Methodological developments for combining data

Modelling non-random missing data in longitudinal studies: how can information from additional sources help?

Alexina Mason

Department of Epidemiology and Public Health
Imperial College, London

July 2008

with thanks to Nicky Best, Ian Plewis and Sylvia Richardson

This work was supported by an ESRC PhD studentship.

Outline

1. Motivation
   - introduction
   - MCS income example

2. Building a Bayesian joint model which combines data
   - model of interest
   - covariate model of missingness
   - response model of missingness

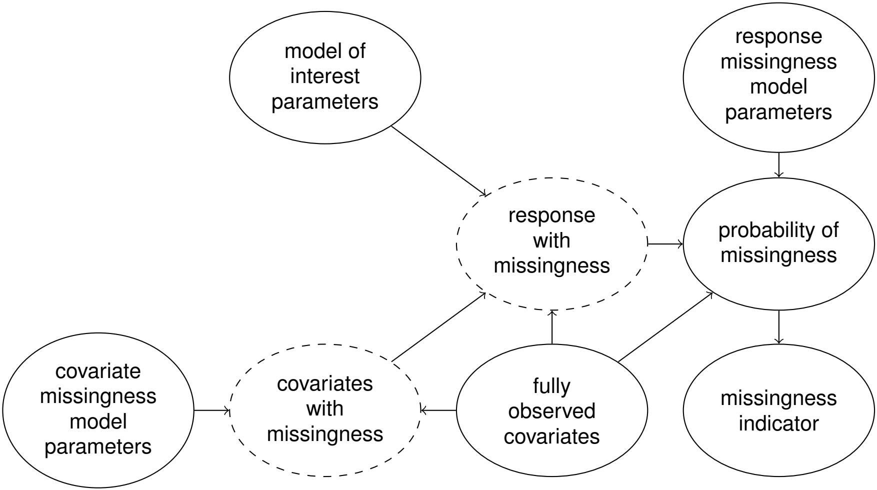
3. Results

Why combine data?

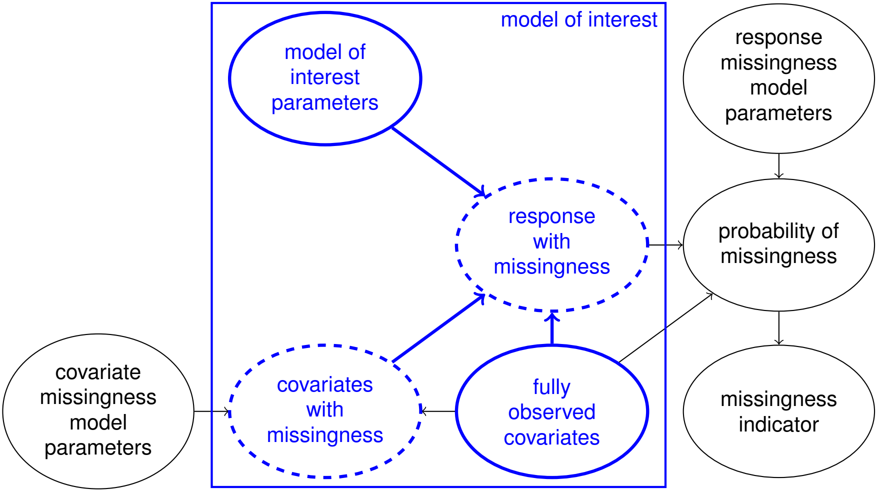
- missing data adds complexity to Bayesian models for analysing longitudinal studies
- typically, they will include a number of sub-models, e.g.
  - model for the question of interest
  - model(s) to impute the missing values
- the estimation of some parameters in the imputation models can be difficult, particularly where information is limited
- but, we can increase the amount of information by incorporating data from other sources, e.g.
  - data from other studies
  - expert opinion

we now look at the general model set-up diagrammatically

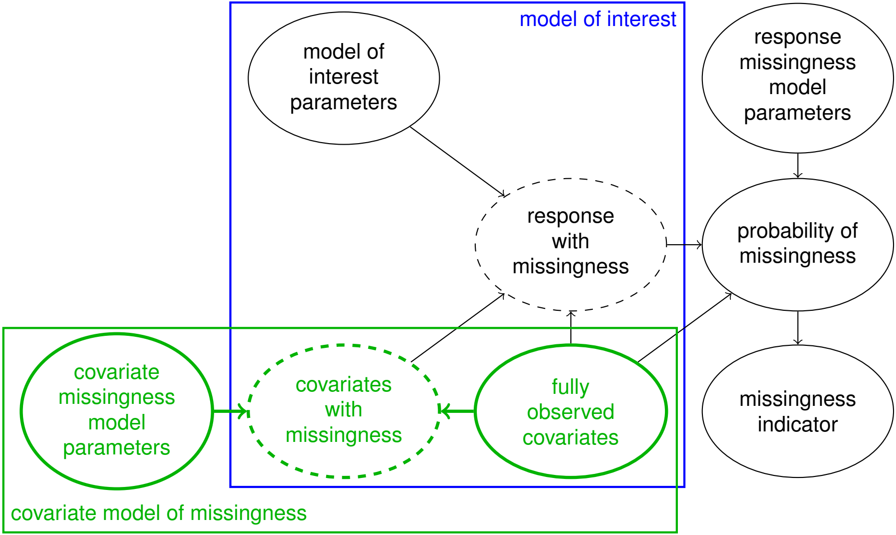
Schematic Diagram


model of interest parameters
response missingness model parameters
response with missingness
probability of missingness
covariate missingness model parameters
covariates with missingness
fully observed covariates
missingness indicator

