

Eliciting expert opinion about missing data in longitudinal studies

Alexina Mason

with thanks to Nicky Best, Sylvia Richardson and Ian Plewis

1 July 2011

Outline

Introduction

- Missing data

- Bayesian analysis

Principles of Elicitation (illustration using simplistic assumptions)

- Eliciting information from experts

- Results and implications

Incorporating Realistic Assumptions

- Allow dependence between explanatory variables

- Allow non-linear relationships

Introduction

- Missing data are common in longitudinal studies
- Potentially distort results of scientific investigation
- Important to think carefully about the causes of missingness
- Require method allowing incorporation of realistic assumptions
- Experts may be able to provide valuable information about the reasons for missingness
- To utilise expert knowledge requires
 - a method of eliciting expert opinion
 - a way of incorporating this information into a statistical model

Illustrative example: income data from MCS

- The Millennium Cohort Study (MCS) has 18,000+ cohort members born in the UK
- Using sweeps 1 and 2, our example predicts income for main respondents (mothers) meeting the criteria:
 - single in sweep 1
 - in work
 - not self-employed
- Motivating question: does change in partnership status affect income?

Missingness in the MCS income dataset

- Initial dataset has 559 records

sweep 1 missingness

		N	%
pay	observed	505	90%
	missing	54	10%

Missingness in the MCS income dataset

- Initial dataset has 559 records

sweep 1 missingness

		N	%
pay	observed	505	90%
	missing	54	10%

- Restrict dataset to individuals fully observed in sweep 1
sweep 2 missingness for remaining 505 individuals

		N	%
pay	observed	320	63%
	missing	185	37%

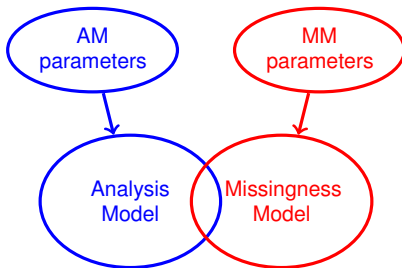
Complete Case (CC) analysis

- If we had complete data, we could just fit an appropriate regression model
- However, the missing data complicates this analysis
- One approach is to
 - discard individuals with incomplete information
 - analyse complete cases only
- Advantage: simple
- Disadvantages:
 - loss of precision, as not using all available information
 - often introduces bias (see next slide)
- CC suggests that if an individual earning £8 an hour gains a partner, their hourly pay will **decrease** by £0.62 (£1.14, £0.09)

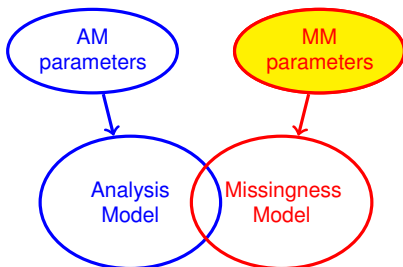
Assumptions about missing income

- Does CC give biased answers?
 - In general, depends on the reasons for the missing values
 - For the example, if income is more likely to be missing for lower earners, then CC will introduce bias
 - Previous research suggests income level is associated with willingness to respond
 - Cannot be verified from the data at hand, so this is an **assumption**
- Require a more 'statistically principled' method which allows realistic assumptions about the missing data
- One such method involves fitting a joint model, including
 - an analysis model (e.g. a regression model)
 - a model for the missingness (typically a logistic regression model)

Joint model (selection model)

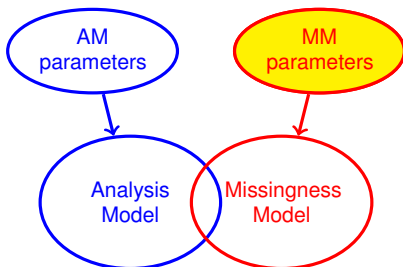


Joint model (selection model)



- The parameters in the missingness model are hard to estimate
- Experts may be able to help
- We can utilise expert information about missingness if we are prepared to be Bayesian

Joint model (selection model)



