Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: http://www.tandfonline.com/loi/cjas20

A Bayesian model for measurement and misclassification errors alongside missing data, with an application to higher education participation in Australia

Harvey Goldstein, William J. Browne & Christopher Charlton

JOURNAL OF APPLIED STATISTICS, 2017
https://doi.org/10.1080/02664763.2017.1322558

A Bayesian model for measurement and misclassification errors alongside missing data, with an application to higher education participation in Australia

Harvey Goldstein, William J. Browne and Christopher Charlton

Centre for Multilevel Modelling, Graduate School of Education, University of Bristol, Bristol, UK

ABSTRACT
In this paper we consider the impact of both missing data and measurement errors on a longitudinal analysis of participation in higher education in Australia. We develop a general method for handling both discrete and continuous measurement errors that also allows for the incorporation of missing values and random effects in both binary and continuous response multilevel models. Measurement errors are allowed to be mutually dependent and their distribution may depend on further covariates. We show that our methodology works via two simple simulation studies. We then consider the impact of our measurement error assumptions on the analysis of the real data set.

ARTICLE HISTORY
Received 3 October 2016
Accepted 16 April 2017

KEYWORDS
Higher education participation; measurement errors; misclassification errors; MCMC; missing data; multilevel

1. Introduction

Measurement errors often occur in many of the variables used in the social and medical sciences. These may arise from unreliable measuring instruments, or, for example, short term fluctuations over time. It is reasonably well known that a failure to deal with measurement errors can lead to biased inferences when the intention is to model data using the 'true' but unknown values. Fuller [10] provides a comprehensive account and there is a more recent literature ([4,6,20,22]) that includes Bayesian approaches. Buonaccorsi [2] provides a comprehensive review of non-Bayesian methods. Muff *et al.* [18] provide a useful overview as well as proposing a Bayesian model using a computationally fast Laplace transformation.

Likewise, missing data values are endemic in observational data, and there is now a considerable literature (see for example, [3]) on how to deal with these, especially when missingness is in predictor variables and is not completely random. The case where data contain both missing values and measurement errors, has received little attention, despite being quite common. The aim of the present paper is to propose an integrated Bayesian approach, using Markov Chain Monte Carlo (MCMC) estimation, to the modelling of such data, where the model for the missing data is viewed as a special case of the model for the measurement errors. Although in the example in this paper our approach focusses on

CONTACT Harvey Goldstein h.goldstein@bristol.ac.uk Centre for Multilevel Modelling, Graduate School of Education, University of Bristol, Bristol BS8 1JA, UK

generalised multilevel linear models, we indicate how our approach can easily be extended to model multivariate data, heteroscedatistic measurement errors, and models that include nonlinear and interaction terms.

We begin by describing briefly data where there are both measurement errors and missing data and then outline the MCMC methodology required. We then carry out two simulations to illustrate the approach on normal response and binary response models, as well as a more detailed analysis of the effects of both missing data and measurement errors on the modelling of our example data set. We end with a discussion in which we also describe various extensions to more complex data structures.

2. Example data set: the longitudinal study of Australian youth

Our procedures will be applied to the 'Longitudinal Study of Australian Youth' (LSAY) data set, a longitudinal study with up to 12 waves of data collection. This is a study that was designed to track the pathways of young Australians as they move from school to further study, work and other destinations. Data were collected on variables related to education, training, work, financial matters, health, social activities and attitudes as well as background family characteristics such as Socio Economic Status (SES). A description of the variables is given by Cumming and Goldstein [7]. LSAY started in 1995 by sampling Year 9 school students, with an average age 14.5 years, in Australian secondary schools and subsequently is following them up every year on a further 11 occasions. Cumming and Goldstein [7] studied year 9 predictors of the probability of being in full time or part time education 6 years after the study start at wave 6 of data collection (a binary response), when the students had a modal age of 20.5 years. The sample suffers attrition of just under 50% over this period, in addition to item missing data. The number of pupils available for analysis, after excluding those students with no wave 6 information is 6901 and the data has a 2 level structure with the pupils each belonging to one of 296 schools.

While Cumming and Goldstein [7] did account for attrition and missing data they did not allow for test score unreliability. Such unreliability is a general feature of educational tests and arises from several sources of variation including the choice (sampling) of test questions, conditions of test administration and short term fluctuations within students. Ecob and Goldstein [8] provide a detailed discussion. We describe the specific model for these data in a later section but now build up the modelling approach in stages while referring to earlier published work for many of the algorithm details.

3. Model specification

In an ideal scenario where we have no measurement errors and missing data then we will fit a binary response model to our indicator of educational attendance at wave 6 and relate this response to various predictor variables whilst also factoring in the 2 level structure via school level random effects. Our final model of interest (MOI) will therefore be of the form:

$$p_{ij} \sim \text{Bin}(1, \ \pi_{ij}),$$

$$\text{probit}(\pi_{ij}) = \beta_0 + \sum_{k=1}^{n} \beta_k X_{kij} + u_j,$$

$$u_j \sim N(0,\ \sigma_u^2).$$

Here, we have deliberately chosen to use a probit link function as opposed to a logistic regression since our modelling approach, that incorporates measurement errors and missing data, will be adapted from an approach used for normal response models using latent variable approaches (see examples later).

