Journal of Official Statistics, Vol. 28, No. 2, 2012, pp. 279–302

Strategy for Modelling Nonrandom Missing Data Mechanisms in Observational Studies Using Bayesian Methods

Alexina Mason[1], *Sylvia Richardson*[1], *Ian Plewis*[2], *and Nicky Best*[1]

Observational studies inevitably suffer from nonresponses and missing values. Bayesian full probability modelling provides a flexible approach for analysing such data, allowing a plausible model to be built which can then be adapted to carry out a range of sensitivity analyses. In this context, we propose a strategy for using Bayesian methods for a "statistically principled" investigation of data which contains missing covariates and missing responses, likely to be nonrandom.

The first part of this strategy entails constructing a "base model" by selecting a model of interest, then adding a submodel to impute the missing covariates followed by a submodel to allow informative missingness in the response. The second part involves running a series of sensitivity analyses to check the robustness of the conclusions. We implement our strategy to investigate some typical research questions relating to the prediction of income, using data from the UK Millennium Cohort Study.

Key words: Longitudinal analysis; cross-sectional analysis; sensitivity analysis; Millennium Cohort Study; income; nonresponse; attrition.

1. Introduction

Social science data typically suffer from nonresponse and missing values, which often render standard analyses misleading. Cross-sectional studies tend to be rife with missing data problems, and studies which are longitudinal inevitably lose members over time in addition to other sources of missingness. As a consequence, researchers generally face the problem of analysing datasets complicated by missing covariates and missing responses. The appropriateness of a particular analytic approach is dependent on the mechanism that led to the missing data, which cannot be determined from the data at hand. Given this uncertainty, researchers are forced to make assumptions about the missingness mechanism and are strongly recommended to check the robustness of their conclusions to alternative

[1] Imperial College London, School of Public Health, Faculty of Medicine, Department of Epidemiology and Biostatistics, LONDON W2 1PG, UK. Emails: a.mason05@imperial.ac.uk, sylvia.richardson@imperial.ac.uk and n.best@imperial.ac.uk

[2] University of Manchester, Social Statistics, Oxford Road Manchester M13 9PL, UK.
Email: ian.plewis@manchester.ac.uk

Acknowledgments: Financial support: this work was supported by an ESRC PhD studentship (Alexina Mason). Nicky Best and Sylvia Richardson would like to acknowledge support from ESRC: RES-576-25-5003 and RES-576-25-0015. Sylvia Richardson and Nicky Best are investigators in the MRC-HPA Centre for Environment and Health. The authors are grateful to the Centre for Longitudinal Studies, Institute of Education for the use of MCS data and to the UK Data Archive and Economic and Social Data Service for making them available. However, they bear no responsibility for the analysis or interpretation of these data.

plausible assumptions. A number of different approaches to this task have been proposed and determining a way forward can be daunting for the analyst.

An extensive literature has built up on the topic of missing data, with the various methods, covering both cross-sectional and longitudinal studies, catalogued and reviewed in papers (Schafer and Graham 2002; Ibrahim et al. 2005), as well as detailed in comprehensive textbooks (Schafer 1997; Little and Rubin 2002; Molenberghs and Kenward 2007; Daniels and Hogan 2008). Broadly speaking, there are two types of methods for handling missing data: ad hoc methods and "statistically principled" methods. Ad hoc methods, such as complete case analysis or single imputation, are generally not recommended because, although they may have the advantage of relative simplicity, they usually introduce bias and do not reflect statistical uncertainty. By contrast, so-called "statistically principled" or "model-based" methods combine the available information in the observed data with explicit assumptions about the missing value mechanism, accounting for the uncertainty introduced by the missing data. These include maximum likelihood methods which are typically implemented by the EM algorithm, weighting methods, multiple imputation and Bayesian full probability modelling.

In this article, we provide guidance to the analyst on the practicalities of modelling incomplete data using Bayesian full probability modelling. We propose a modelling strategy and apply this to investigate two questions relating to the prediction of mother's income, using data from the first two sweeps of the most recent British birth cohort study, the Millennium Cohort Study (MCS). Specifically, for mothers who are single at the start of the study, we look at (i) the income gains from higher education and (ii) changes in pay rates associated with acquiring a partner. In Section 2, we introduce some of the key definitions relating to missing data and briefly describe a Bayesian approach to modelling data with missing values. Our proposed modelling strategy is then described in Section 3, and is compared with alternative modelling strategies in Section 4. In Section 5, we apply this strategy to our illustrative example, discuss possible modifications and the circumstances where these would be necessary in Section 6, and conclude in Section 7.

2. Bayesian Full Probability Modelling of Missing Data

The appropriateness of a particular missing data method is dependent on the mechanism that leads to the missing data and the pattern of the missing data. From a modelling perspective, it also makes a difference whether we are dealing with missing response, missing covariates or missingness in both the response and covariates. Following Rubin (Rubin 1976), missing data are generally classified into three types: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Informally, MCAR occurs when the missingness does not depend on observed or unobserved data, in the less restrictive MAR it depends only on the observed data, and when neither MCAR or MAR hold, the data are MNAR.

In longitudinal studies, nonresponse can take three forms: unit nonresponse (sampled individuals are absent from the outset of the study), wave nonresponse (where an individual does not respond in a particular wave but reenters the study at a later stage) and attrition or drop-out (where an individual is permanently lost as the study proceeds), and these may have different characteristics (Hawkes and Plewis 2006).

Also, different kinds of nonresponse can often be distinguished, typically not located, not contacted and refusal. Missing data patterns may be further complicated by data missing on particular items (item nonresponse) or on a complete group of questions (domain nonresponse).

Bayesian full probability modelling provides a flexible method of incorporating different assumptions about the missing data mechanism and accommodating different patterns of missing data. A full probability model is a joint probability distribution relating all the observed quantities (observed data) and unobserved quantities (including statistical parameters, latent variables and missing data) in a problem (Gelman et al. 2004). For analysing data with missing values, it entails building a joint model consisting of a model of interest and one or more models to describe the missingness mechanism and to impute the missing values, and such models can be implemented using Markov Chain Monte Carlo (MCMC) methods. By estimating the unknown parameters and the missing data simultaneously, this method ensures that their estimation is internally consistent.

Since the required joint models are built in a modular way, they are easy to adapt, facilitating sensitivity analysis which is crucial when the missing data mechanism is unknown. The Bayesian formulation also has the advantage of allowing the incorporation of additional information through informative priors when relevant.

Suppose the data for our research consist of a univariate outcome y_i and a vector of covariates $x_{1i}, \ldots, x_{pi}$, for $i = 1, \ldots, n$ individuals, and we wish to model this data using a linear regression model assuming normal errors. Then the Bayesian formulation of our model of interest, $f(\boldsymbol{y}|\beta, \sigma)$ is

$$\begin{aligned} &y_i \sim N(\mu_i, \sigma^2) \\ &\mu_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ki} \qquad (1) \\ &\beta_0, \beta_1, \ldots, \beta_p, \sigma^2 \sim \text{prior distribution} \end{aligned}$$

where N denotes a normal distribution. Suppose also that the response contains missing values such that $\boldsymbol{y}$ can be partitioned into observed, $\boldsymbol{y}_{obs}$, and missing, $\boldsymbol{y}_{mis}$, values, i.e., $\boldsymbol{y} = (\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis})$. Now define $\boldsymbol{m} = (m_i)$ to be a binary indicator variable such that

$$m_i = \begin{cases} 0: & y_i \quad \text{observed} \\ 1: & y_i \quad \text{missing} \end{cases}$$

and let θ denote the unknown parameters of the missingness function. The joint distribution of the full data, $f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \boldsymbol{m}|\beta, \sigma, \theta)$, can be factorised as

$$f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \boldsymbol{m}|\beta, \sigma, \theta) = f(\boldsymbol{m}|\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \theta) f(\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}|\beta, \sigma) \qquad (2)$$

suppressing the dependence on the covariates, and assuming that $(\boldsymbol{m}|\boldsymbol{y}, \theta)$ is conditionally independent of (β, σ), and $(\boldsymbol{y}|\beta, \sigma)$ is conditionally independent of θ, which is usually reasonable in practice. This factorisation of the joint distribution is known as a selection model (Schafer and Graham 2002). The missing data mechanism is termed *ignorable* (Little and Rubin 2002) for a Bayesian inference about (β, σ) if

- the missing data are MAR, i.e., $f(\boldsymbol{m}|\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \theta) = f(\boldsymbol{m}|\boldsymbol{y}_{obs}, \theta)$
- the parameters of the data model, (β, σ), and the missingness mechanism, θ, are distinct, and
- the priors for (β, σ) and θ are independent.

