(A version of this paper was published as: Goldstein, H. and Woodhouse, G. (1996). *Multilevel models with missing data.* Eleventh International workshop on statistical modelling., Orvieto, Italy).

Efficient estimation with missing data in multilevel models

by

Harvey Goldstein

and

Geoffrey Woodhouse

Institute of Education

20 Bedford Way

London, WC1H 0AL, U.K.

Summary

This paper proposes a procedure for producing consistent and asymptotically efficient moment-based estimators for a multilevel model with randomly missing explanatory or response variables. The procedure is essentially an extension of existing procedures based upon multiple imputation but is computationally efficient and avoids the need to generate multiple data sets to which multilevel models are fitted. It is also shown that the procedure copes effectively with informatively missing data values when the missingness mechanism is incorporated into the estimation procedure. It has close parallels with moment-based estimators for errors in variables models.

Some key words

Imputation, missing data, multilevel.

Acknowledgements

This work was supported by the Economic and Social Research Council under its programme for the Analysis of Large and Complex Datasets. Versions of this paper were presented to the 11*th* International Workshop on Statistical Modelling, Orvieto, July 1996 and at the annual Nordic statistical meeting in Lahti, August 1996.

1. Introduction

A characteristic of many data sets is that some of the intended measurements are unavailable. For example, this may occur through chance or because certain questions are left unanswered by particular groups of respondents. An important distinction is made between situations where the existence of a missing data item can be considered a random event and those situations where it is informative and the result of a non-random mechanism. Randomly missing data may be missing ‘completely at random’ or ‘at random’ conditionally on the values of other measurements. This paper will be concerned with these two types of random event. Where data cannot be assumed to be missing at random, one approach is to attempt to model the missingness mechanism, and then to impute the missing values from this model. Such imputations or predictions can be treated in similar ways to those described below.

We consider the problem of missing data in two parts. First we describe a procedure, appropriate for multilevel data, for predicting data values which are missing. Then we study ways of obtaining model parameter estimates from the resulting ‘filled-in’ or ‘completed’ data set.

Two common procedures for dealing with missing data are either to exclude all data records where any variable value is missing (listwise deletion) or to substitute ‘plausible’ (imputed) values for the missing items. In multiple regression modelling missing data are sometimes handled by computing the required covariance matrix from all pairwise non-missing data. This procedure, however, can lead to inconsistencies. Where data items are missing completely at random, the listwise deletion procedure will produce unbiased estimates but these will be inefficient when a high proportion of records is excluded. Methods based upon the use of plausible or imputed values underlie most recommended procedures and detailed discussions are given by Rubin (1987) and Little (1992). The method to be described in the present paper is an elaboration of these procedures and is designed to provide consistent estimators which are also computationally more efficient. We are interested in multilevel models but first review the problem in relation to a single-level model for simplicity.

The procedure has two stages. In common with other imputation procedures the first stage involves the creation of a completed data set based on a model which produces predicted values for the missing items. The second stage produces the required parameter estimates based upon the results of the first-stage model.

2. Creating and using a completed data set

Consider the ordinary single-level univariate linear model for the *i*-th unit

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i \qquad (1)$$

Suppose that some of the X_{1i} are missing completely at random (MCAR) or conditionally missing at random (MAR) conditional on X_2 and suppose for the moment that there are no missing values of Y_i: these are allowed in the general model described in Section 7. Label the unknown values X_{1i}^*. The first stage of our procedure considers the estimation of these by predicting them from the remaining observations and the parameter set θ for the prediction model

$$x_{1i} = E(X_{1i}^* | X_1, X_2, Y, \theta) \qquad (2)$$

If the data are MAR conditional on X_2 then this variable should be conditioned on when obtaining the predictions as well as being incorporated in (1) and we shall assume in such cases that X_2 has no missing values. Where we have no missing values on Y or on X_2 then the prediction (2) is simply the linear regression of X_1 on X_2, Y. Several procedures, including the EM algorithm, are available for obtaining estimates of the missing values (Little, 1992). Little and Schluchter (1985) also consider the case of discrete variables. By considering the joint distribution of all the (non-conditioned) variables, in particular assuming multivariate Normality, these procedures will also yield maximum likelihood (ML) estimates of the mean vector and covariance matrix. In the ordinary linear model case these are sufficient statistics and can be used to obtain estimates and inferences for models such as (1). In this paper we propose an alternative method which can be used with multilevel models where the simplicity of a single level model does not apply.