With this in mind we begin by considering a simpler example of how one can incorporate measurement errors in predictor variables into statistical models more generally by considering normal response models as the extension to other models will be straightforward.

So consider the linear regression model:

$$Y = X\beta + e \tag{1}$$

$$e \sim N(0,\ \sigma_e^2).$$

In this scenario it makes sense to differentiate between the predictor variables that contain measurement variables which we will label X_1 and those that do not, X_2. Here we use capital letters to represent the true values and we have $X = [X_1\ X_2\ Z]$ where for complete generality we could include $Z = f(X_1,\ X_2)$, thus allowing for interactions between variables and nonlinear effects. The model (1) is therefore a standard model that relates a normal response to the true values of a set of variables. As some of these true values are not available due to measurement errors we therefore also require notation for the observed predictors. Thus corresponding to the true values, X_1, we denote by x_1 the observed values of the variables with measurement error. Although the variables in Z will also not be observed, they are simply functions of the other true values and so thus we do not need a corresponding z for their observed equivalents.

To complete the model we require therefore a measurement error distribution to relate x_1 to X_1 and here we assume that information about the distribution of the measurement errors is available and in particular their variability is known. For continuous variables we assume therefore that the errors are jointly normally distributed with known variances. It is possible in the Bayesian framework to extend the modelling in practice to allow the measurement error variance to be unknown and use instead a prior distribution of possible values but in practice this doesn't add much to the analysis and often one prefers to answer 'what if' questions in terms of the size of the measurement errors as a sensitivity analysis. For simplicity of exposition we begin with the case of a single variable having measurement error. We write the measurement error component of our full model as

$$x_1 = X_1 + \gamma_1 \tag{2}$$

$$\begin{pmatrix} X_1 \\ \gamma_1 \end{pmatrix} \sim N\begin{pmatrix} \sigma_{X_1}^2 & \\ 0 & \sigma_\gamma^2 \end{pmatrix},$$

where we assume independent normal distributions for both the true values (X_1) and the measurement errors (γ_1) and also assume that the measurement errors are independent of the true values of all predictors. Such a formulation is known as 'classical' measurement error modelling which is generally the approach used with observational data (see, e.g. [18]).

We can therefore use our two model components (1) and (2) to form the complete model for the data

$$p(Y,\ x_1,\ X_1,\ X_2) = p(Y|x_1,\ X_1,\ X_2)p(x_1,\ X_1,\ X_2).$$

Since $x_1 = X_1 + \gamma_1$ and we assume that γ_1 is independent of $X_1,\ \ X_2$ and Y we can write the first term as

$$p(Y|x_1,\ X_1,\ X_2) = p(Y|X_1,\ X_2).$$

We can also decompose the second term as

$$p(x_1,\ X_1,\ X_2) = p(x_1|X_1,\ X_2)p(X_1,\ X_2) = p(x_1|X_1,\ X_2)p(X_1|X_2)p(X_2),$$

where again using the formula for x_1 and assuming independence of γ_1 and X_2 we have

$$p(x_1|X_1,\ X_2) = p(x_1|X_1).$$

So that we have

$$p(Y,\ x_1,\ X_1,\ X_2) = p(x_1|X_1)p(X_1|X_2)p(Y|X_1,\ X_2)p(X_2).$$

Since X_2 are all known data we can drop the final term in the above function.

The above expression corresponds to model (3a)–(3c) below:

The three components can be written as the full model:

$$x_1 = X_1 + \gamma_1, \tag{3a}$$

$$X_1 = X_2\alpha + \gamma_2, \tag{3b}$$

$$Y = X\beta + e, \tag{3c}$$

$$\gamma_1 \sim N(0,\ \sigma^2_{\gamma_1}), \qquad \gamma_2 \sim N(0,\ \sigma^2_{\gamma_2}), \qquad e \sim N(0,\ \sigma^2_e).$$

We shall consider generalisations of our simple model in later sections.

As they stand (3a)–(3c) do not provide identifiability for the individual parameters. As is commonly done, the measurement error variance $\sigma^2_{\gamma_1}$ is therefore assumed known so that $(\sigma^2_{\gamma_2},\ \sigma^2_e,\ \alpha,\ \beta)$ are the parameters to be estimated. In our example data analysis we carry out sensitivity analysis on the measurement error variances using some assumed values derived from existing research, since little information is available for the actual data themselves. To complete the Bayesian formulation uniform priors are included for each of these four sets of parameters. In common with standard usage we define the reliability of the observed variable x_1 as $R = \sigma^2_{X_1}/\sigma^2_{x_1}$. Here we calculate $\sigma^2_{x_1}$ directly from the observed variable and then we can estimate $\sigma^2_{X_1} = \sigma^2_{x_1} - \sigma^2_{\gamma_1}$ in other words the variability in the observed response not explained by measurement error.