For a response with missing values, we do not need a missingness model, $f(\boldsymbol{m}|\boldsymbol{y}, \theta)$, provided we can assume that the missing data mechanism is ignorable. In this case the imputation of $\boldsymbol{y}_{mis}$ is unnecessary for valid inference about β and σ. However, if we cannot assume that the missing data mechanism is ignorable, then we need to specify a response model of missingness and impute the missing *ys*.

The situation is different for covariates with missing values. In this case, an imputation model for the missing data is required to fully exploit all the available data, regardless of our assumptions about the missingness process (see Section 3). As for response variables, if we cannot assume that the missing covariates were generated by an ignorable missing data mechanism, then an appropriate missingness indicator must also be modelled via a third model part.

Our proposed modelling strategy for missing data, described in the next section, allows for missing values in both the response and the covariates. Taking a Bayesian approach has a number of advantages over other commonly used "statistically principled" methods. Firstly, Bayesian models are formulated in a modular way, which lends itself to the iterative modelling strategy, building and then modifying a base model, that we propose. Secondly, uncertainty about the imputed missing values is automatically and coherently propagated through the model and reflected in the parameter estimates of interest. Thirdly, Bayesian models provide scope for including extra data or other information, which can be particularly useful when dealing with suspected nonignorable missingness.

3. Proposed Modelling Strategy

The basic steps in our general strategy for analysing longitudinal or cross-sectional data with missing values are shown in Figure 1. This approach allows the uncertainty from the missing data to be taken into account, and a range of relevant sources of information relating to the question under investigation to be utilised. It can be implemented using currently available software for the Bayesian analysis of complex statistical models, such as WinBUGS (Spiegelhalter et al. 2003).

This strategy consists of two parts: 1) constructing a base model and 2) assessing conclusions from this base model against a selection of well-chosen sensitivity analyses. Each of these is now discussed, drawing attention to the key decisions based on our experience. The proposed strategy allows informative missingness in the response, but assumes that the covariates are MAR. We defer discussion of adaptations, extensions and limitations until Section 6.

3.1. Construct a Base Model

In essence, this part involves building a joint model by starting with a model of interest, and then adding a covariate imputation model followed by a model of response missingness. For each submodel, we recommend that plausible alternative assumptions are noted for use in selecting the sensitivity analyses in the second part. The estimation of some parameters in the two models of missingness can be difficult when there is limited information, but the amount of available information can be increased by incorporating data from other sources and/or expert knowledge. We now look at each step in more detail.

Step 1. Select an initial model of interest (MoI) based on complete cases.

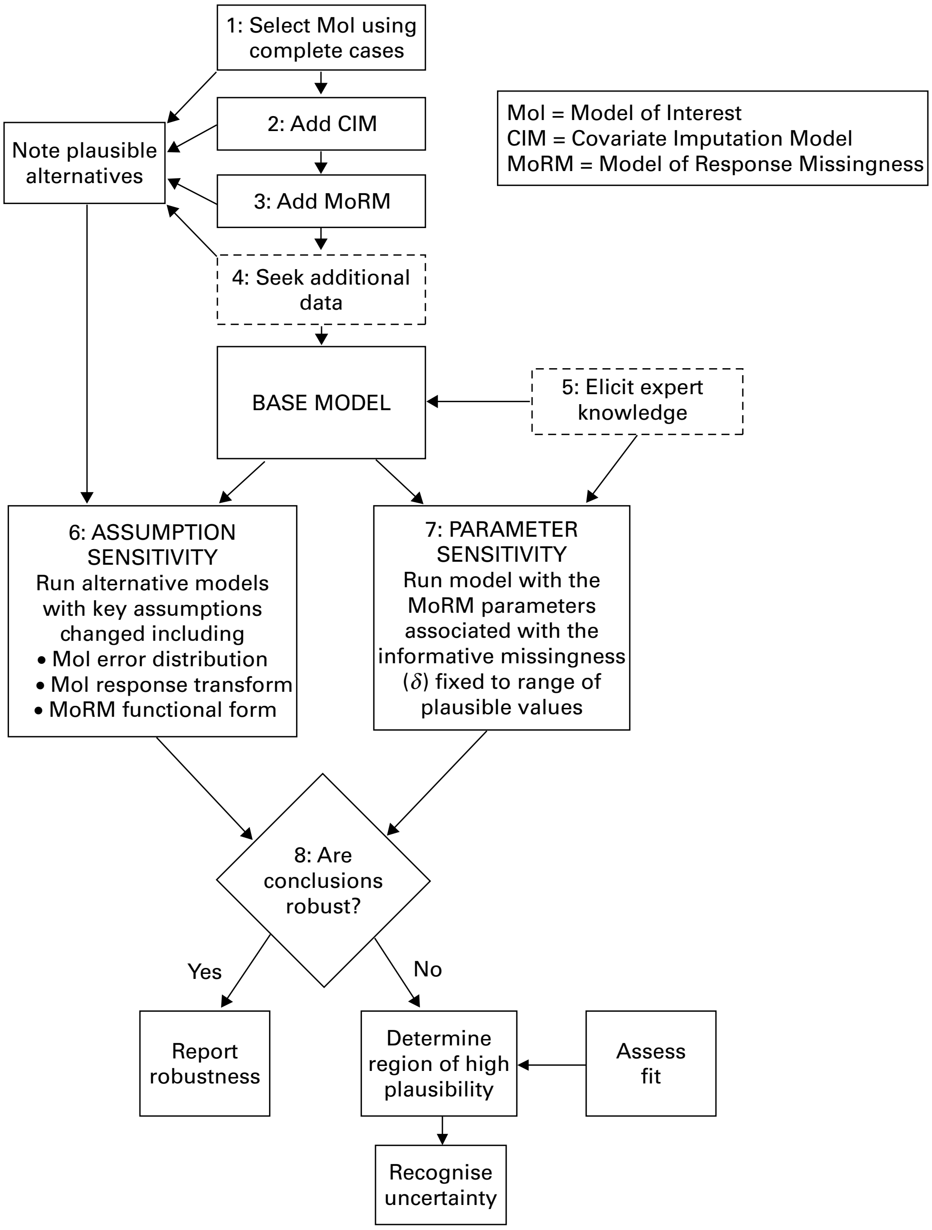
1: Select MoI using complete cases
2: Add CIM
3: Add MoRM
4: Seek additional data
Note plausible alternatives
MoI = Model of Interest
CIM = Covariate Imputation Model
MoRM = Model of Response Missingness
BASE MODEL
5: Elicit expert knowledge
6: ASSUMPTION SENSITIVITY Run alternative models with key assumptions changed including
• MoI error distribution
• MoI response transform
• MoRM functional form
7: PARAMETER SENSITIVITY Run model with the MoRM parameters associated with the informative missingness (δ) fixed to range of plausible values
8: Are conclusions robust?
Yes
No
Report robustness
Determine region of high plausibility
Assess fit
Recognise uncertainty

Fig. 1. Strategy for Bayesian modelling of missing data. The numbers relate to the steps described in Section 3. Dashed boxes indicate optional steps.

The process of building a base model starts with the formation of an initial model of interest using only complete cases and previous knowledge, as for example specified by Equation 1. This includes choosing a transform for the response, model structure and a set of explanatory variables. It may also allow for hierarchical structure and/or other data complexities. In our experience, the most critical assumption is the error distribution of the model of interest, whose misspecification can adversely affect the performance of a selection model (Mason 2009 Ch. 4).

