Homework 3:
Sampling Methods

Instructions:

1. Questions 2-5 in this assignment require coding in python. Submit a jupyter notebook with both your code and your answers to the questions.
2. For question 1, you may either include that in your jupyter notebook or in a separate file.
3. You may discuss this assignment with other students in the class, but you must submit your own answers to the questions below.
4. Include an honor pledge with your submission.
5. Submit on-line.
6. This homework is worth 100 points and the point totals for each question are shown in parentheses with the possibility for 5 extra credit points.

Assignment:

1. (10)
   (a) Suppose whether it is sunny or not in Charlottesville depends on the weather of the last three days. Show how this can be modeled as a Markov chain. How many states are needed?
   (b) Explain why a Markov chain with the transition matrix shown below does not have an equilibrium distribution.

$$\text{Transition Matrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

2. (20) Let $X \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\mu} = (1, 1)$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$

   Use numpy and scipy to implement a Gibbs sampling algorithm for this 2D Gaussian. Show three runs of you algorithm with 100 samples past burn-in as 1D scatter plots in each of the two dimensions overlaid on the plots of the actual marginal distributions.

3. (25) With the CHD data set (CHDdata.csv) from Homework 2 develop a sampling based estimate for the posterior distributions of the parameters in a main effects logistic regression model. Use all predictor variables in the data set and standardize all of the numeric, continuous predictors using the mean and standard deviation.

- Use Gaussian or Cauchy priors with hyperparameters appropriate for uninformed priors and show the graphical representation of your model (you should use graphviz for this).
- Show the summary table and trace plots from the sampling and briefly say what they mean for your results.
- Provide forest plots of the parameters and discuss what these results imply for the relevant predictor variables and the overall model.

4. (20) With the data set bangladesh.csv develop a pooled, a no-pooled, and a hierarchical model for all districts to predict contraception usage. Use only district and age.centered as predictor variables. Plot each of these predictions with age.centered on the x-axis and the expected proportion of women using contraception on the y-Axis with overlaid plots for the districts. Briefly explain these results.

5. (25) For the CHD data in problem 3, develop two other models with fewer predictor variables and/or with a nonlinear transformation of one of the the predictor variables.

   - Use WAIC to recommend which of these models to use.
   - Use WAIC to produce a Bayesian model average result. Show the kernel density plots for the sample posterior predictions of each of the models and the Bayesian model average.
   - (5 points extra credit) Show the overlaid plots of the predictions for the data points in the CHDdata.csv file with age on the x-axis and the predicted probability on the y-axis. Overlay these results for each model and for the Bayesian model average of the models.