GR5065 Assignment 5

Due by 4PM EST on April 16, 2020

1 American Family Survey

Download the 2018 American Family Survey (once)

http://csed.byu.edu/wp-content/uploads/2019/10/Data-for-Release-2018.zip

to your working directory and then call

```
unzip("Data-for-Release-2018.zip")
library(haven)
AFS <- as_factor(read_dta(file.path("Data for Release 2018",
                                     "Data for Release 2018.DTA")))
```

The codebook is available from

http://csed.byu.edu/wp-content/uploads/2019/10/2018-Codebook.pdf

although you may need to recode some variables. The outcome variable is `app_dtrmp`, which measures job approval of President Trump. All of your models should be hierarchical and allow at least some of the parameters to vary by `inputstate`, which is the state that the person lives in.

1.1 Adjusting for Party Identification or Related Variables

One of the most controversial topics in polling is whether models should adjust for which political party (if any) the respondent is registered with, which can be changed fairly easily. Using `brms::brm`, estimate an ordinal model (using `family = cumulative`) of `app_dtrmp` that does not condition on `pid3`, `pid7`, `ideo5`, `presvote16post`, or similar variables. Then, use PSISLOOCV to decide which of the five models is expected to best predict future data as you add (perhaps some recoding of) `pid3`, `pid7`, `ideo5`, or `presvote16post` as a predictor. How much state-level heterogeneity in the data-generating process is there in the preferred model after conditioning on this extra variable? Finally, explain why ELPD is or is not a good criterion to decide which of the five model is best.

1.2 Binary or Ordinal

Many political scientists would choose to recode the outcome variable as binary with "Strongly / Somewhat approve" being 1 and "Strongly / Somewhat disapprove" or "Not sure" being 0 (or perhaps dropping the people who responded "Not sure").

```
AFS$app_dtrmp_binary <- AFS$app_dtrmp %in% c("Strongly approve", "Somewhat approve")
```

Take your preferred model from the previous subproblem and call `pp_expect` on it to return an array whose dimensions are simulations × observations × outcome categories and whose elements are probabilities that sum to 1 across the third dimension. If the result is assigned to `Pr`, you then call

```
ll <- dbinom(x = AFS$app_dtrmp_binary, size = 1, log = TRUE,
             prob = apply(Pr, MARGIN = 1:2, FUN = function(p) p[1] + p[2]))
```

you get a simulations × observations matrix whose cells are the log-likelihood of the ordinal model's parameters evaluated at the `AFS$app_dtrmp_binary`. You can then use the `loo` function in the loo package to estimate the ELPD.

Then, estimate the same model but use `app_dtrmp_binary` as the outcome variable and `family = bernoulli`. Compare the ELPD of that model to the ELPD you just calculated to decide if political scientists are justified in estimating the simpler Bernoulli model.

1.3 Predicting States

Suppose the people in the `AFS` dataset are representative of the voters in their state (which is in no way ensured by the research design). Call `posterior_predict` on the preferred model from the previous subproblem. Further suppose that everyone who the *model* posterior predicts would Strongly or Somewhat approve of Trump will vote for Trump over Joe Biden in the 2020 election, while all others will vote for Biden over Trump. For each state, what is the posterior probability that Trump wins a plurality of the voters? Explain why these state-level forecasts are likely to be better than using the *observed* data on approval of President Trump to forecast the state?

2 Discrimination in Police Stops

Read

https://projecteuclid.org/download/pdfview_1/euclid.aoas/1507168827

and download the north_carolina.rds file (once) from Canvas -> Files -> Assignments

to your working directory. Then, load it with

```
north_carolina <- readRDS("north_carolina.rds")
```

which contains the number of stops for each police department by the race of the driver.

2.1 Prior Predictive Distribution

Write a Stan function in a file called `NC_rng.stan` to draw *once* from the prior predictive distribution of both "searches" and "hits" by police department and race for the data-generating process described in section 2.3 and figure 2. Your function should start like

```
functions {
  int[ , , ] NC_rng(int D, int R, int[] Asian, int[] Black,
                    int[] Hispanic, int[] White) {
    int draws[D, R, 2] = rep_array(0, D, R, 2); // initialize with zeros
    // count stored as Departments by Race by {Searches, Hits}
    // fill in draws using many loops
    return draws;
  }
}
```

The function denoted logit^{-1} in the paper is called `inv_logit` in Stan, which is just the standard logistic CDF $\frac{1}{1+e^{-x}}$. In general, remember that you can utilize the `rstan::lookup` function to find the Stan function that corresponds to a R function or regular expression, like

```
head(rstan::lookup("_rng$")) # all Stan functions that end in _rng
```

```
##            StanFunction                         Arguments ReturnType
## 34 bernoulli_logit_rng                     (reals alpha)          R
## 37       bernoulli_rng                     (reals theta)          R
## 45   beta_binomial_rng (ints N, reals alpha, reals beta)          R
## 54 beta_proportion_rng           (reals mu, reals kappa)          R
## 56            beta_rng         (reals alpha, reals beta)          R
## 66        binomial_rng             (ints N, reals theta)          R
```

Note that the beta distribution in Stan (and R) is parameterized in terms of two shape parameters rather than an expectation (μ) and a "total count" (λ) but you can utilize the fact that

- $\alpha_{rd} = \mu_{rd}\lambda_{rd}$
- $\beta_{rd} = (1 - \mu_{rd})\,\lambda_{rd}$

to go from the parameterization in the paper to the parameterization in Stan. Also, note that although the individual outcomes are Bernoulli random variables, you should store the total of them (for each police deparment and race), which are binomial random variables. Finally, note that the `north_carolina` dataset is sorted such that the largest department is in the first row.

Then execute (once)

```r
rstan::expose_stan_functions("NC_rng.stan")
PPD <- NC_rng(D = nrow(north_carolina), R = ncol(north_carolina),
              Asian = north_carolina[ , 1], Black = north_carolina[ , 2],
              Hispanic = north_carolina[ , 3], White = north_carolina[ , 4])
```

to make sure that your function is working correctly.

2.2 Legal Analysis

The authors state that

> Having formally described our estimation strategy, we conclude by offering some additional intuition for our approach. Each race-department pair has three key parameters: the threshol t_{rd} and two parameters (ϕ_{rd} and λ_{rd}) that define the beta signal distribution. Our model is thus in total governed by $3DR$ terms. However, we only effectively observe $2DR$ outcomes, the search and hit rates for each race-department pair. We overcome this information deficit in two ways. First, we restrict the form of the signal distributions according to equations (1) and (2), representing the collection of DR signal distributions with $2(D + R - 1)$ parameters. With this restriction, the process is now fully specified by $2(D + R - 1) + DR$ total terms, which is fewer than the $2DR$ observations when $R \geq 3$ and $D \geq 5$. Second, we regularize the parameters via hierarchical priors, which lets us efficiently pool information across races and departments. In this way, we leverage heterogeneity across jurisdictions to simultaneously infer signal distributions and thresholds for all race-department pairs. (1201 – 1202)

Suppose that the posterior distribution were used as evidence for a legal claim of discrimination. The defense attorney argues that hierarchical models should not be admissible in court because they pool information across police departments. What are the pros and cons of the defense attorney's argument? Does it matter whether the state of North Carolina or an individual police department is the one defending against the legal claim?