

# Bayesian Approaches to Handling Missing Data

Nicky Best and Alexina Mason

BIAS Short Course, Jan 30, 2012

# Lecture 1.

## Introduction to Missing Data

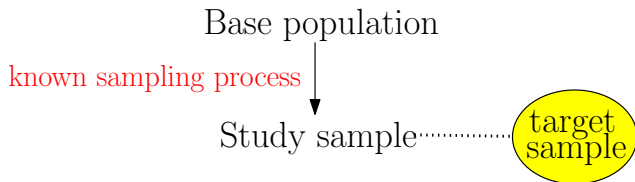
# Lecture Outline

- Introduction
- Motivating example
- Types of missing data
- Ad hoc methods
- ‘Statistically principled’ methods
- Bayesian concepts and MCMC
- Advantages of Bayesian models for missing data

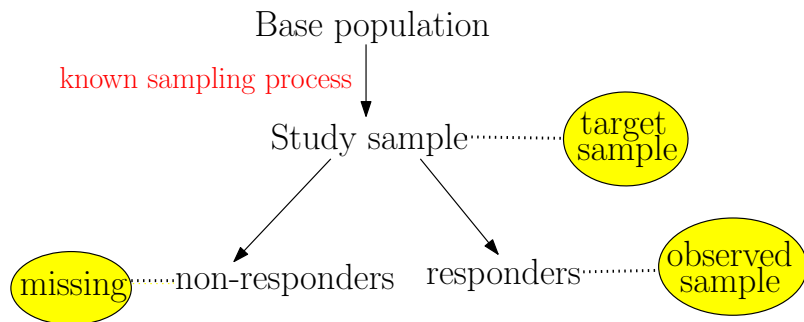
# Introduction

- Missing data are common!
- Usually inadequately handled in both observational and experimental research
- For example, Wood et al. (2004) reviewed 71 recently published BMJ, JAMA, Lancet and NEJM papers
  - ▶ 89% had partly missing outcome data
  - ▶ In 37 trials with repeated outcome measures, 46% performed complete case analysis
  - ▶ Only 21% reported sensitivity analysis
- Sterne et al. (2009) reviewed articles using Multiple Imputation in BMJ, JAMA, Lancet and NEJM from 2002 to 2007
  - ▶ 59 articles found, with use doubling over 6 year period
  - ▶ However, the reporting was almost always inadequate

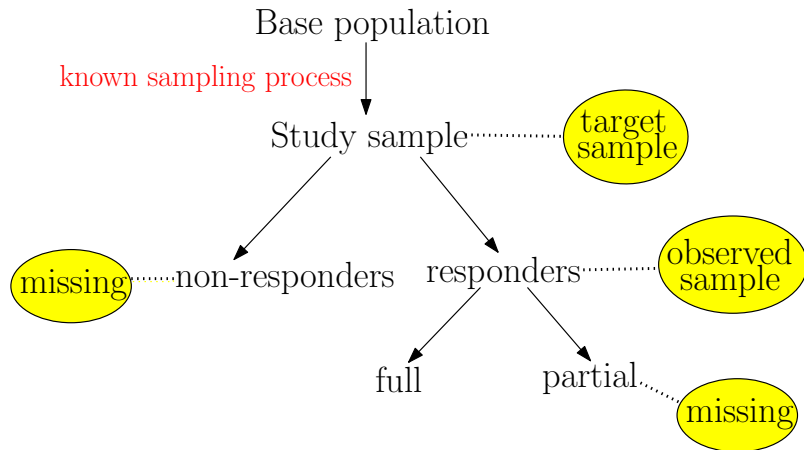
# How does missing data arise?



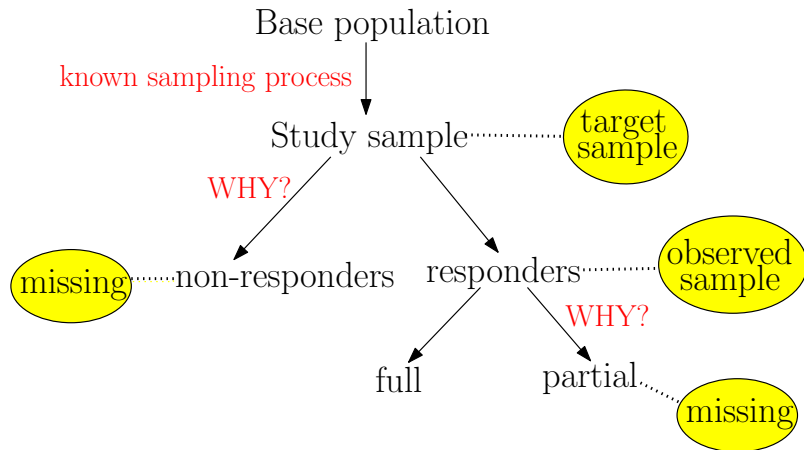
# How does missing data arise?



# How does missing data arise?



# How does missing data arise?



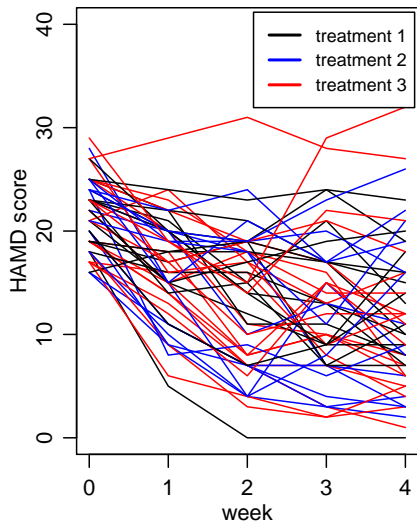
# Motivating example: antidepressant clinical trial

- 6 centre clinical trial, comparing 3 treatments of depression
- 367 subjects randomised to one of 3 treatments
- Subjects rated on Hamilton depression score (HAMD) on 5 weekly visits
  - ▶ week 0 before treatment
  - ▶ weeks 1-4 during treatment
- HAMD score takes values 0-50
  - ▶ the higher the score, the more severe the depression
- **Subjects drop-out from week 2 onwards** (246 complete cases)
- Data were previously analysed by Diggle and Kenward (1994)

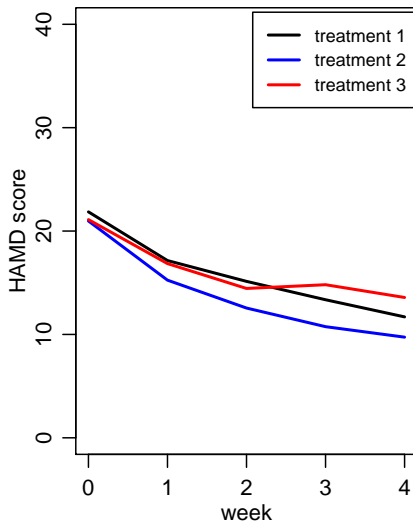
**Study objective: are there any differences in the effects of the 3 treatments on the change in HAMD score over time?**

# HAMD example: complete cases

## 50 Individual Profiles



## Mean Response Profiles



# HAMD example: analysis model

- Use the variables
  - ▶  $y$ , Hamilton depression (HAMD) score
  - ▶  $t$ , treatment
  - ▶  $w$ , week
- and for simplicity
  - ▶ ignore any centre effects
  - ▶ assume linear relationships
- A suitable analysis model would be a **hierarchical model with random intercepts and slopes** (we will look at the details of this later)
- If we had no missing data, we could just fit this model
- However ...

## HAMD example: exploring the missingness

- Subjects drop-out of the study from week 2 onwards
  - ⇒ HAMD score is missing for some individuals for 1, 2 or 3 weeks
- Before analysing the data, we should consider
  - ▶ how many subjects dropped out of the study each week?
  - ▶ is the level and pattern of drop-out consistent across treatments?

Percentage of missingness by treatment and week

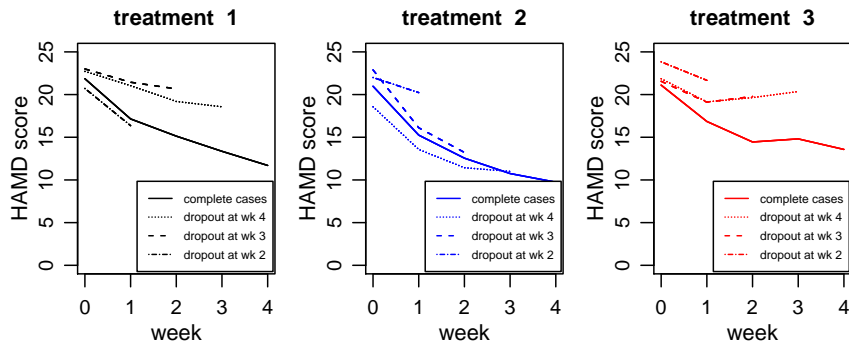
	treat. 1	treat. 2	treat. 3	all treatments
week 2	11.7	22.0	9.3	14.2
week 3	19.2	29.7	16.3	21.5
week 4	36.7	35.6	27.1	33.0

- Fewer subjects drop-out of treatment 3
- Drop-out occurs earlier for treatment 2

# HAMD example: implications of drop-out

- It is also very important to think about possible reasons for the missingness
- Why do you think some subjects dropped out of the HAMD study?
- Do you think that subjects who dropped out have similar characteristics to individuals who remained in the study?
- What do you think are the potential problems with only analysing complete cases?

# HAMD example: missing scores



- Do individuals have similar profiles whether they dropped out or remained in the study?
- Do all the treatments show the same patterns?

# HAMD example: model of missingness

- The full data for the HAMD example includes:
  - ▶  $y$ , HAMD score (observed and missing values)
  - ▶  $t$ , treatment
  - ▶  $m$ , a missing data indicator for  $y$ , s.t.

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

- So we can model
  - ▶  $y$  (random effects model)
  - ▶ and the probability that  $y$  is missing using:

$$m_{iw} \sim \text{Bernoulli}(p_{iw})$$

We now consider different possibilities for  $p_{iw}$ , which depend on the assumptions we make about the missing data mechanism

# Types of missing data

Following Rubin, missing data are generally classified into 3 types

Consider the mechanism that led to the missing HAMD scores ( $y$ )  
recall:  $p_{iw}$  is the probability  $y_{iw}$  is missing for individual  $i$  in week  $w$

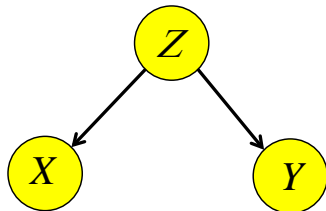
- Missing Completely At Random (MCAR)
  - ▶ missingness does not depend on observed or unobserved data
  - ▶ e.g.  $\text{logit}(p_{iw}) = \theta_0$
- Missing At Random (MAR)
  - ▶ missingness depends only on observed data
  - ▶ e.g.  $\text{logit}(p_{iw}) = \theta_0 + \theta_1 t_i$  or  $\text{logit}(p_{iw}) = \theta_0 + \theta_2 y_{i0}$   
note: drop-out from week 2; weeks 0 and 1 completely observed
- Missing Not At Random (MNAR)
  - ▶ neither MCAR or MAR hold
  - ▶ e.g.  $\text{logit}(p_{iw}) = \theta_0 + \theta_3 y_{iw}$

# Graphical models to represent different missing data mechanisms

- Graphical models can be a helpful way to visualise different missing data mechanisms and understand their implications for analysis
- More generally, graphical models are a useful tool for building complex Bayesian models
- Ingredients of a Bayesian graphical model:
  - ▶ Nodes: all random quantities (data and parameters)
  - ▶ Edges: associations between the nodes
- Used to represent a set of conditional independence statements about system of interest

See Spiegelhalter, Thomas and Best (1996); Spiegelhalter (1998); Richardson and Best (2003)

## Graphical models: a simple example

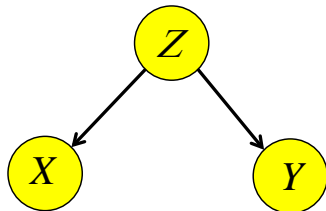


- $Z$  = genotypes of parents;  $X$ ,  $Y$  = genotypes of 2 children
- If we know the genotypes of the parents, then the children's genotypes are **conditionally independent**, i.e.

$$X \perp\!\!\!\perp Y|Z \quad \text{or} \quad Pr(X, Y|Z) = Pr(X|Z)Pr(Y|Z)$$

- Factorization thm: Jt distribution,  $Pr(\mathbf{V}) = \prod Pr(v|\text{parents}[v])$

## Graphical models: a simple example



- If parents' genotypes ( $Z$ ) are **unknown**, then the children's genotypes ( $X, Y$ ) are **marginally dependent**
  - ▶ Knowing  $X$  tells you something about the likely values of  $Y$
- Same principles apply if  $X, Y, Z$  represent random variables (data and parameters) in a statistical model

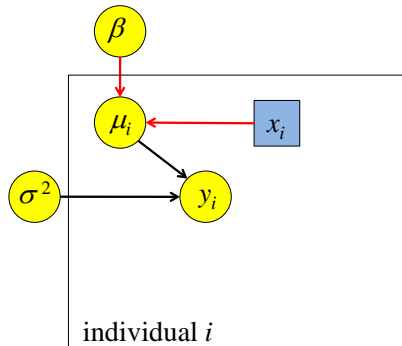
# Bayesian graphical models: notation

A typical regression model of interest

$$y_i \sim \text{Normal}(\mu_i, \sigma^2), \quad i = 1, \dots, N$$

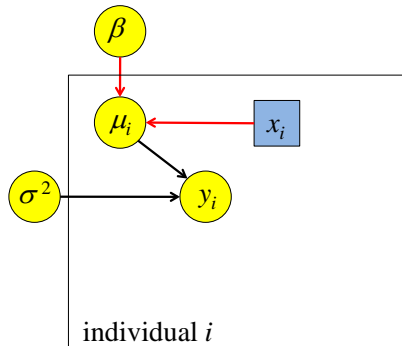
$$\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

$$\boldsymbol{\beta} \sim \text{fully specified prior}$$



# Bayesian graphical models: notation

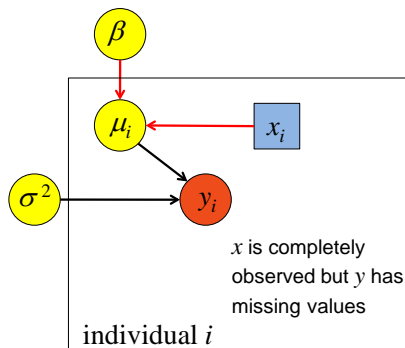
- yellow circles = random variables (data and parameters)
- blue squares = fixed constants (e.g. covariates, denominators)
- black arrows = stochastic dependence
- red arrows = logical dependence
- large rectangles = repeated structures (loops)



Directed Acyclic Graph (DAG) — contains only directed links (arrows) and no cycles

# Bayesian graphical models: notation

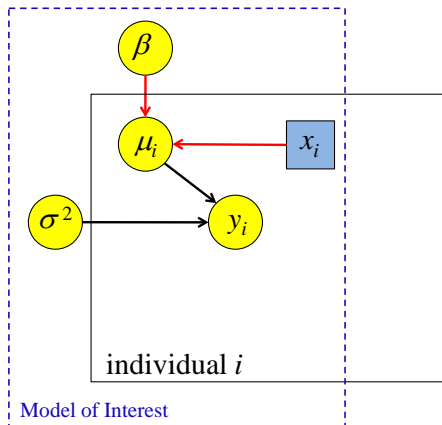
- yellow circles = random variables (data and parameters)
- blue squares = fixed constants (e.g. covariates, denominators)
- black arrows = stochastic dependence
- red arrows = logical dependence
- large rectangles = repeated structures (loops)



- We usually make no distinction in the graph between random variables representing data or parameters
- However, for clarity, we will denote a random variable representing a data node with **missing values** by an **orange circle**

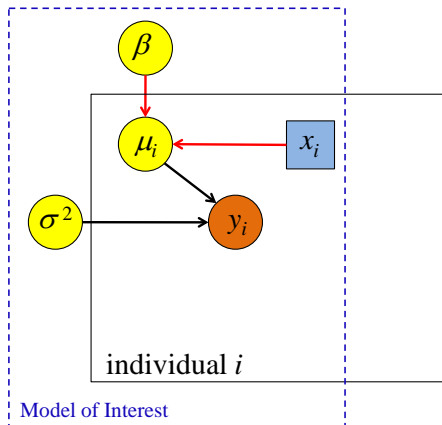
# Using DAGs to represent missing data mechanisms