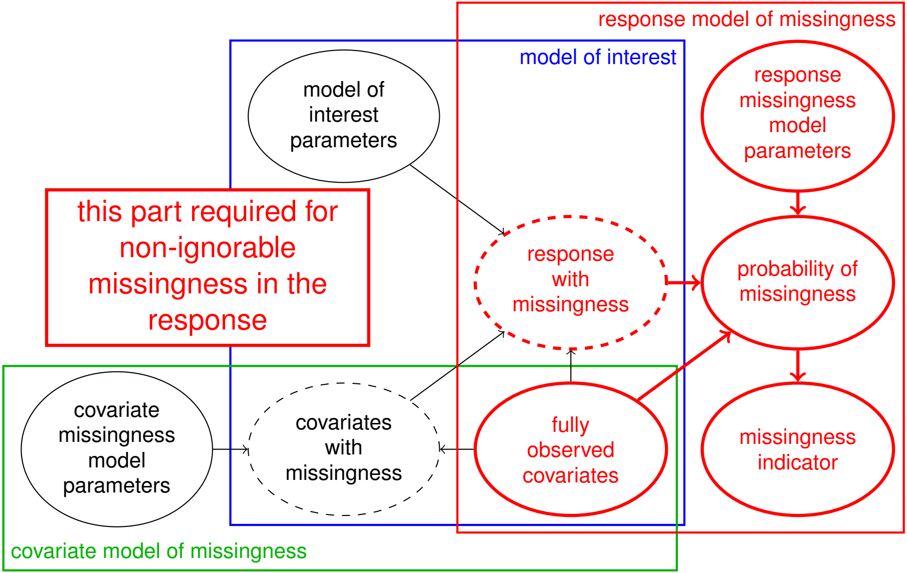
Schematic Diagram


model of interest
model of interest parameters
response with missingness
covariates with missingness
fully observed covariates
response missingness model parameters
probability of missingness
missingness indicator
covariate missingness model parameters

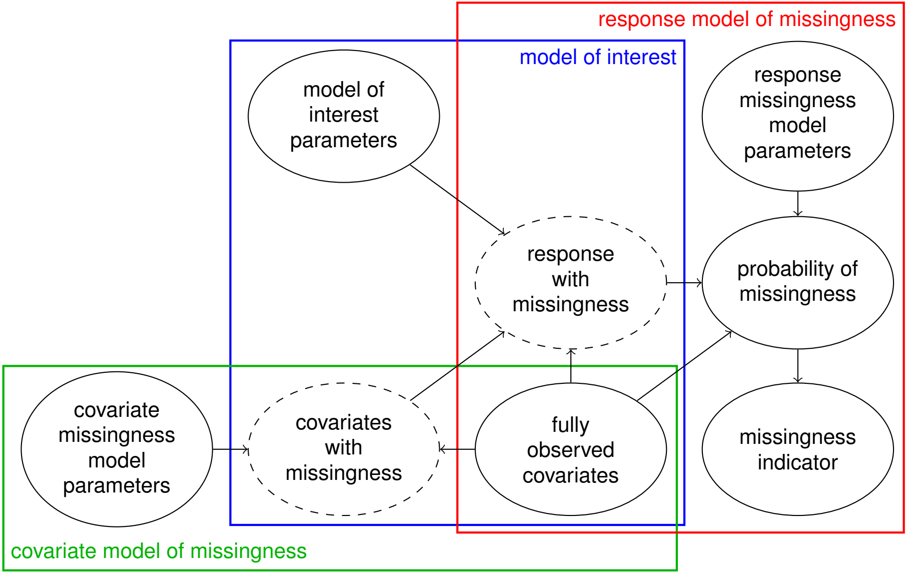
Schematic Diagram


model of interest
model of interest parameters
response missingness model parameters
response with missingness
probability of missingness
missingness indicator
covariate missingness model parameters
covariates with missingness
fully observed covariates
covariate model of missingness

