Homework 4

PSTAT 115, Winter 2024

Due on March 11, 2024 at 11:59 pm

Note: If you are working with a partner, please submit only one homework per group with both names and whether you are taking the course for graduate credit or not. Submit your Rmarkdown (.Rmd) and the compiled pdf on Gauchospace.

Problem 1. Stan Warmup: goal Scoring in the Women's World Cup

Let's take another look at scoring in soccer. The Chinese Women's soccer team recently won the AFC Women's Asian Cup. Suppose you are interested in studying the World Cup performance of this soccer team. Let λ be the be the average number of goals scored by the team. We will analyze λ using the Poisson model where data Y_i is the observed number of goals scored in the ith World Cup game, ie. we have $Y_i|\lambda \sim Pois(\lambda)$.

One nice thing about MCMC is we're no longer constrained to work with conjugate priors. In this case, *a priori*, we expect the rate of goal scoring to be $\lambda \sim Lognormal(\mu, \sigma)$ (not Gamma!) where μ and σ are the mean and standard deviation of $\log(\lambda)$. According to a sports analyst, they believe that $\log(\lambda)$ has mean $\mu = 1$ and standard deviation 0.25.

1a. Create a new Stan file by selecting "Stan file" and name it `women_cup.stan`. Encode the Poisson model with the Lognormal prior for λ in Stan. (Here)[https://mc-stan.org/docs/2_21/functions-reference/lognormal.html] is some useful information on the lognormal in stan. Use `cmdstanr` to generate Monte Carlo samples of λ. Create a histogram of the posterior distribution of λ and add a vertical line at the posterior mean.

```r
y <- c(4, 7, 3, 2, 3) # Number of goals in each game

## Create "women_cup.stan" yourself and fill in the model
soccer_model <- cmdstan_model("women_cup.stan")

## This fits the model to data y
## All parameter samples are stored in a data frame called "samples"
stan_fit <- soccer_model$sample(data=list(Y = y), refresh=0, show_messages = FALSE)
samples <- stan_fit$draws(format="df")

## Compute the posterior mean of the lambda samples
post_mean <- NULL # YOUR CODE HERE
```

```r
. = ottr::check("tests/q1a.R")
```

```r
# YOUR CODE HERE
```

1b. Use the Monte Carlo samples from Stan to compute the probability that the Chinese women's soccer team scores more than 4 goals in the next game, $P(\tilde{Y} > 4|y_1, ...y_n)$.

```r
prob_more_than_four <- mean(rpois(length(samples$lambda), samples$lambda) > 4)
print(prob_more_than_four)
```

```
. = ottr::check("tests/q1b.R")
```

Problem 2. Bayesian inference for the normal distribution in Stan.

Consider the IQ example used from class. Scoring on IQ tests is designed to yield a N(100, 15) distribution for the general population. We observe IQ scores for a sample of n individuals from a particular town, $y_1, \ldots y_n \sim N(\mu, \sigma^2)$. Our goal is to estimate the population mean in the town. Assume the $p(\mu, \sigma) = p(\mu \mid \sigma)p(\sigma)$, where $p(\mu \mid \sigma)$ is $N(\mu_0, \sigma/\sqrt{\kappa_0})$. Before you administer the IQ test you believe the town is no different than the rest of the population, so you assume a prior mean for μ of $\mu_0 = 100$, but you aren't to sure about this a priori and so you set $\kappa_0 = 1$ (the effective number of pseudo-observations).

```
y <- c(70, 85, 111, 111, 115, 120, 123)
n <- length(y)
```

2a. Assume $\sigma = 6$. Write out the theoretical posterior mean of $E[\mu | y_1, ...y_n]$ and theoretical upper and lower bounds for the 95% quantile based credible intervals for μ (hint: use the fact that for a normal distribution 95% of the posterior mass is within ±1.96 standard deviations of the mean). Plug in the observed data and report the posterior mean estimate and values for the bounds on the credible interval.

Type your answer here, replacing this text.

2b Now we'll relax the assumption that σ is known. Let $p(\sigma)$ be Gamma(a, b) and assume that σ has a prior mean of 15 (to match the intended standard deviation of the IQ test). We'll use $a = 15$ and $b = 1$ (remember, the mean of a Gamma is a/b) as our prior parameters.

Create a new Stan file and name it `IQ_model.stan`. Make the necessary modifications to the template example in the default Stan file for this problem based on the specified priors. Use the MCMC samples to make a scatter plot of the posterior distribution of the mean, μ, and the precision, $1/\sigma^2$. Put μ on the x-axis and $1/\sigma^2$ on the y-axis. What is the posterior relationshipμ and $1/\sigma^2$? Why does this make sense? *Hint:* review the lecture notes.

```
normal_stan_model <- cmdstan_model("IQ_model.stan")

# Run rstan and extract the samples
# YOUR CODE HERE

mu_samples <- NULL # YOUR CODE HERE
sigma_samples <- NULL # YOUR CODE HERE
precision_samples <- NULL # YOUR CODE HERE

## Make the plot
# YOUR CODE HERE
```

Type your answer here, replacing this text.

```
. = ottr::check("tests/q2b.R")
```

2c. You are interested in whether the mean IQ in the town is greater than the mean IQ in the overall population. Use Stan to find the posterior probability that μ is greater than 100.

```
# YOUR CODE HERE
```

Type your answer here, replacing this text.

```
. = ottr::check("tests/q2c.R")
```

2d. The coefficient of variation, $c_v = \sigma/\mu$ is defined as the standard deviation over the mean. Make a histogram of $p(c_v \mid y)$ from Monte Carlo samples and report the posterior mean and the lower and upper endpoints of the 95% quantile based interval.

```
# YOUR CODE HERE
```

Problem 3. Logistic regression for toxicity data

Logistic regression for pesticide toxicity data.

