S&DS 238a/538a, Fall, 2019
Problem set #9, Due Fri Nov. 22

This looks long but I think that is at least partially an illusion because a lot of the length is an attempt to make problem statements very clear and to give some hints. Still, "very clear" does not necessarily imply "very easy to understand" and you may find the writing a bit dense, in which case slow and careful reading and re-reading can be worthwhile.

1. Suppose Y is a linear function of X plus noise, by which we mean:

   - X is a random variable with mean μ_X and variance σ_X^2,
   - ε is a random variable that we can interpret as "noise" or "measurement error," with ε independent of X, and ε having mean 0 and variance σ_ε^2,
   - $Y = \alpha + \beta X + \varepsilon$ for some (nonrandom) numbers α and β.

   Let ρ denote the correlation between X and Y. Use the definitions and properties of mean, covariance, and correlation to show that

   (a) $\alpha = \mu_Y - \rho \frac{\sigma_Y}{\sigma_X}\mu_X$ and $\beta = \rho\sigma_Y/\sigma_X$.

   (b) $\sigma_\varepsilon^2 = (1-\rho^2)\sigma_Y^2$.

   (c) Letting $\tilde{X} = \frac{X-\mu_X}{\sigma_X}$ and $\tilde{Y} = \frac{Y-\mu_Y}{\sigma_Y}$ denote the standardized X and Y variables, we can write $\tilde{Y} = \rho\tilde{X} + \tilde{\varepsilon}$, where $\tilde{\varepsilon}$ is another noise random variable (that is, it is independent of X) that has variance $1-\rho^2$.
   [[We could interpret the variance $1-\rho^2$ as quantifying how, as ρ gets close to 1 or close to -1, Y becomes more closely predictable (that is less variable) given X.]]

2. Let H and W denote the heights of the husband and wife in a married couple chosen randomly from a particular population. Assume that the marginal distributions of H and W are both Normal, with $H \sim N(70.0, 4.0^2)$ and $W \sim N(65.0, 3.5^2)$. Suppose the correlation between the heights of husbands and wives in married couples is 0.4, and assume that the joint distribution of (H,W) is *bivariate Normal*, which implies that all conditional distributions, such as the distribution of W given $H = h$, are Normal.

   (a) What is the probability of a random wife's height being at least 68.0 inches?

   (b) What is the equation of the regression line giving the expected wife's height as a function of the husband's height?

   (c) What is the conditional distribution of the height of a wife conditional on her husband's height being 74.0 inches?

   (d) What is the conditional probability of a wife's height being at least 68.0 inches, given that the husband's wife is 74.0 inches?

   [[Hints: An interpretation of the previous problem is that, conditional on $X = x$, the expected value of Y is the "y" on the regression line $\left(\frac{y-\mu_Y}{\sigma_Y}\right) = \rho\left(\frac{x-\mu_X}{\sigma_X}\right)$ and the conditional variance of Y is $(1-\rho^2)\sigma_Y^2$. Also, this is a lot like Example (5.4) on page 212). I would suggest you do your calculations of cumulative Normal probabilities with a Normal table* for practice for an exam or being stranded on a desert island, and you can check with R if you'd like.]]

3. [[Prediction intervals, and how George W. Bush got to be president]] The data here contain the numbers of votes for George W. Bush and Pat Buchanan in all 67 counties in Florida for the 2000 presidential election. It is easy to find information to review the interesting historical context; for example you can look at this wikipedia article, and in particular the first 3 paragraphs that show the "butterfly ballot" and explain what is interesting about the votes for Pat Buchanan. In this problem we will use the Bush vote to predict the Buchanan vote. We'll see that in Palm Beach County, Buchanan got surprisingly many votes, and we'll attempt to quantify how many more votes Buchanan got than would be plausibly expected under ordinary circumstances.

   (a) Draw a scatterplot of the votes, with the Buchanan vote on the vertical axis. Call attention to the Palm Beach point by using a different color (`col`) or plotting character (`pch`).

   (b) Do the same for the `log` of the Bush and Buchanan votes. Discuss how taking logarithms results in data that visually appears more suitable for analysis with the standard linear regression model.

