Forecasting the 2024 U.S. Presidential Election*

Kamala Harris Projected to Defeat Donald Trump 48.3% to 47.1% in the Popular Vote and 270 to 268 in the Electoral College Based on Poll of Polls and Bayesian Modeling

Talia Fabregas Aliza Mithwani Fatimah Yunusa

November 4, 2024

The U.S. Presidential election will take place on Tuesday, November 5th with Vice President Kamala Harris and former President Donald Trump in a close race for the White House. In this paper, we used the poll-of-polls method and applied a Bayesian model to estimate the winner of the popular vote and the winner of each of the seven battleground states: Arizona, Georgia, Nevada, North Carolina, Michigan, Wisconsin, and Pennsylvania, and Nebraska's second congressional district. Using the results of our poll-of-polls analysis, we predict that Vice President Harris will win the popular vote, 48.3% to 47.1% and the electoral college 270 to 268 by winning three of the seven battleground states, Wisconsin, Michigan, and Pennsylvania. Our analysis shows that the race is extremely tight and former President Trump winning the popular vote, electoral college, or both is well within the margin of error. Our results show a statistical tie when we account for margin of error, bias, weaknesses, and limitations.

1 Introduction

On Tuesday November 5, 2024, Americans will head to the polls to elect their 47th president. Polling data has shown a tight race between Vice President Kamala Harris and former President Donald Trump since President Joe Biden made the historic and unprecedented decision to end his re-election campaign on July 21, 2024. President Biden immediately endorsed Vice President Harris on July 21, making her the presumptive Democratic nominee. Vice President Harris officially became the Democratic nominee on August 2, 2024 following a virtual rollcall. This paper uses presidential polling data from after President Joe Biden ended his re-election campaign and Vice President Harris became the presumptive Democratic nominee

*The code and data used to perform this presidential election forecast can be found at: https://github.com/taliafabs/USPresidentialPollingForecast2024.git.

and Bayesian models to estimate the percentages of voters supporting Vice President Harris and former President Trump.

We are interested in the effects of time since Biden ended his re-election campaign, state, pollster, and pollscore on the percentages of poll respondents supporting Vice President Harris and former President Trump at the national level, in each of the seven battleground states (Arizona, Nevada, Georgia, North Carolina, Wisconsin, Michigan, and Pennsylvania), and in Nebraska's second congressional district. Our estimand, which we can never know for sure, is the true effect of time, state, pollster, and pollscore on the percentages of voters supporting Harris and Trump (Alexander 2024).

Based on the results of applying our Bayesian models, we estimate that Vice President Harris will receive 48.3% of the popular vote and win the swing states of Michigan, Wisconsin, Pennsylvania, and Nebraska CD-2, and former President Trump will receive 47.1% of the popular vote and win the swing states of Arizona, Nevada, North Carolina, and Georgia. We use a baseline of 222 electoral votes for Vice President Harris from safe and likely Democratic states that were won by both Hillary Clinton and President Biden in 2016 and 2020, and 219 electoral votes for former President Trump from safe and likely Republican states that he won in 2016 and 2020 (270 to Win 2024). U.S. elections are decided by the electoral college, not the popular vote; the candidate who wins the popular vote is not guaranteed to win the election (270 to Win 2024). We estimate that Vice President Harris will defeat former President Trump in the electoral college 270 to 268.

Political polarization in the United States is at an all-time high. Vice President Harris and former President Trump have presented staunchly different policies and visions for the future of the United States. Vice President Harris has campaigned on reproductive rights, supporting small businesses, building more affordable housing, expanding medicare coverage, and cutting taxes for the middle class. Former President Trump has campaigned on tariffs, sales taxes ending foreign aid to Ukraine, and securing the Southern border despite instructing his allies in the U.S. Senate to oppose a bipartisan border bill. He is has not yet acknowledged that he lost the 2020 election to President Joe Biden. Women make up 50% of the U.S. population, but all 46 presidents so far have been men. Only one person of color, Barack Obama, has ever been elected president.

The remainder of this paper is structured as follows. Section 2 contains an overview of the polling dataset from FiveThirtyEight (2024) that was used, visualizations of different variables, and summary statistics. Section 3 contains the Bayesian regression models used to predict the percentages of voters that will support Vice President Harris and former President Trump. Section 4 contains tables and visualizations that present the national popular vote and state-level results after applying the Bayesian regression model. Section 5 contains detailed discussions about our results, why either candidate could win this election, past polling misses, and how our forecast can be improved. Appendix A contains a thorough discussion about an idealized methodology that we would use if given a $100,000 budget to build a survey to forecast the 2024 U.S. election and our idealized survey questions. Appendix B contains a deep-dive into and evaluation of the methodologies used in the late October 2024 Siena College/New York

Times battleground poll. The statistical programming language R (R Core Team 2023) and the `tidyverse`, `janitor`, `ggplot`, `kableExtra`, `arrow`, `rstanarm`, and `spline` packages were used to perform this election forecast, clean the dataset, build the Bayesian regression models, create data visualizations and tables, and apply the model.

2 Data

2.1 Overview

The data was downloaded on October 29, 2024; polling data released after this date was not considered anywhere in this paper. The presidential polls dataset from FiveThirtyEight contains national polls and state-level polls for each of the 50 states and congressional districts in Maine and Nebraska (FiveThirtyEight 2024). The polls are conducted by various pollsters, including YouGov, Siena/NYT, CES/YouGov, Marquette Law School, The Washington Post, and McCourtney Institute/YouGov. We analyze national data and state-level data for the seven swing states that are expected to determine the results of this election: Michigan, Wisconsin, Pennsylvania, North Carolina, Georgia, Nevada, and Arizona.

We cleaned this dataset to only include high-quality polls at the national and state-level conducted on or after July 21, 2024, where the `population` is likely voters. We determined which polls were high-quality based on the numeric grade, and cleaned the dataset to include only polls with a numeric grade of at least 3.0. Polls with a rating (`numeric_grade`) of 3.0 are conducted by the best pollsters in the United States and score in the 99th percentile or better for accuracy and transparency (Morris 2024b). If a poll in the dataset from FiveThirtyEight (2024) includes multiple populations, the narrowest one is used (likely voters over registered voters and registered voters over all American adults) (Morris 2024a). Additional data cleaning details are available in Appendix C.1.