A environmental agency is testing the effects of a pesticide that can cause acute poisoning in bees, the world's most important pollinator of food crops. The environmental agency collects data on exposure to different levels of the pestidicide in parts per million (ppm). The agency also identifies collapsed beehives, which they expect could be due to acute pesticide poisoning. In the data they collect, each observation is pair (x_i, y_i), where x_i represents the dosage of the pollutant and y_i represents whether or not the hive survived. Take $y_i = 1$ means that the beehive has collapsed from poisoning and $y_i = 0$ means the beehive survived. The agency collects data at several different sites, each of which was exposed to a different dosages. The resulting data can be seen below:

```
x <- c(1.06, 1.41, 1.85, 1.5, 0.46, 1.21, 1.25, 1.09,
       1.76, 1.75, 1.47, 1.03, 1.1, 1.41, 1.83, 1.17,
       1.5, 1.64, 1.34, 1.31)

y <- c(0, 1, 1, 1, 0, 1, 1, 1, 1, 1,
       1, 0, 0, 1, 1, 0, 0, 1, 1, 0)
```

Assume that beehive collapse, y_i, given pollutant exposure level x_i, is $Y_i \sim \text{Bernoulli}(\theta(x_i))$, where $\theta(x_i)$ is the probability of death given dosage x_i. We will assume that $\text{logit}(\theta_i(x_i)) = \alpha + \beta x_i$ where $\text{logit}(\theta)$ is defined as $\log(\theta/(1-\theta))$. This model is known as *logistic regression* and is one of the most common methods for modeling probabilities of binary events.

3a. Solve for $\theta_i(x_i)$ as a function of α and β by inverting the logit function. If you haven't seen logistic regression before (it is covered in more detail in PSTAT 127 and PSTAT131), it is essentially a generalization of linear regression for binary outcomes. The inverse-logit function maps the linear part, $\alpha + \beta x_i$, which can be any real-valued number into the interval [0, 1] (since we are modeling probabilities of binary outcome, we need the mean outcome to be confined to this range).

Type your answer here, replacing this text.

3b The dose at which there is a 50% chance of beehvive collapse, $\theta(x_i) = 0.5$, is known as LD50 ("lethal dose 50%"), and is often of interest in toxicology studies. Solve for LD50 as a function of α and β.

Type your answer here, replacing this text.

3c Implement the logistic regression model in stan by reproducing the stan model described here: https://mc-stan.org/docs/2_18/stan-users-guide/logistic-probit-regression-section.html. In this model, we assume the improper prior $p(\alpha, \beta) \propto \text{const}$. Run the stan model on the beehive data to get Monte Carlo samples. Compute Monte Carlo samples of the LD50 by applying the function derived in the previous part to your α and β samples. Report and estimate of the posterior mean of the LD50 by computing the sample average of all Monte Carlo samples of LD50. Make a trace plot of the first 500 iterations of LD50 from your markov chain. What was the effective sample size from the first 500 iterations (you can use the `effectiveSize` function)?

```
# YOUR CODE HERE

logistic_fit <- NULL # YOUR CODE HERE
logistic_samples <- NULL # YOUR CODE HERE

alpha_samples <- logistic_samples$alpha
```

```
beta_samples <- logistic_samples$beta

ld50 <- NULL # YOUR CODE HERE
print(ld50)
```

Fill in the compute curve function, which computes the probability of hive collapse for each value of x in `xgrid`. Then run the code below to make a plot showing both 50% and 95% confidence band for the probability of a hive collapse as a function of pollutant exposure, $\Pr(y = 1 \mid \alpha, \beta, x)$. This will plot your predicted hive collapse probabilities for dosages from $x = 0$ to 2. Verify that you computed the LD50 correctly by identifying the x-value at which the posterior mean crosses 0.5.

```
xgrid <- seq(0, 2, by=0.1)

## Evaluate probability on the xgrid for one alpha, beta sample
compute_curve <- function(sample) {
  alpha <- sample[1]
  beta <- sample[2]

  prob <- NULL # YOUR CODE HERE
  prob

}

predictions <- apply(cbind(alpha_samples, beta_samples), 1, compute_curve)

quantiles <- apply(predictions, 1, function(x) quantile(x, c(0.025, 0.25, 0.75, 0.975)))
posterior_mean <- rowMeans(predictions)

tibble(x=xgrid,
       q025=quantiles[1, ],
       q25=quantiles[2, ],
       q75=quantiles[3,],
       q975=quantiles[4, ],
       mean=posterior_mean) %>%
  ggplot() +
  geom_ribbon(aes(x=xgrid, ymin=q025, ymax=q975), alpha=0.2) +
  geom_ribbon(aes(x=xgrid, ymin=q25, ymax=q75), alpha=0.5) +
  geom_line(aes(x=xgrid, y=posterior_mean), size=1) +
  geom_vline(xintercept = ld50, linetype="dashed") +
  geom_hline(yintercept = 0.5, linetype="dashed") +
  theme_bw(base_size=16) + ylab("Probability of hive collapse") + xlab("Dosage")
```