Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and nonlinear terms.

by

Harvey Goldstein*
James R. Carpenter**
and
William J. Browne*

Abstract

This paper extends existing models for multilevel multivariate data with mixed response types to handle quite general types and patterns of missing data values in a wide range of multilevel generalised linear models. It proposes an efficient Bayesian modelling approach that allows missing values in covariates, including models where there are interactions or other functions of covariates such as polynomials. The procedure can also be used to produce multiply imputed complete datasets. A simulation study is presented as well as the analysis of a longitudinal dataset. The paper also shows how existing multi-process models for handling endogeneity can be extended by the proposed framework.

Keywords

Missing data, multilevel, multivariate, latent normal model, endogeneity, multiple imputation, MCMC, multi-process model.

Address for correspondence

Professor H. Goldstein
Graduate School of Education
University of Bristol
Bristol, BS8 1JA
UK
h.goldstein@bristol.ac.uk

*University of Bristol. ** London School of Hygiene and Tropical Medicine.

1. Introduction

Estimation and inference for statistical models fitted to partially observed data has been a subject of considerable interest with a large literature and different approaches implemented in a range of software packages. Following the work of Rubin (1987), multiple imputation has become a particularly popular procedure. In very large datasets, such as census data, this may be done non-parametrically, using hot-deck or related donor methods. However, in typical social science settings imputation using parametric methods is likely to be much more efficient and avoids issues that arise with small donor pools. Furthermore, parametric imputation methods are relatively easy to implement, at least in their basic form. More generally, a key attraction of multiple imputation is that after imputation we fit our model of interest to the imputed 'complete' data using standard software.

Two versions of multiple imputation are widely used. The first, developed by Rubin (1987) uses the joint posterior distribution of all the variables with missing data when sampling values to fill in the data gaps. The second, (van Buuren, 2007) known as imputation by chained equations uses the conditional distribution for each variable in turn, conditioned on the remaining variables in the model of interest. One advantage of the latter approach is that non-normal variables, in particular discrete variables are readily handled. Goldstein, Carpenter, Kenward and Levin, ((GCKL), 2009) and Carpenter, Goldstein and Kenward, (2011) show how such variables can also be handled by using a latent normal approach that links the different response types through an underlying multivariate normal distribution, that is thus consistent with a range of response distributions. Imputation by chained equations also has the drawback that for multilevel data structures it cannot readily handle variables measured at higher levels. The GCKL approach is specifically designed to handle such cases.

Neither of these procedures has been able to properly handle interaction terms, including polynomials. Data analysts have tended to adopt different approaches. In one commonly used approach, which has come to be known as 'passive imputation', all non-linear and interaction terms are omitted from the imputation algorithm, and re-created from the imputed values immediately prior to fitting the model of interest. This approach was criticised by Von Hippel (2009), who instead proposed simply treating each interaction or polynomial term as 'just another variable'. However, Seaman, Bartlett and White (2012) show that under the missing at random (MAR) assumption this performs poorly, especially for binary responses.

In the present paper we extend the GCKL approach to handle covariates in the model of interest which can be quite general functions of other covariates. This includes interactions and polynomial terms, and the covariates can be continuous or discrete. A fully Bayesian modelling procedure is proposed which can also be used to produce multiply imputed datasets for analysis in standard software. We shall refer to this procedure as 'missing covariate data' (MCD).

One motivation for this approach was an analysis of the 1958 birth cohort data (Carpenter and Plewis, 2011), exploring the effects of early life and socio-economic variables on educational achievement in early adulthood. Attrition and wave non-response mean that only around 65% of the records are complete, making this a natural candidate for analysis via imputation. Preliminary analysis of the complete records indicates a potentially important non-linear effect of mother's age, which further interacts with social housing provision. However, as discussed above, existing imputation procedures do not perform acceptably in this setting. After presenting and evaluating our proposal, we return to this example in Section 8.

In the next section we introduce a simple version of our model for normally distributed covariates. This is followed by an extension to a general framework that allows for categorical covariates and responses and a discussion of endogeneity. A simulation is presented, followed by an analysis of data from the National Child Development Study. We conclude with a summary and discussion.