2.2 Measurement

Our primary measurement task is to capture public opinion and translate it into a 2024 U.S. presidential election prediction. Surveys are a common instrument for measuring public opinion (Alexander 2024). During presidential election cycles, pollsters conduct surveys to measure public opinion and candidate preferences. These surveys aim to represent the U.S. electorate by sampling likely or registered voters and asking them questions about demographics, partisan affiliation, candidate preferences, and stances on issues. Each survey response reflects an individual's voting preference, which pollsters sum up and adjust to represent the population This includes weighting by state, demographic factors (e.g. age, education, race, gender) and accounting for the likelihood of voting (Office of Institutional Reseaerch 2024). These adjustments turn raw opinions into a projected percentage of support for each candidate and make it possible to predict potential election outcomes.

Our dataset from FiveThirtyEight (2024) is a collection of presidential polls from different pollsters that were conducted during the 2024 presidential election cycle. Each entry represents the percentage of respondents to a unique poll supporting Vice President Harris (after July 21, 2024) or President Joe Biden (before July 21, 2024), former President Trump, and third-party candidates. Unique polls are identified by a poll id and each entry contains information about the poll, such as the pollster that conducted it, its population, sample size, and the methodology that was used and information about its quality and accuracy, including its numeric grade, pollscore, and transparency score(FiveThirtyEight 2024).

The transformation from an individual opinion to an entry in our dataset follows three steps to translate voter preferences into structured data that allows us to look at trends and predict the outcome of an election

- Survey: selected voters respond to a survey.
- Adjustment: survey responses are aggregated and weighted to estimate support for each candidate.
- Reporting: the results from the adjustment step are recorded as dataset entries, which serve as snapshots of public opinion over time.

2.3 Outcome and predictor variables

We will use `end_date` (the date that a poll was completed), `state`, `pollster`, and `pollscore` to predict support for Vice President Harris and former President Trump at the national level, at the state level for each of the seven battleground states, and in Nebraska's second congressional district in Section 3. The tables and visualizations below present possible relationships between the predictor variables and support for Vice President Harris and former President Trump.

2.3.1 Variation in support for Harris and Trump by state

The percentage of voters that support Vice President Harris and former President Trump varies by state; in some states support for Vice President Harris is higher than the national average and in others, it is lower than the national average. The 2024 presidential election is expected to be decided by seven swing states: Arizona, Georgia, Nevada, North Carolina, Wisconsin, Michigan, and Pennsylvania and Nebraska's second congressional district (270 to Win 2024). Two states, Maine and Nebraska, award one electoral vote to the popular vote winner in each congressional district and an additional two electoral votes to the statewide popular vote winner (270 to Win 2024).

Table 1: Polling averages for Harris and Trump at the national level and at the state level for the states included in the polling dataset show a narrow popular vote lead for Vice President Harris and extremely close races in the 7 battleground states (Arizona, Georgia, Nevada, North Carolina, Michigan, Pennsylvania, Wisconsin as of October 29, 2024.

State	Harris %	Trump %
Popular Vote	50.5	48.0
Arizona	47.0	51.0
Georgia	46.0	51.0
Nevada	51.0	47.0
Pennsylvania	49.0	49.0
Michigan	51.0	46.0
Wisconsin	50.0	47.0
North Carolina	48.0	50.0
Florida	46.0	52.0
Minnesota	53.0	43.0
Missouri	41.0	54.0
Montana	39.5	56.5
Nebraska	39.5	54.0
Nebraska CD-2	53.5	41.5
New Hampshire	52.0	45.0
Ohio	45.0	52.0
Texas	41.0	51.5
Virginia	52.0	44.0

Based on our presidential polling data from FiveThirtyEight (2024) Vice President Harris leads former President Trump in the popular vote 50.5% to 48.0%. The margins in the seven battleground states are tight, with Vice President Harris leading in Nevada, Michigan, Wisconsin, and Nebraska's second congressional district and former President Trump leading in Arizona, Georgia, North Carolina. The two candidates are tied in Pennsylvania. The data set also includes state-level polls from likely Democratic states (Minnesota, New Hampshire, Virginia) and likely Republican states (Florida, Missouri, Montana, Nebraska, Ohio, Texas) that are not expected to determine the winner of the election (270 to Win 2024). Support percentages for Harris and Trump are closer, but still have some variation among the seven battleground states. Harris' support in five of the seven battleground states is lower than her national support, while Trump's support in four of the seven battleground states is higher than his national support.

Since President Biden ended his re-election campaign and Vice President Harris became the Democratic Presidential nominee, the polls have shown a close race between Vice President

Harris and former President Trump. Polling averages for the six months leading up to election day, including from before President Biden withdrew on July 21, 2024 can be found in Appendix C.2. Figure 1 shows national polling averages for Harris and Trump since July 21 and Figure 2 shows state-level polling averages fo Harris and Trump in the seven battleground states and Nebraska's second congressional district.

Vice President Harris surpassed former President Trump in popular vote polls shortly after becoming the presumptive Democratic nominee in late July, but her lead narrowed in mid-August and the polls have been neck-and-neck since late August.

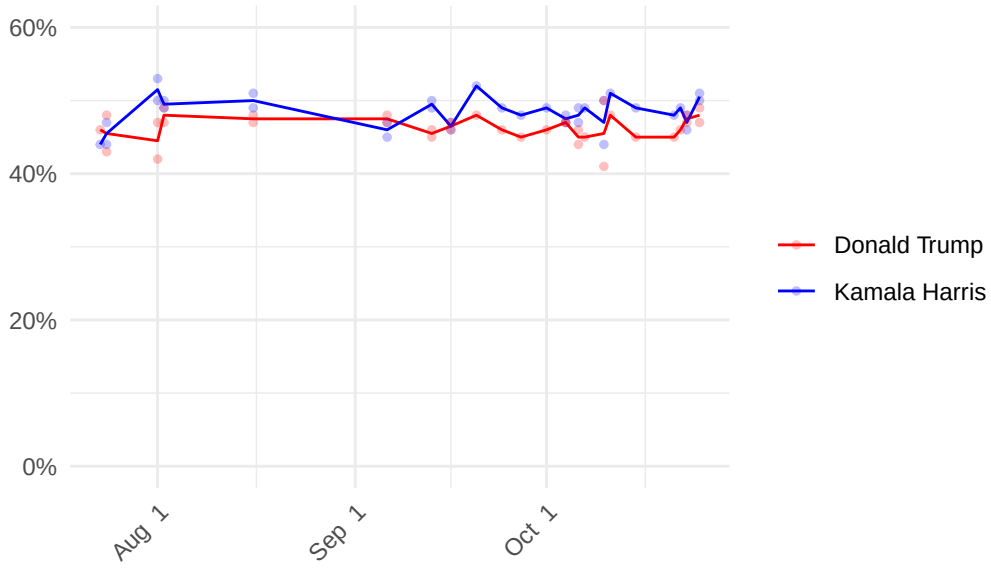


Figure 1: Harris leads Trump in national popular vote polling. The lines are moving poll averages and the points are individual high-quality polls. Color shading of points indicates which candidate won the poll.