A typical regression model of interest



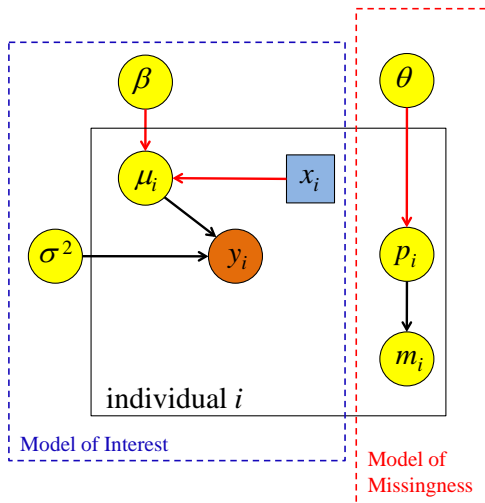
# Using DAGs to represent missing data mechanisms

Now suppose  $\mathbf{x}$  is completely observed, but  $y$  has missing values



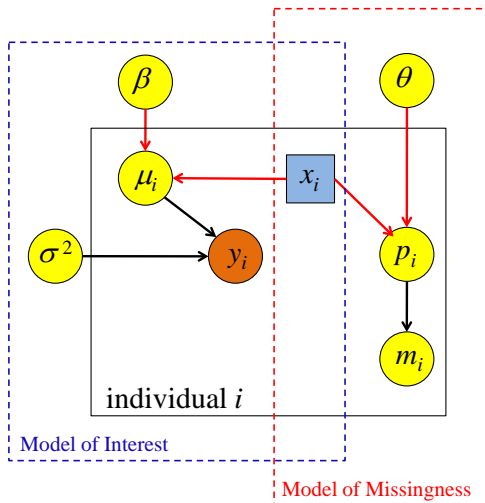
# DAG: Missing Completely At Random (MCAR)

A regression model of interest + model for probability of  $y$  missing



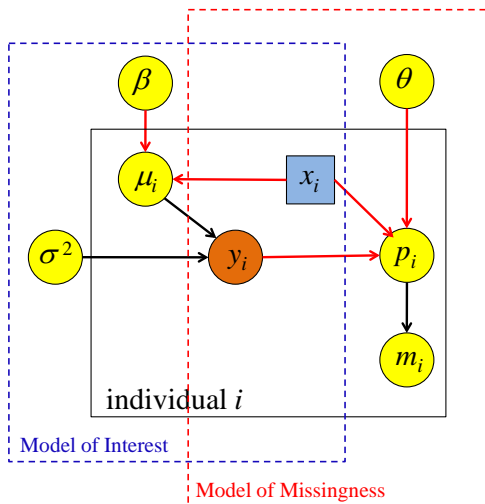
# DAG: Missing At Random (MAR)

A regression model of interest + model for probability of  $y$  missing



# DAG: Missing Not At Random (MNAR)

A regression model of interest + model for probability of  $y$  missing



## Joint model: general notation

We now use general notation, but in the HAMD example  $\mathbf{z} = (y, t)$

- Let  $\mathbf{z} = (z_{ij})$  denote a rectangular data set  
 $i = 1, \dots, n$  individuals and  $j = 1, \dots, k$  variables
- Partition  $\mathbf{z}$  into observed and missing values,  $\mathbf{z} = (\mathbf{z}^{obs}, \mathbf{z}^{mis})$
- Let  $\mathbf{m} = (m_{ij})$  be a binary indicator variable such that

$$m_{ij} = \begin{cases} 0: & z_{ij} \text{ observed} \\ 1: & z_{ij} \text{ missing} \end{cases}$$

- Let  $\beta$  and  $\theta$  denote vectors of unknown parameters
- Then the joint model (likelihood) of the full data is

$$f(\mathbf{z}, \mathbf{m} | \beta, \theta) = f(\mathbf{z}^{obs}, \mathbf{z}^{mis}, \mathbf{m} | \beta, \theta)$$

# Joint model: integrating out the missingness

- The joint model,  $f(\mathbf{z}^{obs}, \mathbf{z}^{mis}, \mathbf{m}|\beta, \theta)$ , cannot be evaluated in the usual way because it depends on missing data
- However, the marginal distribution of the observed data can be obtained by integrating out the missing data,

$$f(\mathbf{z}^{obs}, \mathbf{m}|\beta, \theta) = \int f(\mathbf{z}^{obs}, \mathbf{z}^{mis}, \mathbf{m}|\beta, \theta) d\mathbf{z}^{mis}$$

We now look at factorising the joint model

# Joint model: factorisations

- Two factorisations of the joint model are commonly used

- 1 selection models

$$f(\mathbf{z}^{obs}, \mathbf{z}^{mis}, \mathbf{m} | \beta, \theta) = f(\mathbf{m} | \mathbf{z}^{obs}, \mathbf{z}^{mis}, \beta, \theta) f(\mathbf{z}^{obs}, \mathbf{z}^{mis} | \beta, \theta)$$

- 2 pattern mixture models

$$f(\mathbf{z}^{obs}, \mathbf{z}^{mis}, \mathbf{m} | \beta, \theta) = f(\mathbf{z}^{obs}, \mathbf{z}^{mis} | \mathbf{m}, \beta, \theta) f(\mathbf{m} | \beta, \theta)$$

- An advantage of the selection model factorisation is that it includes the **model of interest** term directly
- On the other hand, the pattern mixture model corresponds more directly to what is actually observed (i.e. the **distribution of the data within subgroups having different missing data patterns**)

**We will focus on the selection model factorisation in this workshop**

# Joint model: selection model factorisation

- The selection model

$$f(\mathbf{z}^{obs}, \mathbf{z}^{mis}, \mathbf{m} | \beta, \theta) = f(\mathbf{m} | \mathbf{z}^{obs}, \mathbf{z}^{mis}, \beta, \theta) f(\mathbf{z}^{obs}, \mathbf{z}^{mis} | \beta, \theta)$$

simplifies with appropriate conditional independence assumptions

$$f(\mathbf{z}^{obs}, \mathbf{z}^{mis}, \mathbf{m} | \beta, \theta) = f(\mathbf{m} | \mathbf{z}^{obs}, \mathbf{z}^{mis}, \theta) f(\mathbf{z}^{obs}, \mathbf{z}^{mis} | \beta)$$

- $f(\mathbf{z}^{obs}, \mathbf{z}^{mis} | \beta)$  is the usual **likelihood** you would specify if all the data had been observed
- $f(\mathbf{m} | \mathbf{z}^{obs}, \mathbf{z}^{mis}, \theta)$  represents the **missing data mechanism** and describes the way in which the probability of an observation being missing depends on other variables (measured or not) and on its own value
- For some types of missing data, the form of the conditional distribution of  $\mathbf{m}$  can be simplified

# Simplifying the factorisation for MAR and MCAR

- Recall we wish to integrate out the missingness

$$\begin{aligned}f(\mathbf{z}^{obs}, \mathbf{m}|\beta, \theta) &= \int f(\mathbf{z}^{obs}, \mathbf{z}^{mis}, \mathbf{m}|\beta, \theta) d\mathbf{z}^{mis} \\ &= \int f(\mathbf{m}|\mathbf{z}^{obs}, \mathbf{z}^{mis}, \theta) f(\mathbf{z}^{obs}, \mathbf{z}^{mis}|\beta) d\mathbf{z}^{mis}\end{aligned}$$

- MAR missingness depends only on observed data, i.e.

$$f(\mathbf{m}|\mathbf{z}^{obs}, \mathbf{z}^{mis}, \theta) = f(\mathbf{m}|\mathbf{z}^{obs}, \theta)$$

$$\begin{aligned}\text{So, } f(\mathbf{z}^{obs}, \mathbf{m}|\beta, \theta) &= f(\mathbf{m}|\mathbf{z}^{obs}, \theta) \int f(\mathbf{z}^{obs}, \mathbf{z}^{mis}|\beta) d\mathbf{z}^{mis} \\ &= f(\mathbf{m}|\mathbf{z}^{obs}, \theta) f(\mathbf{z}^{obs}|\beta)\end{aligned}$$

- MCAR missingness is a special case of MAR that does not even depend on the observed data

$$f(\mathbf{m}|\mathbf{z}^{obs}, \mathbf{z}^{mis}, \theta) = f(\mathbf{m}|\theta)$$

$$\text{So, } f(\mathbf{z}^{obs}, \mathbf{m}|\beta, \theta) = f(\mathbf{m}|\theta) f(\mathbf{z}^{obs}|\beta)$$

# Ignorable/Nonignorable missingness

The missing data mechanism is termed **ignorable** if

- 1 the missing data are MCAR or MAR and
- 2 the parameters  $\beta$  and  $\theta$  are distinct

In the Bayesian setup, an additional condition is

- 3 the priors on  $\beta$  and  $\theta$  are independent

**'Ignorable' means we can ignore the model of missingness, but does not necessarily mean we can ignore the missing data!**

**However if the missing data mechanism is nonignorable, then we cannot ignore the model of missingness**

# Assumptions

- In contrast with the sampling process, which is often known, the missingness mechanism is usually unknown
- The data alone cannot usually definitively tell us the sampling process
  - ▶ But with fully observed data, we can usually check the plausibility of any assumptions about the sampling process e.g. using residuals and other diagnostics
- Likewise, the missingness pattern, and its relationship to the observations, cannot definitively identify the missingness mechanism
  - ▶ Unfortunately, the assumptions we make about the missingness mechanism **cannot** be definitively checked from the data at hand

# Sensitivity analysis

- The issues surrounding the analysis of data sets with missing values therefore centre on assumptions
- We have to
  - ▶ decide which assumptions are reasonable and sensible in any given setting - contextual/subject matter information will be central to this
  - ▶ ensure that the assumptions are transparent
  - ▶ explore the sensitivity of inferences/conclusions to the assumptions

# 'Ad-hoc' methods: complete case analysis (CC)

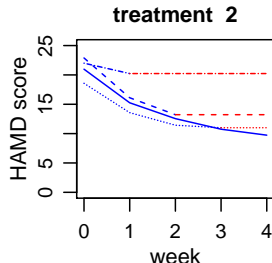
- Missing data are ignored and only complete cases analysed
- Many computer packages do this by default
- Advantage of CC
  - ▶ simple
- Disadvantages of CC
  - ▶ often introduces bias
  - ▶ inefficient as data from incomplete cases discarded

# 'Ad-hoc' methods: single imputation

- Aim to 'fill in' missing values to create single 'completed' (imputed) dataset, which is then analysed using standard methods
- Types of single imputation include
  - ▶ mean imputation
  - ▶ last observation carried forward (LOCF)
- Computationally simple but makes strong assumptions about missingness mechanism (usually MCAR)
- Ignores uncertainty about imputed missing values

# Last observation carried forward (LOCF)

- Widely used in clinical trial settings
- Assumption: **all unseen measurements = last seen measurement**
- Is not even valid under the strong assumption of MCAR
- In the HAMD example
  - ▶ if the treatment is working but individuals drop-out early, we will underestimate the treatment effect
  - ▶ LOCF changes the shape of the profile, as drop-outs likely to have improved at least a little
  - ▶ this effect will vary by treatment, and hence inference about treatment difference will be misleading



**Ad hoc methods are frequently used, but not recommended**

# ‘Statistically principled’ methods

- Generate statistical (stochastic) information about the missing values and/or the missingness mechanism, e.g.
  - ▶ **Multiple imputation (MI)** — generate  $K > 1$  imputed values for missing observations from appropriate probability distribution
  - ▶ **Fully model based (e.g. Bayesian)** — write down statistical model for full data (including missingness mechanism) and base analysis on this model
- Make weaker assumptions, but more computationally complex to implement
- In contrast to ad-hoc methods, principled methods are:
  - ▶ based on a **well-defined statistical model** for the complete data, and explicit assumptions about the missing value mechanism
  - ▶ the subsequent analysis, inferences and conclusions are **valid under these assumptions**
  - ▶ doesn't mean the assumptions are necessarily true but it does allow the dependence of the conclusions on these assumptions to be investigated

# Bayesian inference

Makes fundamental distinction between

- Observable quantities,  $\mathbf{D}$ , (i.e. the data)
- Unknown quantities,  $\Omega$ , (i.e. statistical parameters)
  - ▶ parameters are treated as **random variables**
  - ▶ in the **Bayesian framework**, we make probability statements about model parameters
  - ▶ in the **frequentist framework**, parameters are fixed non-random quantities and the probability statements concern the data

## Bayesian inference (continued)

As with any analysis, we start by positing a model,  $p(\mathbf{D} \mid \Omega)$

This is the sampling distribution of the data, equivalent to the **likelihood** in classical analyses

From a Bayesian point of view

- $\Omega$  is unknown so should have a **probability distribution** reflecting our uncertainty about it before seeing the data  
→ need to specify a **prior distribution**  $p(\Omega)$
- $\mathbf{D}$  is known so we should **condition** on it  
→ use Bayes theorem to obtain conditional probability distribution for unobserved quantities of interest given the data:

$$p(\Omega \mid \mathbf{D}) \propto p(\Omega) p(\mathbf{D} \mid \Omega)$$

This is the **posterior distribution**

# Bayesian computational methods

- Bayesian inference centres around the posterior distribution

$$p(\boldsymbol{\Omega}|\mathbf{D}) \propto p(\mathbf{D}|\boldsymbol{\Omega}) \times p(\boldsymbol{\Omega})$$

where  $\boldsymbol{\Omega}$  is typically a large vector  $\boldsymbol{\Omega} = \{\Omega_1, \Omega_2, \dots, \Omega_k\}$

- $p(\mathbf{D}|\boldsymbol{\Omega})$  and  $p(\boldsymbol{\Omega})$  will often be available in closed form, but  $p(\boldsymbol{\Omega}|\mathbf{D})$  is usually not analytically tractable, and we want to
  - ▶ obtain marginal posteriors

$$p(\Omega_i|\mathbf{D}) = \int \int \dots \int p(\boldsymbol{\Omega}|\mathbf{D}) d\boldsymbol{\Omega}_{(-i)}$$

where  $\boldsymbol{\Omega}_{(-i)}$  denotes the vector of  $\boldsymbol{\Omega}$ 's excluding  $\Omega_i$

- ▶ calculate properties of  $p(\Omega_i|\mathbf{D})$ , such as

$$E(\Omega_i|\mathbf{D}) = \int \Omega_i p(\Omega_i|\mathbf{D}) d\Omega_i$$

$$Pr(\Omega_i > T|\mathbf{D}) = \int_T^\infty p(\Omega_i|\mathbf{D}) d\Omega_i$$

→ **numerical integration** becomes vital

# Markov Chain Monte Carlo (MCMC) integration

- Suppose we can draw samples from the joint posterior distribution for  $\Omega$ , *i.e.*

$$(\Omega_1^{(1)}, \dots, \Omega_k^{(1)}), (\Omega_1^{(2)}, \dots, \Omega_k^{(2)}), \dots, (\Omega_1^{(N)}, \dots, \Omega_k^{(N)}) \sim p(\Omega | \mathbf{D})$$

- Then

- ▶  $\theta_1^{(1)}, \dots, \theta_1^{(N)}$  are a sample from the marginal posterior  $p(\Omega_1 | \mathbf{D})$
- ▶  $E(g(\Omega_1)) = \int g(\Omega_1)p(\Omega_1 | \mathbf{D})d\Omega_1 \approx \frac{1}{N} \sum_{i=1}^N g(\Omega_1^{(i)})$

→ this is **Monte Carlo integration**

- *Independent* sampling from  $p(\Omega | \mathbf{D})$  may be difficult
- **BUT** *dependent* sampling from a *Markov chain* with  $p(\Omega | \mathbf{D})$  as its stationary (equilibrium) distribution is easier
- Theorems exist which prove convergence in limit as  $N \rightarrow \infty$  even if the sample is dependent → this is **Markov Chain Monte Carlo integration**

# Bayesian methods for handling missing data

- Bayesian approach treats missing data as additional unknown quantities for which a posterior distribution can be estimated
  - ▶ no fundamental distinction between missing data and unknown parameters
  - ▶ In our previous notation

$$\begin{aligned} \mathbf{D} &= \{\mathbf{z}^{obs}, \mathbf{m}\} \\ \Omega &= \{\mathbf{z}^{mis}, \beta, \theta\} \end{aligned}$$

- Inferential machinery available for Bayesian parameter estimation extends automatically to models with missing data
  - ‘Just’ need to specify appropriate joint model for the observed and missing data and model parameters, and estimate in usual way using MCMC
- Fully model-based approach to missing data

# Advantages of Bayesian methods

- Fully Bayesian models
  - ▶ are theoretically sound
  - ▶ enable coherent model estimation
  - ▶ allow uncertainty to be fully propagated
- Bayesian models can easily be adapted to
  - ▶ include partially observed cases
  - ▶ incorporate realistic assumptions about the reasons for the missingness

