Strategy for modelling non-random missing data mechanisms in observational studies using Bayesian methods

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

[1]Department of Epidemiology and Biostatistics, Imperial College London, UK

[2]Social Statistics, University of Manchester, UK

Abstract

Observational studies are notoriously full of non-responses 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 non-random.

The first part of this strategy entails constructing a 'base model' by selecting a model of interest, then adding a sub-model to impute the missing covariates followed by a sub-model 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 Millennium Cohort Study.

Key words: longitudinal analysis; cross sectional analysis; sensitivity analysis; Millennium Cohort Study; income; non-response; attrition

Acknowledgements: 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.

1 Introduction

Social science data typically suffer from non-response 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 com-

plicated by missing covariates and missing responses. The appropriateness of a particular analytic approach is dependent on the mechanism that leads 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 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 paper, 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 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 the income gains from higher education and 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. In Section 4 we apply this strategy to an extended example, discuss possible modifications and the circumstances where these would be necessary in Section 5 and conclude in Section 6.

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, non-response can take three forms: unit non-response (sampled individuals are absent from the outset of the study), wave non-response (where an individual does not respond in a particular wave but re-enters 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 non-response 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 non-response) or on a complete group of questions (domain non-response).

Bayesian full probability modelling provides a flexible method of incorporating different assumptions about the missing data mechanism and accommodating different patterns of missing data. It entails building a joint model consisting of a model of interest and one or more models of missingness, 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.

Suppose the data for our research consists 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, $f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma)$,

is

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

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

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

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

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

suppressing the dependence on the covariates, and assuming that $(\boldsymbol{m}|\boldsymbol{y}, \boldsymbol{\theta})$ is conditionally independent of $(\boldsymbol{\beta}, \sigma)$, and $(\boldsymbol{y}|\boldsymbol{\beta}, \sigma)$ is conditionally independent of $\boldsymbol{\theta}$, 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* for Bayesian inference about $(\boldsymbol{\beta}, \sigma)$ if the missing data are MAR ($f(\boldsymbol{m}|\boldsymbol{y}_{obs}, \boldsymbol{y}_{mis}, \boldsymbol{\theta}) = f(\boldsymbol{m}|\boldsymbol{y}_{obs}, \boldsymbol{\theta})$) and the parameters of the data model, $(\boldsymbol{\beta}, \sigma)$, and the missingness mechanism, $\boldsymbol{\theta}$, are distinct and the priors for $(\boldsymbol{\beta}, \sigma)$ and $\boldsymbol{\theta}$ are independent (Little and Rubin, 2002).

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

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). If they are assumed to be generated by a nonignorable missing data mechanism, then an appropriate missingness indicator will also need to be modelled.

The data for our application contains missing values for some covariates and for the response, income. Survey methodology literature has shown that income non-response is usually non-ignorable

(Yan *et al.*, 2010). Plewis *et al.* (2008) investigate predictors of sample loss in the MCS with particular reference to residential mobility, distinguishing between different types of non-response, and conclude that assuming MAR is perhaps too strong. More specifically, the patterns and correlates of missing income data in the MCS are analysed by Hawkes and Plewis (2008). Before proceeding to the application, we introduce our proposed modelling strategy for missing data.

3 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).

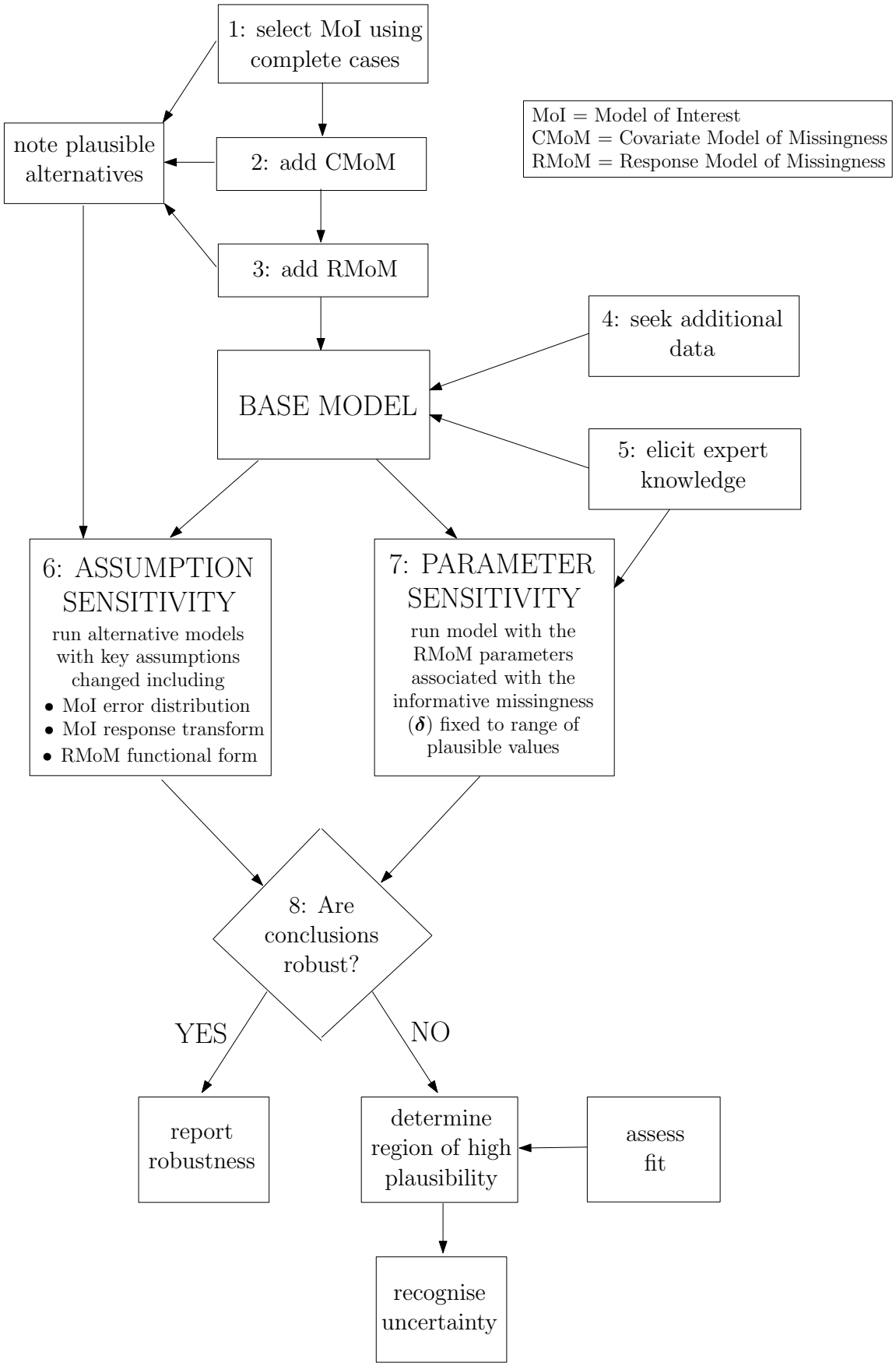
This strategy can be thought of as consisting of two parts: 1) constructing a base model and 2) assessing conclusions from this base model against a selection of well chosen sensitivity analyses. Each of these is now discussed, drawing attention to the key decisions based on our experience. Our proposed strategy allows informative missingness in the response, but assumes that the covariates are MAR. We defer discussion of adaptations, extensions and limitations until Section 5.

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 model of missingness followed by a response model of missingness. For each sub-model, 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.

1 **Select an initial model of interest (MoI) based on complete cases.** The process of building a base model starts with the formation of an initial model of interest using only complete cases and previous knowledge. This includes choosing a transform for the response, model structure and a set of explanatory variables. The most critical assumption is the error distribu-

Figure 1: Strategy for Bayesian modelling of missing data



The numbers relate to the steps described in Section 3.

tion of the model of interest, whose mis-specification can adversely affect the performance of a selection model (Mason *et al.*, 2010).

