Statistical Modeling, Winter 2025

Homework 6

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

I prefer a pdf, or a .Rmd / .Qmd / .R file. Please name the file with your last name and hw number (e.g., elahi_hw1).

Complete the following:

- read chapter 12.3 in *Statistical Rethinking* (McElreath 2020).
- watch Ordered Categories

Note for users that are knitting to pdf / html

Please use the following code in your `ulam` code chunks, as well as your code chunks that are loading libraries:

`{r, warning=FALSE, message=FALSE, results='hide'}`

https://rmarkdown.rstudio.com/lesson-3.html

Questions

From McElreath

1. The data contained in `library(MASS);data(eagles)` are records of salmon pirating attempts by Bald Eagles in Washington State. See `?eagles` for details. While one eagle feeds, sometime another will swoop in and try to steal the salmon from it. Call the feeding eagle the "victim" and the thief the "pirate." Use the available data to build a binomial GLM of successful pirating attempts.

   (a) Consider the following model:

$$
\begin{aligned}
y_i &\sim \text{Binomial}(n_i, p_i) \\
\text{logit}(p_i) &= \alpha + \beta_P P_i + \beta_V V_i + \beta_A A_i \\
\alpha &\sim \text{Normal}(0, 1.5) \\
\beta_P, \beta_V, \beta_A &\sim \text{Normal}(0, 0.5)
\end{aligned}
$$

   where y is the number of successful attempts, n is the total number of attempts, P is a dummy variable indicating whether or not the pirate had large body size, V is a dummy variable indicating whether or not the victim had large body size, and finally A is a dummy variable indicating whether or not the pirate was an adult. Fit the model above to the `eagles` data, using both `quap` and `ulam`. Is the quadratic approximation okay?

(b) Now interpret the estimates. If the quadratic approximation turned out okay, then it's okay to use the `quap` estimates. Otherwise stick to `ulam` estimates. Then plot the posterior predictions. Compute and display both (1) the predicted **probability** of success and its 89% interval for each row (i) in the data, as well as (2) the predicted success **count** and its 89% interval. What different information does each type of posterior prediction provide?

(c) Now try to improve the model. Consider an interaction between the pirate's size and age (immature or adult). Compare this model to the previous one, using WAIC. Interpret.

2. The data contained in `data(salamanders)` are counts of salamanders (*Plethodon elongatus*) from 47 different 49-m^2 plots in northern California. The column `SALAMAN` is the count in each plot, and the columns `PCTCOVER` and `FORESTAGE` are percent of ground cover and age of trees in the plot, respectively. You will model `SALAMAN` as a Poisson variable.

   (a) Model the relationship between density and percent cover, using a log-link (same as the example in the book and lecture). Use weakly informative priors of your choosing. Check the quadratic approximation again, by comparing `quap` to `ulam`. Then plot the expected counts and their 89% interval against percent cover. In which ways does the model do a good job? A bad job?

   (b) Can you improve the model by using the other predictor, `FORESTAGE`? Try any models you think useful. Can you explain why `FORESTAGE` helps or does not help with prediction?

Project

You are going to revisit your project, now that you have learned a little bit more. Some of these questions are repetitive from previous assignments. We are going to go simple now - choose just the single most important predictor for which you would like a causal estimate.

1. Plot the response (y) against the predictor (x). Include units on the axes.
2. Will you model y using a normal, binomial, Poisson, or multinomial distribution? Why? If you are using a normal, do you need to transform y? It is ok to proceed if your distribution is not perfect; let's see what happens.
3. Write out the mathematical expression for a linear model.
4. Translate the math into code. Plot predictive distributions (i.e., lines) from your priors. Justify your priors.
5. Run your model using `ulam`. Diagnose your chains.
6. Plot the data, along with the 89% interval for the average response, and the 89% interval for new predictions (see SR2 Fig. 4.10 for a Gaussian example). If you are using a binomial / Poisson interval please backtransform your model predictions to the original scale (i.e., probability, or counts).

Answers

References

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