Data Science 2 Final project instructions

For your final projects you will need to fit a generalized linear model in STAN using either pymc3 in Python or Rethinking’s ulam function in R. The primary goal of this model will either be inference (i.e. estimating the relationship or causal effect of a predictor variable on an outcome variable, controlling for confounds), or prediction. The final report will consist of:

**1. Introduction**

- Introduce the question you are asking, the background of the dataset, and the purpose of the analysis.

**2. Data**

- Explain in detail the dataset: where does it come from, what variables does it include?

- What is the outcome variable of interest and how is it measured? What is the primary predictor variable?

- Include some descriptive statistics or graphs. (mean and standard deviation of the outcome variable and main predictor, histograms/bar plots of the principal variables of interest).

**3. Analysis:**

- What probability distribution are you going to use to model your data? (Gaussian, Logistic, Poisson, Beta …). Justify your choice. What link function does this suggest?

- If you are estimating an effect, what variables might confound this effect? Do you have a way of controlling for them? (You could include a DAG of the assumed causal relationship between variables here).

- If you a predicting an outcome, what is your strategy for avoiding over-fitting?

- Do you think that there may be any non-linear relationships? How might you go about modeling these?

- Do you think that there could be any interactions?

- Are you using random intercepts to control for clustering?

- Center/scale/standardize your variables appropriately.

- What priors are you using? Are they incorporating any knowledge you have about the data? Are you regularizing to avoid overfitting? Sample from your priors to verify that there are no major issues.

- Please choose at least two equivalent models to compare (same dataset + same probability distribution, but with different parameters e.g. with or without interaction, with or without quadratic, with or without some variable in your dataset). Compare out-of-sample fit with WAIC or LOOCV and draw conclusions

**4. Results**

If focused on prediction: what is the accuracy and calibration of your final model? What are the most relevant predictors, and which predictors dropped out of the model after comparing out-of-sample predictive accuracy?

If focused on inference: what does your model say about the effect of the predictor variable (+ relevant interactions) on the outcome? Present a counterfactual prediction plot fixing all control variables at their mean, with a mean HDI and a prediction interval.

Present a forest plot of your parameter estimates and interpret the most relevant results.

**5. Discussion**

What conclusion do you draw from your model with respect to the question you set up in your introduction?