2 **Add a covariate model of missingness (CMoM).** 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 model of missingness to produce realistic imputations of any missing covariates simultaneously with the analysis of the model of interest. If we have a single covariate, x, there are two obvious ways of building this sub-model: i) specify a distribution, e.g. if x is a continuous covariate, then could 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. This sub-model 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 as necessary. A latent variable approach can be used for binary or categorical variables, as discussed in Section 4.2.2. The reasonableness of this model can be checked by comparing the pattern of the imputed values with the observed values.

3 **Add a response model of missingness (RMoM).** Next, add a response model of missingness to allow informative missingness in the response. 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 *et al.*, 2010). The estimation of the parameters associated with the response 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 with vague priors, and motivate the *parameter sensitivity* described in step 7.

4 **Seek additional data.** Additional data can be incorporated into the various sub-models to help with parameter estimation where there is limited information. 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.

5 **Elicit expert knowledge.** Expert knowledge can also be incorporated into one or more of the sub-models using informative priors. Information relating to the response model of missingness has the potential to make the biggest impact, in particular regarding its functional form.

At each step, checks of model fit should be carried out to ensure that the models are plausible.

3.2 Perform sensitivity analysis

When modelling missing data, some form of sensitivity analysis is essential because we are forced to make assumptions which 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 our strategy is encapsulated by the following steps.

6 **Assumption sensitivity.** For the assumption sensitivity, form a number of alternative models from the base model by changing key assumptions. These should include, but not be limited to, changes in the model of interest error distribution, the transformation of the model of interest response and the functional form of the response model of missingness.

7 **Parameter sensitivity.** The parameter sensitivity involves running the base model with the response model of 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.

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 is recommended (examples are provided in Section 4.5). If the conclusions are robust, this should be reported. Otherwise a range of diagnostics, for example the fit of a validation sample and Bayesian measures of fit (as discussed in Section 4.4.3) 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.

4 Application of strategy to MCS income data

We now provide two examples of the application of this strategy, fitting all the described models using the WinBUGS software (the code for a base model is provided as an Appendix). All the models are run with two chains initialised using diffuse starting values, and converged. 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 plots is satisfactory.

4.1 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, 2007*a*). 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. Non-response is discussed by Plewis (2007*b*), Ketende (2008) and Calderwood *et al.* (2008).

Analysis of data from the MCS needs to take account of its design, as it is clustered geographically, and disproportionately stratified (Plewis, 2007*a*). The population is stratified by UK country (England, Wales, Scotland and Northern Ireland), with England further stratified into three strata (ethnic minority, disadvantaged and advantaged) and the other three countries into two strata (disadvantaged and advantaged). For each stratum individuals are clustered by electoral ward. However, allowing for this clustering is unnecessary for our applications, as we use a subset of the MCS cohort in which most wards contain a single individual.

Using data from sweeps 1 and 2, we investigate two questions relating to the income from paid work of single mothers. We consider the gains 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. Research has found that marital splits are associated with declines in income for separating women and children (Jenkins, 2008), but it is not obvious what we should expect to find if a mother gains a partner. On the one hand we might hypothesise that on gaining a partner, maximising income from their work is no longer such a priority for mothers and they can afford to settle for a lower paid job which is more attractive on other counts. It is also possible that the mother may switch from full-time to part-time work which is less well paid, or do less overtime to boost pay. Under these scenarios we would find a decrease in hourly pay associated with gaining a partner. On the other hand, we could hypothesise that the greater stability of family life resulting from acquiring a partner allows mothers to find a better paid job. There is a lot of uncertainty surrounding this question, and the relationship could be in either direction.

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. We also exclude those who are known to be self-employed or not working in sweep 2, and four records with extreme pay values which look suspicious, leaving 559 individuals.

By definition we are looking at a set of individuals who are the mothers of very young children, so it is hardly surprising that many are working part-time. To simplify our models, we choose hourly net pay as our response variable. Hourly net pay, *hpay*, is calculated by dividing annual pay by number of hours worked in a year, and we find that the distribution of the observed *hpay* is positively skewed.

Drawing on existing literature, we select potential covariates with our motivating questions and the structure of the survey in mind. Our dataset also includes variables which may help to explain the missingness (Hawkes and Plewis, 2008). All these variables are detailed briefly in Table 1. Our educational level variable, *edu*, indicates whether or not an individual has a degree and is based on the level of National Vocational Qualification (NVQ) equivalence of the main respondent's highest academic or vocational educational qualification. The main respondent's social class, *sc*, uses the National Statistics Socio-Economic Classification (NS-SEC) grouped into 5 categories, but since we have excluded the self-employed from our dataset, there are no individuals in category 3 and *sc* has 4 levels. Our partnership status variable, *sing*, is always 1 in sweep 1 from the definition of our 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*. We restrict our analysis of this dataset to modelling the missingness in sweep 2.

Some sweep 2 data was collected from individuals who were originally non-contacts or refusals in sweep 2, after they were re-issued by the fieldwork agency. In our dataset, seven individuals have a complete set of sweep 2 variables as a result of these re-issues. We set these data to missing for the purpose of fitting our models, so they can be used subsequently for model checking.

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	educational level	2 levels[b] (1 = no degree; 2 = degree)
sing[e]	single/partner	2 levels (1 = single; 2 = partner)
sc	social class	4 levels[c] (NS-SEC 5 classes with 3 omitted)[d]
ctry	country	1 = England; 2 = Wales; 3 = Scotland; 4 = Northern Ireland
stratum	country by ward type	9 levels[f]

[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 individual's highest academic or vocational educational qualification. We regard individuals with only other or overseas qualifications as missing.
[c] 1 = managerial and professional occupations; 2 = intermediate occupations; 3 = lower supervisory and technical occupations; 4 = semi-routine and routine occupations
[d] NS-SEC 3 is small employers and own account workers, and these individuals are excluded by definition
[e] always 1 for sweep 1 by dataset definition
[f] three strata for England (advantaged, disadvantaged and ethnic minority); two strata for Wales, Scotland and Northern Ireland (advantaged and disadvantaged)

4.2 MCS example - constructing a base model

4.2.1 Choice of model of interest

For our application, we assume that based on previous work we have available a satisfactory proposed model of interest 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 and the correlation between the two data points for each individual are taken into account using stratum specific intercepts and individual random effects. Hence our model of interest is given by the equations

$$\begin{aligned} y_{it} &\sim t_4(\mu_{it}, \sigma^2) \\ \mu_{it} &= \alpha_i + \gamma_{s(i)} + \sum_{k=1}^{p} \beta_k x_{kit} \end{aligned} \tag{3}$$

for $t = 1, 2$ sweeps, $i = 1, \ldots, n$ individuals and $s = 1, \ldots, 9$ strata. α_i are the individual random effects, s.t. $\alpha_i \sim N(0, \varsigma^2)$ and its hyperparameter, ς, has a vague $N(0, 10000^2)I(0,)$ prior ($N(mean, variance)I(0,)$ denotes 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 $N(0, 10000^2)$ priors and the precision, $\frac{1}{\sigma^2}$, 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 $\boldsymbol{x}$ covariates. For the Education Question we also include *edu* (no degree/degree), whereas for the Partner Question we add *sing* (single/partner). 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.

Plausible alternative model of interest assumptions

The areas of chief concern, based on the insights gained through simulations by Mason *et al.* (2010), are: 1) choice of error distribution; 2) choice of explanatory variables and 3) choice of transform for the response. For each of these, there are a number of possible alternatives to the choice incorporated in the model of interest. We identify the alternatives we consider most plausible for sensitivity analysis to the base model: 1) assume Normal rather than t_4 errors, 2) include age^2 and for the Education Question $age \times edu$ interaction terms and 3) use a cube root rather than log transform of the response.

4.2.2 Choice of covariate model of missingness

To include the incomplete cases, we need to impute the missing sweep 2 values for *age*, *reg* and *edu* or *sing*. For *reg* and *age* we do not use a statistical model, but set their missing values prior to the analysis using simple rules. Missing *reg* are set to their sweep 1 values, which seems reasonable as amongst the 348 individuals with observed *reg* in sweep 2, only three moved from London to another region and just one moved into London. The missing values of *age* are set to the individual's sweep 1 age plus the mean difference in ages between sweeps 1 and 2 for individuals with observed age at both sweeps.

