Biometrika (2012), **99**, 1, *pp.* 151–165
doi: 10.1093/biomet/asr076

Printed in Great Britain
Advance Access publication 1 February 2012

A functional generalized method of moments approach for longitudinal studies with missing responses and covariate measurement error

By GRACE Y. YI

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

yyi@uwaterloo.ca

YANYUAN MA and RAYMOND J. CARROLL

Department of Statistics, Texas A&M University, TAMU 3143, College Station, Texas 77843-3143, U.S.A.

ma@stat.tamu.edu carroll@stat.tamu.edu

Summary

Covariate measurement error and missing responses are typical features in longitudinal data analysis. There has been extensive research on either covariate measurement error or missing responses, but relatively little work has been done to address both simultaneously. In this paper, we propose a simple method for the marginal analysis of longitudinal data with time-varying covariates, some of which are measured with error, while the response is subject to missingness. Our method has a number of appealing properties: assumptions on the model are minimal, with none needed about the distribution of the mismeasured covariate; implementation is straightforward and its applicability is broad. We provide both theoretical justification and numerical results.

Some key words: Functional measurement error; Generalized method of moments; Inverse probability weighting; Longitudinal data; Measurement error; Missing response; Structural measurement error.

1. Introduction

Longitudinal data analysis has attracted considerable research interest and a large number of inference methods have been proposed in the literature. Their validity relies on the important requirements that variables are perfectly measured and data are complete. In practice, however, these conditions are commonly violated, and ignoring this could result in seriously biased results. Although there has been discussion of inference methods to address longitudinal data with either missing values or measurement error, relatively little attention has been directed to accounting simultaneously for both features, which is partly attributable to complexity in modelling and computation. In this paper, we propose a simple method for the marginal analysis of longitudinal data with time-varying covariates, some of which are measured with error, while the response is subject to missingness.

Our method has these important features. First, our model is marginal: we do not specify a full distribution for the complete data, but instead we use estimating equations based on moment restrictions. Secondly, our method does not require the assumption that $E(Y_{ij} \mid X_i, Z_i) = E(Y_{ij} \mid X_{ij}, Z_{ij})$, which has been widely used in marginal analysis, especially when using the generalized estimating equation approach, e.g., Pepe & Anderson (1994) and Lai & Small (2007). Here Y_{ij} and (X_{ij}, Z_{ij}) represent the response and covariates for subject i at time j, and (X_i, Z_i) are the vector-valued subject level covariates. This is useful because in applications it is often difficult to assess this assumption. Moreover, this relaxation extends the scope of applicability of our method. Thirdly, we develop a functional measurement error method, i.e., nothing is assumed about the distribution of the variables measured with error.

In this paper, we explore how covariate measurement error may affect the association structure between the response and the covariates. We show that a conditional independence relationship between Y_{ij} and the true covariates (X_{ik}, Z_{ik}) for $j \neq k$ may be distorted when replacing X_{ik} by its observed surrogate W_{ik}. In the presence of missing values, we show that measurement error will usually alter the missing data mechanism. For example, a missing-at-random mechanism in error-prone covariates X_i is not retained when X_i is replaced by its surrogate W_i. To adjust for missing responses, we use inverse probability weighting based on the observed data. However, the development of a model for the probability of missingness is tricky. See § 2·4, where we discuss an obvious method that leads to difficulties and then a simple method that is completely transparent. In our development, the generalized method of moments, discussed by Hansen (1982), is used to combine unbiased estimating functions in the observed data.

There is considerable work on the measurement error problem in longitudinal data using the structural approach, namely when the distribution of the error-prone covariates is specified. References include Wang et al. (1998), Palta & Lin (1999), Lin & Carroll (2006), Liang (2009), Zhou & Liang (2009) and Xiao et al. (2010). Pan et al. (2009) investigate a transition model and apply the conditional and sufficient score approach of Stefanski & Carroll (1987) to analyse it. Similar uses of the conditional and sufficient score approach are seen in Li et al. (2004, 2007).

Papers that address all three facets of the problem, namely missing responses, longitudinal data and mismeasured covariates include Liu & Wu (2007), Wang et al. (2008), Yi (2005, 2008) and Yi et al. (2011). Liu & Wu (2007) and Yi et al. (2011) take a mixed model framework for the response process, and likelihood-based inferential procedures are used for which the missingness probability is modelled as a function of the true covariate and response variables, while Wang et al. (2008) and Yi (2005, 2008) use a marginal model framework. Except for the latter, these methods take a structural approach and assume that there is a model for the distribution of the mismeasured covariate X given error-free covariates. In modelling the missing data process, Yi (2005) and Wang et al. (2008) assume that the missingness probability is covariate free. Yi (2008) takes a functional approach to relax the need to model the covariate process, and allows the missingness probability to depend on covariates, but the resultant estimator is not exactly consistent due to the use of the simulation-extrapolation method (Cook & Stefanski, 1995).

2. RESPONSE MODEL AND ITS ASSUMPTIONS

2·1. *No measurement error*

Our data set-up is the following. There are $i = 1, \ldots, n$ independent individuals, with the possibility of $j = 1, \ldots, m$ visits. At visit j, the complete data are (Y_{ij}, X_{ij}, Z_{ij}), where Y_{ij} is the response, Z_{ij} are covariates observed exactly, and X_{ij} are error-prone covariates whose true values are unobserved. Denote $Y_i = (Y_{i1}, \ldots, Y_{im})^{\mathrm{T}}$, $X_i = (X_{i1}^{\mathrm{T}}, \ldots, X_{im}^{\mathrm{T}})^{\mathrm{T}}$ and

$Z_i = (Z_{i1}^{\mathrm{T}}, \ldots, Z_{im}^{\mathrm{T}})^{\mathrm{T}}$. Let $\mu_{ij} = E(Y_{ij} \mid X_{ij}, Z_{ij})$ and $v_{ij} = \mathrm{var}(Y_{ij} \mid X_{ij}, Z_{ij})$ be the time-specific conditional expectation and variance of Y_{ij}, respectively, given X_{ij} and Z_{ij}.

Consider the regression model

$$g(\mu_{ij}) = X_{ij}^{\mathrm{T}}\beta_x + Z_{ij}^{\mathrm{T}}\beta_z, \tag{1}$$

where $g(\cdot)$ is a known monotone function, and $\mathcal{B} = (\beta_x^{\mathrm{T}}, \beta_z^{\mathrm{T}})^{\mathrm{T}}$ is the vector of regression parameters. If necessary, an intercept may be included in β_z by including unity in the covariate vector Z_{ij}. Further, assume $v_{ij} = h(\mu_{ij}, \phi)$, where $h(\cdot)$ is a known function and ϕ is a dispersion parameter that is known or may be estimated. For instance, for binary data there is no ϕ and $v_{ij} = \mu_{ij}(1 - \mu_{ij})$. We treat ϕ as known here with emphasis on estimation of $\mathcal{B}$.

2·2. *Measurement error process*

Let W_{ij} be the observed version of the covariate X_{ij}. Denote $W_i = (W_{i1}^{\mathrm{T}}, \ldots, W_{im}^{\mathrm{T}})^{\mathrm{T}}$. We assume that X_{ij} and W_{ij} follow a classical additive measurement error model, i.e., given (X_{ij}, Z_{ij}, Y_{ij}),

$$W_{ij} = X_{ij} + e_{ij}, \tag{2}$$

where $e_{ij} \sim N(0, \Sigma_j)$ is independent of X_i, Z_i and Y_i.

Here we merely model marginal characteristics of the measurement error process for each time-point, and employ a functional modelling strategy with the distribution of X_{ij} left unspecified. Model (2) accommodates the general scenario of dependent measurement errors. We assume that the Σ_j are known or estimated from replication experiments (Carroll et al., 2006).

2·3. *Working independence and evolving covariates*

In our approach, we build separate unbiased estimating functions for $\mathcal{B}$ at each time-point $j = 1, \ldots, m$, eventually combining them via the generalized method of moments. This section describes why such an approach is practical in the context of measurement error.

Model (1) for the response postulates the population mean at each time-point as a function of time-specific covariates. Because of the measurement error, attention must be paid to how the true covariates evolve over time, and how they are related to responses at other time-points.