- The parameters in the missingness model are hard to estimate
- Experts may be able to help
- We can utilise expert information about missingness if we are prepared to be Bayesian

We now introduce Bayesian analysis

Bayesian inference

- Bayesian inference distinguishes between
 - observable quantities, i.e. observed data
 - unobserved quantities (e.g. statistical parameters, missing data)
- Unobserved quantities are viewed as unknown with an associated probability distribution
- So Bayesian methods are a very natural way of handling missing data
 - a probability distribution is estimated for each missing value
 - allows uncertainty to be adequately captured

3 components of Bayesian inference

- The prior distribution
 - reflects the plausibility of different values of the unknowns before the data is seen
- The likelihood
 - expresses support for different values of the unknowns based solely on the data
 - also used in classical inference
- The posterior distribution
 - combines information in the prior distribution with the likelihood using Bayes theorem
 - is the basis of all Bayesian inference
 - expresses uncertainty about the unknowns after seeing the data

Priors

- The prior can be used to incorporate additional information/knowledge into Bayesian models
- Additional information can come from
 - other studies
 - experts - elicitation
- If extra information exists, it is helpful to use it
- If not, 'vague' priors can be used to reflect no knowledge
- Specifying a good prior is non-trivial, but invaluable
- Sensitivity to different priors should always be explored

Priors

- The prior can be used to incorporate additional information/knowledge into Bayesian models
- Additional information can come from
 - other studies
 - experts - elicitation
- If extra information exists, it is helpful to use it
- If not, 'vague' priors can be used to reflect no knowledge
- Specifying a good prior is non-trivial, but invaluable
- Sensitivity to different priors should always be explored

We now look at how to incorporate expert knowledge about missing data through informative priors

Outline

Introduction

Missing data

Bayesian analysis

Principles of Elicitation (illustration using simplistic assumptions)

Eliciting information from experts

Results and implications

Incorporating Realistic Assumptions

Allow dependence between explanatory variables

Allow non-linear relationships

Eliciting informative priors on parameters

- Objective: to create informative priors for the parameters of the model of missingness
- Difficult to elicit information about parameters in statistical models directly
- In general, considered good practice to ask for judgements about observable quantities rather than parameters
- So, better to
 1. elicit information about the probability of response
 2. convert this information to informative priors

Elicitation Overview

The process of elicitation can be summarised as follows:

1. determine the variables that explain the income missingness and their functional form
2. determine the reference category/level for each variable
3. choose design points for any continuous variables
4. elicit median response probabilities and intervals
5. convert this information into informative priors
6. provide feedback and revisit elicited values as required

Elicitation Overview

The process of elicitation can be summarised as follows:

1. determine the variables that explain the income missingness and their functional form
2. determine the reference category/level for each variable
3. choose design points for any continuous variables
4. elicit median response probabilities and intervals
5. convert this information into informative priors
6. provide feedback and revisit elicited values as required

We now consider each step in more detail

Step 1: choice of explanatory variables

- Agree explanatory variables with expert
 - level of hourly pay in sweep 1 (*level*)
 - change in hourly pay between sweeps (*change*)

Step 1: choice of explanatory variables

- Agree explanatory variables with expert
 - level of hourly pay in sweep 1 (*level*)
 - change in hourly pay between sweeps (*change*)
- Agree form of relationship with expert
 - linear (overly simplistic, but better for illustration)

Step 1: choice of explanatory variables

- Agree explanatory variables with expert
 - level of hourly pay in sweep 1 (*level*)
 - change in hourly pay between sweeps (*change*)
- Agree form of relationship with expert
 - linear (overly simplistic, but better for illustration)
- Missingness model will take the form

$$r_i \sim \text{Bernoulli}(p_i),$$

$$\text{logit}(p_i) = \alpha + \theta \text{level}_i + \delta \text{change}_i$$

- r_i is a binary indicator of response
- p_i is the probability that individual i responds
- θ = log odds ratio of response per £1 increase/decrease in *level*
- δ = log odds ratio of response per £1 increase/decrease in *change*