Step 2. Add a covariate imputation model (CIM).
The model of interest will run with missing responses, but not with missing covariates, so to incorporate the incomplete cases the next step is to add a covariate imputation model to produce realistic imputations of any missing covariates simultaneously with the analysis of the model of interest. If there is a single covariate, *x,* there are two obvious ways of building this submodel: i) specify a distribution, e.g., if *x* is a continuous covariate, then specify $x_i \sim N(\nu, \varsigma^2)$ and assume vague priors for ν and ς^2 or ii) build a regression model relating x_i to other observed covariates. For example, if *x* is binary, then it may take the form

$$\begin{aligned} & x_i \sim Bernoulli(q_i) \\ & q_i = \phi_0 + \sum_{k=1}^{s} \phi_r z_{si} \qquad (3) \\ & \phi_0, \phi_1, \ldots, \phi_s \sim \text{prior distribution} \end{aligned}$$

where $z_{1i}, \ldots, z_{si}$ is a vector of fully observed covariates which should include the other covariates which appear in the model of interest, and other variables which are associated with the missingness or explain a considerable amount of variance in *x.* Whether this model is adequate should be checked by comparing the pattern of the imputed values with the observed values.

This submodel will be more complicated when there is more than one covariate with missing values, as is usually the case with real data, and should allow for possible correlation between covariates. A latent variable approach can be used for binary or categorical variables, which can be implemented using a multivariate probit model for binary covariates (Chib and Greenberg 1998) with extensions to ordered categorical variables (Albert and Chib 1993) as required. By creating an underlying set of latent variables, models for mixtures of binary, categorical and continuous variables can be developed (Dunson 2000; Goldstein et al. 2009). Molitor et al. (2009) provide an example of this approach for two binary covariates.

Step 3. Add a model of response missingness (MoRM).
Next, add a model of response missingness to allow informative missingness in the response. Before defining this part of the model, it is important to think about the process that led to the missingness, gathering as much information as possible from the literature and those involved in the data collection process. Then, these findings have to be translated into a statistical model, e.g.,

$$\begin{aligned} & m_i \sim Bernoulli(p_i) \\ & logit(p_i) = \theta_0 + \sum_{k=1}^{r} \theta_k w_{ki} + \delta y_i \qquad (4) \\ & \theta_0, \theta_1, \ldots, \theta_s, \delta \sim \text{prior distribution} \end{aligned}$$

where $w_{1i}, \ldots, w_{ri}$ is a vector of variables which are predictive of the response missingness. It is the inclusion of the response, *y,* (includes the missing values) in this submodel that changes the assumption about the missing responses from MAR to MNAR and provides the link with the rest of the model. The choice of the shape of the relationship between the response and the probability of missingness is important. We need to consider carefully whether a linear relationship is adequate or whether a more complex shape such as that allowed by a piecewise linear functional form would be better. For some datasets, it may well be intuitively plausible that the response is more likely to be missing if it takes high or low values. Thought should also be given to the most appropriate way of including the response; for example, with longitudinal data, change in response from one sweep to the next is an alternative option that is worth exploring.

In the absence of any prior knowledge, the recommended strategy is to assume a linear relationship between the probability of missingness and the response or change in response (Mason 2009, Ch. 4). The estimation of the parameters associated with the response (δ) can be difficult, as it is reliant on limited information from assumptions about other parts of the model. These estimation difficulties increase for more complex models of missingness, and motivate the *parameter sensitivity* described in Step 7.

The δ parameters are identified by the parametric assumptions in both the model of interest and the model of response missingness. The missing responses are imputed in a way that is consistent with the distributional assumptions in the model of interest given their covariates, thus δ are identified by the observed data. Daniels and Hogan (2008, Sec. 8.3.2) provide two simple examples which show clearly how this works. Unfortunately the model of interest distribution is unverifiable from the observed data when the response is MNAR. Since different model of interest distributions will lead to different results, the model of interest distribution is a key assumption to explore in the sensitivity analysis.

Step 4. Seek additional data.

Additional data can be incorporated into the various submodels to help with parameter estimation where there is limited information in the study itself. This may come from another study on individuals with similar characteristics to those being modelled or, in the case of longitudinal data, be provided by earlier/later sweeps not under investigation. We discuss an example in Section 5.2.5.

Step 5. Elicit expert knowledge.

Expert knowledge can be elicited (O'Hagan et al. 2006) and incorporated into one or more of the submodels using informative priors. Information relating to the model of response missingness has the potential to make a substantial impact, as there is little information in the model for estimating some of its parameters. However, it is difficult to elicit priors on parameters directly and a better strategy is to elicit information about the probability of response and convert this into informative priors. Elicitation effort should concentrate on the parameters which are not well identified by the data, in particular those associated with the degree of departure from MAR, and the process should allow for correlation between variables. If a comprehensive elicitation is impractical, extracting information about the functional form of important parameters from experts or the literature is worthwhile. For example, an expert may be able to advise whether a linear or piecewise linear relationship is more appropriate. If piecewise linear seems a better option, then including prior information on the position of the change points and signs of the gradients can be beneficial.

At each step, checks of model fit should be carried out to ensure that the models are plausible, and some suggestions are provided in Section 3.3.

3.2. Perform Sensitivity Analysis

When modelling missing data, some form of sensitivity analysis is essential because assumptions must be made that are untestable from the data. There are many possible options, and to some extent the choice will be determined by the problem at hand. We propose that two types of sensitivity analysis should be carried out, an *assumption sensitivity* and a *parameter sensitivity.* This part of the strategy is encapsulated by the following steps.

Step 6. Assumption sensitivity.

For the assumption sensitivity, form a number of alternative models from the base model by changing the assumptions in the different submodels. Key assumptions that should be explored include the model of interest error distribution, the transformation of the model of interest response and the functional form of the model of response missingness. It could also involve varying the explanatory variables. Initially, each of the chosen sensitivity analyses should vary from the base model in a single aspect so their individual effects can be assessed. A second stage of sensitivity analysis could combine several changes which are shown to have a sizeable impact on results.

Step 7. Parameter sensitivity

The parameter sensitivity involves running the base model with the model of response missingness parameters controlling the extent of the departure from MAR fixed to values in a plausible range. Expert knowledge can help with setting up the parameter sensitivity range.

Step 8. Determine robustness of conclusions

The results of both sets of sensitivity analyses should then be examined to establish how much the quantities of interest vary. A range of plots, providing complementary views of the analysis can help (Mason 2009 Ch. 8). If the conclusions are robust, this should be reported. Otherwise a range of diagnostics should be used to determine a region of high plausibility, and the uncertainty in the results recognised. The sensitivity analysis may also suggest that the base model should be reconsidered, or more external information sought from experts or related studies.

3.3. Checking Model Fit

One option for assessing model fit is to use a set of data not used in the model estimation. In surveys, sometimes data is collected from individuals who are originally noncontacts or refusals. These individuals can be treated as missing in the main analysis, and then the predicted values for the covariates and/or response compared with the actual values measured. Using such individuals for comparing model fit is particularly attractive as they are likely to be similar to individuals who have missing data.

Another option is the Deviance Information Criterion (DIC), which is widely used for Bayesian model comparison for complete data. However, with missing data, DIC can be constructed in different ways (Celeux et al. 2006; Daniels and Hogan 2008; Mason et al. 2012), and its use and interpretation are not straightforward. For example, when the response is missing, Mason et al. (2012) propose a strategy for comparing selection models by combining information from two measures taken from different constructions of the DIC. A DIC based on the observed data likelihood (integrated over the missing data) is used to compare joint models with different models of interest but the same model of missingness, and a comparison of models with the same model of interest but different models of missingness is carried out using a partial DIC which treats the missing data as additional parameters in the likelihood.

4. Comparison with Alternative Modelling Strategies

Although there is an extensive literature on missing data, published papers tend to take the form of either a wide-ranging review or focus on a specific type of missing data problem. The review papers generally outline a broad range of different missing data techniques including Bayesian methods, discussing their pros and cons and the circumstances in which each is appropriate. Their aim is to provide the reader with an overview of possible ways of dealing with missing data, rather than a formal strategy.

Much of the more specific missing data literature discusses the implementation of methods that assume MAR. Acknowledging that it is not possible to distinguish between MAR and MNAR from the observed data alone, authors typically argue that the MAR assumption can be made more plausible by collecting and incorporating more explanatory variables and auxiliary information into the analysis. This is good advice, and our strategy also encourages incorporating extra information where possible (Steps 4 and 5), but also allows for the possibility that we still have informative missingness.