We assume normality for purposes of exposition, but other distributional assumptions are possible with corresponding changes to the MCMC steps described. In particular we shall later deal with the binary case.

For a multilevel model the only change is that (3c) will incorporate random effects. Thus, for a variance components model with a single random effect (3c) would become

$$Y = X\beta + u + e, \qquad u \sim N(0,\ \sigma_u^2) \tag{3d}$$

and extra steps to sample the $u,\ \ \sigma_u^2$ are inserted [19].

In some cases the distribution of the measurement errors may depend on other variables, some of which may be in the MOI. Denoting these by X_4 and assuming that they are measured without error, the term $\sigma_{\gamma_1}^2$ becomes $\sigma_{\gamma_1}^2 D$ where D is a known $(n \times n)$ diagonal scaling matrix with n the sample size. For example, if the measurement error variance is different for males and females, say $\sigma_{\text{em}}^2,\ \ \sigma_{\text{ef}}^2$ then if sample record j is for a male the jth element of D would be σ_{em}^2 and if female, σ_{ef}^2.

4. MCMC estimation for a continuous predictor

Consider first the step in our algorithm where we propose a new true value, say X_{1i}, for record i, where for simplicity we assume that random effects are already incorporated in the response The joint log posterior from Equations (3a), (3b) and (3c) is thus proportional to the sum of the following components:

$$-\frac{0.5(x_{1i} - X_{1i})^2}{\sigma_{\gamma_1}^2}, \qquad -\frac{0.5(X_{1i} - X_{2i}^T\alpha)^2}{\sigma_{\gamma_2}^2}, \qquad -\frac{0.5(\tilde{y}_i)^2}{\sigma_e^2},$$

where $\tilde{y}_i = y_i - X_i\beta$.

When sampling a new value of X_{1i} we use a Metropolis step and for a proposal distribution we suggest a form of independence sampler

$$p(X_{1i}|x_{1i}) \sim N(x_{1i}R,\ R(1-R)\sigma_{x_{1i}}^2), \tag{4}$$

where R is the reliability defined above. Model (3) is similar to the formulation by Richardson and Gilks [20] where they have a 'gold standard validation' sample that provides the information associated with (3a).

For the case where we have more than 1 variable with measurement error we can propose the set of values defined using Equation (3) for each variable separately or look at the joint proposal distribution:

$$f(\boldsymbol{X}_1|x_1) \sim \text{MVN}(\boldsymbol{X}_1\Omega_{x_1}^{-1}\Omega_{X_1},\ \Omega_{X_1} - \Omega_{X_1}\Omega_{x_1}^{-1}\Omega_{X_1}),$$

where $\Omega_{x_1},\ \ \Omega_{X_1}$ are, respectively, the covariance matrices for the observed and true values (the multivariate analogues of $\sigma_{x_1}^2$ and $\sigma_{X_1}^2$, respectively). Other MCMC steps for the MOI (3c) are standard conjugate Gibbs sampling steps as are the steps for the parameters in Equation (3b) as conditional on deriving the true predictor values the model is a standard linear model.

In fact this model, with $\alpha = 0$ in Equation (3b), that is, the model with measurement errors being unrelated to other predictors, is essentially the existing implementation of Goldstein *et al.* [15] based upon Browne *et al.* [1]. Here however they use a Gibbs rather than Metropolis step for the X_{1i}. We note, however, that the use of such a simplified formulation is only really appropriate in the case when X_1 and X_2 are orthogonal.

Where the MOI is a generalised linear model with the response as a binary, multi-category or count we can use a latent normal model for (3c) (for MCMC implementations for such models without measurement errors see [13] for categorical responses and [14] for count models). In these cases an extra sampling step is inserted that samples one or more assumed underlying standard normal variates for each of the observed discrete values.

5. Misclassification errors

We have demonstrated in Section 4 an MCMC algorithm for estimating the true value of a continuous predictor that is measured with error. We next consider binary predictor variables where errors are often described as misclassifications rather than measurement errors. We will then discuss briefly the extension to multicategory predictors.

Consider the case where a new true predictor variable X_3 is binary with corresponding observed value x_3. We now rewrite model (3) as

$$p(x_3 = a|X_3 = b) \; = p_{ab}, \qquad \text{for } a,\; b = (0,\; 1), \tag{5a}$$

$$X_3 = f(X_2\alpha), \tag{5b}$$

$$Y = X\beta + e, \qquad e \sim N(0,\; \sigma_e^2). \tag{5c}$$

We shall choose f as the probit function for convenience to obtain a conditional normal distribution that therefore implies full multivariate normality, so that we can use the steps described in Section 4. We assume that all four of the p_{ab} are known fixed values, although again sensitivity analyses for different values can be carried out. Here we now have $X = [X_3\; X_2\; Z_3]$ where for complete generality, for example in order to fit interactions, we include $Z_3 = f(X_3,\; X_2)$ and X_2 are the variables that do not contain measurement errors as before.