By contrast, for *edu* or *sing*, the variables which particularly interest us, we define a simple Bernoulli model. For the Education Question, if an individual has a degree in sweep 1, they must also have a degree in sweep 2 so there is no need to include them in the imputation model. Hence the covariate model of missingness can be defined as

$$\begin{aligned} edu_{j2} &\sim Bernoulli(q) \\ q &\sim Uniform(0,1) \end{aligned} \tag{4}$$

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

	complete cases		MAR response		MNAR response									
	MoI		MoI.CMoM		BASE		AS1		AS2		AS3		AS4[b]	
Education Question														
MoI: β_{age}	0.12	(0.08,0.16)	0.11	(0.08,0.14)	0.11	(0.08,0.15)	0.10	(0.07,0.14)	0.10	(0.07,0.14)	**0.40**	(0.14,0.67)	0.07	(0.05,0.09)
MoI: β_{age^2}											-0.32	(-0.58,-0.05)		
MoI: β_{edu}	0.23	(0.15,0.31)	**0.18**	(0.10,0.26)	0.22	(0.14,0.29)	**0.19**	(0.11,0.27)	0.23	(0.15,0.31)	**0.16**	(0.08,0.24)	0.15	(0.10,0.20)
MoI: $\beta_{age \times edu}$											0.11	(0.04,0.19)		
MoI: β_{reg}	**-0.16**	(-0.28,-0.04)	**-0.11**	(-0.22,0.00)	-0.13	(-0.24,-0.01)	**-0.11**	(-0.22,-0.01)	-0.15	(-0.26,-0.03)	-0.14	(-0.25,-0.03)	-0.08	(-0.16,-0.01)
RMoM: θ_0					2.66	(1.76,3.73)	**0.02**	(-1.59,1.57)	**3.10**	(2.16,4.22)	2.60	(1.73,3.63)	2.62	(1.71,3.70)
RMoM: $\theta_{level[1]}$					0.24	(0.06,0.43)	**0.14**	(-0.03,0.32)	**0.27**	(0.07,0.49)	**0.21**	(0.05,0.40)	0.25	(0.07,0.44)
RMoM: $\theta_{level[2]}$					0.54	(0.25,0.86)			**0.66**	(0.38,0.98)	0.50	(0.23,0.81)	0.54	(0.26,0.84)
RMoM: $\delta_{change[1]}$					0.62	(0.35,0.92)	**0.17**	(-0.02,0.38)	**0.75**	(0.50,1.03)	0.59	(0.33,0.87)	0.62	(0.36,0.90)
RMoM: $\delta_{change[2]}$					-0.26	(-0.40,-0.12)			**-0.34**	(-0.49,-0.21)	-0.26	(-0.40,-0.12)	-0.24	(-0.39,-0.08)
RMoM: $\theta_{ctry[2:\text{Wales}]}$					-0.15	(-0.81,0.51)	-0.15	(-0.73,0.43)	-0.15	(-0.87,0.58)	-0.16	(-0.82,0.51)	-0.16	(-0.81,0.50)
RMoM: $\theta_{ctry[3:\text{Scotland}]}$					0.18	(-0.44,0.83)	**0.24**	(-0.32,0.81)	0.17	(-0.51,0.86)	0.19	(-0.44,0.84)	0.19	(-0.45,0.83)
RMoM: $\theta_{ctry[4:\text{NI}]}$					0.28	(-0.37,0.98)	**0.21**	(-0.39,0.80)	0.30	(-0.39,1.01)	0.28	(-0.37,0.95)	0.28	(-0.37,0.96)
RMoM: θ_{eth}					-1.07	(-1.76,-0.42)	-0.98	(-1.61,-0.38)	-1.11	(-1.83,-0.39)	-1.05	(-1.73,-0.39)	-1.11	(-1.80,-0.45)
RMoM: $\theta_{sc[2]}$					-0.04	(-0.80,0.73)	**-0.38**	(-1.05,0.29)	**0.04**	(-0.78,0.85)	**-0.09**	(-0.84,0.66)	0.00	(-0.77,0.75)
RMoM: $\theta_{sc[3]}$					-0.20	(-1.24,0.87)	**-0.41**	(-1.35,0.56)	**-0.17**	(-1.28,0.95)	**-0.26**	(-1.28,0.78)	-0.15	(-1.18,0.90)
RMoM: $\theta_{sc[4]}$					0.02	(-0.74,0.80)	**-0.43**	(-1.15,0.29)	**0.13**	(-0.67,0.94)	**-0.05**	(-0.81,0.71)	0.06	(-0.69,0.83)
Partner Question														
MoI: β_{age}	0.15	(0.11,0.18)	0.13	(0.10,0.17)	0.14	(0.10,0.17)	0.13	(0.10,0.16)	0.13	(0.09,0.16)	**0.40**	(0.15,0.65)	0.09	(0.07,0.11)
MoI: β_{age^2}											-0.27	(-0.51,-0.02)		
MoI: β_{sing}	**-0.08**	(-0.15,-0.01)	**-0.07**	(-0.14,0.00)	-0.12	(-0.22,-0.02)	**-0.15**	(-0.25,-0.06)	-0.13	(-0.24,-0.01)	-0.12	(-0.22,-0.02)	-0.09	(-0.16,-0.03)
MoI: β_{reg}	**-0.18**	(-0.31,-0.05)	-0.13	(-0.24,-0.01)	-0.14	(-0.25,-0.02)	**-0.12**	(-0.23,-0.01)	**-0.17**	(-0.29,-0.05)	-0.15	(-0.26,-0.03)	-0.08	(-0.16,-0.01)
RMoM: θ_0					2.34	(1.46,3.42)	**-1.09**	(-2.87,0.69)	**3.28**	(2.25,4.57)	2.40	(1.52,3.44)	2.41	(1.45,3.61)
RMoM: $\theta_{level[1]}$					0.23	(0.06,0.45)	**0.28**	(0.07,0.52)	**0.34**	(0.11,0.59)	0.24	(0.06,0.45)	0.29	(0.09,0.55)
RMoM: $\theta_{level[2]}$					0.46	(0.17,0.81)			**0.74**	(0.43,1.10)	0.48	(0.17,0.79)	0.51	(0.20,0.87)
RMoM: $\delta_{change[1]}$					0.54	(0.26,0.86)	**0.35**	(0.10,0.62)	**0.82**	(0.54,1.14)	0.56	(0.26,0.85)	0.59	(0.28,0.93)
RMoM: $\delta_{change[2]}$					-0.15	(-0.32,0.17)			**-0.31**	(-0.47,-0.16)	**-0.17**	(-0.33,0.05)	-0.08	(-0.29,0.32)

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

[a] Percentage difference $> 10\%$ (and absolute difference > 0.02).

[b] The parameters for AS4 are not directly comparable with the other models, since AS4 uses a $\sqrt[3]{}$ rather than *log* transform of the response, so differences are not highlighted.

for $j = 1, \ldots, m$ individuals who have missing *edu* values in sweep 2 and who do not have a degree in sweep 1. A similar model is defined for all individuals who have sweep 2 missing *sing* values for the Partner Question. In this joint model consisting of our proposed model of interest and covariate model of missingness (MoI.CMoM), not only are the covariates assumed to be MAR, but the response is also assumed to be MAR.

From Table 2 we see that there are small changes in some of the model of interest parameters from fitting MoI.CMoM compared to the complete case analysis (MoI). We are interested in the imputations of the missing covariates and compare the imputed and observed covariates. 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*, 35.8% of those with observed *sing* at sweep 2 gained a partner, compared to 33.6% of those with missing sweep 2 *sing*.

This covariate model of missingness could be expanded to include the imputation of the other two covariates with missing values, *age* and *reg*. Such an extension, which would impute the missing values for multiple covariates allowing for correlation, can be implemented using a multivariate probit model approach for binary covariates (Chib and Greenberg, 1998) and its extension to ordered categorical variables (Albert and Chib, 1993) as appropriate. By creating an underlying set of latent variables in this way, models for mixtures of binary, categorical and continuous variables can be developed (Dunson, 2000; Goldstein *et al.*, 2008). Molitor *et al.* (2009) provide an example of this approach for two binary covariates.

