

# Bayesian approaches to handling missing data: Practical Exercises

## 1 Practical A

*Thanks to James Carpenter and Jonathan Bartlett who developed the exercise on which this practical is based (funded by ESRC).*

In this practical you will explore the three types of missingness MCAR, MAR and MNAR. It would be helpful to have a calculator handy for answering some of the questions.

Suppose that you wish to investigate the cognitive function of a group of 100 males aged 90. You devise a way of classifying their cognitive function into ‘good’, ‘moderate’ or ‘poor’ based on the scores attained in some standard test of cognitive function (*Cog90*). Further, you have available a classification of the same individuals into ‘high’ and ‘low’ cognitive function for a different test undertaken when they were aged 85 (*Cog85*). No other data is available to you.

Table 1 provides the data that you would observe if all the individuals take the test.

1. Use the data in Table 1 to answer the following questions, which are about marginal and conditional distributions.
  - (a) What is the overall mean cognitive function score at age 90 (*Cog90*)?
  - (b) The marginal distribution of *Cog85* is 72% in category ‘high’ and 28% in category ‘low’. What is the marginal distribution of *Cog90* (i.e. the percentage in each category of *Cog90*)?
  - (c) What is the conditional distribution of *Cog90* given ‘high’ *Cog85*? What is the conditional distribution of *Cog90* given ‘low’ *Cog85*? Are they different? Does this make sense?

Now suppose that not all the individuals take the test. Tables 2-4 provide three different realisations (A, B and C) of the data that you actually observe. Each realisation corresponds to a different type of missing data mechanism.

2. What does it mean to say
  - (a) Cognitive function at age 90 is Missing Completely At Random (MCAR)?
  - (b) Cognitive function at age 90 is Missing At Random (MAR)?
  - (c) Cognitive function at age 90 is Missing Not At Random (MNAR)?
3. For each of the three observed data tables (A, B, C):
  - (a) Estimate the mean cognitive function score at age 90 from the observed data. Is this an unbiased estimate?
  - (b) Is the marginal estimate of the cognitive function at age 90 distribution, calculated from the observed data, unbiased?
  - (c) Is the conditional estimate of the cognitive function score at age 90 given ‘high’ cognitive function at age 85, calculated from the observed data, unbiased?
  - (d) Is the conditional estimate of the cognitive function score at age 85 given ‘poor’ cognitive function at age 90, calculated from the observed data, unbiased?

- (e) Which type of missing data mechanism do you think produced the missing values (MCAR, MAR or MNAR). Hint: using the table of true values, calculate the proportion of missing data in each cell (i.e. the probability that an observation in that cell is missing).
4. For each of the three observed data tables, decide if you can obtain an unbiased estimate of the overall mean cognitive function score at age 90 from the observed data? If so, write down the formula. Hint: this may be a weighted average of the scores in different cells.

Table 1: Cognitive function of 100 males (true data)

Cognitive function at age 85	Cognitive function at age 90			age 90 margin
	good (mean = 20)	moderate (mean = 10)	poor (mean = 5)	
high	44	24	4	72
low	4	12	12	28
age 85 margin	48	36	16	100

Table 2: Cognitive function of 100 males (observed data A)

Cognitive function at age 85	Cognitive function at age 90			age 90 margin
	good (mean = 20)	moderate (mean = 10)	poor (mean = 5)	
high	33	12	1	46
low	3	6	3	12
age 85 margin	36	18	4	58

Table 3: Cognitive function of 100 males (observed data B)

Cognitive function at age 85	Cognitive function at age 90			age 90 margin
	good (mean = 20)	moderate (mean = 10)	poor (mean = 5)	
high	22	12	2	36
low	2	6	6	14
age 85 margin	24	18	8	50

Table 4: Cognitive function of 100 males (observed data C)

Cognitive function at age 85	Cognitive function at age 90			age 90 margin
	good (mean = 20)	moderate (mean = 10)	poor (mean = 5)	
high	33	18	3	54
low	1	3	3	7
age 85 margin	34	21	6	61

## 2 Practical B

In this practical you will construct a series of graphical models that represent the Bayesian models that you would build to answer the questions below.

*Materials:* You should have been given a graphical model toolkit consisting of:

- a magnetic white board
- yellow magnetic circles: use these to represent random variables
- orange magnetic circles: use these to represent random variables with missing values
- blue magnetic squares: use these to represent quantities that are regarded as constants, for example fully observed covariates
- coloured marker pens — use for labelling your nodes (you can write on the coloured circles/squares) and drawing arrows to link nodes together:
  - use the black pen to draw stochastic links to which a probability relationship is attached, i.e. any relationship between two variables specified by “ $\sim$ ” in an equation;
  - use the red pen to draw deterministic relationships, i.e. any relationship between two variables specified by “ $=$ ” in an equation.
- a dry-wipe eraser

*Research question:* Suppose you are planning to conduct an analysis to investigate what is the effect of ethnicity on an individual’s income?

