CHL 5223H Applied Bayesian Methods

Homework 2
Tuesday, February 5, 2019
DUE: MONDAY, March 4, 2019

1. (Data from Kleinbaum, Kupper, Muller, and Nizam, pg108) A biologist wished to explore the relationship of temperature on the growth process of a certain type of tissue. Using the same parent batch, the scientist cultured 5 different cell lines for 4 different temperatures (for a total of 20 different lines). The total number of cells sampled for the different temperatures are given in the following table: (Note: numbers are $\times 10^6$ after 7 days.)

Temp=40	Temp=60	Temp=80	Temp=100
1.13	1.75	2.30	3.18
1.20	1.45	2.15	3.10
1.00	1.55	2.25	3.28
0.91	1.64	2.40	3.35
1.05	1.60	2.49	3.12

For this homework, please assume we have the following model. Let the observation for the j-th sample of the i-th temperture level be Y_{ij} (Label the temperature levels: 1, 2, 3, and 4). Let the these observations be normally distributed about the i-th temperature effect mean μ_i with a common precision τ. Let the temperature effects means be normally distributed about a grand mean μ_0 with precision τ_0. Assume that the precisions τ and τ_0 have gamma priors with parameter values $\alpha = 1$ and $\beta = 1^2$. Also, let μ_0 have a normal prior with mean 0 and precision 1^2. Therefore, we have the following model:

$$\begin{aligned} Y_{ij}|\mu_i, \tau &\sim \text{Normal}(\mu_i, \tau) \\ \mu_i|\mu_0, \tau_0 &\sim \text{Normal}(\mu_0, \tau_0) \\ \mu_0 &\sim \text{Normal}(0, 1^2) \\ \tau &\sim \text{Gamma}(1, 1^2) \\ \tau_0 &\sim \text{Gamma}(1, 1^2). \end{aligned}$$

Do the following:

(a) Please give the distribution of:
   i. $f(\mu_i|\text{DATA}, \mu_0, \tau, \tau_0)$
   ii. $f(\mu_0|\text{DATA}, \mu_1, \ldots, \mu_4, \tau, \tau_0)$
   iii. $f(\tau|\text{DATA}, \mu_1, \ldots, \mu_4, \mu_0, \tau_0)$
   iv. $f(\tau_0|\text{DATA}, \mu_1, \ldots, \mu_4, \mu_0, \tau)$

(b) Use R to run a Markov chain Monte Carlo algorithm on this data. (DO NOT USE WinBugs, Openbugs, Jags or any of their relatives. Th purpose is for you to do some basic programing.) You only need to run the MCMC for about 20,000 iterations for this example. Please supply your computer code.

(c) Please give the posterior distribution (a plot) for the mean of each temperature effect (the μ_i's) as well as some of the summary statistics of these posterior distributions for the drug effect means (mean, standard deviation, and some type of 95% credible region is sufficient).

(d) What is the posterior distribution of the difference in number of cells grown at a temperature of 40 versus 80? What is the posterior probability that there will be more cells grown at a temperature of 40 versus 80?

2. The following table gives the number of lung cancers for different age groups and different histories of smoking. Also, given is the number of person years where a potential cases was exposed (was eligible to be identified by the study as having lung cancer) and was in a particular age group-smoker class. (Data is from page 495, Selvin, *Practical Biostatistical Methods*, 1995, (Belmont, CA: Wadsworth).) Do the following with this data.

(a) Do a bayesian poisson regression on this data. Model the logarithm of the rate of deaths (per person-years) as a linear function of age and smoking category. Model both the smoking category and the age groupings as discrete variables. Please give the WinBugs code for this data. Provide the univariate posterior distributions of each level of smoking and age and also give the posterior distribution of the precisions used in your model. (If you wish, you may report the "standard deviations" instead of the precisions.)

(b) Provide a brief justification that your model has converged and that you have "burned-in" your simulation enough. (Note: you may give "thinned" plots of your simulations to avoid printing out plots which are suppose to represent several thousands or hundreds of thousand points. For the purpose of this question, you can provide simple diagnostics like trace/history plots and autocorrelation plots.)

(c) Provide your belief as to the increase probability of death for someone who smokes $>$ 20 cigarettes per day versus a nonsmoker. (That is, you would say that you believe that a smoker who smokes $>$ 20 cigarettes per day has an increase risk of xxx times the risk of a nonsmoker.) Please, include both a "point" estimate of this increase risk and also state some interval estimate. Please identify the type of point and interval estimate that you are using.

Age/Smoking	Never smoker	Past smoker	≤ 20 cigarettes per day	> 20 cigarettes per day
<45				
Deaths	11	6	4	5
Person-years	114616	58259	19482	12947
45-54				
Deaths	20	13	2	6
Person-years	101015	64836	11641	7450
55-64				
Deaths	31	19	4	11
Person-years	78405	48952	8295	6823
65-74				
Deaths	40	17	8	16
Person-years	49216	30296	5031	4024
>74				
Deaths	93	50	20	31
Person-years	58269	31854	6401	5032

3. (There is first preamble and background section. This is followed by the things to do for this problem.)