We note in passing that in the MAR case, if X_2 is treated as a ‘response’ belonging to the multivariate Normal structure then the resulting estimates are not consistent. Nevertheless, the incorporation of X_2 into the prediction stage may be expected to improve the estimates, and we shall illustrate this with an example.

Having obtained a completed data set as the first stage of the procedure, one possibility would be to use this directly in the usual way to estimate the parameters of (1) in a second stage. Unfortunately, this gives biased parameter estimates because the filled-in data are attenuated, having too small a variability. A standard alternative procedure is to correct for this by generating random variables which are added to the predicted values so that they have the correct (on average) distributional properties. These are generated from the residual covariance matrix estimated from the multivariate regression given by (2). One cycle of this procedure generates a new complete data set which is analysed in the usual way. Several cycles are executed and suitable averages of the resulting estimates are used for inference, and this is known as multiple imputation (Rubin, 1987).

Multiple imputation in practice has certain drawbacks, especially in the context of complex multilevel data structures. Computational time can be an important factor and the need to carry out several analyses for each model is a disadvantage. This creates difficulties for model exploration, and may be a particular problem with secondary data where different analysts, often with limited resources, wish to work on the same data set. Instead we propose for the second stage a procedure which produces consistent interval and point estimates in a single analysis of the data set.

As we have already pointed out, in the single level Normal linear model, imputation procedures are unnecessary since we can obtain sufficient statistics using any suitable ML procedure such as an EM algorithm. In the general multilevel case we can also, in principle, obtain sufficient statistics for use during each cycle of an iterative algorithm, for example one based upon an existing EM algorithm or an iterative generalised least squares (IGLS) algorithm (Bryk and Raudenbush, 1992, Goldstein 1995). In practice, however, this approach has limitations. There is a combinatorial increase in the number of sufficient statistics beyond models with 3 levels which creates severe computational problems. A similar difficulty occurs if a general complex variance structure is fitted at the lowest level. We do not, therefore, pursue this, but instead concentrate on the procedure outlined below.

For our simple example the second stage of the procedure considers a model relating the unknown missing value to its expectation given by (2) as

$$X_{1i} = x_{1i} + m_{1i} \tag{3}$$

where the 'residuals' m_{1i} are unknown but where we have estimates of their variances and covariances obtained from the multivariate regression defined by (2) and are assumed to have zero means. This model is similar to the basic model used for errors of measurement (Fuller, 1987) except that the role of X_{1i} is now that of the 'true' value which is unknown. If we assume that the two terms on the right-hand side of (3) are uncorrelated, then we have

$$\text{var}(X_1) = \text{var}(x_1) + \text{var}(m_1) \tag{4}$$

In effect, multiple random imputation works by generating sets of values of the m_{1i} from their estimated distribution. As opposed to the standard multiple imputation procedure, we propose to work directly with estimates of the quantities in (4), generalised to the multilevel case and extended to further variables, to produce consistent estimators, and the exposition in Section 7 parallels closely that for obtaining consistent estimators for the errors in variables model.

3. The multilevel model

We give here a brief exposition of the multilevel model, since this is used in both stages of the imputation procedure. For simplicity we describe the 2-level model, and to make it concrete we can think of the data as consisting of pupils nested within schools. All our results generalise straightforwardly to three or more levels. For a general introduction to the topic of multilevel models see Bryk and Raudenbush (1992), Longford (1993) or Goldstein (1995). These models are now routinely applied to the analysis of hierarchically structured data such as arise in the human sciences.

Consider the following 2-level model:

$$\begin{aligned}
&Y = X\beta + E\\
&Y = \{y_{ij}\},\ \ X = \{X_{ij}\},\ \ X_{ij} = \{x_{0ij}, x_{1ij}, \ldots . x_{pij}\}\\
&E = E_1 + E_2,\quad E_1 = \{e^{(1)}_{\mathrm{ij}}\},\quad E_2 = \{e^{(2)}_{\mathrm{j}}\},\\
&e^{(1)}_{ij} = \sum_{h=0}^{q_1} z^{(1)}_{hij} e^{(1)}_{hij},\ \ e^{(2)}_{j} = \sum_{h=0}^{q_2} z^{(2)}_{hij} e^{(2)}_{hj}\\
&Y \text{ is } N \times 1,\ \ X \text{ is } N \times (p+1),\ \ N = \sum_j n_j
\end{aligned} \qquad (5)$$