   (c) Apparently anomalous aspects of voting in Palm Beach County led to suspicion that the butterfly ballot (used only in Palm Beach) was at least partially to blame. Since we are questioning the validity of the Buchanan vote count for Palm Beach, let's do a regression of log Buchanan votes on log Bush votes, without Palm Beach (that is, take the Palm Beach point out of the data set). In addition to the obvious variables to include in your regression model, also create an additional variable to represent a randomly generated new value for the log(Buchanan) vote in Palm Beach. A way to do this is to define a new variable, say `logPBPB` (for "log Pat Buchanan in Palm Beach"), by adding a line to your JAGS model like this:

```
logPBPB ~ dnorm(alpha + beta*log(152846), tau)
```

*This means if you have a cumulative Normal probability $\Phi(z)$ to calculate, you would approximate it by rounding off z to two decimal places and then look up the cumulative probability in a Normal table.

and then include `logPBPB` among the variables you monitor in your `coda.samples` command. The monitored values for `logPBPB` will be a sample of random log Buchanan vote counts drawn from the conditional distribution given the rest of the data. Find the 95th and 99th percentile of these `logPBPB` values, and then use `exp` to transform these logged quantities back to the vote scale. Call these values `PBPB95` and `PBPB99`.

(d) Subtract `PBPB95` and `PBPB99` from the actual observed Buchanan vote in Palm Beach to find 95% and 99% lower confidence bounds for the extra votes Buchanan received in Palm Beach over what he would be predicted to receive given the rest of the data. One explanation for any such extra votes could involve confusion over the butterfly ballot. As usual in data analysis, try to end with a summary and conclusions. A possibly relevant tidbit to keep in mind is that the official result in Florida was a victory for Bush by a margin of 537 votes over Al Gore, who many believe was the intended recipient of votes cast mistakenly for Buchanan because of confusion due to design of the butterfly ballot.

4. [[Martian heights: A mixture model. Here you will be monitoring and estimating hundreds of unknown quantities, not just a few parameters.]] You are the Statistician for the newly-formed United Planets Organization. The delegate from Mars tells you that Martians have heights that are Normally distributed and gives you a random sample of 400 Martian heights. These heights are in a vector `y` that you can create in your R session with the commands

```
dat <- read.csv("http://www.stat.yale.edu/~jtc5/238/data/martianHeights.csv")
y <- dat$y
```

The heights have been sorted in ascending order.

(a) You draw a histogram of the heights and note that the distribution does not appear Normal. Show the histogram.

Upon further inquiry, the delegate replies, "Oh, I meant that heights of male Martians have a Normal distribution, as do heights of female Martians, but the two distributions are quite different; for example, males have a larger mean." The data `y` consists of a randomly chosen mix of males and females, but, as noted above, the vector `y` has been sorted. You wish that you could know which of the y's are from males and which are from females, but this information is not available. You are told that the Martian sex ratio is 1:1, just like on Earth; that is, new births have probability 0.5 each for female and male.

(b) Shown below is a JAGS model, except it has been censored, in that there are some lines containing "???" that are left for you to complete. (Also there are other ways to express the same model, so you could consider deviating from the given structure if you have other ideas you would like to try.) Show your completed model.

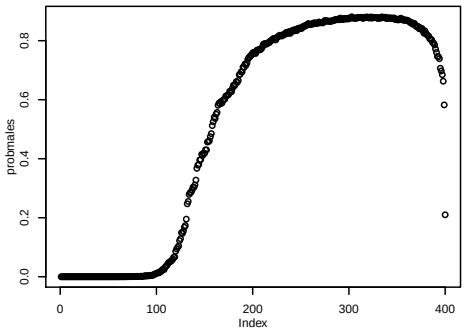
Tips: Let's say the variable gender is interpreted as follows: `gender[i]=1` means individual `i` is female, and `gender[i]=2` means individual `i` is male. You will want to put a prior distribution on `deltaMu` that can only take nonnegative values.

```
model{
  for(i in 1 : n) {
    y[i] ~ dnorm(mu[i], tau[i])
    mu[i] <- mus[gender[i]]
    tau[i] <- ???
    temp[i] ~ dbern(0.5)
    gender[i] <- ???
  }
  mus[1] ~ ???
  mus[2] <- mus[1] + deltaMu
  deltaMu ~ ???
  taus[1] ~ ???
  taus[2] ~ ???
  sigmas[1] <- 1 / sqrt(taus[1])
  sigmas[2] <- 1 / sqrt(taus[2])
}
```

(c) Run JAGS. In addition to the four parameters (the mean and SD for males and also the mean and SD for females), also monitor the `gender` vector as you run JAGS, so that you will get a sample from the posterior distribution of `gender` for each of the 400 martians. For each of the four parameters, draw a histogram and give an interval that contain approximately 95% posterior probability.

(d) What is your estimate of the posterior probability that the shortest Martian is a male (that is, the probability that `gender[1]` is 2)? How about the 200th tallest (that is, the probability that `gender[200]=2`)? The tallest (probability that `gender[400]=2`)? In view of the fact that males have the larger mean, the last probability should seem a bit strange at first; can you give an intuitive explanation of this last probability?

