University of Edinburgh
School of Mathematics

Bayesian Data Analysis

Computer lab 3

Hierarchical logistic regression with missing data: Childhood malaria in the Gambia.

Description

This question uses data from Diggle et al (2002). The study was designed to assess the effectiveness of the National Impregnated Bednet Programme in reducing child morbidity and mortality from malaria. Data were collected on 2035 children living in 65 different villages in the Gambia. Blood samples were taken for each child and measured for the presence of malarial parasites. Covariate data included the child's age, whether or not they regularly slept under a bed net, and if so, whether this was treated (with permethrin insecticide). Two village-level covariates were also measured: an indicator of whether the village belonged to the Primary Health Care system, and a measure of the greenness of the village environment. In addition to estimating the effect of bed net use on risk of malaria, interest also focuses on whether there is evidence of extra-binomial variation (which would indicate the presence of unmeasured factors leading to residual differences between villages in risk of malaria after adjusting for the effects of the measured covariates).

The original data file contains 1 record per child. However, we have aggregated the data to give counts of the number of children with each combination of covariate categories (age has been grouped into 4 categories: 0-2 yrs, 2-3 yrs, 3-4 yrs and 4-5 yrs; since the greenness index is a village level covariate and village is a 65-level category, the greenness index has been left as a continuous variable). This reduces the size of the data set from 2035 individual-level binary observations to 409 grouped binomial observations, which speeds up the MCMC sampling.

Variables:

- POP number of children with each combination of covariates
- MALARIA number of children in each covariate combination with malarial parasites in their blood
- AGE age group (1 = 0 to 2 yrs; 2 = 2 to 3 yrs; 3 = 3 to 4 yrs; 4 = 4 to 5 yrs)
- BEDNET bed net use (1 = doesn't sleep under bed net; 2 = sleeps under untreated bed net; 3 = sleeps under bed net impregnated with permethrin insecticide)
- GREEN continuous measure of greenness of the village environment
- PHC binary indicator: 1/0 = village does/doesn't belong to Primary Health Care system
- VILLAGE integer between 1 and 65 indicating which village the children live in

We will use BEDNET as a 2-level variable (i.e. ignoring whether or not the bed net was treated with insecticide), introducing

BEDNET0 to be a binary indictor (0 = doesn't sleep under bed net; 1 = sleeps under bed net).

Logistic regression model: $Y_i \mid p_i \sim Bern(p_i)$ independently, $i = 1, \ldots, n$, where

$$logit(p_i) = \alpha + \beta_{bednet} BEDNET0[i] + + \beta_{green} GREEN[i] + \beta_{PHC} PHC[i] + \beta_{AGE[i]}$$

1. Fit a non-hierarchical logistic regression model to these data to estimate the odds ratio for malaria associated with treated and untreated bed net use (**the model is in file malaria-model.odc**). The **data are contained in file malaria-data.odc**, and are given in the 'rectangular' format rather than the usual 'R' format.

   To improve convergence, the continuous covariate (GREEN) is centred around its mean.

   *Binary covariates coded as 0/1 (such as PHC) are be included in the linear predictor as a simple product of a scalar regression coefficient and the covariate value, e.g. beta.phc * PHC[i].*

   Categorical covariates with more than 2 levels are coded as 1, 2, 3, ... etc., and are included in the linear predictor using the 'double indexing' syntax. For example, AGE is a 4-level factor, and can be included in the linear predictor as the term beta.age[AGE[i]] where beta.age is a vector of length 4. If AGE[i] = 1 then beta.age[1] will be included in the linear predictor, if AGE[i] = 2 then beta.age[2] will be included in the linear predictor, and so on. One of the covariate levels of should be treated as the baseline level and hence the corresponding coefficient should be set to zero (for identifiability).

   Since the first element of the vector beta.age is fixed at zero rather than being treated as an unknown parameter, it does not require an initial value. Hence in the initial values file, an NA should be used for the first element of the vector of initial values for beta.age, e.g. beta.age = c(NA, 0.23, -0.65, 1.21). Have a look at the solutions file practical-3-malaria-solutions if you are not sure about any of this!

   The following variables are included in the model:

   - prevalence of malaria in baseline group (i.e. child in age group 1 (< 2yrs), sleeps without bednet, and lives in a village with average greenness index and not in the health care system):
     baseline.prev= $\exp(\alpha)/(1+\exp(\alpha))$
   - the odds ratio of malaria associated with each covariate (i.e. the exponential of each regression coefficient):
     - odds ratio of malaria for age group k vs age group 1:
       OR.age[k] = exp(*beta.age*[k])
     - odds ratio of malaria for children using bednets vs children not using bednets:
       OR.bednet = exp(*beta.bednet*)
     - odds ratio of malaria per unit increase in greenness index of village:
       OR.green = exp(*beta.green*)
     - odds ratio of malaria for children living in villages belonging to the primary health care system versus children living in villages not in the health care system: OR.phc = exp(*beta.phc*)
   - the posterior probability that using a bed net vs no bed net reduces risk of malaria: $P(OR.bednet < 1 \mid data)$.

   Run the model; once the simulations have converged, run sufficient further iterations to obtain a sample from the joint posterior distribution of all the parameters (setting burin and thinning as necessary). Record the posterior summaries of the baseline prevalence, odds ratios and posterior probability.

Address the following questions:

(a) Does the use of bednet reduce the risk of malaria?

(b) Comment whether the other considered factors affect malaria or not.

(c) How can the current model be improved for the given data?

2. Non-hierarchical to data with missing observations.

Here some of the values of the BEDNET0 variable are artificially replaced with missing values. The data is given at the individual-level, rather than group-level data because of the missing values. The new data can be found in **malaria-missing-data.odc.** The code to fit a logistic regression model to these data, with the missing bed net data, is given in file **malaria-missing-model.odc.** Two sets of initial values and a script to run the model can be found in malaria-missing-inits1.odc, malaria-missing-inits2.odc and malaria-missing-script.odc.

(a) Fit the model for this data and re-produce the summary statistics of the baseline prevalence, odds ratios and posterior probability. How do they compare to the ones obtained using complete data?

(b) In addition, monitor and obtain summary statistics for the posterior distribution of q (the probability that a child sleeps under a bednet) and a subset of the imputed values for BEDNET0 (say for children 41-50 and 131-140). Can you explain the differences between the posterior distributions of BEDNET0 for the different children?

(Hint: have a look at the value of the response, Y for these children (use Info $->$ Node info $->$ values). Note also that, out of the children with observed values of BEDNET0, 28% sleep under a bed net).