STAT 535 — Bayesian Data Analysis
Test 2 — Spring 2024

Important: Note: For this midterm exam, you are not allowed to receive help from anyone except me on the exams. For example, you may not talk to other students about the exam problems, and you may not look at other students' exams. Violations of this policy may result in a 0 on the exam, an F for the course, and/or punishment by the USC Office of Academic Integrity.

I will answer queries asking for clarification about the exam questions. Since this is an exam (and not homework), I will probably decline to provide very much help in solving the problems, but I am happy to clarify questions if necessary.

In addition, the computer programming/coding and writing of the answers on this exam must be done entirely by you — not with the help of any other individual or any AI program such as ChatGPT. You are welcome to use the textbook, course website, and other STAT 535 materials as aids in doing the problems. If you use other background sources (it is unlikely that you will need to do so), then you must cite the sources you used.

All data are given on my personal course website.

For all questions, show as much work as possible/appropriate! If you use R to solve a problem, please include the R code (e.g., put the code in an appendix). You should upload a Word document or pdf with your answers into Blackboard by Monday, April 8 at 11:59 p.m.

1. Let X be a discrete random variable with support $\{0, 1, 2, \ldots, N\}$ and let Y be a continuous random variable with support on the interval $[0, 1]$. If the joint density function of X and Y is proportional to

$$\binom{N}{x} y^{x+\alpha-1}(1-y)^{N-x+\beta-1}$$

then it can be shown that the conditional distributions are:

$$X|y \sim \text{Binom}(N, y)$$

and

$$Y|x \sim \text{Beta}(x+\alpha, N-x+\beta)$$

Suppose $N = 10$, $\alpha = 2$, and $\beta = 1$. Write a simple Gibbs sampler to sample 10000 times from the conditional distributions of $X|y$ and $Y|x$. Choose any sensible initial values for your Gibbs sampler, and then remove the initial values from the sample. Use the resulting 10000 pairs of (X, Y) values to approximate the following quantities: $E(X)$, $E(Y)$, $E(XY)$, the correlation $cor(X, Y)$, and the probability $P[X > 10Y]$.

Repeat the entire process with $N = 10$, $\alpha = 3$, and $\beta = 1$, and report your approximations to the quantities of interest. Based on your results, suggest general formulas for $E(X)$ and $E(Y)$.

2. Our goal is to build a regression model for an insurance company to relate the claim amount for an automobile accident (Y, in thousands of dollars) to the age of the damaged vehicle (X, in years). We believe that for a specific value of the vehicle age x, the claim amount follows a Gamma distribution with shape parameter

$$e^{\beta_0+\beta_1/\sqrt{x}}$$

and rate parameter 1.

(a) The unknown parameters of interest are β_0 and β_1. For these suppose we use independent priors, each being normal with mean 0 and variance 225. Explain what this choice of prior says about our prior knowledge.

(b) Claim amounts and vehicle age values from a sample of automobile accidents are given on the course website; suppose these constitute a random sample from the population of interest. Carry out a Metropolis-Hastings algorithm to sample values from the posterior distribution of $\boldsymbol{\beta} = (\beta_0, \beta_1)$.

(c) Explain how you could alter the algorithm if the acceptance rate was too high or too low.

(d) Give a Bayesian point estimate and interval estimate for β_0 and β_1. If appropriate, interpret what your estimates of β_0 and β_1 imply, in the context of the variables in the problem.

(e) Plot an estimate of the function $E(Y|x)$ as a function of x, for values of x from 0.5 to 15. Explain what this estimated model implies about the relationship between the claim amount and the age of the vehicle.

(f) Use your estimated model to estimate the expected claim amount for a damaged vehicle that is 10.5 years old.

3. A regression analysis was undertaken to study the relationship between the number of calories (X) of a collection of high-fiber oatmeals and the amount of sugar (Y, in grams). The analyst decided to fit (based on data for 12 oatmeals) a linear regression having the model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \;\; \epsilon_i \sim N(0, \sigma^2), \;\; i = 1, \ldots, 12.$$

```
y <- c(17.7, 24.1, 18.2, 21.0, 18.5, 19.6, 22.6, 16.1, 17.2, 12.4, 11.3, 11.8)
x1 <- c(200, 210, 160, 190, 200, 185, 220, 205, 210, 195, 185, 204)
```

(a) Before looking at the data, the analyst asks an expert to give a guess for the expected sugar amounts for two hypothetical oatmeals: one with 170 calories and another with 210 calories. The expert guessed expected sugar amounts of 14 g and 20 g, respectively. Explain why merely these two "hypothetical prior observations" are sufficient to obtain the necessary prior information on $\boldsymbol{\beta}$ needed for the conjugate analysis we studied in class.

(b) Based on the information given, suggest a prior mean vector for the prior on $\boldsymbol{\beta}$.