And n_j is the number of level 1 units in the *j*-th level 2 unit. We also write

$$\begin{aligned}
&e_{hij} = e^{(1)}_{hij},\ \ u_{hj} = e^{(2)}_{hj}\\
&e_{ij} = \{e_{hij}\}_{(q_1 \times 1)},\ \ u_j = \{u_{hj}\}_{(q_2 \times 1)}\\
&z^{(1)}_{ij} = \{z^{(1)}_{hij}\}^{\mathrm{T}}{}_{(1 \times q_1)},\ \ z^{(2)}_{ij} = \{z^{(2)}_{hij}\}^{\mathrm{T}}{}_{(1 \times q_2)}\\
&Y = X\beta + Z^{(2)}u + Z^{(1)}e
\end{aligned} \qquad (6)$$

where $Z^{(2)}, Z^{(1)}$ are explanatory variable matrices at levels 2 and 1, with (*i*,*j*)th rows $z^{(2)}_{ij}, z^{(1)}_{ij}$ respectively, and the vectors *u* and *e* take the values u_j and e_{ij} for the (*i*,*j*)th unit.

The residual matrices E_1, E_2 have expectation zero with

$$\begin{aligned}
&\mathrm{E}[E_1 E_1^{\mathrm{T}}] = V_{(1)},\quad \mathrm{E}[E_2 E_2^{\mathrm{T}}] = V_{(2)}\\
&\mathrm{E}[E_1 E_2^{\mathrm{T}}] = 0,\qquad V = V_{(1)} + V_{(2)}
\end{aligned} \qquad (7)$$

Further details are given in Goldstein (1995, Appendix 2.1)

In the usual way we can form posterior estimates of the random variables and their covariance matrix

$$\hat{u}_{hj} | Y, \beta, \Omega_u, \Omega_e, \qquad \mathrm{cov}(\hat{u}_{hj} - u_{hj}) \tag{8}$$

and likewise for those at other levels. The IGLS algorithm is a convenient procedure for producing ML or REML estimates under multivariate Normality, since it can efficiently handle multivariate data, many levels and complex level 1 variance structures with a high computational burden. In the standard case, however, with just 2 or 3 levels and a single variance describing level 1 variation other algorithms such as EM can be used and may be more efficient.

4. Discrete variables with missing data

Suppose we have one or more categorical explanatory variables as well as continuous variables where both types may have missing values. The categorical variables will normally be converted to a series of dummy or indicator variables, one fewer in number than the number of categories. For each categorical variable we need to obtain, in the first stage of the procedure, the predicted probabilities of belonging to each category, corresponding to each dummy variable used in the second stage. These are then substituted for the missing values to form the completed data set. Note that where the proportions in each category are not too extreme, we may be able to obtain satisfactory predictions simply by assuming multivariate Normality and we study this in our examples.

More generally, we can use models which consider the joint distribution of categorical and discrete responses (Goldstein, 1995, Chapter 7), but that is not pursued here.

5. Estimation for the multilevel model with imputed data

In this section we give estimation details for the second stage of the procedure following the prediction of the missing values. We give details for the 2-level model; extensions to higher-level models are straightforward. We write

$$
\begin{gathered}
Y_{ij} = y_{ij} + q_{ij} \\
X_{hij} = x_{hij} + m_{hij} \\
m_{hij} = m_{e(h)ij} + m_{u(h)j} \\
\mathrm{cov}(x_{hij} m_{e(h)ij}) = \mathrm{cov}(x_{hij} m_{u(h)j}) = 0 \\
\mathrm{cov}(q_{ij} q_{i'j}) = \mathrm{cov}(m_{e(h)ij} m_{e(h)i'j}) = 0 \\
\mathrm{cov}(m_{e(h)ij} m_{u(h)j}) = \mathrm{cov}(m_{u(h)j} m_{u(h)j'}) = 0 \\
\mathrm{E}(q_{ij}) = \mathrm{E}(m_{e(h)ij}) = \mathrm{E}(m_{u(h)j}) = 0 \\
\mathrm{var}(m_{e(h)ij}) = \sigma^{ij}_{e(h)m} \\
\mathrm{cov}(m_{e(h_1)ij} m_{e(h_2)ij}) = \sigma^{ij}_{e(h_1h_2)m} \\
\mathrm{var}(m_{u(h)j}) = \sigma^{j}_{u(h)m} \\
\mathrm{cov}(m_{u(h_1)j} m_{u(h_2)j}) = \sigma^{j}_{u(h_1h_2)m} \\
\mathrm{var}(q_{ij}) = \sigma^{ij}_{q} \\
\mathrm{cov}(q_{ij} m_{hij}) = \sigma^{ij}_{hq}
\end{gathered}
\tag{9}
$$

for the h-th explanatory variable $(h = 1,2,\ldots,p)$ with vector $m_h = \{m_{hij}\}$ of *discrepancies* at levels 1 and 2, and with q as the discrepancy vector for the response. We are here using the term ‘discrepancy’ to refer to the difference between the true unknown value and the (first stage) imputed value. We use upper case for the true values and lower case for the predicted values, which are equal to the true values if the variable is not missing.