# Coming up

- It is helpful to distinguish between
  - ▶ missing response and missing covariate data (regression context) i.e. we let  $\mathbf{z} = (y, \mathbf{x})$  where  $y$  is the response of interest and  $\mathbf{x}$  is a set of covariates
  - ▶ ignorable and non-ignorable missingness mechanisms, since when mechanism is ignorable, specifying the joint model reduces to specifying  $f(\mathbf{z}^{obs}, \mathbf{z}^{mis} | \beta)$ , and ignoring  $f(\mathbf{m} | \mathbf{z}^{obs}, \theta)$
- In Lecture 2, we will look at missing response data
- Missing covariate data is the subject of Lecture 3
- We will then discuss a general strategy for carrying out Bayesian regression analysis with missing data, including a range of recommended sensitivity analyses (Lecture 4)
- In Lecture 5, we will compare fully Bayesian models with Multiple Imputation and end with a general discussion

# Lecture 2.

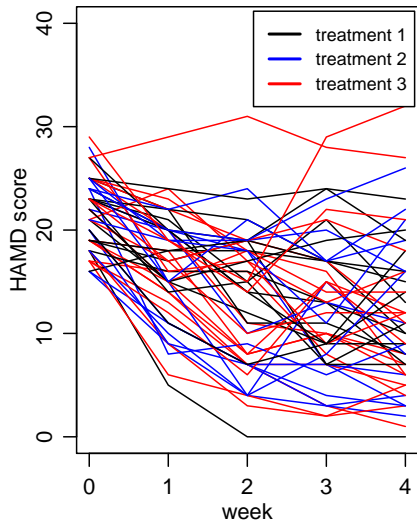
## Bayesian methods for missing response data

# Lecture Outline

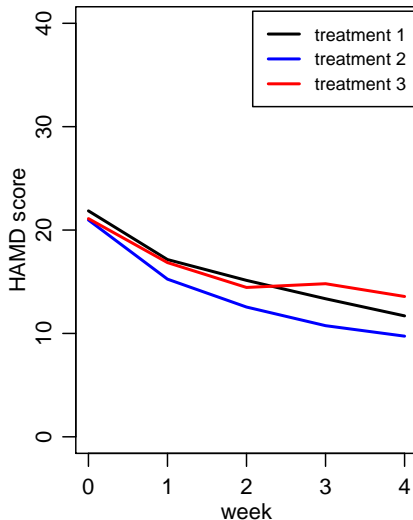
- Missing responses and missing covariates pose different problems
- In this lecture we will focus on analysing data with missing responses
- Using the HAMD example introduced in Lecture 1, we will see that dealing with missing responses in a ‘principled’ way within a Bayesian framework
  - ▶ is trivial when the missing data mechanism is ignorable
  - ▶ but is very dependent on assumptions when the missing data mechanism is informative
- This dependence on assumptions makes sensitivity analysis crucial

# HAMD example: recall exploratory analysis

## 50 Individual Profiles



## Mean Response Profiles



# HAMD example: complete case analysis (CC)

- We start by carrying out a complete case analysis
  - ▶ to provide a comparison with preferred methods
  - ▶ as a building block for more complex models
- CC requires the specification of an analysis model
- For the HAMD data, CC means
  - ▶ discarding partial data from 121 out of 367 subjects
  - ▶ specifying an analysis model which takes account of the repeated structure (observations are nested within individuals)
- We will specify a hierarchical model with random intercepts and random slopes, and for simplicity
  - ▶ ignore any centre effects
  - ▶ assume linear relationships

# HAMD example: analysis model

Specify a random effects (hierarchical) model:

$$y_{iw} \sim \text{Normal}(\mu_{iw}, \sigma^2)$$

$$\mu_{iw} = \alpha_i + \beta_{(\text{treat}(i),i)} \mathbf{w}$$

$$\alpha_i \sim \text{Normal}(\mu_\alpha, \sigma_\alpha^2) \text{ — random intercepts}$$

$$\beta_{(1,i)} \sim \text{Normal}(\mu_{\beta_1}, \sigma_{\beta_1}^2) \text{ — random slopes}$$

$\beta_{(2,i)}, \beta_{(3,i)}$  — follow common prior distributions similar to  $\beta_{(1,i)}$

$$\mu_\alpha, \mu_{\beta_1}, \mu_{\beta_2}, \mu_{\beta_3} \sim \text{Normal}(0, 10000) \text{ — vague prior on hyperparameters}$$

$$\sigma_\alpha, \sigma_{\beta_1}, \sigma_{\beta_2}, \sigma_{\beta_3} \sim \text{Uniform}(0, 100) \text{ — vague prior on hyperparameters}$$

$$\frac{1}{\sigma^2} \sim \text{Gamma}(0.001, 0.001) \text{ — vague prior on precision}$$

$y_{iw}$  = HAMD score for individual  $i$  in week  $w$  (weeks 0, ..., 4)

$\text{treat}(i)$  = treatment indicator of individual  $i$  (takes values 1, 2 or 3)

$w$  = week of the visit, takes value 0 for visit before treatment  
and values 1-4 for follow-up visits

# HAMD example: interpretation of results

- Study objective: are there any differences in the effects of the 3 treatments on the change in HAMD score over time?
- So we are particularly interested in the differences in the slope parameters (contrasts), i.e.

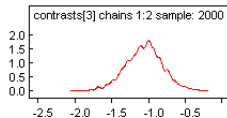
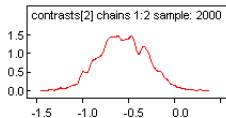
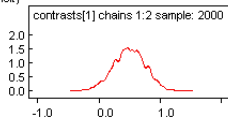
1  $\mu_{\beta_1} - \mu_{\beta_2}$

2  $\mu_{\beta_1} - \mu_{\beta_3}$

3  $\mu_{\beta_2} - \mu_{\beta_3}$

- Results for complete case analysis

Kernel density



## Recap: selection model

Recall, in general, the joint model for a set of data ( $\mathbf{z}$ ) and missing data indicator ( $\mathbf{m}$ ) can be factorised as

$$f(\mathbf{z}^{obs}, \mathbf{z}^{mis}, \mathbf{m} | \beta, \theta) = f(\mathbf{m} | f(\mathbf{z}^{obs}, \mathbf{z}^{mis}, \theta)) f(\mathbf{z}^{obs}, \mathbf{z}^{mis} | \beta)$$

model of missingness analysis model

- When the missing data mechanism is ignorable, we can ignore the model of missingness and just fit the analysis model
- So we need only specify  $f(\mathbf{z}^{obs}, \mathbf{z}^{mis} | \beta)$
- In the regression context, we let  $\mathbf{z} = (y, \mathbf{x})$  where
  - ▶  $y$  is the response of interest
  - ▶  $\mathbf{x}$  is a set of covariates

# Missing response data

- assuming missing data mechanism is ignorable (I)

In this case,  $\mathbf{z}^{mis} = \mathbf{y}^{mis}$  and  $\mathbf{z}^{obs} = (\mathbf{y}^{obs}, \mathbf{x})$

- Usually treat fully observed covariates as fixed constants rather than random variables with a distribution
- Model  $f(\mathbf{z}^{obs}, \mathbf{z}^{mis} | \beta)$  reduces to specification of  $f(\mathbf{y}^{obs}, \mathbf{y}^{mis} | \mathbf{x}, \beta)$
- $f(\mathbf{y}^{obs}, \mathbf{y}^{mis} | \mathbf{x}, \beta)$  is just the usual likelihood we would specify for fully observed response  $y$
- Estimating the missing responses  $\mathbf{y}^{mis}$  is equivalent to posterior prediction from the model fitted to the observed data
  - ⇒ Imputing missing response data **under an ignorable mechanism** will not affect estimates of model parameters

# Missing response data

## - assuming missing data mechanism is ignorable (II)

### In WinBUGS

- denote missing response values by `NA` in the data file
- specify response distribution (likelihood) as you would for complete data
- missing data are treated as additional unknown parameters
  - ⇒ WinBUGS will automatically simulate values for the missing observations according to the specified likelihood distribution, conditional on the current values of all relevant unknown parameters

# HAMD example: ignorable missing data mechanism

- Assume the missing data mechanism is ignorable for the HAMD example
  - ▶ the probability of the score being missing is not related to the current score (or change in score since the previous week)
- Use the same model (and WinBUGS code) as for the complete case analysis
- But the data now includes incomplete records (with the missing values denoted by NA)

# HAMD example: impact on treatment comparisons

**Table:** posterior mean (95% credible interval) for the contrasts (treatment comparisons) from random effects models fitted to the HAMD data

treatments	complete cases <sup>*</sup>		all cases <sup>†</sup>	
1 v 2	0.50	(-0.03,1.00)	0.74	(0.25,1.23)
1 v 3	-0.56	(-1.06,-0.04)	-0.51	(-1.01,-0.01)
2 v 3	-1.06	(-1.56,-0.55)	-1.25	(-1.73,-0.77)

\* individuals with missing scores ignored

† individuals with missing scores included under the assumption that the missingness mechanism is ignorable

Including all the partially observed cases in the analysis provides stronger evidence that:

- treatment 2 is more effective than treatment 1
- treatment 2 is more effective than treatment 3

# HAMD example: non-ignorable missing data mechanism

- Before defining a model for the missing data mechanism
  - ▶ think about the process that led to the missingness
  - ▶ gather information from the literature
  - ▶ seek insight from those involved in the data collection process
- For the HAMD model
  - ▶ patients for whom the treatment is successful and get better may decide not to continue in the study
  - ▶ conversely, if they are not showing any improvement or feeling worse, they may seek alternative treatment and drop-out of the study
  - ▶ in either case, the probability of obtaining a HAMD score is dependent on the change in the patient's depression over the previous week, i.e. since the last HAMD score

# Model for the missing data mechanism

- Then, translate these findings into a statistical model, e.g.

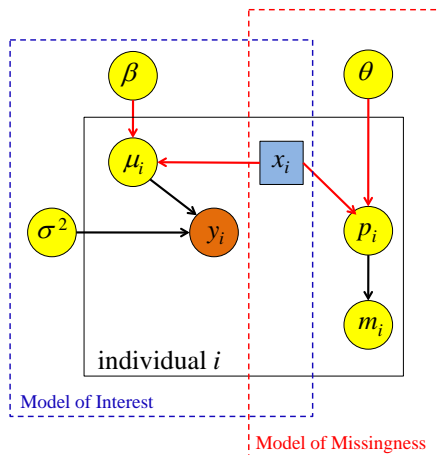
$$m_i \sim \text{Bernoulli}(p_i)$$

$$\text{logit}(p_i) = \theta_0 + \sum_{k=1}^r \theta_k w_{ki} + \delta y_i$$

- $w_{1i}, \dots, w_{ri}$  is a vector of variables which are predictive of the response missingness
- The inclusion of the response,  $y$ 
  - ▶ changes the missingness assumption from MAR to MNAR
  - ▶ provides the link with the analysis model

# Recall: DAG for Missing At Random (MAR)

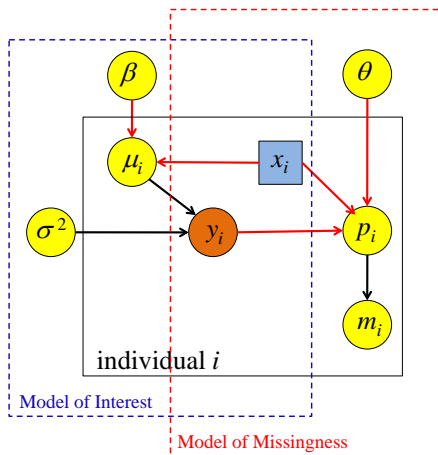
A regression model of interest + model for probability of  $y$  missing



For MAR, the only common element ( $x$ ) in the two models is fully observed

## Recall: DAG for Missing Not At Random (MNAR)

A regression model of interest + model for probability of  $y$  missing



For MNAR, one of the common elements ( $y$ ) has missing values which are estimated

# Important choices

- The shape of the relationship between the response and the probability of missingness
  - ▶ is a linear relationship adequate?
  - ▶ would a more complex shape, e.g. piecewise linear functional form, be better?
  - ▶ for some datasets, it may be intuitively plausible that the response is more likely to be missing if it takes high or low values
- The most appropriate way of including the response
  - ▶ with longitudinal data, could use change in response between current and previous values
- In the absence of any prior knowledge, recommended strategy is to assume a linear relationship between the probability of missingness and the response or change in response (Mason(2009), Chapter 4)

# Estimating parameters associated with the response

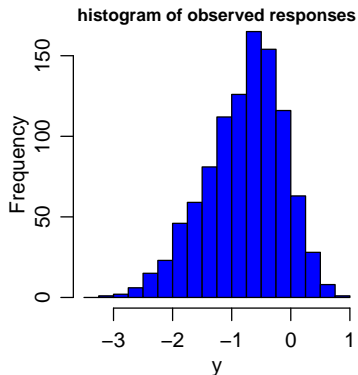
- The parameters associated with the response,  $\delta$ , are identified by the parametric assumptions in
  - ▶ the analysis model (AM)
  - ▶ the model of missingness (MoM)
- Missing responses are imputed in a way that is consistent with the distributional assumptions in the AM given their covariates in the AM
- Thus  $\delta$  are (at least weakly) identified by the observed data in combination with the other model assumptions
- Estimation difficulties increase for more complex models of missingness
- Advisable to use informative priors if possible

# How the AM distributional assumptions are used

Illustrative example (Daniels & Hogan (2008), Section 8.3.2)

- Consider a cross-sectional setting with
  - ▶ a single response
  - ▶ no covariates
- Suppose we specify a linear MoM,

$$\text{logit}(p_i) = \theta_0 + \delta y_i$$



- If we assume the AM follows a normal distribution,  $y_i \sim N(\mu_i, \sigma^2)$ 
  - ▶ must fill in the right tail  $\Rightarrow \delta > 0$
- If we assume the AM follows a skew-normal distribution
  - ▶  $\Rightarrow \delta = 0$

# Uncertainty in the AM distributional assumptions

- Inference about  $\delta$  is heavily dependent on the AM distributional assumptions about the residuals in combination with the choice and functional form of the covariates
- Unfortunately the AM distribution is unverifiable from the observed data when the response is MNAR
- Different AM distributions lead to different results
- Hence sensitivity analysis required to explore impact of different plausible AM distributions

# HAMD example: non-ignorable missing data mechanism

- If we think the probability of the score being missing might be related to the current score, then we must jointly fit
  - 1 an analysis model - already defined
  - 2 a model for the missing data mechanism
- Assume that the probability of score being missing is related to its current value, then model the missing response indicator as

$$m_{iw} \sim \text{Bernoulli}(p_{iw})$$
$$\text{logit}(p_{iw}) = \theta_0 + \delta(y_{iw} - \bar{y})$$
$$\theta_0, \delta \sim \text{mildly informative priors}$$

where  $\bar{y}$  is the mean score of observed  $y$ s

# HAMD example: MAR v MNAR

**Table:** posterior mean (95% credible interval) for the contrasts (treatment comparisons) from random effects models fitted to the HAMD data

treatments	complete cases <sup>1</sup>	all cases (mar) <sup>2</sup>	all cases (mnar) <sup>3</sup>
1 v 2	0.50 (-0.03,1.00)	0.74 (0.25,1.23)	0.75 (0.26,1.24)
1 v 3	-0.56 (-1.06,-0.04)	-0.51 (-1.01,-0.01)	-0.47 (-0.98,0.05)
2 v 3	-1.06 (-1.56,-0.55)	-1.25 (-1.73,-0.77)	-1.22 (-1.70,-0.75)

<sup>1</sup> individuals with missing scores ignored

<sup>2</sup> individuals with missing scores included under the assumption that the missingness mechanism is ignorable

<sup>3</sup> individuals with missing scores included under the assumption that the missingness mechanism is non-ignorable

Allowing for informative missingness with dependence on the current HAMD score:

- has a slight impact on the treatment comparisons
- yields a 95% interval comparing treatments 1 & 3 that includes 0

# HAMD example: sensitivity analysis - MoM

- Since the true missingness mechanism is unknown and cannot be checked, sensitivity analysis is essential
- We have already assessed the results of
  - ▶ assuming the missingness mechanism is ignorable
  - ▶ using an informative missingness mechanism of the form

$$\text{logit}(p_i) = \theta_0 + \delta(y_{iw} - \bar{y})$$

- However, we should also look at alternative informative missingness mechanisms, e.g.
  - ▶ allow drop-out probability to be dependent on the *change* in score

$$\text{logit}(p_i) = \theta_0 + \theta_1(y_{i(w-1)} - \bar{y}) + \delta_2(y_{iw} - y_{i(w-1)})$$

- ▶ allow different  $\theta$  and  $\delta$  for each treatment
- ▶ use different prior distributions

