Likelihood inference in generalized linear mixed measurement error models

Mahmoud Torabi

Department of Community Health Sciences, University of Manitoba, MB, R3E 0W3, Canada

Abstract

The generalized linear mixed models (GLMMs) for clustered data are studied when covariates are measured with error. The most conventional measurement error models are based on either linear mixed models (LMMs) or GLMMs. Even without the measurement error, the frequentist analysis of LMM, and particularly of GLMM, is computationally difficult. On the other hand, Bayesian analysis of LMM and GLMM is computationally convenient in both cases without and with the measurement error. Recent introduction of the method of data cloning has made frequentist analysis of mixed models also equally computationally convenient. As an application of data cloning, we conduct a frequentist analysis of GLMM with covariates subject to measurement error model. The performance of the proposed approach which yields to maximum likelihood estimation is evaluated by two important real data types, Normal and logistic linear mixed measurement error models, and also through simulation studies.

Keywords: Bayesian computation; Exponential family; Hierarchical

Email address: `torabi@cc.umanitoba.ca; Fax: +1-204-789-3905` (Mahmoud Torabi)

models; Measurement error; Random effects

1. Introduction

Generalized linear mixed models (GLMMs) have become very popular for analyzing correlated and overdispersed data (Breslow and Clayton, 1993). Depending on the data, either linear mixed models (LMMs) (Searle et al., 1992; Torabi et al., 2009; Datta et al., 2010; Torabi, 2011, 2012a) or GLMMs (McCulloch et al., 2008) are used. Parameters of the LMM can be estimated using either maximum likelihood (ML) or restricted ML (REML), among other approaches. A potential difficulty in making inference in GLMMs is that a full-likelihood analysis is burdened by often intractable numerical integration (McCulloch, 1997). Parameter estimation under GLMM is then extremely difficult using the frequentist approach. The Bayesian approach, especially the non-informative Bayesian approach, has become quite popular because of its computational convenience. Implementation of the non-informative Bayesian approach, however, requires substantial care. The inferences may also depend on the choice of the prior.

A common problem for analyzing correlated data is also the presence of covariate subject to measurement error (Carroll et al., 2006). The measurement error with independent observations has been extensively reviewed in the literature in the context of linear models (Fuller, 1987). The GLMMs and non-linear models with covariates subject to measurement error have been also studied in the literature (Carroll et al., 2006). In frequentist approach, there are a few studies on the effects of measurement error to analyze the clustered data which are mainly based on some approximation approaches such as regression calibration (RCA) and simulation based such as simulation extrapolation (SIMEX) (Wang et al., 1998). However, likewise GLMMs, the Bayesian approach especially the non-informative Bayesian approach, has become also quite popular in generalized linear mixed measurement error models (GLMMeMs) because of its computational convenience (Carroll et al., 2006).

Recently, Lele et al. (2007) introduced an alternative approach, called data cloning (DC), to compute the ML estimates and their standard errors for general hierarchical models. Similar to the Bayesian approach, data cloning avoids high dimensional numerical integration and requires neither maximization nor differentiation of a function. Extending this work to GLMM situation, Lele et al. (2010) described an approach to compute prediction and prediction intervals for the random effects. Torabi (2012b) also extended the DC approach to the GLMM with two components of dispersion. The data cloning approach was also extended to cross-sectional and time-series data in the context of small area estimation (Torabi and Shokoohi, 2012).

The data cloning approach, thus, is well suited to offer a frequentist approach in the context of measurement error. The advantages of DC are that the model parameters estimate are independent of the choice of priors, non-estimable parameters are flagged automatically and possibility of improper posterior distribution is completely avoided.

In this paper, we use data cloning in the context of GLMMs with covariates subject to measurement error. We first describe the measurement error problem in general (Section 2). We, then, describe how data cloning can be used to analyze GLMMeMs (Section 3). In Section 4, we compare the performance of data cloning which yields to maximum likelihood estimation (MLE) with the non-informative Bayesian approach for two real datasets when the responses are Normal and binary. In Section 5, the performance of data cloning is also studied through two simulation studies. Finally, some concluding remarks are given in Section 6.

2. Generalized linear mixed measurement error model

The basic model in measurement error can be described as follows. Let y_{ij} be the variable of interest for the jth unit within ith cluster $(i = 1, ..., m; j = 1, ..., n_i)$. The y_{ij} are assumed to be conditionally independent with exponential family p.d.f.

$$f(y_{ij}|\theta_{ij}, \phi_{ij}) = \exp\{(y_{ij}\theta_{ij} - c(\theta_{ij}))/\phi_{ij} + d(y_{ij}, \phi_{ij})\}, \tag{2.1}$$

$(i = 1, ..., m; j = 1, ..., n_i)$. The density (2.1) is parameterized with respect to the canonical parameters θ_{ij}, known scale parameters ϕ_{ij} and functions $c(\cdot)$ and $d(\cdot)$. The exponential family (2.1) covers well-known distributions including Normal, binomial and Poisson distributions. The GLMM of y_{ij} given $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ is constructed by assuming that the conditional mean θ_{ij} is related to $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ through a generalized linear model,

$$g(\theta_{ij}) = \beta_0 + \boldsymbol{x}'_{ij}\boldsymbol{\beta}_x + \boldsymbol{z}'_{ij}\boldsymbol{\beta}_z + u_i \ (i = 1, ..., m; j = 1, ..., n_i), \tag{2.2}$$

where $g(\cdot)$ is a monotonic differentiable link function, $\boldsymbol{x}_{ij}(p_1 \times 1)$ are unobserved true covariates, $\boldsymbol{z}_{ij}(p_2 \times 1)$ are observed covariates, $\beta_0, \boldsymbol{\beta}_x(p_1 \times 1)$, and $\boldsymbol{\beta}_z(p_2 \times 1)$ are vector of unknown regression coefficients, and u_i are random effects with $u_i \overset{i.i.d.}{\sim} N(0, \sigma_u^2)$. We also need to define the measurement error structure. In general, there are two different modeling approaches that differ according to distributional assumptions made on error-prone covariates: structural and functional approaches to modeling GLMMeMs. In structural approach, the specific assumptions are made about the distributional structure of the unobserved covariates, while in functional approach nothing is assumed about the unobserved covariates.

