PROBLEM SET # 8

EC/ACM/CS 112: Bayesian Statistics

Submission instructions	>> Create a pdf of the R Notebook with your solutions (details below) >> Submit in Canvas
Additional files included in the problem set package	++ solutions template (.Rmd file) ++ dataset *BattingAverage.csv*

QUESTION 1. ESTIMATING A SIMPLE MULTI-LEVEL HIERARCHICAL MODEL OF BATTING AVERAGES

LEARNING GOALS

- Deepen your understanding of hierarchical models by estimating a multilevel hierarchical model of batting averages.
- Practice using posterior predictive checks to identify limitations of statistical models and directions for model improvement

NOTE

- This problem is based on an example in Kruschke, *Doing Bayesian Data Analysis*, 2015. You may <u>not</u> consult the book, or related on-line materials.

DATASET

- The dataset for this problem, *BattingAverages.csv*, is included with the package for this problem set.
- The dataset contains information on 948 players for the 2012 season of Major League Baseball.
- Each row contains data for a different player.
- Columns include the following variables:
  ++ Player = player name
  ++ PriPos = player's position out of 9 categories (e.g., pitcher, catcher, etc.)

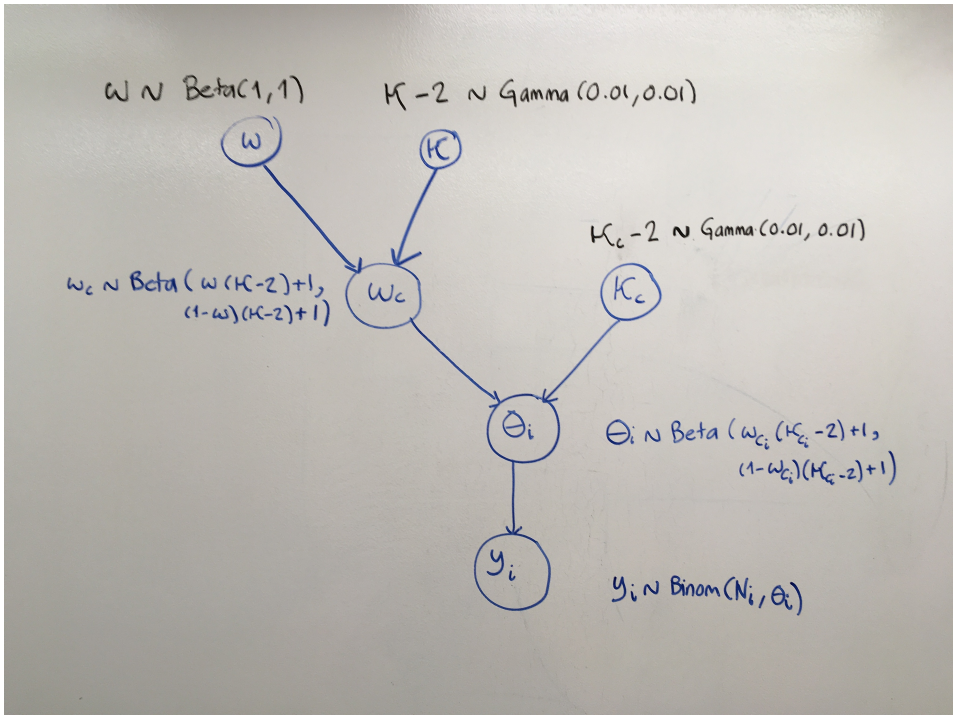
++ AtBats = number of times the player was at bat
++ Hits = number of hits

- Note that the dataset excludes all players with zero at-bats, as well as other variables that you can ignore.

STEPS

Step 1. Use Stan to fit the following model, where:

- The hyperpriors are denoted in black
- The subscript c_i denotes the player i's field position (i.e., *c = pitcher, catcher, …*).
- y_i and N_i denote, respectively, the number of hits and the number of times at bat for player *i*.
- ω_c and κ_c denote, respectively, the mode and concentration of the distribution of hit rates for group c.



