

Bayesian approaches to handling missing data: Solutions to Practical Exercises

1 Practical A

1. (a) The true overall mean cognitive function score at age 90 is 14.
(b) The true marginal distribution of $Cog90$ is 48% in category 'good', 36% in category 'moderate' and 16% in category 'poor' (add up to 100%).
(c) Given 'high' $Cog85$, the distribution of $Cog90$ is 61% in category 'good', 33% in category 'moderate' and 6% in category 'poor' (add up to 100%).
Given 'low' $Cog85$, the distribution of $Cog90$ is 14% in category 'good', 43% in category 'moderate' and 43% in category 'poor' (add up to 100%).
As we would expect, with low cognitive function score at age 85, we tend to have poorer cognitive function score at age 90, which makes intuitive sense.
2. (a) $Cog90$ is MCAR if the probability of observing $Cog90$ does not depend on the values of $Cog90$ or $Cog85$.
(b) $Cog90$ is MAR if **given** $Cog85$ (a fully observed variable) the probability of observing $Cog90$ does not depend on the value of $Cog90$.
(c) $Cog90$ is MNAR if even given fully observed $Cog85$, the probability of observing $Cog90$ depends on the (possibly unseen) value of $Cog90$.

3A. Realisation A

- (a) The mean $Cog90$ from the observed data is 15.9, a biased estimate.
- (b) The marginal distribution of $Cog90$ from the observed data is 62% in category 'good', 31% in category 'moderate' and 7% in category 'poor'. Biased (differs from the true marginal distribution).
- (c) Biased, given 'high' $Cog85$, the conditional distribution of $Cog90$ from the observed data is 72% in category 'good', 26% in category 'moderate' and 2% in category 'poor'.
- (d) Unbiased, given 'poor' $Cog90$, the conditional distribution of $Cog85$ from the observed data is 25% in category 'high' and 75% in category 'low'.
- (e) MNAR, the proportion of missing values depends on the value of $Cog90$ (see red values in brackets in Table 2 below).

3B. Realisation B

- (a) The mean $Cog90$ from the observed data is 14, a unbiased estimate.
- (b) The marginal distribution of $Cog90$ from the observed data is 48% in category 'good', 36% in category 'moderate' and 16% in category 'poor'. Unbiased (the same as the true marginal distribution).
- (c) Unbiased, given 'high' $Cog85$, the conditional distribution of $Cog90$ from the observed data is 61% in category 'good', 33% in category 'moderate' and 6% in category 'poor'.
- (d) Unbiased, given 'poor' $Cog90$, the conditional distribution of $Cog85$ from the observed data is 25% in category 'high' and 75% in category 'low'.

- (e) MCAR, the proportion of missing values is the same regardless of the value of *Cog90* or *Cog85* (see red values in brackets in Table 3 below).

3C. Realisation C

- (a) The mean *Cog90* from the observed data is 15.1, a biased estimate.
- (b) The marginal distribution of *Cog90* from the observed data is 56% in category ‘good’, 34% in category ‘moderate’ and 10% in category ‘poor’. Biased (differs from the true marginal distribution).
- (c) Unbiased, given ‘high’ *Cog85*, the conditional distribution of *Cog90* from the observed data is 61% in category ‘good’, 33% in category ‘moderate’ and 6% in category ‘poor’.
- (d) Biased, given ‘poor’ *Cog90*, the conditional distribution of *Cog85* from the observed data is 50% in category ‘high’ and 50% in category ‘low’.
- (e) MAR, the proportion of missing values depends on the value of *Cog85*, but not *Cog90* (see red values in brackets in Table 4 below).
4. For Table A, it is NOT possible to calculate an unbiased estimate of the mean using only the observed data, since the data are MNAR. We would need to know the proportion of missing values in each cell, which with real data we cannot calculate.
 For Table B, the mean of the observed data is an unbiased estimate since the data are MCAR.
 For Table C, we can calculate an unbiased estimate of the mean using a weighted average of the means in the different categories of *Cog85*, the variable that determines the missingness. First calculate the mean of the observed *Cog90* in each of the *Cog85* categories. Then calculate a weighted average of these, using the number of individuals in each *Cog85* category (*Cog85* is fully observed, so we know whether individuals with missing *Cog90* are in the ‘high’ or ‘low’ category of *Cog85*).

