Applied Bayesian modeling - Exam 2, fall 2022

General information

General instructions, shared prior to the exam as well

- You have 24 hours to complete the exam. The start time of the exam is determined by when you download the exam from Moodle (hence do NOT download the exam before you can start it). Your end time is given by the time stamp of your earliest Moodle download (which I have access to) + 24 hours.

- Submission materials for the exam are: (i) Rmd and knitted Rmd with answers, including the code to produce the answers, and (ii) Exam statement: I uploaded a word document for you to fill out to state that you worked independently and did not consult with anyone but Leontine Alkema. Please use your student ID (not name) as title of the Rmd files.

- You have to hand in the submission materials for the exam within 24 hours by submitting it on moodle under the "Exam 2 submission" section. Exams that are submitted late are subject to severe penalization so make sure that you submit it in time. If you have not yet submitted your exam within 23 hours, you are required to upload a version then (so at the 23 hours deadline) to avoid last-minute submission issues. If you are concerned about technical issues, you may also email me your exam, in addition to submitting it on moodle.
- I am reachable by email for clarifying questions (only) during the exam. I will respond asap during regular office hours (830-5 on W-F). I will try to respond asap to emails received during other times as well but cannot guarantee a fast response time.
- You have to work on this exam independently and you are not allowed to consult with anyone except me (Leontine Alkema). This means that you are NOT allowed to seek input regarding the exam in any form from anyone, i.e. you are not allowed to discuss the exam with anyone whether in person or in some virtual form. You are NOT allowed to post on the course slack. You are also NOT allowed to share any information regarding the exam with anyone until December 23. I am taking this very seriously (as it's the only way I'm able to allow a take home exam). Any evidence of breaking this rule will be reported to the graduate school.

Grading for this exam

This exam contains a total of 6 questions, including one question for extra credit (final question 6). The number of points assigned to each question is added with the question. The total score excluding extra credit is 35, the extra credit question is worth 5 points.

Info about the data and outcome of interest

In this exam, we examine the same outcome of interest as in Exam 1 but consider a different data set and questions/models.

We consider data y_i for $i = 1, .., n$, where y_i refers to a health score calculated for an individual i. The health score can be any value, more negative health scores indicate poorer health while more positive health scores indicate better health. In addition to an individual i's health score y_i, the available data sets also includes individuals' age a_i (with ages ranging from 15 to 65) and their county of residence, denoted by index $j[i]$.

The data set is saved in `dat_exam2_fall2022.csv`, where `y` refers to the health score, `county` to county, and `age` to age.

Question 1 (10 points)

Consider the following Bayesian model, referred to in the remainder as model 1:

$$\begin{aligned} y_i|\alpha_{j[i]}, \beta, \sigma_y &\overset{i.i.d}{\sim} N(\alpha_{j[i]} + \beta(a_i - 30), \sigma_y^2), \\ \alpha_j|\mu_\alpha, \sigma_\alpha &\overset{i.i.d}{\sim} N(\mu_\alpha, \sigma_\alpha^2), \end{aligned}$$

with brm-default priors on model parameters $\beta, \sigma_y, \mu_\alpha, \sigma_\alpha$.

Fit model 1 to the data set and check MCMC-related diagnostics including Rhat and effective samples sizes. If these diagnostics suggests issues, check for coding errors and/or change MCMC-related parameters such that the resulting fit can be used for inference.

To hand in:
- Code to do model fitting and printed summary of model fit for the model that you want to use for inference.
- Report the lowest values of Rhat and the lowest effective sample sizes among the parameters $\beta, \sigma_y, \mu_\alpha, \sigma_\alpha$, and discuss briefly whether these values indicate issues or not.

Question 2 (5 points)

Continuing with model fit 1, provide a point estimate, 50% credible interval (NOT 95%), and interpretation of the estimates, for each of the following parameters: $\sigma_y, \sigma_\alpha, \mu_\alpha, \beta, \alpha_2$. Provide a context-specific interpretation of the parameters, do NOT use the terms intercept or slope in your interpretation.