The probit function can be written as $p(X_{3i} = 1) = \int_{-X_{2i}\alpha}^{\infty} \phi(t)\,\mathrm{d}t = \int_{-\infty}^{X_{2i}\alpha} \phi(t)\,\mathrm{d}t$ where $\phi(t)$ is the standard normal distribution. We first sample, therefore, a set of latent normal variables, X_{3i}^*, according to the current values of X_{3i}, for example, for a value of 1 we sample from the upper tail of this standard normal distribution and for value 0, the lower tail. Thus, we can now rewrite (5b) as the normal linear model

$$X_{3i}^* = X_{2i}\alpha + \gamma_{2i} \tag{5d}$$

so that we can update the α parameters in a standard MCMC step as for continuous predictors. We also note that here the $\gamma_{2i} \sim N(0,\; 1)$, which is fixed by the probit function, and so we do not have a variance parameter to estimate.

To update the X_{3i}, we carry out a Metropolis step so that if the existing value is $X_{3i} = 0$ we propose a new value $X_{3i} = 1$ and vice versa. The joint likelihood contains the same component for Equation (5c) as before in Equation (3c). For Equation (5a) with observed value $x_3 = a$ and proposed true value $X_3 = b$, the component is simply p_{ab}. For Equation (5b) for proposed true value b, we evaluate the probit function at current parameter values (α), using Equation (5d).

This can be readily extended to ordered categories and also unordered categories, using appropriate latent normal transformations (see [13]).

6. Measurement errors and misclassification errors

For the case where there are both measurement errors and binary misclassification errors we assume that these are independent of each other. It will also often be reasonable in applications to assume that the binary misclassification errors are mutually independent, as in the exposition below. If we denote the misclassification error variables by the true values X_3, the joint distribution can now be written as

$$\begin{aligned} &p(Y,\ x_1,\ X_1,\ X_2,\ x_3,\ X_3) \\ &\quad = p(Y|X_1,\ X_2,\ X_3)p(x_1|X_1)p(x_3|X_3)p(X_1|X_2,\ X_3)p(X_3|X_2)p(X_2). \end{aligned} \tag{6}$$

Thus, when updating X_1 we use the equivalent to Equation (3), namely

$$x_1 = X_1 + \gamma_1, \tag{7a}$$

$$X_1 = X_2\alpha_2 + X_3\alpha_3 + \gamma_2, \tag{7b}$$

$$Y = X\beta + e. \tag{7c}$$

When updating each variable in X_3 we now have four components for the likelihood, Equations (7b) and (7c) above and additionally:

$$p(x_3 = a|X_3 = b) \ = p_{ab}, \qquad \text{for } a,\ b \in (0, 1), \tag{7d}$$

$$X_3 = f(X_2^{\mathrm{T}}\alpha_4). \tag{7e}$$

In Equation (7e), for convenience, we may use a probit function for X_3, with assumed known values for the p_{ab}. We note that the decomposition (6) implies no dependence of X_3 on X_1. Where the missclassification errors are not independent (7d) could be extended to incorporate the joint distribution of several binary variables.

7. Incorporating missing data values

Goldstein *et al.* [12] present a Bayesian MCMC algorithm for fitting models with missing covariate data values that extends the traditional joint modelling approach based upon multiple imputation. We can write a simple model with missing data on covariates as follows:

$$Y = X\beta + e, \tag{8a}$$

$$X_1 = X_2\alpha + \gamma_2. \tag{8b}$$

Here X_1 now consists of those variables within X that have missing values and X_2 those that don't. In the update step for the missing data, for each record where there are missing values we propose a new set X_1 using a proposal distribution based on $f(X_1|X_2)$ and then perform a Metropolis step. Thus the only real difference from the measurement error case is that, as seen in Equation (3a), there is an additional component in the posterior for X_1 as we have an observed value x_1. Where we have both variables with measurement errors and missing values, X_1 can include all variables that either have measurement errors or missing values, or both. Where a variable with measurement errors has missing data values these are updated in the step for updating the missing values.

Formally, for a predictor variable with missing values, say $W_j \in (X)$, we note that

$$f(X) = f(W|X_{-W})f(X_{-W}) \tag{9}$$

and we have the additional step for the missing value conditional on the current true values.

The missing values are updated based upon the updated true values, using Equations (8a) and (8b), and when updating a variable's true values, any imputed (true) missing values for this variable will be ignored. This is conveniently carried out by using the current (imputed) value in both the numerator and denominator of the Metropolis ratio so that it has no effect on the acceptance probability. In the simulations and example below, we first carry out, for each data record, a Metropolis step jointly for all the variables with measurement errors. Once the sampling for the true values has taken place the next step carries out imputation where there are missing values, one variable and one record at a time, conditioning on the current true values. This uses the algorithm described in Goldstein *et al.* [12].

