BIOS 902: Bayesian Statistics
Byron J. Gajewski
KU Department of Biostatistics
Due 11/6/18

Homework #4

1. Exercise 4.1, p. 217, Carlin & Louis (data set is on the shared drive).

1. Steensma et al. (2005) presented the data in Table 4.4, from a randomized controlled trial comparing two dosing schedules for the drug erythropoiten. Serum hemoglobin (HGB, in g/dL) was recorded for $N = 365$ cancer patients with anemia over $T = 22$ weeks. The full data file is available at `www.biostat.umn.edu/~brad/data/HGB_data.txt`. We wish to fit a hierarchical simple linear regression model of the form

$$Y_{ij} = \beta_{0i} + \beta_{1i}(X_j - \mu_X) + \epsilon_{ij},\ i = 1, \ldots, 365,\ j = 1, \ldots, 22, \quad (4.43)$$

where $X_j = j$, the week index, $\epsilon_{ij} \overset{iid}{\sim} N(0, \tau)$, $\beta_{0i} \overset{iid}{\sim} N(\mu_0, \tau_0)$, and $\beta_{1i} \overset{iid}{\sim} N(\mu_1, \tau_1)$. That is, as in Example 2.13 (and Example 7.2 in Chapter 7), we allow each subject's HGB trajectory to have its own slope and intercept, but borrow strength from the ensemble by treating these as normal random effects. Note that we also center the week index around its own mean, $\mu_X = 11.5$.

   (a) Use `WinBUGS` to fit the above model, assuming vague priors for the hyperparameters τ, μ_0, τ_0, μ_1, and τ_1. Run multiple chains to assess convergence, and interpret the resulting posterior distributions for the grand slope μ_1 and the individual-specific slopes β_{1i}. How do the interpretations of these parameters differ? Use the `compare` function in `WinBUGS` to examine these for all participants. Are all participants' HGB measurements improving over time?

   (b) Note that many of the Y_{ij} are missing; under the assumption that they are missing at random, `WinBUGS` can impute them according to the fitted model. Monitor the hemoglobin values estimated for participant

Patient	HGB Measurements by Week j, $j = 1, \ldots, 22$					
i	$Y_{i,1}$	$Y_{i,2}$	$Y_{i,3}$	$\cdots$	$Y_{i,21}$	$Y_{i,22}$
1	10.2	9.8	10.3	$\cdots$	NA	11.4
2	10.1	9.4	9.1	$\cdots$	NA	11.6
3	9.9	10	10.2	$\cdots$	NA	12.4
4	10.7	9.7	11.4	$\cdots$	NA	NA
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
365	10.7	10.5	10.4	$\cdots$	14.2	NA

Table 4.4 $T = 22$ *weekly hemoglobin (HGB) measurements.*

10, who is missing data from weeks 5 and 16 through 22. Compare the standard deviation of the estimate at week 5 to those at the end of the study, and explain any differences.

(c) As you've already noticed, the `WinBUGS` data file actually contains information on one more variable, `newarm`, a binary variable indicating which dosing schedule (or *treatment arm*, 1 or 2) was used for each patient. Modify model (4.43) above to allow different grand intercepts and/or slopes for the two treatment arms. Discuss any resulting problems with MCMC convergence, how they might arise, and how they can be handled. Are the two groups significantly different in any respect? Draw a quick plot in `R` comparing the fitted grand means in the two groups. Would a DIC comparison of this "full" model and the "reduced" model in part (a) be sensible here?

2. **Parts "a" and "c" of Exercise 4.5 p. 220, Carlin & Louis (data set is on the shared drive).**

5. Consider `www.biostat.umn.edu/~brad/data/copresence_data.txt`, a data set for which a few records are shown in Table 4.5. Here, Y is a binary variable indicating co-presence of two species in a particular forest at $n = 603$ sampled locations. The lone predictor variable, X, is the log of the distance of each location to the forest edge. Suppose we use a logistic model for p, the probability of co-presence, namely

$$logit(p_i) = \beta_0 + \beta_1 X_i,\ i = 1, \ldots, n.$$

(a) Again following the model of Example 4.4, fit this model in `WinBUGS`, using vague priors. Is proximity to the forest edge a significant predictor of species co-presence?

(c) Replace the logit link above with the complementary log-log link, $\log[-\log(1 - p_i)]$, and compare the two posteriors for β_1. Also plot the two fitted curves as in Figure 4.6, and compare the models more formally using DIC or some other Bayesian model choice statistic. Does the choice of link function matter much for these data?