HOMEWORK 11, STAT 251

In this homework we will compare a frequentist regression analysis to a Bayesian. The first few problems of this homework should be a review from the simple linear regression component of your introductory stat course. Remember to submit your R code along with the completed homework.

(1) A researcher measured heart rate (x) and oxygen uptake (y) for one person under varying exercise conditions. He wishes to determine if heart rate, which is easier to measure, can be used to predict oxygen uptake. If so, then the estimated oxygen uptake based on the measured heart rate can be used in place of the measured oxygen uptake for later experiments on the individual. Below I provide the heart rate and oxygen uptake measurements take on the subject

```
heart.rate <- c(94, 96, 94, 95, 104, 106, 108, 113, 115, 121, 131)
oxygen.uptake <- c(0.47, 0.75, 0.83, 0.98, 1.18, 1.29, 1.40, 1.60, 1.75, 1.90, 2.23)
```

To learn about this assocation the researcher decides to employ a simple linear regression model which is the following

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{with} \quad \epsilon_i \sim N(0, \sigma^2)$$

(a) List the assumptions that accompany the simple linear regression model.

(b) Plot oxygen uptake (y) versus heart rate (x) using a scatterplot with both axis correctly labeled.

(c) Using the `lm` function in `R` fit the simple linear regression model and plot it. (Hint: you fit the model using `lm(y ~ x)`)

(d) Perform the hypothesis test H_0: $\beta_1 = 0$ vs H_1: $\beta_1 > 0$. What conclusions can you make?

(e) You should always explore the appropriateness of the assumptions that accompany the model, but I won't have you explore them for this homework.

Now we are going to perform a Bayesian analysis. To do this we reexpress the simple linear regression model as

$$y_i | \beta_0, \beta_1, \sigma^2, x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

and assume the following prior distributions

- $\beta_0 \sim N(m_0, v_0)$
- $\beta_1 \sim N(m_1, v_1)$
- $\sigma^2 \sim IG(\text{shape} = a, \text{rate} = b)$

(f) What are the assumptions associated with the Bayesian simple linear regression model?

To "fit" the model from a Bayesian perspective we need to derive the joint posterior distribution for β_0, β_1, and σ^2. As we discussed in class, this joint distribution does not have a form of a well known probability distribution and as a result is not easily sampled from (e.g., there does not exist any "`r`" function in `R`). Because of this, we resort to constructing a Gibbs-sampler that is able to take or sample draws from the joint posterior distribution. To build a Gibbs-sampling algorithm, we need the full conditions for β_0, β_1, and σ^2.

From class recall that

$$\pi(\beta_0|\beta_1,\sigma^2,y_1,\ldots,y_n) = N\left(\frac{\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\beta_1 x_i)+\frac{1}{v_0}m_0}{\frac{n}{\sigma^2}+\frac{1}{v_0}}, \frac{1}{\frac{n}{\sigma^2}+\frac{1}{v_0}}\right)$$

$$\pi(\beta_1|\beta_0,\sigma^2,y_1,\ldots,y_n) = N\left(\frac{\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i(y_i-\beta_0)+\frac{1}{v_1}m_1}{\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i^2+\frac{1}{v_1}}, \frac{1}{\frac{1}{\sigma^2}\sum_{i=1}^{n}x_i^2+\frac{1}{v_1}}\right)$$

$$\pi(\sigma^2|\beta_0,\beta_1,y_1,\ldots,y_n) = IG\left(\frac{1}{2}n+a, \frac{1}{2}\sum_{i=1}^{n}(y_i-[\beta_0+\beta_1 x_i])^2+b)\right)$$

(g) Show that full conditional for σ^2 is what I listed above. Showing this should in theory consist of three or four steps.

(h) Write a computer program that employs a Gibbs-sampler to sample from the joint posterior distribution of β_0, β_1, and σ^2. For prior distribution parameters use $m_0 = m_1 = 0$, $v_0 = v_1 = 100$, and $a = b = 1$. Comment very briefly on if you think these values produce reasonable prior distributions.

(i) Argue that the algorithm has converged using trace plots and discuss briefly the "mixing" of the chain using autocorrelation plots (e.g., use `acf` function in R)

(j) Plot the posterior (after removing enough burn-in draws) and prior distribution in the same graph for each parameter (this part should consist of three plots). Make sure to appropriately label the figures.

(k) Create a scatter plot that contains the data, the least squares line from part (c) and the the fitted Bayesian regression line. Use the color red for the least squares line and blue for the Bayesian line. Comment briefly on the differences you see between the frequentist fit and Bayesian fit.

(l) Test the hypothesis that H_0: $\beta_1 = 0$ vs H_1: $\beta_1 > 0$. We did not discuss how to do this type of hypothesis test from a Bayesian perspectve. As you might imagine, testing the hypothsis will require using the posterior distribution of β_1. Try your best to come up with a testing procedure based on the posterior distribution for β_1 that permits you to make a conclusion associated with the competing hypothses. Ask yourself, "What would convince me (based on the posterior distribution of β_1) that the hypothsis is wrong?"

(m) For the last problem of this homework, we will consider the posterior predictive in a regression setting. For a new individual with a heart-rate of 95 collect the same number of draws from the posterior predictive distribution as you collected from the joint posterior distribution of β_0, β_1, and σ^2 (after burn-in). What is the support of the posterior predictive distribution in a regression setting? Plot the distribution and briefly describe the information that it provides.