# HAMD example: sensitivity analysis results

**Table:** posterior mean (95% credible interval) for the contrasts (treatment comparisons) from random effects models fitted to all the HAMD data

treatments	mar		mnar1*		mnar2†	
1 v 2	0.74	(0.25,1.23)	0.75	(0.26,1.24)	0.72	(0.23,1.22)
1 v 3	-0.51	(-1.01,-0.01)	-0.47	(-0.98,0.05)	-0.60	(-1.09,-0.11)
2 v 3	-1.25	(-1.73,-0.77)	-1.22	(-1.70,-0.75)	-1.32	(-1.80,-0.84)

\* probability of missingness dependent on current score

† probability of missingness dependent on change in score

- This is a sensitivity analysis, we do NOT choose the “best” model
- Model comparison with missing data is very tricky
  - ▶ we cannot use the DIC automatically generated by WinBUGS on its own (Mason et al., 2012a)
- The range of results should be presented

# HAMD example: sensitivity analysis - AM

- Sensitivity to the assumptions about the AM should also be explored
- For HAMD data, possibilities include

- ▶ allow for non-linearity by including a quadratic term

$$\mu_{iw} = \alpha_i + \beta_{treat(i)} W + \gamma_{treat(i)} W^2$$

- ▶ include centre effects by allowing a different intercept for each centre
  - ▶ instead of using random effects, account for the repeated structure using an autoregressive model to explicitly model the autocorrelation between weekly visits for each individual
- A comprehensive sensitivity analysis will pair different combinations of AMs and MoMs

# Summary and extensions

- Missing response data is trivial to handle in the Bayesian framework under the assumption of an **ignorable** missing data mechanism
  - ▶ equivalent to posterior prediction from the model fitted to the observed data
- Using the HAMD example, we have shown a general framework for modelling **informative** missing response mechanisms by jointly modelling
  - ▶ the analysis model of interest
  - ▶ a model for the missing data indicator, which is a function of the missing response values (and possibly other observed covariates)

## Summary and extensions

- The most appropriate way of modelling the missing data indicator will be problem specific
- Some elaboration may be required to accommodate different types of drop-out (e.g. death and recovery)
  - ▶ multinomial regression model for categorical missing data indicator
- Some datasets may have informative drop-in (e.g. medics start to monitor patients if they become more unwell)
- For longitudinal studies with drop-out, an alternative is to replace the missingness indicator with a variable representing the time to drop-out and model this using (discrete or continuous time) survival techniques
- Similar selection model approaches are available in likelihood settings, but their implementation may require non-standard maximisation and numerical integration algorithms

# Lecture 3.

## Bayesian methods for missing covariate data

# Lecture Outline

- We have now looked at Bayesian methods for missing responses
- Missing covariates pose some additional problems
- In this lecture, we look at methods for
  - ▶ a single covariate  
(assuming missing data mechanism is ignorable)
  - ▶ multiple covariates  
(assuming missing data mechanism is ignorable)
  - ▶ allowing the missing data mechanism to be informative

# Missing responses v missing covariates

- For missing responses, recall that
  - ▶ missing values are automatically simulated from the analysis model,  $f(y^{obs}, y^{mis} | \mathbf{x}, \beta)$
  - ▶ assuming ignorable missingness, no additional sub-models are required
- For covariates, we will see that
  - ▶ we must build an imputation model to predict their missing values
  - ▶ regardless of the missingness mechanism, at least one additional sub-model is always required

# Treatment of covariates with missing values

Recall that we wish to evaluate the joint model  $f(\mathbf{z}^{obs}, \mathbf{z}^{mis}, \mathbf{m} | \beta, \theta)$

In this case,  $\mathbf{z}^{mis} = \mathbf{x}^{mis}$  and  $\mathbf{z}^{obs} = (y, \mathbf{x}^{obs})$

- To include records with missing covariates (associated response and other covariates observed) we
  - ▶ now have to treat covariates as **random variables** rather than fixed constants
  - ▶ must build an **imputation model** to predict their missing values
- Typically, the joint model (ignoring the missingness mechanism),  $f(\mathbf{z}^{obs}, \mathbf{z}^{mis} | \beta) = f(y, \mathbf{x}^{obs}, \mathbf{x}^{mis} | \beta)$  is factorised as

$$f(y, \mathbf{x}^{obs}, \mathbf{x}^{mis} | \beta) = f(y | \mathbf{x}^{obs}, \mathbf{x}^{mis}, \beta_y) f(\mathbf{x}^{obs}, \mathbf{x}^{mis} | \beta_x)$$

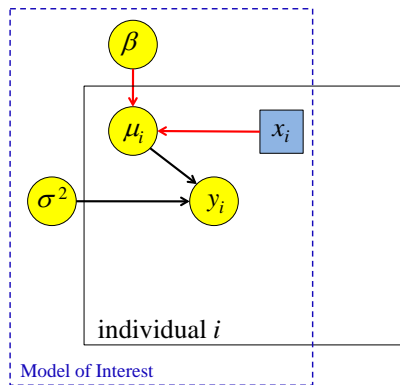
where  $\beta$  is partitioned into conditionally independent subsets  $(\beta_y, \beta_x)$

# Analysis model and covariate imputation model (CIM)

- The first term in the joint model factorisation,  $f(y|\mathbf{x}^{obs}, \mathbf{x}^{mis}, \beta_y)$ , is the usual **likelihood** for the response given fully observed covariates (analysis model)
- The second term,  $f(\mathbf{x}^{obs}, \mathbf{x}^{mis}|\beta_x)$  can be thought of as a **'prior model'** for the covariates (which are treated as random variables, not fixed constants), e.g.
  - ▶ joint prior distribution, say MVN
  - ▶ regression model for each variable with missing values
- It is not necessary to explicitly include response,  $y$ , as a predictor in the prior imputation model for the covariates, as its association with  $\mathbf{x}$  is already accounted for by the first term in the joint model factorisation (unlike multiple imputation)

# A typical DAG

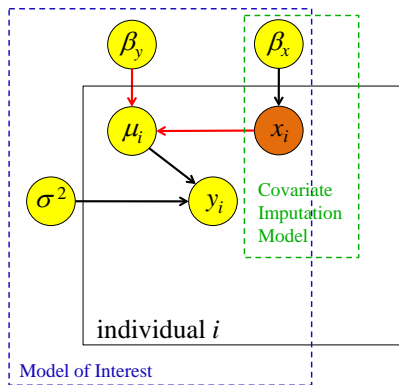
A regression model of interest with fully observed covariates



Note:  $\mathbf{x}$  and  $\mathbf{y}$  are completely observed

# DAG with Missing Covariates

A regression model of interest + prior model for covariates



Note:  $y$  is completely observed, but  $x$  has missing values (assumed to have ignorable missingness mechanism)

# Implementation

## In WinBUGS

- Denote missing covariate values by `NA` in the data file
- Specify **usual regression analysis model**, which will depend on partially observed covariates
- In addition, specify **prior distribution or regression model** for the covariate(s) with missing values
- WinBUGS will automatically simulate values from the posterior distribution of the missing covariates (which will depend on the prior model for the covariates and the likelihood contribution from the corresponding response variable)
- **Uncertainty** about the covariate imputations is automatically accounted for in the analysis model

# Single covariate with missing values ( $x$ )

There are 2 obvious ways of building the covariate imputation model

1 Specify a distribution, e.g. if  $x$  is continuous

- ▶ specify  $x_i \sim N(\nu, \zeta^2)$
- ▶ assume vague priors for  $\nu$  and  $\zeta^2$

2 Build a regression model relating  $x_i$  to other observed covariates, e.g. if  $x$  is binary

- ▶ specify

$$x_i \sim \text{Bernoulli}(q_i)$$

$$q_i = \phi_0 + \sum_{k=1}^s \phi_k Z_{ki}$$

$$\phi_0, \phi_1, \dots, \phi_s \sim \text{prior distribution}$$

- ▶  $Z_{1i}, \dots, Z_{si}$  is a vector of fully observed covariates
- Compare pattern of imputed values with the observed values to check the adequacy of the covariate imputation model (CIM)

## LBW example: low birth weight data

- Study objective: is there an association between trihalomethane (THM) concentrations and the risk of full term low birth weight?
  - ▶ THM is a by-product of chlorine water disinfection potentially harmful for reproductive outcomes
- The variables we will use are:
  - $Y$ : binary indicator of low birth weight (outcome)
  - $X$ : binary indicator of THM concentrations (exposure of interest)
  - $C$ : mother's age, baby gender, deprivation index (vector of measured confounders)
  - $U$ : smoking (a partially measured confounder)

So  $\mathbf{z}^{mis} = U^{mis}$  and  $\mathbf{z}^{obs} = (Y, X, C, U^{obs})$

- We have data for 8969 individuals, but only 931 have an observed value for smoking
  - ▶ 90% of individuals will be discarded if we use CC

# LBW example: missingness assumptions

- Assume that *smoking* is MAR
  - ▶ probability of smoking being missing does not depend on whether the individual smokes
  - ▶ this assumption is reasonable as the missingness is due to the sample design of the underlying datasets
- Also assume that the other assumptions for ignorable missingness hold (Lecture 1), so we do not need to specify a model for the missingness mechanism
- However, since *smoking* is a covariate, we must specify an imputation model if we wish to include individuals with missing values of *smoking* in our dataset

## LBW example: specification of joint model

- **Analysis model**: logistic regression for outcome, low birth weight

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

$$\text{logit}(p_i) = \beta_0 + \beta_X X_i + \beta_C^T \mathbf{C}_i + \beta_U U_i$$

$$\beta_0, \beta_X, \dots \sim \text{Normal}(0, 10000^2)$$

- **Imputation model (1)**: prior distribution for missing smoking

$$U_i \sim \text{Bernoulli}(q_i)$$

$$q_i \sim \text{Beta}(1, 1)$$

- Key assumption: distribution of missing values of  $U$  is **exchangeable** with observed distribution
- Imputed  $U$ 's will depend on posterior distribution of  $q$  estimated from observed  $U$ 's and on 'feedback' from the analysis model about the relationship between  $U$  and  $Y$  (adjusted for  $X$  and  $C$ )
  - ▶ In a Bayesian joint model, feedback means we don't need to explicitly include the response,  $Y$ , in the imputation model
  - ▶ cf multiple imputation, where we do

## LBW example: Imputation model (2)

- Often, the simple exchangeability assumption for the prior on the missing  $U$  is not reasonable
  - ▶ May want to stratify prior by other fully observed covariates, or build a regression model
  - ▶ Particularly important to account for any variables related to the missing data mechanism, to strengthen plausibility of MAR assumption
- **Imputation model (2)**: logistic regression for missing covariate, smoking

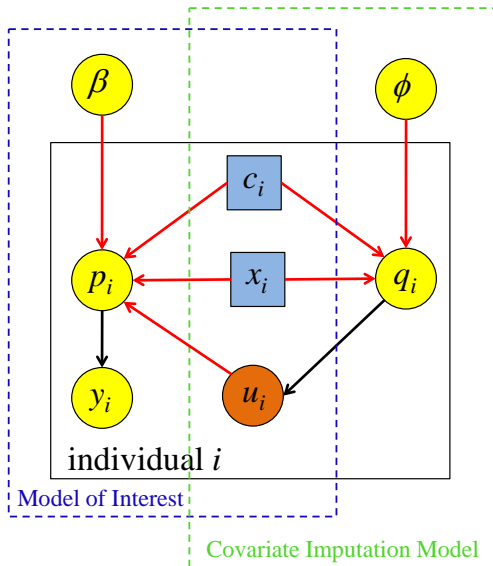
$$U_i \sim \text{Bernoulli}(q_i)$$

$$\text{logit}(q_i) = \phi_0 + \phi_X X_i + \phi_C^T \mathbf{C}_i$$

$$\phi_0, \phi_X, \dots \sim \text{Normal}(0, 10000^2)$$

- Note that study design (missing data mechanism) depends on the deprivation index ( $C$ ), so this has been included in the imputation model

# LBW example: graphical representation



## LBW example: results

		Odds ratio (95% interval)			
		CC (N=931)	All (N=8969)		
<i>X</i>	Trihalomethanes				
	> 60 $\mu$ g/L	2.36	(0.96,4.92)	1.17	(1.01,1.37)
<i>C</i>	Mother's age				
	$\leq 25$	0.89	(0.32,1.93)	1.05	(0.74,1.41)
	25 – 29*		1		1
	30 – 34	0.13	(0.00,0.51)	0.80	(0.55,1.14)
	$\geq 35$	1.53	(0.39,3.80)	1.14	(0.73,1.69)
<i>C</i>	Male baby	0.84	(0.34,1.75)	0.76	(0.58,0.95)
<i>C</i>	Deprivation index	1.74	(1.05,2.90)	1.34	(1.17,1.53)
<i>U</i>	Smoking	1.86	(0.73,3.89)	1.92	(0.80,3.82)

\* Reference group

- CC analysis is very uncertain
- Extra records shrink intervals for *X* coefficient substantially

## LBW example: results

		Odds ratio (95% interval)			
CC (N=931)		All (N=8969)			
<i>X</i>	Trihalomethanes > 60 $\mu$ g/L	2.36	(0.96,4.92)	1.17	(1.01,1.37)
<i>C</i>	Mother's age				
	≤ 25	0.89	(0.32,1.93)	1.05	(0.74,1.41)
	25 – 29*		1		1
	30 – 34	0.13	(0.00,0.51)	0.80	(0.55,1.14)
	≥ 35	1.53	(0.39,3.80)	1.14	(0.73,1.69)
<i>C</i>	Male baby	0.84	(0.34,1.75)	0.76	(0.58,0.95)
<i>C</i>	Deprivation index	1.74	(1.05,2.90)	1.34	(1.17,1.53)
<i>U</i>	Smoking	<b>1.86</b>	<b>(0.73,3.89)</b>	<b>1.92</b>	<b>(0.80,3.82)</b>

\* Reference group

- Little impact on *U* coefficient, reflecting uncertainty in imputations

## Aside: bias in CC

- In regression with missing covariates, the conditions under which CC is biased do not fit neatly into the MCAR/MAR/MNAR categorisation (White & Carlin, 2010)
- For example, consider the case of a single covariate,  $X$ , with missing values and fully observed  $Y$
- CC is unbiased if missingness in  $X$  is
  - ▶ MCAR
  - ▶ MNAR dependent on  $X$
- CC is biased if missingness in  $X$  is
  - ▶ MAR dependent on  $Y$
  - ▶ MNAR dependent on  $X$  and  $Y$

# Multiple covariates with missing values

- The covariate imputation model gets more complex if  $> 1$  missing covariates
  - ▶ typically need to account for **correlation** between missing covariates
  - ▶ could assume multivariate normality if covariates all continuous
  - ▶ for mixed binary, categorical and continuous covariates, could fit latent variable (multivariate probit) model (Chib and Greenberg 1998)

We now extend our LBW example to 2 covariates

## LBW example: two covariates with missing values

- Assume that our set of partially measured confounders ( $\mathbf{U}$ ) contains 2 variables
  - ▶ smoking
  - ▶ ethnicity
- Imputation model: Multivariate Probit for  $P(U|X, \mathbf{C})$

$$\mathbf{U}_i^* \sim MVN(\boldsymbol{\mu}_i, \Sigma)$$

$$\boldsymbol{\mu}_i = \gamma_0 + \gamma_X \mathbf{X}_i + \gamma_C^T \mathbf{C}_i$$

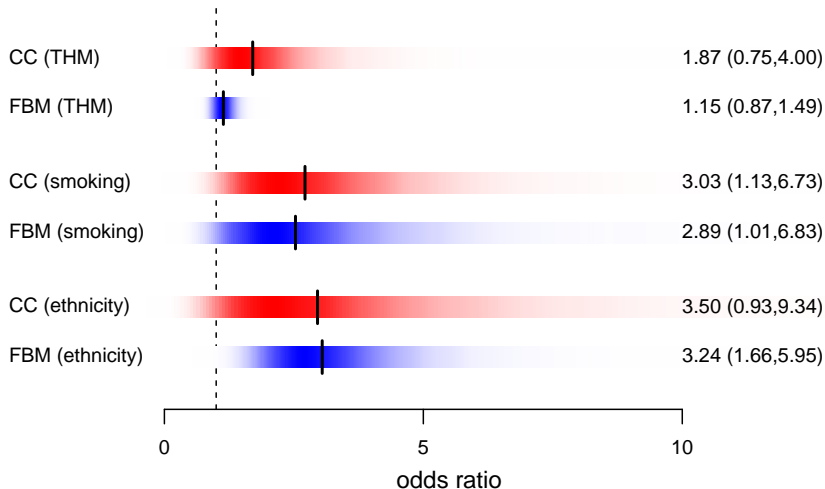
$$U_{ij} = I(U_{ij}^* > 0), j = 1, 2$$

$$\mathbf{U}_i^* = \begin{pmatrix} U_{i1}^* \\ U_{i2}^* \end{pmatrix}, \boldsymbol{\mu}_i = \begin{pmatrix} \mu_{i1} \\ \mu_{i2} \end{pmatrix}, \Sigma = \begin{pmatrix} 1 & \kappa \\ \kappa & 1 \end{pmatrix}$$

$$\kappa \sim \text{Uniform}(-1, 1); \quad \gamma_0, \gamma_X, \gamma_C \sim \text{Normal}(0, 10000)$$

- See Molitor *et al.* (2009) for further details

# LBW example: two missing covariates - results



Plot shows posterior distribution with median marked.  
Posterior mean and 95% interval are shown on the right.

