STAT 535 – Spring 2024 – HW 4

1. Do Exercise 9.1 in the *Bayes Rules!* textbook.

2. Do Exercise 9.3 in the *Bayes Rules!* textbook.

3. 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). (By the way, these are old data; as I know all too well, these days the home prices would be much higher!)

   (a) Suppose that before looking at the data in this study, an expert opined that a hypothetical buyer making \$40,000 per year would be expected to purchase a house worth \$115,000, and a buyer making \$60,000 per year would be expected to purchase a house worth \$150,000. If we assume that this prior knowledge together is worth a **total** equivalent to one sample observation, use the approach we outlined in class to formulate a conjugate prior for the vector of regression coefficients. What is your $\boldsymbol{D}$? What is your prior on $\tilde{\boldsymbol{X}}\boldsymbol{\beta}|\tau$, and thus what is your prior mean for $\boldsymbol{\beta}$ here? (Be as specific as possible.)

   (b) Suppose the prior information on τ is weak, worth only 0.2 sample observations in the opinion of the expert. Use the fact that the expert feels that the highest reasonable house price is \$195,000 for the hypothetical buyer making \$40,000 per year to obtain appropriate parameter values for a gamma prior on τ.

   (c) A random sample of 14 house purchases was obtained. The paired data were:
   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 (a) and (b), find and provide a Bayesian point estimate for τ and for the error *variance* σ^2.

   (d) Based on your priors in (a) and (b), find and provide Bayesian point estimates and 95% posterior interval estimates for the elements of $\boldsymbol{\beta}$. Write the estimated regression model and use it to predict the house price for a new buyer who makes \$60,000 per year.

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

```
ener.data <- read.table("http://people.stat.sc.edu/hitchcock/energybardata.txt",
header=F, col.names = c("price", "calories", "protein", "fat")) ; attach(ener.data)
```

   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 $\boldsymbol{\beta}$ and σ^2. Write the estimated linear regression function for predicting energy bar price.

(b) Based on the posterior inference, which predictor or predictors are most likely to have a strong marginal effect of price?

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

(d) Now consider interaction (cross-product) terms X_1X_2, X_1X_3, X_2X_3 (you could code these as X_4, X_5, X_6 in R) as other candidate predictors. Perform a Bayesian model selection using all six candidate predictors (first-order and interaction terms), using 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 change from the one chosen in part (c)? Explain.

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

```
cer.data <- read.table("http://people.stat.sc.edu/hitchcock/cerealdatabayes.txt", header=T)
# This creates a data frame cer.data with columns named
# Sugar, Sodium, Fiber, Carbohydrates, Potassium

# Alternatively you could create generically named response and predictor variables:
y <- cer.data$Sugar
x1 <- cer.data$Sodium
x2 <- cer.data$Fiber
x3 <- cer.data$Carbohydrates
x4 <- cer.data$Potassium
cer.data.generic <- data.frame(y,x1,x2,x3,x4) #creating a data frame with these variables
```

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. For this problem, you can use whatever prior and Bayesian method to estimate the parameters that you want. Just be sure to clearly explain the prior you are using.

(b) After fitting the model, use the posterior predictive distribution to check the model fit. Do any cereals appear to be outliers? How would you characterize the overall adequacy of the model, based on the posterior predictive distribution analysis?

(c) Predict the sugar content of a cereal having sodium=140, fiber=3.5, carbohydrates=14, and potassium=90. Give a point prediction as well as a 90% prediction interval, using the Bayesian approach.

6. For the cereal data given in the regression example above, define a set of several candidate regression models that you will consider. Using any reasonable model selection method, decide on the "best" model among your candidate models. Carefully explain

how you made your choice, and justify it based on your model selection criteria. Estimate the parameters of your "best" Bayesian linear regression model, and write the fitted regression equation.