STAT 7630: Homework 7
(Due: Tuesday, 11/19/2024)

Note: Show all your work for the necessary steps to receive full credit.

Please turn in the HW on paper, hand-written and/or typed. For computational problems, return only the relevant parts of the output with comments/annotations. Questions taken from the textbook are marked with BR for "Bayes Rules!". From the code you are using to answer the problems, turn in the relevant output, and the figures (if requested), preferably printed from the output. No need to turn in your code or long lists of generated samples.

Please disclose any us of AI in your solutions. Regardless of you use it or not, make sure you submit your own work, not copy from other source(s). Any suspicion of AI use will result in automatic 0 or substantial point loss in any question.

BR Chapter 9:
Do Exercises 9.1 and 9.3.

Additional Questions (AQs):
AQ1. Consider the regression relationship between the yearly income X (in thousands of dollars) of a home buyer and the sale price Y of a home (in thousands of dollars). (Note: these are older data, and home prices today would likely be much higher.)

(a) Suppose that, before examining the data in this study, an expert suggested that a hypothetical buyer with a yearly income of $40,000$ would be expected to purchase a house valued at $115,000$, and a buyer with a yearly income of $60,000$ would be expected to purchase a house valued at $150,000$. If we assume that this prior knowledge is equivalent to one sample observation, use the approach discussed in class to formulate a conjugate prior for the vector of regression coefficients β. Determine the matrix $\mathbf{D}$, and specify the prior distribution of $\tilde{X}\beta|\tau$. Also, provide the prior mean for β based on this information, and be as specific as possible.

(b) Suppose the prior information on τ is weak, valued by the expert as equivalent to only 0.2 sample observations. Given that the expert considers the maximum reasonable house price for a buyer earning $40,000$ per year to be $195,000$, determine appropriate parameter values for a gamma prior on τ.

(c) A random sample of 14 house purchases was obtained, with paired data as follows:

$$X : 28.5, 30, 31.5, 32, 33.5, 35.9, 39, 40.5, 42.5, 45, 54.6, 62.3, 70, 80$$

$$Y : 94, 93.5, 99.5, 105, 110, 116, 125, 130.6, 129.9, 140, 170, 171, 185, 177$$

Based on your priors in parts (a) and (b), find and provide Bayesian point estimates for τ and the error variance σ^2.

(d) Using the priors defined in parts (a) and (b), calculate Bayesian point estimates and 95% posterior interval estimates for the elements of β. Write out the estimated regression model and use it to predict the house price for a new buyer with a yearly income of $60,000$.

AQ2. The "energy bar data set" on the course web page contains data on various types of energy bars. We will develop a regression model to predict the price of a bar using three explanatory variables. The following R code:

```
ener_dat <- read.table("insert your local path here /energybardata.txt",
header=F, col.names = c("price", "calories", "protein", "fat"))
attach(ener_dat)
```

will read in the data correctly.

(a) Estimate the regression model with "price" as the response and (X_1 = calories, X_2 = protein, X_3 = fat) as the predictors, using a Bayesian approach with noninformative priors on β and σ^2. Write the estimated linear regression function for predicting the price of an energy bar.

(b) Based on the posterior inference, identify the predictor(s) most likely to have a strong marginal effect on price.

(c) Adapt the Gibbs sampling model selection code from the course web page to perform a Bayesian model selection based on the response variable Y and the candidate predictor variables X_1, X_2, and X_3. Determine which model(s) appear best based on their posterior probabilities.

(d) Now, consider interaction (cross-product) terms X_1X_2, X_1X_3, and X_2X_3 (which can be coded as X_4, X_5, and X_6 in R) as additional candidate predictors. Perform a Bayesian model selection using all six candidate predictors (first-order and interaction terms), with the convention that no interaction term should appear in the model without each of its component variables appearing as first-order terms. Does the "best" model differ from the one chosen in part (c)? Explain.

AQ3. Consider a regression analysis for the cereal data provided on the course web page. The first column contains the brand names of the cereals and should not be used in the analysis. The response variable Y is "Sugar," and the explanatory variables are "Sodium," "Fiber," "Carbohydrates," and "Potassium." The following R code:

```
cer_dat <- read.table("insert your local path here /cerealdatabayes.txt", header=T)
# This creates a data frame cer_dat with columns named
# Sugar, Sodium, Fiber, Carbohydrates, Potassium
# Alternatively, you could create generically named response and predictor variables:
y <- cer_dat$Sugar
x1 <- cer_dat$Sodium
x2 <- cer_dat$Fiber
x3 <- cer_dat$Carbohydrates
x4 <- cer_dat$Potassium
cer_dat.generic <- data.frame(y, x1, x2, x3, x4) # creating a data frame with these vari
```

will read in the data correctly.

(a) Perform a Bayesian linear regression of Y on X_1, X_2, X_3, and X_4, and write the fitted regression equation. You may use any prior and Bayesian method to estimate the parameters. Be sure to clearly explain the prior used.

(b) After fitting the model, assess the model fit using the posterior predictive distribution. Identify any cereals that appear to be outliers, and characterize the overall adequacy of the model based on the posterior predictive distribution analysis.

(c) Predict the sugar content of a cereal with values: sodium = 140, fiber = 3.5, carbohydrates = 14, and potassium = 90. Provide a point prediction and a 90% prediction interval using the Bayesian approach.

AQ4. For the cereal data provided in the regression example above, define a set of candidate regression models to consider. Use any reasonable model selection method to determine the "best" model among your candidate models. Carefully explain how you made your selection and justify it based on your chosen model selection criteria. Estimate the parameters of your "best" Bayesian linear regression model and write the fitted regression equation.