rejected_ballots_master

```
library("tidyverse")
```

```
## -- Attaching packages ---------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

As the 2020 General Election approaches, some fear that the rejection of absentee ballots may determine the election, specifically in swing states. One quantity of interest is the expected number of rejected ballots as it determines the potential effect on the electoral outcome.

The naive approach involves estimating the product of the number of voters (N), the probability of requesting an absentee ballot (θ), the probability of submitting a received absentee ballot (ψ), and the probability of having one's absentee ballot rejected (κ).

$$p(\text{rejected ballots}|N, \theta, \psi, \kappa) \sim \sum_i p_i(N) * p_i(\theta) * p(\psi|\theta) * p(\kappa|\theta, \psi)$$

Together with an assumption about the expected number of absentee voters, we can calculate the expected number of rejected votes given additional assumptions about turnout, the probability to submit an absentee vote, and to have one's absentee ballot rejected.

Naturally, in the US context, missing data represents a problem such that we would make additional assumptions about the distribution of the missing data.

Motivation

The approach above is naive. It does not consider varying rates of absentee ballot rejections among absentee voters. While rates vary, we know that absentee voters are primarily older and white. If the rate of absentee voters increases, the share of other socio-demographic groups will rise. If their rejection rates differ — being either larger or smaller — then the average rate at which ballots are being rejected will change as well.

Model

While data on the number of rejected ballots exists at the county level, we do not know the personal characteristics of those whose absentee ballots are being rejected. That is, we face an ecological inference problem, wanting to make inference over the behavior of the individual while only having aggregate data. In regression terms we want to estimate

$$y_{ij} = \alpha_j + \beta_j x_{ij} + \epsilon_{ij}$$

but can only estimate

$$\bar{y}_j = \alpha_j + \beta_j \bar{x}_j$$

where x_{ij} is the characteristic of individual i in the group j we care about and $\bar{x}_j$ the averages we have access to.

For binary data — following Gelman et al. (2001) — we can re-express this as

$$\bar{y}_j = \beta_{j1} \bar{x}_j + \beta_{j2}(1 - \bar{x}_j)$$

and fit

$$\bar{y}_j = \beta_1 \bar{x}_j + \beta_2(1 - \bar{x}_j) + \eta_j$$

Fake data and Stan model

To start I simulate some fake data and fit the model using Stan. I start with the individuals ...

```
library(DirichletReg)
```

```
## Loading required package: Formula
```

```
library(boot)
N <- 1e4
J <- 1e2
G <- 4
pr_g <- rdirichlet(1, rep(10, G))
pr_j <- rdirichlet(1, rep(10, J))
g <- sample(seq(1, G), N, replace = TRUE, prob = pr_g)
j <- sample(seq(1, J), N, replace = TRUE, prob = pr_j)
beta_j <- rnorm(J, mean = 0, sd = 0.4)
epsilon_ij <- rnorm(N, 0, 0.2)
y_star <- inv.logit(beta_j[j] * g + epsilon_ij)
y <- as.integer(y_star > 0.5)
```

... and then aggregate.

```
df <- data.frame(y = y, j = j, g = g)
df_bar <- df %>%
  mutate(g1 = ifelse(g == 1, 1, 0),
         g2 = ifelse(g == 2, 1, 0),
         g3 = ifelse(g == 3, 1, 0),
         g4 = ifelse(g == 4, 1, 0)
         ) %>%
  group_by(j) %>%
  summarize(y_bar = mean(y),
            g1_bar = mean(g1),
            g2_bar = mean(g2),
            g3_bar = mean(g3),
            g4_bar = mean(g4))
```