Many authors agree that some form of sensitivity analysis is a crucial ingredient of any modelling strategy, and one proposed approach involves considering a number of different statistical models (Molenberghs and Kenward 2007). We have followed this principle by developing a base model and then creating a neighbourhood of alternative models by varying our base model assumptions. An alternative way would be to simply create and compare a number of models underpinned by a range of plausible assumptions, without assigning special "base case" status to one of them. However, this renders comparison more difficult to organise. We prefer to start by building the model thought to be most plausible, and then exploiting the modular setup of Bayesian models by using this as a starting point for alternative models, having thus implicit "directions" for interpreting the sensitivity.

Since we wish to incorporate the assumptions thought most realistic into the base model, we allow informative missingness in the response and envisage that the subsequent sensitivity analysis will include exploring the assumption of MAR missingness. Most proposals for sensitivity analysis take the opposite stance, start with a MAR model and then explore potential deviations towards a MNAR missingness mechanism, e.g., Troxel et al. (2004). This approach has much to recommend it when using non-Bayesian methods that do not allow for MNAR missingness in such a natural way. Our strategy could easily be modified to adopt this line by performing only Steps 1 and 2 to produce the base model, and carrying out Step 3 as a prelude to the sensitivity analysis. The models analysed would be the same, but the labels would change.

Our strategy is based on a selection model factorisation of the joint model, which has the advantage of specifying the distribution that the analyst is usually interested in, the model of interest, directly. An alternative is the pattern mixture factorisation which allows a different model for *y* for each pattern of missingness. Pattern mixture models are favoured by some because the assumptions about the missing data are more explicit (Daniels and Hogan 2008). A hybrid strategy using both selection models and pattern-mixture models is also possible (Molenberghs and Kenward 2007).

Of the commonly used "statistically principled" methods, multiple imputation is most closely aligned to Bayesian methods. It requires specifying an imputation model, similar to the covariate imputation model in Step 2 of our strategy, to generate multiply imputed data sets. Then each of these "completed data sets" is analysed using a model of interest, as in Step 1, and estimates of the statistics of interest combined. In contrast to the Bayesian approach, it is a two-stage process, whereby the imputation of the missing data is carried out as a distinct phase prior to the analysis. This has the advantage of simplifying computation by dividing the problem, but great care is required to avoid a mismatch between the imputation model and the model of interest (often referred to as the problem of "congeniality" (White et al. 2011)). Assuming that a MAR assumption is appropriate, multiple imputation is now readily implementable using a number of mainstream statistical software packages (Horton and Kleinman 2007), and there are a number of excellent papers providing guidance on its use (Kenward and Carpenter 2007; White et al. 2011).

In principle, multiple imputation can be implemented for MNAR mechanisms, but there are few examples of this and mostly these involve adjusting the imputations in some way. Van Buuren et al. (1999) provide an early example, and Carpenter et al. (2007) obtain an overall MNAR estimate by weighting imputations generated under MAR according to the assumed degree of departure from MAR. The idea behind these relatively crude techniques is to test the robustness of results to departures from MAR, to help decide whether implementing a more sophisticated MNAR modelling technique is worthwhile. More usually, strategies involving multiple imputation focus on ways of turning a MNAR problem into a MAR problem through judicious variable selection for the imputation model.

By contrast, we start by explicitly allowing for informative missingness in the response, modelling this in a principled and coherent manner. Then, we exploit the modularity of Bayesian models to perform a thorough sensitivity analysis, which explores the uncertainty surrounding assumptions that cannot be verified on account of the missing data.

5. Application of Modelling Strategy to MCS Income Data

We now provide two examples of how our strategy can be applied in practice, using data from the MCS (University of London, Institute of Education 2009a; University of London, Institute of Education 2009b) which contains missing values for some covariates and for the response, income. Survey methodology literature has shown that income nonresponse is usually nonignorable (Yan et al. 2010). As this is for demonstration purposes, we make simplifying assumptions for the models and omit details of the checks and analysis that should be carried out (suggestions can be found in Mason 2009). All the models are fitted using the WinBUGS software (the code for a base model is provided as an Appendix), run with two chains initialised using diffuse starting values. Convergence is assumed if the Gelman-Rubin convergence statistic (Brooks and Gelman 1998) for individual parameters is less than 1.05 and a visual inspection of the trace plot for each parameter is satisfactory.

5.1. Description of Data

The MCS was set up to provide information about children living and growing up in each of the four countries of the UK, including information about the children's families, and has over 18,000 cohort members born in the UK between specified dates at the start of the Millennium (Plewis 2007a). Data is collected through interviews and self-completion forms undertaken by a main respondent (usually the cohort member's mother) and a partner respondent (usually the father), and four sweeps of this cohort are now available. Nonresponse is discussed by Plewis (2007b), Ketende (2008) and Calderwood et al. (2008).

Using data from Sweeps 1 (2001/2) and 2 (2004/5), we investigate two questions relating to the income from paid work of single mothers. We consider the benefit from having a degree (which we shall refer to as the *Education Question)* and the changes in a mother's rate of pay related to gaining a partner *(Partner Question).* In line with the literature (Blundell et al. 2000; Zhan and Pandey 2004), we expect that higher pay is related to having a degree. The effect of gaining a partner is less obvious, as various aspects of this change in circumstances can be hypothesised to work in opposite directions. It is also possible that the Education Question and the Partner Question are related, but to keep the models relatively simple for illustration purposes, we look at the two questions separately.

To investigate these questions, we model income for the subset of main respondents who are single in Sweep 1, in paid work and not self-employed, using either education level or partnership status, and other known predictors of income. Those who are known to be self-employed or not working in Sweep 2 are also excluded. By definition we are looking at a set of individuals who are the mothers of very young children, so many are working part-time. To simplify the models, hourly net pay, *hpay,* is chosen as our response variable, and the distribution of the observed *hpay* is positively skewed.

Drawing on existing literature, potential covariates are selected with the motivating questions and the structure of the survey in mind. The dataset also includes variables which may help to explain the missingness (Hawkes and Plewis 2008). All these variables are detailed briefly in Table 1. The key covariates of interest are educational level, *edu,* which indicates whether or not an individual has a degree, and partnership status, *sing,* which is always 1 in Sweep 1 from the definition of the dataset, but is used to indicate whether the individual has acquired a partner by Sweep 2.

Ctry and *stratum* are fully observed by survey design. Of the other variables in the dataset, in Sweep 1, 8% of individuals have missing *hpay,* a very small number have missing *edu* or *sc,* and the remaining variables are completely observed. In Sweep 2 missingness is substantially higher, with 32% of individuals having no Sweep 2 data due to wave missingness, and a small amount of item missingness, predominantly for *hpay.*

Table 1. Description of MCS income dataset variables (these relate to the main respondent)

Name	Description	Details
hpay	hourly net pay	continuous – median = £7, range = (£1, £56)
age	age at interview	continuous[a] – median = 26, range = (15,48)
eth	ethnic group	2 levels (1 = White; 2 = Non-White)
reg	region of country	2 levels (1 = London; 2 = other)
edu[b]	educational level	2 levels (1 = no degree; 2 = degree)
sing[c]	single/partner	2 levels (1 = single; 2 = partner)
sc	social class	4 levels[d] (NS-SEC 5 classes with 3 omitted)
ctry	country	1 = England; 2 = Wales; 3 = Scotland; 4 = Northern Ireland
stratum	country by ward type	9 levels[e]

[a] All continuous covariates are centred and standardised; the median and ranges are for Sweep 1 on the original scale.

[b] Based on the level of National Vocational Qualification (NVQ) equivalence of the main respondent's highest academic or vocational educational qualification. We regard individuals with only other or overseas qualifications as missing.

[c] Always 1 for Sweep 1 by dataset definition.

[d] The main respondent's social class, based on the National Statistics Socio-Economic Classification (NS-SEC) grouped into five categories. NS-SEC 3 is small employers and own account workers, and the self-employed are excluded by definition, so *sc* only has four levels: 1 = managerial and professional occupations; 2 = intermediate occupations; 3 = lower supervisory and technical occupations; 4 = semi-routine and routine occupations.