$$\begin{aligned}
 \text{high.mean} &= ((33 \times 20) + (18 \times 10) + (3 \times 5))/54 \\
 \text{low.mean} &= ((1 \times 20) + (3 \times 10) + (3 \times 5))/7 \\
 \text{unbiased.mean} &= (\text{high.mean} \times 72/100) + (\text{low.mean} \times 28/100)
 \end{aligned}$$

In Tables 2 to 4, the numbers in brackets are the proportion of missing data calculated using the true data in Table 1.

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 (0.75)	12 (0.5)	1 (0.25)	46 (0.64)
low	3 (0.75)	6 (0.5)	3 (0.25)	12 (0.43)
age 85 margin	36 (0.75)	18 (0.5)	4 (0.25)	58 (0.58)

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

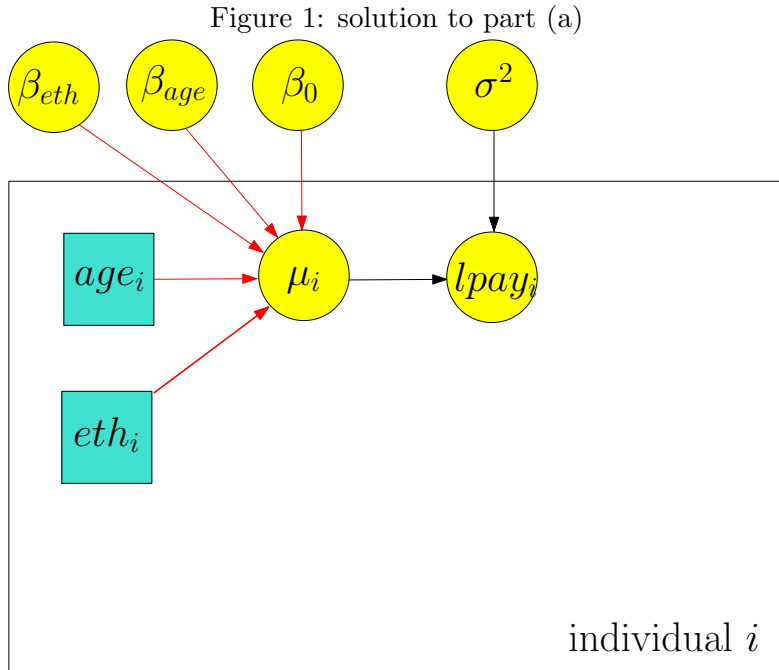
Cognitive function at age 85	Cognitive function at age 90			age 90 margin
	good (mean = 20)	moderate (mean = 10)	poor (mean = 5)	
high	22 (0.5)	12 (0.5)	2 (0.5)	36 (0.5)
low	2 (0.5)	6 (0.5)	6 (0.5)	14 (0.5)
age 85 margin	24 (0.5)	18 (0.5)	8 (0.5)	50 (0.5)

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 (0.75)	18 (0.75)	3 (0.75)	54 (0.75)
low	1 (0.25)	3 (0.25)	3 (0.25)	7 (0.25)
age 85 margin	34 (0.71)	21 (0.58)	6 (0.38)	61 (0.61)

