Bayesian Computing **Sandra Valdés Salas**

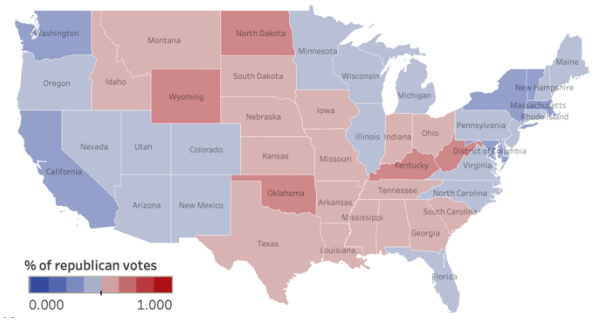
June 25, 2020

A hierarchical Bayesian model to determine the impact of household income at the state-level in the U.S. presidential elections in 2016

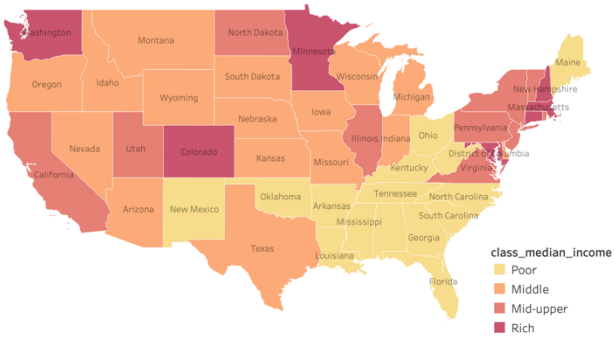
Introduction

Economic polarization is central for the United States presidential elections. According to political scientists, richer people are more likely to vote for Republican candidates while poorer people are more inclined to vote for Democratic candidates. However, in recent decades, "poorer states have voted for conservative Republicans while rich states favor liberal Democrats" (Gelman 2014). As the presidential election in 2016 showed, those states with a majority of Republican votes (map 1) turned out to be those states with the lowest median household income (map 2).

Map 1. Republican votes as a percentage of total votes by state (2016)



Map 2. Mean household income in the U.S. by state (2016)



Sources: Bureau of Economic Analysis, U.S. Department of Commerce, available at: https://apps.bea.gov/regional/downloadzip.cfm; MIT Election Data and Science Lab, 2017, "U.S. President 1976–2016", available at: https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/42MVDX/MFU99O&version=5.0

This paper will address the question if votes varied significantly across the 50 states based on the median household income during the U.S. presidential elections of 2016. A hierarchical Bayesian model is used because, as Kruschke (2015, p. 244) stated, "it is a reasonable way to capture individual differences and group-level tendencies". This project is inspired in the paper "How Bayesian Analysis Cracked the Red-State, Blue-State Problem" by Gelman (2014).

Model

The data consists of 50 rows, each with information about the region, the total of republican votes, the total of democratic votes, the total of other votes, the total of votes, the percentage of votes that were for the Republican party, and the median household income. The income was divided into the following categories: poor (less than 50,000 USD), middle (from 50,000 and 60,000 USD), mid-upper (from 60,000 to 70,000 USD), and rich (more than 70,000 USD).

For the purposes of this paper, θ (theta) is defined as the the number of votes in a state. The outcome of each vote follows a Bernoulli distribution (see equation 1), where y = 1 refers to the number of votes for the Republican candidate (Donald Trump) and y = 0 refers to the number of votes for the Democrat candidate (Hillary Clinton).

$$y_i \sim \text{dbern}(\theta) \tag{1}$$

Particularly, θ is distributed as a beta distribution, with parameters ω (omega) and κ (kappa):

$$\theta \sim \text{dbeta}\left(\omega(\kappa-2)+1,\ (1-\omega)(\kappa-2)+1\right) \tag{2}$$

The states are grouped by different categories according to the median household income in 2016. Each category is assumed to be distributed as a beta density with mode ωc and concentration κc .

$$\omega \sim dbeta(1,1) \tag{3}$$

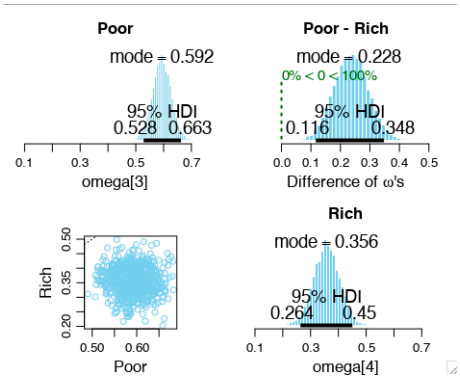
$$\kappa - 2 \sim dgamma(0.01, 0.01) \tag{4}$$

The model also considers an over-arching distribution on the category concentrations κc that has its central tendency and scale estimated. In other words, the model puts each state's estimated probability of voting Republican under a distribution for the income category and puts those distributions under an overarching distribution across all income categories.

The model implemented Markov Chain Monte Carlo (MCMC) to generate a sequence of samples from the posterior distribution of the parameters. The model was run in R, using JAGS package. The code implemented for this model was based on the codes provided by Kruschke (2015, p. 251-253), with minor modifications (see Annex).

