challenging issues associated with longitudinal data.

It is common that longitudinal studies collect a large number of covariates in which some of them have no effect on the response. Including such covariates in modeling and inferential procedures would greatly degrade the quality of the results. Variable selection thus becomes necessary and critical for valid analysis of longitudinal data. With univariate data, there has been a large body of variable selection methods. Miller (2002) provides a comprehensive account on earlier work of variable selection, including variable selection criteria such as BIC (Schwarz 1978) and algorithms for the best subset selection. Recent developments on variable selection techniques including LASSO (Tibshirani 1996), SCAD (Fan and Li 2001) and their extensions are summarized by Fan and Li (2006). Modern regularization methods have been proposed for longitudinal data. Fu (2003) proposed the penalized marginal estimating equation with a bridge penalty for longitudinal data. Fan and Li (2004) extended the SCAD procedure for partially linear models with longitudinal data. Bondell, Krishna and Ghosh (2010) proposed simultaneous selection of the fixed and random factors using a penalized joint log likelihood for the linear mixed model. Ni, Zhang and Zhang (2010) proposed a double-penalized likelihood approach for simultaneous model selection and estimation for semiparametric mixed models.

The validity of these methods, however, relies on two important conditions: (1) data are complete and (2) variables are precisely measured. Nevertheless, these conditions are frequently violated in reality (e.g., Tsiatis, DeGruttola and Wulfsohn 1995; Wang et al. 2008). Although longitudinal studies are designed to collect data on every individual in the studies at each assessment, missing observations arise ubiquitously. With estimation being the primary focus, many authors have shown that seriously biased results can be yielded if missing data are not properly accounted for (e.g., Robins, Rotnitzky and Zhao 1995).

It is typical that measurement error on covariate is present in data collection with various reasons (Carroll et al. 2006). For instance, variables may be difficult to observe precisely for many reasons such as physical location or cost (e.g., exposure to radiation). A covariate may represent an average of a certain quantity over time (e.g. cholesterol level), and any practical way of measuring such a quantity necessarily features measurement error. It has been demonstrated that ignoring covariate measurement error in model selection procedures may give rise to seriously misleading results. Liang and Li (2009) and Ma and Li (2010) proposed the SCAD for the independent and identically distributed (i.i.d) data in the presence of measurement error in covariates. These methods cannot, however, directly apply to the setting of longitudinal data where the i.i.d assumption breaks down as longitudinal measurements are typically correlated, nor can they handle the data with missing observations.

The goal of this paper is to address these practical and important problems. We develop simultaneous variable selection and estimation procedures that can handle “imperfect” longitudinal data with missingness and measurement error. Theoretical properties are established. Although extensive research has been directed towards each of the aspects of model selection, incomplete longitudinal data and measurement error, those methods cannot be simply applied to handle those features at the same time. Simultaneously addressing these issues is more challenging because those characteristics could interactively affect inference about response parameters. Developing asymptotic results becomes more difficult for

combining all the features in the inferential procedures. To the best of our knowledge, there is no precedent work addressing these features simultaneously. Our method is developed for the marginal framework which includes a broad class of mean structures. Our method has several appealing features. Its implementation is straightforward; the method requires minimal model assumptions for the response process; and the distribution of the true covariates is left unspecified, hence protecting us from model misspecification of the response and covariate processes. Our method can be applied to a wide range of settings due to its simplicity.

The remainder is organized as follows. In Section 2 we introduce the basic model setup and notation. Section 3 presents the methodology that allows simultaneous estimation and model selection to handle longitudinal data with incomplete observations and measurement error. Asymptotic results are established in Section 4. We apply the proposed method to analyze the longitudinal data arising from the National Population Health Survey, and the results are reported in Section 5.1. To assess the performance of the proposed method, we conduct simulation studies in Sections 5.2 and 5.3. General remarks are included in the last section.

2 Notation and Framework

We consider the setting where the response variable is subject to missingness and some covariates are error-contaminated. For subject i at time point j , let Y_{ij} be the response variable that is subject to missingness, and $\mathbf{X}_{ij}$ and $\mathbf{Z}_{ij}$ be the associated covariate vectors, where the $\mathbf{X}_{ij}$ are subject to measurement error and the $\mathbf{Z}_{ij}$ are precisely measured, $i = 1, 2, \dots, n$, and $j = 1, 2, \dots, m$. Denote $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{im})^T$, $\mathbf{X}_i = (\mathbf{X}_{i1}^T, \mathbf{X}_{i2}^T, \dots, \mathbf{X}_{im}^T)^T$ and $\mathbf{Z}_i = (\mathbf{Z}_{i1}^T, \mathbf{Z}_{i2}^T, \dots, \mathbf{Z}_{im}^T)^T$.

2.1 Response Process

Let $\mu_{ij} = E(Y_{ij}|\mathbf{X}_i, \mathbf{Z}_i)$ and $v_{ij} = \text{var}(Y_{ij}|\mathbf{X}_i, \mathbf{Z}_i)$ be the conditional expectation and variance of Y_{ij} , respectively, given covariates $\mathbf{X}_i$ and $\mathbf{Z}_i$. Consider a regression model:

$$g(\mu_{ij}) = \mathbf{X}_{ij}^T \boldsymbol{\beta}_x + \mathbf{Z}_{ij}^T \boldsymbol{\beta}_z, i=1, 2, \dots, n; j=1, 2, \dots, m, \quad (1)$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_x^T, \boldsymbol{\beta}_z^T)^T$ is the vector of regression parameters, and $g(\cdot)$ is a link function. If necessary, the intercept may be included in $\boldsymbol{\beta}_z$ by adding a unit vector to covariates $\mathbf{Z}_i$. Furthermore, it is assumed that $v_{ij} = h(\mu_{ij}; \phi)$ where $h(\cdot, \cdot)$ is a given function and ϕ is the dispersion parameter that is known or may be estimated. We treat ϕ as known here with emphasis on estimation of the $\boldsymbol{\beta}$ parameter.

The model setup implicitly implies that

$$E(Y_{ij}|\mathbf{X}_i, \mathbf{Z}_i) = E(Y_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij})$$

and

$$\text{var}(Y_{ij}|\mathbf{X}_i, \mathbf{Z}_i) = \text{var}(Y_{ij}|\mathbf{X}_{ij}, \mathbf{Z}_{ij})$$

for $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$.

Given the model assumptions for the marginal mean and variance structures, it is natural to perform estimation of β using the generalized estimating equations (GEE) (Liang and Zeger 1986). For $i = 1, 2, \dots, n$, let $\mu_i = (\mu_{i1}, \dots, \mu_{im})^T$, and $D_i = \partial \mu_i^T / \partial \beta$ be the matrix of the derivatives of the mean vector μ_i with respect to β . Let $V_i = A_i^{1/2} C_i A_i^{1/2}$ be the (conditional) covariance matrix of $\mathbf{Y}_i$, given covariates $\mathbf{X}_i$ and $\mathbf{Z}_i$, where $A_i = \text{diag}(v_{ij}, j = 1, 2, \dots, m)$, $C_i = [\rho_{ijk}]$ is the correlation matrix with diagonal elements being 1, and ρ_{ijk} is the (conditional) correlation coefficient of response components Y_{ij} and Y_{ik} for $j \neq k$, given covariates $\mathbf{X}_i$ and $\mathbf{Z}_i$.

For $i = 1, 2, \dots, n$, define $\Phi_i(\beta) = D_i V_i^{-1} (\mathbf{Y}_i - \mu_i)$, and

$$\Phi(\beta) = \sum_{i=1}^n \Phi_i(\beta), \quad (2)$$

where the dependence on the parameters of the correlation matrix is suppressed in the notation. If model (1) is correctly specified, then $\Phi(\beta)$ are unbiased estimation functions for β . Under regularity conditions, solving $\Phi(\beta) = \mathbf{0}$ leads to a consistent estimator for β , where the correlation matrix C_i can be replaced by the moment estimate, or alternatively, we may use working independence matrix A_i to replace V_i (e.g., Liang and Zeger 1986).

2.2 Missing Data Process

Let R_{ij} be 1 if Y_{ij} is observed, and 0 otherwise. Let $\mathbf{R}_i = (R_{i1}, R_{i2}, \dots, R_{im})^T$ be the vector of (non-)missing data indicators, $i = 1, 2, \dots, n$. Drop-outs or monotone missing data patterns are considered here. That is, $R_{ij} = 0$ implies $R_{ij'} = 0$ for all $j' > j$. We assume that $R_{i1} = 1$ for every subject i .

To reflect the dynamic nature of the observation process over time, we assume that

$P(R_{ij}=1|\tilde{\mathbf{R}}_{ij}, \mathbf{Y}_i, \mathbf{X}_i, \mathbf{Z}_i) = P(R_{ij}=1|\tilde{\mathbf{R}}_{ij}, \mathbf{Y}_i^{(o)}, \mathbf{X}_i, \mathbf{Z}_i)$, where $\mathbf{Y}_i^{(o)}$ is the vector including observed response components, and $\mathbf{R}_{ij} = \{R_{i1}, \dots, R_{i,j-1}\}$ is the history of the missing data indicators up to (but not including) time j . This assumption implies that the missingness probability depends on the observed responses but not unobserved response components, given covariates, thus leading to a missing at random mechanism (MAR), as defined by Little and Rubin (2002). This assumption is not essential to the development here, but is merely to ensure model identifiability related to the missing data process. More discussion on this is included in Section 6.

