PROBLEM SET # 7

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) ++ pdf of lambert's book ch16 ++ pdf of the rstan package documentation

PROBLEM SET LEARNING GOAL

- Learn how to use Stan to analyze basic models

QUESTION 1. PREPARATORY WORK

Step 1. Start by reading carefully chapter 16 of Lambert's book, which is included with the problem set package. This will provide you with critical basic knowledge about how to fit and analyze models in Stan, and will save you time later on in this and later problem sets.

Step 2. Install the package "rstan" and all of its dependencies. See section 16.4 of the chapter and https://mc-stan.org/users/interfaces/rstan.

You don't need to submit anything related to this question, but you are required to complete these two steps before you proceed.

QUESTION 2. USING STAN TO ANALYZE A FAMILIAR LINEAR REGRESSION MODEL

BACKGROUND

- In problem set 4 we used the grid method to fit a linear regression model of the metabolic rate of tobacco hornworms on body size and caterpillar's life stage.
- Recall that this model was computationally taxing to fit and approached the limits of what can be done using the grid method.
- In this problem set we will carry out the same exercise using Stan. The comparison in performance will be stark.

DATASET. The dataset "MetabolicRate.csv" included with the problem set package was originally compiled by Prof. Itagaki and his students http://www.kenyon.edu/directories/campus-directory/biography/harry-itagaki/).

It contains measures on the following three variables 305 caterpillars:

- BodySize = size of the caterpillar (in grams)
- Instar = number from 1 (smallest) to 5 (largest) indicating the caterpillar's life stage
- mRate = a measurement of the caterpillar's metabolic rate

The dataset has no missing observations.

STEPS

Step 1. Write a Stan script to fit a Gaussian linear regression of log(mRate) on log(BodySize) and Instar.

The model should have the following independent and weakly informative priors:

- The constant and the two slopes have priors $N(0, \sigma^2 = 400)$
- The standard deviation σ has a prior Gamma(1,1)

You don't need to submit this file, since the rest of the steps will not work properly if there is a mistake in the model specification.

Step 2. Use rstan to fit the model and show the summary statistics of the fit using the **print()** command.

You should use at least 4 chains and generate enough samples to get at least 10,000 effective samples for each parameter.

TIP:

- Remember that the .stan file needs to be in the same working directory as your code, or make sure to specify the path to the .stan file.
- You might find the rstan documentation (included as pdf with the problem set package) useful.

Step 3. Do a traceplot of the last 2,000 samples for all parameters and chains, and verify that they are consistent with sampling from the stationary distribution. If this is not the case, there is something wrong with your previous steps that you need to fix.

Step 4. Do a matrix scatterplot showing the posterior samples for the four parameters.

Step 5. Go back to Step 1 and add a generated quantities segment to your Stan model that we will use for posterior predictive checks.

In this segment you should generate a prediction on mRate, call it yPred, for every observation in the dataset.

Step 6. Posterior predictive check 1. Plot the distribution of observed metabolic rates versus the distribution of predicted rates in 100 random iterations.

Note that the generated quantities in Step 5 give you a value of yPred for every observation in every iteration, and that every iteration gives you a predicted distribution of observed mRates.

This step is asking you to plot the predicted distribution of yPred, over all 305 observations, in 100 random iterations of the sampler.

TIPS:

- You can use the command **extract**(fit)[["varName"]], where fit is the name of the rstan object with the results of your sampler, to convert the yPreds into a matrix.
- You might want to use the function **ppc_dens_overlay** from the bayesplot package.

Based on the results of this posterior predictive check, does this simple linear regression model provides a satisfactory account of the data?

Step 7. Posterior predictive check 2. Repeat step 6, separately for observations in each level of the Instar predictor.

Are there differences on the ability of the model to predict the data at different stages of the worms' life cycle?

TIPS:

- You can use the **grid.arrange** function, from the gridExtra package, to easily generate multi-panel plots.