4.2.3 Choice of response model of missingness

Our base model is completed by adding a response model of missingness of the form $m_i \sim Bernoulli(p_i)$, where m_i is a binary missing value indicator for $hpay_{i2}$, set to 1 when hourly pay in sweep 2 for individual i is observed and 0 otherwise. The addition of this sub-model changes our assumption about the missing responses from MAR to MNAR. Before defining this part of the model, we need to think about the process that led to the income missingness, gathering as much information as possible from the literature. Previous work in this area suggests that income is more likely to be missing if it is high or low, or if it has changed substantially between sweeps. Then, our findings have to be translated into a statistical model, and a piecewise linear functional form

is appropriate given our information. So, we define $logit(p_i)$ as

$$\begin{aligned} logit(p_i) &= \theta_0 + Piecewise(level_i) + Piecewise(change_i) \\ &\quad + (\theta_{ctry} \times ctry_{i1}) + (\theta_{eth} \times eth_{i1}) + \sum_{k=1}^{3} (\theta_{sc[k]} \times sc_{[k]i1}) \\ level_i &= hpay_{i1} \\ change_i &= hpay_{i2} - hpay_{i1} \\ Piecewise(level_i) &= \begin{cases} \theta_{level[1]} \times (level_i - 10): & level_i < 10 \\ \theta_{level[2]} \times (level_i - 10): & level_i \geq 10 \end{cases} \\ Piecewise(change_i) &= \begin{cases} \delta_1 \times change_i: & change_i < 0 \\ \delta_2 \times change_i: & change_i \geq 0 \end{cases} \end{aligned} \tag{5}$$

where the second index on the variables indicates sweep, and $sc_{[k]}$ is a binary indicator for sc category k (1 if sc_i is category k, 0 otherwise). The inclusion of variables sc (social class), eth (ethnic group) and $ctry$ (country) as predictors of missing income is based on work on item missingness by Hawkes and Plewis (2008). Note that in the response model of missingness we use an untransformed version of $hpay$. The specification of the piecewise linear functional form for $level$ and $change$ places the knots at £10 and £0 respectively. The priors for the $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ parameters are specified as $\theta_0 \sim Logistic(0, 1)$, $\theta_k \sim N(0, 10000^2)$ and $\delta_1, \delta_2 \sim N(0, 10000^2)$. It is the inclusion of the $\boldsymbol{\delta}$ that allows the response missingness to be MNAR. If $\delta_1 = \delta_2 = 0$, then we are assuming MAR missingness.

Plausible alternative response model of missingness assumptions

We could set up a sensitivity analysis in which the explanatory variables are varied. However, we restrict our attention to varying the functional form of these variables. A linear functional form for *level* and *change* is an obvious alternative. Possible extensions of this sub-model are discussed in Section 5.

4.3 Conclusions from base model

The parameter estimates for the two base models, BASE, are shown in Table 2. The $\boldsymbol{\theta}$ parameters associated with country, ethnicity and social class in the response model of missingness are only shown for the Education Question base model, as those for the Partner Question are similar. As regards our substantive questions, we find strong evidence that having a degree is associated with higher pay and weaker evidence that gaining a partner between sweeps is associated with lower

pay. Compared to the complete case analysis (MoI), the evidence for the association with gaining a partner has strengthened. The covariate imputations are similar to MoI.CMoM (Section 4.2.2). We defer discussion of the response model of missingness parameters until Section 4.4.2.

Our strategy also allows for including additional data (step 4) and an elicitation to provide expert priors (step 5). For example, if we wished to incorporate additional data into the covariate model of missingness part of our base model, one possibility would be to use a subset of data from the 1970 British Cohort Study (BCS70) taken from sweeps 5 and 6. 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.

As regards elicitation (O'Hagan *et al.*, 2006), 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. For our models, *level* and *change* are the key variables. If a comprehensive elicitation is impractical, extracting information about the functional form of important parameters from experts or the literature is worthwhile, and this approach informed our choice of piecewise linear functional form.

4.4 MCS example - assumption sensitivity

We now investigate some of the plausible alternative modelling assumptions noted during the building of the base model by fitting four sensitivity analyses (models AS1-AS4). These sensitivity analyses are by no means exhaustive, but are chosen to demonstrate the type of analyses that can be performed. Each of our chosen sensitivity analyses varies from our 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.

The differences between each of the models AS1-AS4 and BASE are summarised in Table 3. AS1 allows us to explore sensitivity to the choice of functional form in the response model of missingness, while the last three (AS2, AS3 and AS4) investigate different model of interest assumptions.

Table 3: Summary of the differences of the joint models for the sensitivity analyses (AS1-AS4) from the base model (BASE)

model	difference from BASE[a]
AS1	linear functional form for *level* and *change* in the response model of missingness
AS2	Normal error distribution for the model of interest
AS3	additional covariates age^2 and $age \times edu^b$ for the model of interest
AS4	cube root transform of response for the model of interest

All models run with 2 chains for 100,000 iterations of which 50,000 are burn-in and a thinning rate of 5.
[a] The key features of BASE are: model of interest - covariates $\{age, reg, edu$ or $sing\}$, a log transform of the response and a t_4 error distribution; response model of missingness - piecewise linear functional form for *level* and *change*.
[b] Education Question only.

Parameter estimates are given for these four models in Table 2. Non-negligible differences in the parameter estimates of AS1-AS3 from BASE are highlighted in bold, where for this purpose a non-negligible difference is defined as a percentage difference greater than 10% (and an absolute difference greater than 0.02). The parameters for AS4 are not directly comparable with the other models, because AS4 uses a different transform of the response, so differences are not highlighted.

4.4.1 Robustness of the conclusions to substantive questions

All four sensitivity analyses support the conclusions from BASE that gaining a degree is associated with higher pay, although there is some variation in the strength of support. For the question relating to partnership status, AS1 provides stronger evidence that gaining a partner between sweeps is associated with lower pay than BASE or AS2-AS4.

4.4.2 Robustness of the response model of missingness parameter estimates

For both questions, the estimates of the response model of missingness parameters for BASE suggest that sweep 2 pay is more likely to be missing for individuals who are non-white, but social class and country make little difference. Those with low levels of hourly pay in sweep 1 are more likely to be missing, as are those whose pay changes substantially between sweeps. Adding extra covariates to the model of interest (AS3) and using a cube root transform for the response in the model of interest (AS4) provide a consistent message. However, assuming Normal rather than t_4 errors for the model of interest (AS2) increases the magnitude of the posterior means of the parameters associated with *level* and *change*.

It is difficult to compare the response model of missingness parameters for BASE and AS1 directly

because of the change in functional form. However, *level*, *change* and *eth* remain important predictors of missingness. Regardless of functional form, the 95% intervals of the *change* parameters ($\boldsymbol{\delta}$) do not include zero, providing evidence of informative missingness given the model assumptions.

4.4.3 Fit of re-issued individuals

In Section 4.1 we described how data that was collected from seven individuals, as a result of re-issues, was set to missing. Using data collected from individuals who were originally non-contacts or refusals for checking the fit of our models is attractive as such individuals are likely to be similar to those who have missing data, although we must treat our findings with caution as seven is small. For each individual, a *Bayesian p-value* (Gelman and Meng, 1996) can be calculated in WinBUGS as the proportion of iterations in which a higher value than the observed value was imputed. For both BASE models, these do not suggest any great conflict with the proposed model (p-values range from 0.18 to 0.76 for the Education Question and from 0.13 to 0.82 for the Partner Question). Density plots (not shown) of the posterior predictive distribution of hourly pay for each individual imputed by BASE show that hourly pay is estimated reasonably well for all seven individuals, but the predictions for all the individuals are subject to considerable uncertainty.