Denote $\lambda_{ij} = P(R_{ij}=1|\tilde{\mathbf{R}}_{ij}, \mathbf{X}_i, \mathbf{Z}_i, \mathbf{Y}_i)$ and $\pi_{ij} = P(R_{ij}=1|\mathbf{X}_i, \mathbf{Z}_i, \mathbf{Y}_i)$. Note that $\pi_{ij} = \prod_{t=2}^j \lambda_{it}$. Logistic regression models are commonly used to model the drop-out process (e.g., Diggle and Kenward 1994). Namely,

$$\text{logit} \lambda_{ij} = \mathbf{u}_{ij}^T \boldsymbol{\alpha}, \quad (3)$$

where $\mathbf{u}_{ij}$ is the vector consisting of the information of the covariates $\mathbf{X}_i$ and $\mathbf{Z}_i$ and the observed responses $\mathbf{Y}_i^{(o)}$, and $\boldsymbol{\alpha}$ is the vector of regression parameters.

Let M_i be the random drop-out time for subject i , and m_i be a realization, $i = 1, 2, \dots, n$.

Define $L_i(\boldsymbol{\alpha}) = (1 - \lambda_{im_i}) \prod_{t=2}^{m_i-1} \lambda_{it}$, where λ_{it} is determined by model (3). Let

$$\mathbf{S}_i(\boldsymbol{\alpha}) = \partial \log L_i(\boldsymbol{\alpha}) / \partial \boldsymbol{\alpha} \quad (4)$$

be the vector of score functions contributed from subject i . Write $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$ and $d = \dim(\boldsymbol{\theta})$.

2.3 Measurement Error Model

Let $\mathbf{W}_{ij}$ be an observed measurement of the covariate $\mathbf{X}_{ij}$, $i = 1, 2, \dots, n$; $j = 1, 2, \dots, m$. Write $\mathbf{W}_i = (\mathbf{W}_{i1}^T, \mathbf{W}_{i2}^T, \dots, \mathbf{W}_{im}^T)^T$. Suppose the measurement error is postulated by a classical additive measurement error model. That is, conditional on covariates $(\mathbf{X}_i, \mathbf{Z}_i)$ and responses $\mathbf{Y}_i$,

$$\mathbf{W}_{ij} = \mathbf{X}_{ij} + \mathbf{e}_{ij}, j = 1, 2, \dots, m, \quad (5)$$

where the $\mathbf{e}_{ij}$ are assumed to follow $N(\mathbf{0}, \boldsymbol{\Sigma}_e)$ with covariance matrix $\boldsymbol{\Sigma}_e$ (e.g., Carroll et al. 2006).

In the presence of measurement error, one often needs additional data sources, such as a validation subsample or repeated surrogate measurements, to estimate the parameters associated with the measurement error model (Carroll et al. 2006). If none of such data are available, one may conduct sensitivity analyses based on background information about the measurement process to assess the impact of different degrees of measurement error on estimation of $\boldsymbol{\beta}$ (e.g., Yi and Lawless 2007, Yi 2008). To focus the discussion on the parameter $\boldsymbol{\beta}$ of interest, here we treat error distribution parameters as known; extensions to accommodating estimation of such parameters are presented in Section 6.

3 Methodology

In this section we describe an algorithm that can simultaneously conduct model selection and estimation as well as adjust for missingness and measurement error effects. In particular, we use the inverse probability weighted GEE method to accommodate effects of response incompleteness, and employ the simulation-extrapolation (SIMEX) procedure (Cook and Stefanski 1994) to reduce the bias induced from covariate measurement error. The algorithm consists of the following four steps.

Step 1. Simulation

Let K be a given positive integer (say, $K = 500$) and $\Psi = \{\psi_1, \psi_2, \dots, \psi_M\}$ be a sequence of M (say, $M = 10$) non-negative numbers taken from $[0, \psi_M]$ with $\psi_1 = 0$ (say, $\psi_M = 1$). For $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$, generate $\mathbf{e}_{ijk} \sim N(\mathbf{0}, \Sigma_e)$ for $k = 1, 2, \dots, K$, and set

$$\mathbf{W}_{ij}(k, \psi) = \mathbf{W}_{ij} + \sqrt{\psi} \mathbf{e}_{ijk} \text{ for } \psi \in \Psi.$$

Step 2. Estimation

Let $\mathbf{I}_i = \text{diag}\{I(R_{ij} = 1)/\pi_{ij}, j = 1, 2, \dots, m\}$ be the weight matrix accommodating missingness, where $I(\cdot)$ is the indicator function. Define

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

and

$$\mathbf{Q}_i(\boldsymbol{\theta}) = (\mathbf{S}_i^T(\boldsymbol{\alpha}), \mathbf{U}_i^T(\boldsymbol{\theta}))^T,$$

where the $\mathbf{S}_i(\boldsymbol{\alpha})$ are determined by (4).

Let $\mathbf{Q}_i^*(\boldsymbol{\theta}; k, \psi)$ be $\mathbf{Q}_i(\boldsymbol{\theta})$ with $\mathbf{X}_{ij}$ replaced with the simulated covariates $\mathbf{W}_{ij}(k, \psi)$, then solve

$$\sum_{i=1}^n \mathbf{Q}_i^*(\boldsymbol{\theta}; k, \psi) = \mathbf{0}$$

for $\boldsymbol{\theta}$, and let $\hat{\boldsymbol{\theta}}(k, \psi)$ denote the resulting solution. Let

$$\hat{\Gamma}(k, \psi) = \sum_{i=1}^n \{\partial \mathbf{Q}_i^*(\boldsymbol{\theta}; k, \psi) / \partial \boldsymbol{\theta}\} |_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(k, \psi)}, \text{ and}$$

$\hat{\Sigma}(k, \psi) = \sum_{i=1}^n \{\mathbf{Q}_i^*(\boldsymbol{\theta}; k, \psi) \mathbf{Q}_i^{*T}(\boldsymbol{\theta}; k, \psi)\} |_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(k, \psi)}$ then the covariance matrix of $\hat{\boldsymbol{\theta}}(k, \psi)$ is estimated by

$$\hat{\Omega}(k, \psi) = n \cdot \left[\{\hat{\Gamma}(k, \psi)\}^{-1} \cdot \hat{\Sigma}(k, \psi) \cdot \{\hat{\Gamma}(k, \psi)\}^{-1T} \right] |_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(k, \psi)}.$$

Consequently, we calculate

$$\hat{\boldsymbol{\theta}}(\psi) = K^{-1} \sum_{k=1}^K \hat{\boldsymbol{\theta}}(k, \psi),$$

$$\hat{\mathbf{S}}(\psi) = (K - 1)^{-1} \sum_{k=1}^K \{\hat{\boldsymbol{\theta}}(k, \psi) - \hat{\boldsymbol{\theta}}(\psi)\}^T \{\hat{\boldsymbol{\theta}}(k, \psi) - \hat{\boldsymbol{\theta}}(\psi)\},$$

$$\hat{\Omega}(\psi) = K^{-1} \sum_{k=1}^K \hat{\Omega}(k, \psi), \text{ and } \hat{\Sigma}(\psi) = \hat{\Omega}(\psi) - \hat{\mathbf{S}}(\psi).$$

Step 3. Extrapolation

For $r = 1, 2, \dots, d$, let $\theta_r(\psi)$ be the r th element of $\hat{\boldsymbol{\theta}}(\psi)$. Then for each r fit a regression model to the sequences $\{(\psi, \theta_r(\psi)) : \psi \in \Psi\}$ and extrapolate it to $\psi = -1$. Let $\tilde{\theta}_r = \theta_r(-1)$, and $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \dots, \tilde{\theta}_d)^T$.

Next, apply the Cholesky decomposition for each $\hat{\Sigma}(\psi)$: $\hat{\Sigma}(\psi) = \mathbf{L}(\psi)\mathbf{D}(\psi)\mathbf{L}^T(\psi)$, where $\mathbf{L}(\psi)$ is a lower triangular matrix with its diagonal entries being 1, and $\mathbf{D}(\psi)$ is a diagonal matrix. Then we elementwisely fit a regression model to each of the sequences $\{(\psi, \mathbf{L}(\psi)) : \psi \in \Psi\}$ and $\{(\psi, \mathbf{D}(\psi)) : \psi \in \Psi\}$, respectively, and extrapolate it to $\psi = -1$. Let $\hat{\mathbf{L}}(-1)$ and $\hat{\mathbf{D}}(-1)$ respectively denote the lower triangular and diagonal matrices with the elements corresponding to the extrapolated values, and define

$$\tilde{\Sigma} = \hat{\mathbf{L}}(-1)\hat{\mathbf{D}}(-1)\hat{\mathbf{L}}^T(-1)$$

This procedure guarantees the positive definiteness of $\tilde{\Sigma}$.

