STAT452: Bayesian Statistics
Assignment 3

Due 1159pm Friday 10 May 2024

1. Suppose for a set of counties $i \in \{1, \ldots, n\}$ we have information on the population size X_i = number of people in 10,000s and Y_i = number of cancer fatalities. One model for the distribution of cancer fatalities is that, given the cancer rate θ, they are independently distributed with $Y_i \sim \text{Poisson}(\theta X_i)$.

a. Identify the posterior distribution of θ given data $(Y_1, X_1), \ldots, (Y_n, X_n)$ and a $\text{Gamma}(a, b)$ prior distribution.

b. The file `cancer_react.csv` contains population sizes (`x` in 10,000s) and number of cancer fatalities (`y`) for 10 counties that are near nuclear reactors in a given state. The file `cancer_noreact.csv` contains the same data on counties in the same state that are not near nuclear reactors. Consider these data as samples from two populations of counties: one is the population of counties with no neighboring reactors and a fatality rate θ_1 deaths per 10,000, and the other is a population of counties having nearby reactors and a fatality rate of θ_2 deaths per 10 000. We will model beliefs about the rates as independent between the two populations so that $\theta_1 \sim \text{Gamma}(a_1, b_1)$ and $\theta_1 \sim \text{Gamma}(a_2, b_2)$. Using the data in the two files identify the posterior distributions for θ_1 and θ_2.

c. Suppose cancer rates from previous years have been $\theta \approx 2.2$ per 10000 (and note that most counties are not near reactors). For each of the following three prior opinions, use Monte Carlo approximation to compute $E[\theta_1|\mathbf{y_1}, \mathbf{x_1}]$, $E[\theta_2|\mathbf{y_2}, \mathbf{x_2}]$, 95% quantile-based posterior intervals for θ_1 and θ_2, and $\Pr(\theta_1 > \theta_2|\mathbf{y_1}, \mathbf{y_2}, \mathbf{x_1}, \mathbf{x_2})$. Comment on the differences across prior opinions. i. Opinion 1: $(a_1 = a_2 = 2.2 \times 100, b_1 = b_2 = 100)$. Cancer rates for both types of counties are similar to the average rates across all counties from previous years. ii. Opinion 2: $(a_1 = 2.2 \times 100, b_1 = 100, a_2 = 2.2, b_2 = 1)$. Cancer rates in the current year for non-reactor counties are similar to rates in previous years in non-reactor counties. We don't have much information on reactor counties, but perhaps the rates are close to those observed previously in non-reactor counties. iii. Opinion 3: $(a_1 = a_2 = 2.2, b_1 = b_2 = 1)$. Cancer rates in the current year could be different from rates in previous years, for both reactor and non-reactor counties.

d. Using `rjags` and prior Opinion 1, compute an estimate for $\Pr(\theta_1 > \theta_2|\mathbf{y_1}, \mathbf{y_2}, \mathbf{x_1}, \mathbf{x_2})$. Provide MCMC diagnostic check results to demonstrate that your estimate is based on a chain that has converged.

2. Jeffrey's prior: For sampling models expressed in terms of a $d-$dimensional vector ϕ, Jeffreys' prior is defined as $p_J(\phi) \propto \sqrt{|I(\phi)|}$, where $|I(\phi)|$ is the determinant of the $d \times d$ matrix $I(\phi)$ that has entries

$$I(\phi)_{kl} = -E\left[\frac{\partial^2 \log p(Y|\phi)}{\partial\phi_k \partial\phi_l}\right].$$

Show that Jeffreys' prior for the normal model is $p_J(\theta, \sigma^2) \propto (\sigma^2)^{-3/2}$.

3. In this exercise we will analyse data on malaria in children from the Gambia. The data, in the file `gambia.csv`, consist of 2035 children from 65 villages. For child i, the variable `pos` in the dataset is a binary indicator that the child tested positive for malaria (1=child tested positive). There are five covariates in the dataset, but we will focus on the covariate `netuse` defined as:

- `netuse`: indicator variable denoting whether (1) or not (0) the child regularly sleeps under a bed-net. We will analyse the data using a logistic regression model with

$$Y_i \sim \text{Bernoulli}(\pi_i), \text{ and } \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i,$$

where Y represents the variable `pos`, X represents the variable `netuse` and β_0 and β_1 are unknown regression coefficients.

a. Write the likelihood function for the model defined above.
b. Using `rjags` and prior distributions $\beta_j \sim \text{Normal}(\mu = 0,\ \sigma^2 = 100)$ for $j \in \{0, 1\}$ compute posterior means, medians and 95% quantile-based posterior intervals for β_0 and β_1. Provide MCMC diagnostic check results to demonstrate that your estimates are based on a chain that has converged.
c. Based on your MCMC chain in part (b), estimate the median and 95% quantile-based posterior interval for the odds ratio associated with netuse. Use these results to describe the effect of netuse on the risk of a child contracting malaria.