

Bayesian approaches to handling missing data: WinBUGS Practical Exercises

You will be using WinBUGS 1.4.3 for these practicals.

All the data and other files you will need for the practicals are provided in a zip file, which is available on a USB drive, and can also be downloaded from

<http://www.bias-project.org.uk/Missing2012/MissingData.zip>.

Solutions for all the exercises are also provided in the zip files.

The files for this practical include `Methods(ExtraSlice).odc` and `Methods(Original).odc`. To use the slice updater (a more robust updating algorithm), you will need to replace the file `Methods.odc` in the WinBUGS subdirectory

`.../WinBUGS14/Updater/Rsrc`

with the `Methods(ExtraSlice).odc` version.

- Delete file `Methods.odc` from `.../WinBUGS14/Updater/Rsrc`.
- Copy `Methods(ExtraSlice).odc` to `.../WinBUGS14/Updater/Rsrc`.
- Rename `Methods(ExtraSlice).odc` as `Methods.odc`.

Use `Methods(Original).odc` if you wish to return to the original settings after this practical.

Also, **turn the blocking options off** (Selection the Blocking Options from the Options menu in WinBUGS and uncheck fixed effects box).

Practical A: Missing response data: HAMD data example

In this example, the models presented in Lecture 2 will be fitted using WinBUGS. The data used is a slightly modified version of the HAMD data set based on two treatments only. The strategy presented in Lecture 4 will be followed.

Constructing a base case

This part is a class exercise, with each command and result demonstrated. The instructions are also given below.

Step 1: select an analysis model

- 1 Start WinBUGS
- 2 Open `hamdMAR-model.odc`. This is the analysis model with random intercepts and random slopes described in Lecture 2.
- 3 Run this model for the complete cases only using the interactive interface as follows:
 - Open up the Model : Specification window.
 - Highlight model and click on check model (this should also work without highlighting, provided `hamdMAR-model.odc` is the open window).
 - Now open the HAMD data file `hamdMAR-data.odc` and click on load data. Notice that the HAMD scores `hamd` are ordered so that the complete cases are all together at the start of the file, and that `N` is set to 141 so that only the first 141 individuals (i.e. only the complete cases) are used in fitting the model.
 - Type 2 in the num of chains box, to run the model with two chains.
 - Next, click compile.
 - Open the initial values file for chain 1, `hamdMAR-inits1.odc`, and click on load inits.
 - Open the initial values file for chain 2, `hamdMAR-inits2.odc`, and click on load inits.
 - The initial values files contain values for the mean and standard deviance of the random effects, but the program also requires initial values for the individual random effects. Click gen inits to generate some.
 - Open up the Inference : Samples window, type `contrast` in the node window and then click set. This sets the monitor. Repeat for `alpha.mu`, `alpha.sigma`, `beta.mu`, `beta.sigma` and `sigma`. Type * in the node window to indicate all monitored quantities.
 - Open up the Model : Update window and generate 1000 iterations.
 - Click history to generate traces of the simulated values. The model appears to have converged quite quickly (around 250 iterations), but because we are using slice samplers to update some of the parameters, we must discard at least the first 500 iterations since these are used in the adaptive phase of the slice sampler and cannot be used for posterior inference. To be on the safe side generate another 1000 iterations and discard the first thousand.
 - Type 1001 in the begin box (to discard burn-in), then click stats to get summary statistics.
 - Select `contrast` in node and click density to get a density plot of the distribution of `contrast`.

- 4 Now run the same model using both the complete and incomplete cases. Only one change needs to be made to the files used for fitting the complete cases only; `hamdMAR-data.odc` must be edited to set N to 200.

Step 2: add a covariate imputation model

In this example we have no missing covariates, so Step 2 is not needed.

Step 3: add a model of response missingness

To complete the base case, a model of response missingness must be added. `hamdMNAR-model.odc` contains WinBUGS code for a joint model with the analysis model fitted in Step 1 and the model of response missingness described in Lecture 2 which relates the probability of drop-out to the current score.

- Open `hamdMNAR-model.odc`. Notice that the model of response missingness is only fitted for the weeks with drop-out (weeks 2-4) and excludes individuals who have already dropped out (i.e. those whose probability of being missing is clearly 1).
- Highlight model and click on check model in the Model : Specification window.
- Now open the HAMD data file `hamdMNAR-data.odc`. Compare this datafile to the one used just for fitting the analysis model. Click on load data.
- As before, type 2 in the num of chains box and click compile.
- Load the initial values files for chains 1 and 2, `hamdMNAR-inits1.odc` and `hamdMNAR-inits2.odc`, and click on gen inits.
- Set appropriate quantities for monitoring using the Inference : Samples window.
- Generate 1000 iterations. Check convergence and run more iterations as necessary.
- Produce appropriate summary statistics.