We also calculate the mean square error (MSE) of the fit of hourly pay for the seven individuals, and use this as a summary measure of the performance of our models in predicting their sweep 2 hourly pay. Table 4 shows the median and 95% interval MSE of the fit of hourly pay for the seven individuals for our base model, BASE, and the four models run as a sensitivity analysis for both questions. The posterior distribution of this MSE is somewhat skewed, so the median is a better measure than the mean, which is unstable due to outlying values of hourly pay for some individuals. The models with the cube root transform (AS4) and the linear functional form for the response model of missingness (AS1) fit the seven re-issued individuals best, and there are slight improvements over BASE for AS3.

4.5 MCS example - parameter sensitivity

The values of δ_1 and δ_2 control the degree of departure from MAR missingness. We know that these two parameters are difficult for a model with vague priors to estimate, and lack of convergence sometimes provides clear evidence of this. 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 $\boldsymbol{\delta}$.

Table 4: MSE of imputed hourly pay for seven re-issued individuals for models BASE and AS1-AS4

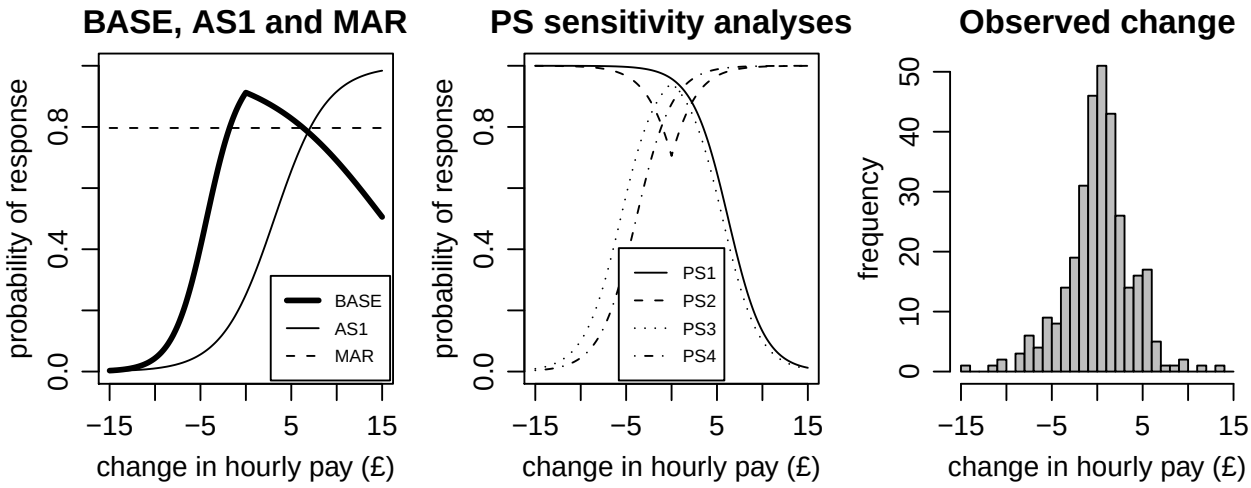
	Education Question		Partner Question	
	median	95% interval	median	95% interval
BASE	23.6	(3.6,386.9)	21.4	(3.6,347.8)
AS1	10.6	(3.1,40.2)	9.5	(3.2,26.1)
AS2	22.3	(3.8,117.5)	29.1	(4.2,154.1)
AS3	16.8	(3.0,340.7)	16.4	(3.1,296.2)
AS4	9.0	(2.1,79.8)	8.0	(2.1,55.2)

In the same spirit, we also carry out a sensitivity analysis in which a series of models is run with these two parameters fixed. We refer to this group of models as PS (Parameter Sensitivity), and it contains eighty-one variants which are formed by combining nine values of δ_1 with each of nine values of δ_2. We use the same set of values, namely $\{-1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1\}$, for both δ_1 and δ_2, which encompasses the estimated values from BASE and AS1 (see $\boldsymbol{\delta_{change}}$ values in Table 2). The design includes nine variants in which the functional form of *change* is linear, i.e. $\delta_1 = \delta_2$, with the $\delta_1 = \delta_2 = 0$ variant equivalent to assuming the response is MAR. In contrast to BASE which estimates δ_1 and δ_2, the PS models fix δ_1 and δ_2 using point priors. An alternative would be to use strongly informative priors.

Before analysing the results, we consider the interpretation of the $\boldsymbol{\delta}$ parameters by looking at the probability of an individual responding. The left and centre plots in Figure 2 show the probability of response as the change in hourly pay varies assuming that all the other covariates are fixed (we use level of pay=£10; country=England; ethnicity=white and social class=1), for some of the scenarios we have analysed for the Partner Question. (Similar plots can be produced for the Education Question, but are not shown.) We see that for the base case the probability of responding reduces to almost 0 for reductions in hourly pay of more than £10. For the parameter sensitivity analyses with $\boldsymbol{\delta}$ of 0.5 or -0.5, we also find that the probability of responding can become close to 0 when pay changes by more than £10 in either direction. However, given their hourly pay levels, we do not expect many of the individuals being modelled to have a change in hourly pay with a magnitude much greater than £5, and this is certainly the case for those with observed change (see right plot in Figure 2). Similar plots for parameter sensitivity analyses with $\boldsymbol{\delta}$ of 1 or -1 (not shown) reveal a further narrowing of the range of non-zero probabilities of responding.

We now examine the robustness of our conclusions regarding the substantive questions to changes in the $\boldsymbol{\delta}$ parameters, by analysing PS. The range of results, in terms of the proportional increase in

Figure 2: Probability of response under different scenarios for the Partner Question and frequency of observing different levels of change in hourly pay



All other covariates fixed: level of pay=£10; country=England; ethnicity=white and social class=1.
PS1: $\delta_1 = \delta_2 = -0.5$; PS2: $\delta_1 = -0.5$, $\delta_2 = 0.5$; PS3: $\delta_1 = 0.5$, $\delta_2 = -0.5$; PS4 $\delta_1 = \delta_2 = 0.5$.

hourly pay associated with having a degree or gaining a partner is shown in Table 5. The effect of gaining a partner between sweeps is most sensitive to the different values of $\boldsymbol{\delta}$. If all the PS variants are plausible, then we cannot even be sure about the direction of this effect, as the models suggest a range of conclusions from strong evidence of a positive effect to strong evidence of a negative effect. The results from the MAR analysis lie between the two extremes, close to the base model (BASE). We now look at ways of presenting the PS results in more detail.

Table 5: Proportional increase in pay associated with having a degree or gaining a partner for PS variants compared with base model (BASE)