The most convenient structure is additive error, so that

$$\boldsymbol{X}_{ij} = \boldsymbol{x}_{ij} + \boldsymbol{V}_{ij}, \tag{2.3}$$

where $\boldsymbol{V}_{ij}$ are independent $N(0, \Sigma_{vv})$ which are also independent of $\boldsymbol{x}_{ij}$. In this paper, we consider the structural approach to modeling GLMMeMs which $\boldsymbol{x}_{ij}$ are not observed, but $\boldsymbol{X}_{ij}$ are observed instead, with $\boldsymbol{x}_{ij} \sim N(\mu_x, \Sigma_{xx})$. When $\boldsymbol{X}$ and $\boldsymbol{x}$ are scalar, we write the measurement variance Σ_{vv} simply as σ_v^2. Define $\boldsymbol{y}_i = (y_{i1}, ..., y_{in_i})^T, \boldsymbol{x}_i = (\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{in_i})^T$, and similarly for $\boldsymbol{z}_i, \boldsymbol{X}_i$, and $\boldsymbol{V}_i$. The joint integrated likelihood in the ith cluster is

$$L_i(\boldsymbol{y}_i, \boldsymbol{X}_i|\boldsymbol{z}_i) = \int L_i(\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{z}_i)L_i(\boldsymbol{X}_i|\boldsymbol{x}_i, \boldsymbol{z}_i)L_i(\boldsymbol{x}_i|\boldsymbol{z}_i)d\boldsymbol{x}_i,$$

where $L_i(\boldsymbol{x}_i|\boldsymbol{z}_i)$ is the likelihood function of $\boldsymbol{x}_i$ and $L_i(\boldsymbol{X}_i|\boldsymbol{x}_i, \boldsymbol{z}_i)$ is the error distribution, which is often assumed to be independent of $\boldsymbol{z}_i$ as in (2.3). The dependence of the likelihood on the within-cluster conditional distribution of the unobserved $\boldsymbol{x}$'s leads to issue of model robustness. In particular, the problem of model misspecification may occur in most applications including but not limited to LMM or GLMM and it involves not only the distribution of the unknown $\boldsymbol{x}$, but also a more general problem starting with linear model in $\boldsymbol{y}$ and involving all the components of the likelihood function (Roeder et al., 1996; Carroll et al., 1999a, 1999b; Guolo, 2008a, 2008b).

It has been shown that ignoring the measurement error may result in misspecifying the structure of both fixed and random effects. The cluster size n_i also plays an important role in the asymptotic bias in MLE under a misspecified model as $m \to \infty$ (Wang et al., 1998). Suppose that x_{ij} is scalar, and define the vector $\boldsymbol{x}_i = (x_{i1}, ..., x_{in_i})'$, and $\boldsymbol{z}_i, \boldsymbol{X}_i, \boldsymbol{V}_i$ are defined similarly, and also assume that

$$\boldsymbol{x}_i = \mathbf{1}_{n_i}\eta_0 + \boldsymbol{z}_i\boldsymbol{\eta}_z + \boldsymbol{e}_{xi},$$

where $\mathbf{1}_{n_i}$ is an $n_i \times 1$ vector of ones and $\boldsymbol{e}_{xi}$ given $\boldsymbol{z}_i$ is Normal variate with mean 0 and variance-covariance Σ_{xxi}. We also define the reliability matrix by $\Lambda_i = \Sigma_{xxi}\{\Sigma_{xxi} + cov(\boldsymbol{V}_i)\}^{-1}$. Then

$$\boldsymbol{x}_i = (\boldsymbol{I}_i - \Lambda_i)(\mathbf{1}_{n_i}\eta_0 + \boldsymbol{z}_i\eta_z) + \Lambda_i\boldsymbol{X}_i + \mathbf{1}_{n_i}u_i^*,$$

or

$$x_{ij} = \alpha_{0j} + \boldsymbol{\eta}_z^{'}\boldsymbol{z}_i^{'}\boldsymbol{\alpha}_{zj} + \boldsymbol{X}_i^{'}\boldsymbol{\alpha}_{wj} + u_i^*,$$

for some $\alpha_{0j}, \boldsymbol{\alpha}_{zj}, \boldsymbol{\alpha}_{wj}$, and u_i^* is independent of u_i and $\boldsymbol{X}_i$ (Wang et al., 1998). It then follows from (2.2) that

$$g(\theta_{ij}) = \beta_0 + (\alpha_{0j} + \boldsymbol{X}_i'\boldsymbol{\alpha}_{wj} + \boldsymbol{\eta}_z^{'}\boldsymbol{z}_i^{'}\boldsymbol{\alpha}_{zj} + u_i^*)\beta_x + z_{ij}\beta_z + u_i. \tag{2.4}$$

It is clear by comparing (2.2) and (2.4) that the variance structure of the x_{ij} plays an important role when properly handling a GLMMeM.