Sensitivity analysis

This part is an individual exercise.

Step 6: assumption sensitivity

Implement a series of modifications to the analysis model and response model of missingness part of the joint model `hamdMNAR-model.odc`, rerun and compare the results. (Hint: you may need to modify the model and initial value files. There is no need to change the data file.)

- 1 assume t_4 errors instead of Normal errors for the response in the analysis model;
- 2 allow drop-out probability in the MoRM to be dependent on the *change* in score. (Hint: see lecture 2 for formula);
- 3 allow drop-out probability in the MoRM to be treatment dependent by using different θ and δ for each treatment, as well as dependent on the *change* in score.

Step 7: parameter sensitivity

Fix δ to different values and explore the impact on **contrast**. (Hint: use the base model as a starting point.)

Step 8: determine robustness of conclusions

What would you report about this trial?

Practical B: Missing response data: Income example

This example is similar to the example discussed in Lecture 4, and uses a small subset of sweep 1 data taken from the Millennium Cohort Study (MCS), relating to 200 single main respondents (usually mothers) who are in work and not self-employed. The motivating question concerns how much extra an individual earns if they have a higher level of education. Some of the values of the HPAY variable have been artificially replaced with missing values, and this data can be found in `income-data.odc`. This file contains the following variables:

HPAY: a continuous variable containing the log of hourly pay (the distribution of hourly pay is skewed, so we use a log transform to achieve approximate normality).

STRATUM: a 9-level categorical variable, required to take account of the structure in the data as MCS is stratified by UK country (England, Wales, Scotland and Northern Ireland), with England further stratified into three strata (ethnic minority, disadvantaged and advantaged) and the other three countries into two strata (disadvantaged and advantaged).

AGE: a continuous variable denoting the respondent's age.

REG: a binary indicator for region (1 = London; 2 = other).

EDU: a 3-level categorical variable indicating the level of National Vocational Qualification (NVQ) equivalence such that 1 = none or NVQ 1; 2 = NVQ 2/3 (0/A-levels) and 3 = NVQ 4/5 (degree).

The code to fit a regression model to this data is given in file `income-model.odc`, and two sets of initial values can be found in `income-inits1.odc` and `income-inits2.odc`. Note that we assume t_4 errors for robustness to outliers.

- 1 Fit the regression model and produce summary statistics using complete cases only. (Hint: use the first 121 cases only). Is income level affected by education level?
- 2 Now fit the regression model and produce summary statistics using the incomplete cases as well. (Hint: use all the records by changing 121 to 200.) Whereas for missing covariates you must add a sub-model to take account of missingness, you do not need to do this for missing responses, so there is no need to edit `income-model.odc`. However, because you are not adding a model of missingness for the response, you are making a strong assumption that the missingness is 'ignorable'. Do your conclusions about the effect of higher education levels change?
- 3 Now extend the code to add a selection model for the missingness, which corresponds to assuming that the missingness mechanism is not ignorable. You will need to include a binary indicator for missing pay, `PAYID`, (0 = observed, 1 = missing) in your data, so now use file `income-missing-data2.odc`. Specify a logistic regression model for `PAYID`, and include `HPAY` as a covariate to predict the probability of missingness. The following weakly informative priors are recommended for the coefficients in the missingness model: $\theta \sim dlogis(0, 1)$ and $\delta \sim dnorm(0, 1.48)$, where θ is the intercept parameter and δ the parameter associated with `HPAY`. These priors incorporate a 95% prior belief that the change in the odds of being missing is between 1/5 and 5 for a one unit change in `HPAY`, which we believe should not be too restrictive. You will also need to amend the initial value files, for example `theta=0, delta=-0.5` and `theta=0, delta=0.5`. Run this model and obtain posterior estimates of the parameters of the analysis model and model of response missingness. What does this model suggest about education level? Do you think low or high hourly pay rates are more likely to be missing?

- 4 What happens if you place an informative prior on δ to force its sign to be positive? (Hint, use the $I(,)$ notation in WinBUGS to impose a suitable constraint on the prior distribution)
- 5 What sensitivity analysis do you think will be appropriate? (Hint: think about fixing parameters.)
If time permits, try implementing some of these models.