The changes in Harris and Trump's polling averages over time at the state level has varied among the seven battleground states and Nebraska's second congressional district. Trump has had a narrow lead over Harris in Arizona since early September. He has been leading over Harris in Georgia since she entered the race. Harris and Trump were tied in Michigan until early October, when Harris took the lead. There is not a lot of polling data for Nebraska's second congressional district, but Harris has had a wider lead there than either candidate has ever had in any of the seven battleground states since late September. Harris and Trump were virtually tied in North Carolina until Trump took a very narrow lead in October, and they are now virtually tied in Pennsylvania. Harris has had a narrow lead in Wisconsin since August.

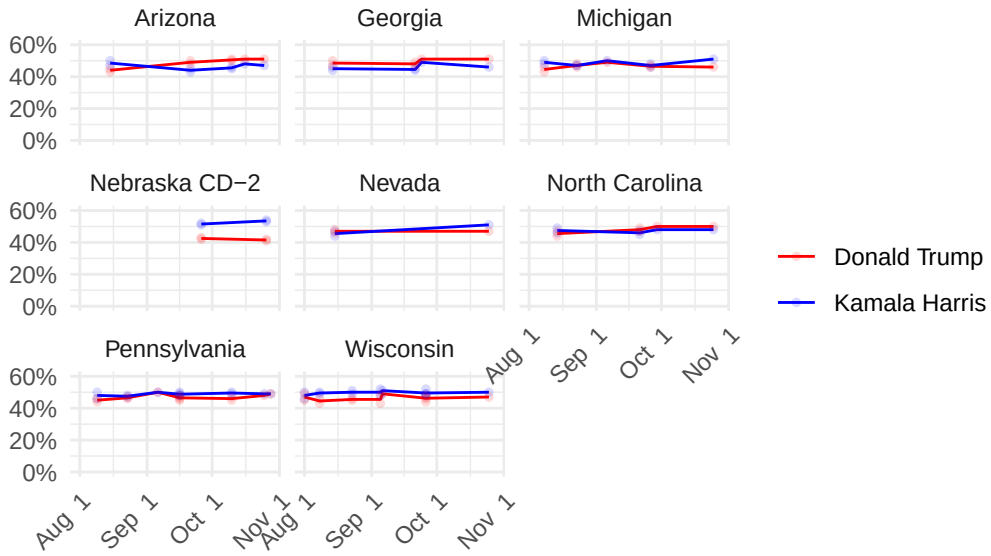


Figure 2: Harris and Trump have been in a dead heat in swing state polls since July 2024. The lines are moving poll averages for each candidate in the seven battleground states. The points are individual state-level polls. The color of each point indicates which candidate lead in the poll.

2.3.2 Variation in support for Harris and Trump by pollster and pollscore

Individual pollsters can produce results that are skewed in favor of either candidate. This can be caused by differences in methodology, respondent recruitment, sample size, or how non-responses are handled (The New York Times 2024c). Overall, the different pollsters within our dataset have shown very close polling averages for Harris and Trump, but there is some variation. Figure 3 shows that Harris and Trump national polling averages within our dataset vary by pollster. Siena/NYT polls showed Harris and Trump virtually tied until mid-September when Harris took a narrow lead, but have been virtually tied again since mid-October. YouGov polls have shown a narrow lead for Harris since mid-August and Marquette Law School polls have shown a shrinking lead for Harris since she became the Democratic nominee. Support for Harris and Trump within a poll can be affected by the pollster that conducted it. We "pool the polls", or average the results from different polls to balance out pollster biases (PASEK 2015). There are fewer CES/YouGov and McCourtney Institute/YouGov polls in our dataset, but they have both shown a narrow lead for Harris.

Pollscore indicates whether a pollscore is more accurate than a theoretical replacement-level pollscore that polled the exact same election and a negative pollscore is better (Morris 2024b). Our dataset only includes high-quality polls with a numeric grade of 3, which means they have negative pollscores, excellent transparency, and high accuracy (Morris 2024b). The pollster in our dataset with the best pollscore is Siena/NYT and its polling averages for Harris and Trump

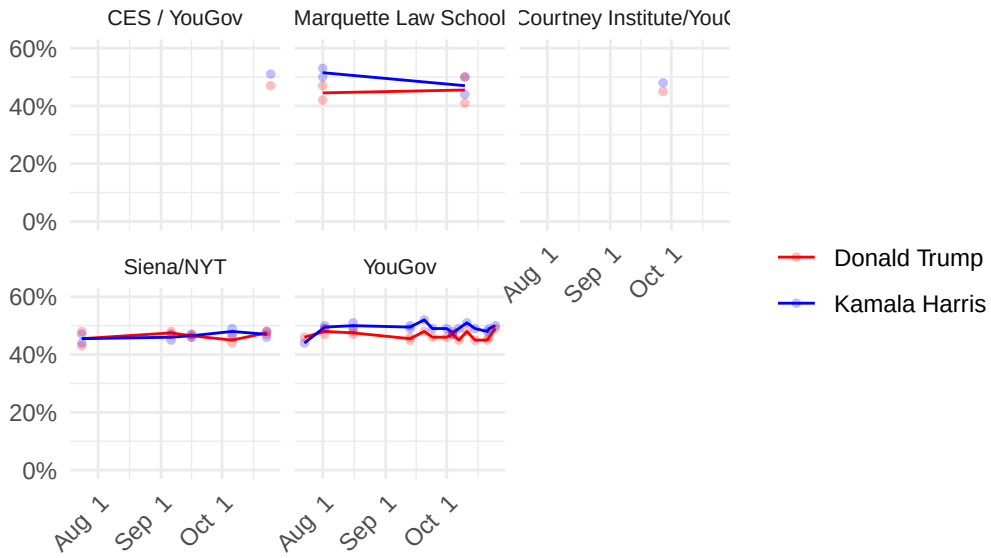


Figure 3: Harris and Trump polling averages vary by pollster. The lines are moving poll averages for each candidate in the seven battleground states. The points are individual polls. The color of each point indicates which candidate lead in the poll.

have varied from other pollsters. We will consider the effects of pollscore on the percentage of respondents to a poll that support Harris and Trump.

Table 2: Only high-quality datasets with a numeric grade of at least 3.0 were included, so the polls included all have good pollscores. Negative/lower pollscores are better.