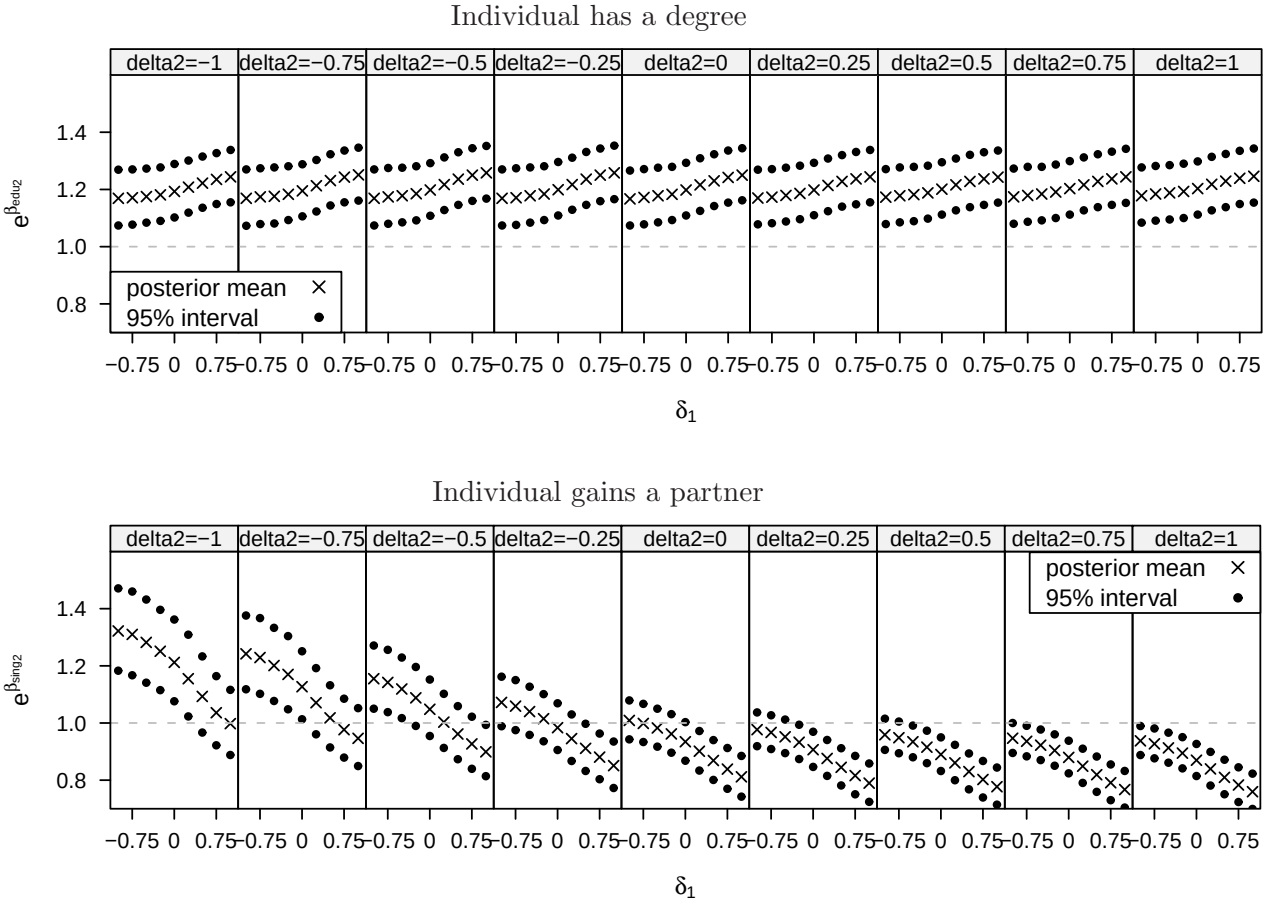
	minimum	min δ_1[a]	min δ_2[a]	maximum	max δ_1[a]	max δ_2[a]	MAR[b]	BASE
degree	1.17	-1	0	1.26	1	-0.5	1.20	1.24
	(1.07,1.27)			(1.17,1.35)			(1.09,1.28)	(1.15,1.34)
partner	0.76	1	1	1.32	-1	-1	0.93	0.89
	(0.70,0.82)			(1.18,1.47)			(0.88,0.99)	(0.80,0.98)

Table shows the posterior mean, with the 95% interval in brackets.
[a] min δ_1, min δ_2 and max δ_1, max δ_2 are the values of δ_1 and δ_2 corresponding to the PS variant with the parameter's lowest and highest posterior mean respectively.
[b] $\delta_1 = 0$ and $\delta_2 = 0$ is MAR.

Yun *et al.* (2007) plot a quantity of interest against a sensitivity parameter, for a set of models in which this sensitivity parameter is fixed to a range of values, and use the resulting sensitivity plot to show the dependence of their conclusions to assumptions about the missingness. We extend this idea to two sensitivity parameters, using trellis graphs to demonstrate the level of robustness of our

quantities of interest to changes in δ_1 and δ_2. Figure 3 shows the posterior mean and 95% interval of the proportional increase in pay associated with having a degree and gaining a partner. Our conclusions regarding the Partner Question are clearly dependent on the $\boldsymbol{\delta}$ values, but are more robust for the Education Question.

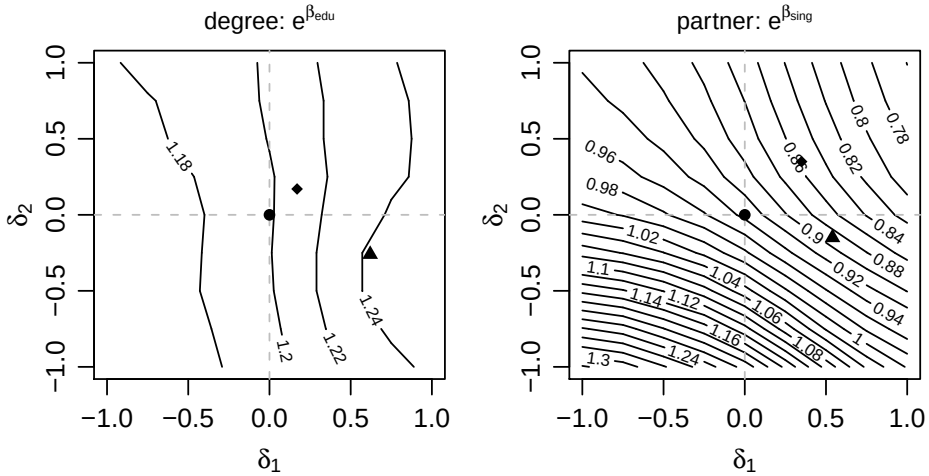
Figure 3: Estimated proportional increase in pay associated with having a degree and gaining a partner versus δ_1 conditional on δ_2 from PS variants



For an alternative presentation of the results, we plot the posterior means of the estimated proportional change in hourly pay associated with a particular covariate as a series of contour lines (Figure 4). The points at the $\boldsymbol{\delta}$ values relating to MAR missingness, our base model (BASE) and the response model of missingness linear functional form sensitivity analysis (AS1) are marked with a circle, triangle and diamond respectively. The closer the contour lines, the greater the variation in the proportional change in pay associated with a selected covariate as δ_1 and δ_2 change. So the sparsity of lines in the left plot of Figure 4 indicates the relative robustness of the results relating to gaining a degree. By contrast, the dense contours in the right plot show that the proportional increase of 1.3 in pay associated with gaining a partner when δ_1 and δ_2 are at their most nega-

tive ($\delta_1 = \delta_2 = -1$), reduces as either δ_1 or δ_2 increases, until there is a substantial proportional decrease (0.78) when δ_1 and δ_2 are at their most positive ($\delta_1 = \delta_2 = 1$).

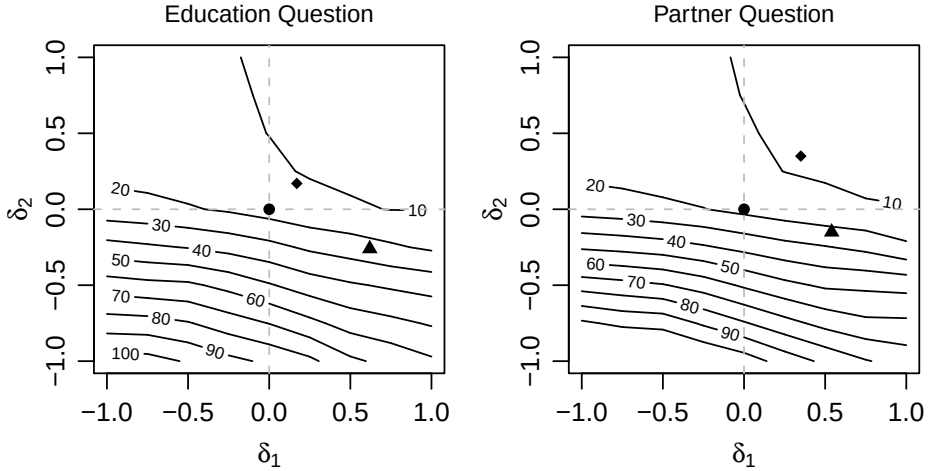
Figure 4: Posterior mean of proportional change in pay associated with having a degree and gaining a partner versus δ_1 and δ_2 from PS variants



The points at the $\boldsymbol{\delta}$ values relating to MAR missingness, BASE and the RMoM linear functional form sensitivity analysis (AS1) are marked with a circle, triangle and diamond respectively.

As with the assumption sensitivity analyses, we use the mean square error (MSE) of the fit of hourly pay for the seven re-issued individuals as a measure of model fit. Figure 5, which uses the same format as Figure 4, suggests that a plausible range of values for our quantities of interest should be based on models which fall in the upper right quadrant, i.e. the quadrant with positive δ_1 and positive δ_2.