2. Normal responses and predictors: a simple model

Consider first a simple linear regression model for Y on X where the joint distribution of (Y,X) is bivariate normal and their joint posterior distribution is

$$f(Y,X|priors) = f(Y|X,priors)f(X|priors) \quad (1)$$

We can write the model as

$$Y = \beta_0 + \beta_1 X + e_{Y|X}$$

$$X = \alpha_0 + e_X \quad (2)$$

We assume missing data are MAR, so we do not need to model the missingness mechanism. We additionally assume $e_{Y|X}, e_X$ are independent. If this is not the case (often known as endogeneity since it implies that the predictor X in the model is correlated with the residual) then the model lacks identifiability: we return to this in Section 7 below.

The case where we have missing values in Y is standard, assuming MAR. We sample any missing values in Y by drawing them from the posterior predictive distribution. Within a Bayesian framework we have a Gibbs step at each iteration that samples the missing value from $N(\beta_0 + \beta_1 X, \sigma^2_{e_{Y|X}})$ using the current MCMC chain parameter values. GCKL (2009) extend this to the case where Y is multivariate, and includes models where the responses are of different types, for example normal and binary. For a wide range of models of interest with missing data, including multilevel models, GCKL show that we can carry out the multiple imputation by setting up an imputation model where any X variables that have missing values are treated as responses, imputations are carried out and a set of complete, imputed, datasets are returned.

The X variables can be normal or categorical (ordered or unordered) or non-normal continuous where a transformation to normality, such as the Box-Cox transformation, is available. The MCMC estimation algorithm, at each cycle, samples a 'latent' normally distributed variable for each non-normal variable in such a way as to ensure multivariate normality, and then randomly imputes normally distributed missing values. For the original non-normal missing values the imputed values are then obtained on the original scales. Each of the imputed datasets is then used to estimate the model of interest and the resulting parameter estimates are combined according to Rubin's rules (Rubin, 1987) to form final estimates with associated standard errors. For model (2) this imputation model can be written

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \Omega\right), \quad \Omega = \begin{pmatrix} \sigma^2_Y & \\ \sigma_{XY} & \sigma^2_X \end{pmatrix} \quad (3)$$

While (3) is satisfactory in this and other cases, it cannot deal with models where there are polynomial terms in the model of interest. Likewise, where there are several covariates, model (3) cannot deal with interaction terms. If the original joint distribution of covariates is multivariate normal, as in the GCKL procedure, then the joint distribution that includes polynomial or interaction terms will not have a multivariate normal distribution, so that the

imputation model assumptions are not satisfied. Likewise, as pointed out in the introduction, other procedures, such as imputation using chained equations, give unsatisfactory results for such models. In this paper we propose an extension to the GCKL procedure that does allow such terms to be included and which provides estimates that are unbiased as well as making efficient use of the data through multiple imputation.

3. A fully model based approach for missing data in explanatory variables

We shall illustrate our approach using (1) & (2) first and then describe how this generalises.

From (1) the likelihood for the *i*th data record in the sample can be written as

$$L(X_i) = f(Y_i|X_i;\theta_Y)f(X_i;\theta_X) \propto \{(2\pi\sigma^2_{Y|X})^{-0.5}\exp(-\frac{0.5(Y_i-\beta_0-X_i\beta_1)^2}{\sigma^2_{Y|X}})\}\{(2\pi\sigma^2_X)^{-0.5}\exp(-\frac{0.5(X_i-\alpha_0)^2}{\sigma^2_X})\} \quad (4)$$

where (θ_Y, θ_X) are parameter vectors and we assume independence as described above.

We note that the two components of (4) can be combined into a single, normal likelihood term that will allow a Gibbs sampling step for drawing a missing X value, but this does not extend to the case where there are interactions, since we do not have a normal likelihood corresponding to the term $\{(2\pi\sigma^2_X)^{-0.5}\exp(-\frac{0.5(X_i-\alpha_0)^2}{\sigma^2_X})\}$ since this will now include both X and the interactions with X. We shall therefore utilise a Metropolis-Hastings (MH) step. Thus, using, for example, a suitable normal proposal distribution $N(X_{current}, c\sigma^2_X)$ centered on the current value, we would accept a proposed value with probability

$$\min\left\{1, \frac{L(X^*)}{L(X)}\right\} \quad (5)$$

which is a Metropolis random walk step and where * denotes the new proposed value. We shall use a scaling factor c=1 in our examples.

In this formulation we have specified a distribution for X with a mean and variance, and we therefore need to incorporate steps to update these parameters also; in the present case this can be done using standard Gibbs steps.

