A Bayesian Hierarchical Model for US Election Data

Benjamin Osafo Agyare, Connor Dayton, Jaucelyn Canfield

Abstract

We compare federal election results for each state versus the USA in every second year from 1992 to 2016 to model partisan lean of each state and its dependence on the nationwide popular vote. For each state, we model both its current partisan lean and its rate of change, as well as sensitivity of state results with respect to the nationwide popular vote, using a Bayesian Hierarchical Model. We then apply this model to predict and compare results with the actual values for the 2018 election

Background and Data

Our data is taken from the Federal Election Commission and the House Clerk web page for the year intervals 2000-2016 and 1992-1998, respectively. These make up 14 election events; each containing vote counts for each State for Democrats and Republicans spanning at least two of Senate, House and Presidential elections. For the purpose of this study, the data is transformed into variables as:

$$x_t = ln\frac{d(e)}{r(e)} \quad \text{and } y_{st} = average(ln\frac{d_s(e)}{r_s(e)})$$

for election events, e, and election year, $t \in \{-14, -13, \ldots, -2, -1\}$ corresponding to 1992, 1994, . . . , 2018.

Exploratory Data Analysis (EDA)

The plots of y_{st} against t are shown below. It can be observed that North Dakota exhibits an outlier at data point t = − 11 (1996). However, this data is still included as is in the model. It is also not a surprise that all data points are clustered around the line y = 0 since log as a function shrinks the magnitude of quantities