# Informative missingness mechanism for covariates

- If we assume that *smoking* is MNAR, then we must add a third part to the model
  - ▶ a model of missingness with a missingness indicator variable ( $n_i$ ) for smoking as the response
  - ▶ e.g.

$$n_i \sim \text{Bernoulli}(r_i)$$

$$\text{logit}(r_i) = \theta_0 + \theta_W^T \mathbf{W}_i + \delta U_i$$

- ▶  $\mathbf{W}$  is a vector of observed variables predictive of the covariate missingness and  $U$  is the (possibly missing) smoking variable
- This Model of Covariate Missingness (MoCM) is similar to the Model of Response Missingness (MoRM) discussed in Lecture 2.

## Summary and extensions

- Excluding missing covariates can be biased and potentially very inefficient
- Handling missing covariates in a Bayesian framework requires specification of an additional sub-model to impute the covariates, even under MCAR or MAR assumptions
- Recommendations for building the covariate imputation model are broadly the same as for multiple imputation
  - ▶ include variables that are related to the imputed variable
  - ▶ include variables that are related to the missingness mechanism
  - ▶ account for correlation structure between multiple variables with missing values
  - ▶ account for hierarchical structure in the covariates if appropriate
- A key difference from multiple imputation is that it is not necessary to include the response in the covariate imputation model
- For covariates that are missing not at random (NMAR), a model of missingness must also be specified

# Summary and extensions

## Missing responses and covariates

- Bayesian models with both missing responses and covariates have the potential to become quite complicated, particularly if we cannot assume ignorable missingness
- However, the Bayesian framework is well suited to the process of building complex models, linking smaller sub-models as a coherent joint model
- A typical model may consist of 3 (or more) parts:
  - 1 analysis model
  - 2 covariate imputation model
  - 3 model(s) of missingness (for response and/or covariates)
- In the next lecture, we will look at a strategy for building complex Bayesian models for analysing incomplete data

# Lecture 4.

## A general strategy for modelling missing data using Bayesian methods

# Lecture Outline

- Modelling data with missing values can be complicated
- Guidance on the practicalities of approaching this task might help
- We provide this by proposing a general strategy
- In this lecture, we
  - ▶ present a strategy for a ‘statistically principled’ investigation of data with missing covariates and/or responses
  - ▶ use an illustrative example to demonstrate each step
  - ▶ discuss adaptations and extensions
- Although this strategy was designed for Bayesian modelling, the framework could be adopted for other inference paradigms.

# Illustrative example

- **Research Question:** is a single mother's rate of pay affected by gaining a partner?
- **Data:** taken from the Millennium Cohort Study (MCS), which
  - ▶ follows 18,000+ children born in the UK at the start of the Millennium
  - ▶ includes information about the children's families
- We use data from sweeps 1 and 2 on the cohort member's mother, with inclusion criteria:
  - ▶ single in sweep 1
  - ▶ in work
  - ▶ not self-employed

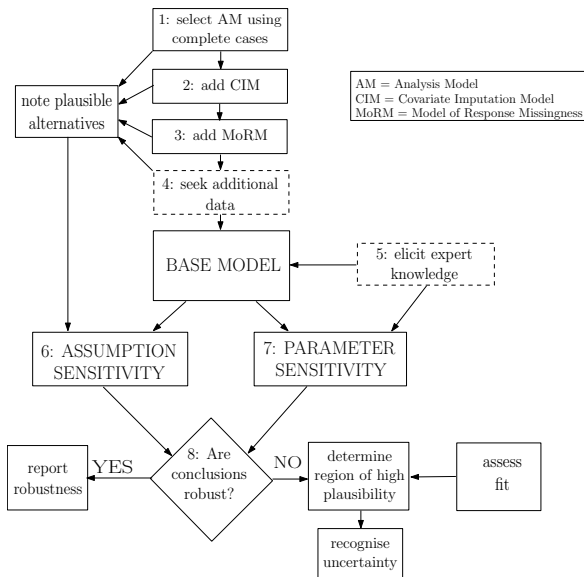
# Strategy Overview

- The strategy consists of two parts
  - ▶ constructing a base model
  - ▶ assessing conclusions from this base model against a selection of well chosen sensitivity analyses
- It allows
  - ▶ the uncertainty from the missing data to be taken into account
  - ▶ additional sources of information to be utilised
- It can be implemented using currently available software, e.g. WinBUGS

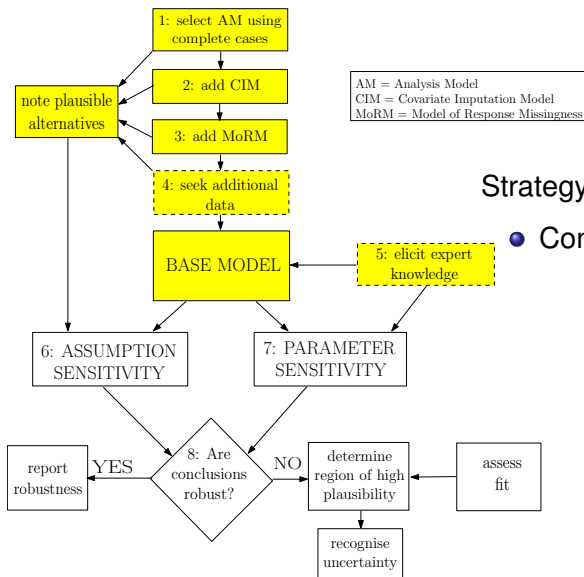
# Advantages of a Bayesian framework

- Bayesian models are formulated in a modular way
  - ▶ ideal for iteratively building and modifying a base model
- Uncertainty about the imputed missing values is
  - ▶ automatically and coherently propagated through the model
  - ▶ reflected in the estimates of interest
- Provides scope for including extra data or other information through
  - ▶ additional submodels
  - ▶ informative priors

# Schematic Diagram



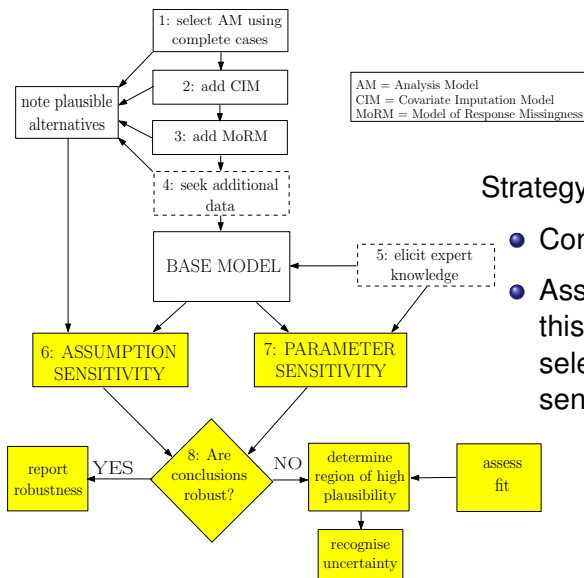
# Schematic Diagram: constructing a base model



Strategy consists of two parts:

- Constructing a base model

# Schematic Diagram: sensitivity analysis



Strategy consists of two parts:

- Constructing a base model
- Assessing conclusions from this base model against a selection of well chosen sensitivity analyses

## Before the strategy: step 0

- The strategy consists of a series of model building steps
- Before starting, the missingness should be explored to determine
  - ▶ which steps are required?
  - ▶ are any other modifications needed?
- In particular
  - ▶ which variables have missing values?
  - ▶ what is the extent and pattern of missingness?
  - ▶ think about plausible explanations for the missingness?

## Illustrative example: step 0

- For simplification, we restrict our MCS dataset to individuals fully observed in sweep 1
- Dataset has 505 individuals, with 37% missing response (pay) and 33% missing covariates

sweep 2 data

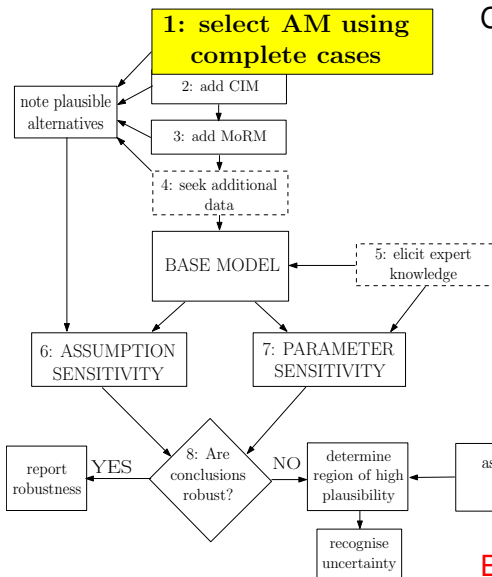
		covariates	
		observed	missing
pay	observed	320	0
	missing	19	166

- Survey methodology literature has shown that income non-response is usually non-ignorable

## Part 1 (steps 1-5): constructing a base model

- This part involves building a joint model as follows:
  - 1 choose an analysis model
  - 2 add a covariate imputation model
  - 3 add a model of response missingness
- Optionally, the amount of available information can be increased by incorporating data from other sources and/or expert knowledge
- The strategy
  - ▶ allows informative missingness in the response
  - ▶ but assumes that the covariates are MAR
- However, it can be adapted to reflect alternative assumptions

# Step 1 - analysis model



## Constructing a base model

- Form an initial Analysis Model (AM) based on
  - ▶ complete cases
  - ▶ previous knowledge
- Includes choosing
  - ▶ transform for the response
  - ▶ model structure
  - ▶ set of explanatory variables
- Allow for
  - ▶ hierarchical structure
  - ▶ other data complexities

**Error distribution is key assumption**

## Illustrative example: step 1 - AM description

- Response: log of hourly net pay
- 4 explanatory variables:

Name	Description	Details
<i>age</i>	age at interview	continuous <sup>a</sup>
<i>reg</i>	region of country	1 = London; 2 = other
<i>sing</i> <sup>b</sup>	single/partner	1 = single; 2 = partner
<i>stratum</i>	country by ward type	9 levels <sup>c</sup>

<sup>a</sup> centred and standardised

<sup>b</sup> always 1 in sweep 1 by definition

<sup>c</sup> 3 strata for England ('advantaged', 'disadvantaged' and 'ethnic minority'); 2 strata for Wales, Scotland and Northern Ireland ('advantaged' and 'disadvantaged')

- Error distribution: t with 4 degrees of freedom ( $t_4$ )
  - ▶ provides robustness to outliers

# Illustrative example: step 1 - AM equations

**Alternative:  
cube root transform**

log of hourly pay (*hpay*)

**Alternative:  
normal errors**

robustness to outliers

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

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

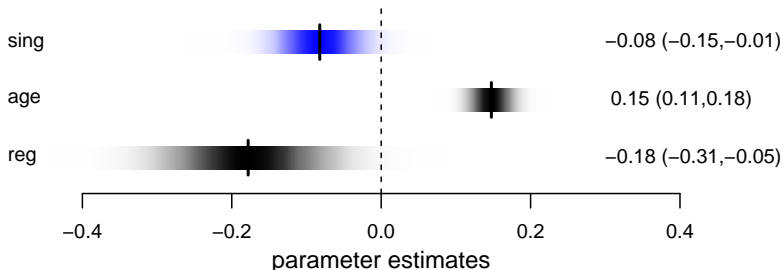
individual  
random effects  
stratum specific  
intercepts

*age* (mother's age)  
*reg* (London/other)  
*sing* (single/partner)

& vague priors e.g.  $\beta_k \sim N(0, 10000^2)$

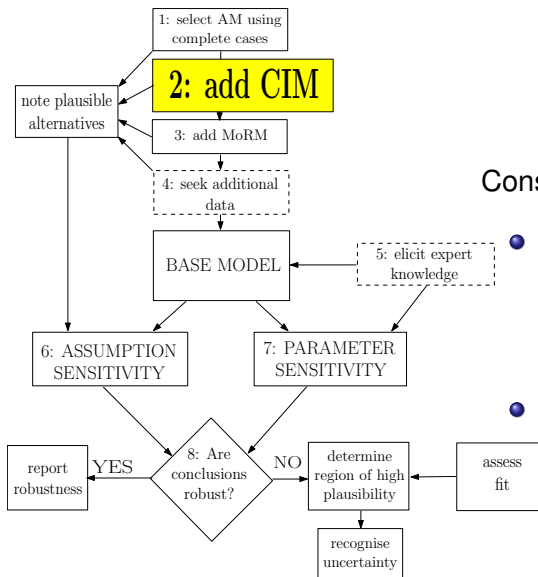
## Illustrative example: step 1 - AM results

- Lower pay is associated with
  - ▶ gaining a partner between sweeps
  - ▶ living outside London
- Higher pay is associated with
  - ▶ increasing age



Plot shows posterior distribution with median marked.  
Posterior mean and 95% interval are shown on the right.

## Step 2 - covariate imputation model



### Constructing a base model

- Add a Covariate Imputation Model (CIM) to produce realistic imputations of any missing covariates
- See Lecture 3 for details

## Illustrative example: step 2 - CIM description

- Assume covariates are missing at random (MAR)
- *stratum* does not change between sweeps
- Imputation of missing sweep 2 values required for *age*, *reg* and *sing*
- For simplicity, missing values of *age* and *reg*
  - ▶ set before analysis using simple rules
  - ▶ *reg*: assign sweep 1 value
  - ▶ *age*: sweep 1 *age* + mean of observed differences in *age* between sweeps
- For *sing* we set up a statistical model

## Illustrative example: step 2 - CIM equations

$$sing_{i2} \sim Bernoulli(q_i)$$

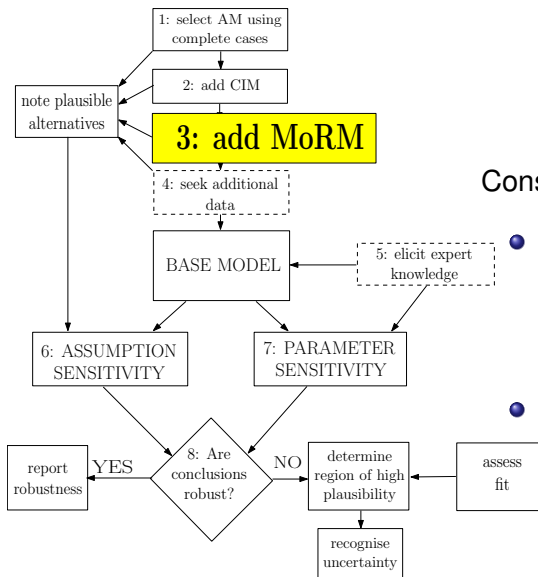
$$q_i = \phi_{s(i)} + (\phi_{age} \times age_{i2}) + (\phi_{reg} \times reg_{i2})$$

stratum specific intercepts

& vague priors e.g.  $\phi_k \sim N(0, 10000^2)$

- As a reasonableness check, we compare the distribution of the imputed and observed values of  $sing_2$ 
  - ▶ observed values: 34% gained a partner
  - ▶ imputed values: 36% (27%,45%) gained a partner
- CIM could be expanded to include  $age$  and  $reg$  (see Lecture 3)

## Step 3 - model of response missingness



### Constructing a base model

- Add a Model of Response Missingness (MoRM) to allow informative missingness in the response
- See Lecture 2 for details

## Illustrative example: step 3 - MoRM description

- Allow informative missingness in the response by modelling a missing value indicator ( $m_i$ ) for sweep 2 pay ( $hpay_{i2}$ ) s.t.

