Applied Bayesian modeling - Exam 1, fall 2022. Solutions

General information

Grading: This exam contains 4 questions that have 8 sub-questions in total, including one sub-question for extra credit (final sub-question 4b). Each sub-question is worth 10 points. Hence the total score is 70 points, plus an additional 10 points extra credit.

This is a closed-book exam. You may use that when $y|\mu,\sigma^2 \sim N(\mu,\sigma^2)$, then $p(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-1}{2\sigma^2}(\mu-y)^2\right)$.

Info about the data and outcome of interest

In this exam, 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, some available data sets also include individuals' age a_i (with ages ranging from 15 to 65) and their county of residence, denoted by index $j[i]$.

Each of the 4 questions below concerns a different model to describe the outcome of interest. You may assume model assumptions hold true when answering questions about specific model fits.

The first rows of the data sets used in questions 1 and 3 are as follows:

```
head(dat1)
```

```
## # A tibble: 6 x 1
##        y
##    <dbl>
## 1  2.93
## 2  3.55
## 3 -0.547
## 4 -2.27
## 5  3.03
## 6 -9.09
```

```
head(dat3)
```

```
## # A tibble: 6 x 2
##         y county
##     <dbl>  <int>
## 1  0.402       1
## 2 -1.49        1
## 3  0.0908      1
## 4  0.410       1
## 5 -9.88        1
## 6 -5.53        1
```

Question 1

Suppose you have fitted the following model, referred to as model 1, to the data set `dat1`:

$$y_i|\mu,\sigma^2 \overset{i.i.d}{\sim} N(\mu,\sigma^2), \text{ for } i = 1,2,\ldots,n, \tag{1}$$

with vague priors on μ and σ. The `brm`-function call and summary output are included at the end of the exercise.

Suppose that you have to present results from model fit 1 to other researchers. The researchers understand the likelihood function used in model 1. They are also familiar with probability density functions and Monte Carlo approximations. However, they have never heard about Bayesian inference or MCMC algorithms.

Based on this information, answer questions a and b below. Your answers may be informal as long as they introduce relevant information on the specific terms. Please be brief in your answers. Points may be subtracted if additional incorrect statements are added to otherwise correct answers.

a. Explain to the researchers what a posterior distribution for μ is, and how it relates to a prior distribution and likelihood function. In your explanation, introduce the terms "prior density" and "posterior density" and "Bayes rule".
b. In brm-output, ESS refers to effective sample size. Explain to the researchers why effective sample size for model parameters is lower than the total number of samples obtained (for example, the largest ESS is 1622, while the total number of samples obtained is 2000). In your answer, introduce the terms "MCMC algorithm", and "autocorrelation" to explain "effective sample size".