Y_{st} vs t for all 50 States

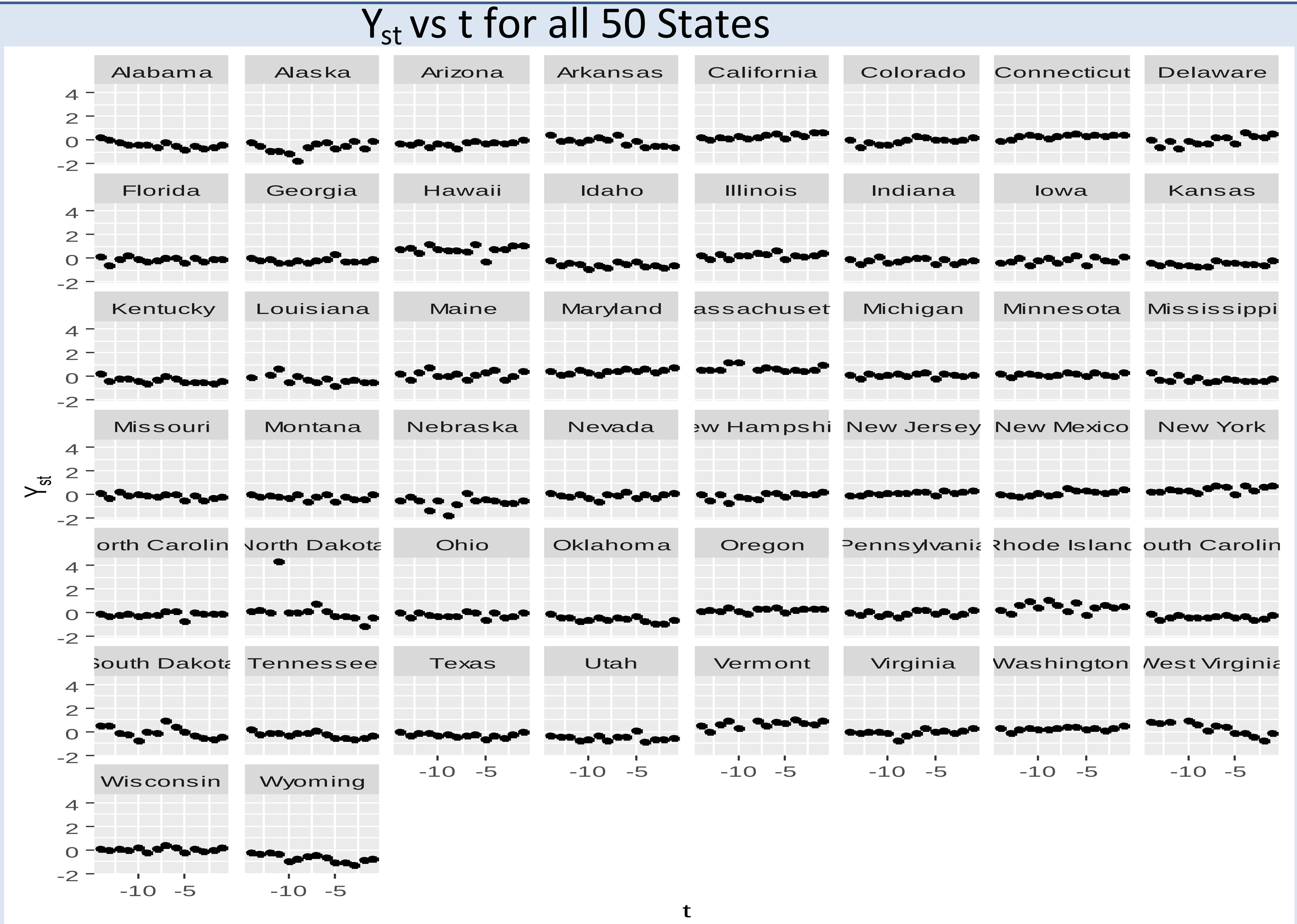

Fig. 1: Scatterplot of Y_{st} vs t for all 50 state (from 1992 to 2018)

Model Specification

We fit a random-intercept, random slope Bayesian Hierarchical Model (BHM) as follows:

$$\begin{aligned} Y_{st} &\sim \mathcal{N}(\mu, \sigma) \\ \mu &= \alpha + \alpha_s + (\beta + \beta_s)X_t + (\gamma + \gamma_s)t \\ \alpha &\sim \mathcal{N}(0,5) \\ \alpha_s &\sim \mathcal{N}(0,5) \\ \beta, \gamma &\sim \mathcal{N}(0,10) \\ \beta_s, \gamma_s &\sim \mathcal{N}(0,10) \end{aligned}$$

where α and α_s represent the current national and state partisan lean, β and β_s are the national and state elasticity: its responsiveness to changes in the national environment (measured by x_t), and γ and γ_s are the national and state partisan lean rate of change. National parameters are modeled as fixed effects and state as the random effects in the Hierarchical model. *Note: distributions are priors.*