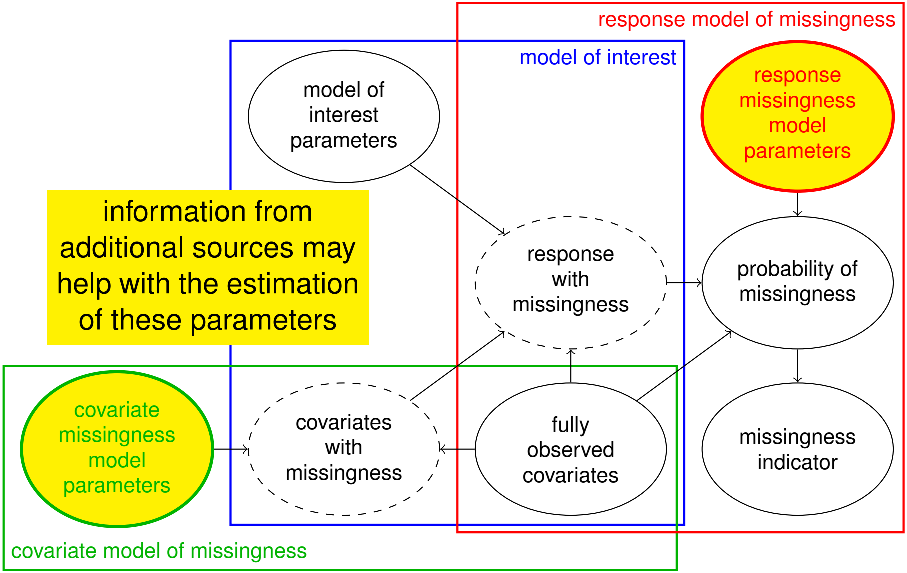
Schematic Diagram


response model of missingness
model of interest
model of interest parameters
response missingness model parameters
this part required for non-ignorable missingness in the response
response with missingness
probability of missingness
covariate missingness model parameters
covariates with missingness
fully observed covariates
missingness indicator
covariate model of missingness

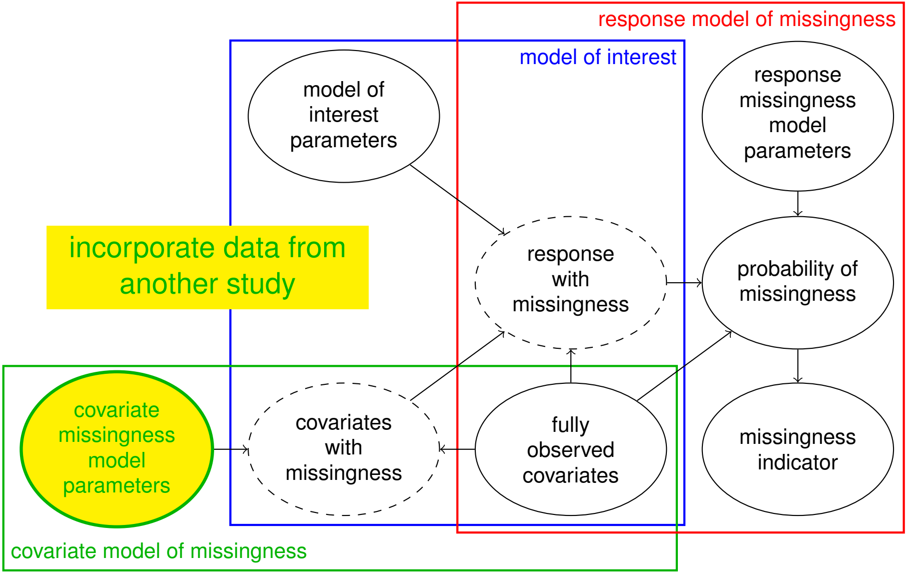
Schematic Diagram


response model of missingness
model of interest
model of interest parameters
response missingness model parameters
response with missingness
probability of missingness
covariate missingness model parameters
covariates with missingness
fully observed covariates
missingness indicator
covariate model of missingness

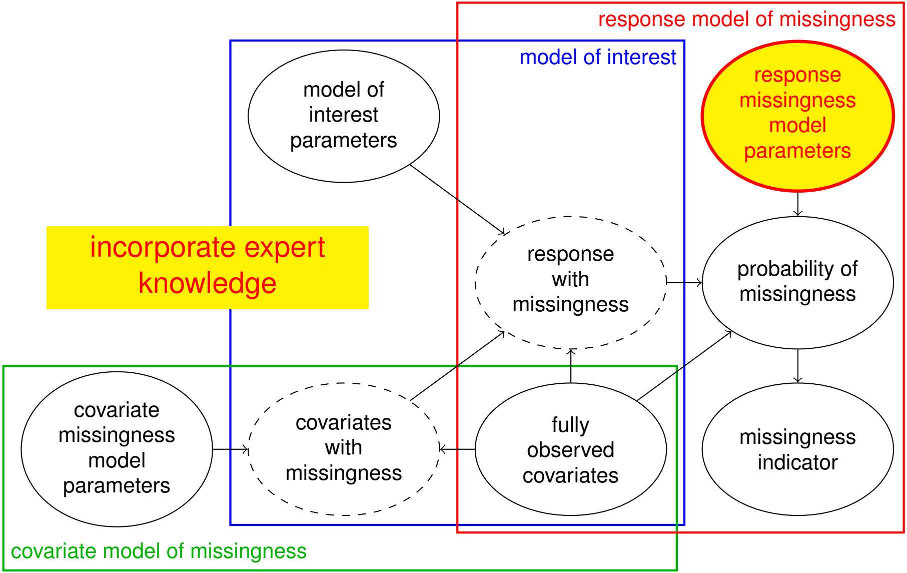
Schematic Diagram


response model of missingness
model of interest
model of interest parameters
response missingness model parameters
information from additional sources may help with the estimation of these parameters
response with missingness
probability of missingness
covariate missingness model parameters
covariates with missingness
fully observed covariates
missingness indicator
covariate model of missingness

