Homework 6
ISyE 6420: Fall 2024

Question 1

A longitudinal study was conducted to understand the effect of age and sex on the orthodontic distance (y). Measurements on 27 children are given in the file ortho.csv. There are a total of 16 boys and 11 girls, which are identified in the dataset using the column Subject. Consider the following random effects model:

$$\begin{aligned} y_{ij} \mid \beta_0, \beta_1, \beta_2, u_i, \sigma_\epsilon^2 &\sim^{\text{ind.}} N\left(\beta_0 + \beta_1 age_{ij} + \beta_2 sex_i + u_i, \sigma_\epsilon^2\right) \\ u_i \mid \sigma_u^2 &\sim^{iid} N\left(0, \sigma_u^2\right) \end{aligned}$$

for $i = 1, \cdots, 27$ and $j = 1, \cdots, 4$. Here u_i represents the random effect of the i th subject. The sex variable should be coded as -1 for female and 1 for male. Assume the following prior distributions:

$$\begin{aligned} \beta_k &\sim^{\text{iid}} N\left(0, \sigma^2 = 10^8\right), k = 0, 1, 2 \\ \tau_\epsilon &\sim \text{Gamma}(.01, .01) \\ \tau_u &\sim \text{Gamma}(.01, .01) \end{aligned}$$

where $\tau = \frac{1}{\sigma^2}$

1. Fit the random effects model and plot the posterior densities of the five parameters $\beta_0, \beta_1, \beta_2, \sigma_\epsilon^2$, and σ_u^2. (use 100,000 samples with 10,000 burn-in.)
2. The intraclass correlation coefficient is defined as

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$$

Plot the posterior density of ρ. Does it appear to be significantly different from 0 ?

3. Fit the model ignoring the random effects (that is, set all the u_i 's to be 0) and plot the posterior densities of the four parameters $\beta_0, \beta_1, \beta_2$, and σ_ϵ^2. What differences do you see from the previous analysis using random effects (compare the posterior means and credible intervals of the four parameters)?

Question 2

The dataset gala.csv contains features of the 30 Galapagos islands. The relationship between the number of plant species and several geographic variables is of interest. Answer the following questions.

1. We are interested in modeling the Species with respect to the five predictor variables using a Poisson Generalized Linear Model with log-link. Obtain 10,000 MCMC samples with 1,000 burn-in. Note that the variable "Elevation" has missing values. Perform multiple imputation for the missing values assuming an exponential distribution with mean 425. Provide the mean and 95% credible intervals of the coefficients corresponding to the five variables (standardize the five variables).
2. Which variables appear to be significant?

Question 3

An exercise in the book Pagano and Gauvreau (2000) [1] features data on 86 patients who after surgery were assigned to placebo or chemotherapy (thiopeta). Endpoint was the time to cancer recurrence (in months).

Variables are: `time`, `group` (0 - placebo, 1- chemotherapy), and `observed` (0 - recurrence not observed, 1 - recurrence observed). This data is given in files `bladerc.csv|dat`.

Assume that observed times are exponentially distributed with the rate parameter λ_i depending on the covariate `group`, as

$$\lambda_i = \exp\{\beta_0 + \beta_1 \times \text{group}_i\}$$

After β_0 and β_1 are estimated, since the variable `group` takes values 0 or 1, the means for the placebo and treatment times become

$$\mu_0 = \frac{1}{\exp\{\beta_0\}} = \exp\{-\beta_0\}$$

$$\mu_1 = \frac{1}{\exp\{\beta_0 + \beta_1\}} = \exp\{-\beta_0 - \beta_1\},$$

respectively. The censored data are modeled as exponentials left truncated by the censoring time. Use noninformative priors on β_0 anmd β_1.

(a) Is the 95% Credible Set for $\mu_1 - \mu_0$ all positive?

(b) What is the posterior probability of hypothesis $H : \mu_1 > \mu_0$?

(c) Comment on the benefits of the treatment (a paragraph).

[1] Bladder cancer data from M Pagano and K Gauvreau, "Principles of Biostatistics, 2nd Ed. Duxbury 2000. Chapter 21, Exercise 9, page 512.