Parameter Estimates

Ordered Caterpillar Plot for the States' Partisan Lean Estimates

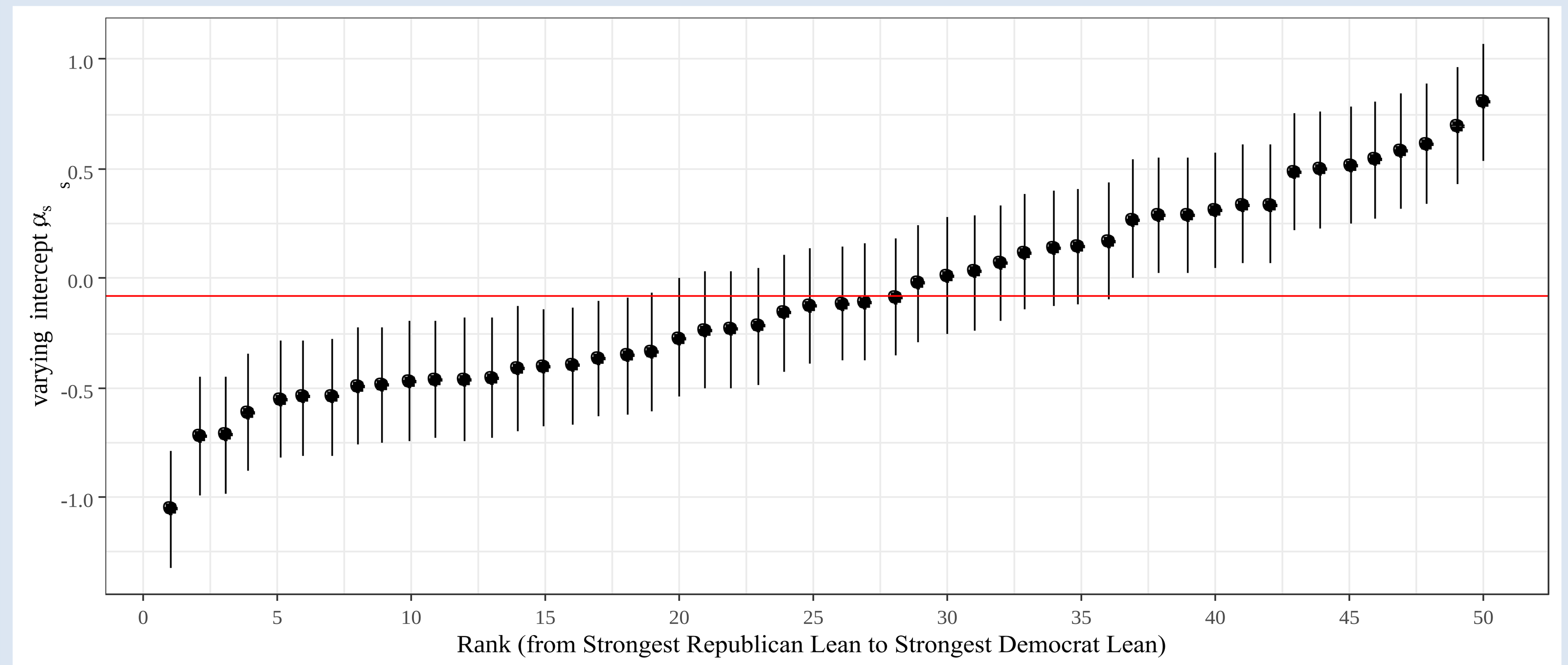

Fig. 2: Caterpillar Plot of α_s

Interpretation

The posterior estimates of the random effects parameters indicate that:

- The reddest state (rank 1 above), measuring by α_s, is Wyoming, α_s = -0.969
- The bluest state (rank 50) is Vermont, α_s = 0.889
- The most neutral state (with α_s closest to 0) is Nevada, α_s = -0.006
- The most rapidly bluing state is Delaware, γ_s = 0.052
- The most rapidly reddening state is North Dakota, γ_s = -0.089
- The state with the smallest rate of change in partisanship is Maine, γ_s = 0.00068
- The state which is most sensitive to the national environment is North Dakota, β_s = 0.049
- The state which is the least sensitive is New Jersey, β_s = 0.00011
- The state with the most negative sensitivity is Alaska, β_s = -0.026

Model Diagnostics

The first diagnostic plot **(Fig 3)** shows the posterior simulation of the distribution of the draws of about 20000 samples. Since the distribution of samples **(y_{rep})** approximates that of the independent variable **y**, it can be said that the model performs very well. It can be observed from the posterior predictive *mean vs sd* plot **(Fig 4)** that the mean-sd pair lies right at the center of all the posterior samples. This also indicates that the model performs very well. From **(Fig 5)** below, it can be observed that the chains are very consistent since about 85% of the chains are well below 1.01 (consistency means $\hat{R} \leq 1.1$) and this clearly shows that the chains are performing well. Finally, the ratio of effective sample size to the total sample size tells how fast and adaptive the chains are. Per literature, if the ratio is less than 0.1%, then there is an indication that the chains are exploring slowly. From **(Fig 6)**, all ratios are well above 0.1%, indicating that the chains are performing well and sufficiently fast, hence the model is good for subsequent analysis.