The conventional assumption when there is no measurement error and no missing data is that $E(Y_{ij} \mid X_i, Z_i) = E(Y_{ij} \mid X_{ij}, Z_{ij})$, which is effectively the same thing as stating that given (X_{ij}, Z_{ij}), Y_{ij} is independent of (X_{ik}, Z_{ik}) for $j \neq k$. Covariates satisfying this condition are called Type I by Lai & Small (2007). Thus, for example, assume that there is no Z, $m = 2$, $\mathrm{cov}(X_{i1}, X_{i2}) = \Sigma_x$, $Y_{ij} = \mu(X_{ij}, \mathcal{B}) + \epsilon_{ij}$ and $\mathrm{cov}(\epsilon_{i1}, \epsilon_{i2}) = \Sigma_\epsilon$. Under the Type I assumption, an unbiased estimating function for $\mathcal{B}$ is

$$\{\mu_{\mathcal{B}}(X_{i1}, \mathcal{B}), \mu_{\mathcal{B}}(X_{i2}, \mathcal{B})\}\Sigma_\epsilon^{-1}[(Y_{i1}, Y_{i2})^{\mathrm{T}} - \{\mu(X_{i1}, \mathcal{B}), \mu(X_{i2}, \mathcal{B})\}^{\mathrm{T}}],$$

where the subscript $\mathcal{B}$ denotes differentiation with respect to $\mathcal{B}$. Accounting for the correlation among the ϵ_{ij} leads to more efficient estimation of $\mathcal{B}$ than ignoring the correlation. It is an unbiased estimating function because, under the Type I assumption, for any function $G(\cdot, \cdot)$, $E[G(X_{ik}, X_{ij})\{Y_{ij} - \mu(X_{ij}, \mathcal{B})\}] = 0$ for $j \neq k$. However, Pepe & Anderson (1994), Pepe & Couper (1997) and Lai & Small (2007) point out that if X_{i2} is not independent of Y_{i1} given X_{i1}, then this generalized estimating function is not unbiased, i.e., does not have mean zero when evaluated at the model, so the resulting estimates may be inconsistent. This fact has led many authors to assume that Σ_ϵ is diagonal, the so-called working independence assumption that leads to consistent estimators. This is an interesting debate, and indeed in cases that the X_{ij} can be observed, Lai & Small (2007) show how to test the Type I assumption.

We are dealing with a very different problem. Suppose that the Type I assumption holds in the underlying data, so that for any j, given (X_{ij}, Z_{ij}), Y_{ij} is independent of (X_{ik}, Z_{ik}) for $k \neq j$. We can show that in the presence of measurement error, the observed data, Y_{ij} and (W_{ik}, Z_{ik}), given (W_{ij}, Z_{ij}) for $j \neq k$, are generally dependent. Thus, testing the Type I assumption based on observed, error-prone data is likely to be difficult in practice. We sum this up in the following result.

LEMMA 1. *Suppose the Type* I *assumption of* Lai & Small (2007) *holds in the complete data* (Y_{ij}, X_{ij}, Z_{ij}), *and thus that* Y_{ij} *and* (X_{ik}, Z_{ik}) *are independent given* (X_{ij}, Z_{ij}) *for* $j \neq k$. *In general, it is not the case that the Type* I *assumption holds in the observed data* (Y_{ij}, W_{ij}, Z_{ij}).

See Appendix A1 for a proof of Lemma 1. As illustrated there, with dependent errors, the dependence of Y_{ij} on the observed W_{ik} may not be fully captured by that of Y_{ij} on W_{ij}. Even when the true covariates X_{ij} and X_{ik} are independent, the dependence between the errors e_{ij} and e_{ik} induces correlation between the observed W_{ij} and W_{ik}, which may distort the relationship between the response and the true covariates.

2·4. *Missing data process*

First we examine how measurement error may change the missing data mechanism classification in the error-free context. In general, the model structure between the missing data indicator and the covariates is not preserved when the true covariates X_i are replaced with their surrogate W_i. In particular, Appendix A2 sketches an illustration of the following result. Let $R_{ij} = 1$ if Y_{ij} is observed and $R_{ij} = 0$ otherwise. Let $\tilde{R}_{ij} = \{R_{i1}, \ldots, R_{i,j-1}\}$ be the history of the missing data indicator at time-point j, and let $Y_i^{(o)}$ contain the observed response measurements of Y_i.

LEMMA 2. *If* $\mathrm{pr}(R_{ij} = 1 \mid \tilde{R}_{ij}, Y_i, X_i, Z_i) = \mathrm{pr}(R_{ij} = 1 \mid \tilde{R}_{ij}, Y_i^{(o)}, X_i, Z_i)$, *then it is possible that* $\mathrm{pr}(R_{ij} = 1 \mid \tilde{R}_{ij}, Y_i, W_i, Z_i) \neq \mathrm{pr}(R_{ij} = 1 \mid \tilde{R}_{ij}, Y_i^{(o)}, W_i, Z_i)$. *That is, the missingness process could be missing at random in* X_i, *but not missing at random in* W_i.

Next, we discuss how and why we model the missing data process. There are two obvious ways to do so. In the first case, one would take a measurement error view, and model the missing data process in terms of the underlying unobserved X_{ij} data, and then adjust this in terms of the observed W_{ij} data. This method is needlessly complex because it involves two stages:

Stage I. Fit a possibly complex logistic measurement error model with R_{ij} being the response and with covariates involving the unobserved X_{ij}, forming in the simplest case, for example, a model with probabilities $\pi(X_{ij}, Z_{ij}, \alpha) = \mathrm{pr}(R_{ij} = 1 \mid X_{ij}, Z_{ij})$, where α is a parameter.

Stage II. Obtain the observed probabilities by regressing $\pi(X_{ij}, Z_{ij}, \alpha)$ on (W_{ij}, Z_{ij}). This almost inevitably requires a model for X_{ij}, which goes against the philosophy of functional measurement error models, and seems an indirect way to estimate probabilities in the observed data space.

In the second case, one simply models the missing data process directly in terms of the observed data. The first case is needlessly complex, even if there are no repeated measures, and thus for the reasons described above we use the second approach.

Because the problems considered here involve covariate measurement error, it is not sensible to stick to the usual classification of missing data mechanisms that are defined in the error-free context. Therefore, we abandon the usual modelling scheme of postulating the dependence of the missing data indicator on the true covariates X_i along with other variables, but modulate the missingness probability based on the observed surrogates W_i. This treatment of the missing data process enables us to build a more sensible model and allows more transparent interpretation of model parameters.

To reflect the dynamic nature of the observation process over time, we assume that $\text{pr}(R_{ij} = 1 \mid \tilde{R}_{ij}, Y_i, W_i, Z_i) = \text{pr}(R_{ij} = 1 \mid \tilde{R}_{ij}, Y_i^{(o)}, W_i, Z_i)$. This assumption is analogous to the missing-at-random mechanism in the error-free context, and it says that the missingness probability depends on the observed data. As required in marginal analysis within the error-free context, inclusion of such an assumption is merely to ensure model identifiability related to the missing data process, a necessary condition for estimating the associated parameters. This assumption is not essential to the development here, and it can be removed when conducting sensitivity analyses is of interest.

Because subjects are assessed sequentially over time, it is natural to make a further assumption to reflect inherent ordering in time, i.e.,

$$\text{pr}(R_{ij} = 1 \mid \tilde{R}_{ij}, Y_i^{(o)}, W_i, Z_i) = \text{pr}(R_{ij} = 1 \mid \tilde{R}_{ij}, \tilde{Y}_{ij}^{(o)}, \tilde{W}_{ij}, \tilde{Z}_{ij}),$$

where $\tilde{S}_{ij} = \{S_{i1}, S_{i2}, \ldots, S_{i,j-1}\}$ represents the history of variable S_{ij} at time-point j for $S_{ij} = Y_{ij}^{(o)}$, W_{ij} and Z_{ij}.

Let $\pi_{ij} = \text{pr}(R_{ij} = 1 \mid \tilde{R}_{ij}, \tilde{Y}_{ij}^{(o)}, \tilde{W}_{ij}, \tilde{Z}_{ij})$. One may use a logistic regression model to posit the missing data process, i.e., $\text{logit}(\pi_{ij}) = u_{ij}^{\mathrm{T}}\alpha$, where u_{ij} is the vector consisting of the information on the history of surrogates W_{ij}, covariates Z_{ij} and the observed responses $Y_{ij}^{(o)}$ as well as the missing data indicator R_{ij}. Here α is the vector of regression parameters.

