BST 701: Bayesian Modeling in Biostatistics
Breheny

Assignment 2
Due: Thursday, February 14 ♡

1. Write an `R` function called `gibbs` that carries out Gibbs sampling for the following model:

$$\begin{aligned} X_i &\overset{\text{iid}}{\sim} \text{N}(\mu_x, \sigma^2) \\ Y_i &\overset{\text{iid}}{\sim} \text{N}(\mu_y, \sigma^2) \\ \mu_x &\sim \text{N}(\mu_{x,0}, \omega_{x,0}^{-1}) \\ \mu_y &\sim \text{N}(\mu_{y,0}, \omega_{y,0}^{-1}) \\ \text{RSS}_0\tau &\sim \chi^2(n_0), \end{aligned}$$

   where $\tau = 1/\sigma^2$. In particular, note that we are assuming equal variances for X and Y. The function should accept the following arguments:

   - `x`: The observed vector of X values
   - `y`: The observed vector of Y values
   - `N`: The requested number of samples from the posterior
   - `mu.x0`: The prior mean for μ_x
   - `mu.y0`: The prior mean for μ_y
   - `omega.x0`: The prior precision for μ_x
   - `omega.y0`: The prior precision for μ_y
   - `n0`: The prior equivalent sample size for τ
   - `RSS0`: The prior equivalent RSS for τ

   You may wish to include default values for all arguments following `x` and `y`. The function should return a matrix or data frame with `N` rows and three columns, labeled `mu.x`, `mu.y`, and `tau`, containing draws from the posterior $p(\mu_x, \mu_y, \tau|\mathbf{x}, \mathbf{y})$. The function must actually carry out the sampling from within `R` – having `gibbs` call JAGS, for example, is not allowed.

2. Suppose that a study is to be carried out in which an intervention intended to reduce fatal collisions between traffic and pedestrians will be applied at the county level, with some counties receiving the intervention and other counties serving as controls. Suppose we intend to model the data from this study using the following model:

$$\begin{aligned} \text{Intervention:} \quad X_i &\overset{\text{iid}}{\sim} \text{Pois}(\lambda_1) \\ \text{Control:} \quad Y_i &\overset{\text{iid}}{\sim} \text{Pois}(\lambda_2). \end{aligned}$$

   You are eliciting an informative prior from the principal investigator, who believes that "this intervention is likely to reduce the rate of fatal collisions by 30%. I think it's quite unlikely that it will have no effect, and quite unlikely that it will cut the rate in half." The investigator

does not wish to put an informative prior on the overall (average) fatal collision rate. You wish to parameterize the prior in terms of $\log \lambda$ rather than λ directly, to allow easy use of the normal distribution without having to worry about negative λ values. Take "quite unlikely" to mean that the investigator believes that there is a 90% chance that the true decrease in collision rate will be between 0% and 50%. What are the appropriate priors[1] for λ_1 and λ_2?

3. Duchenne Muscular Dystrophy (DMD) is a sex-linked genetic disease. Boys with the disease usually die at a young age, while affected girls usually do not suffer symptoms and may unknowingly carry the disease and pass it to their offspring. It is desirable to have some kind of test to detect whether or not a woman is a carrier of the disease. The following data come from a 1981 study attempting to develop such a test[2].

   In the study, 38 women known to be carriers and 82 women known not to carry the DMD-causing allele were given a blood test; the results are below.

		Test result	
		+	-
Carrier	Yes	20	18
status	No	1	81

   (a) Let π_1 and π_2 denote the probability of a woman testing positive given that she is a carrier and given that she is not a carrier, respectively. Write out an appropriate model, including your choice of priors for π_1 and π_2, and report on the posterior distribution for π_1 and π_2 (you must at least provide measures of central tendency and credible intervals; if you want to do more, by all means, do so).

   (b) The outcome of interest in this study is π_3, the probability that a woman is a carrier, given that she tests positive. This probability depends on π_1 and π_2, of course, but also depends on π_4, the unconditional probability of carrying the DMD-causing allele (this is also known as the prevalence). Suppose we use the following non-informative prior for the prevalence:

$$\text{logit}(\pi_4) \sim \text{N}(0, 10^2).$$

   Report and interpret the posterior for π_3, also known as the "positive predictive value" of the test.

   (c) Now assume that we have access to a study on the prevalence of the DMD-causing allele in the population that suggests the following prior:

$$\text{logit}(\pi_4) \sim \text{N}(-8, 1^2).$$

   Note that this prior indicates a 95% probability that the true fraction of the female population carrying the allele is somewhere between 1 in 400 women and 1 in 21,000 women, with the most likely value around 1 in 3,000 women. Using this prior, report and interpret the posterior of π_3.

   (d) Comparing the models in (a), (b), and (c), does the prior we use for π_4 affect the posterior distribution for π_1 and π_2? Why or why not?

[1] Note that the priors for λ_1 and λ_2 may be implicit (*i.e.*, specified in terms of other parameters that induce a probability distribution on $\boldsymbol{\lambda}$)

[2] Advances in DNA technology have since led to more accurate and sophisticated genetic testing

4. This problem continues with the analysis of the alcohol metabolism data set that we started looking at in class.

   (a) Our in-class analysis did not include any – *i.e.*, the effect of dehydrogenase was assumed to be the same in males as it is in females, and the same in alcoholics as it is in non-alcoholics, and so on. Carry out an exploratory analysis to get a sense of whether this assumption seems to hold. You can look at the data however you like, but I'll suggest using the `lattice` package in `R`, which allows you to do things like:

   ```
   require(lattice)
   xyplot(y~x|a+b)
   ```

   which plots `y` versus `x`, conditioning on `a` and `b`. Provide some sort of exploratory plot that illustrates a possible interaction (or lack thereof) and comment on it.

   (b) Decide on a Bayesian linear regression model for analyzing this data set (again, with `Metabol` as the outcome variable. Describe the likelihood as well as the priors for all parameters. In particular, this model should include interactions, and should place a somewhat skeptical prior on them (*i.e.*, the model should allow for interactions, but be somewhat doubtful about them prior to seeing the data). "How skeptical" is for you to decide, but the prior on any interactions should be more skeptical than the priors on the main effects.

   (c) Decide on several quantities that represent objects of interest in this study. These might be parameters of the model in (b), or they might be functions (differences, sums, ratios, etc.) of those parameters. Select at least 8 such quantities, and provide both a short mathematical description (*e.g.*, $\beta_1 - \beta_2$) and a verbal description (*e.g.*, "Intercept: Males"). You might want to put this information in a table.

   (d) Fit the model in (b) and report the posterior for all quantities in (c).

   (e) Briefly, in words, summarize your main conclusions from analyzing this data set. (These conclusions should follow from the results in (d), of course, but there is no need to repeat your results here; just state your conclusions in words).