Question 3 (5 points)

Continuing with model fit 1, obtain the posterior predictive probability that a yet-to-be-sampled individual with age $a = 20$ in a yet-to-be-sampled county has a health outcome greater than 10.

In your answer, in additon to producing and reporting the outcome of interest, also introduce notation and give an expression for the probability using the samples of model parameters, or, if needed, using samples obtained in additional sampling steps. If using additional sampling steps, explain with additional equations how those samples are obtained.

Question 4 (10 points)

Consider the following Bayesian model, referred to in the remainder as model 2:

$$y_i|\alpha_{j[i]}, \sigma_y, \beta_{j[i]} \overset{i.i.d}{\sim} N(\alpha_{j[i]} + \beta_{j[i]} \cdot (a_i - 30), \sigma_y^2), \quad (1)$$

$$(\alpha_j, \beta_j)|\boldsymbol{\mu}, \boldsymbol{\Sigma} \overset{i.i.d}{\sim} N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (2)$$

where $\boldsymbol{\mu} = (\mu_\alpha, \mu_\beta)$ and $\Sigma_{11} = \sigma_\alpha^2$, $\Sigma_{12} = \Sigma_{21} = \rho\sigma_\alpha\sigma_\beta, \Sigma_{22} = \sigma_\beta^2$, using brm-default priors on model parameters $\sigma_y, \beta, \mu_\alpha, \mu_\beta, \sigma_\alpha, \sigma_\beta, \rho$.

In case it's helpful, note that an equivalent way of writing the model for (α_j, β_j) is as follows:

$$\alpha_j = \mu_\alpha + \eta_{0,j}, \quad (3)$$

$$\beta_j = \mu_\beta + \eta_{1,j}, \quad (4)$$

$$(\eta_{0,j}, \eta_{1,j})|\boldsymbol{\Sigma} \overset{i.i.d}{\sim} N_2(\mathbf{0}, \boldsymbol{\Sigma}). \quad (5)$$

Fit model 2 to the data set and check MCMC-related diagnostics including Rhat and effective samples sizes. If these diagnostics suggests issues, check for coding errors and/or change MCMC-related parameters such that the resulting fit can be used for inference.

To hand in:
- Code to do model fitting and printed summary of model fit for the model that you want to use for inference.
- A visualization of the relation between age and health scores in the first 4 counties. In particular, create a plot for each county, that shows the data for that county, the estimated county-specific relation between age and health scores, and a regression line given by $\mu_\alpha + \mu_\beta \cdot (a - 30)$.

Note: if you are unable to produce a fit for model 2, you may do the visualization based on model 1. In that case, instead of plotting the regression line given by $\mu_\alpha + \mu_\beta \cdot (a - 30)$, plot the line given by $\mu_\alpha + \beta \cdot (a - 30)$.

Question 5 (5 points)

Suppose that one of the outcomes of interest in this study is the variability in health scores at higher ages, say ages 50+. Carry out a posterior predictive check to verify whether model 1 and/or model 2 reasonably capture the variability in health scores at age 50+.

In your answer, make sure to specify a summary statistic (in an equation) and visualize and calculate how extreme the summary statistic is when evaluated for the real data, as compared to the statistic as evaluated in replicated data sets.

Note: if you were unable to produce a fit for model 2, produce the results for model 1 only and explain what, in comparison to your finding for model 1, would be a worse model outcome.

Question 6 (extra credit, 5 points)

Continuing with model fit 2, for a yet-to-be-sampled county, obtain the probability that a yet-to-be-sampled individual with age $a = 60$ has a greater health score than a yet-to-be-sampled individual with age $a = 20$.

In your answer, in additon to producing and reporting the outcome of interest, also introduce notation and give an expression for the probability using the samples of model parameters, or, if needed, using samples obtained in additional sampling steps. If using additional sampling steps, explain with additional equations how those samples are obtained.

Note: If you were unable to produce a model fit for model 2, produce the results for model 1, and add the equations for model 1 as well as model 2 for full score.