Pollscore	Number of Polls
-1.5	112
-1.2	8
-1.1	114

Table 3: All the pollsters included in our analysis dataset have a numeric grade of 3, but Siena/NYT has the best pollscore at -1.5.

Pollster	Pollscore	Numeric Grade
CES / YouGov	-1.1	3
Marquette Law School	-1.1	3
McCourtney Institute/YouGov	-1.1	3
YouGov	-1.1	3
The Washington Post	-1.2	3
Siena/NYT	-1.5	3

Table 3: All the pollsters included in our analysis dataset have a numeric grade of 3, but Siena/NYT has the best pollscore at -1.5.

Pollster	Pollscore	Numeric Grade

3 Model

The goal of our modeling strategy is to use a Bayesian model to investigate the relationship between the percentage of voters in a poll who support Vice President Harris or Former President Trump and the date the poll was conducted, the state (or if it was a national poll), the pollster who conducted the poll, and the pollscore it received. Model details, validation, checking, and diagnostics are presented in Appendix D. We use two Bayesian regression models, one each to model the percentage of voters supporting Vice President Harris, *pct_harris* and the percentage of voters supporting former President Trump, *pct_trump*.

The models to predict *pct_harris* and *pct_trump* both use the following predictors:

- `end_date_num`: the number of days since July 21, 2024, when President Biden ended his re-election campaign and endorsed Vice President Harris. This is the spline term; it uses an spline with 5 degrees of freedom to model changes in *pct_harris* (in the Harris model) and *pct_trump* (in the Trump model) over time.
- `state`: accounts for the change in *pct_harris* and *pct_trump* at the state-level for a particular state or at the national level.
- `pollster`: accounts for the differences in *pct_harris* and *pct_trump* among different pollsters.
- `pollscore`:adjusts for the pollscore rating, which is the Predictive Optimization of Latent skill Level in Surveys, Considering Overall Record, Empirically, calculated by averaging the predictive error and predictive bias of a poll (Morris 2024a).

3.1 Model set-up

Define pct_harris_i as the percentage of voters supporting Vice President Harris and pct_trump_i as the percentage of voters supporting former President Trump in the poll with unique `poll_id`=i.

Harris model:

$$
\begin{aligned}
pct\_harris_i &= \beta_0 + \beta_1 \cdot ns(end\_date\_num_i, df = 5) + \beta_2 \cdot state_i + \beta_3 \cdot pollster_i + \beta_4 pollscore_i \\
\beta_0 &\sim \text{Normal}(50, 10) \\
\beta_1 &\sim \text{Normal}(0, 5) \\
\beta_2 &\sim \text{Normal}(0, 5) \\
\beta_3 &\sim \text{Normal}(0, 5) \\
\beta_4 &\sim \text{Normal}(0, 5)
\end{aligned}
$$

Trump model:

$$
\begin{aligned}
pct\_trump_i &= \beta_0 + \beta_1 \cdot ns(end\_date\_num_i, df = 5) + \beta_2 \cdot state_i + \beta_3 \cdot pollster_i + \beta_4 pollscore_i \\
\beta_0 &\sim \text{Normal}(50, 10) \\
\beta_1 &\sim \text{Normal}(0, 5) \\
\beta_2 &\sim \text{Normal}(0, 5) \\
\beta_3 &\sim \text{Normal}(0, 5) \\
\beta_4 &\sim \text{Normal}(0, 5)
\end{aligned}
$$

The models are run in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). The default priors from `rstanarm` are used for both GLM Bayesian models. The intercept normal prior with $\mu = 50$ and $\sigma = 10$ reflects the central tendencies of Harris and Trump's polling percentages, influenced by prior knowledge and the predictors use a normal prior with $\mu = 0$ and $\sigma = 5$ (Goodrich et al. 2022).

3.1.1 Model justification

We use separate Bayesian regression models with the same predictors to estimate *pct_harris* and *pct_trump*. This allows us to predict the percentage of voters supporting Vice President Harris and former President Trump, using the same high quality polls from FiveThirtyEight (2024) and the same predictors. We use `end_date_num` (number of days since President Biden ended his campaign on July 21, 2024) as the spline term because we want our Harris model to account for changes *pct_harris*, in and our Trump model to account for changes in *pct_trump* over time. We use state, pollster, and pollscore as predictors to account for changes in *pct_harris* and *pct_trump* at the national or state-level, across different pollsters, and different pollscores. The Harris model is trained on Vice President Harris' polling data and the Trump model is trained on former President Trump's polling data from FiveThirtyEight (2024).

We used the default priors from `rstanarm` in both our models because using a normal prior with $\mu = 50$ and $\sigma = 10$ for the intercept of each model allows us to reflect the central tendencies of *pct_harris* and *pct_trump* and the prior knowledge from the FiveThirtyEight (2024) presidential polls that show a close race between Vice President Harris at the national popular vote level and in each of the seven battleground states (Goodrich et al. 2022).

Initially, we considered using a single Bayesian regression model to predict *pct_harris*, using `end_days_num` (spline term), `state`, `pollster`, and `pollscore` as predictors. If we used this single Bayesian regression model, we would have calculated *pct_trump* as 100% - *pct_harris*. Calculating the percentage of voters supporting Trump as 100% minus the percentage of voters for Harris would have been insufficient because there are third-party candidates who will receive a small percentage of the vote nationally and in each of the seven battleground states. This may have produced inaccurate or misleading results because either Harris or Trump could win the popular vote or a swing state with less than 50% of the vote (FiveThirtyEight 2024). We would have assumed that if Harris received less than 50% of the vote, she would have lost, even though this is not always the case. President Biden won the swing states of Arizona, Georgia, and Wisconsin in 2020 with less than 50% of the vote (CNN 2016a). Former President Trump won Wisconsin, Michigan, Pennsylvania, and Arizona in 2016 with less than 50% of the vote (CNN 2016b). With many presidential polls this election cycle showing a statistical tie between Harris and Trump in the popular vote and electoral college, calculating Trump's support this way could overestimate it by a few percentage points and produce an inaccurate or misleading result (Silver 2024).

