LSTAT2130 - Introduction to Bayesian Statistics
Project 2017-18

The file `absences.txt` reveals information on a simple random sample of 100 persons working in a big company hiring several thousands of employees: it includes `Sex`, `Age` and the number of days of absence during the preceding year. The information for the first two variables is sometimes missing and coded as 'NA' in the dataset.

Log-linear models are often used to describe how the expected value of a count variable Y changes with covariates.

Denote by

– Y_i: the number of days of absence,

– Z_i: the gender indicator (0: male ; 1: female),

– X_i: the age (in years),

for subject i $(i = 1, \ldots, 100)$. The model hypotheses are the following:

$$\begin{aligned} (Y_i|z_i, x_i) &\overset{\text{ind.}}{\sim} \text{Pois}(\mu_i(z_i, x_i)) \qquad (i = 1, \ldots, 100) \\ \log \mu_i(z_i, x_i) &= (\beta_0 + \alpha z_i) + (\beta_1 + \tau z_i)(x_i - 39) \end{aligned}$$

QUESTIONS

1. Provide a meaningful population interpretation to each of the model parameters.
2. Based on the subjects for which no missing data occur for Z_i or X_i, provide a theoretical expression for the model likelihood. Code its logarithm as a function in R. Then, compute the Maximum Likelihood Estimators (MLEs) using the `R` function `nlm`.
3. Specify non informative priors for each of the model parameters. Write a function in `R` giving the joint log posterior for a given value of the parameter vector (while still discarding subjects with missing data).
4. Still using R (and without using specialized libraries),
   (a) Implement the Metropolis algorithm (in its element-wise version) to generate a random sample from the joint posterior of the model parameters. [Target an acceptance rate of 40% for each parameter].

(b) Using two different criteria, check the convergence of your chains using the `R` package `coda`.

(c) Based on that sample, provide point estimates and 95% credible intervals for the model parameters.

5. Make a similar exercise by building chains using `JAGS` and compare your results with the preceding ones.

6. Based on the generated chains, what can you say about the qualitative and quantitative effects of `sex` and `Age` on the number of days of absence in that company ?

7. Subjects with missing values can easily be handled in a Bayesian framework. This can be done by substituting a parameter to each missing data (here: 8 covariate values are missing), thereby adding (here: 8) extra parameters to the model. Consider the following Uniform priors for the missing $\tilde{x}_i$ and $\tilde{z}_i$,

$$\tilde{x}_i \sim \mathsf{U}_{(18,65)} \;\; ; \;\; \tilde{z}_i \sim \mathsf{Bernoulli}(.50)$$

(a) For the full dataset at hand, implement an MCMC algorithm using `JAGS` to sample the joint posterior for the model parameters (including the missing data).

(b) Based on the chains generated for the missing data, what can you say, for each of the concerned subjects, about his/her plausible values for the missing covariate ?

(c) Compare your conclusions based on the whole dataset on the qualitative and quantitative effects of `Sex` and `Age` on the response Y with the preceding ones (discarding subjects with missing data).

INSTRUCTIONS

- By **Thursday 31st May 2018 at 13:00**, each group of 3 students must transmit to Vincent BREMHORST:

  1. In his mailbox (on the 1st floor of the Stats Institute building): a **printed version** of their report detailing the answers to all the questions. The software code must be in appendix and referred clearly in the main text.
  2. By email (`vincent.bremhorst@uclouvain.be`): the software code (R and JAGS only) in a single zipped file enabling to reproduce the claimed results.

  There is no second chance for this report for a later exam session, see below.

- Each group of 3 students must work independently !! Any detected fraud will lead to a severe penalty.
- Any change in the group composition (see the appended `pdf` file) will lead to a zero score for your project.
- Each member of a group must work on all aspects of the project (no "specialization").

YOUR FINAL MARK FOR THIS COURSE:

Your final mark (E: max 20 points) for LSTAT2130 will be calculated as follows from your results at

- the written exam (W: max 15 points) in June or August-September ;
- the project (P: max 5 points) (written report in May, no second chance for a later session):

1. Case 1: $W \geq 7.5$: $E = W + P$ rounded down to the closest integer ;
2. Case 2: $W < 7.5$: $E = W/15 * 20$ rounded down to the closest integer.

In particular, it means that you must succeed the written exam to take advantage of your project results.