Step 4. Variable Selection

We define a *quadratic loss function*:

$$\ell(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^T \mathbf{V}_n(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \quad (6)$$

where $\mathbf{V}_n$ is a pre-specified weight matrix.

The choice of matrix $\mathbf{V}_n$ is not unique. In principle, any positive definite matrix can be considered as a weight matrix in forming (6). But in application, certain matrices are more suitable than others. When $\tilde{\Sigma}$ is not singular, we set $\mathbf{V}_n$ to be $\tilde{\Sigma}^{-1}$ in the formulation of (6). If $\tilde{\Sigma}$ is near singular, we consider a ridge-type weight matrix by setting $\mathbf{V}_n = (\tilde{\Sigma} + \tau_n I_d)^{-1}$, where τ_n is a ridge parameter, and I_d is a $d \times d$ unit matrix.

The quadratic loss function is a second-order approximation to the negative of the true log-likelihood function, and plays a similar role to that of the log-likelihood function for the penalized likelihood method. Given a tuning parameter λ , let $p_\lambda(\cdot)$ be a penalty function with tuning parameter λ . Define a *penalized quadratic loss function* to be

$$\ell_p(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - n \sum_{j=1}^d p_\lambda(|\theta_j|).$$

Motivated by the one-step sparse maximum likelihood estimate proposed by Zou and Li (2008), we use a weighted L_1 -penalty

$$p_{\lambda}(|\theta_j|) = p'_{\lambda}(|\tilde{\theta}_j|)|\theta_j|,$$

where $\tilde{\theta}_j$ is the j th element of $\tilde{\boldsymbol{\theta}}$, and $p'_{\lambda}(u)$ is the first order derivative of the penalty function $p_{\lambda}(u)$. Thus, the penalized quadratic loss function $\ell_p(\boldsymbol{\theta})$ is convex, and the resulting minimization procedure can be carried out easily by employing existing algorithms, such as the LARS algorithm (Efron et al. 2004). With a proper choice of the tuning parameter, the minimizer of $\ell_p(\boldsymbol{\theta})$ with a weighted L_1 penalty automatically has some zero components, hence achieving the purpose of variable selection.

In our following numerical studies, we will focus on the LASSO penalty (Tibshirani 1996), $p_{\lambda}(|\theta_j|) = \lambda|\theta_j|$; and the SCAD penalty whose first order derivative is defined to be

$$p'_{\lambda}(u) = \lambda \{I(|u| \leq \lambda) + \frac{(a\lambda - |u|)_+}{(a-1)\lambda} \cdot I(|u| > \lambda)\} \text{sign}(u),$$

where $\text{sign}(u)$ is the sign function of u , and $a = 3.7$ (Fan and Li 2001).

As demonstrated in the usual context without missing data nor measurement error, it is critical to choose a suitable value of tuning parameter λ in order to achieve satisfactory performance of the selection procedure. Now we describe an algorithm for choosing an optimal tuning parameter within a given set of candidates. To emphasize the dependence of λ , we denote $\hat{\boldsymbol{\theta}}(\lambda) = \arg \min_{\boldsymbol{\theta}} \ell_p(\boldsymbol{\theta})$, and let df_{λ} be the number of nonzero elements in $\hat{\boldsymbol{\theta}}(\lambda)$. Define

$$BIC(\lambda) = -2\ell(\hat{\boldsymbol{\theta}}(\lambda)) + 2\log n \cdot df_{\lambda}.$$

Then, the optimal tuning parameter λ^* is chosen as the one that has the smallest $BIC(\lambda)$, and the corresponding estimate $\hat{\boldsymbol{\theta}}(\lambda^*)$, denoted $\hat{\boldsymbol{\theta}}$, is taken as the estimate of parameter $\boldsymbol{\theta}$. The BIC tuning parameter selector can select a proper tuning parameter to achieve oracle property in penalized least squares for linear models (Wang, Li and Tsai 2007) and for penalized likelihood for generalized linear models (Zhang, Li and Tsai 2010).

Conceptually, the proposed method has the appeal of robustness; it requires minimal model assumptions for the response process for which only the marginal mean structure needs to be modulated; and moreover, the distribution of the true covariates $\mathbf{X}_i$ is left unspecified. The proposed method extends the ordinary SIMEX methods by including additional steps of missingness correction and model selection. Our approach is ready to implement but requires specification of values of K , M and ψ_M . In principle, specifying large values of K and M can help reduce Monte Carlo sampling error induced in the simulation step. Another source of uncertainty comes from choosing a suitable regression function in the extrapolation step. Experience shows that the proposed method often works reasonably well by specifying ψ_M to be 1 or 2 and taking a quadratic regression function for the extrapolation step.

4 Asymptotic Results

To establish asymptotic results for the penalized quadratic loss estimator $\hat{\boldsymbol{\theta}}$, certain conditions are required. Since the penalized quadratic loss function $\ell_p(\boldsymbol{\theta})$ involves a weight matrix $\mathbf{V}_n$, a penalty function $p_\lambda(\cdot)$ and the SIMEX estimator $\tilde{\boldsymbol{\theta}}$, required conditions should pertain to these quantities. Specifically, in addition to standard conditions, the following assumptions are imposed for establishing the asymptotic properties of $\hat{\boldsymbol{\theta}}$:

- A. As $n \rightarrow \infty$, $n^{-1}\mathbf{V}_n \xrightarrow{p} \mathbf{V}$, where $\mathbf{V}$ is a positive definite matrix;
- B. As $n \rightarrow \infty$, $\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a positive definite matrix;
- C. Define

$$a_n = \max\{|p'_{\lambda_n}(|\theta_{j0}|): \theta_{j0} \neq 0\}, \text{ and } b_n = \max\{|p''_{\lambda_n}(|\theta_{j0}|): \theta_{j0} \neq 0\},$$

where $\boldsymbol{\theta}_0 = (\theta_{10}, \dots, \theta_{d0})^T$ is the true value of $\boldsymbol{\theta}$, and the dependence of λ on sample size n is explicitly indicated by λ_n . Assume that both a_n and b_n tend to 0 as $n \rightarrow \infty$.

Assumption A ensures proper behavior of the weight matrix, and it is easily satisfied by a suitable choice of $\mathbf{V}_n$. Assumption C is a standard condition for model selection and is assumed by many authors (e.g., Fan and Li 2001, Zou and Li 2008). Assumption B requires the SIMEX estimator $\tilde{\boldsymbol{\theta}}$ to have an asymptotic normal distribution. This assumption is not as stringent as it appears. In fact, Carroll et al. (1996) established the asymptotic normality for the SIMEX estimator under the condition that the extrapolation function is known.

Using the techniques by Fan and Li (2001), we can show that with probability tending to 1, estimator $\hat{\boldsymbol{\theta}}$ satisfies

$$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(n^{-1/2} + a_n).$$

This indicates how the convergence rate of estimator $\hat{\boldsymbol{\theta}}$ depends on the tuning parameter λ_n as well as the penalty function. To ensure the $\sqrt{n}$ -consistency of the estimator, one needs to make $a_n = O_p(n^{-1/2})$, often by specifying a very small tuning parameter λ_n .

We next derive the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ and establish the oracle property for $\hat{\boldsymbol{\theta}}$.

Suppose some components of $\boldsymbol{\theta}_0$ are zero, and we write $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{I_0}^T, \boldsymbol{\theta}_{I_{I_0}}^T)^T$ so that $\boldsymbol{\theta}_{I_0}$ contains all non-zero elements and $\boldsymbol{\theta}_{I_{I_0}}$ includes all zero elements. For ease of exposition, we assume $\boldsymbol{\theta}_{I_0}$ contains the first d_1 elements and $\boldsymbol{\theta}_{I_{I_0}}$ contains the last d_2 components with $d_1 + d_2 = d$. We partition $\mathbf{V}$ into 2 by 2 blocks with $\mathbf{V}_{kl}$ being a $d_k \times d_l$ matrix for $k, l = 1, 2$, and define $\mathbf{V}_1 = [\mathbf{V}_{11}, \mathbf{V}_{12}]$.

Theorem 1

Assume the preceding regularity conditions. If

$$\liminf_{n \rightarrow \infty} \liminf_{u \rightarrow 0^+} \sqrt{n} p'_{\lambda_n}(u) \rightarrow \infty,$$

then with probability tending to 1, any $\sqrt{n}$ -convergent estimator $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_I^T, \hat{\boldsymbol{\theta}}_{II}^T)^T$ obtained from the foregoing algorithm must have the properties that

1. $\hat{\boldsymbol{\theta}}_{II} = \mathbf{0}$;
2. $\sqrt{n}(\hat{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_{I0}) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_{11}^{-1} \mathbf{V}_1 \boldsymbol{\Sigma} \mathbf{V}_1^T \mathbf{V}_{11}^{-1})$, as $n \rightarrow \infty$.