As in (6) we write the ‘true’ model in the general form

$$Y_{ij} = (X\beta)_{ij} + (Z^{(2)}u)_{ij} + (Z^{(1)}e)_{ij} \tag{10}$$

where $(X\beta)_{ij}$ is the (*i,j*)th element of $X\beta$, etc., and Y, X, $Z^{(2)}$, $Z^{(1)}$, u and e are as defined for equations (5) and (6). We assume that the $Z^{(1)}$, $Z^{(2)}$ are known, as is the case for variance component models. Where $Z^{(1)}$, $Z^{(2)}$ contain variables that have missing data the estimation becomes considerably more complicated (see below) and this is the subject of further research.

For the predicted values we have

$$\begin{aligned} &y_{ij} = (m\beta)_{ij} - q_{ij} + (x\beta)_{ij} + (Z^{(2)}u)_{ij} + (Z^{(1)}e)_{ij} \\ &m = \{m_h\} \end{aligned} \tag{11}$$

For the true values write

$$\begin{aligned} &M_{XX} = X^{\mathrm{T}}V^{-1}X, \quad M_{XY} = X^{\mathrm{T}}V^{-1}Y \\ &\hat{\beta} = M_{XX}^{-1}M_{YY} \end{aligned} \tag{12}$$

with *V* as defined in (7) and for the moment assumed known (Goldstein, 1995, Appendix 2.1).

We have

$$\begin{aligned} X^{\mathrm{T}}V^{-1}X &= (x+m)^{\mathrm{T}}V^{-1}(x+m) \\ &= x^{\mathrm{T}}V^{-1}x + m^{\mathrm{T}}V^{-1}x + x^{\mathrm{T}}V^{-1}m + m^{\mathrm{T}}V^{-1}m \end{aligned} \tag{13}$$

$$\begin{aligned} X^{\mathrm{T}}V^{-1}Y &= (x+m)^{\mathrm{T}}V^{-1}(y+q) \\ &= x^{\mathrm{T}}V^{-1}y + m^{\mathrm{T}}V^{-1}y + x^{\mathrm{T}}V^{-1}q + m^{\mathrm{T}}V^{-1}q \end{aligned}$$

We assume that the discrepancies are random variables with finite fourth moments and uncorrelated with the predicted values, and so

$$\begin{aligned}\mathrm{E}(X^{\mathrm{T}}V^{-1}X) &= x^{\mathrm{T}}V^{-1}x+\mathrm{E}(m^{\mathrm{T}}V^{-1}m)\\ \mathrm{E}(X^{\mathrm{T}}V^{-1}Y) &= x^{\mathrm{T}}V^{-1}y+\mathrm{E}(m^{\mathrm{T}}V^{-1}q)\end{aligned} \tag{14}$$

Thus, to estimate the fixed parameters we require the expectations on the right hand side of (14) and we now consider how to obtain these where data are missing at both level 1 and level 2. We then consider the problem of obtaining estimates of the random parameters required to form *V*. We now make use of our assumption that $Z^{(1)}, Z^{(2)}$ contain no missing data so that there are no discrepancies in *V* which need to be considered when taking expectations.

For level-1 imputed estimates of explanatory variables based upon residuals estimated from the first stage of the procedure, the (h_1,h_2)th element of $\mathrm{E}(m^{\mathrm{T}}V^{-1}m)$ is

$$\begin{gathered}\sum_{ij=1}^{N}\sigma^{ij}\sigma^{ij}_{e(h_1h_2)m}\\ \text{with }\ C_{\Omega_1}=\{\sum_{ij}\sigma^{ij}\sigma^{ij}_{e(h_1h_2)m}\}_{\ (p\times p)}\end{gathered} \tag{15}$$

where N is the total number of level 1 units and σ^{ij}, $\sigma^{ij}_{e(h_1h_2)m}$ respectively are the (*i*,*j*)th diagonal element of V^{-1} and the estimated covariance between the level-1 discrepancies for variables h_1, h_2 for the (*i*,*j*)th unit.

