

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

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

Description of Data

The data for our application is taken from the first two sweeps of the Millennium Cohort Study, and we model the subset of main respondents (usually the cohort member's mother) who are single in sweep 1, in paid work and not self-employed. Analysis is complicated by missing covariates and missing responses, and the missing responses are thought to be generated by an informative missingness mechanism. As a simplification we model only the sweep 2 missingness. Using this example, we demonstrate a strategy using Bayesian methods that provides a flexible way of carrying out a 'statistically principled' investigation of such data.

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 hierarchical structure of the data.

log of hourly pay ($hpay$)

Alternative (AS1):
cube root transform

Alternative (AS2): Normal errors

$$y_{it} \sim t_4(\mu_{it}, \sigma^2)$$

$$\mu_{it} = \alpha_i + \gamma_{s(i)} + \sum_{k=1}^p \beta_k x_{kit} + \sum_{k=p+1}^q \beta_k z_{ki}$$

robustness to outliers

individual random effects

stratum specific intercepts

age (main respondent's age)

edu (educational level)

reg (London/other)

sing (single/partner)

eth (ethnic group)

Alternative (AS3): include age^2 and $age \times edu$ interaction terms

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 *age*, *edu* and *sing* are jointly imputed using latent variables for categorical *edu* and binary *sing*. Missing sweep 2 values for *reg* are set to their sweep 1 values. We compare the pattern of the imputed values with observed values to check the model is reasonable.

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.

$$m_i = \begin{cases} 1: & y_{i2} \text{ observed} \\ 0: & y_{i2} \text{ missing} \end{cases} \quad \begin{matrix} eth \text{ (ethnic group)} \\ sc \text{ (social class)} \\ ctry \text{ (country)} \end{matrix}$$

$$hpay_{i1} \quad hpay_{i2} - hpay_{i1}$$

$$m_i \sim \text{Bernoulli}(p_i)$$

$$\text{logit}(p_i) = \theta_0 + \text{Piecewise}(\text{level}_i) + \text{Piecewise}(\text{change}_i) + \sum_k \theta_k w_{ki}$$

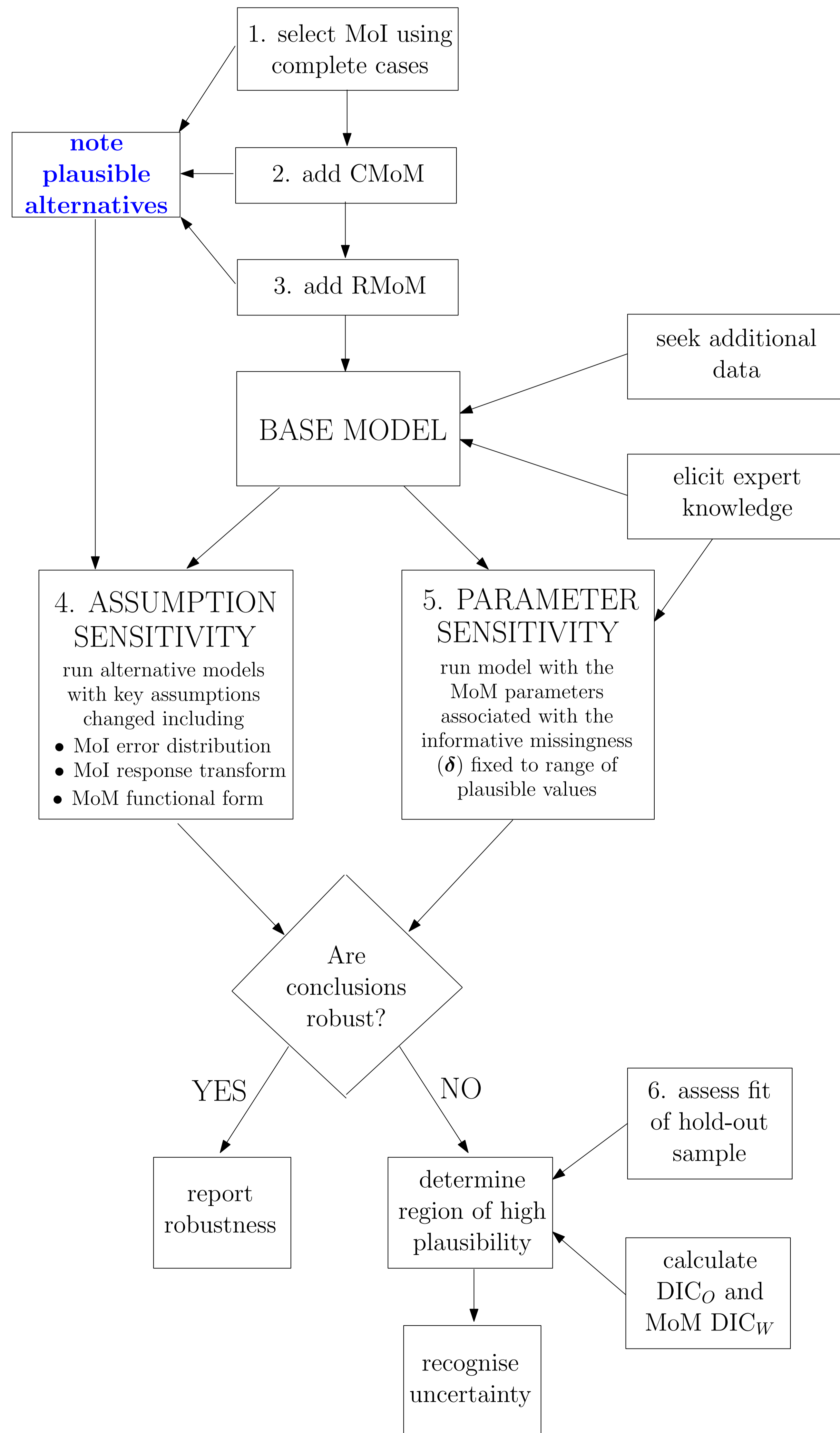
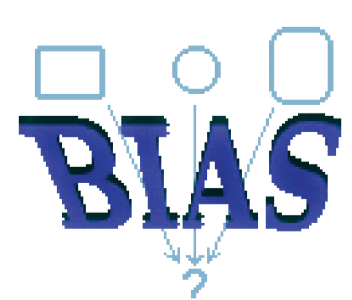
choice of functional form and position of knots are based on expert knowledge

