Methodology

October 2022

About

The ORACLE (Overall Results of an Analytical Consideration of the Looming Elections) of Blair is an election model developed entirely by senior students at Montgomery Blair High School in Silver Spring, Maryland. Under the supervision of statistics teacher David Stein, we created this model during the fall semester to predict the outcomes of the upcoming 2022 Senate and Gubernatorial elections. This is the fourth iteration of election modeling at Blair; previous classes have also developed the Oracle to forecast the 2016 presidential election, 2018 congressional elections, and 2020 presidential election.

In the spirit of transparency and education, we describe in detail exactly how we came up with all of the numbers in our simulation. You can read about our reasoning and methods for constructing the model in the following sections. The code that implements our model is stored on Github. All of the decisions in creating this model were made by the students in the class, and we take full responsibility for this model's methods and predictions. If you are interested in politics, statistics, education, or our model, please consider spreading the word about the work that we've done.

Overview

All of our calculations in this model are based on the two-party vote percentage, which represents the ratio of votes cast for the Democratic candidate to the total votes cast for either the Republican candidate or the Democratic candidate. The two-party vote percentage differs from the actual vote percentage as votes cast for a third-party or an independent candidate are not counted. Positive margins favor the Democratic Party, while negative margins favor the Republican Party.

All of the polls that we use in this model are taken from FiveThirtyEight, a website owned by ABC News that provides data-driven political news and analysis. It was created in 2008 as a polling aggregation website and blog by analyst Nate Silver.

Priors

Blair Partisan Index (BPI)

A state's historic voting tendencies can give us important insight into their future voting behavior. The Blair Partisan Index (BPI) is a metric we use to quantify how a state has voted in past elections. We calculate BPI by taking the weighted average of the two-party vote percentage earned by the Democratic candidate in the following elections, according to these respective weights:

Election	Weight
2014 Senate	0.05
2014 Governor	0.04
2016 Presidential	0.042
2016 House	0.035
2016 Senate	0.075
2016 Governor	0.06
2014 House	0.0435
2018 Senate	0.10
2018 Governor	0.10
2020 Presidential	0.15
2020 House	0.0945
2020 Senate	0.11
2020 Governor	0.10

For a given election, there is a Democratic two-party vote percentage D and weight w. Each weight was determined by the class based on how influential we thought each election was for this year's cycle. The BPI is calculated by taking the weighted average of these previous elections.

$$\text{BPI} = \frac{\sum D_i \cdot w_i}{\sum w_i}$$

The BPI serves as a starting point in our model for estimating the two-party vote percentage that the Democratic candidate will receive.

National Mood (Bigmood)

We incorporated a national mood shift (Bigmood) into our predictions for the senate races. We decided that since gubernatorial races dealt with issues on a more local level, the national mood did not affect voters'

opinions of gubernatorial candidates. The national mood refers to the general feelings of the American people toward the country's issues and our policymakers' actions. We evaluate national mood through generic ballot polls, which ask people which party (Democratic or Republican) they would support in the election if it were held today.

To calculate the national mood, we take the weighted average of generic ballot polls that have earned at least a C- grade or higher on FiveThirtyEight. We weighted polls based on how long ago they were conducted so that more recent polls are weighted more heavily in our model. The weight w for each poll is

$$w = e^{-0.05d}$$

where d represents the number of days since the poll was conducted. This allows more recent polls to have an exponentially greater effect on national mood than earlier polls. Bigmood is calculated by taking the weighted average of the Democratic two-party vote percentage D for each poll.

$$\text{Bigmood} = \frac{\sum D_i \cdot w_i}{\sum w_i}$$

We also calculate a variance for Bigmood, which we will use in our simulation as explained in a later section. If the sampling variance of a poll σ^2 is defined by

$$\sigma^2 = \frac{D(1-D)}{n}$$

where n is the sample size of the poll, then the variance of Bigmood $\sigma^2_{\text{Bigmood}}$ is the weighted average of each of the sampling variances for each poll using the same weight in our weighted average of Bigmood above.

$$\sigma^2_{\text{Bigmood}} = \frac{\sum \sigma_i^2 \cdot w_i}{\sum w_i}$$

BPI and Bigmood On Our Network (BABOON)

BABOON (BPI and Bigmood On Our Network) is our method for combining our BPI value with the Bigmood into a single estimate for the prior two-party vote percentage for senate races. For gubernatorial races, we simply use BPI for BABOON. For each iteration of our simulation, we generate a random value from the normal distribution $X \sim \mathcal{N}(\text{Bigmood}, \sigma^2_{\text{Bigmood}})$. We use X to calculate BABOON for each iteration by multiplying the difference between X and the event in which the Democratic and Republican candidates receive the same number of votes by 0.15 and adding it to the BPI.