Most importantly, we now have a procedure that in the present case allows us automatically to handle terms that are functions of X such as powers, since these are incorporated only in the likelihood contribution from the model of interest (i.e. the likelihood for the model of Y|X in our notation above).

4. Several covariates

Consider now the following model with several covariates some or all of which may have missing values. We assume the missing data are MAR. We shall, for simplicity, consider a single normal response and a single level model of interest, since the fitting of multiple responses of different types and further levels of nesting with random effects introduces no new considerations. GCKL (2009) show how this can be done for models without interaction terms and the same additional steps apply for the present model.

Suppose we have p covariates, and that the model of interest is

$$Y_i = f(X_i;\beta) + e_i, \quad \text{where } e_i \text{ are } i.i.d\ N\left(0, \sigma^2_{Y|X}\right), \quad i = 1, \ldots, n \quad (6)$$

and X_i is a $1 \times p$ vector of covariate values for unit i. We further assume X is multivariate normal.

We can allow any pattern of missing values in X across units. Suppose unit i is missing an observation, X_{li}, on covariate l. We factor the joint distribution of X as

$$f(X_1 \ldots . X_p) = f(X_l|X_1, \ldots . X_{l-1}, X_{l+1}, \ldots . X_p) f(X_1, \ldots . X_{l-1}, X_{l+1}, \ldots . X_p)$$

and again use a Metropolis-Hastings step, adopting a symmetric normal proposal distribution for X_{li}. The MH likelihood ratio (5) now only involves the conditional distribution $f(X_l|X_1, \ldots . X_{l-1}, X_{l+1}, \ldots . X_p)$ and the likelihood of the response model derived from (6), since the joint distribution of the remaining X's cancels from the ratio. In particular, as before, we can handle general functions of the covariates including interaction terms and polynomials since these only occur in the likelihood contribution from the model of interest where they are covariates.

We apply this procedure to each missing X_{li} in turn. Then, we update all the other parameters in the model using the appropriate Gibbs sampling steps. Thus, adopting suitable priors, we have a fully Bayesian procedure for estimating the parameters in (6) (as well as the parameters of the joint distribution of X). As with all such Bayesian procedures, we obtain both the posterior distribution of the parameters of interest and the draws of the missing data given the observed, which can be used to form imputed datasets, by taking the values every nth iteration, for use with multiple imputation. Clearly, if (6) is the model of interest the use of multiple imputation is not necessary. However, we may have models such as (6) which include variables that we do not wish to condition our model of interest on, or where we are only imputing a part of the dataset (say a specific treatment or exposure group). In such cases it will often be convenient to generate imputed datasets and to adopt a multiple imputation approach. This may also be preferred if completed datasets are to be provided for secondary analysis excluding sensitive variables, or where data analysts have access only to standard computational software.

5. Non-normal covariates

As pointed out above, GCKL (2009) show how the joint distribution of a set of variables of different types can be sampled using MCMC steps to produce a (latent) multivariate normal distribution. We apply the same approach to partially observed covariates of different types in the present model. Once we have a latent normal structure, missing values on the latent normal scale can be imputed and then converted back to the original scale, and used to evaluate the model of interest component of the likelihood in the Metropolis Hastings ratio. The particular advantage of the latent normal model is that the specification of the conditional distributions in (6) is straightforward.

We illustrate this for a binary covariate, say Z_1 corresponding to a latent normal variable $X_1 \sim N(\mu_X, 1)$. The steps involved are as follows, and further details are given by GCKL (2009):

1. Given the current estimate of the covariance matrix Ω_X and known values of Z_1, draw a sample value x_{1i}^* from
   $f(X_1 | X_2, \ldots . X_p; Z_1; \Omega_X) = f(X_1^*) \sim N(\mu_{X_1^*}, \sigma_{X_1^*}^2)$
   where x_{1i}^* is sampled from the truncated normal distribution on $[0, \infty]$ if $z_{1i} = 1$ and the truncated normal $[-\infty, 0]$ if $z_{1i} = 0$. The parameters $\mu_{X_1^*}, \sigma_{X_1^*}^2$ are updated at each iteration, along with all the other parameters in the model. Alternatively, we can sample x_{1i} from $N(\mu_{X_1^*}, \sigma_{X_1^*}^2)$ and accept if $x_{1i} > 0$ and $z_{1i} = 1$ or if $x_{1i} \leq 0$ and $z_{1i} = 0$, otherwise retain current value.