`brm`-function call and summary output:

```
fit1  <- brm(y ~ 1, data = dat1,
          chains = 4, iter = 1000, warmup = 500, cores = getOption("mc.cores", 4),
            seed = 12345, file_refit = "on_change", file = "output/exam_fit1")
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: y ~ 1
##    Data: dat1 (Number of observations: 100)
##   Draws: 4 chains, each with iter = 1000; warmup = 500; thin = 1;
##          total post-warmup draws = 2000
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     1.24      0.53     0.16     2.25 1.00     1565     1177
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     5.60      0.41     4.89     6.46 1.00     1622     1150
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Solution

Answers:

a. In Bayesian inference, we use densities to reflect/quantify our state of knowledge about a parameter of interest. The prior density refers to the density of μ, $p(\mu)$, prior to observing data. The likelihood function $p(y|\mu)$ states how the data relate to the parameter of interest. The posterior density $p(mu|y)$ is based on prior information and data; Bayes rule states that $p(\mu|y) \propto p(\mu)p(y|\mu)$.

b. In the brm-fit, we obtain samples from the posterior distribution using a Markov chain Monte Carlo algorithm. In this algorithm, we do not draw independent samples but instead, each sample depends on the previous one (it is a Markov Chain). This can result in autocorrelation in the sampled values. Effective sample size refers to the number of independent samples that give the same precision as obtained from the autocorrelated MCMC samples. Because of autocorrelation, MCMC samples typically result in fewer independent samples as compared to the total number of samples obtained.

Question 2

A second model, referred to as Model 2, is fitted to data on health scores:

$$y_i|\alpha, \sigma^2, \beta \overset{i.i.d}{\sim} N(\alpha + \beta \cdot (a_i - 30), \sigma^2), \text{ for } i = 1, 2, \ldots, n, \tag{2}$$

with vague priors on α, σ^2 and β.

Suppose you are given posterior samples of all model 2 parameters and assume model assumptions hold true. Explain how you would obtain the following probabilities based on the model 2 fit:

a. Posterior probability that α is greater than 0.
b. Posterior predictive probability that a yet-to-be-sampled individual with age $a = 25$ has a health outcome greater than 0.

In your answers, 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.

Solution

Yes we can obtain both probabilities as follows:

a: Let $\alpha^{(s)}$ refer to posterior samples, i.e. $\alpha^{(s)} \sim p(\alpha|\boldsymbol{y})$. We have obtained these samples from the model fit. We can use the samples to approximate the probability as follows: $P(\alpha > 0|\boldsymbol{y}) \approx 1/S \sum_s 1(\alpha^{(s)} > 0)$.

b: We sample $\tilde{y}^{(s)}|\alpha^{(s)}, \beta^{(s)}, \sigma^{(s)} \sim N(\alpha^{(s)} + \beta^{(s)} \cdot (25 - 30), \sigma^2)$, where $\alpha^{(s)}, \beta^{(s)}, \sigma^{(s)} \sim p(\alpha, \beta, \sigma|\boldsymbol{y})$. Then $P(\tilde{y} > 0|\boldsymbol{y}) \approx 1/S \sum 1(\tilde{y}^{(s)} > 0)$.

Question 3

Suppose you have fitted the following model, referred to as model 3, to the data set `dat3`:

$$y_i|\alpha_{j[i]}, \sigma_y \overset{i.i.d}{\sim} N(\alpha_{j[i]}, \sigma_y^2), \tag{3}$$

$$\alpha_j|\mu_\alpha, \sigma_\alpha \overset{i.i.d}{\sim} N(\mu_\alpha, \sigma_\alpha^2), \tag{4}$$

with vague priors on μ_α, σ_y, σ_α. The `brm`-function call and summary output are included at the end of the exercise.

a. Based on the model 3 `brm`-based `fit3`, what is $\hat{\sigma}_\alpha$? Interpret its outcome in the context of the outcome of interest and data set (i.e., do not have names of model parameters in your interpretation, just words that refer to the context).

b. Suppose we simulated data sets A and B from the following model:

$$\alpha_j \overset{i.i.d}{\sim} N(\hat{\mu}_\alpha, \hat{\sigma}_\alpha^2), \text{ for } j = 1, \ldots, J \tag{5}$$

$$y_i|\alpha_{j[i]} \overset{i.i.d}{\sim} N(\alpha_{j[i]}, \hat{\sigma}_y^2), \text{ for } i = 1, \ldots, n, \tag{6}$$

with

- the same number of counties and observations per county as in `dat3`,
- $\hat{\sigma}_y, \hat{\mu}_\alpha$ fixed at the outcomes from `fit3`,
- $\hat{\sigma}_\alpha$ fixed at the outcome of `fit3` for data set A while for data set B, $\hat{\sigma}_\alpha$ is fixed at a very large value, i.e., 1000 times the estimated value from `fit3`.

(i) First explain briefly (in 1 sentence) how the data are expected to differ systematically between the two simulated data sets.

(ii) Suppose the simulated data sets each have some county j^* with small sample size n_{j^*} and a large difference between $\bar{y}_{j^*}$ and $\bar{y}$. When fitting model 3 to simulated data set A and separately, fitting model 3 to simulated data set B, how would shrinkage of $\hat{\alpha}_{j^*}$ from $\bar{y}_{j^*}$ toward $\bar{y}$ or $\hat{\mu}_\alpha$ in the data-set-specific model fits differ between the two settings? I.e., would there be more shrinkage of the estimate when using data from simulation A or simulation B?

`brm`-function call and summary output:

```
fit3  <- brm(y ~ (1|county), data = dat3,
          chains = 4, iter = 2000, warmup = 1000, cores = getOption("mc.cores", 4),
          seed = 123456, file_refit = "on_change", file = "output/exam_fit3")
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: y ~ (1 | county)
##    Data: dat3 (Number of observations: 2000)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Group-Level Effects:
## ~county (Number of levels: 50)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     1.59      0.20     1.24     2.02 1.00     1519     2162
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     0.44      0.25    -0.07     0.94 1.00     1595     2209
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     4.56      0.07     4.42     4.70 1.00     6966     3146
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Solution

Answers:

a. $\hat{\sigma}_\alpha = 1.59$, this parameter refers to the standard deviation of the county mean outcomes for the health outcome.

b. (i) When simulating data from model B, we get increased variability in group means α_j, and thus in (the means of) observed data, as compared to model A.

c. (ii) We get more shrinkage in data set A because variability between group means is much smaller in that data set. You can refer to the full conditional of α_j to see that the weight associated with μ_α decreases as σ_α increases. Hence, when keeping other aspects constant (sample size and σ_y), there is lower weight associated with μ_α in data set B as compared to data set B.