3. Inference using data cloning

Let $\boldsymbol{y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_m)'$ be the observed data vector and, conditionally on the random effects, $\boldsymbol{b}$, assume that the elements of $\boldsymbol{y}$ are independent and drawn from a distribution in exponential family with parameters $\boldsymbol{\alpha}_1$. It is also assumed that distribution for $\boldsymbol{b}$ depends on parameters $\boldsymbol{\alpha}_2$:

$$\begin{aligned} \boldsymbol{y}_i|\boldsymbol{b} &\sim f_{\boldsymbol{y}_i|\boldsymbol{b}}(\boldsymbol{y}_i|\boldsymbol{b}, \boldsymbol{\alpha}_1) \\ \boldsymbol{b} &\sim g_{\boldsymbol{b}}(\boldsymbol{b}|\boldsymbol{\alpha}_2). \end{aligned} \tag{3.1}$$

The goal of the analysis is to estimate the model parameters $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)'$ and predict the random effects $\boldsymbol{b}$ or its function. The likelihood of $\boldsymbol{y}_i$ given $\boldsymbol{\alpha}$ is given by

$$L(\boldsymbol{\alpha}; \boldsymbol{y}) = \int \prod_{i=1}^{m} f_{\boldsymbol{y}_i|\boldsymbol{b}}(\boldsymbol{y}_i|\boldsymbol{b}, \boldsymbol{\alpha}_1) g_{\boldsymbol{b}}(\boldsymbol{b}|\boldsymbol{\alpha}_2) d\boldsymbol{b}.$$

To illustrate the DC approach, we start with standard Bayesian approach to inference for hierarchical models. Denote $\pi(\boldsymbol{\alpha})$ as prior distribution on the parameter space. The posterior distribution $\pi(\boldsymbol{\alpha}|\boldsymbol{y})$ is given by

$$\pi(\boldsymbol{\alpha}|\boldsymbol{y}) = \frac{L(\boldsymbol{\alpha}; \boldsymbol{y})\pi(\boldsymbol{\alpha})}{C(\boldsymbol{y})}, \tag{3.2}$$

where $C(\boldsymbol{y}) = \int L(\boldsymbol{\alpha}; \boldsymbol{y})\pi(\boldsymbol{\alpha}) d\boldsymbol{\alpha}$ is the normalizing constant. There are computational tools, Markov chain Monte Carlo (MCMC) algorithms, that facilitate generation of random variates from the posterior distribution $\pi(\boldsymbol{\alpha}|\boldsymbol{y})$ without computing the integrals in the numerator or the denominator of (3.2)(Gilks et al., 1996; Spiegelhalter et al., 2004).

The DC method uses the Bayesian computational approach for frequentist purposes. In DC, one pretends that the observations $\boldsymbol{y}$ are repeated independently by K different individuals and all these individuals obtain exactly the same set of observations $\boldsymbol{y}$ called $\boldsymbol{y}^{(K)} = (\boldsymbol{y}, \boldsymbol{y}, ..., \boldsymbol{y})$. The posterior distribution of $\boldsymbol{\alpha}$ conditional on the data $\boldsymbol{y}^{(K)}$ is then given by

$$\pi_K(\boldsymbol{\alpha}|\boldsymbol{y}^{(K)}) = \frac{\{L(\boldsymbol{\alpha}; \boldsymbol{y})\}^K \pi(\boldsymbol{\alpha})}{C(\boldsymbol{y}^{(K)})}, \tag{3.3}$$

where $C(\boldsymbol{y}^{(K)}) = \int \{L(\boldsymbol{\alpha}; \boldsymbol{y})\}^K \pi(\boldsymbol{\alpha}) d\boldsymbol{\alpha}$ is the normalizing constant. The expression $\{L(\boldsymbol{\alpha}; \boldsymbol{y})\}^K$ is the likelihood for K copies of the original data. Lele et al. (2007, 2010) showed that, for K large enough, $\pi_K(\boldsymbol{\alpha}|\boldsymbol{y}^{(K)})$ converges to a multivariate Normal distribution with mean equal to the MLE of the model parameters and variance-covariance matrix equal to $1/K$ times the inverse of the Fisher information matrix for the MLE. Hence, this distribution is nearly degenerated at the MLE $\boldsymbol{\alpha}$ for large K. Moreover, the sample mean vector of the generated random numbers from (3.3) provides the MLE of the model parameters, and K times their sample variance-covariance matrix is an estimate of the asymptotic variance-covariance matrix for the MLE $\hat{\boldsymbol{\alpha}}$. Lele et al. (2010) also provided various checks to determine the adequate number of clones K. For instance, one may plot the largest eigenvalue of the posterior variance as a function of the number of clones K to determine if the posterior distribution has become nearly degenerate. As another criterion, it is approximately true that with increasing number of clones, we have

$$(\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}})' \boldsymbol{V}^{-1} (\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}) \sim \chi^2_p, \tag{3.4}$$

where $\boldsymbol{V}$ is the variance of the posterior distribution and p is the dimension of $\boldsymbol{\alpha}$. One may compute the following two statistics: a) $\zeta = \frac{1}{B} \sum_{b=1}^{B} (O_b - E_b)^2$, where O_b and E_b are observed and estimated quantiles for χ^2_p random variable, and b) $\tilde{r}^2 = 1 - \rho^2$, where ρ is the correlation between (O_b, E_b).

If these statistics are close to zero, it indicates that the approximation (3.4) is reasonable.