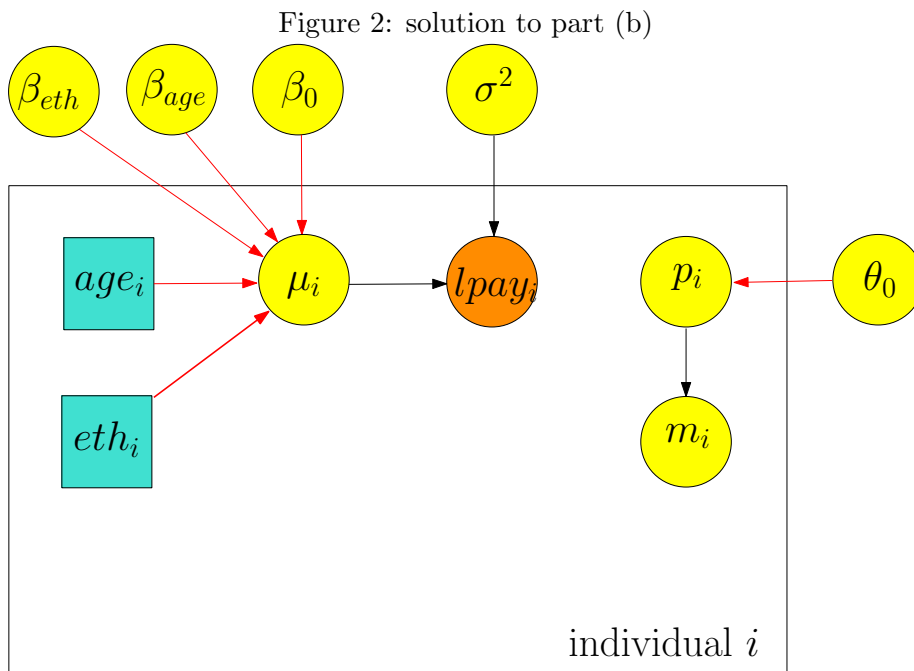
2 Practical B

- (a) The DAG for the Analysis Model (AM) should look something like the following:



- (b) Assume the missingness in $lpay$ is MCAR and construct a Model of Response Missingness (MoRM) which is not connected to the AM. In this case ignorable missingness is assumed, and the AM and MoRM models are disjoint and can be fitted separately (which effectively means that the MoRM doesn't need to be fitted at all if interest is only in the parameters of the AM). A typical specification for the MoRM might be something like

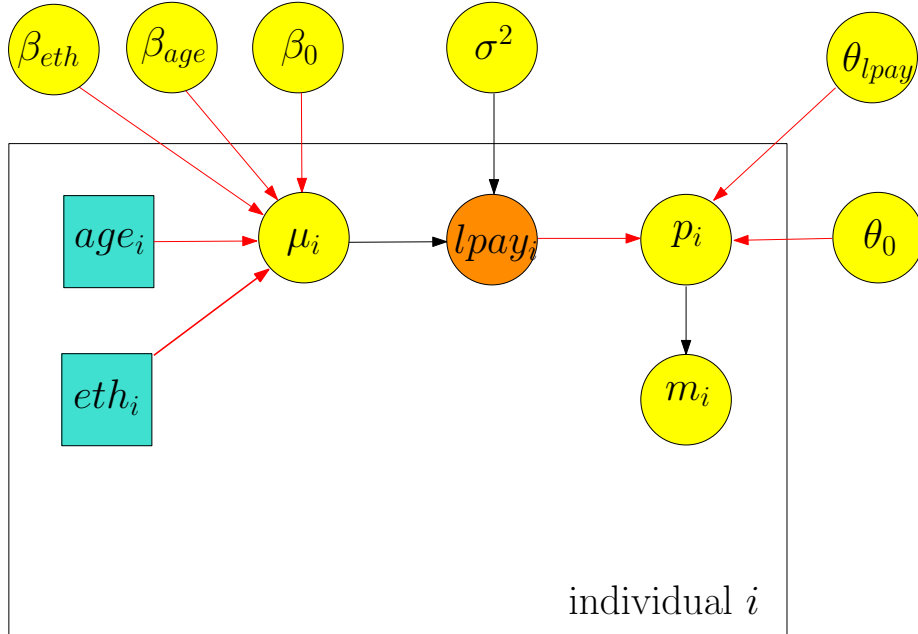
$$m_i \sim \text{Bernoulli}(p_i); \text{logit}(p_i) = \theta_0.$$



- (c) It is plausible that the reason for refusing to respond to the income question is related to the value of the individual's income. This corresponds to a missing not at random (MNAR) mechanism, and hence the missing data indicator m in the graph is dependent on $lpay$. The MoRM should be connected to the AM via the $lpay$ node, and the two sub-models must be fitted simultaneously. A typical specification for the MoRM might be something like