MCMC Diagnostic Plots

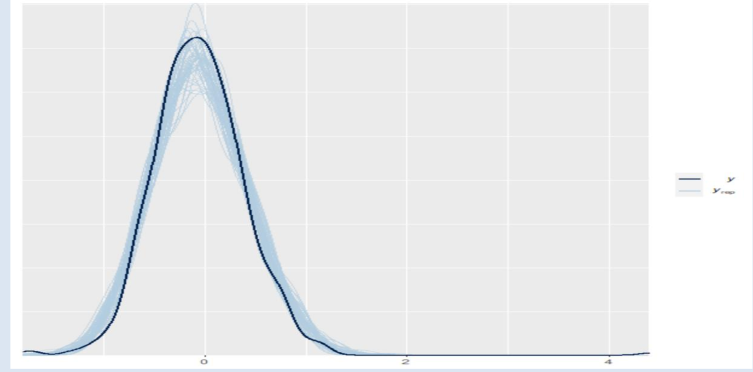

Fig. 3: Posterior Predictive Density Plot

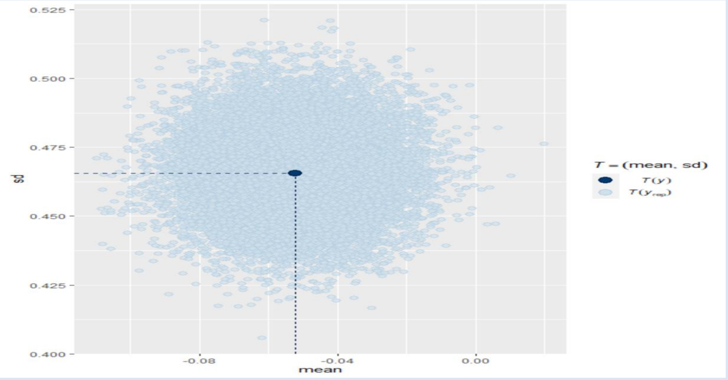

Fig. 4 Posterior Predictive Mean - SD Scatterplot

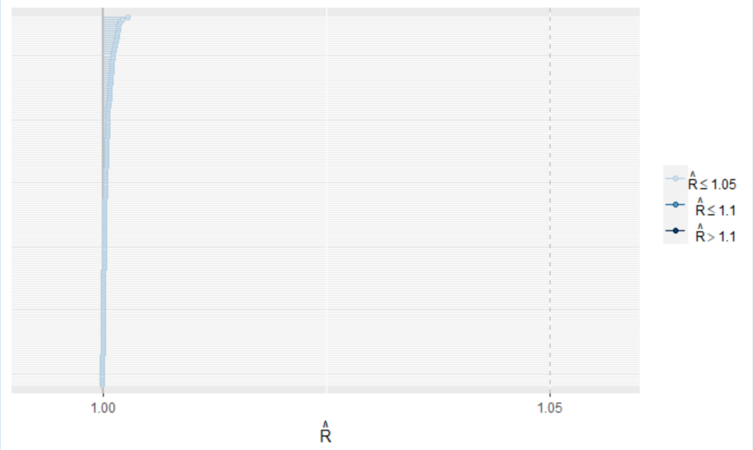

Fig. 5: $\hat{R}$ plot across parameters

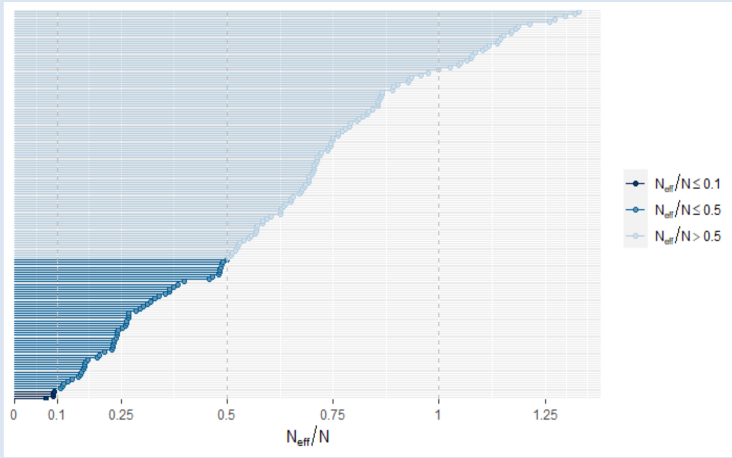

Fig. 6: A Histogram of the ratio of effective sample size to total sample size

Predictions and Results

We show the output of the posterior prediction of the elections for the test data (2018) for the first 5 states. This output includes the means, medians and 90% Bayesian credible intervals, as well as the actual values.

State	Actual	Post. Median	Post. Mean	90% Credible Interval	
				Lower	Upper
Alabama	-0.3630	-0.5189	-0.5198	-1.0569	0.0112
Alaska	-0.1324	0.3416	-0.3409	-0.8673	0.1883
Arizona	0.0413	-0.0799	-0.0787	-0.6008	0.4453
Arkansas	-0.5752	-0.2067	-0.2082	-0.7386	0.3131
California	0.6746	0.6295	0.2671	0.1073	1.1534

Visualization of Posterior Predictions

Fig 7 below gives a visualization of the posterior with the Bayesian credible intervals for all 50 states. The actual values are colored in **blue** on the caterpillar plot below. It can be observed that the actual value for each state lies within its respective credible interval.