Schematic Diagram


response model of missingness
model of interest
model of interest parameters
response missingness model parameters
incorporate data from another study
response with missingness
probability of missingness
covariate missingness model parameters
covariates with missingness
fully observed covariates
missingness indicator
covariate model of missingness

Schematic Diagram


response model of missingness
model of interest
model of interest parameters
response missingness model parameters
incorporate expert knowledge
response with missingness
probability of missingness
covariate missingness model parameters
covariates with missingness
fully observed covariates
missingness indicator
covariate model of missingness

Millennium Cohort Study (MCS) example

- MCS has 18,000+ cohort members born in the UK at the beginning of the Millennium
- using sweeps 1 and 2, our example predicts income for main respondents meeting the criteria:
    - single in sweep 1
    - in work
    - not self-employed
- motivating questions about income include:
    - how much extra do individuals earn if they have a degree?
    - does change in partnership status affect income?
    - does ethnicity affect rate of pay?

Missingness in the MCS income dataset

- initial dataset has 559 records

sweep 1 missingness

		covariates	
		observed	missing
pay	observed	505	7
	missing	43	4

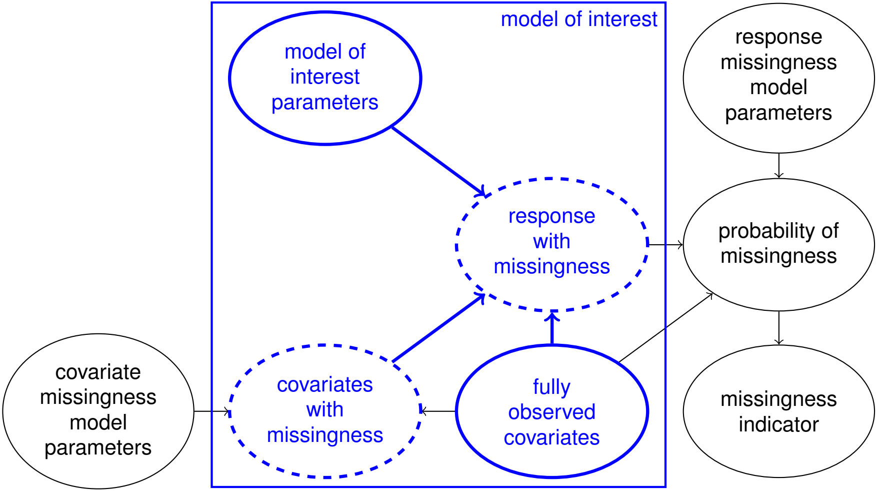
- restrict dataset to individuals fully observed in sweep 1

sweep 2 missingness for remaining 505 individuals

		covariates	
		observed	missing
pay	observed	320	0
	missing	19	166

- don’t distinguish between item and sweep non-response
- all the covariate missingness comes from sweep non-response

Schematic Diagram


model of interest
model of interest parameters
response missingness model parameters
response with missingness
probability of missingness
covariate missingness model parameters
covariates with missingness
fully observed covariates
missingness indicator

Model of interest

- we choose log of hourly net pay as our response
- and 6 explanatory variables

Description of explanatory variables

short name	description	details
age		continuous[a]
edu	educational level	3 levels (1=none/NVQ1; 2=NVQ2/3; 3=NVQ4/5)[b]
eth	ethnic group	2 levels (1=white; 2=non-white)
sing[c]	single/partner	2 levels (1=single; 2=partner)
reg	region of country	2 levels (1=London; 2=other)
stratum	ward type by country[d]	9 levels

[a] centred and standardised

[b] the level of National Vocational Qualification (NVQ) equivalence of the individual's highest academic or vocational educational qualification (level 3 has a degree)

[c] always single in sweep 1

[d] three strata for England (advantaged, disadvantaged and ethnic minority); two strata for Wales, Scotland and Northern Ireland (advantaged and disadvantaged)

