STAT 237: HW4

DUE ELECTRONICALLY VIA THE RSTUDIO SERVER FRIDAY 05/06/22 BY 11:59PM

1. In the 1980s there was a famous study done of the "hot hand" in basketball and people have been studying this ever since.

   In the 2009-2010 season, Kevin Durant made back-to-back free throw attempts 332 times. In 290 instances he made the first shot. In 42 instances he missed the first shot.

   Let Y_{n1} represent the outcome of the nth "first shot" attempt (0 being a miss and 1 a success), and Y_{n2} represent the outcome of the nth "second shot" attempt, for $n = 1, \ldots, N$ (here, $N = 332$). Define

   $$\theta_0 := P(Y_{n2} = 1 \mid Y_{n1} = 0)$$
   $$\theta_1 := P(Y_{n2} = 1 \mid Y_{n1} = 1),$$

   that is, the probability that Durant makes the second shot given that he missed the first, and the probability that Durant makes the second shot given that he made the first, respectively.

   Consider a hierarchical model in which θ_0 and θ_1 share a prior, and where the parameters of that prior themselves have priors.

   (i) Sketch a "circle and arrow" diagram to represent the dependencies among the Ys, the θs, and any additional parameters needed to specify priors.

   (ii) Write out (conditional) probability distributions for each parameter and variable in the model.

   (iii) Write a Stan program called `hot_hand.stan` (Hot Hand Stan would be a great mobster nickname, imo) to encode this model. Feel free to start with one of the example models from class and modify it.

   (iv) Using a prior that reflects an expectated success rate of 80%, perform a prior predictive check to tune the other hyperparameters so that most of the θs sampled from the prior are above 0.60, and where $-0.20 \leq \theta_1 - \theta_0 \leq 0.20$.

Date: Last Revised April 26, 2022.

This will involve some iterative trial and error. Describe the process by which you adjusted your hyperparameters after seeing some simulated success rates.

(v) Of the 290 instances when he made the first shot, he also made the second shot 266 times (and missed the second shot 24 times). Of the 42 instances when he missed the first shot, he made the second shot 36 times (and missed the second shot 6 times). Create a data file called `durant_free_throws.csv` to represent this data. You may want to use the "count" format that the baseball hitting data was in. If you do, make sure you distinguish between the individual throw variables and the corresponding count variables, and make sure your Stan code matches the data format.

(vi) Run the Stan sampler to obtain posterior samples for the parameters of this model given the data. Find the prior and posterior probabilities that $\theta_1 > \theta_0$.

(vii) Repeat the prior predictive check to find a weaker prior for which (θ_1, θ_0) is fairly uniform in density over the $[0, 1]\times[0, 1]$ square. Re-run the sampler with the new hyperparameters. What are the prior and posterior probabilities that $\theta_1 - \theta_0 > 0$ now?

(viii) If Durant making the first shot is predictive of him making the second, one explanation could be that he is more confident after experiencing a success, and this confidence puts him more "in the zone". Another is that he has "on" and "off" days, and making the first shot is an indicator that that day is more of an "on" day, which in turn makes it more likely that he will make the second shot. How could we modify our model to help us distinguish these possibilities? Would we need additional data variables?

2. A survey of residents of various U.S. states asked whether each person had used cocaine in the past year. The aggregated results for seven states are shown below.

State	Respondents	NumUsedCocaine
California	1151	67
Oregon	296	10
Washington	280	14
Alabama	286	14
Florida	1212	64
Georgia	301	10
South Carolina	325	12

Consider a model in which the probability that a given person has used cocaine in the past year varies by state, where states in the same region (West vs South, where the first three states are in the West and the other four are in the South) are expected to be more similar to each other than states in different regions.

(i) Define variables and parameters to describe the data, the state-specific probabilities that a person has used cocaine in the last year, and any other parameters needed for priors on these probabilities. Sketch a "circle and arrow" diagram for this model.

(ii) Write out (conditional) probability distributions for each parameter and variable in this model.

(iii) Select hyperparameters to reflect the prior expectation that about 10% of individuals across states will have used cocaine in the past year, that most states will have use rates between 1% and 20%, and where the variation in use rates across regions is similar to the variation in use rates across states within a region. This will involve some iterative trial and error, using prior predictive checks to examine the expectations that a given set of hyperparameters encodes. Describe the process by which you adjusted your hyperparameters after seeing some simulated use rates.

(iv) Once you settle on a set of hyperparameters, create a model file called `cocaine.stan` (Cocaine Stan being Hot Hand Stan's less well adjusted counterpart), and a data file called `cocaine.csv`, and run the Stan sampler to obtain posterior samples for the parameters.

(v) Graph and report a 95% credible interval for the posterior expected value of the difference between the mean use rate in the West vs the South

(vi) Which Western state has the highest posterior mean use rate? Which one has the lowest? Find the probability that the latter has a higher use rate than the former.