LSTAT2130 - Introduction to Bayesian Statistics
Project 2021-22

Data presentation

A survey of 1200 people (selected by simple random sampling among 20-59 year olds) was conducted in 2013 to determine the profile of cannabis users. Some people felt uncomfortable with the interviewer when asked about this topic, so a device involving the rolling of two (balanced) dice was developed. When asked the question ”Have you recently used cannabis?”, the respondent is asked to roll two dice. The instructions are as follows:

- If you get a double 6, answer ”yes”.
- If not, answer the question honestly.

Of course, the interviewer does not know the result of the dice roll. In this way, the interviewee avoids the embarrassment that a positive answer might cause, since the answer might simply be dictated by the result of the roll.
The file `cannabis.txt` contains the answer ($y = 1$ if yes, 0 otherwise), the gender (`male`=1 if male, 0 otherwise) and the age of the respondents. Here are the first 10 lines of the file:

```
y  male age
0     1  33
1     1  23
0     1  21
0     1  35
0     1  36
0     0  48
0     0  41
0     1  56
0     0  47
1     1  38
etc.
```

Questions

1. If π is the proportion of people in the population who have recently smoked cannabis, calculate the probability γ that a randomly selected person in this population will answer "yes" to the question asked.

2. Given γ, what is the distribution of the number Y of "yes" answers in such a survey? From this, deduce the likelihood function and the posterior distribution for π.

3. **-a-** Using the Metropolis algorithm and the R software (without using JAGS or specific packages), construct a random sample of the posterior distribution for π. Evaluate the convergence of your algorithm.

   -b- From the previous result, give a set of plausible values for π.

   -c- What is the (posterior) probability that the proportion of recent cannabis users is at least 10% among 20-59 year olds?

4. Repeat the same steps separately for men and women and evaluate the posterior plausibility that proportionally more men than women have recently smoked cannabis.

5. Consider a logistic regression model for the proportion π of cannabis users in the population with `male` and age as explanatory variables. More specifically, if age=x, suppose that

$$\pi_x = \frac{1}{1+\exp(-\eta_x)} \quad \text{with} \quad \eta_x = \begin{cases} \alpha_0 + \alpha_1(x-40) & \text{for a male} \\ \beta_0 + \beta_1(x-40) & \text{for a female.} \end{cases}$$

   -a- Starting from non-informative priors, develop R code (calling JAGS) to sample the joint posterior for $(\alpha_0, \alpha_1, \beta_0, \beta_1)$.

   -b- Based on the so-generated MCMC chains, provide a set of plausible values for α_1, β_1 and $(\alpha_1 - \beta_1)$. What can you say about the evolution of the log odds η_x of recent cannabis use with age ? Is it significantly different for men and women ?

   -c- Visualise the posterior distribution of the probability of recent cannabis use for a 25 year old male. Give a set of plausible values for this probability.

INSTRUCTIONS

- By **Friday 27th May 2022 at 13:00**, each group of 3 students must transmit their results to Hortense DOMS by uploading the following 2 documents to the MoodleUCL platform:

  1. A single PDF file containing the report detailing the answers to all the preceding questions (using the same structure and numbering as within the questionnaire). The software code must be in appendix and referred clearly in the main text.

     The file should be named using your family names in a row as in the following example:

     Smith-Jones-Brown_Report.pdf

  2. A single text file containing the commented software code (R and JAGS only) enabling to reproduce the claimed results. Its subdivision must follow the same structure as in the questionnaire.

     The file should be named using your family names in a row as in the following example:

     Smith-Jones-Brown_Rcode.R

  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 to the agreed group composition 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 obtained by rounding to the closet integer the sum of the 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):