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

- Use a logit model for response missingness, i.e.

$$m_i \sim \text{Bernoulli}(p_i); \text{ logit}(p_i) = ?$$

- Previous work in this area informs choice of predictors of missing income
- For simplicity, linear relationships are assumed
- Untransformed hourly pay used in this sub-model

## Illustrative example: step 3 - MoRM equations

missing value indicator  
for sweep 2 hourly pay

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

$$\text{logit}(p_i) = \theta_0 + \sum_{k=1}^r \theta_k w_{ki} + \kappa \times \text{hpay}_{i1} + \delta \times (\text{hpay}_{i2} - \text{hpay}_{i1})$$

*sc* (social class)  
*eth* (ethnic group)  
*ctry* (country)

probability of sweep 2  
hourly pay being missing

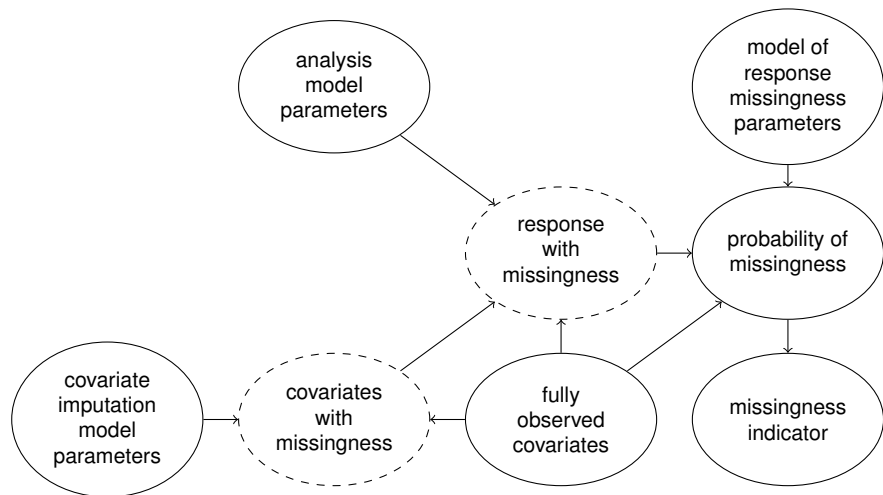
level of pay at sweep 1

change in pay  
between sweeps

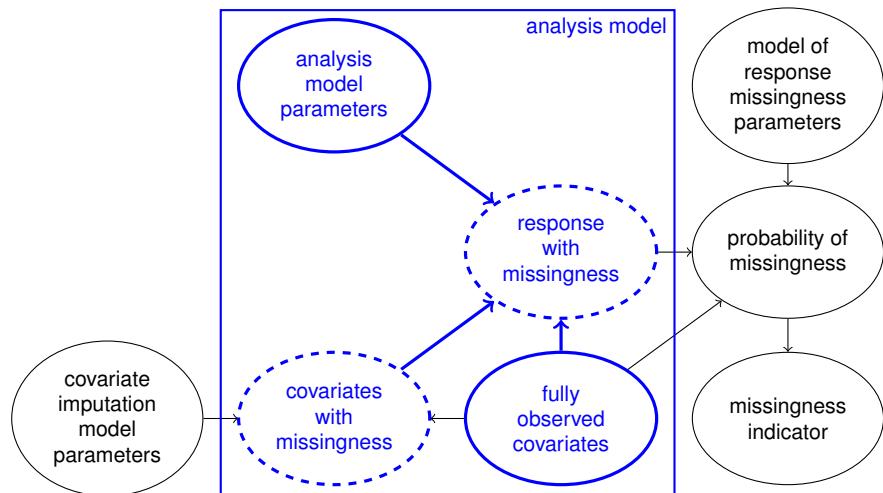
& vague priors

- The inclusion of the term  $\delta \times (\text{hpay}_{i2} - \text{hpay}_{i1})$  allows the response missingness to be MNAR
- If  $\delta = 0$ , then we have MAR missingness

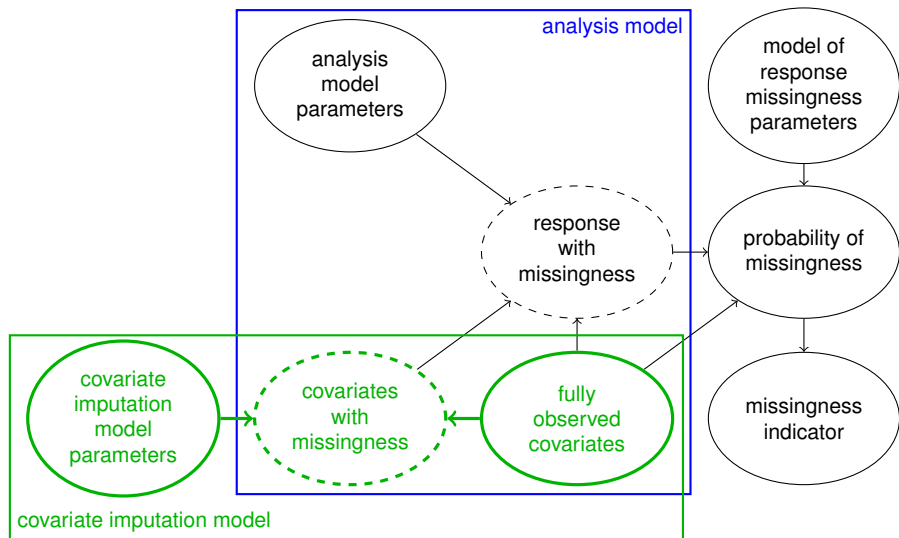
# The joint model (steps 1-3): schematic diagram



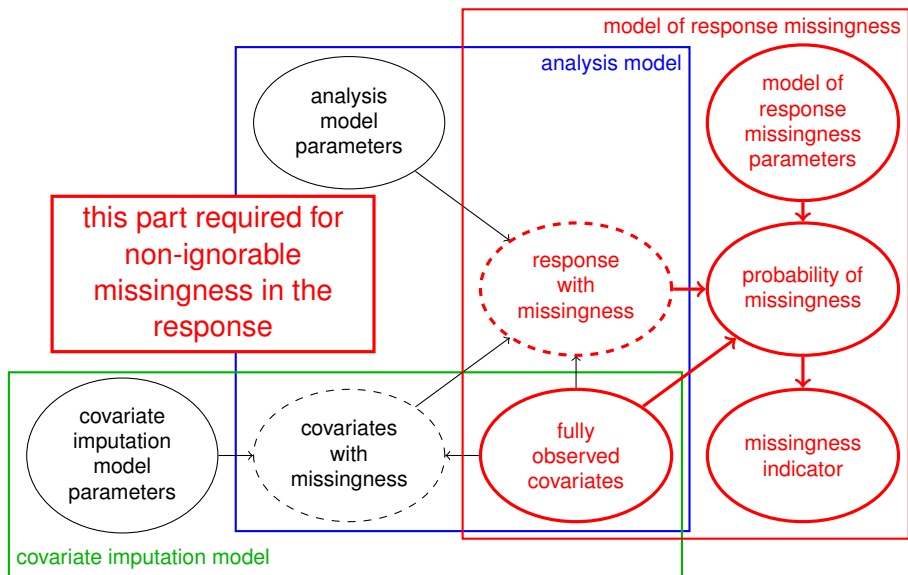
# The joint model (steps 1-3): schematic diagram



# The joint model (steps 1-3): schematic diagram

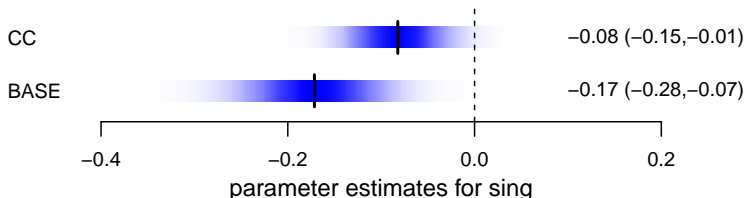


# The joint model (steps 1-3): schematic diagram

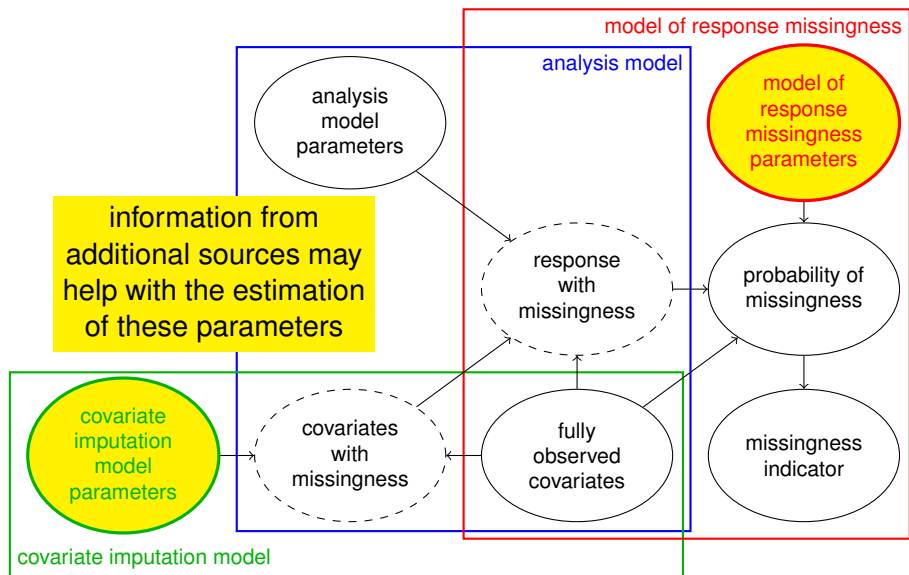


## Illustrative example: results from base model

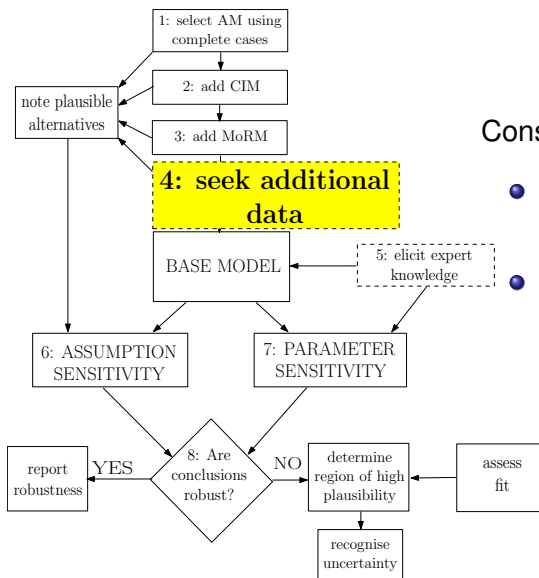
- A greater proportion of individuals are imputed to gain a partner
  - ▶ observed values: 34%
  - ▶ CC imputed values: 36% (27%,45%)
  - ▶ BASE imputed values: 44% (33%,57%)
- Individuals whose pay decreases substantially between sweeps are more likely to be missing
  - ▶  $\delta$ : -0.43 (-0.76,-0.13)
- The evidence that gaining a partner is associated with lower pay has strengthened



# The joint model (steps 1-3): schematic diagram



## Step 4 - seek additional data



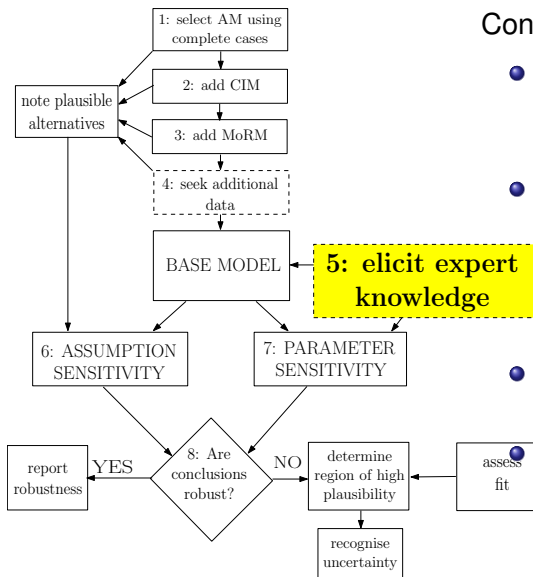
### Constructing a base model

- Additional data can help with parameter estimation
- Possible sources include
  - ▶ earlier/later sweeps of longitudinal study not under investigation
  - ▶ another study on individuals with similar characteristics

## Example: step 4 - incorporating additional data

- Seek another study with individuals with similar characteristics which includes variables of interest
- Expand CIM to simultaneously model data from original study and additional study by
  - ▶ fitting 2 sets of equations with common coefficients
  - ▶ 1 set for imputing the missing covariates in the original study
  - ▶ 1 set for modelling the data from the additional study
- The extra data allows the parameters in the CIM to be estimated with greater accuracy

# Step 5 - elicit expert knowledge



## Constructing a base model

- Expert knowledge can be elicited and incorporated using informative priors
- Focus on parameters not well identified by the data
  - ▶ particularly those associated with the degree of departure from MAR
- Eliciting priors on parameters directly is difficult

A better strategy is

- ▶ elicit information about the probability of response
- ▶ convert to informative priors

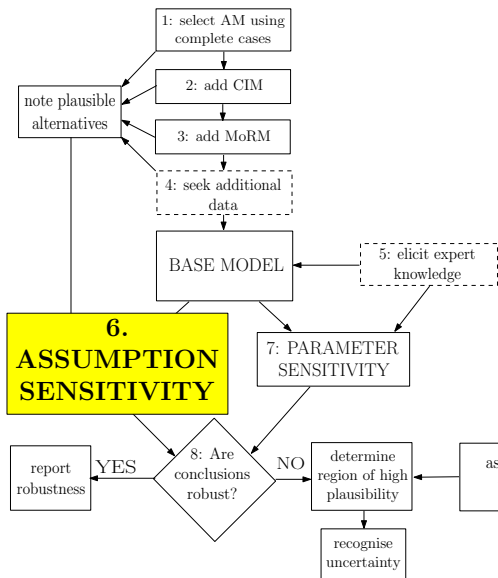
## Example: step 5 - elicit expert knowledge

- The key variables in the MoRM are:
  - ▶ the level of income
  - ▶ the change in income
- Consult survey methodology experts about how these variables are likely to influence the probability of non-response
- In particular, review the assumption of linear relationships
- Individuals may be less inclined to disclose their income if
  - ▶ the level is either low or high
  - ▶ it has changed substantially in either direction

## Part 2 (steps 6-8): sensitivity analysis

- Sensitivity analysis is essential because assumptions are untestable from the data
- There are many possible options, and the appropriate choice is problem dependent
- We propose two types of sensitivity analysis:
  - 1 an assumption sensitivity
  - 2 a parameter sensitivity

# Step 6 - assumption sensitivity

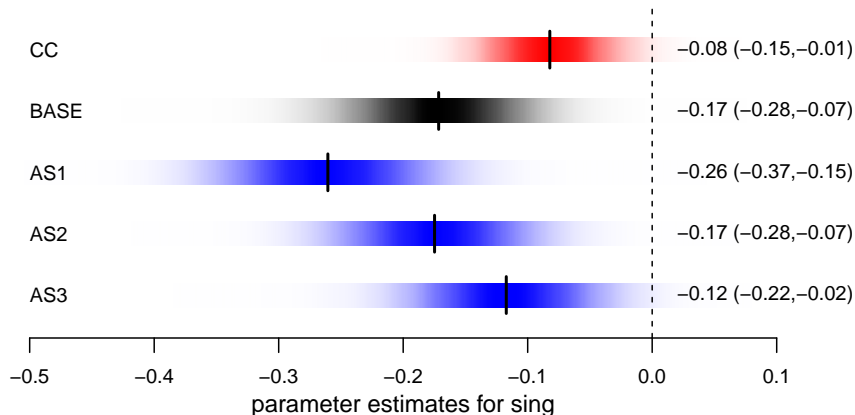


- Assumption sensitivity forms alternative models by changing the assumptions in the different base sub-models
- Key assumptions include:
  - ▶ AM error distribution
  - ▶ transformation of the AM response
  - ▶ functional form of the MoRM
- Could also vary explanatory variables
- Stage 1: change single aspect to assess effect
- Stage 2: combine several changes

# Illustrative example: step 6 - assumption sensitivity

- BASE CASE key features:
  - ▶ AM:  $t_4$  error distribution
  - ▶ AM: covariates  $\{age, reg, sing\}$
  - ▶ AM: log transform of the response
  - ▶ MoRM: linear functional form for *level* and *change*
- Examples of assumption sensitivity analysis options (differences from BASE CASE):
  - ▶ AM: normal error distribution
  - ▶ AM: additional covariate  $age^2$
  - ▶ AM: cube root transform of response
  - ▶ MoRM: piecewise linear functional form for *level* and *change* in hourly pay

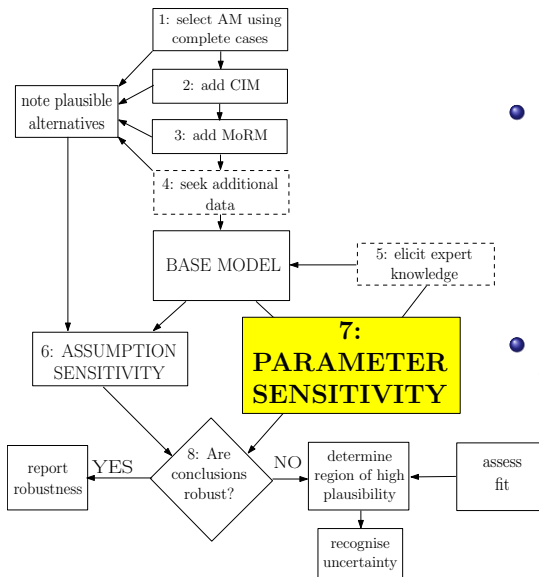
## Example: results from assumption sensitivity