The *h*-th element of the vector $\mathrm{E}(m^{\mathrm{T}}V^{-1}q)$ is

$$\begin{gathered}\sum_{ij}\sigma^{ij}\sigma^{ij}_{hq}\\ \text{with }\ C_{\Omega_{1q}}=\{\sum_{ij}\sigma^{ij}\sigma^{ij}_{hq}\}_{\ (p\times 1)}\end{gathered} \tag{16}$$

For the level-2 discrepancies we have the corresponding correction matrix for the explanatory variables given by

$$C_{\Omega_2} = \sum_j \{(J^{*\mathrm{T}}_{n_j(h_1h_2)} V_j^{-1} J^{*}_{n_j(h_1h_2)})\sigma^j_{u(h_1h_2)m}\}_{(p\times p)}, \qquad (17)$$

The $(n_j \times 1)$ vector $J^{*}_{n_j(h_1h_2)}$, defined for each level-2 unit and for each pair of variables $h_1,\ h_2$, has *i*-th element equal to 1 if for the *i*-th level-1 unit variables h_1, h_2 are both missing, and zero otherwise. The term $\sigma^j_{u(h_1h_2)m}$ is, as in (15), the estimated covariance (or variance) of the level-2 discrepancies for variables h_1, h_2, and V_j^{-1} is the *j*-th block of V^{-1}.

We now form the corrected matrices and estimate the fixed parameters

$$\begin{aligned} \hat{M}_{XX} &= x^{\mathrm{T}}V^{-1}x + C_{\Omega_1} + C_{\Omega_2} \\ \hat{M}_{XY} &= x^{\mathrm{T}}V^{-1}y + C_{\Omega_{1q}} \\ \hat{\beta} &= \hat{M}_{XX}^{-1}\hat{M}_{XY} \end{aligned} \qquad (18)$$

The discrepancy covariance matrices used in the formation of C_{Ω_1}, C_{Ω_2} and $C_{\Omega_{1q}}$ are those given by (9) and are termed the ‘conditional’ ones by Goldstein (1995, Appendix 2.2). We note the similarity with the errors in measurement case where the discrepancy matrices are subtracted rather than added in (18).

In the single level case where there is just a single explanatory variable with missing data, these results reduce to the following. Order the completed data so that the imputed observations are grouped together first. Then, ignoring any correction for sampling variation, the estimate $\hat{M}_{XX}$ becomes

$$(X^T X) + \begin{pmatrix} n_1 \hat{\sigma}_m^2 & 0 \\ 0 & 0 \end{pmatrix}$$

where there are n_1 predicted values. This is very similar to the correction described by Beale and Little (1975), although these authors use an estimate based upon the actual observed residuals calculated from the complete data cases and they approximate the covariance matrix by $\hat{M}_{xx}^{-1}$ (see below).

We now consider the estimation of the parameters of *V*. We write

$$\begin{aligned} \tilde{y}_{ij} &= y_{ij} - (x\beta)_{ij} = (Z^{(2)}u)_{ij} + (Z^{(1)}e)_{ij} + (m\beta)_{ij} - q_{ij} \\ \tilde{y} &= \{\tilde{y}_{ij}\}, \quad Z_u = \{(Z^{(2)}u)_{ij}\}, \quad Z_e = \{(Z^{(1)}e)_{ij}\} \end{aligned}$$

Unlike the measurement error case the discrepancies *m* are correlated with the true values and hence also with the residuals Z_u, Z_e which are defined by (16) in terms of the true values. The discrepancies are uncorrelated with the observed and predicted values so that we can write

$$\begin{aligned} 0 &= \mathrm{E}[(y-(x\beta))(\beta^{\mathrm{T}} m^{\mathrm{T}})] = \mathrm{E}[Z_u(\beta^{\mathrm{T}} m^{\mathrm{T}})] + \mathrm{E}[Z_e(\beta^{\mathrm{T}} m^{\mathrm{T}})] + \mathrm{E}[m\beta\beta^{\mathrm{T}} m^{\mathrm{T}}] - \mathrm{E}[q\beta^{\mathrm{T}} m^{\mathrm{T}}] \\ 0 &= \mathrm{E}[(y-(x\beta))q^{\mathrm{T}}] = \mathrm{E}[Z_u q^{\mathrm{T}}] + \mathrm{E}[Z_e q^{\mathrm{T}}] + \mathrm{E}[m\beta q^{\mathrm{T}}] - \mathrm{E}[qq^{\mathrm{T}}] \end{aligned}$$