Steps 2 and 3: reference and design points

- The reference point is defined to be the point where $\text{logit}(p_i) = \alpha$
- This occurs when
 - *level* = £8 (mean value) - variable is centered
 - *change* = £0

Steps 2 and 3: reference and design points

- The reference point is defined to be the point where $\text{logit}(p_i) = \alpha$
- This occurs when
 - *level* = £8 (mean value) - variable is centered
 - *change* = £0
- The design points specify other values of the variables at which the response probability will be elicited
- Two design points chosen for each variable
 - *level*: £4 & £16
 - *change*: -£5 & £5

Step 4: elicit point estimate for α

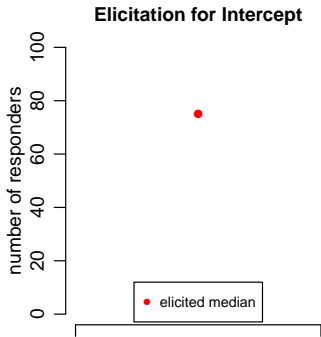
$$\text{logit}(p_i) = \alpha + \theta \text{level}_i + \delta \text{change}_i$$

- Suppose there are 100 individuals with
 - mean hourly pay at sweep 1 (£8)
 - no change in pay between sweeps
- Then, $\text{logit}(p_i) = \alpha$

Step 4: elicit point estimate for α

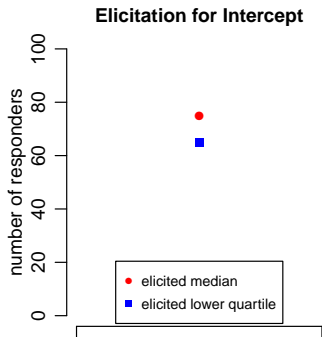
$$\text{logit}(p_i) = \alpha + \theta \text{level}_i + \delta \text{change}_i$$

- Suppose there are 100 individuals with
 - mean hourly pay at sweep 1 (£8)
 - no change in pay between sweeps
- Then, $\text{logit}(p_i) = \alpha$
- Elicit median probability of response
- Q: How many individuals would you expect to respond to the income question?
- A: 75
- $\alpha = \text{logit}(75) = 1.1$



Step 4: elicit uncertainty for α

- Elicit uncertainty interval
- Lower Quartile:
Suppose the true number responding is **less** than 75
Choose a value such that the true number responding is equally likely to be above or below this value
- A: **65**

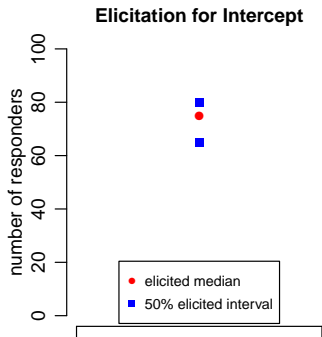


$$sd1 = \frac{\text{logit}(65) - \text{logit}(75)}{\Phi^{-1}(0.25)} = 0.71$$

where $\Phi^{-1}(0.25)$ is the 0.25 quantile of a standard normal distribution

Step 4: elicit uncertainty for α

- Elicit uncertainty interval
- Upper Quartile:
Suppose the true number responding is **more** than 75
Choose a value such that the true number responding is equally likely to be above or below this value
- A: **80**

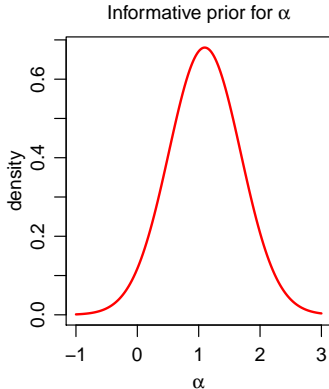
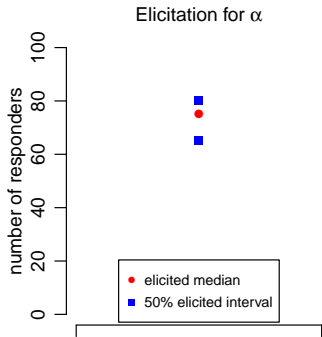


