BST 701: Bayesian Modeling in Biostatistics
Breheny

Assignment 3
Due: Thursday, March 7

1. The data set `donner.txt` contains information on the survival of adult members of the ill-fated Donner Party[1] of pioneers migrating to California in 1846:

   - `Age`
   - `Sex`
   - `Status`: either `Died` or `Survived`

   Fit a the following logistic regression model to this data:

   $$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2\texttt{Age} + \beta_3\texttt{Sex} + \beta_4\texttt{Age}\cdot\texttt{Sex}$$

   For each of the following quantities of interest below, report its posterior:

   - The probability of death for a 20-year old female
   - The probability of death for a 60-year old female
   - The relative risk[2] of death comparing a 40-year old female to a 20-year old female
   - The relative risk of death comparing a 20-year old male to a 20-year old female

2. In this problem, we continue to analyze the Donner Party data with logistic regression, but from a model selection standpoint. As you may have noticed in the previous problem, the posterior intervals are quite wide, suggesting that perhaps we are overfitting the data. For (a)-(e) below, consider the following four models, each with prior $P(M) = 1/4$:

   - Model 1: Intercept only
   - Model 2: `Age`
   - Model 3: `Age` + `Sex`
   - Model 4: `Age` + `Sex` + `Age`·`Sex`

   Before fitting the models, scale `Sex` and `Age` to have mean 0 and variance 1 (for example, using the `scale` function in `R`). For each model, assume a t_3 distribution for each regression coefficient with mean 0 and scaling parameter $\sigma = 2$ (*i.e.*, $\tau = 1/4$).

   (a) For each of the four models, report (i) the mean posterior deviance, (ii) the estimated degrees of freedom (you may use whichever one of p_D, p_V, or p_{opt} you prefer), and (iii) the DIC. Summarize this information in a table.

[1] The tale of the Donner Party is an interesting one; see Wikipedia or various other sources for additional background information

[2] Ratio of probabilities; see the 1-31 notes for a formal definition

(b) Implement a trans-dimensional MCMC model that is capable of jumping between the four models above, and use this approach to calculate the posterior probability of each model.

(c) What is the Bayes factor for comparing Model 4 to Model 3?

(d) Which model is optimal according to DIC? Which model has the highest posterior probability?

(e) Provide a model-averaged posterior for the probability of death for a 20-year old female. How does it compare with the posterior in problem 1?

(f) In a famous paper, Gideon Schwarz derived an approximation to the Bayes factor known as the BIC, or *Bayesian Information Criterion.* In particular, he showed that

$$P(M_j|\mathbf{y}) \approx \frac{\exp(-0.5\mathrm{BIC}_j)}{\sum_k \exp(-0.5\mathrm{BIC}_k)},$$

where the sum is over the models under consideration. Fit models 1-4 using ordinary least squares and use BIC to approximate the probabilities in (b)[3]. How do they compare with the probabilities in (b)?

[3] `BIC(fit)`, if using `R`