Third Homework Assignment

Bayesian Modeling

Due: Wednesday, April 30, 11:59pm

Context and assignment

In this assignment, you will work with data from the Florida voter record (`voter_data`). For each voter, identified by the voter ID (`voter`), the dataset includes some demographic information (`sex, age_group, race, county`), the voter's party preference (`party`), and whether the voter turned out to vote at a specific election, i.e., cast a vote in the election (`1`) or abstained (`0`, `NA` means not eligible to vote). Indicators for turnout are available for primary (`voted_pri_2006 to voted_pri_2018`) and general elections (`voted_gen_2006 to voted_gen_2018`) from 2006 to 2018. If you are not familiar with elections in the United States, the purpose of primary elections is to decide on a candidate for ones party for the general elections. Typically, primary elections are attended by more engaged voters, whereas general elections also attract the more occasional voter. Note further that elections in the years 2006, 2010, 2014, and 2018 were mid-term elections, which attract less voters than the high-stakes presidential elections in 2008, 2012, 2016.

You are going to rely on this dataset, specifically the observed individual participation in several elections of different types and years, to measure voters underlying propensity to exercise their voting rights, i.e., the individual voting propensity using a Bayesian hierarchical regression model. The required dataset including all variables is available on Moodle from the homework assignment page. Read all the tasks thoroughly before starting on the assignment.

Task 1 (3 points):

Given the (`voter_data`) and your goal to estimate individual voting propensities, decide on the appropriate reflective measurement model. Explain your choice of measurement model. Write the model down formally, fully specifying the likelihood and all priors. Interpret and explain the meaning of the core parameters of your model, i.e., those that make up the linear predictor, in the context of your measurement goal. Upload the answers to this task in a .pdf file.

Task 2 (4 points):

In R, import the `voter_data` dataset. Reshape the data to align with the requirements of the model you specified in task 1, then process the dataset in such a way that you can use it for full Bayesian inference with Stan. Next, specify the full Stan program that corresponds to your measurement model, store the Stan file on your computer and compile your Stan program using the cmdstanr R package. Given the Stan program and data you prepared, sample from the posterior. Estimates should be directly stored on your computer.

Task 3 (2 points):

Based on the posterior estimates for the parameters of your model, visualize the predicted probability of turnout/casting a vote along the full scale of potential values of voting propensities. Prepare the predictions for each indicator. You do not need to include estimation uncertainty in this visualization. Posterior means or medians suffice.

Task 4 (1 points):

You want to extend your measurement model by incorporating all the other variables that are available in `voter_data`, i.e., the voter demographics and party preferences. Prepare a graphical depiction of your measurement model that shows the relationship between the latent trait, the indicators, and the additional data that enters the model. A simple sketch on paper, scanned and uploaded as a .pdf file, suffices.

Optional bonus task (1 point):

Alter your Stan program from task 3 to incorporate the additional data mentioned in task 4. Note that this model will require additional identification constraints. There are different ways to implement the constraint. Check out the Stan case studies, the chapter in Gelman and Hill assigned for session 10, or the Stan forum for various ideas on how to solve this.

Submission:

Submit your R script, commented throughout, the .pdf files and all other files on Moodle by the deadline given above.