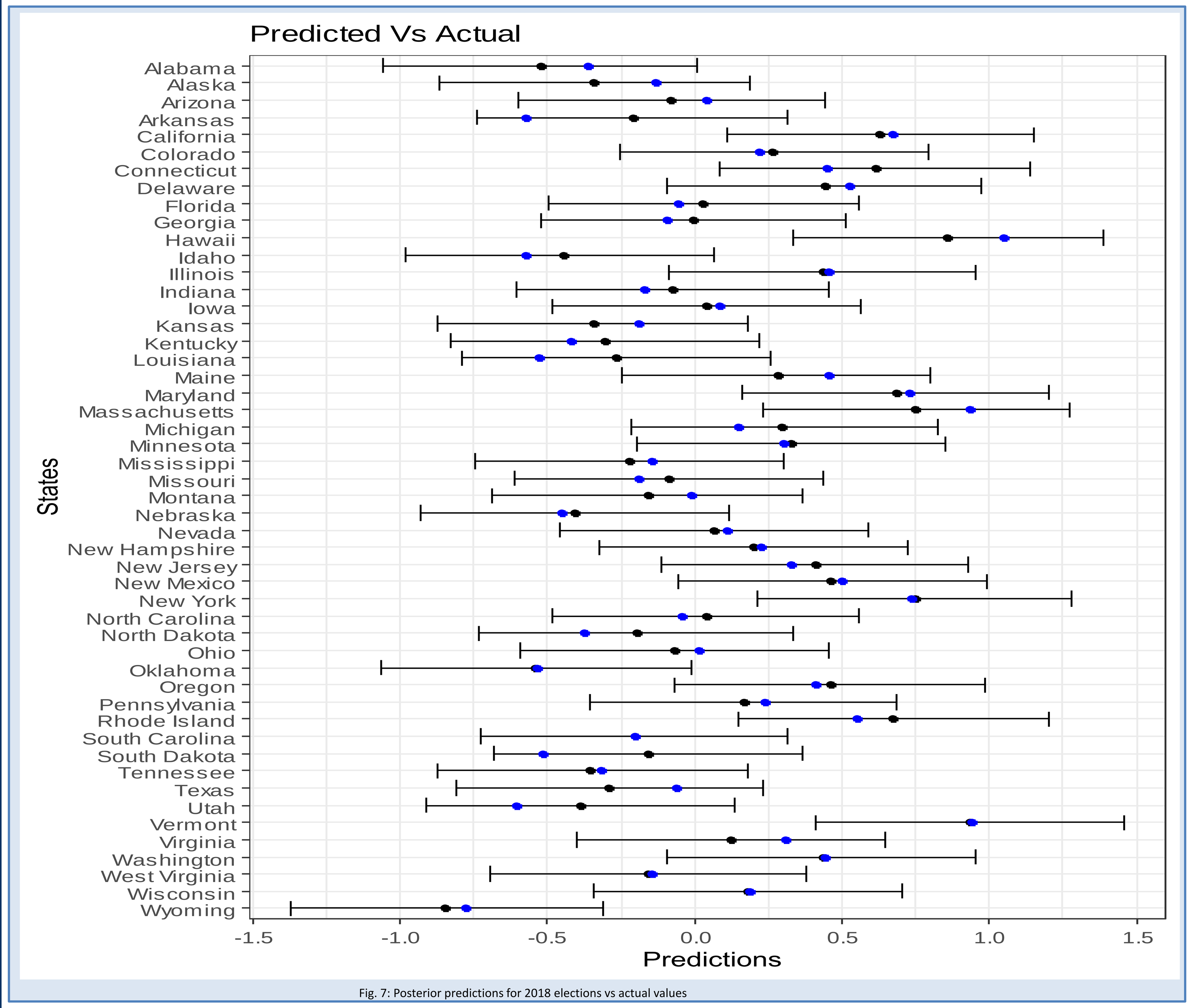

Fig. 7: Posterior predictions for 2018 elections vs actual values

Discussion

As observed from the caterpillar plot above, the model fits the data credibly. Other competing models like the classical Multilevel modeling could also be fitted to the data to assess the quality of its fit. However, we leave that for future research work. This model can also be used to determine if the Electoral College is biased towards one of the major parties. This may be carried out by (for instance) simulating the Electoral College outcome in 2020, given even (equal) nationwide popular vote, as well as the actual 2016, 2008, and 2004 nationwide popular vote. Once again, we leave this for future studies.

References

- https://mc-stan.org/users/documentation/case-studies/tutorial_rstanarm.html#bayesian-inference-for-model-1
- https://mc-stan.org/rstanarm/articles/rstanarm.html#step-2-draw-from-the-posterior-distribution
- https://transition.fec.gov/pubrec/electionresults.shtml
- Andrew Gelman and Jennifer Hill (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge