Final Project
From April 20th 9 am to April 23rd 11:59 pm

1 **For each hour of delay in project report submission, a penalty of 10 points will be applied.**

2 You must include your R codes in your project reports.

3 Make sure your answers to the questions are neat and easy to understand. Scribbled handwriting is not acceptable!

4 The answer to each question must be followed with the R codes for addressing the question.

5 Your answers to questions have to be well organized.

6 You have to show every step of your derivations clearly, including the prior, the joint posterior distribution, conditional posterior distributions and their distribution parameters, simulation steps, the distribution simulated in the MCMC sampling algorithm, posterior inferences, results, and interpretations.

7 If you decide to handwrite your mathematical derivations, be sure to provide neat and well-organized answers. That is, the mathematical derivations should be next to your answers.

8 The answer pages you upload to the course website have to be in the correct order.

The following problem is motivated from the paper by Brown (2008).

"Batting average is one of the principle performance measures for an individual baseball player. It is the percentage of successful attempts, Hits, as a proportion of the total number of qualifying attempts, At- Bats. In symbols, the batting average of the ith player may be written as $BA_i = H_i/AB_i$. This situation, with Hits as a number of successes within a qualifying number of attempts, makes it natural to statistically model each player's batting average as a binomial variable outcome, with a given value of AB_i and a true (but unknown) value of p_i that represents the player's latent ability."

"We will look at batting records for each Major League player over the course of a single season (around 6 months). We use the batting records from the first half of the season (e.g., the first 3 months) in order to estimate the batter's latent ability, p_i, and consequently, to predict their BA performance for the remainder of the season. Since we are using a season that has already concluded, we can then validate the performance of our estimator by comparing the predicted values to the actual values for the remainder of the season."

Bat.dat file contains data of AB and H for 929 players in the two periods of the season (the first 3 months and the latter 3 months). The column "AB.1" contains the total numbers of qualifying attempts, At-Bats, for the players in the first period of the season; The column "H.1" contains the total numbers of hits for the players in the first period of the season. Similarly, The column "AB.2" contains the total numbers of qualifying attempts, At-Bats, for the players in the second period of season; The column "H.2" contains the total numbers of hits for the players in the second period of season. The column "Pitcher" indicates whether the player is a "Pitcher" (corresponding to value 1).

Let H_{ij} and N_{ij} denote the observed number of hits and at-bats, respectively, for player i within period j, $j = 1, 2$. We assume that each H_{ij} is a binomial random variable with an unobserved parameter p_i corresponding to the player i's hitting ability. Thus, for data involving P players over two halves of the season, we write

$$H_{ij} \overset{indep}{\sim} \text{Binomial}(N_{ij}, p_i), j = 1, 2,\ i = 1, \ldots, P.$$

1. To provide reliable prediction results, **in the following analysis**, exclude (from the file **Bat.dat**) the players whose at bats, AB, either in the first or second period, are no larger than 10 ($N_{i1} \leq 10$ or $N_{i2} \leq 10$). Calculate how many players n are left in the analysis.

2. For the players left in the analysis from 1, develop a **Bayesian method** to estimate these players' latent ability, p_i, **based on their data in the first period** (N_{i1}, H_{i1}). Explain your Bayesian model and how you obtain your estimates.

3. Following Brown (2008), we can transform H_{ij}s to normally distributed random variables X_{ij}s:

$$X_{ij} = \arcsin \sqrt{\frac{H_{ij}+1/4}{N_{ij}+1/2}}. \tag{0.1}$$

$$X_{ij} \sim \text{Normal}(\theta_i, \sigma_{ij}^2),\ \text{where } \theta_i = \arcsin \sqrt{p_i} \text{ and } \sigma_{ij}^2 = \frac{1}{4N_{ij}} \tag{0.2}$$

Construct a Bayesian **hierarchical** model for X_{i1}, $i = 1, \ldots, n$, and use the model to estimate p_i (you need to transform estimated θ_i back to obtain estimated p_i.)

4. Treating H_{i2}/N_{i2} for $i = 1, \ldots, n$ in the second period as target values to be predicted. What's your estimates of H_{i2}/N_{i2} from 2 and 3?

5. Denote the estimate (based on the data in the first period) of H_{i2}/N_{i2} by $\tilde{p}_i$, $i = 1, \ldots, n$. Compare the mean squared error (MSE) of estimated $\tilde{p}_i$s

$$\text{MSE} = \sum_{i=1}^{n} (H_{i2}/N_{i2} - \tilde{p}_i)^2/n$$

using the maximum likelihood estimates and the two estimates you developed in 2 and 3. Explain the reasons for the differences in the MSEs of different estimators.

6. The players fall into two subgroups: nonpitchers and pitchers. Then we can estimate H_{i2}/N_{i2}s for two subgroups separately. Perform procedures 2-5 for the data of nonpitchers and pitchers separately. Compare the MSEs of the two groups with those in 5. Discuss the results and explain the reasons for the differences.

Brown, L. D. (2008). In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies. The *Annals of Applied Statistics*, 113-152.