3.1. Prediction of random effects

Prediction of random effects, particularly from the frequentist viewpoint, is usually problematic. If the parameters $\boldsymbol{\alpha}$ are known, then one can clearly use the conditional distribution of $\boldsymbol{b}$, the latent variables, given the observed data. That is, one can use $\pi(\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\alpha}^*)$ where $\boldsymbol{\alpha}^*$ is the true value of the parameter. A naive approach, when $\boldsymbol{\alpha}$ is estimated using the data, is to use $\pi(\boldsymbol{b}|\boldsymbol{y}, \hat{\boldsymbol{\alpha}})$. However, this approach does not take into account the variability introduced by the model parameter estimates. An approach that has been suggested in the literature (e.g., Hamilton, 1986; Lele et al., 2010) to take into account the variation of the estimator is to use the density:

$$\pi(\boldsymbol{b}|\boldsymbol{y}) = \frac{\int f(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\alpha}_1) g(\boldsymbol{b}|\boldsymbol{\alpha}_2) \phi(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}, I^{-1}(\hat{\boldsymbol{\alpha}})) d\boldsymbol{\alpha}}{C(\boldsymbol{y})}, \tag{3.5}$$

where $\phi(., \mu, \sigma^2)$ denotes Normal density with mean μ and variance σ^2, which are equal to the MLE and the inverse of the Fisher information matrix here. In this paper, we obtain the prediction of $\boldsymbol{b}$ using the density in equation (3.5) along with MCMC sampling.

In this paper, the performance of DC and hierarchical Bayes (HB) is evaluated through data analysis and simulation studies. In general, the vague, but proper, priors are used for regression effects and variance components. In particular, the independent Normal distribution is assigned for regression coefficient with zero mean and variance 10^6, and gamma distribution for the inverse of variance component with shape and scale parameter 0.001. To monitor the convergence of the model parameters, we use several diagnostic methods implemented in the Bayesian output analysis (BOA) program (Smith, 2007), a freely available package created for R. We also use diagnostic methods described in Section 3, implemented in dclone package (Sólymos, 2010) in R, to monitor the convergence of the model parameters in terms of number of clones (K).

4. Data analysis

4.1. OPEN study

We study the performance of data cloning approach by applying to a real dataset from the National Cancer Institute's OPEN Study, which is one of the largest biomarker studies ever done (Subar et al., 2001, 2003; Kipnis et al., 2003). In the National Cancer Institute's OPEN Study, one interest is to measure the logarithm of dietary protein intake. However, true long-term log-intake x can not be observed in practice. Instead, a biomarker of log-protein intake X, namely urinary nitrogen, is measured. In this study, $m = 223$ subjects had replicated urinary nitrogen measurements. There is evidence from feeding studies that the protein biomarker captures true protein intake with added variability. Then, the classical measurement error model (2.3) holds in this context. Let x_i be the true log-protein intake for individual i, and let X_{ij} be the jth biomarker log-protein measurement for individual i. The classical measurement error model is given by

$$X_{ij} = x_i + V_{ij}. \tag{4.1}$$

Each individual completed up to two ($n_i = 2$) food frequency questionnaires (FFQ) which measured reported protein intake, and also up to two biomarkers for protein intake. Let y_{ij} denotes the logarithm of FFQ, the model is then given by

$$y_{ij} = \beta_0 + \beta_1 x_i + u_i + \epsilon_{ij}, \qquad (i = 1, ..., m; j = 1, ..., n_i), \tag{4.2}$$

where $u_i \sim N(0, \sigma_u^2), \epsilon_{ij} \sim N(0, \sigma_\epsilon^2), V_{ij} \sim N(0, \sigma_v^2), x_i \sim N(\mu_{x_i}, \sigma_x^2)$ with $\mu_{x_i} = \alpha_0 + \alpha_1 Age_i + \alpha_2 BMI_i$ where Age_i and BMI_i are age and body mass index (bmi) of individual i.

Table 1 shows the estimates of the model parameters by employing MLE and HB approaches. It seems that for some model parameters the MLE provides better results than HB in terms of efficiency, noting that the MLE results do not also depend on the choice of priors unlike the HB approach. In particular, the relative efficiency of MLE compared to HB, $var(\hat{\boldsymbol{\alpha}}_{HB})/var(\hat{\boldsymbol{\alpha}}_{MLE})$, ranges from %100 to %291. For this specific application, we used $K = 40$ to obtain MLE, and the number of iterations for convergence of the model parameters was $15,000$.

4.2. Framingham Heart Study

We also study the performance of data cloning by applying to a real dataset (Framingham Heart Study) which has been studied by Kannel et al. (1986), among others. Note that the Framingham Heart Study is a logistic measurement error model without random effects, as indicated later in this section, which falls in the class of generalized linear measurement error model. However, since this study is very popular in the context of measurement error,

Table 1: The estimates (and standard errors) of the model parameters in OPEN Study for MLE and HB methods.

Parameter	MLE	HB
β_0	2.45(0.867)	1.57(1.472)
β_1	0.48(0.166)	0.65(0.283)
α_0	5.72(0.367)	5.69(0.367)
α_1	-0.01(0.006)	-0.01(0.006)
α_2	-0.003(0.008)	-0.002(0.008)
σ_x^2	0.27(0.052)	0.23(0.058)
σ_u^2	0.33(0.088)	0.22(0.110)
σ_ϵ^2	0.87(0.083)	0.93(0.100)
σ_v^2	0.46(0.044)	0.50(0.056)

we decided to analyze it using the data cloning approach; noting that more complex models (GLMMeMs) will be studied with an extensive simulation study in Section 5.2.