(e) Calculate such a probability for each Martian, and draw a plot of your estimated probabilities that each of the Martians is male. As a check, I got the plot below (so presumably yours should look similar, although not identical since this is Monte Carlo).



5. [[A "mixed-effects" or "multilevel" model for blood pressures in male vs female Martians]] The data file martian-blood-pressure.csv contains blood pressure measurements taken on 10 randomly chosen male and 10 randomly chosen female Martian patients. For each patient the data file contains between 1 and 5 blood pressure measurements; that is, a single patient could have multiple measurements recorded on different occasions. The file has 3 columns: `id` (the patient id), `male` (0 for female, 1 for male), and `bp` (the blood pressure measurement value). Each row of the file corresponds to one measurement of blood pressure and records who was measured, their gender, and the blood pressure value obtained. For example, if you take a look you will see that patient 1 was measured twice, patient 2 was measured 4 times, and so on, for a total of $N = 61$ blood pressure measurements.

(a) Load the data file into R using `read.csv` as usual. Run[†] the command `table(table(id))` and say in words what the result tells us.

(b) Let n_m and n_f denote the number of rows in the data table from males and females, respectively. Use R to find n_m and n_f.

(c) Suppose we modeled `bp` as consisting of two independent samples of size n_m and n_f from a male and female population of blood pressures (although, as we will see below, this not an appropriate model). Make a JAGS model where you assume the n_m measurements from males come from a $N(\mu_m, \tau_m)$ distribution and the n_f measurements from females are a sample from a $N(\mu_f, \tau_f)$ distribution, where of course μ_m, τ_m, μ_f, and τ_f are unknown parameters (this is similar to the "subliminal" example we did in class in the last part of the R script from Nov. 10, although for this question please put a `dgamma(.01,.01)` prior on the precisions instead of putting an exponential prior on the standard deviations). Give your posterior probability that μ_m is less than μ_f. Draw a histogram for the posterior distribution of the difference $\mu_m - \mu_f$ and state a 95% posterior probability confidence interval for this difference. If you believed in this model and analysis, what would you conclude? For example, would you feel that there is very strong evidence about a difference between population averages of male and female blood pressures?

The model from part (5c) would be appropriate if we had n_m males and n_f females in the study and independent measurements from each individual. But we really have only 10 males and 10 females with repeated measurements from the individual patients. Next you will make a model that attempts to capture this structure; such models are known as "mixed-effects" or "multilevel" models. The model should include 20 variables, say, `true[1]`, `true[2]`, ..., `true[20]`, where `true[i]` represents the unknown "true" value of the blood pressure for patient `i` (that is, the patient whose `id` is `i`). Our model for the `true` vector is that `true[1],...,true[10]` is a sample from a Normal distribution $N(\mu_m, \tau_m)$ and `true[11],...true[20]` is a sample from $N(\mu_f, \tau_f)$. Each measurement of the blood pressure for patient `i` is considered to be an independent draw from a Normal distribution with mean `true[i]` and precision `tau.e`, say. The parameter `tau.e` quantifies the precision of measurement "errors" in blood pressures; it is assumed to be the same for different patients (that's why our model has a single `tau.e` and not 20 `tau.e[j]`'s). To try to help clarify the setup, let's look at patient 7, for example, who has a "true" blood pressure `true[7]`. The measurements `bp[19]`, `bp[20]`, and `bp[21]` are for patient 7, and so they come from a Normal distribution with mean `true[7]` and precision `tau.e`. In this sense, `true[7]` is the expected (mean) bp measurement for patient 7, and the actual blood pressure measurements `bp[k]` belonging to patient 7 (that is, the `bp[k]`'s for `k`$\in \{1,\ldots,61\}$ such that `id[k]` is 7) come from a Normal distribution with mean `true[7]`. In other words: for each `k` $\in \{1,\ldots,61\}$, the measurement `bp[k]` comes from a Normal distribution with mean `true[id[k]]` and precision `tau.e`.

(d) Write a JAGS model to capture the description in the previous paragraph. As part of this you will want to choose priors for μ_m, τ_m, μ_f, σ_f, and `tau.e`.

(e) Answer the same questions as from part (5c): Give your posterior probability that μ_m is less than μ_f. Draw a histogram and state a 95% posterior probability confidence interval for the difference $\mu_m - \mu_f$. From this model and analysis, would you feel that there is very strong evidence for a difference between population averages of male and female blood pressures?

[†] If you get an error because R doesn't know what you mean by `id`, you should be able to fix that.