STAT 431 — Applied Bayesian Analysis — Fall 2022

Homework 6

Due: December 2, 11:59 PM (US Central)

In your submission, please include computer code and output that you used to answer the following problems. All numerical values should be given to at least three significant digits, unless otherwise stated.

1. The comma-separated values (CSV) file `lifeexpdiff.csv` contains, for each US state, the difference (in years) between that state's life expectancy and the overall US life expectancy, for the years 1960, 1970, 1980, 1990, 2000, and 2010.[1]

   With that file in R's current working directory, run the following code:

   ```
   Y <- read.csv("lifeexpdiff.csv", row.names=1)
   ```

   `Y` now contains the life expectancy difference for each state (row) and year (column).

   You will analyze the differences as responses in a Bayesian hierarchical normal linear regression, with (centered) year as the (continuous) predictor, and state as the grouping variable. In this problem, you will use the *univariate* prior formulation (as in the lectures on Random Effects and Hierarchical Models).

   (a) [3 pts] Consider the *univariate* prior formulation. Present (i) your R code for the list structure containing all of the data, and (ii) your JAGS code. Make sure to satisfy the following:

   - Use *centered* predictor values. (You may choose whether to pre-compute the average year in R.)
   - Give the overall regression coefficients β_1 and β_2 a mean of zero and a *precision* of 1×10^{-6} in their normal hyperpriors.
   - Give the response precision parameter τ_y^2 a Gamma$(0.001, 0.001)$ prior.
   - Give the regression coefficient standard deviation hyperparameters σ_{α_1} and σ_{α_2} hyperpriors that are Exponential(0.001).
   - Define a variable (deterministic node) for the standard deviation $\sigma_y = 1/\sqrt{\tau_y^2}$ of the response.

   (b) [3 pts] Run your model, and present a summary statistic table that includes at least the nodes for β_1, β_2, σ_y, σ_{α_1}, and σ_{α_2}. Make sure to satisfy the following:

   - Use at least three chains, with widely-separated starting points.
   - Remember to start with a burn-in period.
   - Use at least 100000 iterations from each chain (after burn-in).
   - DO NOT include any plots or convergence diagnostics in your submission.

[1] Data from: Woolf, S.H., & Schoomaker, H. (2019). Life expectancy and mortality rates in the United States, 1959-2017 [Supplemental material]. *The Journal of the American Medical Association*, *322*(20):1996-2016. `https://doi.org/10.1001/jama.2019.16932`

(c) [1 pt] Your approximate 95% equal-tailed posterior credible intervals for β_1 and β_2 should both contain zero. Explain why this is not surprising, given that the data values are differences between the individual state and overall US life expectancies.

2. Modify your analysis in the previous problem to use the *bivariate* prior formulation (as in the lectures on Random Effects and Hierarchical Models).

   (a) [3 pts] Present (i) your modified R code for the list structure containing all of the data (and possibly hyperprior constants), and (ii) your modified JAGS code. Make sure to satisfy the following:
   - Use centered predictors.
   - Choose whether to specify hyperprior constants in the R `list` (with the data) or directly in your JAGS `model` statement.
   - Give the response precision parameter τ_y^2 a Gamma(0.001, 0.001) prior.
   - Give the overall regression coefficient vector $\boldsymbol{\beta}$ a multivariate normal hyperprior with mean vector $\mathbf{0}$ and *inverse* covariance matrix $10^{-6}\boldsymbol{I}$.
   - Give the *inverse* covariance matrix $\boldsymbol{\Omega}^{-1}$ (for the $\boldsymbol{\alpha}$ vectors) a Wishart prior with

   $$\nu\,\boldsymbol{\Omega}_0 \quad = \quad \begin{bmatrix} 2 & 0 \\ 0 & 0.001 \end{bmatrix} \qquad\qquad \nu = 2$$

     for the arguments used by JAGS.
   - Define variables (deterministic nodes) for the standard deviation $\sigma_y = 1/\sqrt{\tau_y^2}$ of the response and the (population) covariance matrix $\boldsymbol{\Omega}$ of the regression parameter vectors (the $\boldsymbol{\alpha}$ vectors).
   - Also define variables (deterministic nodes) appropriate for answering part (c).

   (b) [4 pts] Run your model, and present a summary statistic table that includes at least the nodes for $\boldsymbol{\beta}$, σ_y, and $\boldsymbol{\Omega}$, and also the node(s) to answer part (c). (See also the instructions from part (b) of the previous problem.)

   (c) [1 pt] Let ρ be the (population) correlation between an intercept parameter α_1 and the corresponding slope parameter α_2. Approximate the posterior probability that ρ is greater than zero.

3. GRADUATE SECTION ONLY

   Consider Bernoulli (0 or 1) random variables Y_1 and Y_2.

   (a) Let X be *any* other discrete random variable, and suppose Y_1 and Y_2 are *conditionally iid* given X. In particular, they are identically distributed given X, so we may unambiguously define

   $$g(x) \quad = \quad \text{Prob}(Y_1 = 1 \mid X = x) \quad = \quad \text{Prob}(Y_2 = 1 \mid X = x)$$

   (i) [2 pts] By the law of total probability,

   $$\text{Prob}(Y_1 = 1, Y_2 = 1) \quad = \quad \sum_x \text{Prob}(Y_1 = 1, Y_2 = 1 \mid X = x)\ \text{Prob}(X = x)$$

   where the sum runs over all possible values x for X. Use this to show that

   $$\text{Prob}(Y_1 = 1, Y_2 = 1) \quad = \quad 0$$

   implies

   $$g(x) = 0 \quad \text{for all } x \text{ such that} \quad \text{Prob}(X = x) > 0$$

(ii) [1 pt] Show that

$$g(x) = 0 \quad \text{for all } x \text{ such that} \quad \text{Prob}(X = x) > 0$$
$$\text{implies}$$
$$\text{Prob}(Y_1 = 1) \;=\; \text{Prob}(Y_2 = 1) \;=\; 0$$

(b) Now suppose that

$$\text{Prob}(Y_1 = 1, Y_2 = 0) \;=\; \text{Prob}(Y_1 = 0, Y_2 = 1) \;=\; 1/2$$

(i) [1 pt] Briefly explain why Y_1 and Y_2 are exchangeable.

(ii) [1 pt] Show that there cannot exist any discrete random variable X such that Y_1 and Y_2 are *conditionally iid* given X. (Hint: Use part (a).)

[Remark: More generally, it can be shown that there cannot exist *any* random variable X (discrete or not) such that Y_1 and Y_2 are *conditionally iid* given X.]