This assignment contains five problems (P1-5).

**P1**. The data contained in *eagles.csv* are records of salmon pirating attempts by Bald Eagles in Washington State. While one eagle feeds, sometimes 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) Fit the following model using MCMC:

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

where y is the number of successful attempts, n is the total number of attempts, P is a dummy variable indicating whether the pirate had large body size, V is a dummy variable indicating whether the victim had large body size, and finally A is a dummy variable indicating whether or not the pirate was an adult.

(b) Now interpret the estimates. Then plot the posterior predictions. Compute and display both (1) the predicted probability of success and its 95% interval for each row (i) in the data, as well as (2) the predicted success count and its 95% 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.

**P2**. The data contained in *salamanders.csv* are counts of salamanders 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 in the lecture). Use weakly informative priors of your choosing. Then plot the expected counts and their 95% 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? Can you explain why FORESTAGE helps or does not help with prediction?

**P3.** The data in NWOGrants.csv are outcomes for scientific funding applications for the Netherlands Organization for Scientific Research (NWO) from 2010–2012 (see van der Lee and Ellemers (2015) for data and context). These data have a very similar

structure to the UCBAdmit data discussed in the lecture. The question is also similar: What are the total and indirect causal effects of gender on grant awards? Consider a mediation path (a pipe) through discipline. Draw the corresponding DAG and then use one or more binomial GLMs to answer the question. What is your causal interpretation?

If NWO's goal is to equalize rates of funding between men and women, what type of intervention would be most effective?

**P4**. Suppose that the NWO Grants sample has an unobserved confound that influences both choice of discipline and the probability of an award. One example of such a confound could be the career stage of each applicant. Suppose that in some disciplines, junior scholars apply for most of the grants. In other disciplines, scholars from all career stages compete. As a result, career stage influences discipline as well as the probability of being awarded a grant. Add these influences to your DAG from the previous problem.

What happens now when you condition on discipline? Does it provide an un-confounded estimate of the direct path from gender to an award? Why or why not? Justify your answer with the backdoor criterion.

If you have trouble thinking this though, try simulating fake data, assuming your DAG is true. Then analyze it using the model from the previous problem. What do you conclude? Is it possible for gender to have a real direct causal influence but for a regression conditioning on both gender and discipline to suggest zero influence?

**P5**. The data in Primates301.csv are 301 primate species and associated measures. In this problem, you will consider how brain size is associated with social learning. There are three parts.

(a) Model the number of observations of social_learning for each species as a function of the log brain size. Use a Poisson distribution for the social_learning outcome variable. Interpret the resulting posterior.

(b) Some species are studied much more than others. So, the number of reported instances of social_learning could be a product of research effort. Use the research_effort variable, specifically its logarithm, as an additional predictor variable. Interpret the coefficient for log research_effort. How does this model differ from the previous one?

(c) Draw a DAG to represent how you think the variables social_learning, brain, and research_effort interact. Justify the DAG with the measured associations in the two models above (and any other models you used).