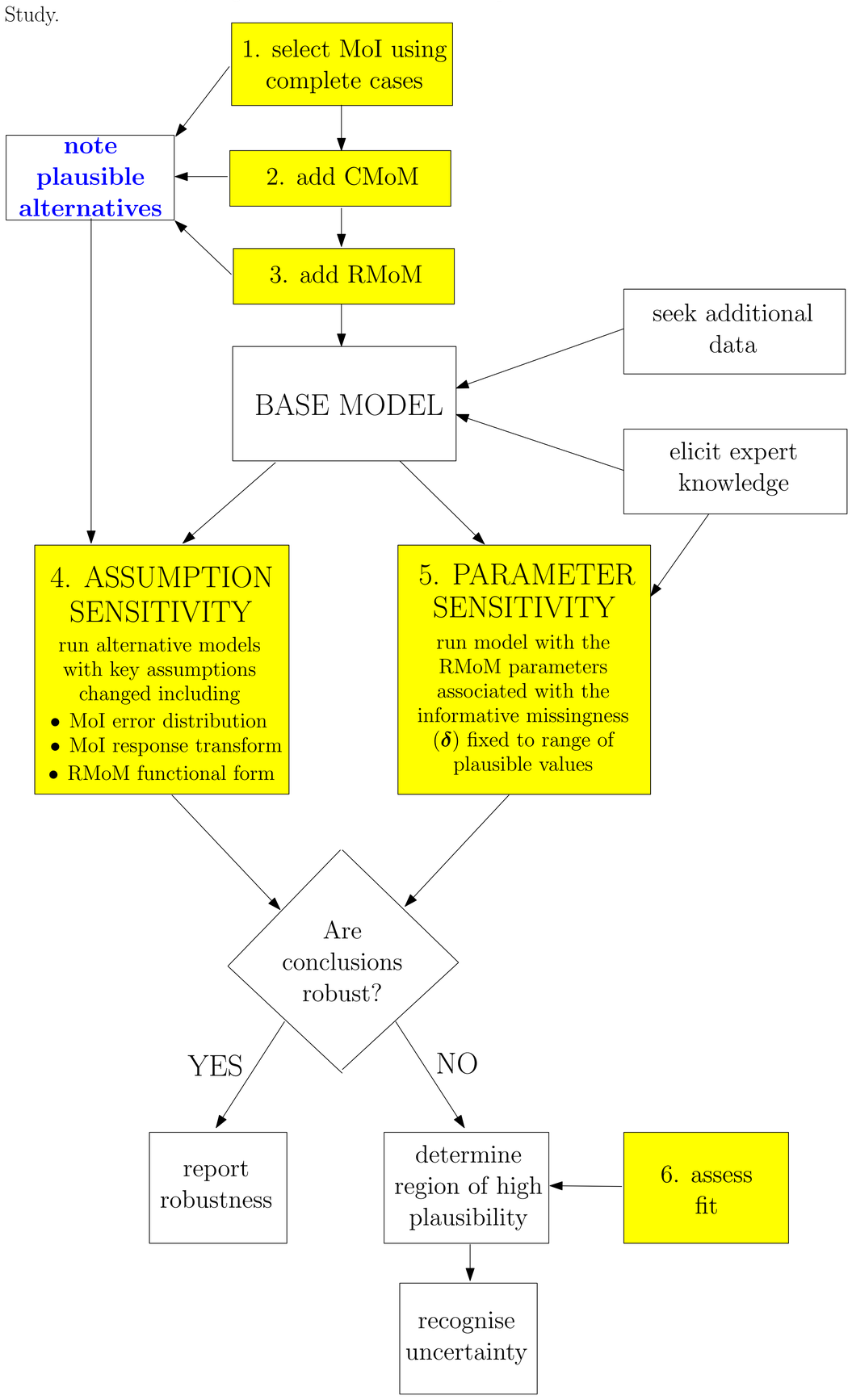
Strategy for modelling non-random missing data mechanisms in longitudinal studies: application to income data from the Millennium Cohort Study

Alexina Mason, Nicky Best, Sylvia Richardson (Imperial College London) and Ian Plewis (University of Manchester)

Technical report available from www.bias-project.org.uk

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 a question relating to the prediction of income, using data from the first two sweeps of the Millennium Cohort Study.



Description of Data

We model mothers who are single in sweep 1, in paid work and not self-employed. The missing covariates are assumed to MAR, but the missing responses are allowed to be MNAR. We model only the sweep 2 missingness.

QUESTION OF INTEREST: does change in partnership status affect income?

1. Select a Model of Interest (MoI) based on complete cases

Our proposed model takes account of the design of the survey and the correlation between the two data points for each individual.

2. Add a Covariate Model of Missingness (CMoM)

The MoI will not run with missing covariates, so we must add a CMoM to incorporate incomplete cases. Missing sweep 2 values for *sing* are imputed using the equations

$$sing_{i2} \sim Bernoulli(q).$$
$$q \sim Uniform(0,1)$$

Missing sweep 2 values for *reg* are set to their sweep 1 values, and for *age* to their sweep 1 values plus the mean of the difference between the values for sweeps 1 and 2 for observed individuals.

