NYU SCHOOL OF GLOBAL PUBLIC HEALTH

Name: _____

Net ID: _____

GPH-GU 2372 Applied Bayesian Analysis in Public Health

FINAL EXAM

INSTRUCTIONS:

This take-home final exam starts at 9:25 PM U.S. Eastern Time on Dec. 12, 2022 and is due at 11:59 PM U.S. Eastern Time on Dec. 18, 2022.

The score of this final exam is capped at 35 points, though there are 5 bonus points. The number of points associated with each question is also noted.

You must submit your final-exam answer file and R code to NYU Brightspace no later than the submission deadline 11:59 PM U.S. Eastern Time on Dec. 18, 2022.

There is a 10-point deduction for each hour the final-exam submission is late.

This exam is "open book," which means you are permitted to use any materials handed out in class, your own notes from the course, the textbook, and anything on the course website.

The exam must be taken completely alone. Showing it or discussing it with anyone is forbidden.

You may not consult with any other person regarding the exam. You may not check your exam answers with any person. You may not discuss any of the materials or concepts in the course with any other person.

Points will be deducted for each question accordingly, if its R code is not available in your submission.

Points will be deducted if it is difficult to find or read your answers.

If any question, please contact the instructor by email.

Good Luck!

Final Exam

1. There is an increasing awareness that we should improve our life style. In Western Europe, a variety of campaigns have been set up in the last decades to give up smoking and to render our diet more healthy, for example by lowering our daily consumption of saturated fat. Around 1990, a dietary survey, the Inter-regional Belgian Bank Employee Nutrition Study (IBBENS) was set up to compare the dietary intake in different geographical areas in Belgium, especially in Flanders. The IBBENS study was performed in eight subsidiaries of one bank situated in seven Dutch-speaking cities in the north and in one French-speaking city in the south of Belgium. The food habits of 371 male and 192 female healthy employees with average age 38.3 years were examined by a 3-day food record with an additional interview. Our goal is to contrast the variability of cholesterol intake (`chol`) in mg/day between the subsidiaries to the variability within the subsidiaries. To achieve this goal, we model the data as follows:

$$\begin{aligned} y_{ij}|\theta_i,\sigma^2 &\overset{ind}{\sim} N(\theta_i,\sigma^2) && \text{for } j=1,\ldots,m_i;\ \ i=1,\ldots,n \\ \theta_i|\mu,\tau^2 &\overset{iid}{\sim} N(\mu,\tau^2) && \text{for } i=1,\ldots,n \\ \sigma^2 &\sim p(\sigma^2) \\ \tau^2 &\sim p(\tau^2) \\ \mu &\sim p(\mu) \end{aligned}$$

   (a) Describe in words what the various components of the hierarchical model represent and discuss how the model could be modified to account for additional sources of variability. (2pts)

   (b) Using the following prior specifications

$$\begin{aligned} p(\sigma^2) &= \text{Inverse Gamma}(\tfrac{\nu_0}{2},\tfrac{\nu_0\sigma_0^2}{2}) = \text{Inverse Gamma}(10^{-3},10^{-3}) \\ p(\tau^2) &= \text{Inverse Gamma}(\tfrac{\eta_0}{2},\tfrac{\eta_0\tau_0^2}{2}) = \text{Inverse Gamma}(10^{-3},10^{-3}) \\ p(\mu) &= \text{N}(0,10^6) \end{aligned}$$

   fit the hierarchical model above to the cholesterol intake data of the 8 subsidiaries using `R`. (The data is in the file "`IBBENS.csv`".)