The Framingham Study is a large study following individuals for the development of coronary heart disease (CHD). It consists of a series of exams taken over two years. We use exam number 3 as the baseline. There are $m = 1615$ men aged between 31 to 65 in this dataset, with the outcome, y, indicating the occurrence of CHD within an eight-year period following exam 3; there were 128 cases of CHD. Predictors employed in this example are the patient's age at exam 2, smoking status at exam 1, serum cholesterol at exam 3, and systolic blood pressure (SBP), the last is the average of two measurements taken by different examiners during the same visit. In this analysis, the error-free covariates z, are age, smoking status, and serum cholesterol. The analysis uses the replicate SBP measurements from exams 2 and 3 for all study participants. The main surrogate X is the measurement of $\log(SBP-50)$. In particular, the transformed data are X_{ij}, where i denotes the individual and $j=1,2$ refers to the transformed SBP at exams 2 and 3, respectively. The overall surrogate is the sample mean for each individual. The model is then given by

$$logit(p_i)=\beta_1\text{Age}_i+\beta_2\text{Smoke}_i+\beta_3\text{Chol}_i+\beta_4 x_i,$$

$$X_{ij}=x_i+V_{ij},$$

where $p_i = Pr(y_i = 1|x_i, z_i), x_i \sim N(\mu_x, \sigma_x^2)$ and $X_{ij}|x_i \sim N(x_i, \sigma_v^2)$. The model parameters in this case are $\boldsymbol{\alpha} = (\beta_1, \beta_2, \beta_3, \beta_4, \sigma_x^2, \sigma_v^2)'$. We then estimate the model parameters and corresponding standard errors by MLE through data cloning as well as HB method. For this specific application, the number of clones was $K = 20$ to obtain MLE with number of iterations $100,000$ for the convergence of the model parameters. We also adopt the results of RCA and SIMEX for this dataset from Carroll et al. (2006). In the following, we briefly describe the procedures of RCA and SIMEX.

The basis of RCA, which was initially suggested by Carroll and Stefanski (1990) and Gleser (1990), is the replacement of $\boldsymbol{x}$ by the regression of $\boldsymbol{x}$ on $(\boldsymbol{z}, \boldsymbol{X})$. One can then perform a standard analysis. Obviously, the simplicity of this algorithm masks its power. Although, RCA tends to be most useful for GLMMs, however, this approach can be rather poor for highly nonlinear models.

The SIMEX, which was initially proposed by Cook and Stefanski (1994), is indeed a simulation-based method of estimating and reducing bias due to measurement error. In this approach, the estimates are obtained by adding additional measurement error to the data in a resampling-like stage, then establishing a trend of measurement error-induced bias versus the variance of the added measurement error, and then extrapolating this trend back to the case of no measurement error. The SIMEX is ideally suited to problems with additive measurement error.

Table 2 shows the inference results for MLE, RCA, SIMEX, and HB methods. It seems that the results in four methods agree reasonably closely on most parameters except the estimate of variance component σ_x^2 which is 0.002 for the MLE compared to 0.013 for for the HB method. Note that also using data cloning, we get the unique MLE based on the likelihood, while RCA is based on an approximate approach and SIMEX is a simulation-based method. Also, note that in HB approach one may get different results than MLE with using different priors, while the inferences in MLE method are invariant to the choice of priors. To show this fact, we also considered uniform distribution $U(0, 1000)$ for σ_x in HB approach and observed that the estimate of variance component σ_x^2 is 0.003 while this value for gamma distribution is 0.013 as shown in Table 2. Note that Carroll et al. (2006) did not estimate the variance component σ_x^2 due to large degrees of freedom for estimating σ_v^2, and also they were not interested in estimating the standard error of $\hat{\sigma}_v^2$ for RCA and SIMEX methods.

As indicated in Section 3, one of the main features of data cloning is the ability to predict the random effects. We provided the 95% prediction bands

Table 2: The estimates (and standard errors) of the model parameters in Framingham Heart Study for MLE, HB, RCA, and SIMEX methods.

Parameter	MLE	HB	RCA	SIMEX
β_1	0.067(0.011)	0.067(0.012)	0.053(0.010)	0.053(0.010)
β_2	0.537(0.245)	0.506(0.252)	0.600(0.250)	0.600(0.240)
β_3	0.008(0.002)	0.008(0.002)	0.008(0.002)	0.008(0.002)
β_4	1.838(0.196)	1.790(0.199)	2.000(0.460)	1.930(0.440)
σ_v^2	0.043(0.002)	0.033(0.004)	0.012	0.006
σ_x^2	0.002(0.002)	0.013(0.003)		

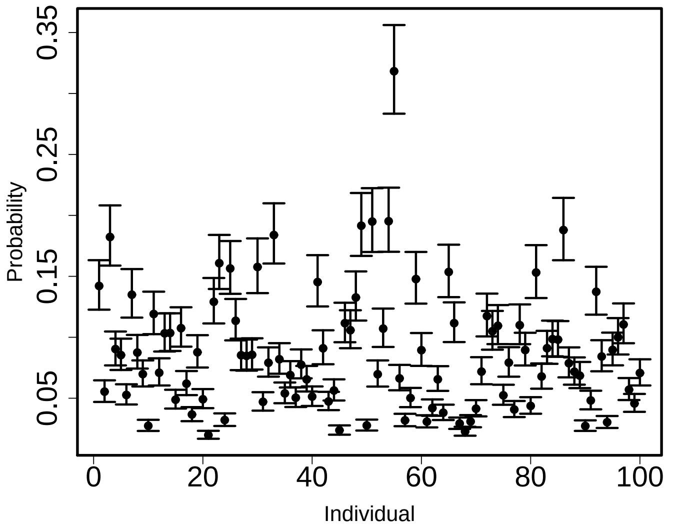

Figure 1: The 95% prediction bands for the first 100 CHD rates in Framingham Heart Study for the MLE approach. The bullet represents prediction rates with corresponding lower and upper prediction bands.

for the first 100 CHD rates (Fig. 1) while the remaining predictions show similar pattern; noting that the CHD rates are predicted not estimated since they are random effects.