AS1: normal error distribution; AS2: additional covariate;  
AS3: piecewise linear functional form for MoRM

- The amount of evidence that gaining a partner is associated with lower pay is very sensitive to different assumptions

## Step 7 - parameter sensitivity



- Parameter sensitivity involves running the base model with the MoRM parameters controlling the extent of the departure from MAR fixed to values in a plausible range
- Expert knowledge can help with setting the parameter sensitivity range

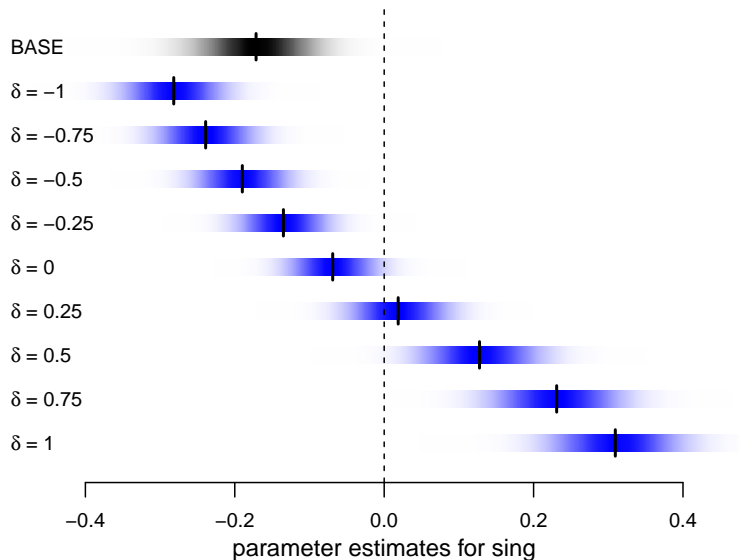
## Illustrative example: step 7 - parameter sensitivity

Recall MoRM equation for Base Case

$$\text{logit}(p_i) = \theta_0 + \sum_{k=1}^r \theta_k w_{ki} + \kappa \times \text{hpay}_{i1} + \delta \times (\text{hpay}_{i2} - \text{hpay}_{i1})$$

- The value of  $\delta$  controls the degree of departure from MAR missingness
- $\delta$  is difficult to estimate for a model with vague prior
- Run a series of models with  $\delta$  fixed using point prior
  - ▶  $\delta$  corresponds to the log odds ratio of a missing response per £1 increase in hourly pay
  - ▶ values outside the -1 to 1 range are intuitively implausible as the probability of response changes from 1 to 0 abruptly
  - ▶ 9 variants: values  $\{-1, -0.75, -0.5, -0.25, 0, 0.25, 0.5, 0.75, 1\}$
  - ▶  $\delta = 0$  variant is equivalent to assuming the response is MAR

# Example: results from parameter sensitivity



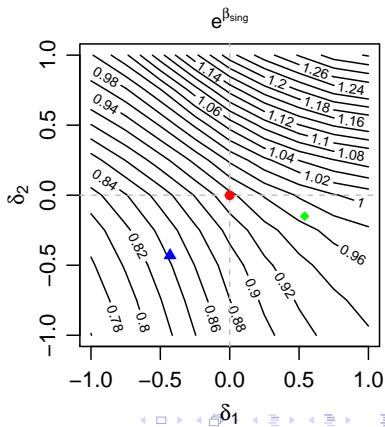
Even the direction of the effect is uncertain

## Illustrative example: parameter sensitivity continued

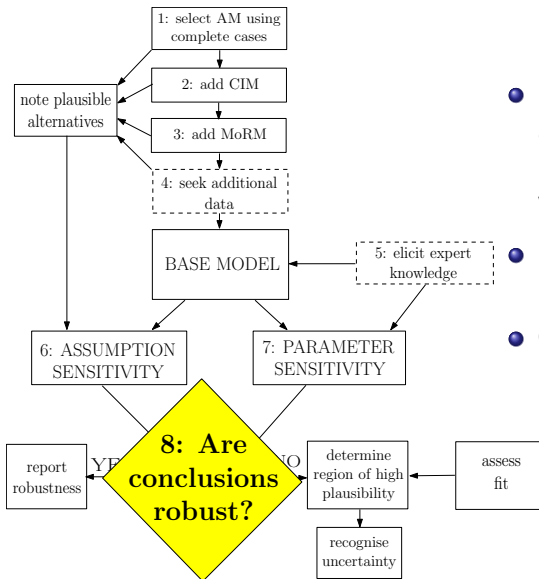
- Suppose a piecewise linear functional form (AS3) was considered more plausible, e.g.

$$\text{logit}(p_i) = \begin{cases} \dots + \delta_1 \text{change}_i + \dots & : \text{change}_i < 0 \\ \dots + \delta_2 \text{change}_i + \dots & : \text{change}_i \geq 0 \end{cases}$$

- Run a parameter sensitivity with both  $\delta_1$  and  $\delta_2$  fixed
- Graph shows proportional change in pay associated with gaining a partner
- Results very sensitive to chosen values of  $\delta_1$  and  $\delta_2$
- red circle: MAR ( $\delta_1 = \delta_2 = 0$ )  
blue triangle: Base Case  
green diamond: AS3



# Step 8 - determine robustness of conclusions



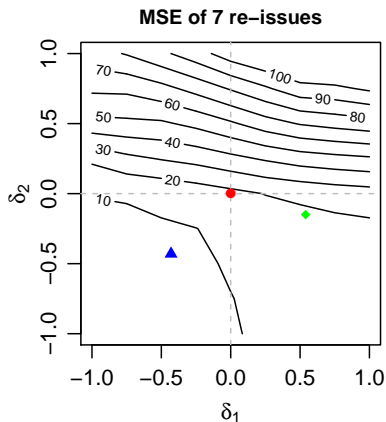
- Examine results of sensitivity analyses to establish how much the quantities of interest vary
- If the conclusions are robust, report this
- Otherwise
  - ▶ seek more information (Steps 4 & 5)
  - ▶ determine a region of high plausibility
  - ▶ recognise uncertainty

# Assessing model fit

- A model's fit to observed data can be assessed
- However, it's fit to unobserved data **cannot** be assessed
  - ⇒ **sensitivity analysis is essential**
- DIC is routinely used by Bayesian statisticians to compare models, but
  - ▶ using DIC in the presence of missing data is not straightforward
  - ▶ the DIC automatically generated by WinBUGS is misleading (Mason et al., 2012a)
- Data not used in model estimation may be helpful in assessing model fit

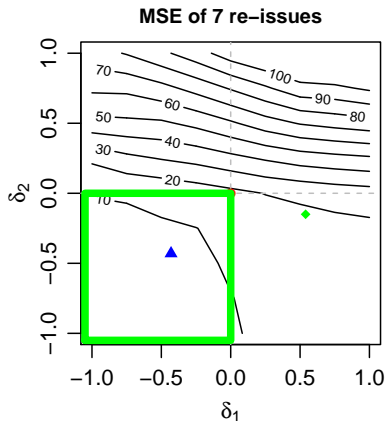
## Illustrative example: assessing model fit

- Data collected from 7 individuals who were originally non-contacts or refusals in sweep 2, after they were re-issued by the fieldwork agency
- Set these data to missing before fitting models, and use for model assessment
- Calculate the mean square error (MSE) of the fit of hourly pay for the 7 re-issues as a summary measure of model performance
- red circle: MAR ( $\delta_1 = \delta_2 = 0$ )  
blue triangle: Base Case  
green diamond: AS3



## Illustrative example: reporting uncertainty

- The MSE plot helps identify a region of high plausibility
- Key points to report are:
  - ▶ There is evidence that gaining a partner is associated with a decrease in hourly pay
  - ▶ However, the magnitude of this decrease is uncertain
  - ▶ Our analysis suggests that the proportional decrease lies in the region 0.76 (0.69,0.82) to 0.93 (0.87,1.00)



- Some models run as part of the 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

# Adaptions and extensions

- There are situations where it may be necessary to adapt this strategy
- Step 2 can be elaborated to allow MNAR covariates
- Steps 3 and 7 can be omitted if informative missingness in the response is implausible
- Could distinguish between different types of non-response
  - ▶ set up a missingness indicator with separate categories for each type of non-response
  - ▶ model using multinomial regression
- Bayesian models have the advantage of being fully coherent, but with large datasets or large numbers of covariates with missingness may be computationally challenging to fit
- In the next lecture, we consider alternative two stage approaches

# Lecture 5.

## Comparison with alternative methods

# Lecture Outline

- In Lectures 1-4 we discussed how to construct a Fully Bayesian Model (FBM) in the presence of missing data
- However, there are circumstances where implementation is difficult in practice
- In this lecture, we compare FBM with alternative methods
  - ▶ review standard multiple imputation
  - ▶ compare performance in practice between the ends of the multiple imputation 'spectrum'
  - ▶ revisit the MCAR/MAR/MNAR paradigm
  - ▶ discuss future directions of missing data research

# Standard Multiple Imputation (MI)

- FBM is one of a number of 'statistically principled' methods for dealing with missing data (Lecture 1)
- Of the alternatives, standard Multiple Imputation is closest in spirit and has a Bayesian justification
- Multiple imputation was developed by Rubin (1996)
  - ▶ Most widely used 'principled' method for handling missing data
  - ▶ Usually assumes missingness mechanism is MAR (can be used for MNAR but more tricky)
  - ▶ Most useful for handling missing **covariates**

# The 3 steps of Multiple Imputation

## 1 Imputation

- ▶ Impute (=fill in) the missing entries of the incomplete data sets, not once, but  $K$  (typically 5-10) times
- ▶ Imputed values are drawn from a distribution (that can be different for each missing variable)
- ▶ This step results in  $K$  complete data sets

## 2 Analysis

- ▶ Analyse each of the  $K$  completed data sets using standard methods

## 3 Pooling

- ▶ Combine the  $K$  analysis results into a single final result
- ▶ Simple rules exist for combining the  $K$  analyses to produce estimates and confidence intervals that incorporate uncertainty about the missing data

# Step 1: Specifying the imputation model

- This is the hardest step of the MI procedure
- Need to specify model for full data  $f(\mathbf{z}^{obs}, \mathbf{z}^{mis} | \beta)$  and then generate values for  $\mathbf{z}^{mis} | \mathbf{z}^{obs}, \beta$
- If only one variable (say  $z_J \in \mathbf{z}$ ) has missing values, this usually amounts to specifying a regression model with  $\mathbf{z}_J$  as the outcome and  $\mathbf{z}_{\setminus J}$  as the predictors
  - ▶ Under MAR, can learn about relationship between  $z_J$  and  $\mathbf{z}_{\setminus J}$  from the cases where  $z_J$  is observed (i.e.  $z_J^{obs}$ ), and then predict the missing cases,  $z_J^{mis}$ , from the fitted regression model
- More tricky when many variables have missing values - there are two main approaches
  - ▶ based on a joint multivariate normal distribution
  - ▶ iterate a series of conditional univariate models

# Step 1: Multivariate regression v 'chained equations'

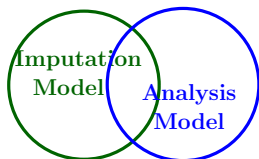
- Multivariate regression techniques
  - ▶ Theoretically sound
  - ▶ Generate multiple imputations assuming an underlying multivariate normal distribution
  - ▶ Parallels with specifying the CIM in FBM, **but must include the response**
- Multivariate imputation by 'chained equations' (MICE)
  - ▶ Specifies separate conditional distributions for the missing data in each incomplete variable
  - ▶ Does not necessarily correspond to a genuine joint distribution
  - ▶ Issue of incompatible conditionals
  - ▶ But does provide a flexible way of implementing complex imputation models
- Both approaches risk a mismatch between the imputation model and the analysis model (often referred to as the problem of 'congeniality')

# Step 1: Selecting predictors for imputation model

- It is generally recommended to include variables that:
  - ▶ appear in the main analysis model, **including the response**
  - ▶ are associated with the missingness mechanism
  - ▶ explain a considerable amount of variance in the variables to be imputed
- Better to err on the side of caution and include predictors rather than leave them out
- Imputation model
  - ▶ typically contains richer set of variables/predictors than are of interest in main model
  - ▶ should reflect any complexity in the analysis model (e.g. non-linear terms, interactions, hierarchical structure)

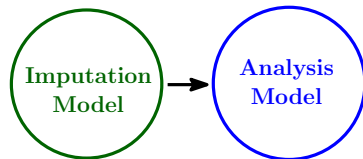
# Comparison of FBM and MICE

## FBM



- 1 stage procedure
  - ▶ fit Imputation and Analysis Models simultaneously
- imputation model uses joint distribution of all missing variables
- uses full posterior distribution of missing values

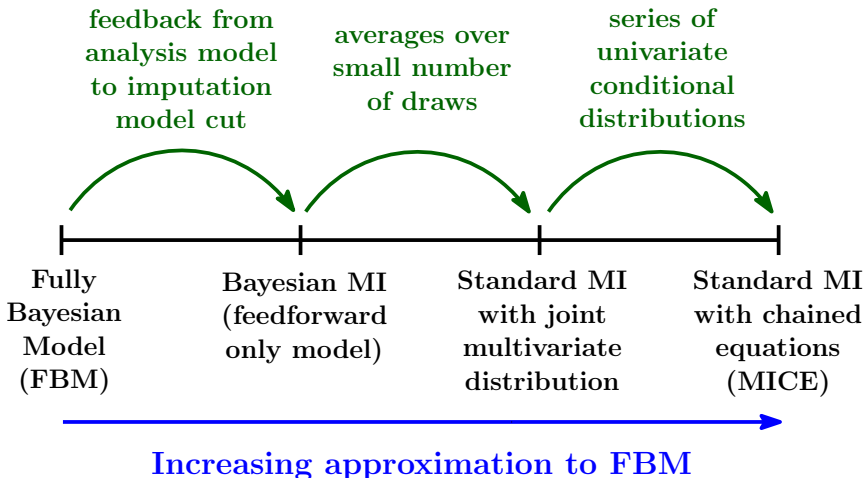
## MICE



- 2 stage procedure
  - 1 fit Imputation Model
  - 2 fit Analysis Model
- imputation model based on a set of univariate conditional distributions
- uses small number of draws of missing values from their predictive distribution

# Multiple Imputation (MI) spectrum

FBM and MICE can be considered the extremes of a MI spectrum



# Pros and cons of FBM and MICE

- FBM - pros
  - ▶ theoretically sound
  - ▶ coherent model estimation
  - ▶ uncertainty fully propagated
  - ▶ can add missingness model to explore informative missingness
- FBM - cons
  - ▶ implementation can be challenging
- MICE - pros
  - ▶ range of readily available packages
  - ▶ speed
- MICE - cons
  - ▶ conditional distributions may not correspond to a joint distribution
  - ▶ difficult to explore informative missingness
- But how do FBM and MICE actually perform in practice?