$$\text{BABOON} = \text{BPI} + 0.15(X - 0.5)$$

Averaging Polls

Polling data is a valuable predictor of people's future voting behavior. In our model, we only include polls that earned at least a C- grade on FiveThirtyEight and have no more than 1 Democratic and 1 Republican candidate with the exception of Alaska. In the case of Alaska we categorized any candidate other than the top Democratic and Republican candidates as third-party candidates. Similar to Bigmood, we weight the polls based on how long ago they were conducted so that more recent polls are weighted more heavily in our model. The weight w for each poll is

$$w = e^{-0.05d}$$

where d represents the number of days since the poll was conducted. The average of the polls is determined by taking the weighted average of the Democratic two-party vote percentage D for each poll.

$$\mu = \frac{\sum D_i \cdot w_i}{\sum w_i}$$

We also calculate the variance for the average of the polls in the same way we calculated it for Bigmood. Using the same weights as above, the average variance σ^2_{polls} is

$$\sigma^2_{\text{polls}} = \frac{\sum \sigma_i^2 \cdot w_i}{\sum w_i}$$

where σ^2 is the sampling variance of each poll.

Lean

To combine our polling average with BABOON into a single estimate for each race, we take a weighted average of the two values. In a world where we are given an infinite number of polls for a race, the vast majority of our estimate should come from the polling averages. Thus we chose the arctangent function to calculate the weight w for the polling average so that as the number of polls approaches infinity, the polling average comprises 95% of our estimate. The weight is calculated by

$$w = \frac{1.9}{\pi} \arctan(1.75n_{30} + 0.05n)$$

where n_{30} and n are the number of polls for that race in the last thirty days and in total respectively. The state lean is calculated by weighting the polling average in this manner.

$$\text{Lean} = w \cdot \text{polls} + (1 - w) \cdot \text{BABOON}$$

Variance

There are many sources of uncertainty in our model. To calculate the overall variance, we begin by finding the weighted sampling variance of the polls and then adding additional variance based on two factors: the number of undecided voters for each race (VIBE) and how wrong a state's polling predictions have been historically (GOOFI).

Variance of Indecisive Ballot Electors (VIBE)

The Variance of Indecisive Ballot Electors (VIBE) is a metric used to add uncertainty to our model based on how many undecided voters there are in each race according to the polls. We define the percentage of undecided voters U as anyone who reported not voting for the top Republican and Democratic candidates, $1 - (D_{\text{actual}} + R_{\text{polls}})$. Note that D_{polls} is not the same as D used in earlier calculations because D is the proportion of voters who voted for the top Democratic candidate amongst those who voted for either the top Democratic or Republican candidates. Using the weights

$$w = e^{-0.05d}$$

we used in Bigmood and polling averages, the average undecided voter percentage A is

$$A = \frac{\sum (1 - (D_i + R_i)) \cdot w_i}{\sum w_i}.$$

We then calculate VIBE using a logarithmic function to standardize the data as A was heavily skewed right.

$$\text{VIBE} = (2 \ln(A))^2$$

Gradient of Ordinary Fixedness Index (GOOFI)

Evaluating how accurately polls have been able to predict election results for states in the past can help us determine uncertainty in the polls for each state in this cycle. We calculate the Gradient of Ordinary Fixedness Index (GOOFI) by looking at how wrong the 2020 ORACLE was about each state's behavior in the 2020 presidential election. The percent error is calculated by

$$\epsilon = \frac{D_{2020} - \hat{D}}{D_{2020}}$$

where D_{2020} was the 2020 Democratic two-party vote percentage and $\hat{D}$ was the 2020 ORACLE prediction. GOOFI is then calculated with an arctangent function because we do not want our variance to explode because the 2020 ORACLE was not very accurate.

$$\text{GOOFI} = \left(\frac{0.08}{\pi} \arctan(0.2\epsilon) \right)^2$$

Overall Variance

We combine VIBE and GOOFI in the same manner we combined BABOON and polling averages using an arctangent function with the same weights. The idea behind this is the same; if we are given an infinite number of polls, the vast majority of our extra variance should come from the undecided voters as opposed to how wrong the 2020 ORACLE was.

$$\sigma^2_{\text{extra}} = w \cdot \text{VIBE} + (1 - w) \cdot \text{GOOFI}$$

The overall variance σ^2 is calculated by adding the sampling variances of the polls with the extra variance.

$$\sigma^2 = \sigma^2_{\text{polls}} + \sigma^2_{\text{extra}}$$

As a caveat, this assumes that the polling averages are independent from undecided voters and the error of the 2020 ORACLE. Although it is safe to assume the 2020 ORACLE error is independent from 2022 polls, there might be some correlation between the proportion of undecided voters A and the Democratic two-party vote percentage D. However, D can take on any number independent of what A is. For example, if the actual Democratic vote percentage is 20%, and A is 20%, the value of D is 25%. But if the actual Democratic vote percentage is 40%, then the value of D becomes 50%. There is likely little correlation between the actual Democratic vote percentage and undecided voter percentage, so the covariance between VIBE and the polling averages should be significantly smaller than the sum of the extra variance and sampling variance.

