University of Edinburgh
School of Mathematics

Bayesian Data Analysis

Assignment 2

Gene expression data analysis 2.

Expression (concentration) of each of 50 genes was observed in the pancreas tissue of two groups of people: in a case group of 12 people who have pancreatic cancer, and in a control group of 10 people who do not have pancreatic cancer (the groups are approximately matched by age and gender). The study was performed in the same hospital.
The question is to determine whether there is a difference between the gene expression in the pancreas tissue between the groups of people with and without the cancer.

Statistical analysis.
Introduce the following random variables. Denote the logarithm of the gene expression of the kth gene for individual i in the case group by X_{ik}, and for the jth individual in the control group by Y_{jk}, $k = 1, 2, \ldots, N = 50$, $i = 1, \ldots, n = 12$, $j = 1, 2, \ldots, m = 10$.
Possible likelihood:

$$X_{ik} \mid \mu_1, \sigma_1 \sim N(\mu_1, \sigma_1^2),\ i = 1, 2, \ldots, n \quad \textit{independently (given } \mu_1, \sigma_1^2),$$
$$Y_{jk} \mid \mu_2, \sigma_2 \sim N(\mu_2, \sigma_2^2),\ j = 1, 2, \ldots, m \quad \textit{independently (given } \mu_2, \sigma_2^2).$$

Also, X_{ik} and Y_{jk} are independent for all i, j, k.
The observed data can be summarised as follows: the sample means are $\bar{x} = 4.03$, $\bar{y} = 2.59$, and the sample variances are $s_X = 0.29$ and $s_Y = 0.11$ where

$$\bar{X} = \frac{1}{nN}\sum_{i=1}^{n}\sum_{k=1}^{N} X_{ik}, \quad S_X^2 = \frac{1}{nN}\sum_{i=1}^{n}\sum_{k=1}^{N}(X_{ik} - \bar{X})^2,$$
$$\bar{Y} = \frac{1}{mN}\sum_{j=1}^{m}\sum_{k=1}^{N} Y_{jk}, \quad S_Y^2 = \frac{1}{mN}\sum_{j=1}^{m}\sum_{k=1}^{N}(Y_{jk} - \bar{Y})^2,$$

and under the likelihood above,

$$\bar{X} \mid \mu_1, \sigma_1 \sim N(\mu_1, \sigma_1^2/(nN)), \quad S_X^2 \mid \sigma_1 \sim \Gamma((nN-1)/2, nN/(2\sigma_1^2)),$$
$$\bar{Y} \mid \mu_2, \sigma_2 \sim N(\mu_2, \sigma_2^2/(mN)), \quad S_Y^2 \mid \sigma_2 \sim \Gamma((mN-1)/2, mN/(2\sigma_2^2)).$$

1. State how you can address the question of interest, whether there is a difference between the gene expression in the pancreas tissue between the groups of people with and without the cancer, in terms of μ_1 and μ_2. **[0.5 mark]**

2. Use proper 'non-informative' independent priors for μ_1 and μ_2, and justify why they can be viewed as 'non-informative'. **[0.5 mark]**

   [*Hint: you can use e.g. $N(0, A^2)$ or $U[-A, A]$ prior with large A*].

3. **Priors for σ_1 and σ_2.**

   (a) Use proper 'non-informative' independent priors for σ_1 and σ_2 (or for their one-to-one transformation); justify why they can be viewed as 'non-informative'. **[0.5 mark]**

   [*Hint: normal distribution in WinBUGS is specified by its mean and precision $\tau = 1/\sigma^2$, so the prior for standard deviation σ is usually stated as a prior for precision τ; e.g. $\Gamma(\epsilon, \epsilon)$ for τ or for σ, with small ϵ.*]

(b) Use a priori information from a previous study that the average precisions are known to be 11.11 for cases and 100 for controls, to specify priors for σ_1 and σ_2 or for their one-to-one transformations. The variability is not known a priori so try two different values of the prior variance (one with a large variance, e.g. 1000, and another with a smaller variance, e.g. 10). **[1 mark]**

[*Hint: typical prior distributions for precision $\tau = \sigma^{-2}$ are Gamma and log Normal distributions. Don't forget to state the expressions for the mean and for the variance of these distributions.*]

4. **Posterior analysis.** For each of the three priors for (σ_1, σ_2) (or, equivalently, for (τ_1, τ_2) where $\tau_r = 1/\sigma_r^2$, $r = 1, 2$), run a WinBUGS model (using at least two chains and running the chains until convergence):

   (a) produce the posterior summaries (mean, median, standard deviation, and two-sided 95% credible interval) for μ_1, μ_2, σ_1, σ_2 and for $\delta = \mu_1 - \mu_2$; **[0.5 mark]**

   (b) produce the density plots of the posterior distribution of $\delta = \mu_1 - \mu_2$, μ_1, μ_2, σ_1, σ_2; **[0.5 mark]**

   (c) use the posterior distribution of δ (or the posterior distributions of μ_1 and μ_2) to address the question of interest (give the conclusion in terms of the original question). **[1 mark]**

5. **Convergence.** For each prior for (σ_1, σ_2),

   (a) Specify the number of chains you used and the initial values for each chain, the number of burnin iterations and the number of iterations used for posterior inference.

   (b) Produce the plot of the Gelman-Rubin statistic for μ_1, μ_2, σ_1, σ_2. What conclusion can you make about convergence of the MCMC? **[0.5 mark]**

   (c) Produce the history plots for each parameter μ_1, μ_2, σ_1, σ_2. Justify the argument that the MCMC converged at the burnin iteration (and hence your use of the number of burnin iterations). **[1 mark]**

   (d) Attach the autocorrelation plots for each parameter μ_1, μ_2, σ_1, σ_2. Comment on the strength of the correlation (and whether it is necessary to thin.)**[0.5 mark]**

   (e) Give the MC error for each parameter $\delta = \mu_1 - \mu_2$, μ_1, μ_2, σ_1, σ_2. Comment how the MC error compares with the standard deviation of the posterior distribution. (You can give the MC error in the posterior summary table in 4(a), and comment on the MC error here.) **[0.5 mark]**

6. **Sensitivity to the prior.**

   (a) Compare the posterior distributions of the parameters for different priors, and comment whether they are affected by the different choices of prior.**[1 mark]**

   (b) Comment whether the main conclusion on the question of interest (4(c)) is affected by the choice of priors. **[1 mark]**

7. Appendix. Attach the code of your WinBUGS models (mark is given for correct models). **[0.5 mark]**

Style of the report (whether it is easy to read): **[0.5 mark]**

Remark. *WinBUGS output can be attached separately. In that case, please label plots and tables and refer to them in the text.*