   Preambel and Background:

   First, there is a description of a data analysis problem which includes a data file and a model for the data. You are asked to fit the model to the data with a Bayesian analysis. Use Openbugs or Jags (or a similar program) to fit the model. Please show the model file used by the software program and provide some preliminary evidence that that MCMC has converged. (That is, it is sufficient to use such methods as trace plot examination and looking at a plot of the auto-correlation function.) Afterwards, there are additional questions about the model that you will need to answer.

   The data in the CSV file named DiabetesDrugEffect.csv which contains data from 12 different clinical trial. The purpose of each of the clinical trials was to see if drug A was better than drug B. (Aside, they were looking to lower blood glucose in diabetic patients.) The outcome results were measured on a continuous scale. The column labeled "diff" was the main result in the study and is the overall average difference between drug A and drug B. The column "Sediff" is the estimate of the standard error of the difference estimate. The column "StudyN" is the number of subjects that were in each of the studies. The studies are labeled with the variable "StudyId" and take on values from 1 to 12. Note that the values from each study are obtained from the individual study publications.

   The goal is to combine the evidence in the 12 different studies. That is, the goal is to use the 12 different studies to obtain an overall belief in the average difference between the two drugs. In the medical literature, this is called a meta-analysis. Here, we make the

assumption that for each study site there are unknown site specific conditions so that each site has a different effect and that these individual effects can be modeled as a normal distribution around an overall average different. That is, the following model is assumed.

- Define θ as the overall "true" effects of the difference between the two drugs. The main goal of this analysis is to learn about this parameter.
- Define δ_i be the true difference between treatments for the i-th study. These study effects are normally distributed around θ with a common variance of σ_0^2. The parameter σ_0^2 is of moderate interest.
- Let Y_i be the observed difference for the i-th study. This is the estimate for δ_i from the original paper from the i-th study.
- Let S_i be the observed standard error of the estimator Y_i of δ_i. It is assumed one can ignore the error in using the observed value S_i for the true value. That is, here one is assuming that the observed S_i is the exact value.

The implies the following model which we call model 1:

$$Y_i|\delta, S_i \sim \text{Normal}(\text{mean} = \delta_i, \text{variance} = S_i^2)$$
$$\delta_i|\theta, \sigma_0^2 \sim \text{Normal}(\text{mean} = \theta, \text{variance} = \sigma_0^2)$$

Sometimes one is not concerned with the latent variables δ_i. So, one can collapse the model into the following collapsed model which we call model 2:

$$[Y_i|\theta, \sigma_0^2] \quad \sim \quad \text{Normal}(\text{mean} = \theta, \text{variance} = S_i^2 + \sigma_0^2)$$

(Note: in Openbugs/Jags, when one is using the normal distribution it is correct that the first parameter is still the mean but the second parameter is the precision which is 1/variance.)

Do the following for question 3:

(a) Fit both model 1 and model 2 with Openbugs/Jags. Provide the model file. Provide evidence that the MCMC has converged. (You don't need to provide advance statistics. Trace plots and plot of the auto-correlation function with the appropriate discussion is sufficient for the purpose of this exam.) Also, provide summary statistics of the posterior distributions. (The statistics such as the posterior mean, posterior standard deviation, and credible region obtain from a print statement is sufficient.)

(b) For model 1 and model 2, compare the estimates of the posterior distribution of θ and σ_0 from the two models. Are they close?

(c) Using model 1, one gets an estimate of the posterior distribution of the individual study effects, δ_i. That is, one can get the posterior mean, standard deviation and credible region. Also, from the original, individual study data, one can use the variables "diff" and "Sediff" to get an estimate of the study effects from the original

studied publications. Therefore, one can use these values to obtained the estimated effect from the study and the confidence interval for the estimate. Compare the estimates from the individual studies to the estimates from the meta-analysis. For each study compare the values of "diff" and the associated confidence interval with the posterior mean of the δ_i estimate and the associated credible region.

i. You will probably find that the estimate of the individual study effect using "diff" will be different than the estimate using the posterior mean of δ. When considering the movement of the estimate from the value "diff" to the posterior mean of δ, what general direction is this movement?

ii. When considering the length of the confidence intervals (using the individual study publications) compared to the credible region (obtained from the Bayesian meta-analysis), what generally happens to the change in the lenght?

iii. When looking at study 2, note that this study had the largest value of "Sediff" which is the standard error of the estimate in the original study. You might notice that there is a large difference between "diff" compared to the posterior mean of δ compared to most of the other studies. What might be causing this large shift. (Look at some other features of this study.)