Alternative (AS4): linear functional form

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

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

& vague priors



4. Assumption Sensitivity

Table 1 compares the proportional change in pay associated with gaining a partner between sweeps ($e^{\beta_{sing}}$) given by the model of interest (MoI), base model (BASE) and four assumption sensitivities (AS1-AS4).

BASE provides some evidence that gaining a partner between sweeps is associated with lower pay. AS4 (linear functional form of RMoM) provides stronger evidence that gaining a partner is associated with lower pay. Our conclusions regarding our question of interest show some sensitivity to our model assumptions.

Table 1: Proportional change in pay associated with gaining a partner

	posterior mean	95% interval
MoI	0.93	(0.87,1.00)
BASE	0.90	(0.82,0.98)
AS1	0.92	(0.87,0.98)
AS2	0.88	(0.79,0.99)
AS3	0.90	(0.82,0.99)
AS4	0.85	(0.78,0.93)

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 δ 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 1 and 2).

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

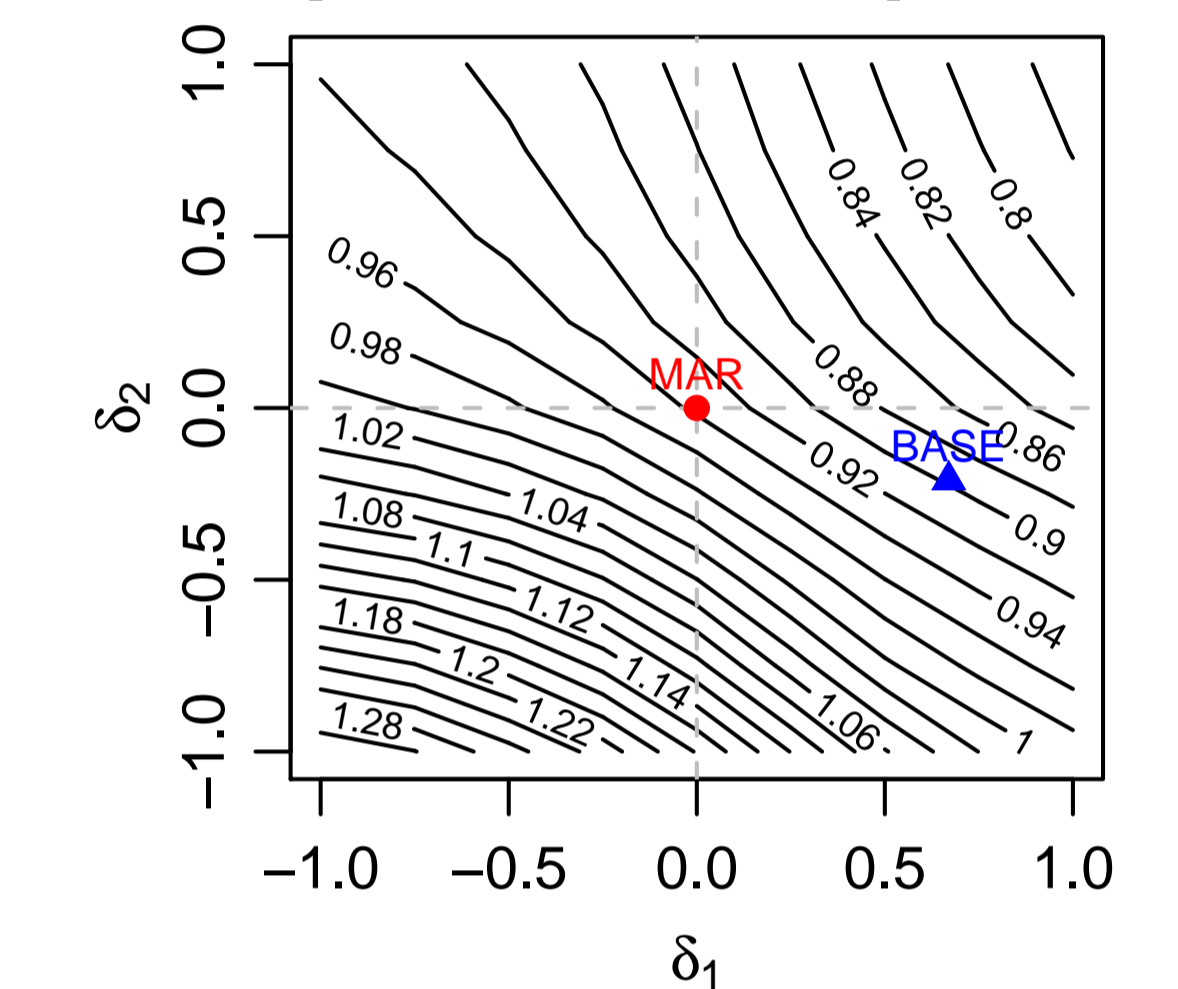
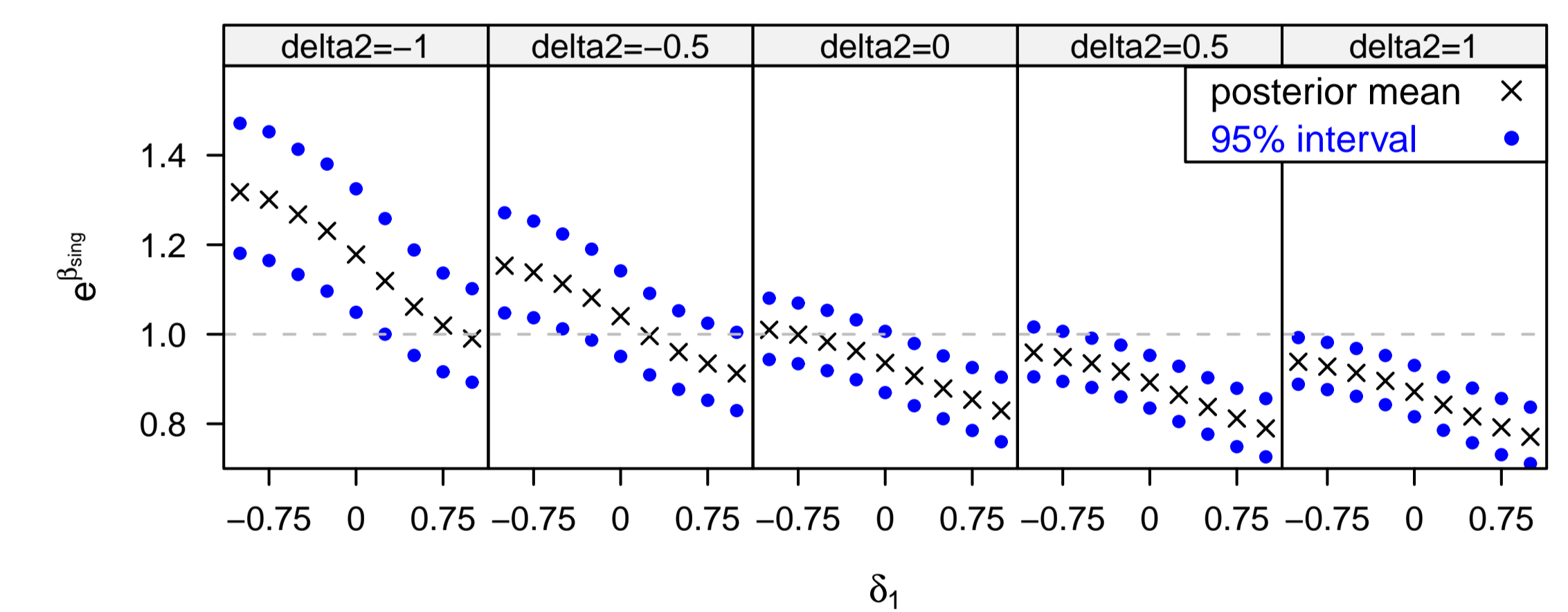


Figure 2: Proportional change in pay associated with gaining a partner between sweeps



6. Assess fit of hold-out sample

Some sweep 2 data was 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 2 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 3 this measure provides greatest support for the models in the upper right quadrant.

CONCLUSION: there is weak evidence that gaining a partner is associated with lower pay.

Table 2: MSE of imputed hourly pay for 7 re-issued individuals

	MSE for re-issues	median	95% interval
BASE	18.7	(3.1,367.0)	
AS1	8.0	(1.9,73.6)	
AS2	16.8	(3.2,108.8)	
AS3	14.2	(2.8,295.3)	
AS4	8.8	(2.9,21.7)	

Figure 3: MSE of imputed hourly pay