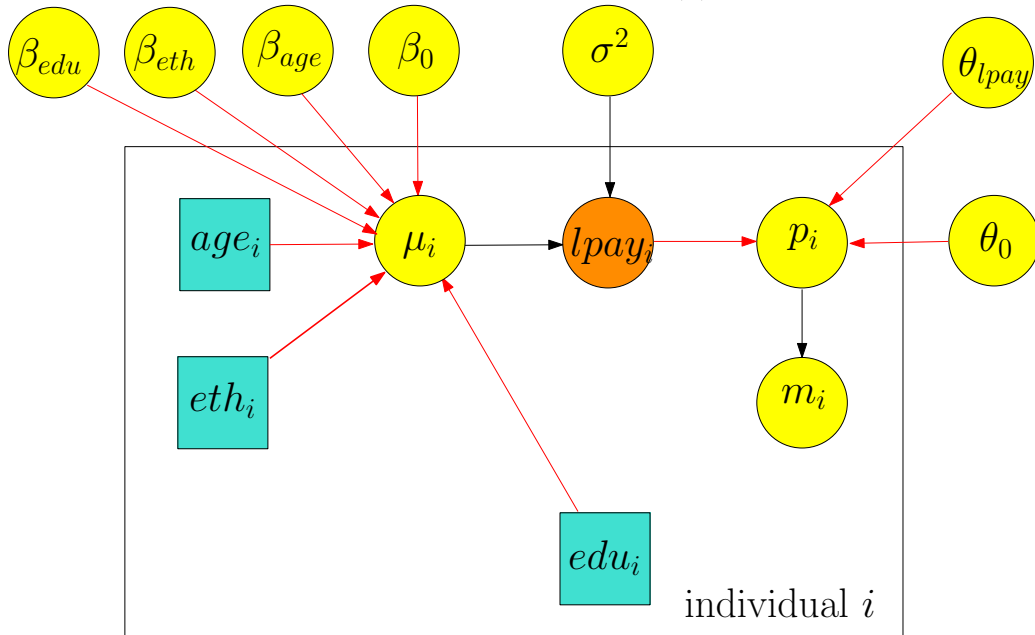
$$m_i \sim \text{Bernoulli}(p_i); \text{logit}(p_i) = \theta_0 + \theta_{lpay} lpay_i.$$

Figure 3: solution to part (c)



- (d) The DAG with edu (initially assumed to be fully observed) included as a covariate in the AM is as follows:

Figure 4: solution to part (d)