2. For missing data values propose a new value using a normal proposal distribution centered on the current value with, say, unit variance.
3. In a Metropolis step compare each new value for X_1 with the existing value, using the Metropolis ratio as described in section 3, evaluated at the proposed and current value. If accepted, derive the corresponding value of Z_1 as $z_{1i} = 1 \; if \; x_{1i} > 0, \;\; z_{1i} = 0 \; if \; x_{1i} \leq 0$.
4. The remaining steps are as for normal data. In our examples below we have chosen, for the normally distributed *X* variables, a normal proposal distribution with variance equal to that estimated from the non-missing values of the variable.

If data are multicategory, ordered or unordered, similar latent normal transformations are available. GCKL also describe the use of Box-Cox transformations for skewed continuous distributions and Goldstein and Kounali (2009) describe a transformation for the Poisson distribution. Thus a very wide class of covariates can be accommodated within the present framework.

6. Non-normal responses for the model of interest

So far we have considered normal response variables for the model of interest. Suppose, however, that we have non-normal responses. For example, if we have a binary response then, assuming a probit link, we can add an extra step at each iteration in the algorithm to sample a latent normal variable, as in GCKL (2009) which then becomes the response. This is as in step 1 in Section 5 above. The steps for sampling any covariate values which are missing are then as described in Sections 4 & 5. Thus the likelihood component for the response in the model of interest, $f(Y|X;\theta_Y)$, is based on the multivariate latent normal distribution of the responses, including any imputed responses and the steps are as for the case with observed multivariate normal responses.

Putting this altogether, we have a multilevel multivariate response model, with responses of different types, with a covariate structure with the flexibility to handle non-linear relationships and interactions, also of different types. Within this framework, missing values can be handled in either responses or covariates.

In the next section we discuss briefly the case where our model is miss-specified with a dependency between the residual terms of the model of interest and the model for the covariates, the 'endogeneity' case.

7. Endogenous models of interest

For simplicity, consider model (2) and the case where $e_{Y|X}, e_X$ are not independent. In terms of the mean and covariance structure we now have a model with 3 fixed coefficients, 2 variances and a covariance, i.e. 6 parameters. In fact there are only 5 free parameters from the joint distribution $f(Y,X)$, two means, two variances and a covariance, so that the separate parameters cannot be identified. If, now, we introduce a further variable so that we have a joint distribution $f(Y,X_1,X_2)$ with the model of interest as before, not depending on X_2, and a model for X_1 then we can write

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + e_{Y|X_1} \\
X_1 &= \alpha_0 + \alpha_1 X_2 + e_{X_{12}} \\
X_2 &= \gamma_0 + e_{X_2} \qquad (7) \\
E\left(e_{Y|X_1} e_{X_{12}}\right) &\neq 0, \quad E\left(e_{X_{12}} e_{X_2}\right) = E\left(e_{Y|X_1} e_{X_2}\right) = 0
\end{aligned}
$$

This yields 5 coefficients, three variances and a covariance for the lack of independence between $e_{Y|X_1}$ and $e_{X_{12}}$, giving 9 parameters, and the joint distribution has 3+6=9 parameters so that we do now have identifiability. This can be extended to include further auxiliary variables to increase efficiency, and these can also have missing values that can be imputed, given that identifiability is assured for the parameters in the model of interest. What we have done is to introduce an 'auxiliary' or 'instrumental' variable into the model for X_1 that does not appear in the model of interest and is uncorrelated with the residuals in the models for Y, X_1. If we have such auxiliary variables then we can estimate models where the assumption of independence between specific covariates and the residual in the model of interest can be relaxed, so called endogenous variable models.

Model (7) is a multi-process model and a special kind of multivariate model that can in fact be fitted directly, for normal as well as non-normal variables, using the methods described in GCKL (2009). The extension described in the present paper allows missing values to occur in any of the variables, including the case where there are interaction terms.

8. Examples

In section 8.1 we describe a simple simulation that demonstrates the properties of our proposed MCD procedure, and in Section 8.2 we then apply it to a real dataset that contains appreciable amounts of missing data in main effects and interaction terms.

8.1 A simulation

We simulate data based loosely upon the structure of a real educational dataset that has been used extensively to illustrate multilevel models – the 'tutorial dataset described by Goldstein (2011, Chapter 3). This consists of 4059 students grouped within 65 schools in Inner London. We simulate from the following two-level model of interest with a zero intercept and two explanatory variables having a bivariate normal distribution:

$$y_{ij} = \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_j + e_{ij} \qquad (8)$$

$$\begin{pmatrix} x_{1ij} \\ x_{2ij} \end{pmatrix} \sim N\left(\begin{matrix} 0 \\ 0 \end{matrix}, \Omega\right), \quad \Omega = \begin{pmatrix} 1 & \\ 0.5 & 1 \end{pmatrix}, \quad u_j \sim N(0, \sigma_u^2), \quad e_{ij} \sim N(0, \sigma_e^2)$$

$$\beta_1 = 0.5, \quad \beta_2 = 0.5, \quad \sigma_u^2 = 0.1, \quad \sigma_e^2 = 0.5$$

where we assume that the residuals are independent.

Default (diffuse) priors are used as follows:

$$p(\beta) \propto 1, \quad p(\Omega) \propto 1,$$

$$p(\sigma_u^2) \sim \Gamma^{-1}(0.001, 0.001), \quad p(\sigma_e^2) \sim \Gamma^{-1}(0.001, 0.001)$$

In the literature there are several commonly used diffuse priors for such models (see Browne and Draper (2006) and Gelman (2006) for discussion on this) but this is not of prime interest in this paper and the dataset is large enough for choice of prior to be less important.

We introduce, at random, 20% missing values in both covariates, giving, on average, 36% of records with at least 1 missing value and 4% with missing values in both variables. There are no missing values for *Y*. The first four columns of table 1 show the means and standard errors of the means for the parameter estimates from 100 replications of fitting the model to data simulated from (8), with each run using a burn in of 500 and chain of 1000 iterations. The

first column shows results of using listwise deletion when a record contains any missing values, and as expected, shows no discernible biases. The results in the second column of this table also show no discernible biases for the MCD procedure. The third column of results provides estimates when we ignore the correlation (0.5) between the covariates and update the missing values using only the marginal covariate distributions. The fourth column shows the results from model (8) where x_2 becomes a binary variable by treating all values greater than zero as 1 and values less than or equal to zero as 0, using a latent normal formulation so that the probability of a 1 is 0.5. In addition we use a binary response (Z) with a probit link function, so that $Z = 0$ if $Y < 0$, else $Z = 1$. The level 1 variance is now set to 1.0 to match the assumption made for the latent normal distribution. The fifth column of Table 1 gives results for the following extension of (8):

$$y_{ij} = \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{1ij} x_{2ij} + \beta_4 x_{2ij}^2 + u_j + e_{ij} \qquad (9)$$

$$\beta_1 = 0.5, \quad \beta_2 = 0.5, \quad \beta_3 = 0.5, \quad \beta_4 = 0.5, \quad \sigma_u^2 = 0.1, \quad \sigma_e^2 = 0.5$$

We also simulated, in column 6, the 'just another variable' procedure where the interaction term and the squared term were given missing values whenever one of the constituent variables were missing and normality is assumed for the interaction and squared terms.

Finally we simulated model (8), with no missing data. Running 100 simulations as in Table 1 we obtain values (standard errors) for $\beta_1, \beta_2, \sigma_u^2$ of 0.498 (0.003), 0.506 (0.004), and 0.104 (0.004) respectively.

For the procedures proposed in the present paper these simulations give results that are not significantly different from the expected population values. In the marginal sampling case there are biases in the fixed coefficients of about 5% that result from assuming uncorrelated covariates and these biases increase as the correlation increases so that for a correlation of 0.9 the biases are of the order of 8%. For the case where listwise deletion is used, we expect the coefficient estimates to be unbiased, but the average standard errors to be higher. Thus for the model in the second column of Table 1 the average standard errors for the fixed coefficients were 0.0385, 0.0140 and 0.0142 and for the listwise deletion model in the first column they were 0.0411, 0.0160 and 0.0161, an average increase of 11%. For the 'just another variable' analysis there are downward biases in the estimates for β_2 and β_3, about 3.5% for the interaction term, and an upward bias of 4% for the level 1 variance estimate.

(**Table 1 here**)

8.2 Data from the National Child Development Study (NCDS)

Our second example uses data from the National Child Development Study (Carpenter and Plewis, 2011).