$$sd1 = 0.71; \quad sd2 = \frac{\text{logit}(80) - \text{logit}(75)}{\Phi^{-1}(0.75)} = 0.43$$

average sd1 and sd2 to get estimate of sd for α

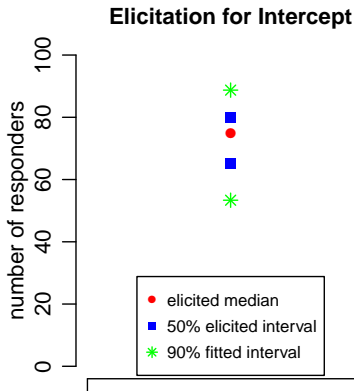
Step 5: convert to informative prior for α

- Use elicited information to form expert prior for α
- Assume prior has a normal distribution
 - mean = elicited median = 1.1
 - sd estimated from elicited uncertainty interval = 0.57



Step 6: intercept feedback

- An important part of the elicitation process is providing feedback
- For example, calculate alternative intervals from expert prior

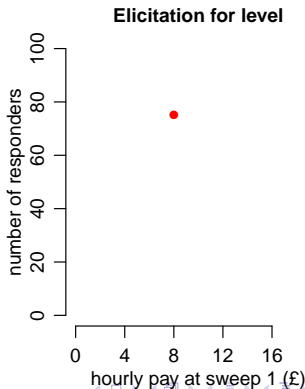


Expert should revisit elicited values if unhappy with feedback

Step 4: elicit point estimate for θ

$$\text{logit}(p_i) = \alpha + \theta \text{level}_i + \delta \text{change}_i$$

- Suppose there are 100 individuals with
 - no change in pay between sweeps
- Then, $\text{logit}(p_i) = \alpha + \theta \text{level}_i$
- We have already elicited the probability of response when $\text{level} = \text{£}8$

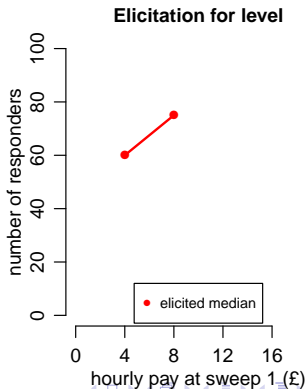


Step 4: elicit point estimate for θ

$$\text{logit}(p_i) = \alpha + \theta \text{level}_i + \delta \text{change}_i$$

- Suppose there are 100 individuals with
 - no change in pay between sweeps
- Then, $\text{logit}(p_i) = \alpha + \theta \text{level}_i$
- We have already elicited the probability of response when $\text{level} = \text{£}8$
- Elicit median probability of response when $\text{level} = \text{£}4$

$$\theta = \frac{\text{logit}(60) - \text{logit}(75)}{-4}$$



Step 4: elicit point estimate for θ

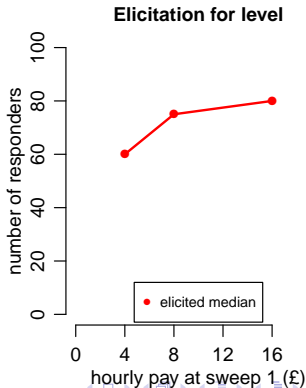
$$\text{logit}(p_i) = \alpha + \theta \text{level}_i + \delta \text{change}_i$$

- Suppose there are 100 individuals with
 - no change in pay between sweeps
- Then, $\text{logit}(p_i) = \alpha + \theta \text{level}_i$

- We have already elicited the probability of response when $\text{level} = \text{£}8$
- Elicit median probability of response when $\text{level} = \text{£}4$

$$\theta = \frac{\text{logit}(60) - \text{logit}(75)}{-4}$$

- Elicit median probability of response when $\text{level} = \text{£}16$



Step 4: elicit point estimate for θ

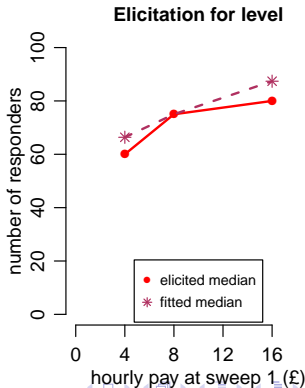
$$\text{logit}(p_i) = \alpha + \theta \text{level}_i + \delta \text{change}_i$$