Estimation of α can proceed using a likelihood-based method. Let $L_i(\alpha) = \prod_{j=1}^{m} \pi_{ij}^{R_{ij}} (1 - \pi_{ij})^{1-R_{ij}}$ be the likelihood contribution from subject i, and $S_i(\alpha) = (\partial/\partial\alpha) \log L_i(\alpha)$. Then solving $\sum_{i=1}^{n} S_i(\alpha) = 0$ leads to a consistent estimator $\hat{\alpha}$ of α.

3. Methodology

3·1. *Overview*

In this section, we propose an inference method based on the time-specific marginal structure of the response process. The method is simple to implement but flexible enough to accommodate a wide class of applications. The key idea is to construct an unbiased estimating function, say $\mathcal{G}_{ij}(Y_{ij}, X_{ij}, Z_{ij}, \mathcal{B})$, for each time-point j under the ideal situation when neither missing responses nor covariate measurement error is present. Then in § 3·2 we correct for the error effects by using the marginal moments for the error process, and denote by $\mathcal{G}_{ij}^{*}(Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B})$ the adjusted estimating functions that are expressed in terms of the responses and the observed covariates. In the next step, in § 3·3 we modify these estimating functions by using inverse probability weights to incorporate missingness effects. Finally, in § 3·4, we use the generalized method of moments to combine those time-specific unbiased estimating functions to formulate one that is efficient in the class of all of their linear combinations.

3·2. *Corrected scores adjusting for error effects*

For each time-point $j = 1, \ldots, m$, let $\mathcal{G}_{ij}(Y_{ij}, X_{ij}, Z_{ij}, \mathcal{B})$ be unbiased estimating functions for $\mathcal{B}$ when there is neither measurement error nor missing observations. Specifically, we take

$$\mathcal{G}_{ij}(Y_{ij}, X_{ij}, Z_{ij}, \mathcal{B}) = \frac{\partial \mu_{ij}}{\partial \mathcal{B}^{\mathrm{T}}} v_{ij}^{-1}(Y_{ij} - \mu_{ij}),$$

which requires the weakest model assumption for the response process as made in § 2·1. For example, a logit link function for binary data yields $\mathcal{G}_{ij}(Y_{ij}, X_{ij}, Z_{ij}, \mathcal{B}) = \{Y_{ij} - H(X_{ij}^{\mathrm{T}}\beta_x + Z_{ij}^{\mathrm{T}}\beta_z)\}(X_{ij}^{\mathrm{T}}, Z_{ij}^{\mathrm{T}})^{\mathrm{T}}$, where $H(t) = \exp(t)/\{1 + \exp(t)\}$ is the logistic distribution function.

In the absence of measurement error and missing observations, $\mathcal{G}_{ij}(Y_{ij}, X_{ij}, Z_{ij}, \mathcal{B})$ are unbiased estimating functions for $\mathcal{B}$, and hence they can be used to estimate $\mathcal{B}$. When X_{ij} is subject to measurement error, however, we cannot directly use $\mathcal{G}_{ij}(Y_{ij}, X_{ij}, Z_{ij}, \mathcal{B})$ by replacing X_{ij} with its observed value W_{ij}, because the resulting estimating functions are not unbiased. One strategy to remedy this is to work on the conditional expectation $E\{\mathcal{G}_{ij}(Y_{ij}, X_{ij}, Z_{ij}, \mathcal{B}) \mid Y_{ij}, W_{ij}, Z_{ij}\}$ to correct for the error effects. However, this approach requires a distributional assumption for X_{ij}, which is difficult to specify in many applications. Alternatively, we proceed with a correction approach which does not require specification of the distribution of X_{ij}. The idea is to construct estimating functions $\mathcal{G}_{ij}^*(Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B})$ such that

$$E\{\mathcal{G}_{ij}^*(Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B}) \mid Y_{ij}, X_{ij}, Z_{ij}\} = \mathcal{G}_{ij}(Y_{ij}, X_{ij}, Z_{ij}, \mathcal{B}). \quad (3)$$

With this construction, unbiasedness of $\mathcal{G}_{ij}(Y_{ij}, X_{ij}, Z_{ij}, \mathcal{B})$ under the response model (1) ensures unbiasedness of $\mathcal{G}_{ij}^*(Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B})$ in the absence of missingness. This strategy has the same spirit as the so-called corrected likelihood method discussed by Nakamura (1990) for generalized linear models, but it applies more generally to estimating functions.

With regression models such as linear regression, Gamma regression, inverse Gaussian regression and Poisson regression, Yi (2005) presents the expressions of the $\mathcal{G}_{ij}^*(Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B})$ functions. However, with logistic regression for binary data under error model (2), there exist no analytical functions $\mathcal{G}_{ij}^*(Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B})$ such that (3) is true (Stefanski, 1989). This nonexistence is basically caused by the terms like $X_{ij}[1 + \exp\{-(X_{ij}^{\mathrm{T}}\beta_x + Z_{ij}^{\mathrm{T}}\beta_z)\}]^{-1}$ involved in the logistic regression.

To develop a more flexible method, we next propose to introduce proper weights for the estimating functions $\mathcal{G}_{ij}(Y_{ij}, X_{ij}, Z_{ij}, \mathcal{B})$ so that a workable version $\mathcal{G}_{ij}^*(Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B})$ is readily constructed. To be specific, let $\eta(X_{ij}, Z_{ij}, \mathcal{B})$ be a function that does not depend on the response Y_{ij} or missing data indicators R_i, but could depend on covariates and parameters. Define

$$\mathcal{G}_{wij}(Y_{ij}, X_{ij}, Z_{ij}, \mathcal{B}) = \eta(X_{ij}, Z_{ij}, \mathcal{B})\mathcal{G}_{ij}(Y_{ij}, X_{ij}, Z_{ij}, \mathcal{B}).$$

If there is a function $\mathcal{G}_{ij}^*(Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B})$ such that

$$E\{\mathcal{G}_{ij}^*(Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B}) \mid Y_{ij}, X_{ij}, Z_{ij}\} = \mathcal{G}_{wij}(Y_{ij}, X_{ij}, Z_{ij}, \mathcal{B}),$$

then $\sum_{i=1}^{n} \mathcal{G}_{ij}^*(Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B})$ is an unbiased estimating function for $\mathcal{B}$ in the absence of missing observations. This scheme can be used in other contexts as well. For instance, with binary data Huang & Wang (2001) use this strategy for the logistic regression model, and in unpublished work, the first author has invoked this technique to construct estimating functions for analysis of interval count data.

In particular, for binary data with logistic regression, setting $\eta(X_{ij}, Z_{ij}, \mathcal{B}) = 1 + \exp\{-(X_{ij}^{\mathrm{T}}\beta_x + Z_{ij}^{\mathrm{T}}\beta_z)\}$ leads to

$$\mathcal{G}_{wij}(Y_{ij}, X_{ij}, Z_{ij}, \mathcal{B}) = \{Y_{ij} H^{-1}(X_{ij}^{\mathrm{T}}\beta_x + Z_{ij}^{\mathrm{T}}\beta_z) - 1\}(X_{ij}^{\mathrm{T}}, Z_{ij}^{\mathrm{T}})^{\mathrm{T}}.$$

By the moment identities associated with the error model (2), $E(W_{ij} \mid X_{ij}) = X_{ij}$, $E\{\exp(W_{ij}^{\mathrm{T}}\beta_x) \mid X_{ij}\} = \exp(X_{ij}^{\mathrm{T}}\beta_x + \beta_x^{\mathrm{T}}\Sigma_j\beta_x/2)$ and $E\{W_{ij}\exp(W_{ij}^{\mathrm{T}}\beta_x) \mid X_{ij}\} = (X_{ij} + \Sigma_j\beta_x)$ $\exp(X_{ij}^{\mathrm{T}}\beta_x + \beta_x^{\mathrm{T}}\Sigma_j\beta_x/2)$, hence we take $\mathcal{G}_{ij}^*(Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B})$ to be $\{Y_{ij}H^{-1}(W_{ij}^{\mathrm{T}}\beta_x + Z_{ij}^{\mathrm{T}}\beta_z + \beta_x^{\mathrm{T}}\Sigma_j\beta_x/2) - 1\}(W_{ij}^{\mathrm{T}}, Z_{ij}^{\mathrm{T}})^{\mathrm{T}} + Y_{ij}\{H^{-1}(W_{ij}^{\mathrm{T}}\beta_x + Z_{ij}^{\mathrm{T}}\beta_z + \beta_x^{\mathrm{T}}\Sigma_j\beta_x/2) - 1\}\{(\Sigma_j\beta_x)^{\mathrm{T}}, 0^{\mathrm{T}}\}^{\mathrm{T}}$, where the column vector 0 has the dimension of Z_{ij}.