Model of Interest: the equations

$$pay_{it} \sim t4(\mu_{it}, \sigma^2)$$

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

$\alpha_i \sim N(0, \varsigma^2)$ individual random effects

$\varsigma \sim N(0, 10000^2)I(0,)$

$\gamma_{s(i)} \sim N(0, 10000^2)$ stratum specific intercepts

$\beta_k \sim N(0, 10000^2)$

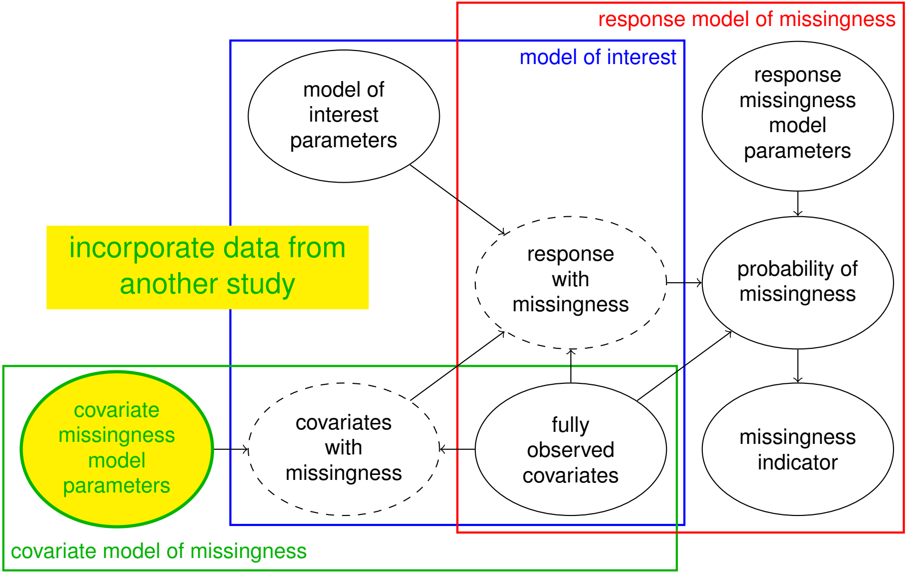
$\frac{1}{\sigma^2} \sim Gamma(0.001, 0.001)$

for t=1,2 sweeps; i=1,. . . ,n individuals;

x={age,edu,$sing$,reg}; z={eth}.

$N(mean, variance)I(0,)$ denotes a half Normal distribution restricted to positive values.

Schematic Diagram


response model of missingness
model of interest
model of interest parameters
response missingness model parameters
incorporate data from another study
response with missingness
probability of missingness
covariate missingness model parameters
covariates with missingness
fully observed covariates
missingness indicator
covariate model of missingness

Covariate model of missingness

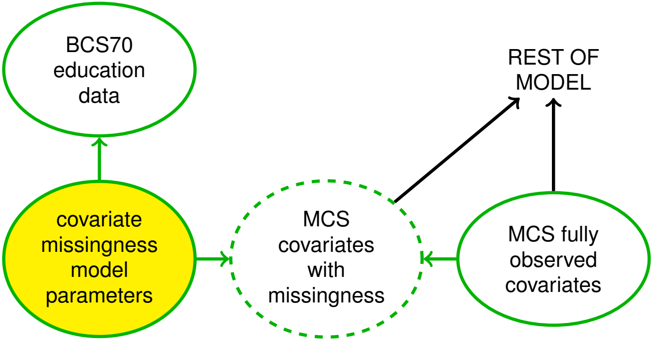
- assume covariates are missing at random (MAR)
- *stratum* and *eth* do not change between sweeps
- imputation of missing values for the other 4 covariates is required
  - *age*: impute age difference between sweeps and add to sweep 1
  - *reg*: assign sweep 1 value
  - *sing*: impute completely randomly to maintain proportion for observed individuals
  - *edu*: impute using a latent variable with fixed cut points 0 and 1, and conditions to prevent education level decreasing
- ignore correlation between covariates for now, but this is investigated as an extension
- imputing *edu* is difficult because few individuals gain qualifications between sweeps - additional data can help here

Additional education data

- additional data taken from a different longitudinal study, the 1970 British Cohort Study (BCS70)
- we assemble education variables at similar time points to MCS using sweeps
  - 5 (1999/2000), cohort members aged 30
  - 6 (2004/2005), cohort members aged 34
- and select individuals with similar characteristics to MCS, i.e.
  - mother
  - single in sweep 5
  - in work
  - not self-employed
- 157 fully observed cohort members meet these criteria

Combining the education data

- we model BCS70 educational level using equations with the same parameters as our equations for imputing *edu*
- so the BCS70 data helps estimate these covariate missingness model parameters


BCS70 education data
REST OF MODEL
covariate missingness model parameters
MCS covariates with missingness
MCS fully observed covariates

edu Model of Missingness: the equations

$mcs.edu_i^\star \sim N(mcs.\nu_i, \Sigma^2)I(mcs.left_i, mcs.right_i)$ latent variable

$mcs.\nu_i = \eta + \kappa_2 mcs.edu_{i1,2} + \kappa_3 mcs.edu_{i1,3} + \phi mcs.age_{i1}$

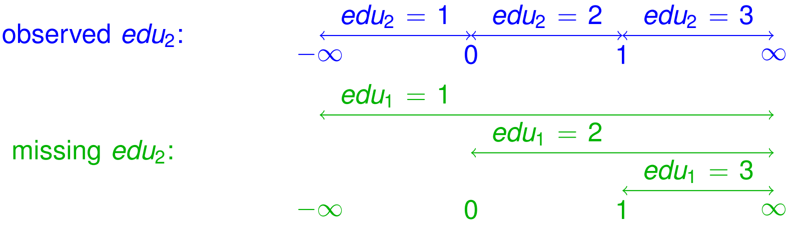
$bcs.edu_j^\star \sim N(bcs.\nu_j, \Sigma^2)I(bcs.left_j, bcs.right_j)$ latent variable

$bcs.\nu_j = \eta + \kappa_2 bcs.edu_{j1,2} + \kappa_3 bcs.edu_{j1,3} + \phi bcs.age_{j1}$

$\eta, \kappa_2, \kappa_3, \phi, \Sigma \sim priors$

$N(mean, variance)I(left, right)$ denotes a restricted Normal distribution.