- Suppose there are 100 individuals with
 - no change in pay between sweeps
- Then, $\text{logit}(p_i) = \alpha + \theta \text{level}_i$

- We have already elicited the probability of response when $\text{level} = \text{£}8$
- Elicit median probability of response when $\text{level} = \text{£}4$

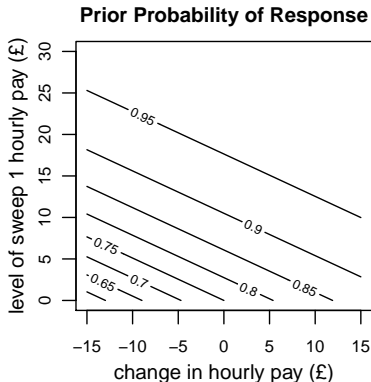
$$\theta = \frac{\text{logit}(60) - \text{logit}(75)}{-4}$$

- Elicit median probability of response when $\text{level} = \text{£}16$

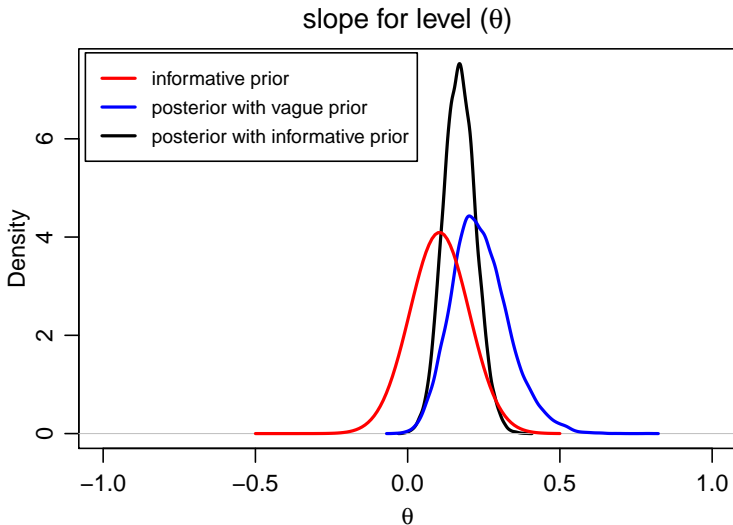


Step 6: providing feedback

- The provision of feedback improves elicitation
- Some feedback already provided as alternative intervals
- Also show implications for combinations of *level* and *change*



Impact of informative prior



Illustrative example: results

Table: posterior mean (95% credible interval) for the change in hourly pay of an individual earning £8 who gains a partner

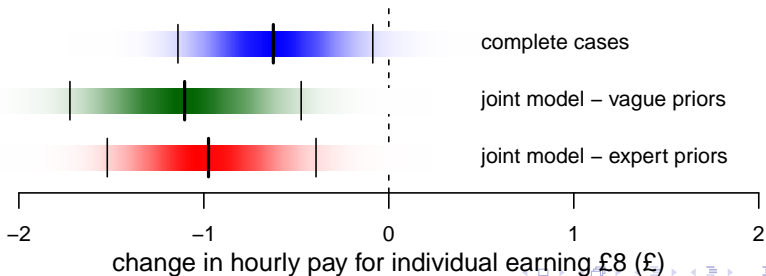
	change in pay (£)	
complete cases	-0.62	(-1.14,-0.09)
joint model - vague priors	-1.10	(-1.72,-0.47)
joint model - expert priors	-0.97	(-1.52,-0.39)