[e] Three strata for England ("advantaged", "disadvantaged" and "ethnic minority"); two strata for Wales, Scotland and Northern Ireland ("advantaged" and "disadvantaged") – see Plewis (2007a) for details.

The analysis of this dataset is restricted to modelling the missingness in Sweep 2 and contains 505 individuals, of which 322 are complete cases.

5.2. MCS Example – Constructing a Base Model

5.2.1. Step 1: Choice of Model of Interest

Based on previous work, we assume that a satisfactory proposed model of interest is available for addressing each question. The models are similar in both cases. Skewness in the response, *hpay,* is dealt with by taking a log transformation, and t_4 errors are used for robustness to outliers. The design of the survey, which is disproportionately stratified (Plewis 2007a), and the correlation between the two data points for each individual are taken into account using stratum-specific intercepts and individual random effects. Hence the model of interest is given by the equations

$$\begin{aligned} y_{it} &\sim t_4(\mu_{it}, \sigma^2) \\ \mu_{it} &= \alpha_i + \gamma_{s(i)} + \sum_{k=1}^{p} \beta_k x_{kit} \end{aligned} \tag{5}$$

for $t = 1,2$ sweeps, $i = 1, \ldots, n$ individuals and $s = 1, \ldots, 9$ strata. The β_ks are not time-dependent, so information from both sweeps contributes to their estimation. However, for the partner question only the values for *sing* in Sweep 2, when some individuals have a partner, inform about the value of β_{sing}. α_i are the individual random effects, such that $\alpha_i \sim N(0, \varsigma^2)$ and its standard deviation, ς, has a vague prior defined as a half normal distribution restricted to positive values. Vague priors are also specified for the other unknown parameters of the model of interest: the stratum-specific intercepts, $\gamma_{s(i)}$ and β_k parameters are assigned $N(0,10{,}000^2)$ priors and the inverse of the sampling (Level 1) variance, $1/\sigma^2$, is given a *Gamma*(0.001,0.001) prior.

For both questions, *age* (main respondent's age) and *reg* (London/other) are included in the set of time-dependent *x* covariates. For the education question *edu* (no degree/degree) is also included, whereas for the partner question *sing* (single/partner) is added. The parameter estimates for these initial models of interest, MoI, based on complete cases only are shown in Table 2 (the other models in this table will be discussed later). They suggest that higher levels of hourly pay are associated with increasing age and having a degree, and lower levels of hourly pay are associated with living outside London and gaining a partner between sweeps.

5.2.2. Step 2: Choice of Covariate Imputation Model

To include the incomplete cases, the missing Sweep 2 values for *age, reg* and either *edu* or *sing* need to be imputed. For simplicity, we do not use a statistical model for *reg* and *age,* but prior to the analysis set their missing values using simple rules. Missing *reg* are set to their Sweep 1 values, and missing *age* are set to the individual's Sweep 1 age plus the mean observed difference in ages between Sweeps 1 and 2. Imputing *edu* or *sing* is sufficient for demonstrating our strategy, although the covariate imputation model could be expanded to include the imputation of *age* and *reg* (ways of doing this are discussed in Section 3, Step 2).

By contrast, for imputing *edu* or *sing,* the variables which particularly interest us, we define a Bernoulli model. For the partner question, we define a covariate imputation model for the individuals who have Sweep 2 missing *sing* values by

$$\begin{aligned} &sing_{i2} \sim Bernoulli(q_i) \\ &q_i = \phi_{s(i)} + (\phi_{age} \times age_{i2}) + (\phi_{reg} \times reg_{i2}) \end{aligned} \tag{6}$$

where $\phi_{s(i)}$ are stratum-specific intercepts. Vague N(0, 10,000^2) priors are assigned to the ϕ_k parameters. For the education question, an individual with a degree in Sweep 1 must also have a degree in Sweep 2, so there is no need to include them in the imputation model. We only impute Sweep 2 *edu* for individuals who have missing *edu* values in Sweep 2 and who do not have a degree in Sweep 1. A simpler covariate imputation model without explanatory variables is used, as too few individuals gain a degree between sweeps to estimate all the parameters in a more complex model. In this joint model consisting of the proposed model of interest and covariate imputation model (MoI.CIM), not only are the covariates assumed to be MAR, but the response is also assumed to be MAR.

From Table 2, there are small changes in some of the model of interest parameters from fitting MoI.CIM compared to the complete case analysis (MoI). Among individuals without a degree in Sweep 1 and observed educational level in Sweep 2, 5 individuals (1.9%) gained a degree by Sweep 2. Based on the posterior means, this is similar to the 2.3% (95% interval from 0% to 5.7%) imputed to gain a degree between sweeps. For *sing,*

Table 2. Comparison of selected parameter estimates from different models (with BASE and non-negligible differences[a] from BASE highlighted in bold)

	Complete cases		MAR response		MNAR response					
	MoI		MoI.CIM		BASE		AS1		AS2	
Education question										
MoI: β_{age}	0.12	(0.08,0.16)	0.11	(0.08,0.14)	**0.10**	**(0.07,0.14)**	0.09	(0.06,0.13)	**0.35**	(0.09,0.61)
MoI: β_{age^2}									− 0.27	(− 0.54, − 0.01)
MoI: β_{edu}	**0.23**	(0.15,0.31)	0.18	(0.10,0.26)	**0.19**	**(0.11,0.27)**	0.21	(0.12,0.29)	**0.14**	(0.06,0.22)
MoI: $\beta_{age \times edu}$									0.11	(0.03,0.19)
MoI: β_{reg}	**− 0.16**	(− 0.28, − 0.04)	− 0.11	(− 0.22,0.00)	**− 0.11**	**(− 0.22,0.00)**	**− 0.13**	(− 0.24, − 0.01)	− 0.12	(− 0.23, − 0.01)
RMoM: δ[b]					**− 0.17**	**(− 0.41,0.05)**	**− 0.23**	(− 0.59,0.05)	**− 0.13**	(− 0.37,0.06)
Partner question										
MoI: β_{age}	0.15	(0.11,0.18)	0.13	(0.10,0.17)	**0.13**	**(0.09,0.16)**	**0.11**	(0.08,0.15)	**0.35**	(0.09,0.60)
MoI: β_{age^2}									− 0.22	(− 0.48,0.03)
MoI: β_{sing}	**− 0.08**	(− 0.15, − 0.01)	**− 0.07**	(− 0.14,0.00)	**− 0.17**	**(− 0.28, − 0.07)**	**− 0.26**	(− 0.37, − 0.15)	− 0.18	(− 0.28, − 0.08)
MoI: β_{reg}	**− 0.18**	(− 0.31, − 0.05)	− 0.13	(− 0.24, − 0.01)	**− 0.12**	**(− 0.23, − 0.01)**	− 0.12	(− 0.24, − 0.01)	− 0.12	(− 0.24, − 0.01)
RMoM: δ[b]					**− 0.43**	**(− 0.76, − 0.13)**	**− 0.81**	(− 1.24, − 0.44)	− 0.42	(− 0.76, − 0.12)

Table shows the posterior mean, with the 95% interval in brackets.

[a] Absolute difference $>$ 0.02 and percentage difference $>$ 10%.

[b] δ is the key parameter in the model controlling the departure from MAR, and multiplies change in pay.

35.5% (95% interval from 26.5% to 44.6%) of those with missing *sing* at Sweep 2 were imputed to gain a partner, compared to 33.6% of those with observed Sweep 2 *sing*.