Figure 5: The mean square error of the fit of hourly pay for the re-issued individuals versus δ_1 and δ_2 from PS variants



The points at the $\boldsymbol{\delta}$ values relating to MAR missingness, BASE and the RMoM linear functional form sensitivity analysis (AS1) are marked with a circle, triangle and diamond respectively.

Given models with $\boldsymbol{\delta}$ values outside the -0.5 and 0.5 range over-narrow the range of plausible response probabilities, the information from all these plots suggests that if our conclusions take account of the PS sensitivity analyses with $\delta_1 = \delta_2 = 0$ (MAR) and $\delta_1 = \delta_2 = 0.5$, then we will be adequately reporting the results from the range of plausible PS models.

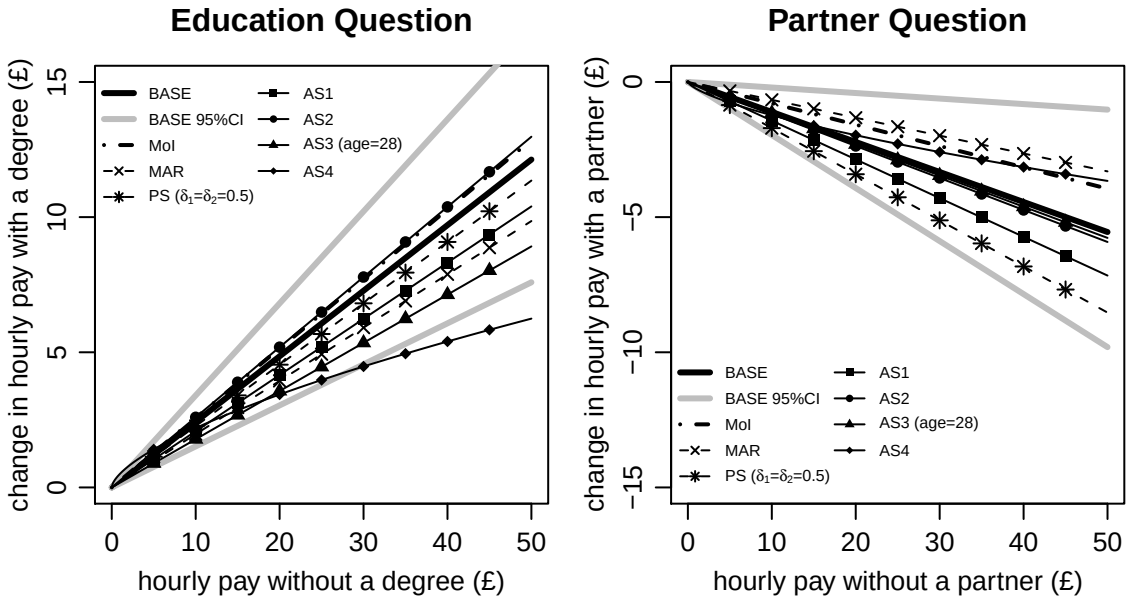
4.6 Reporting conclusions on substantive questions

Our analysis for the Education question clearly suggests that having a degree is associated with higher pay, but there is some uncertainty about the level of this association. If the results had been robust, we could simply report the base case analysis and state that the conclusions are robust to a sensitivity analysis, giving an indication of what this involved. However, given the uncertainty, we report a range of results covering the region of high plausibility which we have identified, comprising the base case, four assumption sensitivity analyses and the parameter sensitivity analyses MAR and $\delta_1 = \delta_2 = 0.5$. To report these in an interpretable manner which enables direct comparison across all the models, we present a graph (left plot in Figure 6) showing the change in hourly pay for an individual with a degree against their hourly pay if they do not have a degree (all other characteristics remain unchanged). We find that most of the sensitivity analyses suggest lower increases from gaining a degree than the base case. The interquartile range for hourly pay in sweep 1 in our dataset is (£5,£10), with just 2 individuals earing over £25. Our analysis suggests that for an individual earning £10 an hour, gaining a degree would make a difference of between £1.78 (£0.87,£2.76)(AS3) and £2.59 (£1.62,£3.67)(AS2) an hour (based on the posterior means, with 95% credible intervals shown in brackets).

In addition to plotting the black lines based on the posterior means of the parameters associated with gaining a degree, we have added grey lines calculated using the 2.5 and 97.5 percentiles of the posterior distribution for the base case to indicate the uncertainty within a particular model. At £10, this suggests an increase between £1.52 and £3.40 is plausible. These 95% interval lines encompass the point estimates for all the scenarios except that using a cube root transform (AS4), where the curved profile suggests that there is less of a gain for those who already have high pay.

We also include the results of the complete case analysis (dashed-dot line labelled MoI) for comparison. This line is almost coincident with the AS2 line, and slightly higher than the base case line and all the other sensitivity analyses shown. Interval lines for the complete case analysis (not shown) are close to their base case counterparts. Using MoI, at £10, gaining a degree results in a

Figure 6: Presentation of results on substantive questions



All lines have been calculated using the posterior means of the parameters associated with gaining a degree or acquiring a partner, apart from the BASE 95% credible interval (BASE 95%CI) lines which use the 2.5 and 97.5 percentiles of the posterior distribution. Note that for the Education Question the MoI and AS2 lines are almost coincident.

£2.57 (£1.61,£3.60) increase in pay.

Similarly, we report our conclusions for the Partner Question using the right plot in Figure 6. This leads us to report that there is weak evidence that gaining a partner is associated with lower pay, and the reduction is likely to be between £0.66 (£1.32,-£0.03)(MAR) and £1.71 (£2.32,£1.07)(PS) an hour for an individual earning £10 an hour. The MAR 95% interval just includes no change. For this question, in contrast to the Education question, the uncertainty generated by the missing data is greater for the parameter sensitivity analysis than the assumption sensitivity analysis. The complete case analysis (£0.79 (£1.42,£0.11) at £10) is further removed from the base case (£1.11 (£1.96,£0.20) at £10) and captures less of the uncertainty as the width of the 95% interval is reduced. We also point out that 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 in the region of high plausibility.

5 Extensions 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 we use separate missingness indicators for the covariates and the response or should we use an overall missingness indicator for attrition? If we use separate indicators, a new sub-model linked to the existing covariate model of missingness is required. In implementing this we would need a different indicator for each covariate pattern of missingness. Alternatively, if we use an overall missingness indicator for attrition, we then also require a method for dealing with any item missingness that occurs in the response or covariates. Although in theory a model allowing MNAR covariates could be designed, it may currently be computationally prohibitive in WinBUGS. Conversely, if we have 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 non-response that can occur, but in our applications modelled the missing data as a homogeneous process. However, the non-response 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 we know that these three types of non-response have different correlates (Plewis, 2007*b*; Plewis *et al.*, 2008), there is considerable scope for expanding the sensitivity analysis to re-specify the response model of missingness to specifically allow for these differences. This could be implemented by modelling a missingness indicator with separate categories for each type of non-response using multinomial regression.

In our application, we do not distinguish between missing income resulting from the entire sweep being missing and item non-response. Also, as we are only using two sweeps of data, we cannot distinguish between wave non-response and attrition. Including a further sweep would allow this distinction to be made. Again, extending the response model of missingness using multiple missingness indicators would allow different predictors to be used for item missingness, wave missingness and attrition. Further, we restricted our analysis to individuals with fully observed data in sweep 1, and individuals with sweep 1 missingness could also be incorporated.

There are situations where it may be necessary to adapt this strategy. For example, if the dataset to be modelled is very large, or there are large numbers of covariates with missingness, then running times may be prohibitive or computational issues encountered. In these circumstances, one option is

to take a two stage approach and impute some or all of the covariates prior to running the other parts of the model. In this case issues surrounding multiple imputation regarding compatibility will apply (Rubin, 1996; Carpenter and Goldstein, 2004), and a combination of Bayesian and non-Bayesian methods are used. It may be possible to identify covariates where using simplistic assumptions to impute their missingness is acceptable (as we have done for age and region here). If not all the covariates are correlated, another option is to split the covariate model of missingness into several smaller sub-models. Although we have implemented our strategy using Bayesian models, there is no reason why the general principles could not be adapted for a non-Bayesian framework.