In what follows we assume [21] that data are missing completely at random (MCAR), or missing at random (MAR). By MAR is meant that it is randomly missing at least conditionally on all the observed values, that is the covariates and Y, where the latter conditioning is implicit since the full likelihood (6) contains the response as well as the covariates. For missing not at random we may be able to additionally condition on auxiliary variables, not in the MOI, by incorporating them in Equation (8b).

8. Simulations

The first simulation study illustrates a normal response model with a mixture of continuous measurement errors and misclassification errors.

Each simulated data set was generated as follows:

$$X_0 = 1, \quad \begin{pmatrix} X_1 \\ Z \\ X_3 \end{pmatrix} \sim N\left[0, \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}\right], \quad X_2 = \begin{Bmatrix} 0 & \text{if } Z < 0 \\ 1 & \text{if } Z \geq 0 \end{Bmatrix} \tag{10}$$

$$x_1 = N(X_1,\ 0.25), \qquad p_{01} = p_{10} = 0.2 \text{ to create } x_2 \text{ from } X_2.$$

The simulation model is

$$Y \sim N(\mu,\ 1), \qquad \mu = X_0 + X_1 + X_2 + X_3. \tag{11}$$

We fit the measurement error model described in Equations (7a)–(7c) and (8a)–(8b) where the MOI is

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e_i. \tag{12}$$

A sample of 1000 such records is generated.

A burn in of 250 iterations followed by 250 stored iterations was used with 100 simulated data sets. The results are as follows in Table 1.

Table 1. Measurement error simulation.

Estimate (true value)	No adjustment	Adjusted for measurement and misclassification errors
β_0 (1.0)	0.95 (0.01)	1.00 (0.01)
β_1 (1.0)	0.74 (0.01)	1.01 (0.01)
β_2 (1.0)	1.12 (0.01)	0.99 (0.01)
β_3 (1.0)	1.11 (0.01)	1.00 (0.01)
σ_e^2 (1.0)	1.16 (0.02)	0.99 (0.02)

Notes: Hundred simulated data sets from model (12). Between – simulation standard errors in brackets. Reliability = 0.8. Burn in = 250, iterations = 250. Sample size 1000.

We see that biases are induced for all the predictors, including those without errors if we do not adjust for measurement error but including the measurement errors and misclassifications in the model removes these biases.

The second simulation study will consider the case of a binary response and includes both continuous measurement errors and data MCAR. Each simulated data set was generated as before, but with no misclassification errors:

$$X_0 = 1, \quad \begin{pmatrix} X_1 \\ Z \\ X_3 \end{pmatrix} \sim N\left[0, \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}\right], \quad X_2 = \left\{\begin{matrix} 0 & \text{if } Z < 0 \\ 1 & \text{if } Z \geq 0 \end{matrix}\right\}$$

$$x_1 = N(X_1,\ 0.25)$$

For X_1 and X_2 20% of values were randomly assigned to be missing, so that on average 36% of records had at least one missing value.

The simulation model, omitting subscripts, is

$$Y \sim N(\mu,\ 1), \qquad \mu = X_0 + X_1 + X_2 + X_3. \tag{13}$$

For a binary response, the observed response Y_{obs} is defined as

$$Y_{\text{obs}} = 1 \quad \text{if } Y > 0, \qquad Y_{\text{obs}} = 0 \quad \text{if } Y \leq 0.$$

We fit the measurement error model described in Equations (7a)–(7c) and (9b)–(9c) where the MOI is now

$$E(Y_{\text{obs},i}) = \pi_i, \qquad \text{probit}(\pi_i) = \ \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}, \tag{14}$$

where we have $E(\pi_i) = 0.74$.

The number of simulated data sets is 200 and three sample sizes are used; 500, 1000 and 4000. The estimates are given in Table 2, along with the estimates resulting from making no adjustment for measurement error where there is no missing data.

We see that biases are induced for all the predictors if we do not adjust for measurement error, with no missing data, with a large average downward bias of 7% for a sample size of 1000. The percentage bias of our procedure, averaged over the four fixed parameters, is 10.3% for a sample size of 500, 2.7% for a sample size of 1000 and 1.7% for a sample size of 4000. We see, therefore, that where we have both missing data and measurement errors there will remain biases for small samples. Further research into this would be welcome.

Table 2. Measurement error simulation.