5.2.3. Step 3: Choice of Model of Response Missingness

The base model is completed by adding a model of response missingness as specified by Equation 4, where m_i is a binary missing value indicator for $hpay_{i2}$, set to 0 when hourly pay in Sweep 2 for individual i is observed and 1 otherwise. The predictors of missing income, $\boldsymbol{w}$, are *sc* (social class), *eth* (ethnic group) and *ctry* (country), and their inclusion is based on work on item missingness by Hawkes and Plewis (2008). The missingness is also allowed to depend on the level of pay at Sweep 1 and the change in pay between sweeps, so

$$logit(p_i) = \theta_0 + \sum_{k=1}^{r} \theta_k w_{ki} + \kappa \times hpay_{i1} + \delta \times (hpay_{i2} - hpay_{i1}) \qquad (7)$$

For simplicity, linear relationships are assumed, and this submodel uses an untransformed version of *hpay*. The priors for the θ, κ, and δ parameters are specified as $\theta_0 \sim Logistic(0,1)$ and θ_k, κ, $\delta \sim \mathrm{N}(0,10{,}000^2)$. It is the inclusion of the term $\delta \times (hpay_{i2} - hpay_{i1})$ that allows the response missingness to be MNAR. If $\delta = 0$, then we have MAR missingness.

5.2.4. Conclusions from Base Model

Selected parameter estimates for the base model for each question, BASE, are shown in Table 2 and can be compared with the models in the left-hand side of this table. As regards our substantive questions, compared to the complete case analysis (MoI), the evidence that having a degree is associated with higher pay is similar and the evidence that gaining a partner is associated with lower pay has strengthened. The covariate imputations are similar to MoI.CIM for the education question (Section 5.2.2), but a greater proportion of individuals are imputed to gain a partner, 43.9% (32.5%,56.6%), once we allow for nonignorable missing responses. For the partner question, individuals whose pay decreases substantially between sweeps are more likely to be missing, but this effect is not so strong for the education question.

5.2.5. Steps 4 and 5: Incorporating Additional Data and Expert Knowledge

The strategy also allows for including additional data (Step 4) and an elicitation to provide expert priors (Step 5). For example, one possibility for incorporating additional data into the covariate imputation model would be to use information on educational qualifications and partner status from women who had recently had children, taken from Sweeps 5 and 6 of the 1970 British Cohort Study (BCS70) (University of London, Institute of Education 2007a; University of London, Institute of Education 2007b). Data from these sweeps would be appropriate as they were carried out at similar times to the MCS Sweeps 1 and 2, when the cohort members were aged 30 and 34. The difference is that the BCS70 data would be on the cohort members themselves rather than their mothers. The BCS70 and MCS data would then be modelled by simultaneously fitting two sets of equations with common parameters, one for each data source, allowing these parameters to be estimated with greater accuracy.

For this application, level and change in income are the key variables in the model of response missingness, and survey methodology experts could be consulted to elicit prior beliefs about how these variables are likely to influence the probability of nonresponse. In particular, the assumption of linear relationships should be reviewed as individuals may be less inclined to disclose their income if it is either low or high, or has changed substantially in either direction. See Mason (2009, Ch. 7) for examples of incorporating data from another study and eliciting expert knowledge.

5.3. Step 6: MCS Example – Assumption Sensitivity

To demonstrate the assumption sensitivity, we fit two sensitivity analyses (AS1 and AS2) to investigate different model of interest assumptions. For AS1 a normal error distribution is used instead of a t_4, and for AS2 additional covariate terms age^2 (for both questions) and $age \times edu$ (for the education question only) are added. The parameter estimates are given for both models in Table 2, with nonnegligible differences from BASE highlighted in bold, where a nonnegligible difference is defined as a percentage difference greater than 10% and an absolute difference greater than 0.02.

5.4. Step 7: MCS Example – Parameter Sensitivity

The value of δ in the model of response missingness controls the degree of departure from MAR missingness. This parameter is difficult to estimate for a model with vague priors. Verbeke et al. (2001) envisage a sensitivity analysis in which the changes in the parameters or functions of interest are studied for different values of δ. In the same spirit, we also carry out a sensitivity analysis in which a series of models is run with this parameter fixed to different values. We refer to this group of models as PS (Parameter Sensitivity), and it contains nine variants in which δ is set to the values $\{-1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1\}$, where δ corresponds to the log odds ratio of a missing response per £1 increase in hourly pay. Values outside this range are intuitively implausible as the probability of response is then assumed to change from 1 to 0 very abruptly. The $\delta = 0$ variant is equivalent to assuming the response is MAR. In contrast to BASE which estimates δ, the PS models fix δ using point priors. An alternative would be to use strongly informative priors.

The estimated proportional increase in hourly pay associated with having a degree, $\exp(\beta_{edu})$ varies from 1.17 (1.08,1.27) when $\delta = 1$ to 1.24 (1.15,1.34) when $\delta = -1$ (based on the posterior means, with 95% credible intervals shown in brackets), indicating that the effect of gaining a degree on income is fairly robust to different settings for δ. By contrast, the range for $\exp(\beta_{sing})$ is 0.76 (0.69,0.82) to 1.36 (1.22,1.51), so the effect of gaining a partner between sweeps on hourly pay is very sensitive to the value of δ. If all the PS variants are plausible, then even the direction of this effect is uncertain.

Given the sensitivity of the results to the assumptions, the plausibility of the range of δ values needs to be considered carefully, as discussed in Step 8. We looked at the fit of a small sample of individuals, who responded after they were reissued by the fieldwork agency and were not included in our analysis, and various measures of DIC. Taken together, this analysis suggests that positive values of δ are unlikely (Mason 2009, Ch. 8).

5.5. Step 8: Robustness of the Conclusions on Substantive Questions

From BASE, gaining a degree would make a difference of £2.08 (£1.18,£3.04) an hour for an individual earning £10 an hour. However, from the sensitivity analysis, this difference could plausibly vary between £1.48 (£0.57,£2.46) (AS2) and £2.44 (£1.54,£3.41) (PS with $\delta = -1$). There is greater uncertainty surrounding the impact of gaining a partner, with a plausible range on the reduction in pay from £0.67 (£1.32, − £0.04)(PS with $\delta = 0$) to £2.46 (£3.06, £l.82)(PS with $\delta = -1$) an hour for an individual earning £10 an hour. BASE suggests a £1.59 (£2.41,£0.72) reduction. Some models run as part of the parameter sensitivity analysis suggest that change in partnership status is associated with an increase in pay, but these models do not fall within the region of high plausibility.

5.6. Comparison with Other Methods

Finally, we compare the results from our modelling strategy with those from other missing data methods, using the partner question and focussing on the change in pay for an individual earning £10 an hour (Table 3). So that the alternative methods can be implemented using readily available packages written for the R software (R Development Core Team 2011), the model of interest used in all the models discussed in this section is fitted with normal rather than t_4 errors.

For Step 1, a complete case analysis can be carried out by fitting the multilevel model of interest using the **lme** function from the **nlme** library (Pinheiro et al. 2011). As vague priors were used in the Bayesian model, the results using maximum likelihood are virtually identical. Last Observation Carried Forward (LOCF) is a single imputation method, which makes the strong and usually unjustifiable assumption that the missing values for an individual are the same as their last seen value. For the example, the missing Sweep 2 values of *hpay* and *sing* are set to their Sweep 1 values, which has the predictable effect of reducing the impact of gaining a partner and underestimating the uncertainty of the estimate, as evidenced by the narrower interval.

Multiple imputation provides a comparison with a "statistically principled" method and can be implemented using functions from the **mice** library (van Buuren and Groothuis-Oudshoorn 2011). So, at Step 2, the missing values for both *hpay* and *sing* can be imputed under an assumption of MAR using the chained equations approach (White et al. 2011) with ten imputations (typical number). Missing values of *sing* are imputed using a logistic regression with all the regressors from the Bayesian covariate imputation model and *hpay*. We add *hpay* because multiple imputation is a two-stage approach, so, unlike in the Bayesian joint model, there is no feedback from the model of interest, and consequently the response must be included directly in the imputation model. The missing values of *hpay* are imputed using a two-level linear model (*2l.norm* method in the *mice* function), to correctly reflect the complexity in the model of interest (Carpenter and Goldstein 2004). Again, the inclusion of an imputation model for *hpay* is necessitated by the separation of the imputation and analysis, whereas in a Bayesian joint model the model of interest is used. This multiple imputation is similar in terms of assumptions to the Bayesian joint model developed at Step 2 (model of interest plus covariate imputation model), but is not mathematically equivalent. It can be viewed as an approximation to the Bayesian joint model, and the multiple imputation gives a slightly higher point estimate and wider 95% interval (see Table 3).