Illustrative example: results

Table: posterior mean (95% credible interval) for the change in hourly pay of an individual earning £8 who gains a partner

	change in pay (£)	
complete cases	-0.62	(-1.14, -0.09)
joint model - vague priors	-1.10	(-1.72, -0.47)
joint model - expert priors	-0.97	(-1.52, -0.39)

Impact on hourly pay of gaining a partner between sweeps



Illustrative example: assumption review

The assumptions used are oversimplistic

1. effect of *change* on non-response is unlikely to be independent of the effect of *level*
 - e.g. effect of a £1 *change* may be larger at low levels of pay
2. linear relationships may be unrealistic
 - e.g. individuals may be less likely to respond if they have high **or** low levels of pay

Outline

Introduction

Missing data

Bayesian analysis

Principles of Elicitation (illustration using simplistic assumptions)

Eliciting information from experts

Results and implications

Incorporating Realistic Assumptions

Allow dependence between explanatory variables

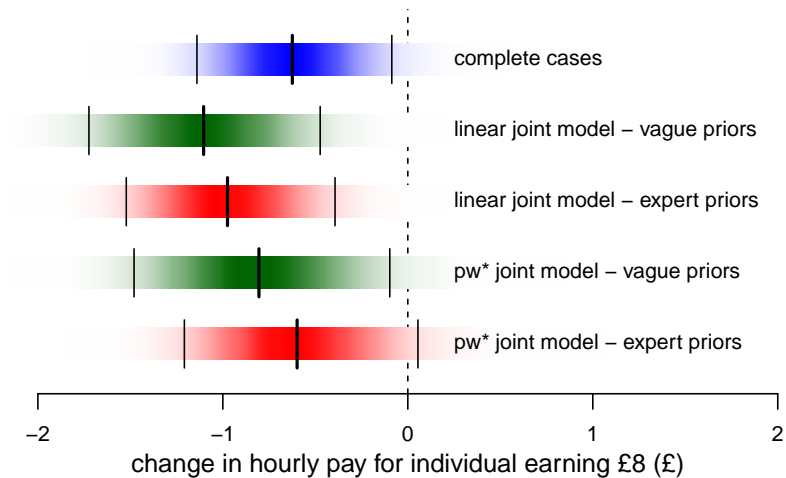
Allow non-linear relationships

Dependence between explanatory variables

- In the illustrative example, the effects of *level* and *change* were assumed to be independent
- If this is unrealistic, possible solutions are:
 1. Transform *level* and/or *change* to make their effects less dependent
 - e.g. we could use *%change* instead of *change*
 2. Elicit a joint prior for *level* and *change* (multivariate normal)
 - elicit *level* as before
 - elicit *change* at both *level* design points (£4 and £16)
 - use additional information to calculate covariance parameter

Result comparison

Impact on hourly pay of gaining a partner between sweeps



Summary

- Bayesian joint models provide a coherent and flexible way of incorporating realistic assumptions about a missingness mechanism
- Subject experts can provide valuable information to help
 - determine plausible assumptions about the missingness
 - identify parameters in the missingness model
- There is a whole wealth of experience about survey methodology, and these techniques provide a way of incorporating it into statistical models

Summary

- Bayesian joint models provide a coherent and flexible way of incorporating realistic assumptions about a missingness mechanism
- Subject experts can provide valuable information to help
 - determine plausible assumptions about the missingness
 - identify parameters in the missingness model
- There is a whole wealth of experience about survey methodology, and these techniques provide a way of incorporating it into statistical models

Sensitivity analysis is crucial

Acknowledgements and References

- The Centre for Longitudinal Studies, Institute of Education for the use of MCS data and the UK Data Archive and Economic and Social Data Service for making the data available
- ESRC for funding
- The BIAS project (PI N Best), based at Imperial College, London, is a node of the Economic and Social Research Council's National Centre for Research Methods (NCRM)
- For papers and technical reports, see our web site
www.bias-project.org.uk

▶ Daniels, M. J. and Hogan, J. W. (2008).

Missing Data In Longitudinal Studies Strategies for Bayesian Modeling and Sensitivity Analysis. Chapman & Hall.

▶ 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.

▶ 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.