3·3. *Inverse probability weights adjusting for missingness effects*

Let $\Phi_{ij}^* = (R_{ij}/\pi_{ij})\mathcal{G}_{ij}^*(Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B})$. Then Φ_{ij}^* is unbiased by the definition of π_{ij}. Indeed, let E_{R_i}, $E_{R_{i1}}$ and $E_{R_{ij}|\tilde{R}_{ij}}$ denote the conditional expectations taken with respect to the conditional densities $f(r_i \mid Y_i, W_i, Z_i)$, $f(r_{i1} \mid Y_i, W_i, Z_i)$ and $f(r_{ij} \mid \tilde{R}_{ij}, Y_i, W_i, Z_i)$. Then

$$\begin{aligned} E_{R_i}(\Phi_{ij}^*) &= E_{R_{i1}}[E_{R_{i2}|\tilde{R}_{i2}} \ldots \{E_{R_{im}|\tilde{R}_{im}}(\Phi_{ij}^*)\}] \\ &= E_{R_{i1}}(E_{R_{i2}|\tilde{R}_{i2}} \ldots [E_{R_{ij}|\tilde{R}_{ij}}\{(R_{ij}/\pi_{ij})\mathcal{G}_{ij}^*(Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B})\}]) \\ &= E_{R_{i1}}(E_{R_{i2}|\tilde{R}_{i2}} \ldots [E_{R_{i,j-1}|\tilde{R}_{i,j-1}}\{\mathcal{G}_{ij}^*(Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B})\}]) \\ &= \mathcal{G}_{ij}^*(Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B}), \end{aligned}$$

where the sign ... represents the evaluation for the sequence of conditional expectations $E_{R_{i3}|\tilde{R}_{i3}}, \ldots, E_{R_{i,j-1}|\tilde{R}_{i,j-1}}$, and the third identity is due to the assumptions made in § 2·4 and the definition of π_{ij}. As a result, $E(\Phi_{ij}^*) = 0$ by the unbiasedness of $\mathcal{G}_{ij}^*(Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B})$ established in § 3·2.

The unbiasedness of Φ_{ij}^* allows us to express all unbiased estimating functions in the form

$$\frac{R_{ij}}{\pi_{ij}}\mathcal{G}_{ij}^*(Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B}) + \frac{R_{ij} - \pi_{ij}}{\pi_{ij}}\mathcal{H}(W_{ij}, Z_{ij}, \mathcal{B}),$$

for some function $\mathcal{H}(\cdot)$. In cases such as ours where $\mathcal{G}_{ij}^*(\cdot)$ is linear in Y_{ij}, so that $\mathcal{G}_{ij}^*(\cdot) = A(W_{ij}, Z_{ij}, \mathcal{B})Y_{ij} - B(W_{ij}, Z_{ij}, \mathcal{B})$ say, we propose setting $\mathcal{H}(W_{ij}, Z_{ij}, \mathcal{B}) = B(W_{ij}, Z_{ij}, \mathcal{B})$. This is algebraically equivalent to replacing argument Y_{ij} with $(R_{ij}/\pi_{ij})Y_{ij}$ in $\mathcal{G}_{ij}^*(\cdot)$, and we thus propose the estimating function

$$\Phi_{ij}(\mathcal{B}) = \mathcal{G}_{ij}^*\{(R_{ij}/\pi_{ij})Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B}\},$$

where dependence of $\Phi_{ij}(\mathcal{B})$ on the parameter α is suppressed.

3·4. *Generalized method of moments*

We now optimally combine the functions $\sum_{i=1}^n \Phi_{ij}(\mathcal{B})$ across $j = 1, \ldots, m$, using the generalized method of moments (Hansen, 1982). Let $\Phi_i(\mathcal{B}) = \{\Phi_{i1}^{\mathrm{T}}(\mathcal{B}), \ldots, \Phi_{im}^{\mathrm{T}}(\mathcal{B})\}^{\mathrm{T}}$ and $\Phi(\mathcal{B}) = \sum_{i=1}^n \Phi_i(\mathcal{B})$. Then a generalized method of moments estimator of $\mathcal{B}$ is obtained by minimizing $\Phi^{\mathrm{T}}(\mathcal{B})G\Phi(\mathcal{B})$, where G is a weight matrix. The asymptotically optimal weight matrix G is the inverse of the covariance matrix of $\Phi_i(\mathcal{B})$. Equivalently, the optimal generalized method of moments estimator solves the estimating equations

$$\sum_{i=1}^{n} \Psi_i(\mathcal{B}) = 0, \tag{4}$$

where $\Psi_i(\mathcal{B}) = DV^{-1}\Phi_i(\mathcal{B})$, $D = E\{(\partial/\partial\mathcal{B})\Phi_i^{\mathrm{T}}(\mathcal{B})\}$ and $V = E\{\Phi_i(\mathcal{B})\Phi_i^{\mathrm{T}}(\mathcal{B})\}$.

In implementing (4), we replace D and V by the consistent estimates $\hat{D} = n^{-1}\sum_{i=1}^{n}(\partial/\partial\mathcal{B})\Phi_i^{\mathrm{T}}(\mathcal{B})$ and $\hat{V} = n^{-1}\sum_{i=1}^{n}\Phi_i(\mathcal{B})\Phi_i^{\mathrm{T}}(\mathcal{B})$. Set $\hat{\Psi}_i(\mathcal{B}) = \hat{D}\hat{V}^{-1}\Phi_i(\mathcal{B})$. Then solving $\sum_{i=1}^{n}\hat{\Psi}_i(\mathcal{B}) = 0$ leads to the estimator $\hat{\mathcal{B}}$ of $\mathcal{B}$.

3·5. *Asymptotic theory*

When α is known to be α_0, say, then under regularity conditions, $n^{1/2}(\hat{\mathcal{B}} - \mathcal{B})$ has an asymptotic multivariate normal distribution with mean zero and covariance matrix $\Gamma_0^{-1}E\{\Psi_i(\mathcal{B},\alpha_0)\Psi_i^{\mathrm{T}}(\mathcal{B},\alpha_0)\}(\Gamma_0^{-1})^{\mathrm{T}}$, where $\Gamma_0 = E\{\partial\Psi_i(\mathcal{B},\alpha_0)/\partial\mathcal{B}^{\mathrm{T}}\}$. However, when α is unspecified and estimated, the variation of $\hat{\alpha}$ must be taken into account. A sketch of the technical details is given in Appendix A3. In this case, under regularity conditions, $n^{1/2}(\hat{\mathcal{B}} - \mathcal{B})$ has an asymptotic multivariate normal distribution with mean zero and covariance matrix $\Gamma^{-1}\Omega(\Gamma^{-1})^{\mathrm{T}}$, where $\Gamma = E\{\partial\Psi_i(\mathcal{B},\alpha)/\partial\mathcal{B}^{\mathrm{T}}\}$, $\Omega = E\{Q_i(\mathcal{B},\alpha)Q_i^{\mathrm{T}}(\mathcal{B},\alpha)\}$ and $Q_i(\mathcal{B},\alpha) = \Psi_i(\mathcal{B},\alpha) - E\{\partial\Psi_i(\mathcal{B},\alpha)/\partial\alpha^{\mathrm{T}}\}[E\{\partial S_i(\alpha)/\partial\alpha^{\mathrm{T}}\}]^{-1}S_i(\alpha)$. Thus, inference on $\mathcal{B}$ can be conducted by replacing the asymptotic covariance matrix with its consistent estimate in the asymptotic normal distribution for $\hat{\mathcal{B}}$. All these quantities can be estimated by method of moments calculations. Specifically, as $n\to\infty$, Γ is estimated by the consistent estimator $\hat{\Gamma} = n^{-1}\sum_{i=1}^{n}\{\partial\Psi_i(\hat{\mathcal{B}},\hat{\alpha})/\partial\mathcal{B}^{\mathrm{T}}\}$, and Ω is estimated by $\hat{\Omega} = n^{-1}\sum_{i=1}^{n}\{\hat{Q}_i(\hat{\mathcal{B}},\hat{\alpha})\hat{Q}_i^{\mathrm{T}}(\hat{\mathcal{B}},\hat{\alpha})\}$, where $\hat{Q}_i(\mathcal{B},\alpha) = \Psi_i(\mathcal{B},\alpha) - [n^{-1}\sum_{i=1}^{n}\{\partial\Psi_i(\mathcal{B},\alpha)/\partial\alpha^{\mathrm{T}}\}][n^{-1}\sum_{i=1}^{n}\{\partial S_i(\alpha)/\partial\alpha^{\mathrm{T}}\}]^{-1}S_i(\alpha)$.