(c) For the diagonal matrix $\mathbf{D}$ that plays a role in the conjugate analysis, choose a diagonal matrix with all 1's along the diagonal. What does such a choice of $\mathbf{D}$ imply about our level of certainty in our prior information about $\boldsymbol{\beta}$?

(d) Choose a gamma prior on the error precision parameter τ that has first parameter $a = 0.5$ and second parameter $b = 6.67$. What does the value $a = 0.5$ imply about our level of certainty in our prior information about τ? [Alternatively, if using `stan_glm`, you could choose a default exponential prior on the error standard deviation, but still answer the question above about $a = 0.5$.]

(e) Estimate the model: Write your fitted model with point estimates for the model coefficients, and provide a 90% credible interval for the coefficient of number of calories. Use the model to predict the sugar amount for an oatmeal with 215 calories.

(f) Check the model fit and predictive accuracy using your favorite techniques. Be clear and complete in your explanations.

4. A multiple regression model is being built to predict the response variable $Y =$ **common (base-10) logarithm** of the survival time in days using a set of 4 candidate predictors (see the file at `https://people.stat.sc.edu/hitchcock/survivaldatabayes.txt` on the course website for data). The data set consists of 54 patients. Perform a Bayesian regression with noninformative priors for $\boldsymbol{\beta}$ and σ^2.

```
x1 = blood-clotting index
x2 = prognostic measurement
x3 = enzyme measurement
x4 = liver function measurement
y = survival time (had been measured in days, before common (base-10) log transformation)
```

(a) Explain why the analyst may have chosen to define the response variable Y as the (common) logarithm of survival days rather than as survival days itself.

(b) Based on your R analysis (specifically the credible intervals for the β's) which of the predictor variables seem to have higher posterior probabilities of being "unimportant?" Explain your answer.

(c) Give a point estimate for the expected survival time (in days) for a patient with blood-clotting index 6.0, prognostic measurement 65, enzyme measurement 72.00, and liver function measurement 1.50. Show how you got your answer.

(d) Give a point estimate for the ratio of expected survival time (in days) for patients with enzyme measurement 73.00 to expected survival time (in days) for patients with enzyme measurement 72.00 (holding other predictors constant). Explain briefly how you got your answer.

(e) Consider selecting a set of predictor variables in the model. If we only consider first-order terms as potential predictors (no interactions), then use any model selection techniques to choose among the class of possible models. What model is chosen as best? How does this relate to your answer in part (b)?

5. The file at `https://people.stat.sc.edu/hitchcock/soccerdataJan2018.txt` contains data on 45 randomly selected soccer games played by the "Cosmos," a professional soccer team. Note that the first column is just the observation number and should not be used in the analysis. The response variable y is the number of goals scored by the Cosmos in each game; x_1 is a computer-generated power rating of their opponent for that game (scaled to be within 0 and 100, with a higher rating indicating a stronger opponent); x_2 is an indicator of whether the game was a home game for the Cosmos (0 = away, 1 = home); x_3 is an indicator of whether the game was a night game (0 = day, 1 = night); x_4 is the number of players on the Cosmos' roster who were injured for that game; and x_5 is the number of players on the opponents' roster who were injured for that game.

(a) Fit a Bayesian Poisson regression model (clearly explaining how whatever prior beliefs you may have had were incorporated into your model) to help shed light on some questions of interest: Which variables affect or predict number of goals scored? What is the apparent relationship between each important predictor and goals scored? Do these relationship(s) depend on the values of the other predictors? Give the regression equation of your "best" model, and justify the choice of model in a few brief comments. Verify whether your model provides a reasonable fit to the data.

(b) Propose another reasonable regression model other than your chosen "best" model from part (a), and use any reasonable model selection criterion to compare those two proposed models.

(c) The Cosmos' next game is a home night game against a team with a power rating of 70.5. The Cosmos have one players injured and the opponent has two players injured. Discuss the posterior predicted number of goals the Cosmos will score in this game, using your "best" model.

(d) For the game in part (c), give the entire posterior predictive probability distribution for the number of goals the Cosmos will score in the game. You may round off the predicted probabilities to two decimal places.

6. For parameter μ, suppose you have a prior model that is Normal with mean 8 and variance 8^2. After seeing the data, the posterior model is Normal with mean 5 and variance 5^2. You wish to test $H_0 : \mu \geq 6$ vs. $H_a : \mu < 6$.

   (a) Give the posterior probability that the alternative hypothesis is true. Also calculate and interpret the posterior odds of the alternative hypothesis.

   (b) Calculate and interpret the prior odds of the alternative hypothesis.

   (c) Calculate the Bayes Factor for the alternative hypothesis. Interpret this in plain English for someone with little familiarity with Bayesian statistics.

   (d) If we had wanted the Bayes Factor for the **null** hypothesis, what value would this be?