3.1.2 Model weaknesses and limitations

The use of two separate models to predict the percentage of voters supporting Vice President Harris and the percentage of voters supporting former President Trump has weaknesses and limitations. We do not use support for Harris as a predictor in the Trump model or support for Trump as a predictor in the Harris model. However, we know that as support for Harris at the national level or in one of the seven battleground states increases, support for Trump decreases (and vice versa). We could improve our model by adding support for Harris as a spline term in our Trump model and support for Trump as a spline term in our Harris model. This would allow us to use changes in support for Harris over time to predict support for Trump and changes in support for Trump over time to predict support for Harris. Our models do not give higher quality polls or polls with a larger sample size more weight. Our dataset only includes polls with a numeric grade of 3, so this is likely not an issue when we apply our models to it. We did not consider sample size when cleaning our data or building our models because we only included high-quality polls in the 99th percentile for accuracy and transparency. If our models were applied to a larger dataset that includes polls with different numeric grades, these could be weaknesses.

4 Results

Model results and model summary are presented in Appendix D.

4.1 National popular vote results

We applied the Harris model and the Trump model defined in Section 3.1 to predict each candidate's share of the national popular vote.

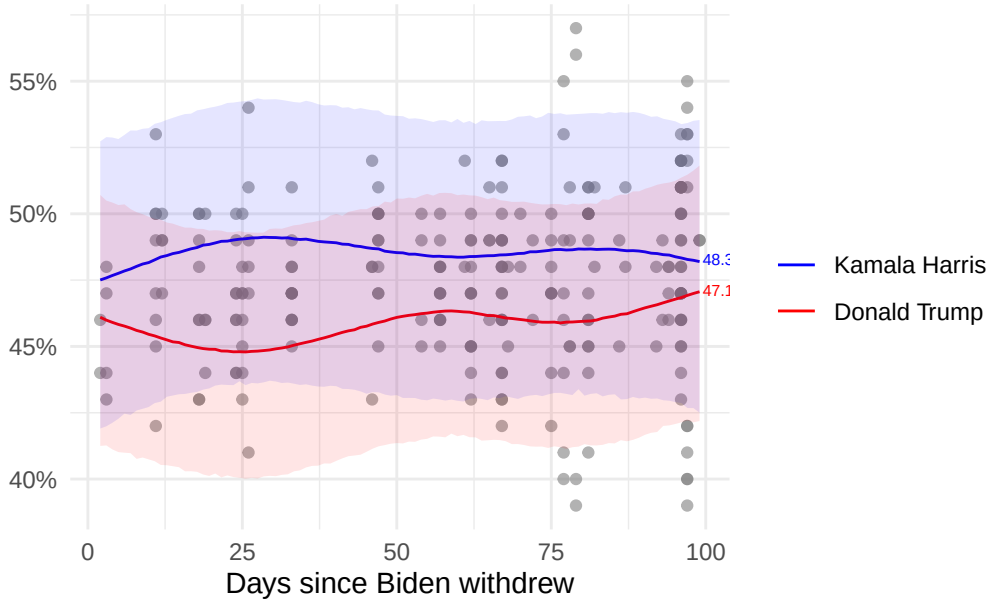


Figure 4: Results of applying Bayesian regression models with spline to predict support for Harris and Trump show that Harris is leading the popular vote 48.3% to 47.1% as of October 29. Lines are moving averages from applying the model, points are individual polls, and the shaded areas show the range of values within the margin of error.

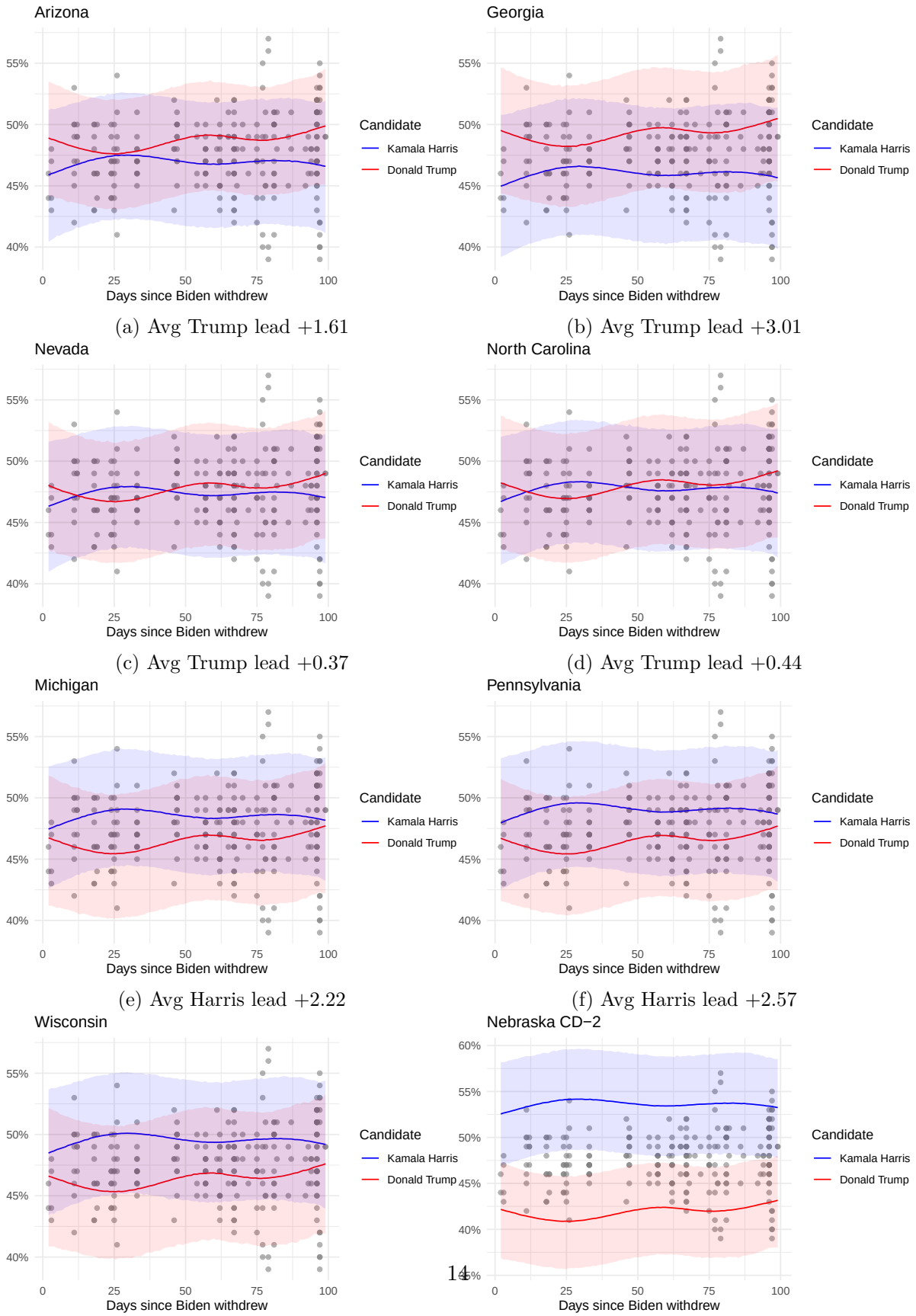
4.2 State-Level results for the seven battleground states and Nebraska's second congressional district