This leads to

$$\begin{aligned} &\mathrm{E}(\tilde{y}\tilde{y}^{\mathrm{T}}) = V - Q - T_1 - T_2 + 2Q_\beta \\ &T_1 = \bigoplus_{ij}(\hat{\beta}^{\mathrm{T}}\Omega_{eijm}\hat{\beta}), \qquad T_{2(h_1h_2)} = \bigoplus_{j}[\sigma^{j}_{u(h_1h_2)m}(\hat{\beta}^{\mathrm{T}}\hat{\beta})(J^{*}_{n_j(h_1h_2)}J^{*\mathrm{T}}_{n_j(h_1h_2)})], \\ &Q_\beta = \bigoplus_{ij}\hat{\beta}^{\mathrm{T}}\sigma_q^{ij}, \quad \sigma_q^{ij} = \{\sigma_{hq}^{ij}\}, \quad \Omega_{eijm} = \{\sigma^{ij}_{e(h_1h_2)m}\} \\ &Q = \bigoplus_{ij}\sigma_q^{ij}, \quad T_2 = \sum_{h_1h_2}\delta_{h_1h_2}T_{2(h_1h_2)}, \quad \delta_{h_1h_2} = \begin{cases} 1 \text{ if } h_1 = h_2 \\ 2 \text{ if } h_1 \neq h_2 \end{cases} \end{aligned} \tag{19}$$

Thus the quantity $V_c = Q + T_1 + T_2 - 2Q_\beta$ should be added to the cross-product matrix $\tilde{y}\tilde{y}^{\mathrm{T}}$ based upon the predicted values at each iteration, to form Y^* as in (8) for the estimation of the random parameters (Goldstein, 1995, Appendix 2.1).

When we estimate the covariance matrix of the estimated coefficients we need to condition on the known values. Our sampling scheme implies that the predicted values and hence the discrepancies are resampled along with the responses. We therefore have the following for the covariance matrix of the fixed parameters

$$\hat{M}_{XX}^{-1}\,\mathrm{E}[(x^{\mathrm{T}}V^{-1}\tilde{y})(x^{\mathrm{T}}V^{-1}\tilde{y})^{\mathrm{T}}]\hat{M}_{XX}^{-1}$$

which gives the covariance matrix as

$$\hat{M}_{XX}^{-1}\,\mathrm{E}[x^{\mathrm{T}}V^{-1}\tilde{y}\tilde{y}^{\mathrm{T}}V^{-1}x]\hat{M}_{XX}^{-1} \tag{20}$$

We can produce so-called sandwich or robust estimators (Goldstein, 1995, Chapter 3) for both fixed and random parameters. For the fixed parameters the sandwich estimators are obtained by substituting directly into (19) the corresponding sample quantities, using the observed cross-product matrix of residuals with contributions from each block. The model-based estimator is

$$\hat{M}_{XX}^{-1}(x^{\mathrm{T}}V^{-1}x - x^{\mathrm{T}}V^{-1}V_cV^{-1}x)\,\hat{M}_{XX}^{-1} \tag{21}$$

This expression ignores the sampling variation associated with the prediction function for the missing values.

For the random parameters the covariance matrix is given by

$$(Z^{*\mathrm{T}}V^{*-1}Z^*)^{-1}Z^{*\mathrm{T}}V^{*-1}\,\mathrm{cov}(Y^{**})V^{*-1}Z^*(Z^{*\mathrm{T}}V^{*-1}Z^*)^{-1}$$

and the sandwich estimator is obtained by substituting for $\mathrm{cov}(Y^{**})$ the observed (corrected) matrix derived from the contributions from each level-2 block indexed by *j*, namely

$$\bigoplus_{j}(\hat{Y}_j^{**}\hat{Y}_j^{**\mathrm{T}})$$
$$\hat{Y}_j^{**}=\mathrm{vec}(\hat{Y}_j^{*}-\mathrm{E}(\hat{Y}_j^{*})),\quad \hat{Y}_j^{*}=\tilde{y}_j\tilde{y}_j^{\mathrm{T}}+V_{c,j},\quad \mathrm{E}(\hat{Y}_j^{*})=V_j \tag{22}$$

6. An example using educational test scores

In this set of analyses we fit the following 2-level variance components model to data from the Junior School Project (JSP), (Mortimore *et al*., 1988).

$$Y_{ij}=\beta_0+\beta_1X_{1ij}+\beta_2X_{2ij}+u_j+e_{ij} \tag{23}$$

where the explanatory variables, respectively, are a Mathematics score taken at 8 years, and social class (non-manual, manual), and the response variable is a Mathematics test score taken when the same students were 11 years old. To carry out our procedure, for the first example analysis, we assume bivariate Normality for Y, X_1 conditional on X_2.