The theorem suggests that the proposed procedure can correctly identify significant variables and remove irrelevant variables. Furthermore, it yields that the proposed estimator $\hat{\boldsymbol{\theta}}$ is often more efficient than the SIMEX estimator $\tilde{\boldsymbol{\theta}}$, which is particularly the case when the limit matrix $\mathbf{V}$ is identical to $\boldsymbol{\Sigma}^{-1}$. In this instance, $\mathbf{V}_1 \boldsymbol{\Sigma} = [I_{d_1}, \mathbf{0}_{d_1 d_2}]$, and hence, $\mathbf{V}_{11}^{-1} \mathbf{V}_1 \boldsymbol{\Sigma} \mathbf{V}_1^T \mathbf{V}_{11}^{-1} = \mathbf{V}_{11}^{-1}$ where $\mathbf{0}_{d_1 d_2}$ is a $d_1 \times d_2$ zero matrix. If we partition $\boldsymbol{\Sigma}$ in the same way as for $\mathbf{V}$, and let $\boldsymbol{\Sigma}_{kl}$ denote the corresponding block matrix for $k, l = 1, 2$, then

$$\mathbf{V}_{11}^{-1} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}.$$

Therefore, $\boldsymbol{\Sigma}_{11} - \mathbf{V}_{11}^{-1}$ is positive definite, implying that $\hat{\boldsymbol{\theta}}_I$ is more efficient than $\tilde{\boldsymbol{\theta}}_I$, where $\tilde{\boldsymbol{\theta}}_I$ is the SIMEX estimator of $\boldsymbol{\theta}_I$.

We conclude this section with a simple but informative example to illustrate the rationale of constructing the quadratic loss function (6) for variable selection.

Example

Suppose that $\{(\mathbf{X}_i, Y_i) : i = 1, \dots, n\}$ is a random sample chosen from the linear model

$$Y_i = \mathbf{X}_i^T \boldsymbol{\theta} + \varepsilon_i$$

where $\boldsymbol{\theta}$ is the regression parameter vector of dimension d , and the error term $\varepsilon_i \sim N(0, \sigma^2)$. Without loss of generality, assume that $\sigma^2 = 1$ for simplicity.

Let $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, and $\mathbf{X} = (\mathbf{X}_1^T, \dots, \mathbf{X}_n^T)^T$ be the $n \times d$ matrix of the covariates. If using the least squares method to estimate $\boldsymbol{\theta}$, then the resulting estimator is $\tilde{\boldsymbol{\theta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$, which has the properties

$$n \text{cov}(\tilde{\boldsymbol{\theta}}) = \left(\frac{1}{n} \mathbf{X}^T \mathbf{X} \right)^{-1} \xrightarrow{p} \{E(\mathbf{X}_i \mathbf{X}_i^T)\}^{-1}$$

and

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \{E(\mathbf{X}_i \mathbf{X}_i^T)\}^{-1}),$$

where θ_0 is the true value of θ .

Note that

$$\frac{1}{2}\|Y - X\theta\|^2 = \frac{1}{2}\|y - X^T\tilde{\theta}\|^2 + \frac{1}{2}(\theta - \tilde{\theta})^T(X^T X)(\theta - \tilde{\theta}). \quad (7)$$

Since the first term $\|y - X^T\tilde{\theta}\|^2$ on the right side of (7) does not involve parameter θ , to perform variable selection, it is natural to approximate the left side of (7) using the second term of the right side of (7), which is the quadratic loss function defined in (6) with the weight matrix taken as $V_n = \{\text{cov}(\tilde{\theta})\}^{-1} = X^T X$.

We then define the penalized quadratic loss function and follow the Step 4 procedure in Section 3 to perform variable selection. Let $\hat{\theta}$ denote the resulting estimator. Then this estimator has the properties stated in Theorem 1. In fact, according to Fan and Li (2001), the oracle estimator is $\hat{\theta}_H = \mathbf{0}$ and $\hat{\theta}_I = (X_I^T X_I)^{-1} X_I^T Y$ with asymptotic normality:

$\sqrt{n}(\hat{\theta}_I - \theta_{I0}) \rightarrow N(0, V_{11}^{-1})$ where X_I is the $n \times d_1$ submatrix of X corresponding to θ_I in the linear model, and V_{11} is the $d_1 \times d_1$ submatrix of $E(X_i X_i^T)$ corresponding to θ_I .

5 Numerical Studies

5.1 Analysis of NPHS Data

The National Population Health Survey (NPHS) is a longitudinal study that was conducted every second year, beginning in 1994/1995. The survey questionnaire is designed to collect health information and related socio-demographic information by following a group of Canadian household residents. The questions for the NPHS include many aspects of in-depth health information such as health status, use of health services, chronic conditions and activity restrictions. Social background questions, including age, sex and income level, are contained in the questionnaire. A primary objective of the NPHS study is to understand how population health is influenced by multiple covariates or risk factors. For instance, distress is an important variable associated with the health profile for an individual. It is measured by Distress Scale - K6 in the NPHS data set, ranging from MIN=0 to MAX=24 with a higher value indicating more severe distress. A binary variable is created to feature the distress status with 0 indicating distress free, and 1 otherwise.

We analyze the data with the first 5 cycles that were carried out in 1994/1995 (Cycle 1), 1996/1997 (Cycle 2), 1998/1999 (Cycle 3), 2000/2001 (Cycle 4), and 2002/2003 (Cycle 5), respectively. The data set contains the measurements for 1089 subjects.

Covariates include: (1) Body Mass Index (BMI); (2) health status, measured by the Health Utilities Index Mark (HUI) from eight attributes: vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain and discomfort. HUI score ranges from -0.36 to 1.00, with 1.00 standing for being completely healthy and -0.36 representing a lowest level of health state; (3) household income (INC), measured by the provincial level of household income. INC ranges from 1 to 10, where 1 denotes household income in a poorest province,

while 10 denotes household income in a riches province; (4) education, characterized by the highest education level for a subject. Dummy variables are used to represent education level, with “Lower than Middle School” as a baseline, EDU1 representing “Middle School” and EDU2 corresponding to “College Level”; (5) marital status, indicated by a binary variable MA, with value 1 for the status of being married or living common-law, and 0 for other situations including being widowed, separated, divorced or single (never married); and (6) age.

Let Y_{ij} be the distress status for subject i at Cycle j , and μ_{ij} be the mean of Y_{ij} conditional on the covariates, $i = 1, \dots, 1089$; $j = 1, \dots, 5$. The responses and covariates are postulated by the logistic regression model

$$\text{logit} \mu_{ij} = \beta_0 + \beta_{x1} X_{ij1} + \beta_{x2} X_{ij2} + \beta_{x3} X_{ij3} + \beta_{z1} Z_{ij1} + \beta_{z2} Z_{ij2} + \beta_{z3} Z_{ij3} + \beta_{z4} Z_{ij4},$$

where for subject i at Cycle j , X_{ij1} is the standardized body mass index, X_{ij2} is the standardized Health utility index, X_{ij3} is the standardized provincial income rank, Z_{ij1} is the marriage status, Z_{ij2} is the high school level education (0-No; 1-Yes), Z_{ij3} is the college level education (0-No; 1-Yes), and z_{ij4} is age.

Among all the covariates, the Z variables are precisely measured, but the X covariates, i.e., BMI, HUI, and INC, are subject to measurement error. Let $\mathbf{X}_{ij} = (X_{ij1}, X_{ij2}, X_{ij3})^T$ denote the vector of true covariate values, and $\mathbf{W}_{ij} = (W_{ij1}, W_{ij2}, W_{ij3})^T$ be the vector of the observed measurements. Assume the error model is given by (5) with covariance matrix $\Sigma_e = \sigma_e^2 \mathbf{C}$, where $\mathbf{C}$ is the correlation matrix with diagonal elements 1. As there is no additional source, such as a validation sample, to be used to estimate the covariance matrix, we conduct sensitivity analyses to understand error effects on model selection. In particular, we take $\Sigma_e = \sigma_e^2 \mathbf{C}$ where σ_e^2 is specified as $(0.15 \times a)^2$, and we take

$$\mathbf{C} = \begin{pmatrix} 1.0 & 0.001 \times c & 0.009 \times c \\ 0.001 \times c & 1.0 & 0.200 \times c \\ 0.009 \times c & 0.200 \times c & 1.0 \end{pmatrix}$$

for different values of $c = 0, 1/4, 1, 4$ and $a = 1/4, 1/2, 1, 2$. When $c = a = 1$, the Σ_e is the empirical estimate obtained from the data.