The development here primarily focuses on accounting for the variation induced by estimation of α, the parameter associated with the missing data process. The dispersion parameter ϕ in the response model and the parameters governing the measurement error process are typically assumed known. One does not, however, have to be restricted by this assumption. It is straightforward to modify the proof in Appendix A3 to accommodate the variability due to estimation of those parameters when necessary. For instance, if there are replicates of W_{ij}, one may use the method of moments to estimate parameters, say σ, for the measurement error model. Now let $U_i(\sigma)$ be the corresponding vector of estimating functions for σ from subject i. Then under regularity conditions, $n^{1/2}(\hat{\mathcal{B}} - \mathcal{B})$ has an asymptotic multivariate normal distribution with mean zero, and its asymptotic covariance matrix assumes the same form as before, except for replacing $S_i(\alpha)$ with $S_i^*(\theta) = \{U_i^{\mathrm{T}}(\sigma), S_i^{\mathrm{T}}(\alpha)\}^{\mathrm{T}}$ and replacing α with $\theta = (\sigma^{\mathrm{T}}, \alpha^{\mathrm{T}})^{\mathrm{T}}$. In the same spirit, if the dispersion parameter ϕ is estimated from an estimating function, one can add this function to $S_i^*(\theta)$ to work out the asymptotic covariance matrix for $n^{1/2}(\hat{\mathcal{B}} - \mathcal{B})$.

4. SIMULATION STUDIES

4·1. *Comparison with the other methods*

In this section, we discuss the results of simulation studies meant to assess the performance of our method, and contrast these with other analyses which ignore measurement error or missingness or both. We set $n = 500$ and $m = 5$ and generated 1000 simulated datasets for each parameter configuration. We generated response measurements Y_{ij} independently from the logistic regression model

$$\mathrm{logit}(\mu_{ij}) = \beta_0 + \beta_{x1} X_{ij1} + \beta_{x2} X_{ij2} + \beta_z Z_i, \tag{5}$$

where Z_i takes values 0 or 1 with probability 0·5. Independent of Z_i, $X_{ij} = (X_{ij1}, X_{ij2})^{\mathrm{T}}$ was generated as $N(\mu_x, \Sigma_x)$ where $\mu_x = (\mu_{x1}, \mu_{x2})^{\mathrm{T}}$, while Σ_x has variances $(\sigma_{x1}^2, \sigma_{x2}^2)$ and correlation ρ_x, with $\mu_{xr} = 0{\cdot}5$ and $\sigma_{xr} = 1$ for $r = 1, 2$. We set $\beta_0 = -0{\cdot}1$, $\beta_{x1} = 0{\cdot}3$, $\beta_{x2} = 0{\cdot}6$ and $\beta_z = 0{\cdot}5$. The surrogate value $W_{ij} = (W_{ij1}, W_{ij2})^{\mathrm{T}}$ was generated as $N(X_{ij}, \Sigma_j)$, where Σ_j has variances (σ_1^2, σ_2^2) and correlation ρ. Various configurations were considered to feature distinct scenarios of measurement error in the covariate X_{ij}. Specifically, we considered $(\sigma_1, \sigma_2) = 0{\cdot}15, 0{\cdot}5, 0{\cdot}75$ to feature minor, moderate and severe marginal measurement error, and $(\rho_x, \rho) = (0, 0)$, $(0, 0{\cdot}5)$, $(0{\cdot}5, 0)$ or $(0{\cdot}5, 0{\cdot}5)$ to allow for various correlations. Since the main conclusions are similar in all cases, we will display only the last case.

We considered a case with drop-outs where the missing data indicator is generated from the model

$$\mathrm{logit}(\pi_{ij}) = \alpha_0 + \alpha_y Y_{i,j-1} + \alpha_w W_{i,j-1,1} + \alpha_z Z_i, \tag{6}$$

where we set $\alpha_0 = -0{\cdot}3$, $\alpha_y = 0{\cdot}5$, $\alpha_w = 0{\cdot}2$ and $\alpha_z = 0{\cdot}2$. This yields about 45% missingness.

Four analyses were conducted: (a) the naive one which ignores both covariate measurement error and missing responses, performed using the usual generalized estimating equations method with an independence correlation matrix employed; (b) the analysis that takes missingness into account but ignores measurement error in covariates; (c) the analysis that accounts for measurement error effects but ignores missingness and (d) our method which accommodates both measurement error and missingness.

In Table 1, we report on the case that $(\rho_x, \rho) = (0{\cdot}5, 0{\cdot}5)$, the other configurations being similar. We displayed the biases of the estimates, their mean squared errors and also the coverage rates of nominal 95% confidence intervals. As expected, the three analyses that do not accommodate measurement error or missingness produce strikingly biased results. Their biases increase as the degree of measurement error increases. Ignoring measurement error or missingness also has a profound impact on coverage rates of confidence intervals, primarily because of the bias, and the coverage rates decrease as the measurement error variance increases. Biases are also affected by the correlations between the true and observed covariates. In contrast, our method is very satisfactory, with coverage rates close to the nominal 95%.

These empirical studies demonstrate that ignoring either missingness or measurement or both could result in visibly biased results. It is important to adjust for effects induced by measurement error and missing data in inferential procedures, as does our method.

4·2. *Sensitivity of our method*

In this section, we assess the sensitivity of our method. In particular, we consider the case that the missing data model is misspecified. To be specific, the missing data indicator for drop-outs is generated from the model which depends on the true covariates X_i through

$$\mathrm{logit}(\pi_{ij}) = \alpha_0 + \alpha_y Y_{i,j-1} + \alpha_w X_{i,j-1,1} + \alpha_z Z_i, \tag{7}$$

where the parameter values are the same as in § 4·1. However, when fitting the data, we use the model (6) that depends on the observed surrogate variable W_i. All other aspects of the data generation process are the same as in § 4·1 with the same parameter values, except when one parameter value is changed intentionally to allow us to study its effect.

Table 1. *Results of the simulation study when both the covariates and the measurement errors have nondiagonal covariance matrices with correlation $\rho_x = \rho = 0{\cdot}5$*

Method		Bias	MSE	CVG (%)		Bias	MSE	CVG (%)
			$\sigma_1 = \sigma_2 = 0{\cdot}15$					
Naive	β_{x1}	−0·13	2·07	23·8	β_{x2}	−0·26	7·05	0·9
	β_z	−0·01	1·21	86·5	β_0	−0·80	64·42	0·0
No error adjusted	β_{x1}	−0·04	0·72	89·8	β_{x2}	−0·08	1·20	80·8
	β_z	0·01	1·45	94·9	β_0	0·04	0·96	91·0
No missing adjusted	β_{x1}	−0·08	1·27	67·4	β_{x2}	−0·15	2·87	39·6
	β_z	−0·03	1·50	87·5	β_0	−0·70	50·10	0·0
Our method	β_{x1}	0·00	0·85	94·3	β_{x2}	0·01	0·99	95·0
	β_z	0·01	1·61	94·8	β_0	0·01	0·96	93·7
			$\sigma_1 = \sigma_2 = 0{\cdot}50$					
Naive	β_{x1}	−0·17	3·24	3·0	β_{x2}	−0·34	12·13	0·0
	β_z	−0·01	1·18	85·8	β_0	−0·73	54·64	0·0
No error adjusted	β_{x1}	−0·11	1·56	58·5	β_{x2}	−0·21	4·85	12·3
	β_z	−0·00	1·40	94·1	β_0	0·13	2·48	64·9
No missing adjusted	β_{x1}	−0·08	3·42	70·8	β_{x2}	−0·14	5·48	49·9
	β_z	−0·03	1·64	87·4	β_0	−0·71	52·36	0·0
Our method	β_{x1}	−0·00	1·20	94·8	β_{x2}	0·01	1·44	95·5
	β_z	0·02	1·82	95·0	β_0	0·01	1·13	93·6
			$\sigma_1 = \sigma_2 = 0{\cdot}75$					
Naive	β_{x1}	−0·19	3·92	0·3	β_{x2}	−0·38	14·88	0·0
	β_z	−0·02	1·10	88·0	β_0	−0·70	50·10	0·0
No error adjusted	β_{x1}	−0·14	2·28	30·6	β_{x2}	−0·27	7·55	0·7
	β_z	−0·02	1·32	95·0	β_0	0·17	3·76	43·1
No missing adjusted	β_{x1}	−0·06	15·99	68·0	β_{x2}	−0·08	18·51	50·8
	β_z	−0·02	4·24	86·9	β_0	−0·72	57·23	0·5
Our method	β_{x1}	−0·02	3·16	95·3	β_{x2}	−0·00	2·35	96·1
	β_z	0·01	1·78	95·7	β_0	0·03	5·01	94·2

MSE, 100× mean squared error; (σ_1^2, σ_2^2), the measurement error variances; BIAS, the bias; CVG, the actual coverage of a nominal 95% confidence interval. Naive ignores both missing data and measurement error; No error adjusted accommodates missing data but ignores the measurement error; No missing adjusted adjusts for measurement error but ignores the missing data; Our method accounts for both features.

