EXERCISE: POSTERIOR DISTRIBUTION OF EDUCATION EFFECTS

We will consider a model in which $Y \mid S \sim N(\mu_S, \sigma_S^2$, where Y is `logwage` in the AK data and S is years of schooling, `educ` in the AK data. The flat-prior posterior on μ_S, σ_S^2 is independent across S and of the usual Normal-Inverse-Gamma form.

We will Form estimates of posterior probabilities of several characteristics of the vector μ_S, $S = 0, \dots, 20$ — for example, the probability that $\mu_S > \mu_{S-1}$ for all S between 1 and 20.

Since the joint distribution of the μ_S's is of known form, we can write down an expression for this probability as an integral, but evaluating the integral analytically is difficult. There are two ways to do it easily by posterior simulation, and three interpretations of what justifies these approaches.

The two approaches are:

(a) Since the sample sizes are quite large even for the least common values of `educ`, treat the σ_S values as known exactly to be the sample standard errors of the means. Draw μ vectors from the corresponding normal posteriors.

(b) Use the full Normal-inverse-gamma: For each draw of the vector μ, first draw a vector of σ_S values from their respective inverse-gamma distributions, then use those σ_S values to generate draws from the corresponding conditional normal joint distributions for μ.

Approach (b) is justified by the standard normal model we have discussed. Approach (a) can be justified as a somewhat simpler approximation to (b) that will differ little from the latter because variances are so sharply estimated in samples of this size. Approach (a) can also be justified by observing that the sample mean of Y in each schooling cell is asymptotically normal, even if Y is not itself normal. So the procedure can be justified as approximately Bayesian based on treating the observed data as the sample means themselves. This interpretation will be discussed in the Tuesday 9/29 lecture.

Evaluate the following probabilities, each by both methods:

- $\mu_S > \mu_{S-1}$ for all $S \in 1, \dots 20$. (An additional year of schooling always corresponds to a higher expected log wage, whatever the year.)
- $\mu_{12} - \mu_{11} > \mu_{11} - \mu_{10}$. (Staying the last year of high school, and thereby graduating, gives a greater boost to expected income than staying the previous year.)
- $\mu_S > \mu_{S-1}$ for all $S \in 5, \dots, 16$. (Every year of schooling from 5th grade through 4 year college graduation increases expected log wage.)
- $\mu_S > \mu_{S-1}$ for all $S \in 17, \dots, 20$ (Every year of schooling beyond college graduation adds to expected log wage.)
- $\mu_{16} - \mu_{12} > \mu_{12} - \mu_8$. (High school adds less to expected income than four years of college does.)

Notes: To draw from the inverse-gamma, draw from the gamma, and invert the result.

Date: September 27, 2020.

In R the **`rgamma()`** and **`rnorm()`** functions will, if you ask for a vector of 20 random draws, allow you to input 20 different values for the distribution parameters.

Probably 1000 draws from the posterior on μ are enough. You can calculate all of the posterior probabilities asked for from the same set of random draws of μ. However, to check whether 1000 is enough, you might also try in one or two cases to see if taking a second sample of 1000 gives nearly the same answer.

It's tempting to interpret these results causally: that is, to think of the difference in mean log wage between S and $S + 1$ years of schooling as what happens to an individual's expected income if he or she stays in school that extra year. The AK paper itself attempted to deal with the problem that people with more "ability" might stay in school longer, so there is much less effect on an individual's earnings than the mean differences would suggest. However, Angrist and Krueger didn't find that accounting for this possibility made much difference to estimates.

Also, our kmeans exercise suggested that there might be some individuals who have low incomes regardless of years of schooling, while others have incomes that grow with schooling. This exercise does not account for such heterogeneity. You might try to think through how you could expand the model to recognize this kind of heterogeneity.

AK used a linear model, implicitly assuming every additional year of schooling had the same effect on log wage. Do your results on this exercise suggest any problems with the AK model linearity?