Practical C: Missing covariate data: childhood malaria example

This question uses data on childhood malaria in the Gambia, and involves dealing with missing covariate data. Data are available for 805 children, in file `malaria-missing-data.odc`, and include the following variables:

Y: a binary indicator of whether the child has malaria,

AGE: a continuous variable indicating the child's age in years,

BEDNET: a binary indicator (0 = doesn't sleep under bed net; 1 = sleeps under bed net),

GREEN: a continuous measure of greenness of the village environment in which the child lives (which is related to mosquito prevalence)

PHC: a binary indicator of whether the village that child lives in belongs to the Primary Health Care system (0 = No, 1 = Yes).

Questions of interest include:

- Does sleeping under a bed net reduce the risk of malaria?
- Do other child or village level covariates help explain the risk of malaria?

Values of the `BEDNET` variable are missing for 317 of these children. We will be assuming that the missing data mechanism is ignorable, so we just need to specify a model to impute the missing `BEDNET` covariate, but do not need to worry about explicitly modelling the missing data mechanism.

- 1 Start by fitting a logistic regression model to the complete case data. The complete cases data are given in file `malaria-cc-data.odc`, the model is given in the file `malaria-model.odc`, and the files `malaria-inits1.odc` and `malaria-inits2.odc` contain two sets of initial values. Run the model to produce summary statistics of the odds ratios of interest.
- 2 Now re-fit the same logistic regression analysis model to all the cases, including those with missing covariate values. You will need to edit the model code to include a covariate imputation model to impute the missing values of `BEDNET`. Model the missing bed net data using a logistic regression with `AGE`, `GREEN` and `PHC` as predictors,

$$\begin{aligned} \text{BEDNET}[i] &\sim \text{Bernoulli}(q[i]) \\ \text{logit}(q[i]) &= \gamma_1 + \gamma_2 \text{AGE}[i] + \gamma_3 \text{GREEN}[i] + \gamma_4 \text{PHC}[i], \end{aligned}$$

and specifying vague priors for γ_1 , γ_2 , γ_3 and γ_4 . Modify `malaria-model.odc` and the two sets of initial values. In addition to the usual parameters, monitor and obtain summary statistics for the posterior distribution of a subset of q and the imputed values for `BEDNET` (say for children 1–10 and 531–540), and of the γ parameters of the imputation model. Note that, out of the children with observed values of `BEDNET`, 70.5% sleep under a bed net. Can you explain the pattern of imputed values of `BEDNET` (hint: look at the posterior estimates of the γ 's to see which variables are most predictive of `BEDNET`; also think about how the values of the response `Y` influence the imputed values of `BEDNET`). Compare the odds ratios of interest with those from the complete case analysis.

- 3 Since the children are clustered within villages, we might want to consider elaborating our analysis model to include a village-level random effect. The data file `malaria-hierarchical-data.odc` contains the same data as before, but with the addition of a variable denoting which village (labelled 1,...,25) each child lives in. Extend your analysis model to include a village level random effect and re-fit the model. How do your estimates of the odds ratios compare to the model without random effects. What happens if you also include a village level random effect in the imputation model?

4 Children whose immune systems are compromised by being HIV positive may be at higher risk of malaria. A binary indicator of whether or not the child is HIV positive ($HIV = 1$ if HIV positive and 0 otherwise) is also available for most children, but is missing for children in some of the villages that are not in the Primary Health Care system. The HIV data are in file `malaria-hiv-data.odc`. Using the **non-hierarchical** analysis model, extend it to include HIV as an additional predictor. This will require you to also extend your covariate imputation model from question 2 to impute the missing values of HIV as well as the missing BEDNET values. For this we will use the multivariate probit model discussed at the end of Lecture 3 to jointly impute the missing values for these two covariates. Implementation of this model is quite tricky, so we have provided part of the model code in file `malaria-mvprobit-model.odc`. You will also need an additional data file called `malaria-hiv2-data.odc`. We will explain this code during the class, and once we have done so you should complete the model specification by specifying the remaining part of the multivariate probit for imputing HIV, and specifying priors for parameters of the multivariate probit imputation model.

- Monitor and obtain summary statistics for the posterior distribution of the parameters of the analysis model, the parameters of the imputation model and a subset of the imputed values for BEDNET (say for children 1–10 and 531–540) and for HIV (say for children 508–512 and 801–805). Is there any evidence that the child's HIV status is predictive of malaria risk?