Statistical Modeling, Winter 2025

Homework 9

Please submit your answers to these questions via Canvas prior to class next Tuesday.

I will accept a pdf only. This means you can knit your `.Rmd` document into a pdf, or you can make a word document and insert relevant images of your code and plots.

Please name the file with your last name and hw number (e.g., elahi_hw1).

Complete the following:

- read chapter 15.2 in *Statistical Rethinking* (McElreath 2020).
- watch Missing Data

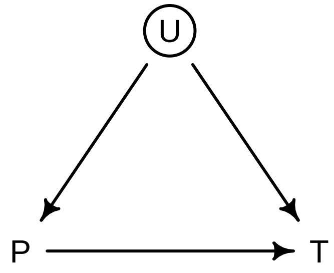
Questions

From McElreath videos 16-18 (SR2 14.5, 15.1, 15.2)

1. Gaussian processes are used to cope with unobserved categorical confounds that arise as a consequence of continuous variables. For example, the number of tool types (T) on oceanic islands is thought to be causally related to population size (P), but proximity to other islands may confound the causal estimate through any number of unobserved mechanisms (U; e.g., geography, tool sharing). In this case, spatial covariation in the distance between islands can 'stand in' for these island-related (categorical) confounds (SR2 14.5). For your study system of interest, can you think of a situation where Gaussian processes may help with such unobserved confounds? Draw the DAG and explain the rationale for the continuous categories used to stand in for unobserved categorical confounds.

```r
library(dagitty)
library(rethinking)
```

```r
dag1 <- dagitty( "dag {
    U [unobserved]
    P -> T
    P <- U -> T }")
coordinates(dag1) <- list(x = c(P = 1, T = 3, U = 2),
                          y = c(P = 2, T = 2, U = 1))
drawdag(dag1)
```



2. Rewrite the Oceanic tools model (from Chapter 11) below so that it assumes measurement error on the log population sizes of each society. You don't need to fit the model to data. Just modify the mathematical formula below.

$$
\begin{aligned}
T_i &\sim \text{Poisson}(\mu_i) \\
\text{log}(\mu_i) &= \alpha + \beta P_i \\
\alpha &\sim \text{Normal}(0, 1.5) \\
\beta &\sim \text{Normal}(0, 1)
\end{aligned}
$$

3. Rewrite the same model so that it allows imputation of missing values for log population. There aren't any missing values in the variable, but you can still write down a model formula that would imply imputation, if any values were missing.

4. What does it mean to say that in the Bayesian framework, the distinction between parameters and data is subtle? How does this idea relate to the mathematical specification of likelihoods and priors in a Bayesian model? You can refer to your answers for the previous two questions as context.

Project

Please answer the following questions. Note that your answers to these questions will transfer immediately to your presentation slides. But you will not submit these slides as part of this homework.

1. Give a **quick** overview of your awesome project.

2. Draw your DAG.

3. Write out your mathematical model (you can hand-write it and take a picture). Or you can try writing it in .Rmd. For example, the markdown syntax below will create the Poisson model above:

```
$$
\begin{aligned}
T_i &\sim \text{Poisson}(\mu_i) \\
\text{log}(\mu_i) &= \alpha + \beta P_i \\
\alpha &\sim \text{Normal}(0, 1.5) \\
\beta &\sim \text{Normal}(0, 1)
\end{aligned}
$$
```

   This website is helpful.

4. Write out the model in `rethinking` code.

5. Diagnose your chains - you don't need traceplots for every parameter, but you should discuss any issues.

6. **ONE** plot of the posterior and prior distribution (together) for one important parameter.

7. Caterpillar plot of the most important parameters (e.g., the plot you get from using `plot(precis)`)

8. **ONE** plot of posterior predictions that help answer your question.

9. What are your conclusions?

References

McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press.