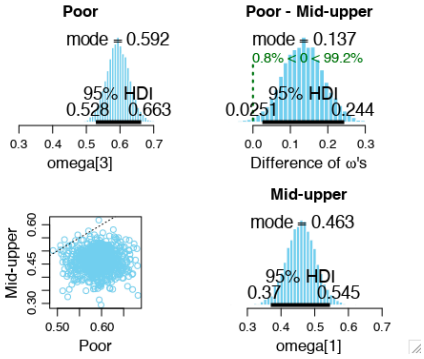
Results

Figure 1. Marginal posterior distribution of Poor – Rich categories



The marginal posterior on the difference of "Poor" and "Rich" categories (upper right graph in Figure 1) indicated that the most credible difference (mode) is larger than zero (0.228 approximately), suggesting that, in effect, states under the "Poor" category voted for the Republican presidential candidate. Also, it is worth noting that the 95% HDI of the marginal posterior difference of "Poor" and "Rich" categories has a value different from zero. This suggests that the difference is statistically significant and not due to randomness.

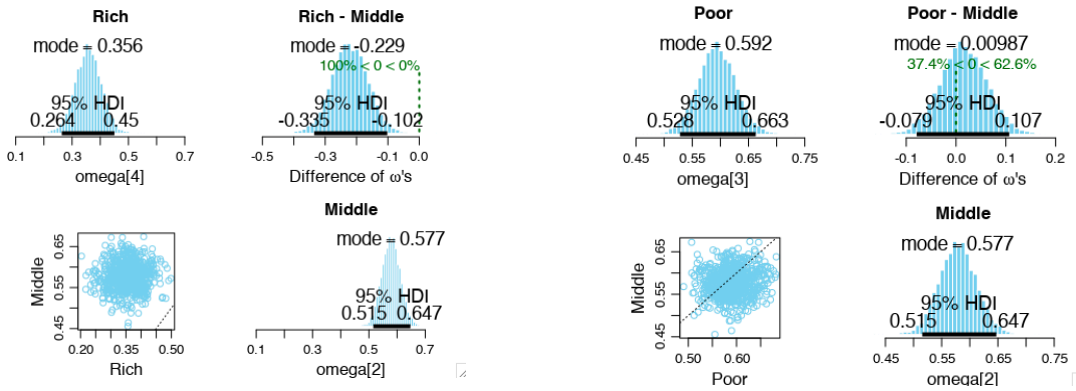
Figure 2. Marginal posterior distribution of Poor – Mid-upper categories



If we take a look at the marginal posterior distribution between "Poor" and "Mid-upper" categories, we can see a similar trend as the above. Figure 2 (upper right graph) indicates that the most credible difference is larger than zero (0.137 approximately) and, therefore, "Poor" states (with respect to "Mid-upper" states) voted for the Republican presidential candidate. This result is statistically significant because the 95% HDI yields a result different from zero.

Finally, by taking a look at trends between "Poor" and "Rich" states with respect to states classified as "Middle" income, there are contrasting results. The posterior distribution between "Rich" and "Middle" states shows a clear trend: Middle states voted for the Republican candidate (left side in Figure 3). However, the right side of Figure 3 shows a posterior distribution between "Poor" and "Middle" states are inconclusive: although the mode is slightly larger than zero, the 95% HDI contains the value zero, suggesting that the difference between these categories can be due to chance.

Figure 3. Marginal posterior distributions for voting data across Poor-Middle and Rich-Middle states



Further research

For making the model simple, the median household income was only considered in this project. However, other predictors are worth considering, such as ethnicity, age and college degree. Therefore, a multiple linear regression will be considered as further research.

References

Bureau of Economic Analysis. *U.S. Department of Commerce*. Retrieved from https://apps.bea.gov/regional/downloadzip.cfm

Gelman, A. (2014). How Bayesian Analysis Cracked the Red-State, Blue-State Problem. *Statistical Science, 29(1)*, 26-35. Retrieved from https://stat.columbia.edu/~gelman/research/published/bayesrb4.pdf

Kruschke, John K. (2015). *Doing Bayesian Data Analysis: A tutorial with R, JAGS, and Stan* (2nd ed.) London: Elsevier.

MIT Election Data and Science Lab. (2017). *U.S. President 1976–2016*. Retrieved from https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/42MVDX/MFU99O&version=5.0

National Election Pool (ABC News, Associated Press, CBS, CNN, Fox News, NBC). *National Election Pool Poll: 2016 National Election Day Exit Poll, 2016* [Dataset]. Roper #31116396, Version 3. Edison Research [producer]. Cornell University, Ithaca, NY: Roper Center for Public Opinion Research [distributor]. Retrieved from https://ropercenter-cornell-edu.proxygw.wrlc.org/ipoll/study/31116396

Annex

```
model {

    for ( i in 1:Nsubj ) {
      z[i] ~ dbin( theta[i] , N[i] )
      theta[i] ~ dbeta( omega[c[i]]*(kappa[c[i]]-2)+1 , (1-
omega[c[i]])*(kappa[c[i]]-2)+1 )
    }
    for ( c in 1:Ncat ) {
      omega[c] ~ dbeta( omegaO*(kappaO-2)+1 , (1-omegaO)*(kappaO-2)+1 )
      kappa[c] <- kappaMinusTwo[c] + 2
      kappaMinusTwo[c] ~ dgamma( kappaS , kappaR )
    }
    # Over-arching distribution of omega
    omegaO ~ dbeta( 1.0 , 1.0 )
    kappaO <- kappaMinusTwoO + 2
    kappaMinusTwoO ~ dgamma( 0.01 , 0.01 )  # mean=1 , sd=10 (generic
vague)

    # over-arching distribution of kappa
    kappaS <- kappaMinusTwoS + 2
    kappaMinusTwoS ~ dgamma( 0.01 , 0.01 ) # mean=1 , sd=10 (generic
vague)

    kappaR <- kappaMinusTwoR + 2
    kappaMinusTwoR ~ dgamma( 0.01 , 0.01 ) # mean=1 , sd=10 (generic
vague)
  }
```