The NCDS target sample size at birth in 1958 was 17,634 and this had reduced, as a result of death and permanent emigration, to 15,885 by age 23. The size of the observed sample at age 23 was 12,044 with 1,837 cases lost from the target sample as a result of attrition and wave non-respondents. Thus 24% of the target sample at age 23 were missing, and the pattern of missing data in the NCDS is not monotone, with more than half the missing cases at age 23 reappearing in the observed sample at later waves. Of the missing cases at age 23, about one-third were due to non-cooperation, while two-thirds were either not located or not contacted.

Of the 15,885 possible cases, 22 have all covariates and the response missing and these have been omitted from the analysis.

Table 2 shows the results of fitting the model given in Table 23.3 of Carpenter and Plewis (2011) using MCMC with diffuse priors as described above.
The model is as follows, using a probit link function

$$pr(y_i = 1|(X\beta)_i) = \int_{(X\beta)_i}^{\infty} \phi(t)dt \qquad (10)$$

where $\phi(t)$ is the standard normal density function. The response y_i for person i is whether or not the respondent has an educational qualification at age 23 and the predictors (X) are measures of 'in care status at age 7' (binary), 'In social housing at age 7' (binary), 'inverse birth weight' (continuous in ounces multiplied by 100), and 'mother's age at birth of cohort member' (continuous years centered at age 28 years). Inverse birthweight has been rescaled (by multiplying it by 100) in order to avoid a very unbalanced covariance matrix for the covariates which can result in very poor chain mixing. For the complete case analysis A, there is an interaction between mother's age and housing, and an interaction between mother's age squared and housing. The response has 24% of values missing and the predictors range from 1.3% to 13.6% having missing values. All these missing values are imputed using an extra MCMC step (on the latent normal scale) at each iteration, as described in Section 6.

(Table 2 here)

The parameter chains all mix well, with moderate first order correlations between successive sampled parameter values.

Of particular interest is the interaction between age and social housing, because this suggests that for mothers in social housing, the linear component of age is removed (coefficient for mother's age -0.016, coefficient for interaction, 0.014).

Carpenter and Plewis (2011) report a number of multiple imputation analyses using a full conditional specification approach. In the first, non-linear and interaction terms are omitted from the imputation model. Resulting coefficients for age, age-squared and their interactions with social housing are all attenuated. In a second analysis, age-squared is included in the imputation model, which alleviates the attenuation of the age coefficients. However, since the interaction involves social housing, which is itself partially observed, imputation congenial with the full interaction structure is problematic under full conditional specification. Instead, Carpenter and Plewis disregarded individuals with missing social housing before imputing separately in each social housing category, including additionally in the imputation model, auxiliary variables recorded while children were at school. Our results from model B are similar, although this does not include auxiliary variables. However, our approach here does not require imputation. Note that if we wish, we can include additional auxiliary variables in the model, generate imputed datasets, and then fit the simpler model of interest to the imputed data, combining the results for inference using Rubin's rules (Rubin, 1987).

In this analysis, there is little difference in the parameter estimates. However, with our fully Bayesian missing data procedure we include information from all individuals with missing data, resulting in generally smaller standard errors, by up to about 30%. In particular, the significance of the interaction of mother age and social housing is confirmed.

9. Discussion

We have shown how quite general multilevel models with missing response and covariate data can be fitted. The model of interest can be multivariate with different types of responses, continuous and discrete, and covariates that likewise can be of mixed types and can include interactions and general functions of variables that have missing data values. The MCD model also allows multiply imputed datasets to be created and fitted in the usual way using suitable combination rules. This may be useful in some circumstances, for example where secondary analysis is to be carried out with only standard software packages, or where the model of interest does not include auxiliary variables used as covariates in the MCD model, which may also be in a causal pathway. Alternatively, we may consider instead including auxiliary variables as additional responses, without conditioning on them as covariates in the model of interest. From the classic two-stage multiple-imputation point of view, our procedure ensures that the model of interest and the imputation model are congenial (Meng, 1994). This is because the imputation model is defined as the analyst's model of interest (ie a conditional distribution of response given covariates) multiplied by the joint distribution of the covariates. Ensuring congeniality in this way avoids the problems of existing approaches, which often attenuate the interaction or non-linear effects, at the expense of overestimating the main effects or residual variance. This is borne out by our simulation study (Table 1), and in our example, where the attenuation reported by Carpenter and Plewis (2011) is avoided without recourse to discarding records with missing social housing values.

