Bayesian Analysis I, FALL 2024, TAKE-HOME Assignment

INSTRUCTIONS: Be sure to include your **codes, output and detailed comments** for each problem in your write-up. When submitting, convert your write-up into a single file (preferably in PDF) that includes your *name* and *student number*. You may use any books, references, notes, but are **not allowed** to discuss these problems with any person other than the instructor until the due date. No credit will be given if any collaboration is detected.

1. We aim to assess the efficacy of a COVID-19 vaccine based on data from the Pfizer Phase III clinical trial. The trial involves two groups: a placebo group and a vaccine group. The number of infected participants in each group is assumed to follow a Binomial distribution.

Group	**Infected Participants**	**Severe Cases**	**Total Participants**
Placebo (Pfizer Study)	162	9	18,325
Vaccine (Pfizer Study)	8	1	18,198

Let θ_p represent the infection probability for the placebo group, and θ_v for the vaccine group. i.e. the probability that a randomly selected individual in the placebo group contracts the virus is θ_p, while the probability that a randomly selected individual in the vaccine group contracts the virus is θ_v.

The efficacy of the vaccine is often defined as:

$$E = 1 - \frac{\theta_v}{\theta_p}$$

Using these data, answer the following questions:

a) With a noninformative conjugate prior distribution for θ_p, obtain the posterior distribution of θ_p. Similarly, derive the posterior distribution of θ_v.

b) Using all patients, conduct a Bayesian analysis of the efficacy of the vaccine, E. Specifically, determine whether the vaccine has efficacy at least 0.70 and quantify the uncertainty in vaccine's efficacy using credible intervals.

c) Are the results from b) sensitive to different priors (i.e. do they change a lot when you change the priors)? Analyze and compare the outcomes using 3-4 distinct priors of your choice. Summarize the results with a clearly-labeled plot and table, and provide comments on your findings.

2. A study was conducted on 32 cars to explore the relationship of the gasoline consumption on the the weight of the car and engine sizes in cylinders.

For each car we have observations on how many miles that car can travel on a gallon of gasoline (mpg), the weight of the car (weight) and two dummy variables that indicates if the car's engine has four cylinders (sixcyl=0 and eightcyl=0) six cylinders (sixcyl=1 and eightcyl=0) or eigth cylinders (sixcyl=0 and eightcyl=1).

mpg: 21.0, 21.0, 22.8, 21.4,18.7,18.1,14.3, 24.4, 22.8,19.2,17.8,16.4,17.3,15.2,10.4,10.4,14.7, 32.4, 30.4, 33.9, 21.5, 15.5,15.2,13.3,19.2, 27.3, 26.0, 30.4,15.8,19.7,15.0, 21.4

weight: 2.620, 2.875, 2.320,3.215,3.440, 3.460, 3.570, 3.190, 3.150, 3.440, 3.440, 4.070, 3.730, 3.780, 5.250, 5.424, 5.345, 2.200, 1.615, 1.835, 2.465, 3.520, 3.435, 3.840, 3.845, 1.935, 2.140, 1.513, 3.170, 2.770, 3.570, 2.780

sixcyl: 1, 1, 0, 1, 0, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0

eightcyl: 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0

We want to sample from the joint posterior distribution in the Normal linear regression:

$$mpg = \beta_0 + \beta_1 * weight + \beta_2 * sixcyl + \beta_3 * eightcyl + error, \quad error \sim N(0, \tau)$$

with conjugate priors

$$\begin{aligned} \beta_i &\sim N(0, 10000), \quad i = 0, ..., 3 \\ 1/\tau &\sim Gamma(0.01, \text{rate} = 0.01) \end{aligned}$$

a) Give the plots of the marginal posterior distributions for the parameters, β_1, β_2, and β_3.

b) Construct 95% equal tail probability intervals for each parameter and interpret them.

c) Investigate if the effect on mpg is different in cars with six cylinders compared to cars with 8 cylinders.

d) Obtain the predictive distribution for a new 4 cylinder car with weight = 3.5. Give comments.

(Extra-credit)
e) It appears that the relationship between weight and mpg may vary across cars or subgroups. Introducing weight as a random effect in the model could provide more flexibility in capturing these variations. Let u_i represent the random effect associated with the weight for each individual car, assumed to be drawn from a Normal distribution:

$$u_i \sim N(0, \sigma_u^2).$$

This random effect term can be added to the model mean to account for variations across individual cars. How do the posterior results change when this random effect is incorporated, compared to a model where the random effect is not included?