STA623 - Bayesian Data analysis - Assignment 1

Marc Henrion

5-9 September 2022

Assigment

Please email your typed or scanned solutions to BOTH `mhenrion@mlw.mw` and `biostat-unima@unima.ac.mw`, before 23:59 on Monday 10 October 2022. Please include `STA623 - Assignment 1` in the subject line.

While we used `R` and JAGS in the classroom, you can use any software package of your liking to fit models for this assignment. Please include both your fitting code and model output and graphs with your solutions.

However, for any code that you are submitting, please explain what each block of code does by including comment lines in your code.

Notation

Please try to use the following notation where possible.

- X, Y, Z - random variables
- x, y, z - measured / observed values
- $\bar{X}$, $\bar{Y}, \bar{Z}$ - sample mean estimators for X, Y, Z
- $\bar{x}$, $\bar{y}, \bar{z}$ - sample mean estimates of X, Y, Z
- $\hat{T}$, $\hat{t}$ - given a statistic T, estimator and estimate of T
- $P(A)$ - probability of an event A occuring
- $f_X(.)$, $f_Y(.), f_Z(.)$ - probability mass / density functions of X, Y, Z; sometimes $p_X(.)$ etc. rather than $f_X(.)$
- p(.) - used as a shorthand notation for pmfs / pdfs if the use of this is unambiguous (i.e. it is clear which is the random variable)
- $X \sim F$ - X distributed according to distribution function F
- $E[X]$, $E[Y]$, $E[Z]$, $E[T]$ - the expectation of X, Y, Z, T respectively

Exercise 1 [60 marks]

Assume you observe some data $y_1, \ldots, y_n$ for exponentially distributed random variables $Y_1, \ldots, Y_n \sim \text{Exp}(\lambda)$.

- Derive the conjugate prior distribution for this sampling model.
- Derive the Jeffreys prior for this sampling model and state whether this is a proper prior distribution or not.
- Assume now a $\Gamma(a, b)$ prior distribution. Derive the posterior predictive distribution for new data $\tilde{Y}$.
- Write computer code, assuming you had a data object `dat` loaded in memory that would allow you to fit the model resulting from a $\Gamma(a, b)$ prior and an $\text{Exp}(\lambda)$ sampling model. Run your model on the following data $y_i, i = 1, \ldots, 20$:

$$0.2223, 0.0041, 0.0278, 0.3343, 0.4345, 0.0494, 0.0728, 0.0134, 0.3412, 0.1844$$

$$0.0597, 0.0002, 0.0516, 0.0022, 0.1461, 0.2334, 0.1828, 0.0171, 0.3775, 0.3179$$

[15 marks each]

Exercise 2 [40 marks]

Download the file `biomarker.csv` the STA623 GitHub page.

This dataset contains information on patients for which a certain biomarker was measured. This biomarker is hypothesized to be upregulated in patients infected with a certain pathogen but there are currently no prior data supporting this. There is however some prior evidence of an effect of biological sex on the marker measurements, with males showing slightly elevated levels. Finally, the biomarker measurement is sensitive to the operator and equipment being used, so that there may be differences depending at which hospital the measurement was taken.

The dataset contains the following columns:

- `pid` - this is just an anonymised patient identification number
- `sex` - this records the biological sex of each patient
- `infection` - test result for presence of the pathogen of interest in the patient (1=infected, 0=non-infected)
- `hospital` - this records an identification code for the hospital where each patient was seen
- `biomarker` - this records the measurement of the biomarker levels in each patient

Use JAGS to the following model, choosing priors of your own choosing for each parameter, writing $Y_{i,j}$ for the biomarker measurement for patient $i = 1, \ldots, n$ seen in hospital $j = 1, \ldots, k$:

$$Y_{i,j} = \beta_0 + \beta_1 \cdot male\_sex_i + \beta_2 \cdot infection_i + \mu_j + \epsilon_i$$

where $\mu_j \sim \mathcal{N}(0, \rho^2), j = 1, \ldots, k$ and $\epsilon_i \sim \mathcal{N}(\prime, \sigma^{\in}), i = 1, \ldots, n$

1. Explain the choice of prior distributions for all model parameters $(\beta_0, \beta_1, \beta_2, \rho^2, \sigma^2)$. [8 marks]
2. Write JAGS model code to fit the model. [16 marks]
3. Show trace plots and histograms for all model parameters and compute the effective sample size and Gelman-Rubin potential scale reduction factors. Discuss other model checks you could do. [16 marks]