calculating *left* and *right*



observed edu_2:

$edu_2 = 1$ $edu_2 = 2$ $edu_2 = 3$

$-\infty$ 0 1 ∞

$edu_1 = 1$

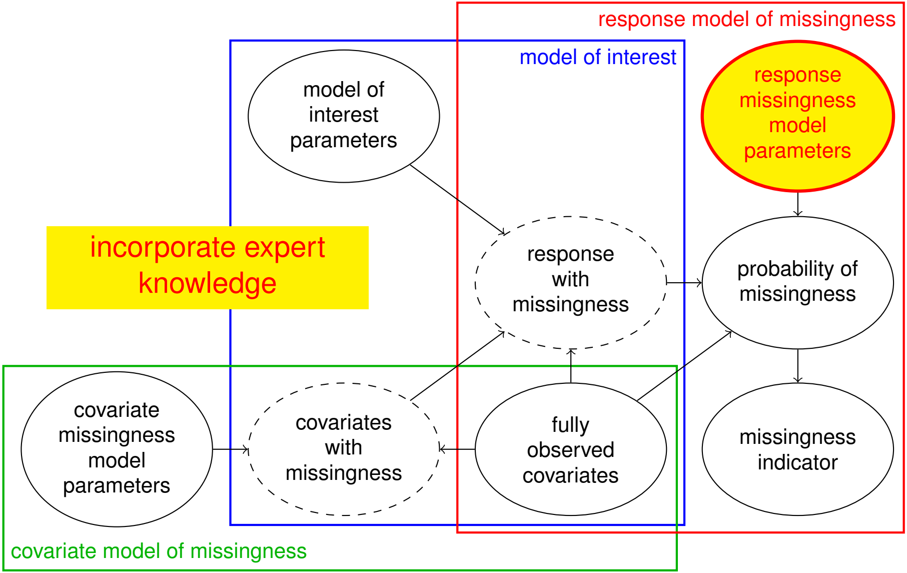
missing edu_2:

$edu_1 = 2$

$edu_1 = 3$

$-\infty$ 0 1 ∞

Schematic Diagram


response model of missingness
model of interest
model of interest parameters
response missingness model parameters
incorporate expert knowledge
response with missingness
probability of missingness
covariate missingness model parameters
covariates with missingness
fully observed covariates
missingness indicator
covariate model of missingness

Response model of missingness (selection model)

We use a logit model for response missingness, i.e.

$$m_i \sim Bernoulli(p_i); \quad logit(p_i) = ?,$$

where m_i is a binary missingness indicator for sweep 2 pay, pay_2

- pay_2 is Missing at Random (MAR) if p_i depends only on observed data, then
  - the logit equation does not include pay_2
  - the response model of missingness and the rest of the model can be estimated separately
- otherwise pay_2 is Missing not at Random (MNAR), then
  - we cannot ignore the model of missingness
  - the two parts of the model must be estimated simultaneously
- we are interested in non-random missing data mechanisms

Response missingness model parameters

- the response missingness model parameters are known to be difficult to estimate
    - there is limited information in a binary indicator of missingness
    - often resulting in a flat likelihood
- we wish to incorporate expert knowledge to help with their estimation
- so, we recruited an expert with
    - general knowledge about missing data in longitudinal studies
    - specific knowledge about missing MCS family income

Eliciting informative priors on parameters

- we want informative priors for the response missingness model parameters
- but, these are difficult to elicit directly
- instead
  1. we elicit information about the probability of response at design points
  2. convert this to informative priors
- we use Mary Kynn's graphical elicitation package, ELICITOR (silmaril.math.sci.qut.edu.au/$\sim$whateley/)

About ELICITOR

- ELICITOR was created to elicit normal prior distributions for Bayesian logistic regression models in ecology
- The process of elicitation can be summarised as follows:
  1. determine the variables to explain the income missingness
  2. determine the category/level that maximises the response probability for each variable
  3. choose design points for any continuous variables
  4. elicit median response probabilities and intervals
  5. provide feedback and revisit elicited values as required
  6. convert this information into informative priors

We now consider each step in more detail.

Elicitation 1: determining explanatory variables

1. Following discussion with our expert, five variables were chosen

Explanatory variables for income missingness

name	description
level	level of hourly pay (sweep 1)
change	change in hourly pay (sweep 2 - sweep 1)
sc	social class (sweep 1)
eth	ethnicity
ctry	country

- $m_i \sim Bernoulli(p_i); \quad logit(p_i) = \theta + \sum_{k=1}^{p} \delta_k x_{ki}$
- we wish to place informative priors on θ and $\boldsymbol{\delta}$
- to illustrate the remaining steps we focus on *change*

Elicitation 2/3: optimum values and design points

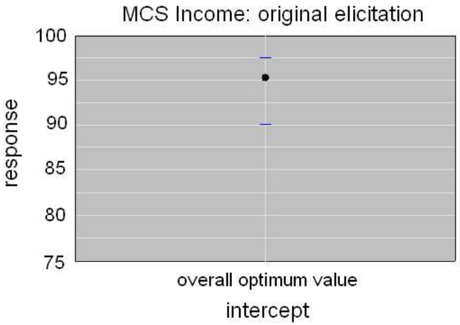
2.
- which value of *change* maximises the response probability?
- our expert decided on £0
- £0 is the optimum value for *change*