In this data set, a portion of subjects have dropped out of the study. Specifically, dropout percentages are 14.1%, 24.7%, 37.4%, and 46.6% at Cycles 2, 3, 4 and 5, respectively. The drop-out process is characterized by the logistic regression model

$$\text{logit} \lambda_{ij} = \alpha_0 + \alpha_y Y_{i,j-1} + \alpha_{x1} X_{i,j-1,1} + \alpha_{x2} X_{i,j-1,2} + \alpha_{x3} X_{i,j-1,3} + \alpha_{z1} Z_{i,j-1,1} + \alpha_{z2} Z_{i,j-1,2} + \alpha_{z3} Z_{i,j-1,3} + \alpha_{z4} Z_{i,j-1,4},$$

where $\alpha_0, \alpha_y, \alpha_{x1}, \alpha_{x2}, \alpha_{x3}, \alpha_{z1}, \alpha_{z2}, \alpha_{z3}$ and α_{z4} are regressions parameters, and the dependence of drop-out probability at Cycle j on the response at the previous cycle features a possible missing at random mechanism.

We analyze the data with the selection and inference method described in Section 3. In particular, we use $K = 200$ and $\Psi = \{0, 0.25, 0.50, 0.75, 1.0, 1.25, 1.50, 1.75, 2.0\}$ which equally divide $[0, 2]$ into 8 subintervals. A quadratic regression function is used in the extrapolation step. We consider both LASSO type and SCAD type penalty functions when applying the proposed method; in contrast, we report on the estimation results obtained from the SIMEX method alone (i.e., the first three steps in Section 3). Table 1 displays the analysis results for the case with $a = c = 1$.

Although the selection methods based on the SCAD and LASSO penalties result in different estimates, they do select the same covariates, HUI X_{ij2} and age Z_{ij4} , for the response model, and the associated standard errors are quite close. The results suggest that health utility index has a positive effect on reducing distress: individuals having higher HUI are less likely to have distress. As subjects become older, they are less likely to have distress. The results obtained from using the SIMEX method alone agree fairly well with the selection results except for the estimate of the marital effect, and which is indicated by insignificance of the estimates for those unselected variables.

Regarding selecting variables for the drop-out process, both the SCAD and LASSO penalties yield the same results, indicating that none of the covariates are significant. Both methods suggest that the dependence on the previous response measurement is not significant, and this implies that data are missing completely at random. Again, the results obtained from the SIMEX method alone are in reasonable agreement with those findings except that there is slight evidence for significance of EDU1.

In Figure 1, we further display the sensitivity analysis results for the HUI covariate. The dark lines represent the estimates; the grey lines denote the corresponding upper and lower bounds of 95% confidence intervals, respectively. The results for the age covariate are not reported to save space, but the pattern is quite similar to that of HUI with less variable magnitudes.

Finally, we comment that measurement error model (5) is used to describe the error involved in the covariate income although this variable is categorical in the data analysis. Since the income variable X_{ij3} is defined by categorizing a latent continuous variable into 10 classes, we have conducted additional simulation studies to assess the impact of using error model (5) for X_{ij3} on inference results. Our simulation studies (not reported here) offer us certain comfort that assuming error model (5) for the rank-type covariate, income, does not noticeably affect model selection and estimation of the covariate effects, although the variance estimate of the income variable seems somewhat conservative.

5.2 Simulation Study: Performance of the Proposed Method

We conduct simulation studies to assess the performance of the proposed methods under various circumstances. The implementation code is available from the second author upon request. We consider the configuration of $m = 5$ and $d_1 = d_2 = 5$. Two sample sizes with $n = 200$ and 500 are considered. One thousand simulations are performed for each parameter setting.

Error-contaminated covariates $\mathbf{X}_{ij}$ are generated from a d_1 -multivariate normal distribution $N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$, and error-free covariates $\mathbf{Z}_{ij}$ are independently generated from a d_2 -multivariate normal distribution $N(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$. Surrogates $\mathbf{W}_{ij}$ are generated by adding an error term to $\mathbf{X}_{ij}$: $\mathbf{W}_{ij} = \mathbf{X}_{ij} + \mathbf{e}_{ij}$ with $\mathbf{e}_{ij} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_e)$. We set $\boldsymbol{\mu}_x = 0.5 \cdot \mathbf{1}_{d_1}$ and $\boldsymbol{\mu}_z = -0.2 \cdot \mathbf{1}_{d_2}$, where $\mathbf{1}_r$ is the $r \times 1$ unit vector. $\boldsymbol{\Sigma}_x$ is taken as a $d_1 \times d_1$ matrix with (i, j) element $\sigma_{xij} = \sigma_x^2 \cdot \rho_x^{|i-j|}$, $\boldsymbol{\Sigma}_z$ is taken as a $d_2 \times d_2$ matrix with (i, j) element $\sigma_{zij} = \sigma_z^2 \cdot \rho_z^{|i-j|}$, and $\boldsymbol{\Sigma}_e$ is taken as a $d_1 \times d_1$ matrix with (i, j) element being $\sigma_{eij} = \sigma_e^2 \cdot \rho_e^{|i-j|}$, where we set $\sigma_x^2 = \sigma_z^2 = 1.0$, and $\rho_x = \rho_z = 0.5$. We set $\sigma_e = 0.15, 0.50, 0.70$, and $\rho_e = 0, 0.1, 0.5$ to feature various scenarios of measurement error.

Binary response vectors $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{im})^T$ are generated from the joint density function

$$f(y_{i1}, y_{i2}, \dots, y_{im}) = \prod_{j=1}^m \mu_{ij}^{y_{ij}} (1 - \mu_{ij})^{1-y_{ij}} \left\{ 1 + \sum_{j < k} \rho_{ijk} \frac{y_{ij} - \mu_{ij}}{\sqrt{v_{ijj}}} \cdot \frac{y_{ik} - \mu_{ik}}{\sqrt{v_{ikk}}} \right\} \quad (8)$$

where ρ_{ijk} is the correlation coefficient of Y_{ij} and $Y_{ik} (j < k)$, given by

$\rho_{ijk} = (\mu_{ijk} - \mu_{ij}\mu_{ik}) / \sqrt{v_{ijj}v_{ikk}}$, the mean $\mu_{ij} = E(Y_{ij}|\mathbf{X}_i, \mathbf{Z}_i)$ is modeled as

$$\text{logit} \mu_{ij} = \beta_0 + \sum_{k=1}^{d_1} \beta_{xk} X_{ijk} + \sum_{k=1}^{d_2} \beta_{zk} Z_{ijk}, \quad (9)$$

the variance $v_{ij} = \mu_{ij}(1 - \mu_{ij})$, and $\mu_{ijk} = P(Y_{ij} = 1, Y_{ik} = 1 | \mathbf{X}_i, \mathbf{Z}_i)$ is determined by the marginal probabilities μ_{ij} and μ_{ik} , together with the odds ratio ψ_{ijk} between Y_{ij} and Y_{ik} which is modeled as

$$\log \psi_{ijk} = \phi. \quad (10)$$

Set $\beta_0 = -1.6$, $\boldsymbol{\beta}_x = (3.0, 0, 0, 1.5, 0)^T$, $\boldsymbol{\beta}_z = (0, 0, -1.5, 0, 0)^T$, and $\phi = 0.1$.

For the missing data process, we generate monotone missing data indicator $\mathbf{R}_i = (R_{i1}, R_{i2}, \dots, R_{im})^T$ according to the conditional probability

$\lambda_{ij} = P(R_{ij} = 1 | R_{i,j-1} = \dots = R_{i1} = 1, \mathbf{Y}_i^{(o)}, \mathbf{X}_i, \mathbf{Z}_i)$. Specifically, we consider the model

$$\text{logit} \lambda_{ij} = \alpha_0 + \alpha_1 Y_{i,j-1} + \sum_{k=1}^{d_1} \alpha_{xk} X_{ij-1,k} + \sum_{k=1}^{d_2} \alpha_{zk} Z_{ij-1,k}, \quad j=2, 3, \dots, m,$$

where parameters are set as $\alpha_0 = 0.5$, $\alpha_1 = 1.5$, $\boldsymbol{\alpha}_x = (0, -1.5, 0, 0, 0)^T$, and $\boldsymbol{\alpha}_z = (3.0, 0, 0, 0, 0)^T$.

In performing the SIMEX approach, we use $K = 200$ and $\Psi = \{0, 0.25, 0.50, 0.75, 1.0, 1.25, 1.50, 1.75, 2.0\}$ which equally divide $[0, 2]$ into 8 subintervals, and a quadratic regression function is used in the extrapolation step.

To evaluate the performance of model selection, we compute the square root mean squared error (RMSE) for both the reduced model (i.e., the model with selected variables only) and the full model using the SIMEX method, and then calculate the ratio of these two RMSE values. The median of the ratios from 1000 simulations is reported under the column “MRMSE”. A smaller value indicates a better performance in terms of parameter estimation. We also calculate the *specificity* and *sensitivity* for each simulation. Specificity is defined as the proportion of zero coefficients that were correctly estimated to be zeros; sensitivity is the proportion of nonzero coefficients that are estimated to be non-zeros. The average of both indices over 1000 simulations are listed under the columns with corresponding headings in the tables. In addition, *under fit* is defined as the proportion of excluding any significant coefficients in 1000 replications, whereas *exact fit* is the proportion of selecting the exact sub-model. The proportions of including all significant variables and some (1, 2, or 3) non-significant variables are also reported in the columns under *over fit*. We report on the results in Tables 2 and 3.

