STA6206 - Bayesian Data analysis - Assignment (Solutions)

Marc Henrion

9-13 September 2019

Assigment

Please email your typed or scanned solutions to `mhenrion@mlw.mw` and `biostat-unima@cc.ac.mw`, before 23:59 on Monday 21 October 2019. Please include `STA6206 - Assignment` 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 [40 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 Jeffrey's 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 $y_i, i = 1, \ldots, 20$:

$$0.2013, 0.0637, 0.1342, 0.0453, 0.2333, 0.2122, 0.1580, 0.0205, 0.3537, 0.5856$$

$$0.1193, 0.0333, 0.1380, 0.0707, 0.5619, 0.3633, 0.0235, 0.5965, 0.0055, 0.2835$$

[10 marks each]

Exercise 2 [60 marks]

Return to the titanic dataset from the STA6103 assignment. (You can download the file `titanic.csv` from either the STA6103 or the STA6206 GitHub page).

It will depend on the exact code you write, but you will almost certainly need to reformat the data into a format that will be useful for model fitting.

First assume a simple binomial sampling model for all $n = 2201$ passengers, writing p for the survival probability:

$$\begin{cases} p & \sim \text{Beta}(a, b) \\ Y|p & \sim \text{Bin}(n, p) \end{cases}$$

State or derive the posterior distribution of $p|Y = k$.

Write computer code to fit this model. Show trace plots and examine the posterior distribution plot for p.

[10 marks]

Now, assume a logistic regression model so that p depends on covariates. Write computer code to fit the following logistic regression model under the Bayesian paradigm:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot class_{2nd} + \beta_2 \cdot class_{3rd} + \beta_3 \cdot class_{crew} + \beta_4 \cdot child + \beta_5 \cdot male$$

[14 marks]

Produce trace plots for model parameters and comment.

[8 marks]

Plot the posterior distribution for model parameters and comment.

[8 marks]

Now add interction terms for class & age and class & sex. Show your computer code. Show trace and posterior distribution plots and identify any potential issues. If you do, how could you remedy this?

[8 marks]

Show a posterior distribution plot for the survival probability for a male crew member.

[4 marks]

Explain how you could decide between the 3 models you fitted.

[4 marks]

How does the last model compare to the same model implemented in a frequentist way?

[4 marks]