PubH 7440: Introduction to Bayesian Analysis
Spring 2020, Homework 4
Due: April 10, 2020

1. Consider again the bike data in HW#2, which are on the number of bicycles at intersections for streets that are and are not bike routes. The goal is to analyze the mean rate of bicycles for each intersection (separately for the bike-route and non-bike-route streets) using the hierarchical model (M1)

$$Y_i|\lambda_i \sim Poisson(\lambda_i)$$

$$\lambda_i \sim Gamma(\alpha, rate = \beta)$$

$$\alpha \sim Gamma(0.01, rate = 0.01)$$

$$\beta \sim Gamma(0.01, rate = 0.01)$$

Where:

$$p(y_i|\lambda_i) = \frac{(\lambda_i)^{y_i} e^{-\lambda_i}}{y_i!}$$

and

$$p(\lambda_i|\alpha, \beta) = \frac{\lambda_i^{\alpha-1} \beta^\alpha e^{-\beta\lambda_i}}{\Gamma(\alpha)}$$

a. Edit your BUGS code in HW#2 to calculate the marginal posterior predictive p-value $P(y_i^{rep} > y_i|\boldsymbol{y})$ for each data point. What do they tell about overdispersion (greater variance than expected) in the data?

b. Compare the hierarchical model specified above with the following two models:

M2: Pooled model (separate for the bike-route and non-bike-route streets)

$$Y_i|\lambda \sim Poisson(\lambda)$$

$$\lambda \sim Gamma(0.01, rate = 0.01)$$

M3: Independent model

$$Y_i|\lambda_i \sim Poisson(\lambda_i)$$

$$\lambda_i \sim Gamma(0.01, rate = 0.01)$$

Calculate the DIC for each of the three models. Which is the best model for these data?

2. (Gelman, Chapter 14, problem 1): Consider the radon measurement data found on the canvas site. The data include the following variables:

    - $\text{County} = (1 = \text{Blue Earth}, 2 = \text{Clay}, 3 = \text{Goodhue})$
    - radon (pCi/L)
    - $\text{first\_floor} (0 = \text{basement}, 1 = \text{first floor})$
    - $\text{N} = \text{the number of observations}$

    a. Fit a linear regression to the logarithms of the radon measurements with indicator variables for the three counties and for whether a measurement was recorded on the first floor with a vague $NIG(0, 10^4 * I_4, 0.01, 0.01)$ prior. Provide the posterior mean, median, and 95% credible interval estimates for the regression parameters.

    b. Suppose another house is sampled at random from Blue Earth County. Provide summaries for the prediction distribution (2.5th percentile, medina, 97.5th percentile) assuming the measurements were completed in the basement and assuming the measurements were completed on the first floor.

    c. Consider an alternative regression model without the binary variable for whether a measurement was recorded on the first floor with the $NIG(0, 10^3 * I_4, 0.01, 0.01)$ prior. Calculate the DIC for each model. Which model would you choose?

    d. Now compare the two models using Bayes Factor. Which model would you choose?

3. (Gelman, Chapter 15, problem 2): Consider the penicillin production data found on the canvas site. The data include the following variables:

    - penicillin_yield
    - $\text{block1} – \text{block5}$ (indicators of the yield coming from blocks 1 through 5)
    - $\text{treatA} – \text{treatD}$ (indicators of the yield coming from treatments A through D)
    - $\text{N} = \text{the number of observations}$

    a. Fit a standard analysis of variance model to the data, that is, a linear regression with a constant term, indicators for all but one of the blocks, and all but one of the treatments. Summarize posterior inference for this model.

    b. Set up a hierarchical extension of the model, in which you have indicators for all five blocks and all five treatments, and the block and treatments are two sets of random effects. Explain why the means for the block and treatment indicator groups should be fixed at zero.

    c. Fit the hierarchical model in b using BUGS and summarize posterior inference.

    d. Calculate DIC for each model and determine which model is a better fit to the data.

4. (Gelman, Chapter 16, problem 5, a - d) Consider the horseshoe crab data on the canvas site. The data provide data relating the number of satellites for a female crab with the weight. The data contain the following variables:

   - $\text{satellite\_num} = (\text{number of satellites})$
   - weight (kg)
   - $\text{N} = \text{the number of observations}$

   a. Fit a standard analysis Poisson regression model relating the log of the expected count linearly to the predictor.

   b. Perform some model checking on the simple model proposed in (a), and see if there is evidence of overdispersion.

   c. Fit a hierarchical model assuming independent normally distributed errors.

   d. Is there evidence that this model provides a better fit to the data?