We investigate using simulations with missing covariates

# General setup of simulations

- Generate 1000 simulated data sets with
  - ▶ 2 correlated explanatory variables,  $x$  and  $u$
  - ▶ response,  $y$ , dependent on  $x$  and  $u$
  - ▶  $\approx 50\%$  or  $90\%$  missingness imposed on  $u$  dependent on  $y$
- Each simulated dataset analysed by a series of models
- Performance of models assessed for
  - ▶ coefficient for  $u$ ,  $\beta_u$ , (true value=-2)
  - ▶ coefficient for  $x$ ,  $\beta_x$ , (true value=1)
- We report
  - ▶ average estimate (average of the posterior means)
  - ▶ bias (average estimate - true value)
  - ▶ coverage rate (proportion of times true value is contained in the 95% interval)
  - ▶ interval width (average width of 95% interval)

# Simulation setup: model descriptions

- We run 5 types of models
  - ▶ GOLD: correct analysis model run on complete datasets
  - ▶ EXU: excludes  $u$  from analysis model
  - ▶ CC: complete case analysis
  - ▶ FBM: Fully Bayesian Model (analysis and imputation models)
  - ▶ MICE: uses 20 imputations
- GOLD provides performance targets
- GOLD, EXU, CC, FBM all fitted using WinBUGS software
- MICE fitted using functions from `mice` package in R software
- FBM and MICE have several variants dependent on scenario
- All models have the 'correct' analysis model

# Non-hierarchical linear simulation

- Data generated
  - ▶ all variables continuous
  - ▶ no hierarchical structure
  - ▶ 1000 individuals
  - ▶  $\approx 90\%$  missingness
- Note: with single covariate with missing values, the approximation of using chained equations disappears
- Full data generated as follows

$$\begin{pmatrix} x \\ u \end{pmatrix} \sim MVN \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right)$$
$$y \sim N(1 + x - 2u, 4^2)$$

MAR missingness imposed on  $u_i$  with probability  $p_i$

$$\text{logit}(p_i) = -7 + y_i$$

## Non-hierarchical linear simulation - results

		average estimate	bias	coverage rate	interval width
$\beta_x$	GOLD	1.00	0.00	0.95	0.58
$\beta_x$	EXU	0.00	-1.00	0.00	0.54
$\beta_x$	CC	0.33	-0.67	0.34	1.11
$\beta_x$	FBM	0.92	-0.08	0.96	1.51
$\beta_x$	MICE	0.92	-0.08	0.95	1.56
$\beta_u$	GOLD	-2.00	0.00	0.94	0.56
$\beta_u$	CC	-0.69	1.31	0.01	1.13
$\beta_u$	FBM	-1.80	0.20	0.95	2.49
$\beta_u$	MICE	-1.82	0.18	0.93	2.63

- EXU: extreme bias and 0 coverage ( $\beta_x$  only)
- CC: serious bias and very low coverage
- FBM & MICE: correct most of bias and achieve nominal coverage

# Non-hierarchical linear simulation - conclusion

- Findings extend to multiple covariates with missingness
- Even with extreme levels of missingness, FBM and MICE have similar performance with non-complex data
- We now look at two types of complexity
  - ▶ hierarchical structure
  - ▶ informative missingness

# Hierarchical linear simulation - description

- Data generated
  - ▶ with hierarchical structure (individuals within clusters)
  - ▶ 10 clusters, each with 100 individuals
  - ▶  $\approx 50\%$  missingness
- FBM models: 3 variants with different imputation models for  $u$ 
  - ▶ no hierarchical structure (no HS)
  - ▶ random intercepts (HS:  $r_i$ )
  - ▶ random intercepts + random slopes on  $x$  (HS:  $r_i + r_s$ )
- MICE: no hierarchical structure in imputation model
  - ▶ in theory could run variants with hierarchical structure
  - ▶ but implementation difficulties

# Hierarchical linear simulation - equations

Generate full data set as follows:

$$\begin{pmatrix} x_c \\ u_c \\ \alpha_c \end{pmatrix} \sim MVN \left( \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 2 & 0.5 & 0.5 \\ 0.5 & 2 & 0.5 \\ 0.5 & 0.5 & 4 \end{pmatrix} \right)$$
$$\begin{pmatrix} x_i \\ u_i \end{pmatrix} \sim MVN \left( \begin{pmatrix} x_c \\ u_c \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right)$$
$$y_i \sim N(\alpha_c + x_i - 2u_i, 1)$$

$c$  indicates cluster level data;  $i$  indicates individual level data

Impose missingness such that  $u_i$  is missing with probability  $p_i$

$$\text{logit}(p_i) = -0.5 + 0.5y_i$$

## Hierarchical linear simulation - $\beta_U$ results

	average estimate	bias	coverage rate	interval width
GOLD	-2.00	0.00	0.93	0.14
CC	-1.92	0.08	0.70	0.21
FBM (no HS)	-1.93	0.07	0.67	0.19
FBM (HS: ri)	-2.00	0.00	0.94	0.19
FBM (HS: ri+rs)	-2.00	0.00	0.94	0.19
MICE (no HS)	-1.36	0.64	0.00	0.33

If hierarchical structure ignored in imputation model

- FBM - slight bias and poor coverage
- MICE - much worse (no feedback from structure in analysis model)

## Hierarchical linear simulation - $\beta_U$ results

	average estimate	bias	coverage rate	interval width
GOLD	-2.00	0.00	0.93	0.14
CC	-1.92	0.08	0.70	0.21
FBM (no HS)	-1.93	0.07	0.67	0.19
FBM (HS: ri)	-2.00	0.00	0.94	0.19
FBM (HS: ri+rs)	-2.00	0.00	0.94	0.19
MICE (no HS)	-1.36	0.64	0.00	0.33

If hierarchical structure incorporated in imputation model

- bias corrected
- nominal coverage rate achieved

## Hierarchical linear simulation - $\beta_x$ results

	average estimate	bias	coverage rate	interval width
GOLD	1.00	-0.00	0.94	0.14
EXU	0.00	-1.00	0.00	0.25
CC	0.96	-0.04	0.89	0.20
FBM (no HS)	0.85	-0.15	0.21	0.19
FBM (HS: ri)	0.99	-0.01	0.94	0.19
FBM (HS: ri+rs)	0.99	-0.01	0.94	0.19
MICE (no HS)	0.53	-0.47	0.01	0.26

Pattern of bias and coverage results similar to  $\beta_u$

# V-shaped informative missingness - description

- Data generated
  - ▶ with no hierarchical structure
  - ▶ 100 individuals
  - ▶ missingness imposed on  $u$  depends on  $y$  and  $u$
  - ▶  $\approx 50\%$  missingness
- FBM models: 4 variants
  - ▶ MAR: no model of covariate missingness
  - ▶ MNAR: assumes linear shape (linear)
  - ▶ MNAR: allows v-shape (v-shape)
  - ▶ MNAR: allows v-shape + priors inform signs of slopes (v-shape+)
- MICE: MAR, no model of covariate missingness
  - ▶ most implementations do not readily extend to MNAR
  - ▶ ad hoc sensitivity analysis to MNAR possible by inflating or deflating imputations (van Buuren and Groothuis-Oudshoorn, 2011)

# V-shaped informative missingness - equations

Generate full data set as follows:

$$\begin{pmatrix} x \\ u \end{pmatrix} \sim MVN \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \right)$$
$$y \sim N(1 + x - 2u, 4^2)$$

Impose missingness such that  $u$  is missing with probability  $p$

$$\text{logit}(p) = -2 + 2|u| + 0.5y$$

## V-shaped informative missingness - $\beta_U$ results

	average estimate	bias	coverage rate	interval width
GOLD	-1.99	0.01	0.95	1.68
CC	-1.66	0.34	0.92	2.63
MAR: FBM	-2.25	-0.25	0.93	3.18
MNAR: FBM (linear)	-2.08	-0.08	0.97	3.76
MNAR: FBM (vshape)	-2.06	-0.06	0.96	3.49
MNAR: FBM (vshape+)	-2.02	-0.02	0.96	3.31
MAR: MICE	-2.25	-0.25	0.90	3.33

- MAR results in bias and slightly reduced coverage
- improvements if allow MNAR, even if wrong form
- further improvements from correct form
- and even better with informative priors

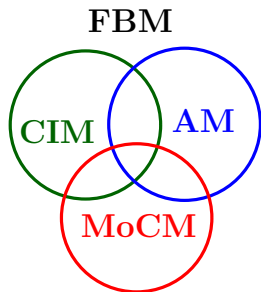
## V-shaped informative missingness - $\beta_x$ results

	average estimate	bias	coverage rate	interval width
GOLD	0.99	-0.01	0.94	1.65
EXU	0.51	-0.49	0.83	1.81
CC	0.70	-0.30	0.91	2.06
MAR: FBM	0.87	-0.13	0.94	1.85
MNAR: FBM (linear)	0.83	-0.17	0.94	1.89
MNAR: FBM (vshape)	0.87	-0.13	0.95	1.91
MNAR: FBM (vshape+)	0.89	-0.11	0.94	1.93
MAR: MICE	0.87	-0.13	0.94	1.88

- MAR results in modest bias (FBM and MICE)
- wrong MNAR (linear) slightly worse than MAR
- little gain in correct MNAR over MAR

# Summary of simulation results

- hierarchical structure
  - ▶  $\beta_U$  and  $\beta_X$  - clear benefits from incorporating structure in FBM imputation model (unable to assess MICE)
- informative missingness
  - ▶  $\beta_U$  - benefits from correct MNAR
  - ▶  $\beta_X$  - no clear benefits from allowing MNAR



AM = Analysis Model

CIM = Covariate Imputation Model

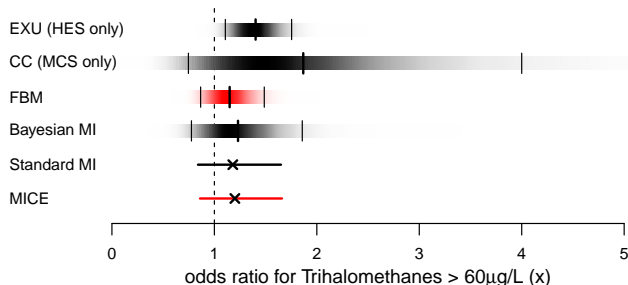
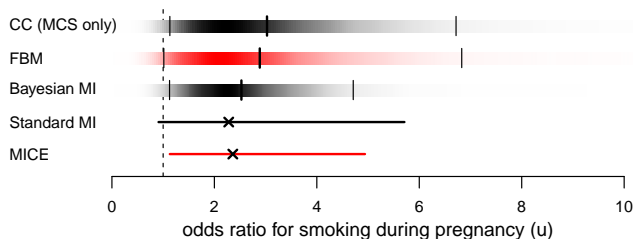
MoCM = Model of Covariate Missingness

With FBM for informative missingness,  
3 linked models fitted simultaneously

## LBW example: low birth weight

- The LBW example was introduced in Lecture 3
- Objective: estimate the association between trihalomethane concentrations (THM) and the risk of full term low birth weight
- Key variables are:
  - $Y$ : binary indicator of low birth weight (outcome)
  - $X$ : binary indicator of THM concentrations (exposure of interest)
  - $U$ : smoking and ethnicity (confounders, over 90% of values missing)
- We compare estimates of odds ratios using models from different points on the MI spectrum
  - ▶ FBM: WinBUGS
  - ▶ Bayesian MI: WinBUGS using *cut* function
  - ▶ Standard MI: REALCOM-IMPUTE/MLwiN
  - ▶ MICE: `mice` package in R

# LBW example: comparison of imputation strategies




Standard MI and MICE: point estimate and 95% confidence interval

Other: posterior distribution (tick marks indicate 2.5% quantile, mean & 97.5% quantile)

# Software for MI

- Many programs now include routines for carrying out MI using a set of univariate conditional distributions (MICE)
  - ▶ R package `mice` (Stef van Buuren and Karin Groothuis-Oudshoorn)
  - ▶ R package `mi` (Masanao Yajima, Yu-Sung Su, M. Grazia Pittau and Andrew Gelman)
  - ▶ Stata
  - ▶ SAS
  - ▶ SPSS
- Implementations based on a joint multivariate distribution are less widely available
  - ▶ REALCOM-IMPUTE: REALCOM macros in conjunction with MLwiN
- Bayesian models can be implemented using BUGS
  - ▶ WinBUGS, OpenBUGS, JAGS

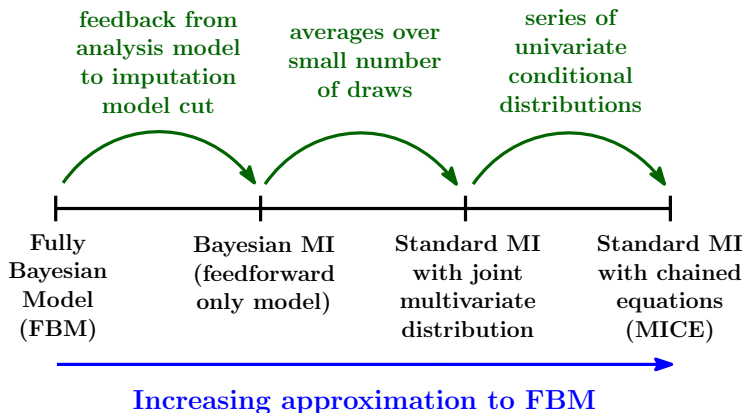
## Practical advice

	“simple”	complex
small dataset and few covariates with missingness	FBM MICE	FBM
large dataset and/or many covariates with missingness	MICE	

# Hybrid approaches

- Neither FBM or MICE are currently well suited to analysis of
  - 1 a large dataset and/or many covariates with missingness and
  - 2 complexity e.g. hierarchical structure or informative missingness
- In these circumstances a hybrid approach may provide a pragmatic alternative
- We need to choose a convenient point on the MI spectrum
- Starting from FBM, consider the following approximations
  - ▶ impute (multiple times) some covariates prior to fitting a Bayesian model
  - ▶ split CIM into several smaller sub-models, based on level of correlation

# Direction for missing data research



- Where should our starting point be?
  - ▶ FBM: improve computational efficiency, more case studies
  - ▶ MICE: robust in simple setup, but more work needed for realistic situations

# Concluding Remarks

# MCAR/MAR/MNAR paradigm

- It is tempting to think that the missingness affecting a variable always slots neatly into one of Rubin's 3 categories
- From a practical viewpoint the reality is more complex
  - ▶ MNAR can take different forms and depart from MAR by different amounts
  - ▶ it depends on what else has been observed!
- If enough additional information is collected, MNAR missingness can be converted (close) to MAR missingness
  - ▶ often used as justification for methods which assume MAR
- However, the MAR or MNAR classification will still be an assumption
  - ⇒ sensitivity analysis is required

**REMEMBER: It is not possible to distinguish between MAR and MNAR from the observed data alone**

# MCAR/MAR/MNAR paradigm and bias in CC (1)

- In Lecture 3 we saw that the conditions under which CC is biased with **missing covariates** cut across the MCAR/MAR/MNAR categorisation
- For the case of a single covariate,  $X$ , with missing values and fully observed  $Y$
- CC is unbiased if missingness in  $X$  is
  - ▶ MCAR
  - ▶ MNAR dependent on  $X$
- CC is biased if missingness in  $X$  is
  - ▶ MAR dependent on  $Y$
  - ▶ MNAR dependent on  $X$  and  $Y$

## MCAR/MAR/MNAR paradigm and bias in CC (2)

- For **missing responses**, distinguish between cross-sectional and longitudinal data
- Cross-sectional data
  - ▶ only MNAR causes bias
- Longitudinal data
  - ▶ if MAR missingness depends on data at previous time-points
  - ▶ excluding partially observed records for CC will convert to MNAR
  - ▶ results in bias
  - ▶ see difference in analysis of complete cases and all cases for HAMD example (Lecture 2)

**Missing data requires a lot of thought!**

# What does the Bayesian approach offer for missing data problems?

Bayesian methods are probably the most powerful and most general methods for dealing with missing data

- Naturally accommodate missing data without requiring new techniques for inference
- Bayesian framework is well suited to building complex models by linking smaller sub-models into a coherent joint model for the full data
- Bayesian approach lends itself naturally to sensitivity analysis through different choices of prior distributions encoding assumptions about the missing data process
- Offers possibility of including informative prior information about missing data process
- But models can become computationally challenging...

Thank you for your attention!

# Bayesian Missing Data Course

## References

- 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.
- Diggle, P. and Kenward, M. G. (1994). Informative Drop-out in Longitudinal Data Analysis (with discussion). *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **43**, (1), 49–93.
- Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, (2nd edn). John Wiley and Sons.
- Mason, A., Richardson, S., and Best, N. (2012a). Two-pronged strategy for using DIC to compare selection models with non-ignorable missing responses. *Bayesian Analysis*. to appear.
- Mason, A., Richardson, S., Plewis, I., and Best, N. (2012b). Strategy for modelling non-random missing data mechanisms in observational studies using Bayesian methods. *Journal of Official Statistics*. to appear.
- Mason, A. J. (2009). *Bayesian methods for modelling non-random missing data mechanisms in longitudinal studies*. PhD thesis, Imperial College London. available at [www.bias-project.org.uk](http://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*, **172**, (3), 615–37.
- Richardson, S. and Best, N. (2003). Bayesian hierarchical models in ecological studies of health-environment effects. *Environmetrics*, **14**, 129–47.
- Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Society*, **91**, (434), 473–89.
- Spiegelhalter, D. J. (1998). Bayesian Graphical Modelling: A Case-Study in Monitoring Health Outcomes. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **47**, (1), 115–33.
- Spiegelhalter, D. J., Thomas, A., and Best, N. G. (1996). Computation on Bayesian Graphical Models. In *Bayesian Statistics 5: Proceedings of the*

*Fifth Valencia International Meeting*, (ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith), pp. 407–25. Claredon Press.

Sterne, J. A. C., Carlin, J. B., Spratt, M., Kenward, M. G., Wood, A. M., and Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal*, **338**, b2393.

White, I. R. and Carlin, J. B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine*, **29**, 2920–31.

Wood, A. M., White, I. R., and Thompson, S. G. (2004). Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials*, **1**, (4), 368–76.