As expected, the model selection performance, in terms of sensitivity, specificity and the correct fit rate, would decay as the degree of measurement error becomes more substantial. The performance of the model selection is also affected by the sample size. We note, in good agreement with the theoretical results, that a larger sample size tends to yield better results. This trend is observed in all situations, regardless that measurement errors are correlated or uncorrelated.

Next, we evaluate how measurement error and missingness may differently affect the performance of model selection. The results are displayed in Table 4 where Setting I refers to scenarios with measurement error only while Setting II corresponds to situations with missingness only. It is seen that although the results for Setting I show the same trends as those observed before, these results are fairly insensitive to the degree of measurement error, and are quite satisfactory. On the other hand, the results for Setting II demonstrate that missingness has more noticeable impact on model selection than measurement error does for the configurations we consider.

5.3 Simulation Study: Comparison of the Proposed Method with a Two-Stage Method

When the dimension of covariates is not very large, one may pursue an alternative method, as pointed out by a referee. This method consists of two steps: in the first step, we apply the SIMEX algorithm to the full dataset with all covariates included; in the second step, remove those covariates with p-value smaller than a threshold, say 0.05, and then re-apply the SIMEX algorithm to the data with the remaining covariates.

We run simulation studies to compare the performance of the proposed method and this two-step algorithm. Using the same data generation procedure as in Section 5.2, we generate data with sample size $n = 200$ and run simulations 300 times for each of the three model configurations: (A) $\rho_e = 0.5$, $\sigma_e = 0.24$; (B) $\rho_e = 0.5$, $\sigma_e = 0.33$; and (C) $\rho_e = 0.5$, $\sigma_e = 0.41$ with both drop-out and measurement errors involved.

We compare the model selection performance between the SIMEX approach and the 2-stage approach. The simulation results are reported in Table 5. In terms of RMSE, the 2-stage

approach seems to outperform the SIMEX approach; it tends to produce smaller RMSE than the SIMEX approach does. But the SIMEX approach performs better than the 2-stage procedure in terms of the rate of exact correct fit. The rate of exact correct fit of the 2-stage approach can be about 30% lower than that of the proposed SIMEX approach. Overall, both methods work generally well.

6 Discussion

Missing observations and measurement error arise frequently in longitudinal studies, and present considerable challenges in inference. In this paper, we develop simultaneous procedures for model selection and estimation for longitudinal data with these features. The development is cast in a unified framework of marginal generalized linear models; this is attractive in that the full distribution of the response process is not needed. In the case where both missing data and measurement error exist, we explore a simulation-based approach that applies for many applications and establish theoretical properties of the resulting estimator. In contrast to the naive analysis of ignoring measurement error or missingness, numerical studies indicate reasonable performance of the proposed method.

In terms of model selection consistency, Leeb and Pötscher (2008) raised the superefficiency issue of sparse estimators under the linear model framework, and implied that a model selection consistent procedure must suffer from the lack of uniform consistency. As pointed out by a referee, in spite of this aspect, it is debatable whether or not model selection consistency is a useful criterion. The Akaike information criterion (AIC) (Akaike 1974) and the Bayesian information criterion (BIC) (Schwarz 1978) are perhaps the most popular classical variable selection criteria. Model selection criteria are commonly classified into two categories: an *efficient* selection, such as the AIC; and a *consistent* selection, such as the BIC. The efficiency criterion selects the model so that the average squared error is asymptotically equivalent to the minimum offered by the candidate models which are assumed to approximate the true model. The consistency criterion identifies the true model with a probability approaching to 1 as the sample size increases under the assumption that a set of candidate models contains the true model. A detailed discussion on efficiency and consistency is provided by Shao (1997).

Generally speaking, an efficient model selection procedure, such as the AIC, leads to an estimate possessing uniform consistency, while a consistent selection criterion (e.g., the BIC) yields an estimate which lacks uniform consistency. The oracle property established in the paper implies that our proposed variable selection procedure is a consistent model selection procedure, and thus, suffers from the lack of uniform consistency. The tuning parameter λ in the penalized quadratic loss function controls the model complexity. It is important to set a suitable range for the tuning parameter so that the selected λ is located around the middle of the range in the implementation of the BIC tuning parameter selector; otherwise, significant risk inflation may arise as observed by Leeb and Pötscher (2008).

In the algorithm described in Section 3, the measurement error covariance matrix Σ_e is assumed to be known. This method is particularly useful for many practical settings. Often, no additional data sources are available to determine the parameters associated with the

measurement error model, and one must conduct sensitivity analyses to evaluate the impact on inference of varying degrees of measurement error. In situations where the parameters in covariance matrix Σ_e can be estimated from additional data, such as a validation subsample or repeated surrogate measurements, the proposed methods can be readily adapted by modifying the simulation step as suggested by Carroll et al. (2006).

In the formulation of the methods, we use the inverse weighting scheme to adjust for bias induced by missing observations, and model the missing data indicators with widely used logistic regression under the MAR assumption. The MAR assumption allows us to express the missingness probability as a function of the observed responses, together with covariates. This assumption is commonly adopted in marginal analysis to ensure model identifiability related to the missing data process, a necessary condition for estimating the associated parameters (e.g., Robins, Rotnitzky and Zhao 1995). The MAR assumption is, however, not essential to the development of our methods here. If the probability of missingness does depend on unobserved response components, nonidentifiability may arise and in this case conducting sensitivity analyses is a viable strategy. Our methods can be applied to handle this scenario as well.

In circumstances where only the marginal structures are modulated, correction of missingness and error effects is commonly carried out sequentially. In this paper our methods are based on correcting first missingness effects and then error effects. This is basically driven by the way of modeling the missing data process for which the missingness probability is model as a function of error-prone covariates $\mathbf{X}_i$, together with responses $\mathbf{Y}_i$ and precisely measured covariates $\mathbf{Z}_i$. Such a modeling scheme for the missing data process is however, not essential for our proposed method. In some applications, the missingness probability is modulated as a function of surrogate measurements $\mathbf{W}_i$, along with $\mathbf{Y}_i$ and $\mathbf{Z}_i$. For example, Yi, Ma and Carroll (2012) considered this modeling scheme and developed marginal methods to first adjust for error effects and then correct missingness effects. Our model selection procedures can be modified to accommodating the case where the missingness probabilities are modulated as functions of surrogate measurements $\mathbf{W}_i$ as well as $\mathbf{Y}_i$ and $\mathbf{Z}_i$. It may be interesting to develop a sensible measure to compare the performance of model selection under these two modeling schemes for missing data processes.

Acknowledgements

The authors thank the referees for their helpful comments to improve the presentation of this work. The authors are grateful to Statistics Canada/The Research Data Centre Network for permission to use the data from the National Population Health Survey (NPHS), and thank Haocheng Li for his help to access the data. Yi's research was supported by the Natural Sciences and Engineering Research Council of Canada. Tan's research was supported by NIDA grants P50-DA10075. Li's research was supported by NSF grant DMS 1512422 and NIDA, NIH grants, P50 DA036107 and P50 DA039838. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIDA, NIH or NSF.

References

1. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 1974; 19:716–723.

2. Bondell HD, Krishna A, Ghosh SK. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*. 2010; 66:1069–1077. [PubMed: 20163404]
3. Carroll RJ, Lombard F, Küchenhoff H, Stefanski LA. Asymptotics for the SIMEX estimator in structural measurement error models. *Journal of the American Statistical Association*. 1996; 91:242–250.
4. Carroll, RJ.; Ruppert, D.; Stefanski, LA.; Crainiceany, CM. *Measurement Error in Nonlinear Models*. 2nd ed.. Chapman & Hall; 2006.
5. Cook JR, Stefanski LA. Simulation-extrapolation in parametric measurement error models. *Journal of the American Statistical Association*. 1994; 89:1314–1328.
6. Diggle P, Kenward MG. Informative drop-out in longitudinal data analysis (with discussion). *Applied Statistics*. 1994; 43:49–93.
7. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Annual of Statistics*. 2004; 32:407–499.
8. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96:1348–1360.
9. Fan J, Li R. New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*. 2004; 99:710–723.
10. Fan, J.; Li, R. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In: Sanz-Sole, M.; Soria, J.; Varona, JL.; Verdera, J., editors. *Proceedings of the International Congress of Mathematicians*. Vol. III. Zurich: European Mathematical Society; 2006. p. 595–622.
11. Fu WJ. Penalized estimating equations. *Biometrics*. 2003; 59:126–132. [PubMed: 12762449]
12. Geyer C. On the asymptotics of constrained M-estimation. *Ann. Statist.* 1994; 22:1993–2010.
13. Knight K, Fu W. Asymptotics for lasso-type estimators. *Ann. Statist.* 2000; 28:1356–1378.
14. Lai TL, Small D. Marginal regression analysis of longitudinal data with time-dependent covariates: a generalized method of moments approach. *Journal of the Royal Statistical Society, Ser. B*. 2007; 69:79–99.
15. Liang H, Li R. Variable selection for partially linear models with measurement Errors. *Journal of American Statistical Association*. 2009; 104:234–248.
16. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986; 73:13–22.
17. Lin X, Carroll RJ. Semiparametric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association*. 2001; 96:1045–1056.
18. Lin X, Carroll RJ. Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society, Ser. B*. 2006; 68:69–88.
19. Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. 2nd Ed.. New York: Wiley; 2002.
20. Ma Y, Li R. Variable selection in measurement error models. *Bernoulli*. 2010; 16:274–300. [PubMed: 20209020]
21. Miller, A. *Subset Selection in Regression*. 2nd edition. Boca Raton, Florida: Chapman & Hall/ CRC; 2002.
22. Ni X, Zhang D, Zhang HH. Variable selection for semiparametric mixed models in longitudinal studies. *Biometrics*. 2010; 66:79–88. [PubMed: 19397585]
23. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*. 1995; 90:106–121.
24. Schwarz G. Estimating the dimension of a model. *Annals of Statistics*. 1978; 6:461–464.
25. Shao J. An asymptotic theory for linear model selection. *Statistica Sinica*. 1997; 7:221–264.
26. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*. 1996; 58:267–288.
27. Tsiatis AA, DeGruttola V, Wulfsohn MS. Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*. 1995; 90:27–37.

