Bayesian Hierarchical Models for the Prediction of American Elections

Brittany Alexander

Previous Research

- Alexander (2018) and Alexander & Ellingson (2019) developed multiple conjugate prior based models to predict the 2008, 2012, and 2016 US Presidential elections
- Both Gaussian and Beta-Binomial models were considered, and the Gaussian models performed better
- These models beat averaging the polls and a non-informative Beta-Binomial model in terms of average error
- For simplicity the models assume known variance but this proved to vastly underestimate uncertainty

Goal of This Project

- The goal of this project was to build upon previous models and find a single ideal model for 2020 prediction
- Both precision in predicting the vote share and realistic uncertainty quantification are valued
- The goal was also to determine how the models perform under different data inclusion criteria

Data Source

Huffington Post Pollster has polling data for 2012-2016. There was a site to get 2008 data before Pollster merged with Huffington Post, but that link is broken. There are 5756 state level polls across the three elections. Most states have multiple polls for each election year. Only polls of Likely Voters and Registered Voters were included in the models.

Data Criteria

- All three models were fit over a grid of possible data inclusion criteria.
- The grid considered all combinations of 5-100 days (in 5 day intervals) before the election and the most recent 5-15 polls
- Limiting the data was found to be beneficial
- The best MSE was at the last 5 polls and the last 25 or 30 days prior to the election
- MSE of restricted models were less than the MSE of non-restricted models
- 30 days was chosen because it is approximately a month
- Washington DC is excluded from average error estimates because it is an outlier and has limited polling data

Normalization

- Poll data contains undecided and occasionally minor candidate as options
- Undecided is not a ballot option
- Polling for minor candidates is inaccurate and inconsistent
- The minor candidates vary greatly from election to election
- To make things simpler the poll and election results were proportionally normalized so that the republican and democratic support summed to one.

Prior Specification

- States are grouped into clusters based on the average margin (democratic – republican) from the past four presidential elections.
- The mean and variance of polls inside a cluster is used for the prior mean and variance
- The classification is based on cutoffs: <-.2 (Strong Red), -.2<m<-.1 (Red), -.1<m<-0.025 (Lean Red), -0.025<m<0.025 (competitive), 0.025<m<.1 (Lean Blue), 0.1<m<0.2 (Blue), >.2 (Strong Blue)
- K means and mixture models were tried but they put every state in the same group and performed worse

Models

Model 1: Gaussian Conjugate Prior assuming unknown mean known variance

Model 2: Gaussian Conjugate Prior assuming unknown mean unknown variance (non-informative inverse gamma prior on variance)

Model 3: Gaussian Conjugate Prior assuming unknown mean known variance with logit transformation (non-informative inverse gamma prior on variance)

Average Error

	Model 1 (no prior on variance)	Model 2 (inverse gamma prior)	Model 3 (inverse gamma prior with logit transformation)	Average of Last 5 polls
2008 All States	0.0200	0.0215	0.0216	0.0343
2012 All States	0.0242	0.0246	0.0246	0.0164
2016 All States	0.0279	0.0286	0.0287	0.0285
Average All States	0.0241	0.0249	0.0249	0.0264
2008 Competitive States	0.0129	0.0155	0.0157	0.0214
2012 Competitive States	0.007	0.0085	0.0089	0.0294
2016 Competitive States	0.0210	0.0214	0.0213	0.0194
Average Competitive States	0.0136	0.0152	0.0153	0.0234

Uncertainty

- How often do 95% credible intervals contain the actual election result for the three models?

	Model 1 (no prior on variance)	Model 2 (inverse gamma prior)	Model 3 (inverse gamma prior with logit transformation)
2008	.529	1	.960
2012	.608	.980	.921
2016	.353	.843	.843
Overall	.497	.941	.908

Conclusion

- Model 2 had the highest coverage credible intervals
- Model 2 had the second highest average error but was within 0.001 of the other models
- All three tested models take a trivial amount of time to run on a laptop computer.
- The models improve upon a average of polls while not increasing model complexity much.

Future Research for 2020

- Prior that is a mixture distribution of polls from other states and historical election trends
- New methods to normalize the vote including a mixture of proportional normalization with normalization based on last's elections vote
- A model that incorporates the correlation between states in the uncertainty estimates

References

Alexander, B. (2019), A Bayesian Model for the Prediction of United States Presidential Elections, SIAM Undergraduate Research Online, 12.

Alexander, B. and L. Ellingson. 2019. Poll-Based Conjugate Prior Models for the Prediction United States Presidential Elections. In *JSM Proceedings*, Scientific and Public Affairs Advisory Committee. Alexandria, VA: American Statistical Association. 112-131.