Final Exam, Due by 6:00 PM MST, Wednesday, April 22nd

STAT 251

Winter 2020

You must email me a copy of the exam that has been knitted to a .pdf or .html file. You must work alone. You are not permitted to receive from nor give any assistance to any mortal in regards to any of the exam content.

On this exam you are permitted to use your notes, your past homeworks/quizzes/exams, and the R help pages. You are also premitted to use other resources on the internet, but not to solicit or give information. That is, you can use the internet passively but not to make or answer active requests for guidance. Please put `set.seed(101)` at the beginning of your code.

1. An experiment was designed to determine whether a mineral supplement was effective in increasing yield in milk. Fifteen pairs of identical twin dairy cows were used as the experimental units. One cow from each pair was randomly assigned to the treatment group that received the supplement. The other cow from the pair was assigned to the control group that did not receive the supplement. The milk yields over the time of the experiment are given below

```
control <- c(35.25,43.21, 47.63, 48.99, 32.34, 34.69, 34.39,
             36.58, 33.85, 32.26, 36.71, 35.01, 38.42, 39.98, 40.04)

tmt <- c(35.40, 44.79, 51.10, 50.66, 31.25, 38.80, 39.65, 38.49,
         34.97, 32.24, 34.18, 33.46, 44.45, 42.86, 39.11)
```

   Assume that the annual yields from cows receiving the treatment (y_{it}) are $y_{it}|\mu_t, \sigma_t^2 \overset{iid}{\sim} N(\mu_t, \sigma_t^2)$, and that the annual yields from the cows in the control group (y_{ic}) are $y_{ic}|\mu_c, \sigma_c^2 \overset{iid}{\sim} N(\mu_c, \sigma_c^2)$. Researchers want to test the hypothesis that H_0: $\mu_c - \mu_t \geq 0$ vs H_1: $\mu_c - \mu_t < 0$

   (a) Using $\mu_t \sim N(40, 10^2)$ (note that 10 is the standard deviation), $\mu_c \sim N(35, 10^2)$ (note that 10 is the standard deviation), $\sigma_t^2 \sim IG(1,1)$, and $\sigma_c^2 \sim IG(1,1)$ collect 10,000 samples from the joint posterior distribution for the treatment and control groups and provide evidence that the algorithm converged.

   (b) Plot the posterior and prior distribution for each parameter on the same graph (you should have four figures). Make sure to include a legend to distinguish the prior from the posterior.

   (c) Using the draws from the posterior distribution, test the hypothesis provided above. Clearly indicate what conclusions you make in the context of the problem

   (d) A separate researcher decided that since the two cows in the same pair share identical genetic background, their responses will be more similar than two cows that were from different pairs. There is a natural pairing. As the samples drawn from the two populations cannot be considered independent of each other, they decided to take differences $d_i = y_{ic} - y_{it}$. The differences will be assumed conditionally $iid\ N(\mu_d, \sigma_d^2)$, where $\mu_d = \mu_c - \mu_t$ and we will assume that $\sigma_d^2 = 7$ is known. The researcher believes that the mean of the differences in yeild between the treatment and control groups is fairly close to zero and is quite certain (say 95%) that the difference is not more than ± 15

      i. Formulate a prior distribution for μ_d according to the information provided by the researcher.

ii. Is the assumption of normality for differences in yield sensible?

iii. Identify the posterior distribution of μ_d (i.e., $\pi(\mu_d|\sigma_d^2, d_1, \ldots, d_n)$) and use it to test the hypothesis that H_0: $\mu_d \geq 0$ vs H_1: $\mu_d < 0$

(e) Which of the two analysis is more appropriate in this scenario? Why? In your answer to this problem focus on how the analysis was carried out, not the fact that in one scenario the population variance is assumed known and in the other it is not.

2. In class we studied the association between BMI of Pima indian women and diastolic blood pressure (dbp). We found that there was a positive assocition between these two variables and we quantified this association using a Bayesian simple linear model. In this part of the exam you will be tasked with answering the question "Is the regression slope between BMI and bloop pressure different for diabetic pima women relative to nondiabetic?" You will answer this question by fitting two simple linear regression models. Let y_{di} and x_{di} be the blood pressure and BMI measurements for the ith subject that is diabetic and y_{di} and x_{di} be the blood pressure and BMI measurements for the ith pacient that is healthy. Now the two models you will fit are

$$y_{di} = \beta_{0d} + \beta_{1d}x_{di} + \epsilon_i \text{ where } \epsilon \sim N(0, \sigma_d^2) \text{ and } i = 1, \ldots, D$$
$$y_{hi} = \beta_{0h} + \beta_{1h}x_{hi} + \epsilon_i \text{ where } \epsilon \sim N(0, \sigma_h^2) \text{ and } i = 1, \ldots, H.$$

Here D and H are the number of diabetic and healty subjects, and (β_{0d}, β_{1d}) are the slope and intercept for the diabetic group while (β_{0h}, β_{1h}) are the slope and intercept for the healthy group. Assume that the prior distribution for intercepts from both groups is $N(m_0, v_0)$ and that the prior distribution for the slope from both groups is $N(m_1, v_1)$ and that the prior for both variances in the error term is $IG(a, b)$. Use the following prior values $m_0 = m_1 = 0$, $v_0 = v_1 = 100^2$, and $a = b = 1$.

To answer the question well, you will need to convince me that the MCMC algorithms for both models have converged and mix well. Then you will need to report a summary of the posterior distribution of $\beta_{1h} - \beta_{1d}$ that provides concrete evidence in support of the answer you supply.