Table 3. Decrease in hourly pay associated with gaining a partner for an individual earning £10 an hour

Bayesian modelling[a]			Alternative methods[b]		
Complete cases (Step 1: MoI)	£0.72	(£1.39,£0.00)	Complete cases	£0.72	(£1.40, − £0.01)
			LOCF	£0.40	(£1.00, − £0.23)
MAR (Step 2: MoI.CMoM)	£0.56	(£l.28, − £0.22)	MAR (mice)	£0.69	(£1.54, − £0.24)
MNAR (Step 3: AS1)	£2.28	(£3.07,£1.37)			
MNAR PS (Step 7: $\delta = -0.5$)	£1.82	(£2.47,£1.11)	MNAR (mice −50%)	£1.80	(£2.78,£0.68)

95% credible or confidence interval shown in brackets.

[a] Models fitted using WinBUGS software and R software for pre and post-processing.

[b] Models fitting using R software.

Although multiple imputation is not restricted to MAR, most implementations do not readily extend to MNAR other than in an ad hoc way. It is not currently possible to fit an equivalent model to the Bayesian joint model produced at Step 3 of our strategy using MICE or similar packages. However, following the suggestions of van Buuren and Groothuis-Oudshoorn (2011), we carry out a sensitivity analysis under MNAR by post-processing the imputations. Sensitivity to the imputed values of income being too low can be tested by multiplying the original imputed values by, say, 1.5 to inflate them by 50%. Similarly a factor of, say, 0.5 can be used to assess the effects of the imputed values being too high. The results for decreasing the *hpay* imputations by 50% on the untransformed scale in this way are shown in Table 3, but we do not look at increasing given our previous conclusion about the implausibility of positive δ values (Section 5.4). This multiple imputation sensitivity analysis is not directly comparable with the parameter sensitivity analysis undertaken at Step 7 of our strategy, but decreasing the imputed *hpay* values has a similar effect to a negative δ which associates missingness with a pay decrease. This multiple imputation adjustment approach and the proposed Bayesian methods both show sensitivity to the assumptions about the missingness process.

6. Extensions and Adaptations of Modelling Strategy

Our proposed strategy assumes that the covariates are MAR, but in principle Step 2 can be elaborated to allow MNAR covariates. This raises a number of questions, for example should separate missingness indicators for the covariates and the response be used or an overall missingness indicator for attrition? If separate indicators are used, a new submodel linked to the existing covariate imputation model is required. To implement this, a different indicator for each covariate pattern of missingness is needed. Alternatively, if an overall missingness indicator for attrition is used, then a method for dealing with any item missingness that occurs in the response or covariates is also required. Although in theory a model allowing MNAR covariates could be designed, it may currently be computationally prohibitive in WinBUGS. Conversely, if there is reason to suspect that the responses are not generated by an informative missingness process, then the strategy can be simplified by omitting Step 3 and restricting the sensitivity analysis to varying the assumptions.

In Section 2 we discussed the different types of nonresponse that can occur, but in the application we modelled the missing data as a homogeneous process. However, the nonresponse in Sweep 2 can result from the failure to trace families who have moved, failure to contact families at a known address and refusal of individuals to continue to cooperate. As these three types of nonresponse have different correlates (Plewis 2007b; Plewis et al. 2008), there is considerable scope for expanding the sensitivity analysis to respecify the model of response missingness to specifically allow for these differences. This could be implemented by modelling a missingness indicator with separate categories for each type of nonresponse using multinomial regression in place of the logistic regression model (Equation 4). In general, the model of response missingness can be extended in a similar way, using multiple missingness indicators, to allow different predictors to be used for item missingness, wave missingness and attrition as appropriate.

There are situations where it may be necessary to adapt the strategy that we recommend. Bayesian models have the advantage of being fully coherent, but pragmatically, with large datasets or large numbers of covariates with missingness they may be computationally challenging to fit. In these circumstances, some sort of hybrid approach is required, whereby some of the covariates are imputed (preferably multiple times) prior to fitting a Bayesian model. As this would be a two-stage process, the usual issues surrounding multiple imputation regarding compatibility must be considered (Rubin 1996; Carpenter and Goldstein 2004). It may be acceptable to impute the missing values of some covariates using simplistic assumptions (as for *age* and *region* in the example). If not all the covariates are correlated, another option is to split the covariate imputation model into several smaller submodels. Although our strategy has been implemented using Bayesian models, there is no reason why the framework could not be adopted for other inference paradigms.

7. Concluding Remarks

Compared to performing a complete case analysis or using some other ad hoc method for dealing with missing data, the implementation of the strategy set out here, which enables a "principled" missing data analysis, is time-consuming in terms of the extra work in designing and implementing a base model and number of sensitivity analyses. However, the time taken to implement this more complex analysis is still likely to be a small fraction of the overall time spent collecting, preparing and analysing the data. In return, realistic assumptions about the missingness mechanism can be thoroughly explored and the uncertainty resulting from the missing data properly reflected in the discussion of results.

The individual elements of the strategy are not novel, but to our knowledge they have not previously been presented as part of a general strategy for analysts to follow. We hope that having an adaptable iterative framework to follow will encourage the analyst to think carefully about the reasons for the missingness in their data, to incorporate realistic assumptions into their models and to acknowledge the uncertainty that missing data adds in the presentation of their results.

Appendix

```
# WinBUGS code for running the base model for the Partner Question
# Model of Interest: hpay logged, covariates {age, sing, reg}, individual random effects,
# stratum-specific intercepts and t4 errors
# Covariate Model of Missingness: imputes missing sing in Sweep 2 using covariates
# {age, reg} and stratum-specific intercepts
# Response Model of Missingness: logit(p) regressed on {change, level, ctry, eth, sc}
  model

  {
    for (i in 1:N) { # N individuals
      for (t in 1:2) {#2 sweeps
        # Model of Interest
        hpay[i,t] ~ dt(mu[i,t],tau,4)
        mu[i,t] < -beta0[i] + beta0.stratum[stratum[i]] + beta.age*age[i,t]
          + beta.sing*sing[i,t] + beta.reg*reg[i,t]
        e.hpay[i,t] < -exp(hpay[i,t]) # unlog hpay for response missingness model
      resid[i,t] < -(hpay[i,t]-mu[i,t])/sigma # calculate residuals
    }
    beta0[i] ~ dnorm(0,beta0.tau) # individual random effects
    # Missingness model for the Response – sweep 2 only
    payid[i] ~ dbern(p[i])
    logit(p[i]) < -theta0 + theta.eth*eth[i]+ theta.ctry[ctry[i]]+ theta.sc[sc[i]]
      + theta.level*level[i] + delta.change*change[i]
    linkp[i] < -theta0 + theta.eth*eth[i]+ theta.ctry[ctry[i]]+ theta.sc[sc[i]]
      + theta.level*level[i] + delta.change*change[i]
    level[i] < -e.hpay[i,1]-mean(e.hpay[,1])
    change[i] < -e.hpay[i,2]-e.hpay[i,1]
  }
  # sing imputation model – sweep 2 only
  for (i in 1:N){
    sing[i,2] ~ dbern(q[i])
    logit(q[i]) < -phi.stratum[stratum[i]] + phi.age*age[i,2] + phi.reg*reg[i,2]
  }
  # Priors for model of interest
  beta0.sigma ~ dnorm(0,0.00000001)l(0,)
  beta0.tau < -1/(beta0.sigma*beta0.sigma)
  for (st in 1:9) { beta0.stratum[st] ~ dnorm(0,0.00000001)}
  # 9 stratum-specifc intercepts
  beta.age ~ dnorm(0,0.00000001)
  beta.sing ~ dnorm(0,0.00000001)
  beta.reg ~ dnorm(0,0.00000001)
  tau ~ dgamma(0.001,0.001)
  sigma < -sqrt(2 / tau) # t errors on 4 degrees of freedom
  # Priors for sing imputation model
  for (st in 1:9) { phi.stratum[st] ~ dnorm(0,0.00000001)}
  # 9 stratum-specifc intercepts
  phi.age ~ dnorm(0,0.00000001)
  phi.reg ~ dnorm(0,0.00000001)
  # Priors for missingness model
  theta0 ~ dlogis(0,l)
  theta.eth ~ dnorm(0,0.00000001)
  theta.sc[1] < −0 # alias first level of sc (social class) beta
  for (s in 2:4) {theta.sc[s] ~ dnorm(0,0.00000001)} # 4 levels of social class
  theta.ctry[1] < −0 # alias first level of ctry (country) beta
  for (y in 2:4) {theta.ctry[y] ~ dnorm(0,0.00000001)} # 4 countries
  theta.level ~ dnorm(0,0.00000001)
  delta.change ~ dnorm(0,0.00000001)
  # Odds ratios
  sing.or < -exp(beta.sing)
}
```

8. References

Albert, J.H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. Journal of the American Statistical Association, 88, 669–679.