Note that the model has 968 unknown parameters!

Print the Stan fits for both hyperparameters, for the field position mode and concentrations, and for the theta of the 1st, 500th, and 948th player in the dataset. You should show the 5%, 50% and 95% of the fits for each parameter.

TIPS:

- Note that the priors for κ and κ_c are defined over $\kappa - 2$ and $\kappa_c - 2$. You can deal with this by defining a hyperparameter **kappaMinusTwo_0** and a vector of parameters **kappaMinusTwo** for the different positions and using them in defining the Stan model. If desired, you can then obtain samples for κ and the κ_cs by a subtracting 2 from the samples for the transformed parameters.
- You want to use at least 4 chains with at least 20,000 iterations each and the default 50% warm-up.
- Your code will run faster if you vectorize your Stan script (see the Lambert chapter included in the previous problem set).

Step 2. Use convergence diagnostics to determine if your Markov chain is sampling properly, and adjust your code and sampler above if required.

With so many parameters, it is difficult to visually inspect all of the converge statistics and diagnostic plots, so proceed as follows:

- Extract the Rhat statistic for the 968 parameters and plot them in a histogram to make sure that there are no convergence issues with any of them.
- Extract the ESS statistic for the 968 parameters and plot them in a histogram to make sure that you have at least 1000 effective samples for each parameter.
- Identify the 10 parameters with the worse Rhat statistic and the 10 parameters with the worse ESS statistic.
- Inspect the trace plots for both hyperparameters, for the field position mode and concentrations, and for the theta of the 1st, 500th, and 948th player in the dataset.

TIPS:

- Useful info on how to access contents of stanfit objects: https://cran.r-project.org/web/packages/rstan/vignettes/stanfit-objects.html

Step 3. Show a histogram of the posterior difference in $\omega_{catcher} - \omega_{pitcher}$, compute the 95% highest density interval (HDI) for this statistic, and compute the posterior probability that $\omega_{catcher} > \omega_{pitcher}$.

Step 4. Show a histogram of the posterior difference in $\omega_{catcher} - \omega_{1st-base}$, compute the 95% HDI for this statistic, and compute the probability that $\omega_{catcher} > \omega_{1st-base}$.

Step 5. Using a single plot, compare the posterior of ω_c for each of the nine positions, as well as a histogram of the posterior for ω. Why do you think the posterior for ω is so much wider than the posteriors for the ω_c?

TIP:

- You can add density lines to a histogram type plot using the command **lines(density(varName), …).**

Step 6. Perform a posterior predictive check of the model using the following steps, separately for each of the nine positions:

- Sample hit probabilities θ_i 1000 new players from the posteriors for the category
- Compare the distribution of observed hit rations (i.e., y_i/N_i) with the distribution of predicted hit probabilities, as given by your samples.
- To make the plot easy to read, plot the data as a standard histogram, and then add a line to the plot describing the density of the samples.

TIP:

- You can generate a 3x3 multipanel plot using the command **par(mfrow=c(3,3))**.

Step 7. What does the posterior predictive check in Step 5 suggest about the model exchangeability assumptions in the model?