Estimate (generating value)	Measurement error, but with no adjustment and no missing data. $N = 1000$	Adjusted for measurement errors and missing data. $N = 500$	Adjusted for measurement errors and missing data. $N = 1000$	Adjusted for measurement errors and missing data. $N = 4000$
$\beta_0(1.0)$	0.868 (0.006)	1.116 (0.011)	1.047 (0.010)	1.025 (0.006)
$\beta_1(1.0)$	0.690 (0.004)	1.147 (0.014)	1.033 (0.012)	1.014 (0.007)
$\beta_2(1.0)$	1.109 (0.010)	1.060 (0.013)	1.002 (0.014)	1.006 (0.009)
$\beta_3(1.0)$	1.033 (0.005)	1.087 (0.008)	1.037 (0.009)	1.022 (0.005)
$\sigma_e^2(1.0)$	1	1	1	1

Notes: Two hundred simulated data sets from model (14). Between – simulation standard errors in brackets. Reliability $= 0.8$. Burn in $= 500$, iterations $= 500$. Sample sizes denoted by N.

Table 3. LSAY data; prediction of the probability of HE participation adjusting for measurement error (ME) with different reliabilities (R) in maths and reading scores, and missing data.

Estimate	Complete cases, no ME adjustment	Adjusting for missing data only	Adjusting for missing data and ME ($R = 0.8$)	Adjusting for missing data and ME ($R = 0.7$)
Intercept	−0.946 (0.091)	−1.010 (0.066)	−1.197 (0.073)	−1.311 (0.080)
Female (male)	0.055 (0.050)	0.051 (0.036)	0.061 (0.037)	0.057 (0.037)
Non-government school	0.202 (0.055)	0.221 (0.049)	0.217 (0.046)	0.219 (0.047)
Maths score year 9	0.499 (0.062)	0.504 (0.046)	0.654 (0.060)	0.753 (0.068)
Reading score year 9	0.235 (0.061)	0.275 (0.044)	0.339 (0.051)	0.390 (0.065)
Non-Australia country of birth of mother (Australia)	0.186 (0.059)	0.183 (0.041)	0.183 (0.042)	0.180 (0.044)
Home language not English (English)	0.412 (0.114)	0.452 (0.072)	0.469 (0.078)	0.485 (0.076)
SES ANU3 score father	0.487 (0.106)	0.438 (0.096)	0.366 (0.106)	0.286 (0.105)
SES ANU3 score mother	0.271 (0.136)	0.185 (0.120)	0.128 (0.112)	0.088 (0.122)
Level 2 variance	0.036 (0.015)	0.038 (0.009)	0.036 (0.011)	0.039 (0.010)

Notes: Probit link. Burn in $= 500$, iterations $= 1000$. Standard errors in brackets. Sample size $= 6901$, complete cases $= 3407$.

9. An example using student participation in higher education

We now return to the data that we described in Section 3. Cumming and Goldstein [7] analyse this data set but considered only the case of missing data and ignored possible measurement errors, using the algorithm described by Goldstein *et al.* [12] to obtain efficient parameter estimates. The first two columns of results in Table 3 replicate these analyses, pooling the two categories of non-Government school (Catholic, Private) that were treated separately by Cumming and Goldstein [7] but in fact showed only a small and non-significant difference and so have been combined in our analysis. The level 2 units are the year 9 schools, and Table 3 lists the predictor variables with full details given by Cumming and Goldstein [7]. Note that the scale of the SES measures has been divided by 100 and the test scores divided by 10 to provide more significant figures for the coefficient estimates.

The specific MOI is a 2-level model with a simple random effect at the school level, and is given by

$$p_{ij} \sim \text{bin}(1,\ \pi_{ij})$$

$$y_{ij} = \text{probit}(\pi_{ij}) = \beta_0 + \sum_{k=1}^{8} \beta_k X_{kij} + u_j + e_{ij} \tag{15}$$

$$u_j \sim N(0,\ \sigma_u^2), \qquad e_{ij} \sim N(0, 1),$$

where for clarity we now utilise the standard double subscript notation for a 2-level model, with $x_1,\ X_1,\ X_2$ defined as in Equations (3a)–(3c). Thus, in terms of our model X_1 are the year 9 test scores having measurement errors and X_2 consists of the remaining predictors without measurement errors. The model contains no discrete covariates with misclassification errors.

The year 9 test scores are each made up 20 binary items, but there appears to be no information about the associated reliabilities. We have therefore carried out a sensitivity analysis using values of 0.8 and 0.7 to study the effect of making adjustments for measurement errors. These values are typical of those found in educational test scores (see, e.g. [9, pp. 351–358]). The correlation between the observed test scores at year 9 is approximately 0.5 and where the true correlation is zero this becomes the correlation between the measurement errors, and can be treated as an upper bound, and we use this value in our analysis. We have also fitted the model assuming a correlation of 0.25 between the measurement errors. The parameter estimates and their standard errors are very similar, as are the standard errors, so that the choice of correlation value is not crucial.