We performed extensive numerical experiments, but here we discuss just one case, when the true covariates X_{ij1} and X_{ij2} are correlated with common variance $\sigma_x^2 = 1$ and correlation 0·5, and the measurement errors have correlation $\rho = 0{\cdot}5$ and common variance σ_u^2.

First we examined the sensitivity of estimation of β relative to the error degree. In particular, we varied the measurement error variance σ_u^2 from 0·2 to 1·6 to represent a wide range of scenarios. Because the covariate variance $\sigma_x^2 = 1$, when $\sigma_u^2 = 0{\cdot}5$, the corresponding error is already nontrivial, while at $\sigma_u^2 = 1$ the error variance is as large as the covariate variance, i.e., the noise and the signal are roughly the same. For $\sigma_u^2 > 1$, the measurement error dominates the signal, so recovering the information in the covariates is not an easy task. The results on β_{x1} are displayed in Fig. 1(a). It is seen that if the measurement error is not too large, our method possesses a surprising robustness property, in that the average estimate is quite close to the true value. As expected, when the measurement error increases, the bias increases.

We also experimented with different true values for β_{x1} in (5) and α_w in (7). Figures 1(b) and (c) contain the corresponding estimates of β_{x1} and their confidence intervals when the α_w or β_{x1} value changes, while all other parameters are kept at the original values. Results in Fig. 1 clearly suggest that the robustness property we observed does not rely on the specific values of the parameters in either model.

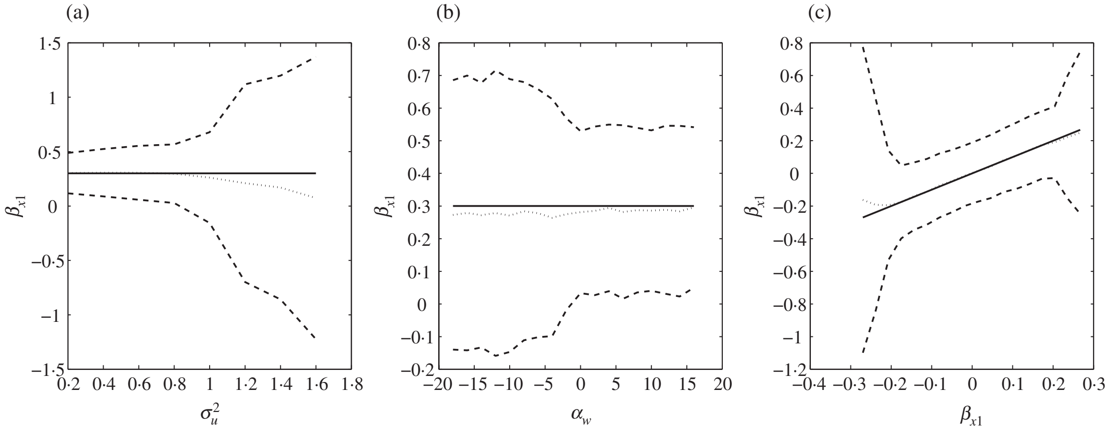

Fig. 1. Sensitivity analysis of our method. The three panels display the average estimates of β_{x1} (dotted curves) and 95% confidence intervals (dashed curves) as functions of σ_u^2 (a), α_w (b) and β_{x1} (c). Solid lines correspond to the true values of β_{x1}.

5. Empirical example

As an illustration, we applied our method to analyse a dataset arising from the Continuing Survey of Food Intake by Individuals (Agricultural Research Service, 1997). The dataset consists of repeated measurements for 1737 individuals with 24 hour recall food intake interviews taken on four different days. Information on age, vitamin A intake, vitamin C intake, total fat intake and total calorie intake is collected at each interview.

Individuals with high levels of fat in their diet have higher risks of outcomes such as obesity and cancers. Let Y be the binary response variable to indicate whether or not an individual's reported percentage of calories exceeds 35%, a threshold that had been previously established, see Food and Nutrition Board (2005, Ch. 8). Here we study how the fat intake changes with age and how it is associated with the intakes of vitamin A and vitamin C.

Often in nutritional epidemiology, 24 hour recalls on caloric intake are treated as missing when the values are physiologically implausible. While there is no universal agreement what values should be taken as implausible (Tooze et al., 2007), Beasley et al. (2008) state that reported intakes of less than 500 calories for women and less than 800 calories for men are implausibly low. Here we took their definition, and this yields about 4% missingness of the Y_{ij} in the data we analysed.

We considered the logistic regression model

$$\text{logit}(Y_{ij}) = \beta_0 + \beta_{x1}X_{ij1} + \beta_{x2}X_{ij2} + \beta_z Z_i \quad (j = 1, \ldots, 4;\ i = 1, \ldots, 1737),$$

where Z_i denotes the age for subject i, and Y_{ij}, X_{ij1} and X_{ij2} represent the response, and intakes of vitamins A and C for subject i at interview j, respectively.

Vitamins A and C are measured with substantial random error. We set W_{ij1} to be the logarithm of 0·005 plus the standardized reported vitamin A intake, W_{ij2} to be similarly defined for reported vitamin C intake and Z_i to be the baseline age in years divided by 100. The transformation on the raw scales of the vitamins A and C values allows us to assume a normal error distribution, with the Kolmogorov–Smirnov test yielding a p-value of around 0·2. However, the study does not have sufficient information to estimate the covariance matrix of the measurement errors directly. To obtain an approximate assessment, we first treated the four measurements of the vitamins A and C intakes as repeated measurements of an average intake value, and obtained that

Table 2. *Results for data analysis of § 5 using (2) as the measurement error structure*

	Naive est	Naive s.e.	Naive p-value	No error adjusted est	No error adjusted s.e.	No error adjusted p-value	No missing adjusted est	No missing adjusted s.e.	No missing adjusted p-value	Our method est	Our method s.e.	Our method p-value
β_0	0·21	0·14	0·14	0·21	0·15	0·14	0·31	0·15	0·04	0·26	0·16	0·09
β_{x1}	0·17	0·03	0·00	0·12	0·03	0·00	0·39	0·07	0·00	0·28	0·07	0·00
β_{x2}	−0·12	0·03	0·00	−0·13	0·03	0·00	−0·31	0·06	0·00	−0·31	0·07	0·00
β_z	0·41	0·39	0·29	0·42	0·39	0·29	0·60	0·40	0·13	0·57	0·41	0·16

est, the estimate; s.e., the standard error of the estimate; p-value, the p-value when testing whether the corresponding coefficient is zero; Naive ignores both missing data and measurement error; No error adjusted accommodates missing data but ignores the measurement error; No missing adjusted adjusts for measurement error but ignores the missing data; Our method accounts for both features.

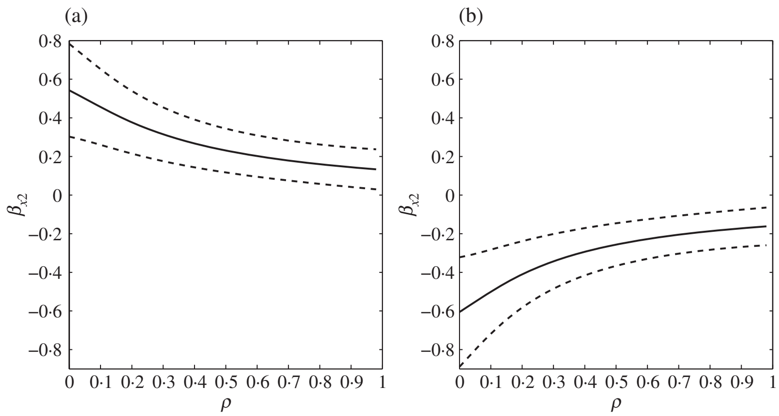

Fig. 2. Sensitivity analyses for the empirical example in § 5. (a) contains estimates of β_{x1} (solid curve) and 95% confidence intervals (dashed curves), displayed as a function of the correlation coefficient ρ. (b) contains the same information for β_{x2}.