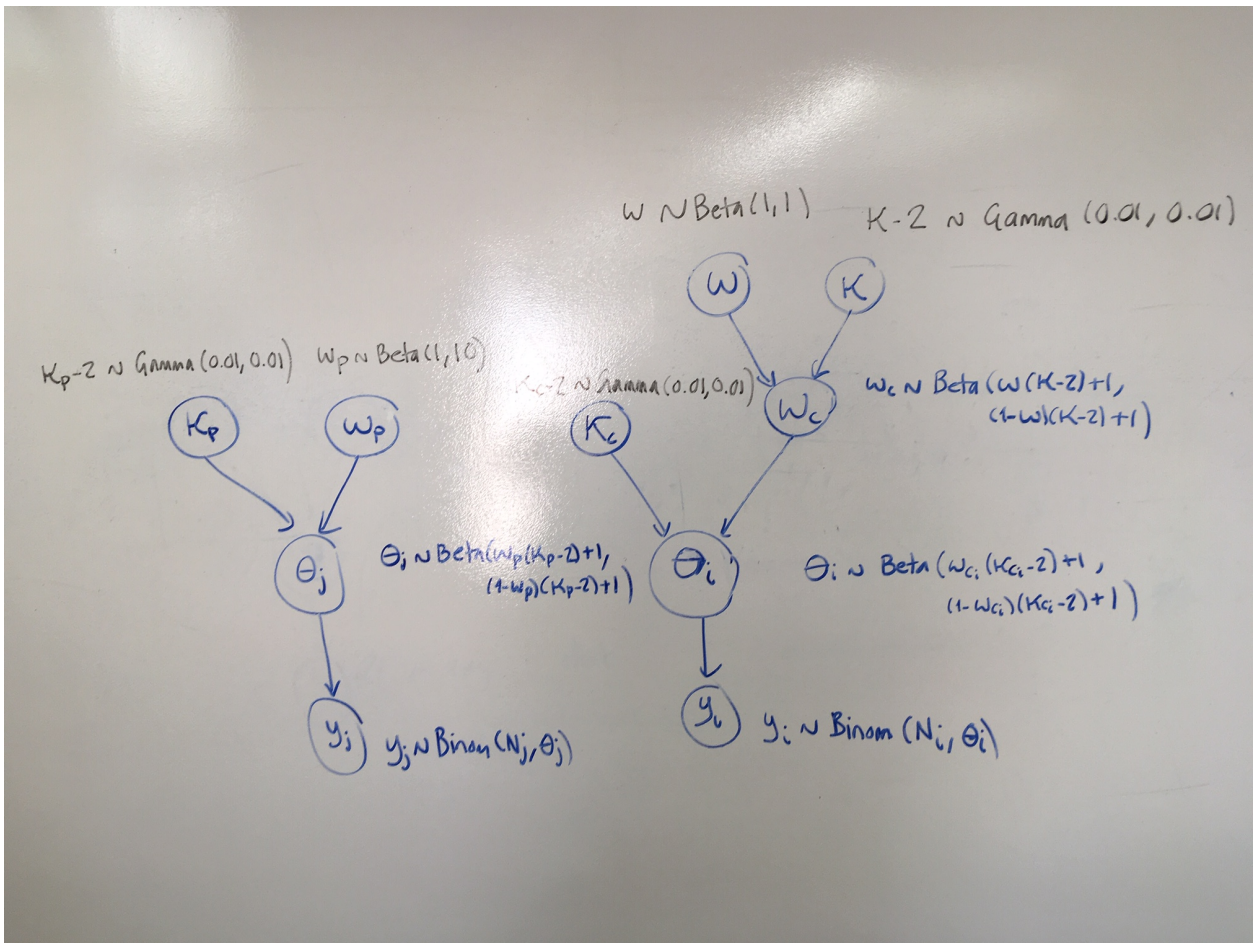
QUESTION 2. ESTIMATING AN ALTERNATIVE SIMPLE MULTI-LEVEL HIERARCHICAL MODEL OF BATTING AVERAGES

LEARNING GOALS

- Deepen your understanding of what shapes the posteriors in multi-level hierarchical models.
- Practice the process of iterative model fitting and improvement.

STEPS

Step 1. Use Stan to fit the following modified version of the model, which is very similar to the previous one, except that pitchers are now treated as a non-exchangeable group with the other positions.



Step 2. Repeat step 5 above, but using the fits from the new model.

Step 3. Repeat the posterior prediction exercise in step 6 above, but using the new model. Plot the predictions of both the old and the new models.

Step 4. Why does the plot is Step 2 changes drastically between the two models, but the posterior predictive checks in Step 3 does not?