3.
- which other values of *change* should be used for elicitation of response probability?
- our expert chose -£5 and £5
- -£5, £0 and £5 are the design points for *change*

Elicitation 4: overall optimum value

4. the overall optimum value occurs when all 5 covariates are set to their optimum values

- first, elicit median: if there were a 100 individuals with all covariates at optimum value, how many would you expect to respond to the income question?
- then, elicit interval: lower and upper quartiles

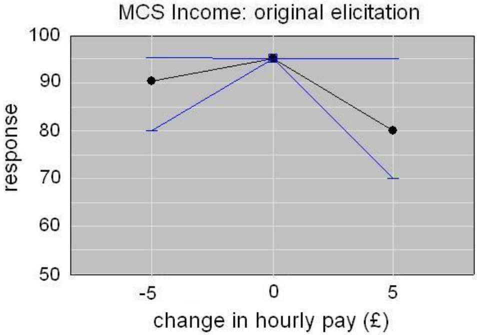


Elicitation 4: design points

4. each explanatory variable considered in turn
   - optimum value probabilities already elicited
   - remaining design points elicited assuming all other variables are at optimum level

Example: change variable

- determine suitable functional form
- piecewise linear selected
- each variable is assumed independent, so covariances are not elicited



Elicitation 5: feedback

5. providing feedback allows the expert to reconsider their assessments
   - ELICITOR enables feedback during the elicitation and any variable can be revisited
   - our expert wished to see the implied median response probability when all the variables are set to their minimum design points, the worst case, and believed this would be $\approx 60\%$
   - running the model produced a worst case median of 1%
   - our expert revisited his original elicited values
   - these changes resulted in a worst case response of 9%

- our worst case is very extreme
- the rate of response rapidly decreases as probabilities are multiplied

giving good intuition about probabilities that are combined is difficult

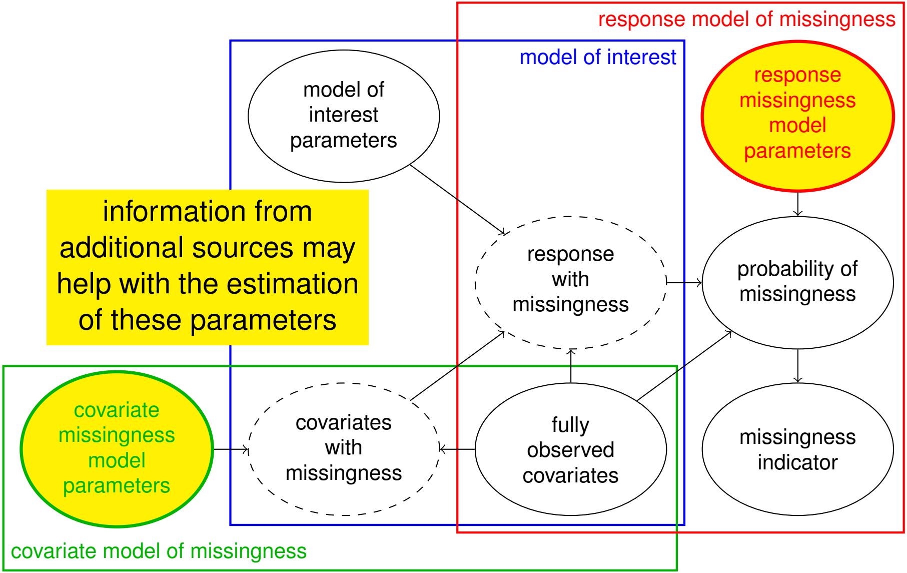
Elicitation 6: conversion

6. ELICITOR converted the elicited means and intervals into a Bayesian model with informative priors

For illustration: the original elicitation with the single explanatory variable, *change*, generates

$$\begin{aligned} logit(p_i) &= \theta + Piecewise(change_i) \\ Piecewise(change_i) &= \begin{cases} \delta_1 change_i : & change_i < 0 \\ \delta_2 change_i : & change_i > 0 \end{cases} \\ \theta &\sim N(3, 1.3) \\ \delta_1 &\sim N(0.15, 0.23) \\ \delta_2 &\sim N(-0.32, 0.32) \end{aligned}$$

Schematic Diagram


response model of missingness
model of interest
model of interest parameters
response missingness model parameters
information from additional sources may help with the estimation of these parameters
response with missingness
probability of missingness
covariate missingness model parameters
covariates with missingness
fully observed covariates
missingness indicator
covariate model of missingness

Model Summary

Summary of Joint Models

Model label	BCS70 data included?	Informative priors[a]	Functional form[b]
A			Linear
B			Piecewise Linear
C	✓		Piecewise Linear
D		✓	Piecewise Linear
E	✓	✓	Piecewise Linear

[a] on the parameters of the response model of missingness

[b] of *level* and *change* in the response model of missingness

Convergence problems were encountered for model B

Missing *edu* imputations

Posterior mean percentages using MCS data only (Model D)