We have carried out simulations by randomly omitting values of the explanatory variables and using the procedures of this paper to estimate the model parameters. In Table 1 we randomly omit 25% of the 8-year Mathematics scores. We compare the model estimates with the 'naive' procedure of using the predicted values without adjusting for prediction error and with the estimates from the full data set.

Table 1. Results of 100 simulations, randomly omitting 25% of 8-year Mathematics scores each time.

Parameter	**A**	**B**	**C**
Fixed			
Intercept	15.25	11.95	15.23
8-year Maths score	0.593 (0.033) (0.043)	0.725 (0.035)	0.596 (0.035) (0.033)
Manual Social class	1.273 (0.44) (0.39)	1.002 (0.43)	1.299 (0.41) (0.38)
Random			
Level 2 variance	4.051 (1.18) (1.05)	4.216 (1.19)	4.034 (1.10) (1.27)
Level 1 variance	27.85 (1.36) (2.51)	25.55 (1.25)	27.86 (1.25) (1.42)

Note: Column A gives results for the full data set. Column B gives results using the ‘naive’ procedure (with model-based standard errors) and column C gives results using the correction procedures of the paper. For analysis A the first set of standard errors are model-based and the second set are sandwich estimates. For analysis C the standard errors in the first parentheses are the means of the sandwich estimates and in the second parentheses are the standard deviations of the simulated parameter estimates. The simulated responses are generated from the known explanatory variable values in the full data set with the same set of missing values for each simulated data set

As pointed out above, the ‘naive’ estimates which use just the completed data set are biased. The increase in the coefficient of the 8-year Maths score is the opposite of what tends to occur in the uncorrected errors in variables case, and this is evident from (18). The full missing data estimates are very close to the full data set estimates and in all cases the latter lie within a 90% confidence interval with respect to the distribution of parameter estimates over replications.

The estimated sampling standard errors of the estimates indicate that the sandwich estimators perform reasonably well for the fixed parameters but tend somewhat to underestimate for the random parameters. Some further exploration of this with other data sets would be useful. When data are missing, the sandwich estimators are based upon residuals which are more stable than they would be if all the data were present, as they are based upon predicted values and the estimators also make more use of model-based quantities than when there is no missing data. This appears to explain the relatively large sandwich standard error for the level-1 variance in analysis A. It should be noted that the standard errors used here for the missing data case do not take account of the variability in the prediction function and will tend to be underestimates. This bias can be estimated from the simulation replications and in our examples is approximately 5%.

Table 2 gives the results for the model of analysis C in table 1, additionally omitting a random 25% of social class values. We now assume multivariate Normality for Y, X_1, X_2, and while this is not strictly true since the latter is discrete, the results suggest that this is a reasonable approximation.

Table 2. Estimates from 100 simulations corresponding to analysis C of Table 1, additionally omitting 25% of Social Class values. Standard errors are sandwich estimates.

Parameter	**Estimate (s.e.)**
Fixed	
Intercept	15.41
8-year Maths score	0.586 (0.034)
Manual Social class	1.317 (0.37)
Random	
Level 2 variance	4.064 (1.18)
Level 1 variance	27.98 (1.21)

The results in Table 2 are similar to those in Table 1 and again the full data set parameter values lie within a 90% confidence interval with respect to the distribution of parameter estimates over replications.

Finally, in Table 3 we present the results of simulating conditionally missing (missing at random) data. The range of the response score is divided into three with approximately a third of the distribution in each subrange. For the lowest third the 8-year maths score is omitted at random with a probability of 0.4: for the middle third with a probability of 0.25 and for the top third with a probability of 0.15, giving an overall proportion missing an 8-year score of approximately 0.25. Such a pattern is not unreasonable in educational data where those who reach low achievement levels are more likely to be missing when testing is carried out. The first column of Table 3 gives the results from 100 simulations where all cases with missing data are excluded, illustrating clearly the biases, especially in the variance estimates, which result from this informatively missing pattern. The second column shows the results from applying the full missing data procedure where the response is included in the prediction function. As pointed out earlier, we do not in general obtain fully consistent estimates in this case. Nevertheless, as before, we do obtain estimates close to the full data set values and for which, apart from the social class coefficient, the 90% confidence intervals include the population values.