3. Add a Response Model of Missingness (RMoM)

This sub-model allows informative missingness in the response, by modelling m_i, a binary missing value indicator for y_{i2}, s.t.

$$m_i = \begin{cases} 1: & y_{i2} \text{ observed} \\ 0: & y_{i2} \text{ missing} \end{cases}$$

$$m_i \sim Bernoulli(p_i)$$
$$logit(p_i) = \theta_0 + Piecewise(level_i) + Piecewise(change_i) + \sum_k \theta_k w_{ki}$$

$hpay_{i1}$; $hpay_{i2} - hpay_{i1}$; *eth* (ethnic group), *sc* (social class), *ctry* (country)

choice of functional form and position of knots are based on expert knowledge

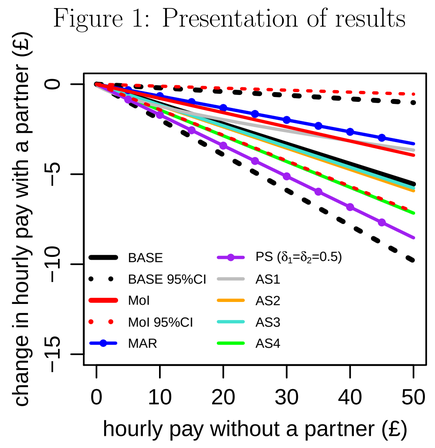
Alternative (AS4): linear functional form

$$Piecewise(level_i) = \begin{cases} \theta_{level[1]} \times (level_i - 10): & level_i < 10 \\ \theta_{level[2]} \times (level_i - 10): & level_i \geq 10 \end{cases}$$

$$Piecewise(change_i) = \begin{cases} \delta_1 \times change_i: & change_i < 0 \\ \delta_2 \times change_i: & change_i \geq 0 \end{cases}$$

& vague priors

Figure 1: Presentation of results



4. Assumption Sensitivity

Figure 1 shows 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) for a range of models.

BASE provides some evidence that gaining a partner between sweeps is associated with lower pay. There is clearly some sensitivity to our model assumptions, and AS4 (linear functional form of RMoM) provides stronger evidence that gaining a partner is associated with lower pay. By comparison, the complete case analysis (MoI) underestimates the decrease and fails to fully capture the uncertainty in the estimates.

5. Parameter Sensitivity

The values of δ_1 and δ_2 control the degree of departure from MAR missingness. So for the parameter sensitivity, a series of models is run with $\boldsymbol{\delta}$ fixed to different values.

Sensitivity of the proportional change in pay associated with gaining a partner between sweeps to the different assumptions can be displayed graphically, and two possibilities are shown (Figures 2 and 3).

If all the PS variants are plausible, then we cannot even be sure about the direction of the effect of change in partnership status on income, as the models suggest a range of conclusions from strong evidence of a positive effect to strong evidence of a negative effect.

Figure 2: Posterior mean of proportional change in pay associated with gaining a partner between sweeps

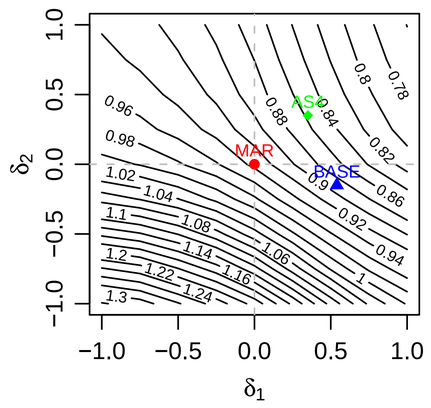
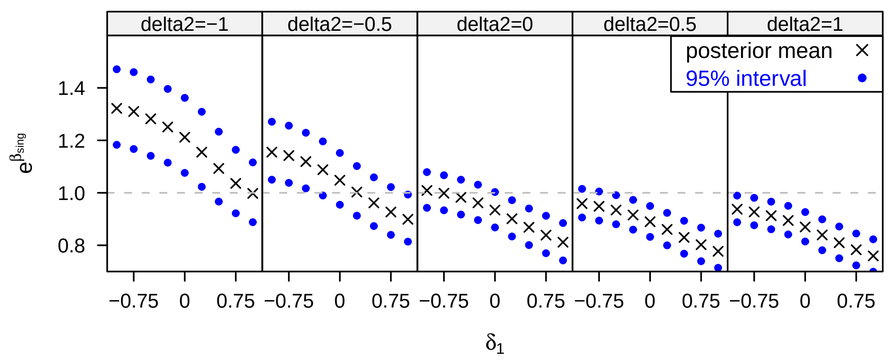


Figure 3: Proportional change in pay associated with gaining a partner between sweeps



6. Assess fit of validation sample

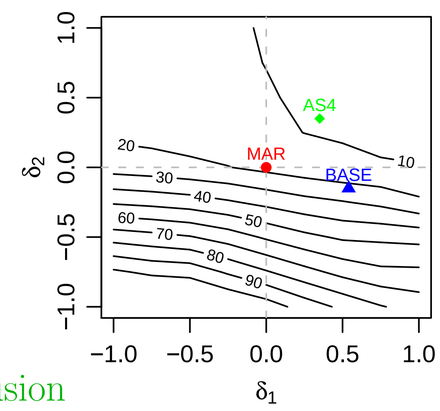
Some sweep 2 data were collected from 7 individuals who were originally non-contacts or refusals in sweep 2, after they were re-issued by the fieldwork agency. We set these data to missing before fitting our models, so they can now be used for model checking.

We calculate the mean square error (MSE) of the fit of hourly pay for these 7 individuals, for use as a summary measure of the performance of our models. For the assumption sensitivities, Table 1 suggests that the models with the linear functional form for the RMoM (AS4) and with the cube root transform (AS1) fit the 7 re-issued individuals best. Regarding the parameter sensitivity, from Figure 4 this measure provides greatest support for the models in the upper right quadrant. The results from two such models are shown in Figure 1.

Table 1: MSE of imputed hourly pay for 7 re-issued individuals

	MSE for re-issues	
	median	95% interval
BASE	21.4	(3.6,347.8)
AS1	8.0	(2.1,55.2)
AS2	29.1	(4.2,154.1)
AS3	16.4	(3.1,296.2)
AS4	9.5	(3.2,26.1)

Figure 4: MSE of imputed hourly pay



Conclusion

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. 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.