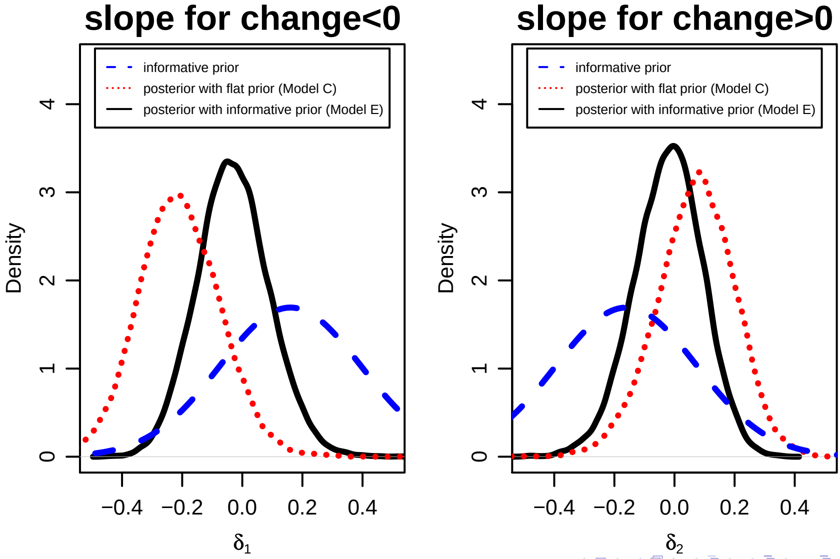
	edu2=1	edu2=2	edu2=3	total
edu1=1	83	17	0	100
edu1=2		96	4	100
edu1=3			100	100

Posterior mean percentages using MCS and BCS70 data (Model E)

	edu2=1	edu2=2	edu2=3	total
edu1=1	82	18	0	100
edu1=2		95	5	100
edu1=3			100	100

Using the BCS70 data results in a slight increase in the percentage of individuals imputed to increase their level of education

Prior and posterior distributions of *change* parameter



Estimates of model of interest parameters (95% interval)

	linear selection model	piecewise selection model		
	A	C	D	E
edu[2]	1.17 (1.08,1.27)	1.17 (1.08,1.28)	1.18 (1.08,1.28)	1.18 (1.08,1.28)
edu[3]	1.37 (1.24,1.51)	1.35 (1.23,1.50)	1.35 (1.23,1.49)	1.36 (1.23,1.50)
eth	0.95 (0.83,1.08)	0.94 (0.83,1.07)	0.94 (0.83,1.07)	0.94 (0.82,1.07)
sing	0.88 (0.81,0.95)	0.93 (0.86,1.00)	0.93 (0.87,1.00)	0.93 (0.87,1.01)

parameters are e^{β}, representing the proportional increase in pay associated with each covariate

- the functional form of the selection model affects *sing*, but otherwise these parameter estimates are similar for all models
- higher education levels are associated with higher pay
- being non-white or gaining a partner between sweeps is associated with lower pay

Comparison with complete case analysis

	model of interest only complete case analysis (CC)	linear selection model A	piecewise selection model E
edu[2]	1.17 (1.07,1.28)	1.17 (1.08,1.27)	1.18 (1.08,1.28)
edu[3]	1.41 (1.27,1.57)	1.37 (1.24,1.51)	1.36 (1.23,1.50)
eth	0.96 (0.84,1.11)	0.95 (0.83,1.08)	0.94 (0.82,1.07)
sing	0.93 (0.87,1.00)	0.88 (0.81,0.95)	0.93 (0.87,1.01)

parameters are e^{β}, representing the proportional increase in pay associated with each covariate

- our model of interest can be run separately if we restrict our dataset to fully observed individuals - a complete case analysis
- the CC edu[3] estimate is slightly higher than for both selection models, but the other parameter estimates are similar to Model E
- the extra information in the joint models has narrowed the 95% intervals compared with CC, except for *sing*

Summary

Modelling non-random missing data in longitudinal studies: how can information from additional sources help?

- by informing weakly or non-identifiable parts of the model
- by allowing more realistic models to be fitted
- by improving the imputations
- by compensating for difficulties in separating different sources of uncertainty, e.g. assumptions about the distributional form and the missing data process

sensitivity analysis is crucial

Relevant literature

- The BIAS project.
  `www.bias-project.org.uk/.`
- Best, N. G., Spiegelhalter, D. J., Thomas, A., and Brayne, C. E. G. (1996).
  Bayesian Analysis of Realistically Complex Models.
  Journal of the Royal Statistical Society, Series A (Statistics in Society), **159**, (2), 323–42.
- Little, R. J. A. and Rubin, D. B. (2002).
  Statistical Analysis with Missing Data, (2nd edn). John Wiley and Sons.
- 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).
  Non-Response in a Birth Cohort Study: The Case of the Millennium Cohort Study.
  International Journal of Social Research Methodology, **10**, (5), 325–34.
- White, I. R., Carpenter, J., Evans, S., and Schroter, S. (2004).
  Eliciting and using expert opinions about dropout bias in randomised controlled trials.
  Technical report, London School of Hygiene and Tropical Medicine.