Table 3. Estimates from 100 simulations corresponding to analysis C of Table 2, with conditionally missing at random values of 8-year score (see text). Analysis A is the result of omitting cases with missing data and analysis B gives results from the full missing data procedure. Standard errors are sandwich estimates.

	A	**B**
Parameter	**Estimate (s.e.)**	**Estimate (s.e.)**
Fixed		
Intercept	16.70	15.26
8-year Maths score	0.557 (0.037)	0.589 (0.036)
Manual Social class	1.209 (0.50)	1.137 (0.43)
Random		
Level 2 variance	3.406 (1.14)	4.075 (1.06)
Level 1 variance	26.13 (1.50)	27.95 (1.26)

7. Discussion

The ability to obtain model estimates with measurements missing at random allows us to make efficient use of survey data where typically many measurements are missing. The requirement for random missingness is important, as with all procedures for dealing with missing data, but a substantive discussion of this is outside the scope of this paper. A review is given by Rubin (1987). From a design point of view, however, our procedures raise the interesting possibility that surveys can be designed with data deliberately missing at random. Such designs are variously known as rotation or matrix designs. The analysis of such designs, however, typically has been restricted to the efficient estimation of means or of model parameters where the ‘missingness’ has been confined to the response variables (see for example Goldstein and James, 1983). Using the methods of the present paper, we are able to analyse models where there is any (random) missing pattern among response or explanatory variables. This is useful, for example, when we wish to restrict the length of a survey by rotating questions, or where some data are available only on a subsample of individuals for reasons of cost or opportunity.

In our example of informatively missing data we obtain estimates which are much improved over those based upon casewise deletion. One important application here is with longitudinal data where, over time, subjects with atypical (for example small) values drop out of a study (Diggle and Kenward, 1994). By including all the data available and predicting for those occasions where data are missing we may expect improved estimates.

Further exploration of data where such informatively missing mechanisms operate is currently being planned. Further work is also underway for the case where there are random coefficients for variables with missing data and for the case of discrete data where probabilities are extreme, using mixed response generalised linear models. It is also important to study models with more complex covariance structures, with different sizes and ratios of variances. The performance of the sandwich estimators also needs exploring and comparing with model based estimators.

All the analyses were carried out using the multilevel software program *MLn* (Rasbash and Woodhouse, 1995).

References

Beale, E. M. L. and Little, R. J. A. (1975). Missing values in multivariate analysis. *J. Royal Statist. Soc.*, B, 37, 129-45.

Breslow, N.E. and Clayton, D. G. (1993). Approximate inference in generalised linear mixed models. *J. American Statistical Association*, 88, 9-25.

Bryk, A. S., and Raudenbush, S. W. (1992). *Hierarchical Linear Models.* Newbury Park, Sage.

Diggle, P. and Kenward, M. G. (1994). Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics*, 43, 49-93

Fuller, W. A. (1987). *Measurement Error Models.* New York, Wiley.

Goldstein, H. (1991). Nonlinear multilevel models with an application to discrete response data. *Biometrika*, 78, 45-51.

Goldstein, H. (1995). *Multilevel Statistical Models*. London, Edward Arnold: New York, Wiley.

Goldstein, H. and James, A. (1983). Efficient estimation for a multiple matrix design. *Brit. J. Math. and Statist. psychol*, 36, 167-74.

Goldstein, H. and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. J. Royal Statistical Society, A, 159, (to appear)

Kass, R. E., & Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models*). J. American Statistical Association*, 84, 717-26.

Little, R. J. A. (1992). Regression with missing X's: a review. *J. American Statistical Association*, 87, 1227-37.

Little, R. J. A. and Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* **72**: 497-512.

Longford, N. T. (1993). *Random Coefficient Models*. Oxford, Clarendon Press.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D. and Ecob, R. (1988). *School Matters*. Wells, Open Books.

Rasbash, J. and Woodhouse, G. (1995). *MLn Command Reference*. London, Institute of Education.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, Wiley.

Woodhouse, G., Yang, M., Goldstein, H., & Rasbash, J. (1996). Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society, A*, 159, 201-212.