5. Simulation study

5.1. Simulation study based on linear mixed measurement error model

We conduct a simulation study on the performance of model parameters estimate for the MLE approach via data cloning. We use a real dataset given in OPEN Study to simulate samples from models (4.1)-(4.2).

We first obtain the estimates of model parameters from the dataset, using for example MLE approach, and then treat them as known for the simulation study. In particular, $m = 223, n_i = 2$, and we choose $\beta_0 = 2.45, \beta_1 = 0.48, \alpha_0 = 5.72, \alpha_1 = -0.01, \alpha_2 = -0.003, \sigma_x^2 = 0.27, \sigma_u^2 = 0.33, \sigma_\epsilon^2 = 0.87$, and $\sigma_v^2 = 0.46$. Using those parameter values, estimates are obtained using MLE and HB analyses of R=1,000 datasets $\{(y_{ij}^{(r)}, X_{ij}^{(r)}, \text{Age}_i, \text{BMI}_i), r = 1, ..., R\}$ generated from models (4.1)-(4.2). For this simulation set up, the average number of clones was $K = 50$ to obtain MLE, and the average number of iterations for convergence of the model parameters was about $20,000$.

Table 3 presents the mean values and variances of the fixed parameters and variance component parameters, as well as corresponding simulated variances of the model parameters estimate for two methods MLE and HB. It seems that the MLE approach produces estimates with small bias for the model parameters, and their variances are comparable with corresponding

Table 3: Mean values and variances (VAR), and simulated VAR of the estimates based on 1,000 simulated datasets in OPEN Study for MLE and HB methods.

	MLE			HB		
Parameter	Mean	VAR	Simulated VAR	Mean	VAR	Simulated VAR
$\beta_0 = 2.45$	2.432	0.620	0.644	1.860	0.764	1.442
$\beta_1 = 0.48$	0.483	0.024	0.024	0.596	0.029	0.055
$\alpha_0 = 5.72$	5.746	0.143	0.148	5.726	0.128	0.150
$\alpha_1 = -0.01$	-0.010	0.00003	0.00003	-0.010	0.00003	0.00003
$\alpha_2 = -0.003$	-0.004	0.0001	0.0001	-0.003	0.0001	0.0001
$\sigma_x^2 = 0.27$	0.266	0.003	0.002	0.236	0.002	0.003
$\sigma_u^2 = 0.33$	0.311	0.008	0.008	0.215	0.008	0.011
$\sigma_\epsilon^2 = 0.87$	0.876	0.007	0.007	0.927	0.009	0.010
$\sigma_v^2 = 0.46$	0.458	0.002	0.002	0.484	0.002	0.002

simulated variances. On the other hand, the bias for the model parameters in HB method is relatively large with also underestimating variances of some model parameters estimate. Overall, it seems that MLE approach provides reasonable point estimates and corresponding variances for this set up, and its performance is much better than HB method.

5.2. Simulation study based on logistic mixed measurement error model

We also conduct a simulation study to evaluate the performance of the proposed MLE approach in the context of logistic mixed measurement error model. Binary observations y_{ij} are generated within each cluster with conditional success probabilities satisfying the following logit model

$$logit\{Pr(y_{ij}|x_i, z_{ij}, u_i)\} = \beta_0 + \beta_1 x_i + \beta_2 z_{ij} + u_i, \tag{5.1}$$

$$X_{ij} = x_i + V_{ij}, \quad (i = 1, ..., m; j = 1, ..., n), \tag{5.2}$$

where $x_i \sim N(\mu_x, \sigma_x^2), u_i \sim N(0, \sigma_u^2), V_{ij} \sim N(0, \sigma_v^2)$. We set $m = 50$ and $n = 3$ which is a common sample size in longitudinal studies, for example. The exactly measured covariate z is generated independently from a standard Normal distribution. Other parameters used to specify $y|x$ and $X|x$ models are $\beta_0 = 0, \beta_1 = 2, \beta_2 = 1, \mu_x = 0, \sigma_x^2 = 1$, and $\sigma_u^2 = 0.5$. We consider different values of $\sigma_v^2 = 0.5, 1.0, 2.0$, and also generate x_i from different distributions: Normal distribution with mean μ_x and variance σ_x^2, and chi-square distribution with one degree of freedom. We generate 1,000 simulations from models (5.1)-(5.2). The results for Normal and chi-square distributions, for different values of σ_v^2, are displayed in Tables 4 and 5, respectively. In general, in the case of Normal distribution, for different values of σ_v^2, the MLE method performs well with very small bias for regression coefficients and variance components, and also their variances are estimated very well and are comparable with the corresponding simulated variances (Table 4). On the other hand, the HB approach performs poorly with relatively large bias particulary for variance component σ_u^2, and also the variances of model parameters estimates are generally underestimated. The variances of model parameters estimates are also increased with increasing the values of σ_v^2 for both DC and HB methods. We are also interested to check the robustness of the DC approach in terms of misspecification of the measurement error x. Table 5 reports means, variances, and simulated variances of model parameters estimates for both DC and HB methods when the measurement error

x are generated from chi-square distribution with one degree of freedom but normal specification of x in the likelihood approach would be appreciated. It seems that we have relatively large bias for μ_x and σ_x^2 for both MLE and HB methods, however, in the case of MLE approach the bias terms for other model parameters estimates are reasonably small and the variances of the model parameters estimates are also estimated well. Similarly, the variances of the model parameters estimates are increased with increasing the values of σ_v^2. Note that for this simulation set up, the average number of clones was $K = 50$ to obtain MLE, and the average number of iterations for convergence of the model parameters was about $20,000$.