$\sigma_1^2 = 0{\cdot}90$, $\sigma_2^2 = 0{\cdot}84$ and their correlation $\rho = 0{\cdot}36$. Considering that the estimated variances we have obtained in fact contain two sources of variability, the variability of the true vitamin intake near the time of the visits and the measurement error variability, yet we do not have sufficient information to separate these two, we decided to allocate half to each, so that the measurement error variances are 0·45 and 0·42. Using this, we performed the corresponding analyses, and the results are reported in Table 2. The analysis shows a significantly positive correlation between vitamin A intake and over-consumption of fat, while this association is negative for vitamin C. Considering that common sources of vitamin A are meat and animal organs, while that of vitamin C are vegetables and fruits, these results are perhaps plausible. The consequence of ignoring the measurement error is attenuation towards zero, while ignoring the missingness seems to result in slightly overestimating the covariate effects.

We also conducted a sensitivity analysis to assess how different degrees of measurement error correlation may affect estimation of the parameters. In this analysis, we fixed the error variances as described above, and let the correlation vary from 0 to 1. Figure 2 shows that the error correlation does have an effect on the quantitative level of the estimation, although its effect is not dramatic and does not alter the qualitative aspects.

6. Extensions

We have emphasized the marginal generalized linear model together with normally distributed measurement error and corrected scores for functional measurement error estimation. However, our basic methodology applies far more generally. For example, suppose that marginally, we propose a parametric model at each time-point for Y_{ij} given (X_{ij}, Z_{ij}) in terms of a parameter $\mathcal{B}$. Suppose further that one has a marginal parametric model for the measurement error in W_{ij} given (X_{ij}, Z_{ij}), in terms of parameters ν_j that are either known or estimated at the $n^{1/2}$-rate: this does not have to be an additive measurement error model, nor does it have to be a normal measurement error model. Then Tsiatis & Ma (2004) show how to construct an unbiased estimating function for $\mathcal{B}$ based on the observed data, generically denoted here as $\mathcal{K}(Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B}, \nu_1, \ldots, \nu_m)$. Replacing $\mathcal{G}^*_{ij}(Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B})$ with $\mathcal{K}(Y_{ij}, W_{ij}, Z_{ij}, \mathcal{B}, \nu_1, \ldots, \nu_m)$, all our developments in § 3 carry through. Hall & Ma (2007) show how to extend this to the case that the error model is additive but estimated nonparametrically through replication. Even more generally, as long as one can construct a marginal, unbiased estimating function, then our approach applies. Furthermore, if a large number of covariates are present and there is need to select a subset from it, the method in Ma & Li (2010) can be used to perform variable selection and estimation simultaneously.

Although extensive research has been directed to analysis of longitudinal data with either missing values or measurement error, those methods cannot be immediately applied to handle data with both features. Simultaneously, addressing missingness and measurement error is more challenging because these two characteristics could interactively affect inference about response parameters. Our method has applications in a broad variety of settings. It can be directly applied to deal with correlated data, such as clustered data and multivariate data, with missing observations and covariate measurement error. Our method can also be readily modified to handle data with more complex association structures, such as longitudinal data arising in clusters or longitudinal multivariate data.

Acknowledgement

The authors thank the referees for their helpful comments. Yi's research was supported by the Natural Sciences and Engineering Research Council of Canada. Ma's research was supported by the National Science Foundation and the National Institute of Neurological Disorders and Stroke. Carroll's research was supported by the National Cancer Institute, the National Institute of Neurological Disorders and Stroke and the King Abdullah University of Science and Technology.

Appendix

A1. *Proof of Lemma 1*

We consider a simple counterexample where $m = 2$ and there is no Z. Suppose that $Y_{ij} = X_{ij}\beta + \epsilon_{ij}$, $j = 1, 2$, $(\epsilon_{i1}, \epsilon_{i2})^{\mathrm{T}} \sim N(0, \Sigma_\epsilon)$ and $(\epsilon_{i1}, \epsilon_{i2})^{\mathrm{T}}$ is independent of $(X_{i1}, X_{i2})^{\mathrm{T}}$. Then the Type I assumption holds for the regression mean of Y_{ij} on (X_{i1}, X_{i2}), $j = 1, 2$. Suppose further that $(X_{i1}, X_{i2}) \sim N(0, I)$ where I is the 2×2 identity matrix, and that $(W_{i1}, W_{i2})^{\mathrm{T}} = (X_{i1}, X_{i2})^{\mathrm{T}} + (e_{i1}, e_{i2})^{\mathrm{T}}$ where $(e_{i1}, e_{i2}) \sim N(0, \Sigma_e)$ with Σ_e being not diagonal, and $(X_{i1}, X_{i2})^{\mathrm{T}}$ and $(e_{i1}, e_{i2})^{\mathrm{T}}$ are independent. Then it is easily seen that as long as Σ_e is not diagonal and $\beta \neq 0$, the Type I assumption fails for the observed data. For example, since X_{ij} has mean zero, $\mathrm{cov}(Y_{i1}, Y_{i2}) = \beta^2 I + \Sigma_\epsilon$, $\mathrm{cov}(W_{i1}, W_{i2}) = I + \Sigma_e$ and $\mathrm{cov}\{(Y_{i1}, Y_{i2}), (W_{i1}, W_{i2})\} = \beta I$. The joint covariance matrix of $(Y_{i1}, Y_{i2}, W_{i1}, W_{i2})$ is

$$\Sigma = \begin{pmatrix} \beta^2 I + \Sigma_\epsilon & \beta I \\ \beta I & I + \Sigma_e \end{pmatrix}.$$

Hence, $E\{(Y_{i1}, Y_{i2})^{\mathrm{T}} \mid (W_{i1}, W_{i2})\} = \beta(I + \Sigma_e)^{-1}(W_{i1}, W_{i2})^{\mathrm{T}}$. If the diagonal elements of Σ_e both equal 1, and if the correlation coefficient of e_{i1} and e_{i2} is ρ_e, then the regression mean of Y_{i1} on (W_{i1}, W_{i2}) is $\beta(2W_{i1} - \rho_e W_{i2})/(4 - \rho_e^2)$, not a function of W_{i1} alone. This proves Lemma 1.

A2. *Error effect on the model for the missing data process*

Lemma 2 follows from the following example that the underlying missingness process is missing-at-random in X_i but not in W_i. There are no Z_i covariates. Consider a case with drop-outs which is modelled by a probit regression model

$$\mathrm{pr}(R_{ij} = 1 \mid Y_i, X_i) = F(\alpha_0 + \alpha_y Y_{i,j-1} + \alpha_x X_{ij}), \tag{A1}$$

where $F(\cdot)$ is the standard normal cumulative distribution function. Conditional on Y_i and X_i, the R_{ij} are assumed to be independent. Assume that conditional on X_i, the Y_{ij} are independent having a $N(\beta_0 + \beta_x X_{ij}, \sigma^2)$ distribution, and that the W_{ij} are independent having a $N(X_{ij}, \sigma_e^2)$ distribution. Assume that the X_{ij} are independent having a marginal distribution $N(\mu_x, \sigma_x^2)$.