$$\sigma_{\text{extra,polls}} \ll \sigma^2_{\text{polls}} + \sigma^2_{\text{extra}}$$

Therefore,

$$\sigma^2 = \sqrt{\sigma^2_{\text{polls}} + \sigma^2_{\text{extra}} + \sigma_{\text{extra,polls}}}$$

simplifies to the independent case of

$$\sigma^2 = \sigma^2_{\text{polls}} + \sigma^2_{\text{extra}}.$$

Correlation

The leans we have now are naive predictions for the outcomes of each race. There is a chance that our predictions are wrong, and if our predictions are wrong for one state, they should be similarly wrong for states with similar demographics. Thus, we must correlate our predictions for states with similar demographics. The demographics of importance that we have decided are:

- the percentage of Black residents
- the percentage of Hispanic residents
- the percentage of Evangelical Christians
- the percentage of rural residents
- the percentage of White residents with a college education
- the median household income

We start by standardizing these values for each state by their z-scores. As an example, we will take state S. We find the correlation coefficients $r_1, r_2, \ldots, r_n$ between the z-scores of state S and the z-scores of all the other states. We then simulate the error between our prediction and what the actual Democratic voter percentage is $t_1, t_2 \ldots, t_n$ with a normal distribution of the same variance as the overall variance for that race. The idea is that since the error is the result of adding many small, independent sources of error together, the total error becomes normal by the central limit theorem, and the variance we already have is a good guess for the distribution's variance. We then take the weighted average of the errors with the correlation coefficients as weights to find the shift due to demographics Δx_{dem}.

$$\Delta x_{\text{dem}} = \frac{\sum_{i=1}^{n} r_i \cdot t_i}{\sum_{i=1}^{n} r_i}$$

This ensures that the bigger the correlation between two states, the larger the effect one state will have on the shift of another. As a note, we are only correlating senator races with each other and governor races with each other for now, and the two Oklahoma senator races are set to have a correlation coefficient of $r = 0.8$ with each other. We repeat and find the Δx_{dem} for every other race.

We also need to account for the correlation between a senator race and a governor race in the same state. We decided that the correlation coefficient for races in the same state was 0.6. Using this, we calculate the shift Δx_{state} by multiplying 0.6 to the simulated errors of the two races t_{senator} and t_{governor} and use that as the shifts for each race. For example, the shift due to the governor race on the senator race in state would be

$$\Delta x_{\text{state}} = 0.6 t_{\text{governor}}$$

and vice versa. If a state has only a senator race or only a governor race, $\Delta x_{\text{state}} = 0$.

Thus, the total shift Δx due to other races on a given race is equal to the sum of its shifts due to demographics and within state elections.

$$\Delta x = \Delta x_{\text{dem}} + \Delta x_{\text{state}}$$

The final estimate μ for a given race is then calculated by

$$\mu = \text{Lean} + \Delta x.$$

Example

Suppose that there are only states S, A, B, and C. The senator race in state S has a lean of 0.55.

- the gubernatorial race in state S has a predicted error of 0.05
- the senator race in state A has a predicted error of 0.01 and has a correlation of 0.9 with state S
- the senator race in state B has a predicted error of -0.05 and has a correlation of 0.1 with state S
- the senator race in state C has a predicted error of 0.1 and has a correlation of -0.5 with state S
- the gubernatorial race in state C has a predicted error of 0.02 but has zero correlation with the senate race in state S

We can perform the calculations outlined above.

$$\begin{aligned} \Delta x_{\text{dem}} &= \frac{0.9 \cdot 0.01 + 0.1 \cdot (-0.05) + (-0.05) \cdot 0.1}{0.9 + 0.1 - 0.5} && = -0.002 \\ \Delta x_{\text{state}} &= 0.6 \cdot 0.05 && = +0.030 \\ \Delta x &= -0.002 + 0.03 && = +0.028 \end{aligned}$$

Therefore the estimated Democratic vote percentage for state S is

$$\mu = 0.55 + 0.028 = 0.578.$$

Simulation

Now that we have a estimated vote percentage and variance for each race, we can now obtain the final prediction for the Democratic two-party vote percentage by taking a random value from the normal distribution $p \sim \mathcal{N}(\mu, \sigma^2)$. Each iteration of our model, we find a random value p for each race. The win probability of each race is calculated by the percentage of the time that the Democratic candidate won that race. The probability of winning the Senate is calculated by the percentage of the time that the Democratic party was able to secure enough seats to control 50 seats in the Senate. Each run of our model runs one million iterations and the results are displayed on our website. The implementation of our model can be found here.