6. Concluding remarks

There are many situations where responses are proportions or counts and covariates are measured with error. Often, for fitting complex models in measurement error context, approximate methods such as RCA or simulation-based approaches such as SIMEX are used in the frequentist paradigm; a potential difficulty in making inference based on MLE is that a full-likelihood analysis is burdened by often intractable numerical integration. However, Bayesian methods are advocated because they are computationally more convenient than ML method. Analysis based on data cloning overcomes the computational difficulties of the ML method. Under the Normal and logistic mixed models, the data cloning, which yields to the MLE, may also lead to better inferential solutions to the model parameters, compared to existing frequentist approaches such as RCA and SIMEX, where the covariates are

Table 4: Mean values and variances (VAR), and simulated VAR of the estimates for different measurement error variances based on 1,000 simulated datasets in logistic mixed measurement error model for MLE and HB methods.

	MLE			HB		
Parameter	Mean	VAR	Simulated VAR	Mean	VAR	Simulated VAR
$\beta_0 = 0$	-0.012	0.086	0.092	-0.014	0.084	0.104
$\beta_1 = 2$	2.119	0.253	0.320	2.376	0.342	0.534
$\beta_2 = 1$	1.049	0.103	0.116	1.072	0.099	0.124
$\mu_x = 0$	0.002	0.022	0.023	0.002	0.020	0.023
$\sigma_x^2 = 1$	0.969	0.052	0.062	0.865	0.043	0.056
$\sigma_u^2 = 0.5$	0.570	0.512	0.620	0.052	0.002	0.0001
$\sigma_v^2 = 0.5$	0.497	0.005	0.008	0.502	0.005	0.008
$\beta_0 = 0$	-0.010	0.104	0.110	-0.012	0.143	0.161
$\beta_1 = 2$	2.131	0.319	0.292	2.809	0.788	0.808
$\beta_2 = 1$	1.042	0.105	0.102	1.116	0.117	0.129
$\mu_x = 0$	0.008	0.026	0.026	0.008	0.022	0.026
$\sigma_x^2 = 1$	0.974	0.070	0.068	0.788	0.055	0.068
$\sigma_u^2 = 0.5$	0.519	0.506	0.614	0.050	0.002	0.00002
$\sigma_v^2 = 1.0$	0.996	0.019	0.019	1.032	0.022	0.023
$\beta_0 = 0$	-0.026	0.128	0.132	-0.035	0.341	0.310
$\beta_1 = 2$	2.013	0.360	0.305	3.470	2.413	1.316
$\beta_2 = 1$	1.015	0.105	0.101	1.120	0.126	0.138
$\mu_x = 0$	0.004	0.034	0.032	0.004	0.027	0.032
$\sigma_x^2 = 1$	1.054	0.118	0.081	0.678	0.085	0.087
$\sigma_u^2 = 0.5$	0.408	0.375	0.509	0.049	0.002	0.00001
σ_v^2=2.0	1.943	0.069	0.062	2.074	0.086	0.080

Table 5: Mean values and variances (VAR), and simulated VAR of the estimates for different measurement error variances based on 1,000 simulated datasets in logistic mixed misspecified measurement error model for MLE and HB methods.

	MLE			HB		
Parameter	Mean	VAR	Simulated VAR	Mean	VAR	Simulated VAR
$\beta_0 = 0$	0.238	0.132	0.142	0.079	0.127	0.165
$\beta_1 = 2$	1.554	0.259	0.220	1.815	0.352	0.420
$\beta_2 = 1$	1.057	0.106	0.105	1.059	0.097	0.106
$\mu_x = 0$	1.003	0.040	0.039	1.003	0.038	0.039
$\sigma_x^2 = 1$	1.949	0.180	0.778	1.797	0.152	0.071
$\sigma_u^2 = 0.5$	0.662	0.574	0.707	0.052	0.002	0.0001
$\sigma_v^2 = 0.5$	0.498	0.005	0.005	0.496	0.005	0.005
$\beta_0 = 0$	0.322	0.147	0.149	0.053	0.188	0.243
$\beta_1 = 2$	1.366	0.222	0.192	1.785	0.426	0.568
$\beta_2 = 1$	1.058	0.108	0.103	1.084	0.106	0.114
$\mu_x = 0$	0.999	0.043	0.043	0.999	0.039	0.043
$\sigma_x^2 = 1$	1.904	0.202	0.773	1.686	0.167	0.719
$\sigma_u^2 = 0.5$	0.701	0.658	0.782	0.052	0.002	0.00003
$\sigma_v^2 = 1.0$	0.996	0.020	0.019	1.011	0.022	0.022
$\beta_0 = 0$	0.395	0.167	0.173	-0.236	0.575	0.804
$\beta_1 = 2$	1.180	0.194	0.178	2.032	0.909	1.307
$\beta_2 = 1$	1.043	0.107	0.104	1.104	0.116	0.125
$\mu_x = 0$	1.010	0.051	0.046	1.010	0.043	0.046
$\sigma_x^2 = 1$	1.950	0.282	0.752	1.527	0.226	0.802
$\sigma_u^2 = 0.5$	0.673	0.674	0.772	0.050	0.002	0.00001
$\sigma_v^2 = 2.0$	1.992	0.078	0.079	2.113	0.103	0.124

measured with error. Under the GLMMeM, data cloning can also provide prediction intervals similar to Bayesian approach with added advantages that the answers are invariant to the choice of prior.

Although, in this paper our focus was on a classical, additive, unbiased and non-differential measurement error model, the data cloning approach can be also easily extended to different situations such as heteroscedastic, Berkson or differential measurement error structures (Carroll et al., 2006). We have planned to develop these models.

Acknowledgments

I would like to thank two referees and an Associate Editor for constructive comments and suggestions. This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. I also thank Amy F. Subar for providing the OPEN Study dataset.