Now we show that although (A1) corresponds to missing at random if X_i were available, the conditional distribution of the missing data indicator given the observed surrogate W_i is not missing-at-random. Indeed, with the assumption analogous to nondifferential error that $\mathrm{pr}(R_{ij} = 1 \mid Y_i, X_i, W_i) = \mathrm{pr}(R_{ij} = 1 \mid Y_i, X_i)$, we obtain

$$\begin{aligned}\mathrm{pr}(R_{ij} = 1 \mid Y_i, W_i) &= \int \mathrm{pr}(R_{ij} = 1 \mid Y_i, X_i) f(X_i \mid Y_i, W_i)\, dX_i \\ &= \int F(\alpha_0 + \alpha_y Y_{i,j-1} + \alpha_x X_{ij}) f(X_{ij} \mid Y_{ij}, W_{ij})\, dX_{ij}.\end{aligned}$$

Since $X_{ij} \mid (Y_{ij}, W_{ij}) \sim N(a + bY_{ij} + cW_{ij}, d^2)$ for some constants a, b, c, d that are determined by $\beta_0, \beta_x, \sigma_x^2, \sigma_e^2, \sigma^2$ and μ_x, then using the identity $\int (1/\gamma) F(\delta + u)\phi(u/\gamma)\, du = F\{\delta/(1 + \gamma^2)^{1/2}\}$ for constants γ and δ, where $\phi(\cdot)$ is the standard normal density function, we can show that $\mathrm{pr}(R_{ij} = 1 \mid Y_i, W_i) = F[\{\alpha_0 + \alpha_y Y_{i,j-1} + \alpha_x(a + bY_{ij} + cW_{ij})\}/(1 + \alpha_x^2 d^2)^{1/2}]$. The dependence of $\mathrm{pr}(R_{ij} = 1 \mid Y_i, W_i)$ on the response measurement Y_{ij} at time-point j shows that the missingness probability depends on a future unobserved response component, suggesting that a missing-at-random structure is not true for the conditional distribution $\mathrm{pr}(R_{ij} = 1 \mid Y_i, W_i)$.

A3. *Sketch of the arguments for the asymptotic theory*

Let $H_i(\mathcal{B}, \alpha) = \{\Psi_i^{\mathrm{T}}(\mathcal{B}, \alpha), S_i^{\mathrm{T}}(\alpha)\}^{\mathrm{T}}$. Then $E\{H_i(\mathcal{B}, \alpha)\} = 0$. By Newey & McFadden (1993, Theorem 3.4), under standard regularity conditions there is a unique solution $(\hat{\mathcal{B}}, \hat{\alpha})$ to the equation $\sum_{i=1}^n H_i(\mathcal{B}, \alpha) = 0$ with probability approaching 1, that satisfies $n^{1/2}\{(\hat{\mathcal{B}} - \mathcal{B})^{\mathrm{T}}, (\hat{\alpha} - \alpha)^{\mathrm{T}}\}^{\mathrm{T}} = -[E\{\partial H_i(\mathcal{B}, \alpha)/\partial(\mathcal{B}^{\mathrm{T}}, \alpha^{\mathrm{T}})\}]^{-1} n^{-1/2} \sum_{i=1}^n H_i(\mathcal{B}, \alpha) + o_p(1)$. For the estimator $\hat{\mathcal{B}}$ of central interest, we have $n^{1/2}(\hat{\mathcal{B}} - \mathcal{B}) = -\Gamma^{-1} n^{-1/2} \sum_{i=1}^n Q_i(\mathcal{B}, \alpha) + o_p(1)$. The central limit theorem then leads to the asymptotic distribution for $n^{1/2}(\hat{\mathcal{B}} - \mathcal{B})$.

REFERENCES

AGRICULTURAL RESEARCH SERVICE (1997). *Design and Operation: The Continuing Survey of Food Intakes by Individuals and the Diet and Health Knowledge Survey, 1994–96*. US Department of Agriculture, Hyattsville, MD.

BEASLEY, J. M., RILEY, W. T., DAVID, A. & SINGH, J. (2008). Evaluation of a PDA-based dietary assessment and intervention program: a randomized controlled trial. *J. Am. College Nutr.* **27**, 280–86.

CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. & CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*, 2nd ed. New York: Chapman & Hall/CRC Press.

COOK, J. & STEFANSKI, L. A. (1995). A simulation extrapolation method for parametric measurement error models. *J. Am. Statist. Assoc.* **89**, 1314–28.

FOOD AND NUTRITION BOARD (2005). *Dietary Reference Intakes for Energy, Carbohydrate, Fiber, Fat, Fatty Acids, Cholesterol, Protein, and Amino Acids (Macronutrients)*. Washington, D.C.: The National Academies Press.

Hall, P. & Ma, Y. (2007). Measurement error models with unknown error structure. *J. R. Statist. Soc.* B **69**, 429–46.
Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029–54.
Huang, Y. & Wang, C. Y. (2001). Consistent functional methods for logistic regression with errors in covariates. *J. Am. Statist. Assoc.* **96**, 1469–82.
Lai, T. L. & Small, D. (2007). Marginal regression analysis of longitudinal data with time-dependent covariates: a generalized method of moments approach. *J. R. Statist. Soc.* B **69**, 79–99.
Li, E., Zhang, D. & Davidian, M. (2004). Conditional estimation for generalized linear models when covariates are subject-specific parameters in a mixed model for longitudinal measurements. *Biometrics* **60**, 1–7.
Li, E., Wang, N. & Wang, N.-Y. (2007). Joint models for a primary endpoint and multiple longitudinal covariate processes. *Biometrics* **63**, 1068–78.
Liang, H. (2009). Generalized partially linear mixed-effects models incorporating mismeasured covariates. *Ann. Inst. Statist. Math.* **61**, 27–46.
Lin, X. & Carroll, R. J. (2006). Semiparametric estimation in general repeated measures problems. *J. R. Statist. Soc.* B **68**, 68–88.
Liu, W. & Wu, L. (2007). Simultaneous inference for semiparametric nonlinear mixed-effects models with covariate measurement errors and missing responses. *Biometrics* **63**, 342–50.
Ma, Y. & Li, R. (2010). Variable selection in measurement error models. *Bernoulli* **16**, 274–300.
Nakamura, T. (1990). Corrected score functions for errors-in-variables models: methodology and application to generalized linear models. *Biometrika* **77**, 127–37.
Newey, W. K. & McFadden, D. (1993). Estimation in large samples. In *Handbook of Econometrics*, vol. 4. Ed. D. McFadden and R. Engler. Amsterdam: North-Holland.
Palta, M. & Lin, C. Y. (1999). Latent variables, measurement error and methods for analysing longitudinal binary and ordinal data. *Statist. Med.* **18**, 385–96.
Pan, W., Zeng, D. & Lin, X. (2009). Estimation in semiparametric transition measurement error models for longitudinal data. *Biometrics* **65**, 728–36.
Pepe, M. S. & Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Commun. Statist.* B **23**, 939–51.
Pepe, M. S. & Couper, D. (1997). Modeling partly conditional means with longitudinal data. *J. Am. Statist. Assoc.* **92**, 991–8.
Stefanski, L. A. (1989). Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Commun. Statist.* A **18**, 4335–58.
Stefanski, L. A. & Carroll, R. J. (1987). Conditional scores and optimal scores in generalized linear measurement error models. *Biometrika* **74**, 703–16.
Tooze, J. A., Vitolins, M. Z., Smith, S. L., Arcury, T. A., Davis, C. C., Bell, R. A., DeVellis, R. F. & Quandt, S. A. (2007). High levels of low energy reporting on 24-hour recalls and three questionnaires in an elderly low-socioeconomic status population. *J. Nutr.* **137**, 1286–93.
Tsiatis, A. A. & Ma, Y. (2004). Locally efficient semiparametric estimators for functional measurement error models. *Biometrika* **91**, 835–48.
Wang, C. Y., Huang, Y., Chao, E. C. & Jeffcoat, M. K. (2008). Expected estimating equations for missing data, measurement error, and misclassification, with application to longitudinal nonignorable missing data. *Biometrics* **64**, 85–95.
Wang, N., Lin, X., Gutierrez, R. G. & Carroll, R. J. (1998). Generalized linear mixed measurement error models. *J. Am. Statist. Assoc.* **93**, 249–61.
Xiao, Z., Shao, J. & Palta, M. (2010). GMM in linear regression for longitudinal data with multiple covariates measured with error. *J. Appl. Statist.* **37**, 791–805.
Yi, G. Y. (2005). Robust methods for incomplete longitudinal data with mismeasured covariates. *Far East J. Theor. Statist.* **16**, 205–34.
Yi, G. Y. (2008). A simulation-based marginal method for longitudinal data with dropout and mismeasured covariates. *Biostatistics* **9**, 501–12.
Yi, G. Y., Liu, W. & Wu, L. (2011). Simultaneous inference and bias analysis for longitudinal data with covariate measurement error and missing responses. *Biometrics* **67**, 67–75.
Zhou, Y. & Liang, H. (2009). Statistical inference for semiparametric varying-coefficient partially linear models with error-prone linear covariates. *Ann. Statist.* **37**, 427–58.

[*Received February* 2011. *Revised October* 2011]