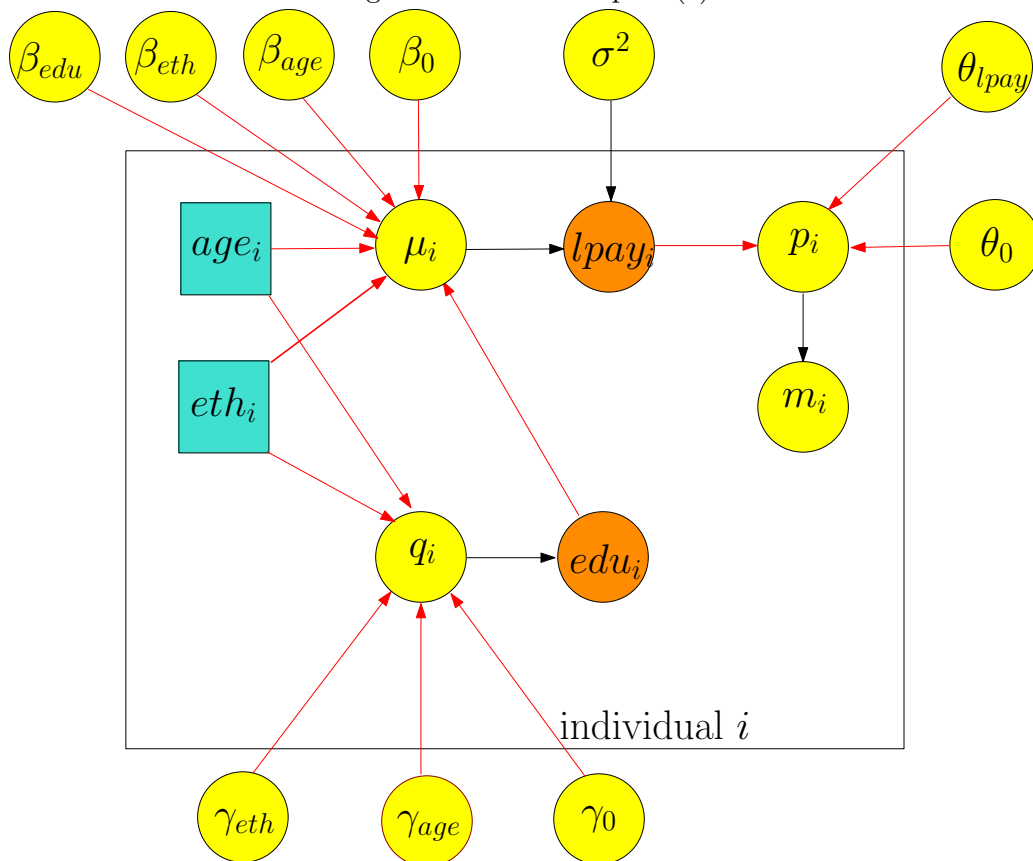


- (e) Assume the missingness in edu can be explained by age and eth (MAR), and add a Covariate Imputation Model (CIM). Information between the CIM and AM flows through the edu node.

Now the three sub-models, AM, MoRM and CIM, must be fitted simultaneously (see DAG below). (Note that when building a CIM, we need to think about including two sets of variables — those that we think are likely to influence the *value* of the missing covariate, and those we think are likely to influence the *reason* the covariate value is missing, i.e. the missingness mechanism. In this example, *age* and *eth* may well influence both). A typical specification for the CIM might be something like

$$edu_i \sim \text{Bernoulli}(q_i); \text{logit}(q_i) = \gamma_0 + \gamma_{eth}eth_i + \gamma_{age}age_i.$$