   Run the MCMC algorithm long enough so that the effective sample size of each parameters is at least 1000. (3pts)

   (c) Make posterior inference on the average cholesterol intake in each of the eight subsidiaries: provide posterior means and 95% credible intervals. (2 pts)

   (d) Determine which subsidiary has the highest posterior probability of having employees with the highest cholesterol intake. Analogously, given the data, determine which subsidiary is more likely to have employees with the lowest cholesterol intake per day. (2 pts)

(e) Make posterior inference on the population average level of cholesterol intake: provide posterior mean and 95% credible interval. (2pts)

(f) Discuss whether the data suggest that there is more "between-subsidiaries" variability than "within-subsidiary" variability: provide posterior means and 95% credible intervals for the appropriate parameters to illustrate this point. (3pts)

(g) Produce an appropriate plot that illustrates the concept of shrinkage and discuss whether your analysis suggests that there is a large or moderate level of shrinkage. (2 pts)

2. In a double-blinded multicentric Randomized Clinical Trial (RCT with 36 centers), a total of 298 sportsmen and elderly people were treated for toenail dermatophyte onychomycosis with either of two oral medications: Itraconazol 250 mg daily (`treat=0`) or Lamisil 250 mg daily (`treat=1`). The patients received treatment for 12 weeks and were evaluated at 0, 1, 2, 3, 6, 9 and 12 months. As response, the unaffected nail length for the big toenail was measured. We want to determine whether the treatment works on average and if there is a difference among the treatments.
We analyze the data (the file "`toenail.txt`") using the linear fixed-effect regression model given by

$$y_{ij} = \beta_0 + \beta_1 \cdot t_{ij} + \beta_2 \cdot (t_{ij} \times \texttt{treat}_i) + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma^2) \tag{1}$$

where y_{ij} is the unaffected toenail length of patient (`id`) i at time $t_{ij} = 0, 1, 2, 3, 6, 9, 12$ months under treatment $\texttt{treat}_i$. No main effect for treatment was included in the model since at baseline the effect of treatment must be zero given the randomized character of the study.

(a) Fit the above fixed effect model using the following priors:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \sim N_3 \left(\begin{pmatrix} 2.5 \\ 0.6 \\ 0 \end{pmatrix}, 100 \times \mathbf{I}_3 \right)$$

$$\sigma^2 \sim \text{Inverse Gamma}(1, 3.7)$$

where $\mathbf{I}_3$ is the identity matrix of dimension 3.
Derive posterior means and 95% credible intervals for the regression coefficients and for σ^2 and determine whether (i) the unaffected toe nail grows significantly over time; (ii) there is a difference between the two treatments over time. (4 pts)

(b) Since each individual has his/her own physiological and biological characteristics and might respond to treatment differently, we repeat the data analysis by fitting the following random intercept and slope model:

$$y_{ij} = \beta_{0i} + \beta_{1i} \cdot t_{ij} + \beta_2 \cdot (t_{ij} \times \texttt{treat}_i) + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma^2) \tag{2}$$

where

$$\begin{aligned}
\begin{pmatrix} \beta_{0i} \\ \beta_{1i} \end{pmatrix} | \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \boldsymbol{\Sigma} &\overset{iid}{\sim} N_2\left(\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \boldsymbol{\Sigma}\right) \qquad i = 1, \ldots, 298 \\
\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} &\sim N_2\left(\begin{pmatrix} 2.5 \\ 0.6 \end{pmatrix}, 100 \times \mathbf{I}_2\right) \\
\boldsymbol{\Sigma} &\sim \text{Inverse Wishart}(4, 0.1 \times \mathrm{I}_2) \\
\beta_2 &\sim N(0, 100) \\
\sigma^2 &\sim \text{Inverse Gamma}(1, 3.7) \qquad (3)
\end{aligned}$$

Before fitting the model in (2), conduct an exploratory analysis by fitting the regression model

$$y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma_i^2)$$

separately for each patient i. Plot the estimated regression lines for each patient grouped by treatment (make two separate plots for the Itraconazol and the Lamisil group). In each plot also, draw the average evolution of toenail length over time by using the means of the estimated regression coefficients for all patients in each treatment group. Discuss what the plots reveal. (2 pts; here just fit the regression by the OLS method.)