*Available data:* Taken from a cross-sectional survey of 1000 individuals. Available variables are:

- $lpay$  = a continuous measure of annual income in £, transformed to the log scale
- $age$  = age, a continuous variable measured in years
- $eth$  = ethnicity, a 2 level categorical variable (0=white; 1=non-white)

Use your graphical models toolkit to construct DAGs to represent suitable Bayesian models to address the questions below.

- (a) You decide to fit a simple Bayesian linear regression model to investigate the effect of ethnicity on income, adjusting for age, as follows:

$$\begin{aligned}lpay_i &\sim \text{Normal}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_{eth}eth_i + \beta_{age}age_i \\ \beta_0, \beta_{eth}, \beta_{age}, \sigma^2 &\sim \text{Fully specified vague priors}\end{aligned}$$

- (i) Construct a DAG to represent this model (which we will call the analysis model of interest). (Leave some space on your white board, as you will be extending this DAG in later parts of the question).
- (b) When the data arrive from the survey company, you discover that 20% of the values for  $lpay$  are missing. The survey company tells you that pay was originally collected from all the individuals in the study, but the answers for the first 200 individuals were subsequently lost due to a data handling error (embarrassingly, no backups had been taken).
- (i) What sort of missing data mechanism does this correspond to?

- (ii) Let  $m$  be a missingness indicator for  $lpay$  (set to 0 when pay is observed and 1 when pay is missing). Construct a DAG for  $m$  to represent a suitable model of response missingness. Does this DAG share any links with the DAG for your analysis model of interest?
- (c) Suppose that the survey company contacts you again to say that they had made a mistake, and in fact they had never had income data for these 200 individuals because they had refused to answer the question.
  - (i) Discuss in your group how this new information might change your assumptions about the missing data mechanism.
  - (ii) Change your DAG to reflect any new assumptions you decide to make about the missing response mechanism.
- (d) Suppose you decide that your analysis could be improved by adding education to your analysis model. The survey company tell you that they can provide you with a binary education variable indicating whether or not each subject had any formal educational qualifications.
  - (i) Modify your DAG to include education (denoted  $edu$ ) as an additional covariate in your analysis model
- (e) When the education data arrive from the survey company, you find that 35% of the values are missing. You do some exploratory analysis of the data and find that the majority of missing values are for young, non-white individuals.
  - (i) Discuss in your group what might be a reasonable assumption about the missing data mechanism for  $edu$ .
  - (ii) Change your DAG to reflect the fact that  $edu$  contains missing values, and how you plan to model this.
- (f) *Optional question if you have time!* You then notice that the 200 individuals with missing income data are a subset of the 350 with missing education data. You further notice that, amongst the subjects with observed income,  $lpay$  tends to be lower for those with missing values for  $edu$ .
  - (i) In your group, discuss the implications of this information in terms of the assumptions you have made about the missing data mechanism for  $edu$ .
  - (ii) If necessary, elaborate your DAG to take account of any new modelling assumptions you make.

### 3 Practical C

In this exercise you will discuss a fictitious analysis of some data relating to HIV status from a demographic and health survey for an African Country.

*Research aim:* to estimate HIV prevalence for men and women among the adult population of country X.

*Available data:* in addition to HIV status ( $h$ ), the survey includes information on age ( $a$ ), gender ( $g$ ), region of residence ( $r$ ), socioeconomic ( $s$ ) and behavioral ( $b$ ) variables. There are no results of HIV testing for 30% of eligible men and 25% of the eligible women, and for those with a HIV test result, 10% have missing values for one or more of the socioeconomic or behavioral variables. In a recent study in another African country, it was found that among people who already know their HIV status, those who are HIV-positive are four times more likely to refuse a test than those who are HIV-negative.

*Proposed analysis:* Your colleague proposes to carry out the following analysis in order to estimate HIV prevalence using these data:

- Discard any individuals with missing covariates
- Using data from the remaining individuals (which includes those with missing HIV status but fully observed covariates) fit the following Bayesian logistic regression model

$$h_i \sim \text{Bernoulli}(p_i)$$
$$\text{logit}(p_i) = \alpha_r + \beta \mathbf{Z}$$

where  $\mathbf{Z} = \{a, g, s, b\}$  and  $\alpha_r$  are region specific intercepts. Then calculate prevalence from the observed and imputed values of  $h$ .

In your group, discuss the following questions:

- (a) What assumption is being made about the type of missingness for the response? Do you consider this assumption to be reasonable?
- (b) What assumption is being made about the type of missingness for the covariates? Do you consider this assumption to be reasonable?
- (c) You persuade your colleague that it is not necessary to discard the records with missing covariates, and that these could be imputed within a joint Bayesian model, along with the missing HIV test results. You plan to use the general strategy for Bayesian analysis with missing data, discussed in Lecture 4.
  - (i) What model would you use as a base case?
  - (ii) What further analyses would you carry out to explore the robustness of the results?
  - (iii) How might you make use of the extra information about refusal rates for HIV testing in the other African country?