Vice President Harris starts out with 225 electoral college votes from states that went to both Hillary Clinton in 2016 and Joe Biden in 2020 (270 to Win 2024). These states are considered safe or likely Democrat and available poling from FiveThirtyEight (2024) shows Vice President Harris with a lead outside the five percent margin of error and in most cases, a double-digit lead. Former President Trump starts out with 219 electoral college votes from states that he won in both 2016 and 2020 (270 to Win 2024). These states are considered safe or likely Republican

and available polling data from FiveThirtyEight (2024) shows former President Trump with a lead outside the five percent margin of error, and in most cases, a double-digit lead. For this reason, the results of the Bayesian models presented in Section 3 focus on estimating support for Vice President Harris and former President Trump in Arizona, Nevada, Georgia, North Carolina, Wisconsin, Michigan, Pennsylvania, and Nebraska's second congressional district, which are worth a total of 94 electoral votes (270 to Win 2024).

We applied the Harris model and the Trump model defined in Section 3.1 at the state-level to predict the percentage of the vote that each candidate will receive in each of the seven battleground states. This will allow us to predict the winner of the electoral college and who will become the 47th President. Figure 5 shows the results of applying the Harris and Trump Bayesian models with spline at the state-level over time.

Figure 5 shows former President Trump leading in Arizona, Nevada, Georgia, and North Carolina and Vice President Harris leading in Michigan, Wisconsin, Pennsylvania, and Nebraska's second congressional district as of October 29, 2024. The margins of error for Trump and Harris have little overlap for Nebraska CD-2, so we can predict that Harris will likely win its single electoral vote. Nebraska CD-2 will not be considered safe or likely Democratic because there is still overlap in the margins of error for Harris and Trump and it was not won by Clinton in 2016. The shaded margins of error for the seven battleground states have a lot of overlap; this means that it is within the margin of error for either candidate to win any of the seven battleground states.



(a) Avg Trump lead +1.61

(b) Avg Trump lead +3.01

(c) Avg Trump lead +0.37

(d) Avg Trump lead +0.44

(e) Avg Harris lead +2.22

(f) Avg Harris lead +2.57

(g) Avg Harris lead +3.31

(h) Avg Harris lead +11.87

Figure 5: The lines are the weighted moving averages resulting from applying the Bayesian model with spline, dots are individual state-level polls, and shaded areas show range of values within the margin of error.

4.3 Electoral College Results

Before the seven key battleground states are decided, Vice President Harris has 225 electoral votes from safe and likely Democratic states and former President Trump has 219 from safe and likely Republican states (270 to Win 2024). We use this starting point and the results from Section 4.2 to predict the electoral college.

Table 4: Electoral votes for the seven battleground states and Nebraska CD-2

State	Electoral Votes
Arizona	11
Georgia	16
Michigan	15
Nebraska-2	1
Nevada	6
North Carolina	16
Pennsylvania	19
Wisconsin	10

If the results presented in Figure 5 hold on election night, and Harris and Trump win their safe and likely states, Harris would defeat Trump in the electoral college, 270 to 268. This is the mean estimate for the number of electoral college votes for Harris and Trump. If Vice President Harris wins all seven battleground states and Nebraska CD-2, she would defeat former President Trump in the electoral college 319 to 219. If former President Trump wins all seven battleground states, he would defeat Vice President Harris in the electoral college 313 to 225.

Table 5: The electoral college outcomes that fall within the margin of error for the seven battleground states.

	Harris	Trump
Harris Upper/Trump Lower Estimate	319	219
Mean Estimate	270	268
Harris Lower/Trump lower Estimate	225	313

5 Discussion

5.1 Polling misses in 2016 and 2020

In 2016 and 2020, pollsters underestimated support for Trump at the national level and in the key battleground states. 8% of national polls overestimated support for Hillary Clinton in 2016, and 93% of national polls overestimated support for President Biden in 2020 (Kennedy et al. 2021). As a result, pollsters predicted that Clinton would win the 2016 election and Biden would win the 2020 election with larger popular vote and swing state margins than he actually did. Silver (2024) points to non-response bias as the main reason why pollsters underestimated support for Trump in key battleground states in two presidential elections.

Responses from Trump voters were missing not at random because there was a relationship between the lack of willingness to respond to a survey and support for Trump (Alexander 2024). Trump voters are more likely to be low-information voters who distrust government and social institutions, which makes them less likely to respond to an election survey (Silver 2024). Durand et al. (2017) found that swing state voters with a college education were more likely to support Hillary Clinton and to respond to a survey, while swing state voters without a college education were more likely to support Trump and less likely to respond to a survey. Silver (2024) suggests that the fact that Democrats were more likely to stay home because of Covid-19 restrictions gave them more time to answer phone calls or respond to online surveys may have contributed to the 2020 polling misses. Pollsters did not reach enough Trump voters in 2016 and 2020 (Silver 2024). They also failed to adjust for the over-representation of college-educated voters in swing states (Durand et al. 2017). Pollsters' samples were not representative of the population that voted in either election and their predictions missed as a result. A survey that is conducted on a non-representative sample will produce skewed and inaccurate results (Alexander 2024).

5.2 Why either Trump or Harris could win

The 2024 U.S. presidential election is a tossup. Vice President Harris and former President Trump have been polling within a few percentage points of each other in national and state-level polls in Arizona, Nevada, Georgia, North Carolina, Wisconsin, Michigan, and Pennsylvania since the summer (Silver 2024). Our results presented in Section 4 show Harris leading Trump 48.3% to 47.1% in the popular vote and 270 to 268 in the electoral college. Our electoral college prediction is based on narrow leads for Harris in Wisconsin, Michigan, and Pennsylvania and narrow leads for Trump in Arizona, Nevada, Georgia, and North Carolina. Harris' popular vote lead and all of Harris' and Trump's leads in the seven battleground states fall well within the margin of error. Fifty-fifty is the only reasonable forecast that we can draw from this (Silver 2024). It is possible that these fifty-fifty forecasts are wrong, and both candidates are equally likely to out perform them and win the election comfortably (Silver 2024). A normal polling error could swing the popular vote and the seven battleground states in either candidate's

direction (Silver 2024). Silver (2024) found that the average polling miss is 3 to 4 percentage points. If either Harris or Trump outperforms their swing state and national polling averages by 3 to 4 percentage points, they would win the popular vote, sweep the seven battleground states, and win the electoral college comfortably (Silver 2024). This would result in Harris winning the electoral college 319 to 219 or Trump winning the electoral college 313 to 226; both of these results are within the margin of error shown in Section 4.