28. Wang CY, Huang Y, Chao EC, Jeffcoat MK. Expected estimating equations for missing data, measurement error, and misclassification, with application to longitudinal nonignorable missing data. *Biometrics*. 2008; 64:85–95. [PubMed: 17608787]
29. Wang H, Li R, Tsai C-L. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*. 2007; 94:553–568. [PubMed: 19343105]
30. Yi GY. A simulation-based marginal method for longitudinal data with dropout and mismeasured covariates. *Biostatistics*. 2008; 9:501–512. [PubMed: 18199691]
31. Yi GY, Lawless JF. A corrected likelihood method for the proportional hazards model with covariates subject to measurement error. *Journal of Statistical Planning and Inference*. 2007; 137:1816–1828.
32. Yi GY, Reid N. A note on mis-specied estimating functions. *Statistica Sinica*. 2010; 20:1749–1769.
33. Yi GY, Ma Y, Carroll RJ. A functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error. *Biometrika*. 2012; 99:151–165.
34. Zhang Y, Li R, Tsai C-L. Regularization parameter selections via generalized information criterion. *Journal of American Statistical Association*. 2010; 105:312–323.
35. Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*. 2008; 36:1509–1533. [PubMed: 19823597]

Appendix

Proof of Theorem 1

We use the related techniques in Zou and Li (2008) to prove Theorem 1. Let $\mathbf{u} = (u_1, \dots, u_d)^T$ and consider $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + (1/\sqrt{n})\mathbf{u}$. Define

$$\delta_n(\mathbf{u}) = \frac{1}{2} \left(\frac{\mathbf{u}}{\sqrt{n}} + \boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}} \right)^T \mathbf{V}_n \left(\frac{\mathbf{u}}{\sqrt{n}} + \boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}} \right) + n \sum_{j=1}^d p'_\lambda(|\tilde{\theta}_j|) \left| \theta_{0j} + \frac{u_j}{\sqrt{n}} \right|.$$

Then

$$\delta_n(\mathbf{u}) - \delta_n(\mathbf{0}) = \frac{1}{2} \frac{\mathbf{u}^T}{\sqrt{n}} \mathbf{V}_n \frac{\mathbf{u}}{\sqrt{n}} + (\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}})^T \mathbf{V}_n \frac{\mathbf{u}}{\sqrt{n}} + n \sum_{j=1}^d p'_\lambda(|\tilde{\theta}_{0j}|) \left(\left| \theta_{0j} + \frac{u_j}{\sqrt{n}} \right| - |\theta_{0j}| \right)$$

Let $\hat{\mathbf{u}}(n) = \arg \min_{\mathbf{u}} [\delta_n(\mathbf{u}) - \delta_n(\mathbf{0})]$, then $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + (1/\sqrt{n})\hat{\mathbf{u}}(n)$.

Let $\mathbf{w}$ denote the limiting distribution of $\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. Thus, $\mathbf{w} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ by Assumption B given in Section 3. Using the related techniques in Zou and Li (2008), we can show that as $n \rightarrow \infty$,

$$T_1 \equiv \frac{1}{2} \frac{\mathbf{u}^T}{\sqrt{n}} \mathbf{V}_n \frac{\mathbf{u}}{\sqrt{n}} \xrightarrow{p} \frac{1}{2} \mathbf{u}^T \mathbf{V} \mathbf{u}$$

$$T_2 \equiv (\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}})^T \mathbf{V}_n \frac{\mathbf{u}}{\sqrt{n}} \xrightarrow{p} -\mathbf{w}^T \mathbf{V} \mathbf{u}$$

$$T_3 \equiv n \sum_{j=1}^d p'_\lambda(|\tilde{\theta}_j|) \left(\left| \theta_{0j} + \frac{u_j}{\sqrt{n}} \right| - |\theta_{0j}| \right) \xrightarrow{p} \begin{cases} 0, & \text{if } \mathbf{u}_2 = \mathbf{0}, \\ \infty, & \text{otherwise,} \end{cases}$$

where $\mathbf{u}_2$ is the d_2 dimensional subvector of $\mathbf{u}$ which is written as $\mathbf{u} = (\mathbf{u}_1^T, \mathbf{u}_2^T)^T$ with $\mathbf{u}_1$ being a d_1 -dimensional subvector.

Noting that $\mathbf{V}_1 = [\mathbf{V}_{11}, \mathbf{V}_{12}]$ and $\delta_n(\mathbf{u}) - \delta_n(\mathbf{0}) = T_1 + T_2 + T_3$, we have that as $n \rightarrow \infty$,

$$\delta_n(\mathbf{u}) - \delta_n(\mathbf{0}) \xrightarrow{p} \delta(\mathbf{u}) \equiv \begin{cases} \frac{1}{2} \mathbf{u}_1^T \mathbf{V}_{11} \mathbf{u}_1 - \mathbf{u}_1^T \mathbf{V}_1 \mathbf{w}, & \text{if } \mathbf{u}_2 = \mathbf{0}, \\ \infty, & \text{otherwise.} \end{cases}$$

The unique minimum of $\delta(\mathbf{u})$ is $\mathbf{u}_1 = \mathbf{V}_{11}^{-1} \mathbf{V}_1 \mathbf{w}$ since $\delta_n(\mathbf{u}) - \delta_n(\mathbf{0})$ is a convex function of $\mathbf{u}$.

Partition $\hat{\mathbf{u}}(n) = (\hat{\mathbf{u}}_1^T(n), \hat{\mathbf{u}}_2^T(n))^T$ with $\hat{\mathbf{u}}_1(n)$ being d_1 -dimensional. By epiconvergence (Geyer 1994; Knight and Fu 2000), we conclude that as $n \rightarrow \infty$,

$$\hat{\mathbf{u}}_1(n) \xrightarrow{d} \mathbf{V}_{11}^{-1} \mathbf{V}_1 \mathbf{w} \text{ and } \hat{\mathbf{u}}_2(n) \xrightarrow{d} \mathbf{0}.$$

By the definition of $\mathbf{w}$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_I - \boldsymbol{\theta}_{I0}) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_{11}^{-1} \mathbf{V}_1 \boldsymbol{\Sigma} \mathbf{V}_1^T \mathbf{V}_{11}^{-1}) \text{ as } n \rightarrow \infty.$$

As $n \rightarrow \infty$, $\hat{\mathbf{u}}_2(n) \xrightarrow{d} \mathbf{0}$, so we have that $\sqrt{n} \hat{\boldsymbol{\theta}}_{II} \xrightarrow{p} \mathbf{0}$. We next show that $\hat{\boldsymbol{\theta}}_{II} = \mathbf{0}$ with probability tending to one. It suffices to prove that for $j = d_1 + 1, \dots, d$,

$$\text{if } \theta_{0j} = 0, \text{ then } P(\hat{\theta}_j \neq 0) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

If $\hat{\theta}_j \neq 0$, by the KKT conditions, then we must have

$$\frac{1}{\sqrt{n}} \{ \mathbf{V}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) \}_j = \sqrt{n} p'_\lambda(|\tilde{\theta}_j|), \quad (11)$$

where $\{\mathbf{a}\}_j$ stands for the j th element of $\mathbf{a}$. By Assumption B, $\tilde{\theta}_j \rightarrow 0$ in probability as $n \rightarrow \infty$. Then the assumption in Theorem 1 implies that the right-hand side of (11) goes to ∞ in probability as $n \rightarrow \infty$. However, the left-hand side of (11) can be shown to be of order $O_p(1)$. Thus,

$$P(\hat{\theta}_j \neq 0) \leq P\left(\frac{1}{\sqrt{n}} \{ \mathbf{V}(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}) \}_j = \sqrt{n} p'_\lambda(|\tilde{\theta}_j|)\right) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Fig. 1. Sensitivity analysis results for the HUI covariate using the SIMEX procedure alone, using the SCAD penalty in combination with the SIMEX procedure, and using LASSO penalty in combination with the SIMEX procedure.