In most situations (Carpenter and Kenward, 2013, Ch 2) we wish the imputation model to be at least as rich as the model of interest, although in certain settings (Mitra and Reiter, 2011) this may not be appropriate. One interesting extension of our approach would be to link with the approach proposed by Mitra and Dunson (2010), who demonstrate potential gains of model averaging over a set of imputation models, possibly specifying appropriate priors for parameters in the imputation model.

We consider that our procedure overcomes a number of major issues that have been discussed in the literature, as follows:

First, as demonstrated by GCKL (2009), there is no need to assume multivariate normality for the variables entering the model. The latent normal formulation can handle mixtures of discrete, normal and non-normal continuous variables in either the responses or covariates. Secondly, in a general multilevel framework, we can have variables defined at any level of a data hierarchy or within cross classifications or multiple membership structures (Goldstein, 2011, Chapter 13). Thirdly, the ability to handle interactions and general functions of variables overcomes existing difficulties in multiple imputation procedures about how to handle these. Generally speaking these currently are either treated as separate variables to be imputed or the basic variables are imputed and functions of them formed from the imputed values. Seaman et al., (2012) have shown that neither approach is satisfactory.

We have also demonstrated in our simulation that it is necessary to utilise the joint distribution of covariates when imputing, rather than using just the marginal distributions.

Finally, we have shown how auxiliary variables can be used to model endogeneity whereby one or more covariates may be correlated with the residual vector in the model of interest.

Our procedures allow such endogeneity models to handle missing values in any of the variables involved in the joint model.

In our simulations we have used both normal and probit models of interest. In fact, our procedure is designed for quite general response distributions in the model of interest. As regards imputation of the covariates, the probit link functions, used by GCKL and in the present paper, facilitate the computations through the use of a latent multivariate normal distribution. The alternative use of logistic link functions could potentially be accommodated through formulating a latent multivariate logistic distribution. This would, however, entail the assumption of a logistic distribution for the continuous variables also, which does not seem to us a useful line to pursue. Furthermore, in practice, imputed data from these two alternatives are practically indistinguishable unless the fitted probabilities are very close to 0 or 1 (Carpenter and Kenward, 2013, p. 95).

The software used is an extension of the freely available REALCOM software (Realcom, 2011) which was designed to fit the models described in GCKL (2009). The procedures described in the present paper are being incorporated in new software, Stat-JR (Charlton et al., 2012), being developed at the Universities of Bristol and Southampton (see http://www.bristol.ac.uk/cmm/software/statjr/index.html for more details).

10. Acknowledgements

James Carpenter was funded by ESRC research fellowship RES-063-27-0257.

Harvey Goldstein was partly funded by ESRC grant RES-062-23-2265 as part of the National Centre for Research Methods.

William Browne was partly funded by ESRC grant RES-062-23-2265 as part of the National Centre for Research methods and ESRC grant RES-149-25-1084 as part of the Digital Social Research programme.

We would like to thank the referees and associate editor for their helpful comments.

11. References

Browne, W. J. and Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models (with discussion). *Bayesian Analysis* 1: 473-550

Carpenter, J.R., Goldstein, H. and Kenward, M.G. (2011). "REALCOM-IMPUTE Software for Multilevel Multiple Imputation with Mixed Response Types." Journal of Statistical Software **45**,4,1-14.

Carpenter, J. R. and Kenward M. G (2013) *Multiple imputation and its application.* Chichester: Wiley.

Carpenter, J. and Plewis, I. (2011). Analysing Longitudinal Studies with Non-response: Issues and Statistical Methods. In: Malcolm Williams and Paul Vogt (Editors). *The SAGE handbook of Innovation in Social Research Methods*. London, Sage.

Charlton, C.M.J., Michaelides, D.T., Cameron,B. Szmaragd, C. Parker, R.M.A., Yang,H. Zhang, Z. and Browne, W.J. (2012) *Stat-JR software*. Center for Multilevel Modelling, University of Bristol and Department of Electronics and Computer Science, University of Southampton.

Gelman, A.E. (2006). Prior distributions for variance parameters in hierarchical models (Comment on article by Browne and Draper). *Bayesian Analysis* 1: 513-533

Goldstein, H. (2011). Multilevel Statistical Models. Fourth Edition. Chichester, Wiley.