Former President Trump could outperform his polling averages and win the election comfortably if the pollsters are not accounting for social desirability bias and have not been able to correct the non-response bias and low response rates that caused them to underestimate support for him in 2016 and 2020 (Silver 2024). The shy Trump voter theory and the combination of the "Bradley effect" and the "Hillary Clinton effect" are rooted in social desirability bias (Silver 2024). According to Silver (2024), Trump voters might be shy to admit that they will vote for him due to the social stigma attached to it and some voters might be unwilling to vote for a woman of color, but shy to admit, it so they tell pollsters that they are undecided.

Vice President Harris could outperform her polling averages and win the election comfortably if the new techniques that pollsters have used to try to correct their 2016 and 2020 misses have gone too far (Cohn 2024). These new techniques include weighting by highest level of education and weighting on recalled vote (Silver 2024). Weighting by highest level of education aims to correct non-response by giving respondents without a college degree more weight because they are less likely to respond to surveys and more likely to support Trump (Silver 2024). Weighting on recalled vote is used to ensure that the sample includes enough Trump voters by accounting for who survey respondents say they voted for in 2020 (Cohn 2024). It aims to account for non-response bias by assuming that some respondents who say they voted for Biden in 2020 are actually Trump voters who do not report their 2020 vote accurately because they would rather say they voted for the winner (Silver 2024). More than two-thirds of presidential polls from October 2024 were weighted on recalled vote (Cohn 2024). Silver (2024) warns that these polls may have over corrected 2016 and 2020 polling misses and overestimated support for Trump by incorrectly labeling respondents who said they voted for Biden in 2020 as Trump voters. Cohn (2024) found that weighting on recalled vote would have led to a larger polling miss and overestimated support for the losing candidate in every presidential election since 2004.

5.3 Weaknesses, limitations, and next steps

Harris and Trump have been deadlocked in national and battleground state-level polls this election cycle. In Section 4, we predicted that Vice President Harris will win the popular vote to and the electoral college 270 to 268. These are razor-thin margins, indicative of a statistical tie. Our approach and results are not without weaknesses and limitations. As previously mentioned in Section 3.1.1, our Harris model does not use support for Trump as a predictor and our Trump model does not use support for Harris as a predictor. This is a weakness in our models because we know that as support for Harris increases, support for

Trump decreases (and vice versa). Not including it in our model could weaken our results, especially in the seven battleground states. Our model and results do not explicitly consider third-party candidates. We used separate models with the same predictors to estimate support for Harris and Trump because it allowed us to have percentages of voters supporting Harris and Trump that do not add up to 100%, and account for the small percentage of voters that will vote for a third-party candidate. In 2016, third-party candidates received more votes than Trump's margins of victory in Wisconsin, Michigan, and Pennsylvania (CNN 2016b). In 2020, third-party candidates received more votes than Biden's margins of victory in Arizona and Georgia (CNN 2016a). Even though third-party candidates only receive a small percentage of the vote in U.S. presidential elections, considering them can help us accurately predict support for Harris and Trump in battleground states that will decide the 2024 election and could be decided by just one percentage-point. A next step is to consider third-party candidates; this could be as a separate model or as a predictor for support for Harris and Trump.

Once the results of the 2024 U.S. presidential election are in, our next step will be to compare the actual results to the polls and our own forecast and study why the polls were accurate or inaccurate. If the results are within the margin of error, we will know that the polls accurately predicted a tight race. If Harris or Trump wins in a landslide — by sweeping all seven battleground states, winning some battleground states by a wide margin, or flipping one of the losing candidate's safe or likely states — we will know that the polls overestimated support for one candidate and underestimated support for the other. Either way, we will go through 2024 pollster methodologies in more depth, compare them to 2016 and 2020, and study why the new techniques did or did not work. In Section 5.1, we discussed why the polls in 2016 and 2020 were inaccurate, so our next step will involve doing the same for the 2024 polls. We will also be able to use this information to discuss what pollsters might be able to do to improve the accuracy of their predictions in the 2028 election cycle.

Appendix

A Idealized methodology

A.1 Survey objectives

This idealized methodology outlines a plan to conduct a survey to forecast the 2024 U.S. presidential election with a budget of $100,000 USD. It includes plans for a sampling approach, respondent recruitment, data validation, and survey design. The Times/Siena polls were within an average of two percentage points of the actual results during the 2022 mid-term election cycle. All polls and survey methodologies have weaknesses and limitations, so we will focus on correcting non-response bias, which led to the 2016 and 2020 polling misses.

A.2 Sampling approach

The target population, defined by Alexander (2024) as the population that we would like to represent in our survey, is likely voters in the United States. The sampling frame, defined by Alexander (2024) as potential respondents from the target population, is voters from the L2 database (L2 2024). We will use the L2 voter database to build a representative sampling frame. The L2 voter database is one of the most trusted sources for enhanced voter data and it includes detailed demographics and voting history data (L2 2024). Using this data set is the first step towards ensuring that our sample aligns with the general electorate and is representative of likely voters so that we can sample precisely.

We will combine probability and non-probability based sampling methods within the strata to make the most of our resources and improve the reach of our survey (Mercer and Lau 2023). Probability based methods ensure that each individual in the sampling frame has a known, non-zero chase of being selected and allow for results to be generalized to the broader population (Alexander 2024). Non-probability based methods select individuals in a way that does not give each member of the sampling frame a known probability if being selected (Alexander 2024).

We will use a stratified sampling (probability-based) approach to closely examine voter demographics. Stratified sampling allows us to look at every stratum and carry out simple random sampling with those strata (Alexander 2024). Its main goal is to ensure that every strata of the population is represented (Neyman 1934). We selected stratified sampling because it will allow us to have representation within the subgroups that we are interested in and it has a reduced sampling error and improved accuracy rate (Alexander 2024). We considered simple random sampling, but ultimately chose stratified sampling instead because it tends to produce more precise estimates when used to forecast U.S. elections (Pew Research Centre 2024).

In 2016 and 2020 presidential election polls underestimated support for Trump because they did not reach out to enough non-college educated, white respondents from rural areas who were more likely to support Trump but less likely to respond to a pollster (Silver 2024). We will use respondent-driven sampling, which is a non-probability based method, to reach hard-to-reach populations and reduce non-response bias (Raifman et al. 2021). Raifman et al. (2021) found that respondent-driven sampling is well equipped to reach members of hard-to-reach populations who are least likely to respond to a survey.