Table 1
Analysis of NPHS Data with Three Methods: Estimate and Standard Errors (a = 1, c = 1)

Response Model									
	Int.	BMI	HUI	INC	Marital	EDU1	EDU2	Age	
SCAD	0.581 (0.091)	0.000	-0.559 (0.065)	0.000	0.000	0.000	0.000	-0.030 (0.006)	
LASSO	0.601 (0.091)	0.000	-0.312 (0.065)	0.000	0.000	0.000	0.000	-0.036 (0.006)	
SIMEX	1.064 (0.168)	-0.016 (0.043)	-0.553 (0.067)	-0.076 (0.047)	-0.309 (0.100)	-0.138 (0.132)	0.125 (0.178)	-0.045 (0.007)	
Drop-out Model									
	Int.	Respons	BMI	HUI	INC	Marital	EDU1	EDU2	Age
SCAD	1.751 (0.047)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
LASSO	1.754 (0.047)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
SIMEX	1.444 (0.200)	0.046 (0.102)	-0.083 (0.049)	0.088 (0.047)	0.112 (0.056)	0.006 (0.113)	0.305 (0.131)	0.332 (0.194)	0.005 (0.008)

Table 2

Simulation Results for Response Model Parameters β

n	ρ_e	σ_e	MRMSE	Specif.	Sensit.	Under-fit	Exact fit	Over-fit		
								1	2	3
200	0.0	.15	0.305	0.963	0.970	0.080	0.771	0.106	0.024	0.019
		.50	0.837	0.887	0.901	0.236	0.459	0.151	0.072	0.083
		.70	1.077	0.881	0.853	0.333	0.347	0.175	0.077	0.069
	0.1	.15	0.298	0.959	0.972	0.070	0.767	0.106	0.036	0.022
		.50	0.950	0.873	0.885	0.248	0.432	0.133	0.072	0.115
		.70	1.106	0.862	0.845	0.363	0.296	0.159	0.077	0.106
	0.5	.15	0.307	0.953	0.973	0.063	0.753	0.119	0.038	0.027
		.50	0.920	0.894	0.869	0.293	0.440	0.118	0.048	0.102
		.70	1.223	0.867	0.829	0.378	0.304	0.140	0.078	0.101
500	0.0	.15	0.200	0.976	0.988	0.028	0.894	0.049	0.010	0.020
		.50	0.727	0.888	0.939	0.153	0.579	0.101	0.052	0.116
		.70	0.980	0.863	0.907	0.232	0.445	0.136	0.061	0.126
	0.1	.15	0.198	0.977	0.988	0.028	0.889	0.056	0.011	0.017
		.50	0.738	0.890	0.933	0.167	0.578	0.089	0.050	0.117
		.70	1.004	0.873	0.903	0.240	0.463	0.115	0.056	0.126
	0.5	.15	0.207	0.977	0.988	0.031	0.889	0.052	0.010	0.018
		.50	0.788	0.899	0.934	0.160	0.604	0.073	0.050	0.113
		.70	1.051	0.900	0.886	0.278	0.501	0.083	0.051	0.087

Table 3

Simulation Results for Drop-out Model Parameters α .

n	p	σ_e	MRMSE	Specif.	Sensit.	Under-fit	Exact fit	Over-fit		
								1	2	3
200	0.0	.15	0.548	0.990	0.964	0.109	0.842	0.036	0.008	0.005
		.50	0.607	0.944	0.943	0.143	0.654	0.096	0.045	0.063
		.70	0.790	0.952	0.944	0.123	0.666	0.125	0.040	0.046
		0.1	.15	0.506	0.989	0.974	0.858	0.056	0.005	0.002
	0.1	.50	0.664	0.943	0.932	0.152	0.630	0.112	0.056	0.050
		.70	0.839	0.938	0.923	0.176	0.588	0.107	0.061	0.069
		.15	0.570	0.987	0.971	0.085	0.840	0.063	0.008	0.005
		.50	0.693	0.952	0.917	0.204	0.625	0.089	0.042	0.040
	0.5	.70	0.915	0.942	0.889	0.269	0.518	0.114	0.053	0.046
500	0.0	.15	0.504	0.993	0.998	0.006	0.970	0.015	0.001	0.009
		.50	0.504	0.929	0.996	0.011	0.815	0.046	0.030	0.098
		.70	0.733	0.926	0.993	0.015	0.775	0.066	0.047	0.098
		0.1	.15	0.510	0.993	0.998	0.968	0.017	0.001	0.009
	0.1	.50	0.510	0.931	0.994	0.015	0.811	0.052	0.030	0.093
		.70	0.707	0.931	0.994	0.015	0.786	0.062	0.038	0.099
		.15	0.518	0.995	0.998	0.005	0.972	0.015	0.003	0.004
		.50	0.511	0.940	0.994	0.017	0.814	0.051	0.040	0.077
	0.5	.70	0.596	0.949	0.983	0.045	0.795	0.066	0.027	0.067

Table 4

Effect on Model Selection: Measurement Error versus Drop-out

Setting	p_e	σ_e	MRMSE	Specif.	Sensit.	Under-fit	Exact fit	Over-fit		
								1	2	3
I ($n = 200$)	0.0	.15	0.338	0.998	1.000	0.000	0.984	0.016	0.000	0.000
		.50	0.726	0.994	1.000	0.000	0.957	0.043	0.000	0.000
		.70	0.866	0.974	1.000	0.000	0.827	0.167	0.004	0.002
	0.1	.15	0.325	0.999	1.000	0.000	0.991	0.008	0.001	0.000
		.50	0.803	0.996	1.000	0.000	0.970	0.029	0.001	0.000
		.70	0.883	0.985	1.000	0.000	0.898	0.097	0.005	0.000
	0.5	.15	0.329	0.998	1.000	0.000	0.990	0.009	0.001	0.000
		.50	1.033	0.999	1.000	0.000	0.991	0.009	0.000	0.000
		.70	1.081	0.998	1.000	0.000	0.986	0.014	0.000	0.000
	II ($n = 200$)		0.247	0.974	0.982	0.046	0.836	0.090	0.020	0.008
I ($n = 500$)	0.0	.15	0.328	0.999	1.000	0.000	0.996	0.004	0.000	0.000
		.50	0.717	0.989	1.000	0.000	0.928	0.070	0.002	0.000
		.70	0.873	0.931	1.000	0.000	0.545	0.426	0.029	0.000
	0.1	.15	0.329	0.999	1.000	0.000	0.996	0.004	0.000	0.000
		.50	0.760	0.994	1.000	0.000	0.960	0.039	0.001	0.000
		.70	0.860	0.959	1.000	0.000	0.731	0.255	0.015	0.000
	0.5	.15	0.333	1.000	1.000	0.000	0.997	0.003	0.000	0.000
		.50	1.030	1.000	1.000	0.000	0.999	0.001	0.000	0.000
		.70	1.038	1.000	1.000	0.000	0.999	0.001	0.000	0.000
	II ($n = 500$)		0.189	0.987	0.993	0.020	0.915	0.062	0.004	0.000

Setting I: only measurement error is present. Setting II: only missingness is present

Table 5

Model Selection: 2-stage versus SIMEX

Method	Part	MRMSE	Specif.	Sensit.	Under-fit	Exact fit	Over-fit		
							1	2	3
Model A: ($\rho_e = 0.5, \sigma_e = 0.24$)									
2-stage	Drop-out	0.461	0.934	0.960	0.125	0.550	0.233	0.071	0.021
SIMEX	Drop-out	0.618	0.990	0.967	0.096	0.854	0.038	0.008	0.004
2-stage	Response	0.525	0.931	0.967	0.096	0.571	0.246	0.063	0.025
SIMEX	Response	0.346	0.943	0.967	0.083	0.700	0.146	0.046	0.025
Model B: ($\rho_e = 0.5, \sigma_e = 0.33$)									
2-stage	Drop-out	0.485	0.935	0.950	0.146	0.542	0.229	0.063	0.021
SIMEX	Drop-out	0.593	0.991	0.963	0.108	0.842	0.042	0.008	0.000
2-stage	Response	0.498	0.931	0.963	0.108	0.563	0.242	0.063	0.025
SIMEX	Response	0.574	0.933	0.963	0.104	0.642	0.142	0.088	0.025
Model C: ($\rho_e = 0.5, \sigma_e = 0.41$)									
2-stage	Drop-out	0.526	0.936	0.940	0.179	0.504	0.242	0.058	0.017
SIMEX	Drop-out	0.652	0.986	0.960	0.125	0.804	0.050	0.021	0.000
2-stage	Response	0.592	0.933	0.957	0.129	0.567	0.200	0.092	0.013
SIMEX	Response	0.768	0.937	0.950	0.129	0.633	0.146	0.075	0.017