(c) Fit the hierachical linear mixed model in (2) using the prior specifications in (3). Run the MCMC algorithm long enough so that all parameters have effective sample sizes of at least 1000. Assess convergence and report the posterior means and 95% credible intervals for $\beta_0, \beta_1, \beta_2$ and σ^2. Interpret the estimates: is there evidence for an increase in the unaffected toenail length over time? Do the two treatments have a different effect over time? (5 pts. Hint: use the "MCMCpack" package.)

(d) Let's compare the fixed effect model vs the mixed effect model: based on your posterior inference of the parameters, is there variability among patients in the way their unaffected toenail grows over time? (2 pts)

(e) Suppose two new patients are enrolled in the study, one taking Itraconzol and the other one taking Lamisil. Predict how long their unaffected toe nail will be at 0, 1, 2, 3, 6, 9 and 12 months since enrollment in the study and provide 95% confidence bands around your predictions (provide a plot of the predictions and the 95% confidence bands for for both patients). (4 pts)

3. Disease rates: The number of occurrences of a rare, nongenetic birth defect in a five-year period for six neighboring counties is $\boldsymbol{y} = (1, 3, 2, 12, 1, 1)$. The counties have populations of $\boldsymbol{x} = (33, 14, 27, 90, 12, 17)$, given in thousands. The second county has higher rates of toxic chemicals (PCBs) present in soil samples, and it is of interest to know if this town has a high disease rate as well. We will use the following hierarchical model to analyze these data:
   - $Y_i|\theta_i, x_i \sim \text{Poisson}(\theta_i x_i)$;
   - $\theta_1, \ldots, \theta_6|a, b \sim \text{gamma}(a, b)$;
   - $a \sim \text{gamma}(1,1)$; $b \sim \text{gamma}(10,1)$.

a) Describe in words what the various components of the hierarchical model represent in terms of observed and expected disease rates. (2 pts)

b) Identify the form of the conditional distribution of $p(\theta_1, \ldots, \theta_6|a, b, \boldsymbol{x}, \boldsymbol{y})$, and from this identify the full conditional distribution of the rate for each county $p(\theta_i|\boldsymbol{\theta}_{-i}, a, b, \boldsymbol{x}, \boldsymbol{y})$. (1 bonus pt)

c) Write out the ratio of the posterior densities comparing a set of proposal values $(a^*, b^*, \boldsymbol{\theta})$ to values $(a, b, \boldsymbol{\theta})$. Note the value of $\boldsymbol{\theta}$, the vector of county-specific rates, is unchanged. (1 bonus pt)

d) Construct a Metropolis-Hastings algorithm which generates samples of $(a, b, \boldsymbol{\theta})$ from the posterior. Do this by iterating the following steps:
   1. Given a current value $(a, b, \boldsymbol{\theta})$, generate a proposal $(a^*, b^*, \boldsymbol{\theta})$ by sampling a^* and b^* from a symmetric proposal distribution centered around a and b, but making sure all proposals are positive (see Exercise 10.1). Accept the proposal with the appropriate probability. (See Exercise 10.1 on textbook page 244)
   2. Sample new values of the θ_j's from their full conditional distributions.

   Make posterior inference on the infection rates using the samples from the Markov chain. In particular,

   i. Compute marginal posterior distributions of $\theta_1, \ldots, \theta_6$ and compare them to $y_1/x_1, \ldots y_6/x_6$. (1 bonus pt)

   ii. Examine the posterior distribution of a/b, and compare it to the corresponding prior distribution as well as to the average of y_i/x_i across the six counties. (1 bonus pt)

   iii. Plot samples of θ_2 versus θ_j for each $j \neq 2$, and draw a 45 degree line on the plot as well. Also estimate $\Pr(\theta_2 > \theta_j|\boldsymbol{x}, \boldsymbol{y})$ for each j and $\Pr(\theta_2 = \max\{\theta_1, \ldots, \theta_6\}|\boldsymbol{x}, \boldsymbol{y})$. Interpret the results of these calculations, and compare them to the conclusions one might obtain if they just examined y_j/x_j for each county j.
   (1 bonus pt)