References

N.E. Breslow, D.G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88, 9-25.

R.J. Carroll, J.D. Maca, D. Ruppert (1999a). Nonparametric regression in the presence of measurement error. *Biometrika* 86, 541-554.

R.J. Carroll, K. Roeder, L. Wasserman (1999b). Flexible parametric measurement error models. *Biometrics* 55, 44-54.

R.J. Carroll, D. Ruppert, L.A. Stefanski, C.M. Crainiceanu (2006). "*Measurement Error in Nonlinear Models: A Modern Perspective,*" Chapman and Hall/CRC, Boca Raton.

R.J. Carroll, L.A. Stefanski (1990). Approximate quasilikelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, 85, 652-663.

J.R. Cook, L.A. Stefanski (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89, 1314-1328.

G.S. Datta, J.N.K. Rao, M. Torabi (2010). Pseudo-empirical Bayes estimation of small area means under a nested linear regression model with functional measurement errors. *Journal of Statistical Planning and Inference*, 140, 2952-2962.

W.A. Fuller (1987). "*Measurement error models*," John Wiley and Sons, New York.

W.R. Gilks, S. Richardson, D.J. Spiegelhalter (ed.) (1996). *Markov chain Monte Carlo in Practice.* Chapman and Hall, London.

L.J. Gleser (1990). Improvements of the naive approach to estimation in non-linear errors-in-variables regression models. In P.J. Brown and W.A. Fuller (Eds.) *Statistical Analysis of Measurement Error Models and Application.* American Mathematics Society, Providence.

A. Guolo (2008a). A flexible approach to measurement error correction in case-control studies. *Biometrics*, 64, 1207-1214.

A. Guolo (2008b). Robust techniques for measurement error correction: a review. *Statistical Methods in Medical Research*, 17, 555-580.

J.D. Hamilton (1986). A standard error for the estimated state vector of a state-space model. *Journal of Econometrics*, 33, 387-397.

W.B. Kannel, J.D. Neaton, D. Wentworth, H.E. Thomas, J. Stamler, S.B. Hulley, M.O. Kjelsberg (1986). Overall and coronary heart disease mortality rates in relation to major risk factors in 325,348 men screened for MRFIT. *American Heart Journal*, 112, 825-836.

V. Kipnis, D. Midthune, L.S. Freedman, S. Bingham, N.E. Day, E. Riboli, R.J. Carroll (2003). Bias in dietary-report instruments and its implications for nutritional epidemiology. *Public Health Nutrition*, 5, 915-923.

S.R. Lele, B. Dennis, F. Lutscher (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecololgy Letters*, 10, 551-563.

S.R. Lele, K. Nadeem, B. Schmuland (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, 105, 1617-1625.

C.E. McCulloch (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92, 162-170.

C.E. McCulloch, S.R. Searle, J.M. Neuhaus (2008). "*Generalized, Linear, and Mixed Models*," 2nd ed., John Wiley and Sons, New York.

R Development Core Team: A language and environment for statistical computing. R foundation for statistical computing. http://www.Rproject.org.

K. Roeder, R.J. Carroll, B.G. Lindsay (1996). A semiparametric mixture approach to case-control studies with errors in covariates. *Journal of the American Statistical Association*, 91, 722-732.

S.R. Searle, G. Casella, C.E. McCulloch (1992). "*Variance Components*," John Wiley and Sons, New York.

B.J. Smith (2007). *BOA user manual (version 1.1.7)*, Department of Biostatistics, College of Public Health, University of Iowa, Ames.

P. Sólymos (2010). dclone: data cloning in R. *The R Journal*, 2, 29-37.

D. Spiegelhalter, A. Thomas, N. Best, D. Lunn (2004). *WinBUGS version 1.4 user manual.* MRC Biostatistics unit, Institute of Public Health, London.

A.F. Subar, V. Kipnis, R.P. Troiano, D. Midthune, D.A. Schoeller, S. Bingham, C.O. Sharbaugh, J. Trabulsi, S. Runswick, R. Ballard-Barbash, J. Sunshine, A. Schatzkin (2003). Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The Observing Protein and Energy Nutrition (OPEN) study. *American Journal of Epidemiology*, 158, 1-13.

A.F. Subar, F.E. Thompson, V. Kipnis, D. Midthune, P. Hurwitz, S. McNutt, A. McIntosh, S. Rosenfeld (2001). Comparative validation of the Block Willett and National Cancer Institute food frequency questionnaires: The Eating at America's Table Study. *American Journal of Epidemiology*, 154, 1089-1099.

M. Torabi (2011). Small area estimation using survey weights with functional measurement error in the covariate. *Australian & New Zealand Journal of Statistics*, 53, 141-155.

M. Torabi (2012a). Small area estimation using survey weights under a nested error linear regression model with structural measurement errors. *Journal of Multivariate Analysis*, 109, 52-60.

M. Torabi (2012b). Likelihood inference in GLMM with two components of dispersion using data cloning. *Computational Statistics and Data Analysis*, 56, 4259-4265.

M. Torabi, G.S. Datta, J.N.K. Rao (2009). Empirical Bayes estimation of small area means under a nested error linear regression model with measurement errors in the covariates. *Scandinavian Journal of Statistics*, 36, 355-368.

M. Torabi, F. Shokoohi (2012). Likelihood inference in small area estimation by combining time-series and cross-sectional data. *Journal of Multivariate Analysis*, 111, 213-221.

N. Wang, X. Lin, R.G. Gutierrez, R.J. Carroll (1998). Bias analysis and SIMEX approach in generalized linear mixed measurement error models. *Journal of the American Statistical Association*, 93, 249-261.

00