In the final two columns of Table 3 we show the results of adjusting for these reliabilities. We note first, that the principal gain in efficiency lies in moving from a complete case analysis to one that uses the full sample with missing data and in fact the additional adjustment for measurement errors generally increases the standard error estimates. The actual parameter estimates, apart from SES, do not change very much here. There is little change in any of the parameter estimates, except as expected, for the test score coefficients but also for SES, when moving from a model with a reliability of 0.8 to one of 0.7. For the SES of the mother this is reduced considerably moving from the complete case analysis to that assuming a reliability of 0.7 and adjusting for missing values, where the estimate is no longer statistically significant at 5%. Interestingly, the estimates for the other covariates associated with the response, appear relatively unaffected by either adjusting for missingness or measurement error. The sensitivity of SES effects to measurement error adjustment is also found in other studies [9,11] and generally reduces the effects associated with SES. It is worth pointing out that these SES effects are conditional on year 9 test scores and these are themselves associated with SES. Since we do not have good estimates for the reliabilities, we cannot be very precise about the 'true' effects for SES. Further analyses exploring these data are currently under consideration. It does seem reasonable, however, to conclude that the coefficients for the remaining variables other than the test scores, are relatively unaffected by our adjustments.

In Cumming and Goldstein [7] it was concluded that the principal effect of adjusting for missing data was a gain in efficiency, with a small increase in the estimate of the difference between Government and non-Government schools, so that there were no important policy implications. Adjusting additionally for measurement error, however, shows a marked reduction in the SES effects and this would seem to have more important implications for policy. As the debate in Feinstein *et al.* [9] shows, in the UK the effect of SES on children's performance is a source of policy discussions. For example, if SES is found to be less 'important' as a result of improved modelling that takes account of measurement error, this would seem to have important implications for resource allocation policies. In our example and also in the data used in Feinstein *et al.* [9] there were no good estimates of the sizes of measurement error variances and this suggests that more effort should routinely be devoted to obtaining good estimates for these.

10. Discussion

There has long been an awareness of the importance of taking account of measurement errors in observed data, but this is not a feature that is generally available in many software packages. One reason for this may be the complexity associated with available adjustment procedures, typically moment-based ones. There has also been an awareness of the need to deal with missing data values, with rather more software available. In the present paper we have presented a fully Bayesian MCMC algorithm, currently using routines written in Matlab [16], and to be incorporated into the StatJR software [5], that allows adjustments for both measurement errors and missing data. We demonstrate through simulations how our procedures remove biases associated with a failure to account for measurement errors and also how we can simultaneously adjust for measurement errors and missing data. We also describe how it can be used for quite general model structures, including multilevel generalised linear models. A note of caution is needed where we have both measurement errors and missing data where our simulations show that with small sample sizes positive biases may be induced in the parameter estimates. We also, in our example, point to the substantive importance of adjusting for measurement errors, where the low reliability of some predictor variables can have large effects on the resulting estimates, at least in the case of educational data, and we would surmise in other areas too.

There are a number of relatively straightforward extensions to the models proposed.

In addition to the implicit latent normal transformations for non-normal variables, we may wish to formulate the additive measurement error component of the joint model (3a) in terms of a transformed variable. Thus, for example, if the measurement error was multiplicative, we could then express (3a) in additive form by writing (3a)–(3c) as

$$x_1 = X_1 e^{\gamma_1}, \qquad \log(x_1) = \log(X_1) + \gamma_1, \tag{16a}$$

$$X_1 = X_2\alpha + \gamma_2, \tag{16b}$$

$$Y = X\beta + e \tag{16c}$$

$$\gamma_1 \sim N(0,\ \sigma^2_{\gamma_1}), \qquad \gamma_2 \sim N(0,\ \sigma^2_{\gamma_2}), \qquad e \sim N(0,\ \sigma^2_e),$$

where $\sigma^2_{\gamma_1}$ is assumed known or derived from a known value of the variance of e^{γ_1}. The formulation (16a) may be useful for skewed data such as income where a transformation may also help to ensure normality. We may also wish to use transformed values of X_1 in Equations (16b) or (16c) or both. We could also choose, for example, a gamma distribution for γ_1 with corresponding modifications to the likelihood and this would be a useful area for further research.

As in the case of jointly modelling variables with missing values we also can introduce auxiliary variables, say X_3, into Equation (3b) to give

$$X_1 = X_2\alpha_2 + X_3\alpha_3 + \gamma_2. \tag{17}$$

This allows us to deal with the case where, for example, X_1 depends on such auxiliary variables that are not in the MOI, whereas the X_2 are in the MOI (see also [17]).

As we showed in our example multilevel models, including those with cross classifications and multiple memberships, are readily incorporated by the addition of the relevant random effects into the MOI (3c), together with the corresponding parameter sampling steps. In fact the example application in Section 8 includes random effects.

For multivariate models (3c) becomes a multivariate model that is updated accordingly. Goldstein ([19], chapter 6) discusses the steps involved. Structural equation models can also be incorporated ([19], chapter 8).