When thinking about this problem, it is also helpful to think about extreme cases. As σ_α goes to infinity, the hierarchical model set up results in less pooling/shrinkage because there is more uncertainty associated with the α_j's. An informal way to think about this is that the density for the group means that is imposed by the hierarchical model becomes more comparable to a vague prior (given the large variance). V.v., when σ_α goes to zero, the α_j become identical, and as an effect, they are all shrunk to the overall mean.

Question 4

Model 4 is specified as follows:

$$y_i|\alpha_{j[i]}, \sigma_y, \beta \overset{i.i.d}{\sim} N(\alpha_{j[i]} + \beta \cdot (a_i - 30), \sigma_y^2), \tag{7}$$

$$\alpha_j|\mu_\alpha, \sigma_\alpha \overset{i.i.d}{\sim} N(\mu_\alpha, \sigma_\alpha^2), \tag{8}$$

with vague priors on $\sigma_y, \beta, \mu_\alpha, \sigma_\alpha$.

a. Suppose that we expect the association between health score and age to vary by group but that data alone are insufficient to estimate group-specific coefficients independently for each group. Can you extend model 4 to incorporate the additional assumption?
If yes, do so and give the model equations. If not, explain why not.

b. (Extra credit question) For model 4 (NOT the model extended in 4a), obtain the full conditional density for $\alpha_j|\boldsymbol{y}, \mu_\alpha, \beta, \sigma_y, \sigma_\alpha$, with the a_i known. You may use that if $y_i|\mu, \sigma^2 \sim N(\mu, \sigma^2)$ (independent) and if $\mu|\gamma, \delta \sim N(\gamma, \delta^2)$, then $\mu|\boldsymbol{y}, \sigma^2, \gamma, \delta \sim N\left(\frac{\gamma/\delta^2 + n\cdot\bar{y}/\sigma^2}{1/\delta^2 + n/\sigma^2}, \frac{1}{1/\delta^2 + n/\sigma^2}\right)$.
Hint: consider introducing $z_i = y_i - \beta(a_i - 30)$.

Solution

a. Yes, we can make the regression coefficient for age group-specific as follows:

$$y_i|\alpha_{j[i]}, \sigma_y, \beta_{j[i]} \overset{i.i.d}{\sim} N(\alpha_{j[i]} + \beta_{j[i]} \cdot (a_i - 30), \sigma_y^2), \tag{9}$$

$$\boldsymbol{\theta}_j|\mu_\alpha, \mu_\beta, \sigma_\alpha, \sigma_\beta, \rho \overset{i.i.d}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{10}$$

where $\boldsymbol{\mu} = (\mu_\alpha, \mu_\beta)$ and Σ is the variance-covariance matrix with σ_α^2 and σ_β^2 on the diagonal and $\rho\sigma_\alpha\sigma_\beta$ off-diagonal.

b. Full conditional:

Let $z_i = y_i - \beta(a_i - 30)$. Then

$$z_i|\alpha_{j[i]}, \sigma_y, \beta \overset{i.i.d}{\sim} N(\alpha_{j[i]}, \sigma_y^2). \tag{11}$$

Using this equation, the general result applied to data from group j, and using that $\alpha_j|\mu_\alpha, \sigma_\alpha \sim N(\mu_\alpha, \sigma_\alpha^2)$, we get that

$$\begin{aligned}
\alpha_j|\boldsymbol{y}, \mu_\alpha, \beta, \sigma_y, \sigma_\alpha &\sim N(m, v), \\
v &= \left(n_j/\sigma_y^2 + 1/\sigma_\alpha^2\right)^{-1}, \\
m &= v \cdot \left(\frac{n_j}{\sigma_y^2}\bar{z}_j + \frac{1}{\sigma_\alpha^2}\mu_\alpha\right) = \frac{\frac{n_j}{\sigma_y^2}(\bar{y}_j - \beta\bar{a}_j) + \frac{1}{\sigma_\alpha^2}\mu_\alpha}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}}.
\end{aligned}$$