Goldstein, H., J. Carpenter, Kenward, M. and Levin, K. (2009). "Multilevel Models with multivariate mixed response types." Statistical Modelling. **9**(3): 173-197.

Goldstein, H. and D. Kounali (2009). "Multivariate multilevel modelling of childhood growth, numbers of growth measurements and adult characteristics." Journal of the Royal Statistical Society, A **172**(3): 599-613.

Meng, X.L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). Statistical Science, 10,538-573.

Mitra, R. and Dunson, D. (2010). Two-Level Stochastic Search Variable Selection in GLMs with Missing Predictors. *International Journal of Biostatistics* 6:. DOI: 10.2202/1557-4679.1173

Mitra, R. and Reiter, J.P. (2011). Estimating propensity scores with missing covariate data using general location mixture models. *Statistics in Medicine* 30: 627-641

Realcom (2011). Developing multilevel models for REAListically COMplex social science data. http://www.bristol.ac.uk/cmm/software/realcom

Rubin, D. B. (1987). Multiple imputation for non response in surveys. Chichester, Wiley.

Seaman, S.R., Bartlett, J.W. and White, I.R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: evaluation of statistical methods. BMC *Methodology*, **12**, 46.

Van Buuren S (2007). Multiple imputation of discrete and continuous data by fully conditional specification" Statistical Methods in Medical Research, 16(3), 219-242

Von Hippel, P.T. (2009). "How to impute interactions, squares and other transformed variables." Sociological methodology **39**(1): 265-291.

	Table 1. Average parameter estimates over 100 replications with data simulated under models described in the text. Burn in=500, iterations=1000. Standard errors across simulations in brackets.					
Parameter (true value)	Listwise deletion from model (8)	Data from Model (8) using conditional sampling	Data from Model (8) using marginal sampling	Data from Model (8) with binary response using conditional sampling, x_2 binary	Data from model (9) with interactions (conditional sampling)	Data from model (9) with interactions (conditional sampling) using 'just another variable'
β_0 (0)	-0.003 (0.005)	0.005 (0.004)	0.010 (0.005)	-0.005 (0.005)	0.000 (0.004)	0.006 (0.005)
β_1 (0.5)	0.498 (0.002)	0.498 (0.001)	0.521 (0.001)	0.501 (0.001)	0.498 (0.001)	0.499 (0.002)
β_2 (0.5)	0.501 (0.002)	0.501 (0.001)	0.524 (0.001)	0.501 (0.003)	0.501 (0.001)	0.490 (0.002)
β_3 (0.5)					0.496 (0.002)	0.483 (0.002)
β_4 (0.5)					0.503 (0.002)	0.501 (0.001)
σ_u^2 (0.1)	0.102 (0.002)	0.101 (0.002)	0.106 (0.002)	0.098 (0.002)	0.100 (0.002)	0.103 (0.002)
σ_e^2 (0.5)	0.502 (0.001)	0.500 (0.001)	0.496 (0.001)	0.501 (0.001)	0.501 (0.001)	0.521 (0.001)

Table 2. Fitting the NCDS data. Burn in =1000, iterations =2500. Probit link function. Standard errors in brackets.

Estimate (% missing)	A	B
Intercept	-1.568 (0.103)	-1.544 (0.074)
In care (12.6)	0.650 (0.096)	0.639 (0.085)
Social Housing (13.6)	0.574 (0.037)	0.572 (0.034)
(100 *) Inverse birth weight (4.5)	0.745 (0.109)	0.737 (0.081)
Mothers age at birth (1.3)	-0.016 (0.004)	-0.016 (0.003)
Mothers age squared	0.0019 (0.0005)	0.0019 (0.0005)
Age * housing	0.013 (0.0053)	0.013 (0.0053)
Age squared * housing	-0.00065 (0.00067)	-0.00078 (0.00066)

Model A is a complete case analysis N=10,279 (65% of sample). Model B uses the MCD model. Response is whether respondent had an educational qualification at age 23 (24% missing values). Note that, as a result of our rescaling, the inverse birthweight coefficient is smaller by a factor of 100 than the value given by Carpenter and Plewis (2011). Carpenter and Plewis use maximum likelihood estimation for the complete records analysis that yields slightly smaller standard error estimates for some parameters.