Figure 5: solution to part (e)



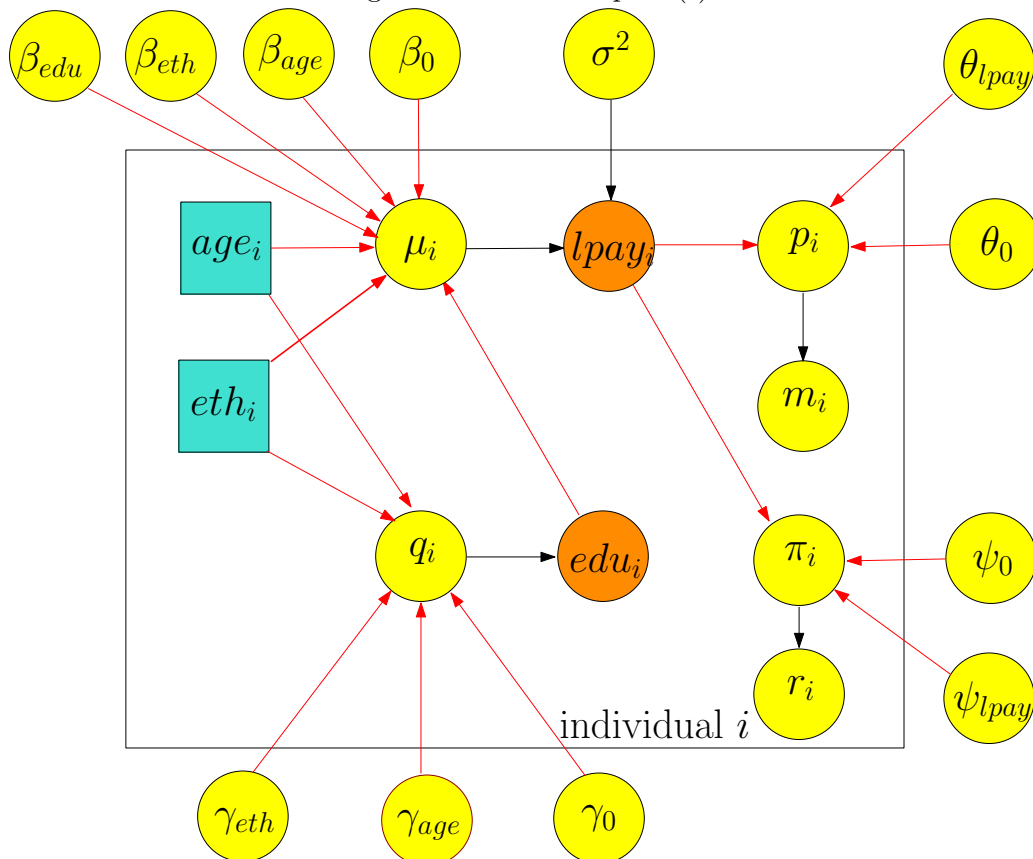
- (f) This is a tricky one! If we assume the missingness in edu also depends on $lpay$, then if $lpay$ were fully observed it would **not** be necessary to further elaborate the CIM, since there is automatic feedback from $lpay$ via the AM in a full Bayesian model. However, since $lpay$ and edu are both missing for some individuals, we should really consider elaborating the model to assume that the missing data mechanism for edu is also MNAR. This requires adding a fourth sub-model, a Model of Covariate Missingness (MoCM), to model the missing covariate indicator (denoted r_i , say) together with a link from $lpay$ to the variable representing the probability that edu is missing. A typical specification for the MoCM might be something like