Blundell, R., Dearden, L., Goodman, A., and Reed, H. (2000). The Returns to Higher Education in Britain: Evidence from a British Cohort. The Economic Journal, 110, F82–F99.

Brooks, S. and Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. Journal of Computational and Graphical Statistics, 7, 434–455.

Calderwood, L., Ketende, S., and MacDonald, J. (2008). Patterns of Longitudinal Participation in the Millennium Cohort Study. Prepared for the Panel Surveys Workshop in Essex July. Available at www.iser.essex.ac.uk.

Carpenter, J.R. and Goldstein, H. (2004). Multiple Imputation in MLwiN. Multilevel Modelling Newsletter, 16.

Carpenter, J.R., Kenward, M.G., and White, I.R. (2007). Sensitivity Analysis after Multiple Imputation under Missing at Random: A Weighting Approach. Statistical Methods in Medical Research, 16, 259–275.

Celeux, G., Forbes, F., Robert, C.P., and Titterington, D.M. (2006). Deviance Information Criteria for Missing Data Models. Bayesian Analysis, 1, 651–674.

Chib, S. and Greenberg, E. (1998). Analysis of Multivariate Probit Models. Biometrika, 85, 347–361.

Daniels, M.J. and Hogan, J.W. (2008). Missing Data In Longitudinal Studies; Strategies for Bayesian Modeling and Sensitivity Analysis. London: Chapman & Hall.

Dunson, D.B. (2000). Bayesian Latent Variable Models for Clustered Mixed Outcomes. Journal of the Royal Statistical Society, Series B, Statistical Methodology, 62, 355–366.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004). Bayesian Data Analysis, (2nd ed.). London: Chapman & Hall.

Goldstein, H., Carpenter, J., Kenward, M.G., and Levin, K.A. (2009). Multilevel Models with Multivariate Mixed Response Types. Statistical Modelling, 9, 173–197.

Hawkes, D. and Plewis, I. (2006). Modelling Non-Response in the National Child Development Study. Journal of the Royal Statistical Society, Series A, Statistics in Society, 169, 479–491.

Hawkes, D. and Plewis, I. (2008). Missing Income Data in the Millenium Cohort Study: Evidence from the First Two Sweeps. CLS Cohort Studies, Working Paper 2008/10, Institute of Education, University of London.

Horton, N.J. and Kleinman, K.P. (2007). Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. The American Statistician, 61, 79–90.

Ibrahim, J.G., Chen, M.-H., Lipsitz, S.R., and Herring, A.H. (2005). Missing-Data Methods for Generalized Linear Models: A Comparative Review. Journal of the American Statistical Association, 100, 332–346.

Kenward, M.G. and Carpenter, J. (2007). Multiple Imputation: Current Perspectives. Statistical Methods in Medical Research, 16, 199–218.

Ketende, S. (2008). Millennium Cohort Study: Technical Report on Response. Technical Report, (2nd ed.). Institute of Education, University of London.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data, (2nd ed.). Hoboken, NJ: John Wiley and Sons.

Mason, A., Richardson, S., and Best, N. (2012). Two-Pronged Strategy for Using DIC to Compare Selection Models with Non-Ignorable Missing Responses. Bayesian Analysis, 7, 109–146.

Mason, A.J. (2009). Bayesian Methods for Modelling Non-Random Missing Data Mechanisms in Longitudinal Studies. PhD thesis, Imperial College London. Available at www.bias-project.org.uk.

Molenberghs, G. and Kenward, M.G. (2007). Missing Data in Clinical Studies, (1st ed.). Hoboken, NJ: John Wiley and Sons.

Molitor, N.-T., Best, N., Jackson, C., and Richardson, S. (2009). Using Bayesian Graphical Models to Model Biases in Observational Studies and to Combine Multiple Data Sources: Application to Low Birth-Weight and Water Disinfection By-Products. Journal of the Royal Statistical Society, Series A, 172, 615–637.

O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., and Rakow, T. (2006). Uncertain Judgements: Eliciting Experts' Probabilities, (1st ed.). Hoboken, NJ: John Wiley and Sons.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Development Core Team, (2011). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-98.

Plewis, I. (2007a). The Millennium Cohort Study: Technical Report on Sampling. Technical Report, (4th ed.). Institute of Education, University of London.

Plewis, I. (2007b). Non-Response in a Birth Cohort Study: The Case of the Millennium Cohort Study. International Journal of Social Research Methodology, 10, 325–334.

Plewis, I., Ketende, S.C., Joshi, H., and Hughes, G. (2008). The Contribution of Residential Mobility to Sample Loss in a Birth Cohort Study: Evidence from the First Two Waves of the UK Millennium Cohort Study. Journal of Official Statistics, 24, 1–22.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, Available at http://www.R-project.org.

Rubin, D.B. (1976). Inference and Missing Data. Biometrika, 63, 581–592.

Rubin, D.B. (1996). Multiple Imputation After 18 + Years. Journal of the American Statistical Society, 91, 473–489.

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data, (1st ed.). London: Chapman & Hall.

Schafer, J.L. and Graham, J.W. (2002). Missing Data: Our View of the State of the Art. Psychological Methods, 7, 147–177.

Spiegelhalter, D.J., Thomas, A., Best, N.G., and Lunn, D. (2003). WinBUGS Version 1.4 User Manual. MRC Biostatistics Unit, Cambridge, Available at www.mrc-bsu.cam.ac.uk/bugs.

Troxel, A.B., Ma, G., and Heitjan, D.F. (2004). An Index of Local Sensitivity to Nonignorability. Statistica Sinica, 14, 1221–1237.

University of London, Institute of Education (2007a). Centre for Longitudinal Studies, 1970 British Cohort Study: Thirty-Four-Year Follow-up, 2004–2005 [computer file]. Colchester, Essex: UK Data Archive [distributor], March. SN: 5585.

University of London, Institute of Education (2007b). Centre for Longitudinal Studies, 1970 British Cohort Study: Twenty-Nine-Year Follow-up, 1999–2000 [computer file]. Joint Centre for Longitudinal Research, [original data producer(s)]. Colchester, Essex: UK Data Archive [distributor], January. SN: 5558.

University of London, Institute of Education (2009a). Centre for Longitudinal Studies, Millennium Cohort Study: First Survey, 2001–2003 [computer file] (8th ed.) Colchester, Essex: UK Data Archive [distributor], March. SN: 4683.

University of London, Institute of Education (2009b). Centre for Longitudinal Studies, Millennium Cohort Study: Second Survey, 2003–2005 [computer file]. (5th ed.) Colchester, Essex: UK Data Archive [distributor], March. SN: 5350.

van Buuren, S., Boshuizen, H.C., and Knook, D.L. (1999). Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis. Statistics in Medicine, 18, 681–694.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). MICE: Multiple Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1–67.

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E., and Kenward, M.G. (2001). Sensitivity Analysis for Nonrandom Dropout: A Local Influence Approach. Biometrics, 57, 7–14.

White, I.R., Royston, P., and Wood, A.M. (2011). Multiple Imputation Using Chained Equations: Issues and Guidance for Practice. Statistics in Medicine, 30, 377–399.

Yan, T., Curtin, R., and Jans, M. (2010). Trends in Income Nonresponse Over Two Decades. Journal of Official Statistics, 26, 145–164.

Zhan, M. and Pandey, S. (2004). Postsecondary Education and Economic Well-Being of Single Mothers and Single Fathers. Journal of Marriage and Family, 66, 661–673.

Received May 2010
Revised December 2011