6 Conclusions

Compared to performing a complete case analysis, the implementation of this strategy 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. The computing for the applications in this paper was reasonably rapid, typically 30-40 minutes for each model on a desktop computer with a dual core 2.4GHz processor and 3.5GB of RAM. 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.

A complete case analysis leads to broadly similar conclusions to our substantive questions as the full sensitivity analysis, i.e. we would report an increase in pay associated with gaining a degree and a decrease in pay associated with gaining a partner. However, the complete case analysis tends to overestimate the magnitude of the increase for the Education Question and underestimate the decrease for the Partner Question. For the second question, the complete case analysis also fails to fully capture the uncertainty in the estimates. In short, the conclusions about a substantive question will be more soundly based and can be reported with greater confidence if our proposed strategy is followed.

Appendices

This appendix contains the WinBUGS code for running the base model for the Education Question. Function *elicitor.piecewise*, written by Mary Kynn, is used to implement the piecewise linear regression in the response model of missingness, and can be downloaded from the WinBUGS development website (http://www.winbugs-development.org.uk/). Alternatively, extra code could be written in WinBUGS to specify this model.

```
# WinBUGS code for running the base model for the Education Question
# Model of Interest: hpay logged, covariates {age, edu, reg}, individual random effects, stratum specific intercepts and t4 errors
# Covariate Model of Missingness: imputes missing edu for individuals who do not have a degree in sweep 1
# Response Model of Missingness: logit(p) regressed on {change, level, ctry, eth, sc}, using piecewise linear functions for
#   change and level
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.edu[edu[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]]+levelPW[i]+changePW[i]
            linkp[i]<- theta0+theta.eth[eth[i]]+theta.ctry[ctry[i]]+theta.sc[sc[i]]+levelPW[i]+changePW[i]
            levelPW[i] <- elicitor.piecewise(e.hpay[i,1],3,1,level.slope[],level.knot[])
            change[i]<-e.hpay[i,2]-e.hpay[i,1]
            changePW[i] <- elicitor.piecewise(change[i],3,1,change.slope[],change.knot[])
        }

        # Education imputation model - sweep 2 only
        for (i in 1:Nlevel1.edu) { # loop through edu with level 1 in sweep 1
            edu.ind[edulevel1.lst[i]]~dbern(q)
        }

        for (i in 1:Nmiss.edu) { # loop through missing covariate edu
            edu[edu.mlst[i],2]<-edu.ind[edu.mlst[i]]+1
        }

        for (i in 1:Nri) { # loop through re-issues who provided data
            ri.hpay[i]<-exp(hpay[ri.lst[i],2])  # unlog hourly pay
            ri.pval[i]<-step(Tpay[i]-ri.hpay[i])
            ri.MSEcontrib[i]<-pow(Tpay[i]-ri.hpay[i],2)
            ril.hpay[i]<-hpay[ri.lst[i],2]  # logged hourly pay
            ril.pval[i]<-step(log(Tpay[i])-ril.hpay[i])
            ril.MSEcontrib[i]<-pow(log(Tpay[i])-ril.hpay[i],2)
            }
        ri.MSE<-sum(ri.MSEcontrib[])/Nri
        ril.MSE<-sum(ril.MSEcontrib[])/Nri
```

```
        # Priors for model of interest
        beta0.sigma~dnorm(0,0.00000001)I(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.edu[1]<-0  # alias first level of edu beta
        beta.edu[2]~dnorm(0,0.00000001)
        beta.reg[1]<-0 # alias first level of reg beta
        beta.reg[2]~dnorm(0,0.00000001)
        tau~dgamma(0.001,0.001)
        sigma<-sqrt(2 / tau)  # t errors on 4 degrees of freedom

        # Priors for Education imputation model
        q ~ dunif(0,1)

        # Priors for response missingness model
        theta0 ~ dlogis(0,1)
        level.knot[1] <- 0
        level.slope[1] ~ dnorm(0,0.00000001) # theta level 1
        level.knot[2] <- 10.0
        level.slope[2] ~ dnorm(0,0.00000001) # theta level 2
        level.knot[3] <- 1000
        change.knot[1] <- -100
        change.slope[1] ~ dnorm(0,0.00000001) # delta 1
        change.knot[2] <- 0.0
        change.slope[2] ~ dnorm(0,0.00000001) # delta 2
        change.knot[3] <- 100
        theta.eth[1] <- 0
        theta.eth[2] ~ dnorm(0,0.00000001)
        theta.sc[1] <- 0
        for (sc in 2:4) { theta.sc[sc]~dnorm(0,0.00000001) }
        theta.sc[2] ~ dnorm(0,0.00000001)
        theta.sc[3] ~ dnorm(0,0.00000001)
        theta.sc[4] ~ dnorm(0,0.00000001)
        theta.ctry[1] <- 0
        for (c in 2:4) { theta.ctry[c]~dnorm(0,0.00000001) }

        # Odds ratios
        edu.or<-exp(beta.edu[2])

        # Covariate contingency tables - sweep 1 can only be 1
        for (i in 1:2)  { # loop through sweep 2 categories
                        # loop through individuals
                        for (j in 1:Nmiss.edu) { edu.count[j,i]<-equals(edu[edu.mlst[j],2],i)}
                        edu.tab[i]<-sum(edu.count[,i])  # counts individuals with missing sweep 2 edu only
                        edu.pctab[i]<-edu.tab[i]/sum(edu.tab[])  # percentages of individuals with missing sweep 2 edu only
                }
}
```

References

Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, **88**, (422), 669–79.

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**, (461), F82–99.

Brooks, S. and Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–55.

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 2008 and posted on `www.iser.essex.ac.uk`.

Carpenter, J. R. and Goldstein, H. (2004). Multiple imputation in MLwiN. *Multilevel modelling newsletter*, **16**, (2).

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, **85**, (2), 347–61.

Daniels, M. J. and Hogan, J. W. (2008). *Missing Data In Longitudinal Studies Strategies for Bayesian Modeling and Sensitivity Analysis*. 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**, (2), 355–66.

Gelman, A. and Meng, X.-L. (1996). *Markov Chain Monte Carlo in Practice*, chapter 11: Model checking and model improvement, pp. 189–201. Chapman & Hall, (1st edn).

Goldstein, H., Carpenter, J., Kenward, M. G., and Levin, K. A. (2008). Multilevel models with multivariate mixed response types. *Statistical Modelling*. To appear.

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**, (3), 479–91.

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.

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**, (469), 332–46.

Jenkins, S. P. (2008). Marital splits and income changes over the longer term. ISER working paper 2008-07, Institute for Social and Economic Research, University of Essex.

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

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, (2nd edn). John Wiley and Sons.

Mason, A., Best, N., Plewis, I., and Richardson, S. (2010). Insights into the use of Bayesian

models for informative missing data. Technical report, Imperial College London. available at www.bias-project.org.uk.

Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Studies*, (1st edn). 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 (Statistics in Society)*. To appear.

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

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

Plewis, I. (2007*b*). Non-Response in a Birth Cohort Study: The Case of the Millennium Cohort Study. *International Journal of Social Research Methodology*, **10**, (5), 325–34.

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**, (3), 1–22.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, **63**, (3), 581–92.

Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Society*, **91**, (434), 473–89.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, (1st edn). Chapman & Hall.

Schafer, J. L. and Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods*, **7**, (2), 147–77.

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

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.

Yan, T., Curtin, R., and Jans, M. (2010). Trends in Income Nonresponse Over Two Decades. *Journal of Official Statistics*, **26**, (1), 145–64.

Yun, S.-C., Lee, Y., and Kenward, M. G. (2007). Using hierarchical likelihood for missing data problems. *Biometrika*, **94**, (4), 905–19.

Zhan, M. and Pandey, S. (2004). Postsecondary Education and Economic Well-Being of Single Mothers and Single Fathers. *Journal of Marriage and Family*, **66**, (3), 661–73.