Biostat 682 Homework 3

Due: Monday, November 3, 2025 (23:59 pm)

Please use JAGS or PyMC to complete this homework.

1. Download the CSV file `swim_time.csv` from Canvas. The data file contains a data matrix on the amount of time, in seconds, it takes each of four high school swimmers to swim 50 yards. Each swimmer has six times, taken on a biweekly basis.

   (a) For each swimmer $j (j = 1, 2, 3, 4)$, fit a Bayesian linear regression model where considers the swimming time as the response variable and week as the explanatory variable. To formulate your prior, use the information that competitive times for this age group generally range from 22 to 24 seconds

   (b) For each swimmer $j (j = 1, 2, 3, 4)$, obtain a posterior predictive distribution for Y_j^*, their time if they were to swim two weeks from the last recorded time.

   (c) The coach of the team has to decide which of the four swimmers will compete in a swimming meet in two weeks. Using your predictive distributions, compute $\Pr(Y_j^* = \min\{Y_1^*, \ldots, Y_4^*\} \mid \mathbf{Y})$ for each swimmer j , and based on this make a recommendation to the coach.

2. Download the CSV file `UScrime.csv` from Canvas. The dataset contains crime rates (y) and data on 15 explanatory variables for 47 U.S. states. A description of the variables is provided in Table 1.

   (a) Fit a Bayesian linear regression model using noninformative priors. Obtain marginal posterior means and 95% credible intervals for coefficients. Describe the relationships between crime and the explanatory variables. Which variables seem strongly predictive of crime rates?

   (b) To test how well regression models can predict crime rates based on the explanatory variables, randomly divide the data roughly in half, into training set and a test set. Use the training dataset to fit the model and generate the posterior predictive median of the crime rates given the explanatory variables in the test dataset. Compare the posterior predictive median and the actural crime rate in the test dataset.

   (c) Repeat Parts (a) and (b) using spike-and-slab priors for regression coefficients.

Table 1: Description of variables in the `UScrime` dataset

Variable	**Description**	**Type / Unit**
`M`	Percentage of males aged 14–24.	Continuous (%)
`So`	Indicator for a Southern state (1 = South, 0 = otherwise).	Binary (0/1)
`Ed`	Mean years of schooling.	Continuous (years)
`Po1`	Police expenditure in 1960.	Continuous (index)
`Po2`	Police expenditure in 1959.	Continuous (index)
`LF`	Labour force participation rate.	Continuous (%)
`M.F`	Number of males per 1000 females.	Continuous (ratio)
`Pop`	State population.	Continuous (scaled)
`NW`	Number of non-whites per 1000 people.	Continuous (per 1000)
`U1`	Unemployment rate of urban males aged 14–24.	Continuous (%)
`U2`	Unemployment rate of urban males aged 35–39.	Continuous (%)
`GDP`	Gross domestic product per head.	Continuous (index)
`Ineq`	Income inequality measure.	Continuous (index)
`Prob`	Probability of imprisonment.	Continuous (0–1)
`Time`	Average time served in state prisons.	Continuous (years)
`y`	Crime rate per head of population in a specific category.	Continuous (rate)

3. Download the CSV file `gambia.csv` from Canvas. The dataset consists of 2,035 children from 65 villages from The Gambia. It contains eight different variables. A description of the variables in Table 2. Let $Y_i \in \{0, 1\}$ (pos) indicate the presence (1) or absence (0) of malaria in a blood sample taken from the child $i (i = 1, \ldots, 2035)$. Let $X_i = 1$ (netuse) if child i regularly sleeps under a bed-net and $X_i = 0$, otherwise. Let $v_i \in \{1, \ldots, 65\}$ denote the village of child i. Note that the dataset only contains the locations of villages instead of the labels.

   Fit the following logistic regression model

$$\text{logit}\{\Pr(Y_i = 1)\} = \alpha_{v_i} + X_i \beta_{v_i},$$

   where α_j and β_j are intercept and slope for village $j (j = 1, \ldots, 65)$. The priors are

$$\alpha_j \sim \text{N}(\mu_a, \sigma_a^2), \qquad \beta_j \sim \text{N}(\mu_b, \sigma_b^2).$$

   Choose noninformative priors for the hyperparameters μ_a, μ_b, σ_a^2 and σ_b^2. Based on your model fitting, address the following questions:

   (a) Scientifically, why might the effect of bed-net vary by village?

   (b) Do you see evidence that the slopes and/or intercepts vary by village? You may consider alternative model fitting and perform model comparisons.

   (c) Which village has the largest intercept? Slope? Does this agree with the data in these villages?

   (d) Are the results sensitive to the priors for the hyperparameters?

Table 2: Description of variables in the `gambia` dataset

Variable	**Description**	**Type / Unit**
`x`	x-coordinate of the village (UTM).	Continuous
`y`	y-coordinate of the village (UTM).	Continuous
`pos`	Presence (1) or absence (0) of malaria in a blood sample taken from the child.	Binary (0/1)
`age`	Age of the child (days).	Continuous
`netuse`	Whether (1) or not (0) the child regularly sleeps under a bed-net.	Binary (0/1)
`treated`	Whether (1) or not (0) the bed-net is treated (0 if `netuse`=0).	Binary (0/1)
`green`	Satellite-derived vegetation greenness near the village.	Continuous
`phc`	Presence (1) or absence (0) of a health center in the village.	Binary (0/1)