Finally, in our example we illustrate the implications of properly allowing for measurement errors. We also highlight the issue of providing good estimates for the distribution of measurement errors, notably the variance. Often, such estimates are known only very approximately and one possibility is to use an informative prior (which we have not investigated here) or as we have done in our example, carried out a sensitivity analysis over a plausible range of values. This highlights those parameter estimates that were relatively unaffected by our adjustment procedures. Ecob and Goldstein [8] explore a number of approaches to the estimation of measurement error distributions, and this is an important area for further work.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] W.J. Browne, H. Goldstein, G. Woodhouse, and M. Yang, *An MCMC algorithm for adjusting for errors in variables in random slopes multilevel models*, Multilevel Model. Newslett. 13 (2001), pp. 4–9.

[2] J.P. Buonaccorsi, *Measurement Error: Models, Methods and Applications*, Chapman and Hall–CRC Press, Boca Raton, 2010, xxvi + 438 pp. ISBN 978-1-420-06656-0.

[3] J.R. Carpenter and M.G. Kenward, *Multiple Imputation and its Application*, Wiley, Chichester, 2013.

[4] R.J. Carroll, D. Ruppert, L.A. Stefanski, and C. Crainiceanu, *Measurement Error in Nonlinear Models: A Modern Perspective*, Chapman & Hall, Boka Raton, FL, 2006.

[5] C.M.J. Charlton, D.T. Michaelides, R.M.A. Parker, B. Cameron, C. Szmaragd, H. Yang, Z. Zhang, A.J. Frazer, H. Goldstein, K. Jones, G. Leckie, L. Moreau, and W.J. Browne, *Stat-JR Version 1.0*, Centre for Multilevel Modelling, University of Bristol & Electronics and Computer Science, University of Southampton, 2013.

[6] D. Clayton, *Models for the analysis of cohort and case-control studies with inaccurately measured exposures*, in *Statistical Models for Longitudinal Studies on Health*, J.H. Dwyer, M. Feinlieb, P. Lippert, and H. Hoffmeister, eds., Oxford University Press, Oxford, 1992, pp. 301–331.

[7] J. Cumming and H. Goldstein, *Handling attrition and non-response in longitudinal data with an application to a study of Australian youth*, Longit. Lifecourse Stud. J. 7 (2016), pp. 53–63. doi:10.14301/llcs.v7i1.342.

[8] R. Ecob and H. Goldstein, *Instrumental variable methods for the estimation of test score reliability*, J. Educ. Statist. 8 (1983), pp. 223–241.

[9] L. Feinstein, J. Jerrim, A. Vignoles, H. Goldstein, R. French, E. Washbrook, R. Lee, and R. Lupton, *Comment and debate: Social class differences in early cognitive development*, Longit. Life Course Stud. 6 (2015). doi:10.14301/llcs.v6i3.361.

[10] W.A. Fuller, *Measurement Error Models*, Wiley, New York, 1987.

[11] H. Goldstein, *Some models for analysing longitudinal data on educational attainment (with discussions)*, J. Roy. Statist. Soc. Ser. A 142 (1979), pp. 407–442.

[12] H. Goldstein, J.R. Carpenter, and W.J. Browne, *Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms*, J. Roy. Statist. Soc. Ser. A 177 (2014), pp. 553–564. doi:10.1111/rssa.12022.

[13] H. Goldstein, J. Carpenter, M. Kenward, and K. Levin, *Multilevel models with multivariate mixed response types*, Statist. Model. 9 (2009), pp. 173–197.

[14] H. Goldstein and D. Kounali, *Multilevel multivariate modelling of childhood growth, numbers of growth measurements and adult characteristics*, J. R. Statist. Soc. A 172(Part 3) (2009), pp. 599–613.

[15] H. Goldstein, D. Kounali, and A. Robinson, *Modelling measurement errors and category misclassifications in multilevel models*, Statist. Model. 8 (2008), pp. 243–261.

[16] MATLAB and Statistics Toolbox Release, The MathWorks, Inc., Natick, Massachusetts, United States, 2007b.

[17] S. Muff and L.F. Keller, *Reverse attenuation in interaction terms due to covariate measurement error*, Biom. J. 57 (2015), pp. 1068–1083. doi:10.1002/bimj.201400157.

[18] S. Muff, A. Riebler, L. Held, H. Rue, and P. Saner, *Bayesian analysis of measurement error models using integrated nested Laplace approximations*, Appl. Statist. 64 (2015), pp. 231–252.

[19] H. Goldstein, *Multilevel Statistical Models*, Wiley, Chichester, 2011.

[20] S. Richardson and W.R. Gilks, *Conditional independence models for epidemiological studies with covariate measurement error*, Stat. Med. 12 (1993), pp. 1703–1722.

[21] D.B. Rubin, *Multiple Imputation for Non Response in Surveys*, Wiley, Chichester, 1987.

[22] A. Skrondal and S. Rabe-Hesketh, *Generalized Latent Variable Modelling: Multilevel*, Longitudinal and Structural Equation Models, Chapman and Hall/CRC, Boca Raton, FL, 2004.