$$r_i \sim \text{Bernoulli}(\pi_i); \text{logit}(\pi_i) = \psi_0 + \psi_{lpay}lpay_i.$$

See DAG below. Notice that there is no link between edu_i and r_i in this MoCM, since we are assuming that the non-ignorable missingness is arising due to dependence of the missingness mechanism on $lpay$. However, we may wish to consider extending the MoCM to depend on edu as well, if we thought that the (possibly unobserved) value of edu also directly influenced the probability that it was missing.

Note that, in practice, a joint model as complicated as this may be quite tricky to fit! Informative priors on some of the parameters of the missing data models may help.

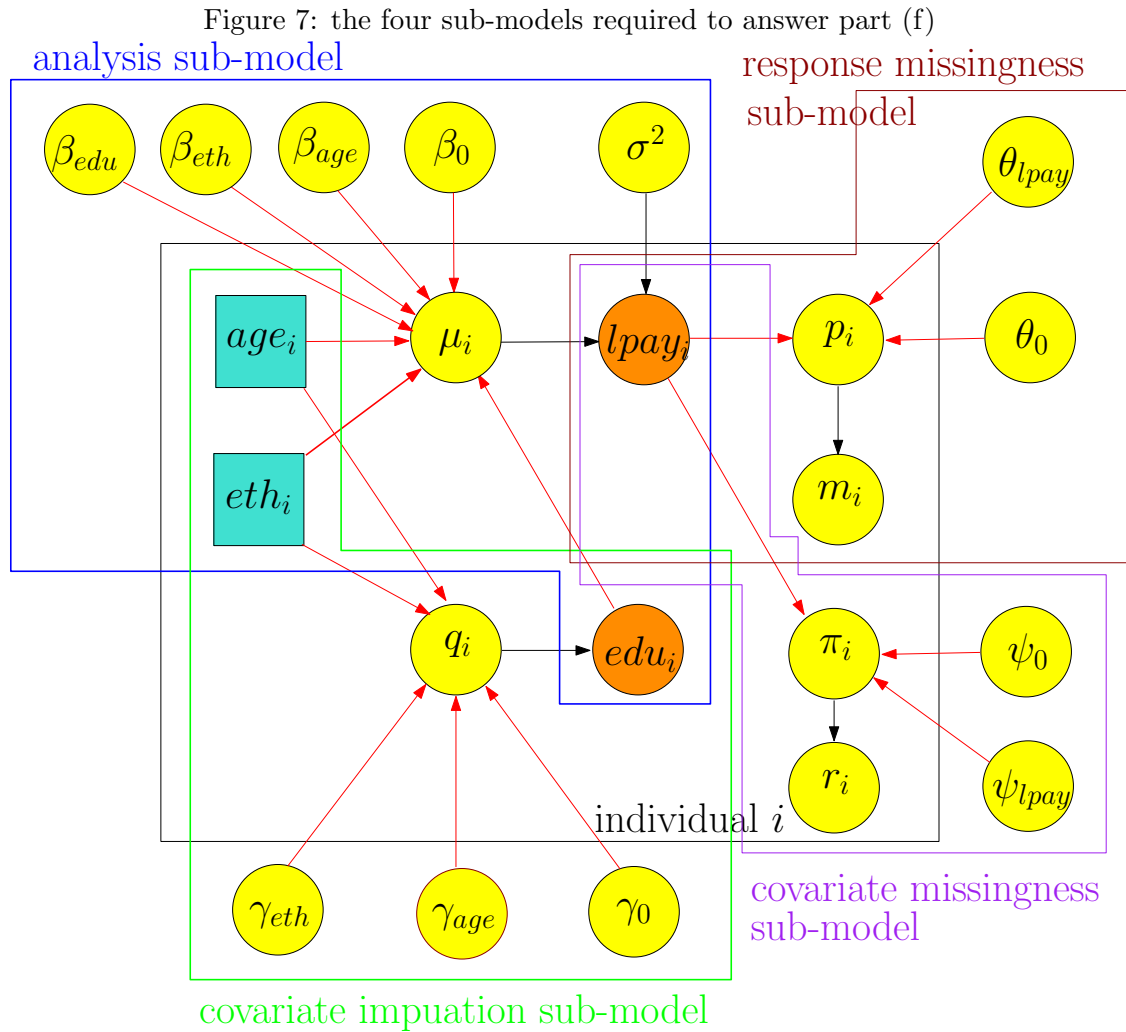
Figure 6: solution to part (f)



The joint model now consists of four submodels:

- Analysis Model (AM): answers the research question,
- Covariate Imputation Model (CIM): imputes the missing *edu* values,
- Model of Response Missingness (MoRM): allows for informative missingness in the response,
- Model of Covariate Missingness (MoCM): allows for informative missingness in the partially observed covariate, *edu*.

The nodes in each submodel are shown in Figure 7.



3 Practical C

- (a) HIV is assumed to have ignorable missingness. This does not seem reasonable given the findings from the other study that suggest that the probability of HIV status being missing is likely to be higher for individuals who are HIV positive.
- (b) The implicit assumption for the socioeconomic or behavioral covariates is that doing a complete case analysis is unbiased (which typically means an assumption of MCAR, see discussion on bias in CC for missing covariates in Lecture 3). It is more plausible that the missingness mechanism for these variables is MAR, dependent on other observed variables such as region, age and gender, and maybe on additional variables collected in the survey that were not initially considered for this analysis.
- (c)
- (i) A suitable base case model would consist of
- an analysis model (AM) — the logistic regression model that your colleague proposed could form the starting point for this;
 - a covariate imputation model (CIM) — some exploratory analysis of the pattern of the missing socioeconomic or behavioral variables would be required to determine how these should be imputed; the CIM should take into account the correlation structure of the covariates;
 - as we have external information suggesting the HIV has informative missingness, we should also include an model of response missingness (MoRM) of the form

$$m_i \sim \text{Bernoulli}(p_i)$$
$$\text{logit}(p_i) = \theta_0 + \delta h_i$$

where m_i is a missingness indicator for *HIV*.

- (ii) There are various possibilities for the sensitivity analyses. For example, place a point prior on δ in the MoRM (parameter sensitivity); treat the region specific intercepts as random effects in the AM or add some interaction terms (assumption sensitivity).
- (iii) The information about non-response rates from the other study could be used to construct an informative prior for δ in the MoRM.