The U.S. election is decided by the electoral college, not the popular vote. We will over sample from the seven battleground states that are expected to decide the 2024 election: Arizona, Nevada, Georgia, North Carolina, Wisconsin, Michigan, and Pennsylvania (FiveThirtyEight 2024). Oversampling from the seven battleground states will allow us to focus on forecasting their results and increase the accuracy of our electoral college estimate. Our target sample size of 100,000 respondents will provide a margin of error of 2% at a 95% confidence level; this will allow us to be precise while still accounting for budgetary constraints (Pew Research Centre 2024).

A.3 Respondent recruitment

We will use multi-channel respondent recruitment, including both phone and digital outreach. Phone outreach will allow us to capture older demographics and individuals who use the internet and social media less. Telephone polls have fared well in recent elections because voter registration files, like L2, provide a clear way to ensure a balance between Democrat and Republican respondents and are a great way to reach a random sample quickly (The New York Times 2024c). Digital outreach, such as targeted social media ads, email lists, and news websites will enable us to recruit a broader sample. Digital outreach channels are cost effective and can reach younger demographics. We will allocate $20,000 of our budget for rewarding phone poll respondents, particularly within harder-to-reach groups such as young people, low-information voters, and voters without a college education in the battleground states to reduce non-response bias. Smith et al. (2019) found that incentives increase survey participation, particularly among hard-to-reach groups and demographics that are unlikely to respond. Siena College/New York Times polls typically have response rates around 2%, which is typical for high-quality telephone polls (The New York Times 2024c). However, a Times/Siena telephone poll in Wisconsin that paid voters up to $25 to respond to it obtained a response rate close to 30%. (The New York Times 2024c).

A.4 Data Validation

Common issues with surveys include incomplete or inconsistent responses, duplicate responses, and fake responses by bots (Alexander 2024). We will use automate validation checks to detect and filter out incomplete or inconsistent responses, IP tracking to prevent duplicate responses, and CAPTCHA technology to minimize bot interference (Zhang et al. 2019).

A.5 Weighting and Data Adjustments

We will use post-stratification weighting to correct demographic imbalances in our final sample (The New York Times 2024a). Kolenikov (2016) defines post-stratification weighting as a statistical adjustment technique where the sample is divided into strata based on demographic characteristics and weights that correspond to how much of the general population they make up are assigned to each group. Post-stratification weighting will help us make our poll an accurate representation of likely voters and address non-response bias by increasing the weight of groups that are underrepresented in our sample (Clinton 2024). A limitation of post-stratification weighting is that it relies on the assumption that the political preferences of the voters in each strata who took the our survey are similar to those who did not take the survey (Clinton 2024)

A.6 Budget

Our $100,000 budget is structured to create a reliable, representative survey. Personnel costs ($30,000) cover management and analysis. Incentives ($20,000) boost response rates, especially among hard-to-reach groups like young adults. Advertising and recruitment ($20,000) funds targeted outreach via digital ads and phone calls. Survey platform costs ($15,000) support a user-friendly platform with necessary features. Data validation ($10,000) ensures quality through automated checks and IP tracking. Finally, poll aggregation software ($5,000) enables effective trend analysis, GPU capacity, and rolling averages.

A.7 Survey design

Our survey is conducted using GoogleForms, which allows for an easy, user-friendly experience and increases respondent likelihood. The survey length is kept within the five to ten minute range, with thirteen questions to balance the need for data collection with the need for respondent engagement (Stantcheva 2023). As suggested by Stantcheva (2023), we used plain language throughout the survey and kept our questions short to reduce the cognitive load and the risk of voters exiting the survey because they find it too tedious to answer. Stantcheva (2023) defines priming as a cognitive-based question order effect; it involves using the question order to make respondents think about a topic related to the following questions and therefore, influencing the respondent's answers to those questions. We placed questions about key issues such as abortion and the economy after the question about preferred 2024 presidential candidate, so our survey does not have priming. We include only one open-ended question in our survey, at the very end, to encourage respondents to think more about and engage with our survey (Stantcheva 2023).

Our survey includes the following sections:

- Demographics: In this section, we collect essential demographic data that we will use for post-survey weighting. This includes gender and ethnicity.
- Candidate evaluation: This section includes direct questions about candidate support, vote choice, and support level.
- Key issues and concerns: This section includes questions that aim to identify the top issues driving voter decisions, such as abortion rights and the economy.

A.8 Tradeoffs and limitations

Our idealized survey is not without trade-offs and limitations. Every question in our survey is required. This increases the quality of responses by ensuring that we have complete information about each respondent's demographics, likeliness to vote, political preferences, and issue evaluations. A trade-off that came with this decision is the risk of receiving fewer responses. Respondents cannot submit incomplete responses, so if they drop out close to the end, we will lose their entire response. If our survey's attrition, defined by Stantcheva (2023) as the rate at which respondents drop out of the survey before completing it, is high, then we will lose out on a lot of partial responses and have a low number of responses to work with. We could lose responses that contain answers to the demographic and vote preference questions.

Our second question, which asks respondents how likely they are to vote in the upcoming election, might be limited by social desirability bias. Social desirability bias happens when a respondent feels inclined to choose the option that they think would make them look good to others (Alexander 2024). Voting is viewed positively, so respondents who have no intention of voting might say that they will vote or are likely to vote. This would cause our sample of likely voters to include individuals who are unlikely to vote.

A.9 Idealized survey questions

Thank you for your interest in our 2024 U.S. Presidential Election Survey. The election will take place on Tuesday, November 5th, 2024 and it will decide the 47th President of the United States. Your participation is 100% voluntary and you can withdraw at any time, for any reason, with no questions asked.

This survey collects information about voters' political views and who they support in the 2024 U.S. Presidential Election. The data collected will not be shared with any external parties and will strictly be used for research purposes only. This survey is completely anonymous and your data will be protected. Any published material regarding the results drawn from this survey will not be traceable back to you. The goal of this survey is to conduct research about the upcoming presidential election. Please answer as accurately as possible.

Contact Information

If you have any questions or concerns about this survey or its methodology, please reach out via email to any of the following individuals:

- Aliza Mithwani: aliza.mithwani@mail.utoronto.ca
- Fatimah Yunusa: fatimah.yunusa@mail.utoronto.ca
- Talia Fabregas: talia.fabregas@mail.utoronto.ca

Correspondence will not be shared with any external parties.

Screening and Consent

By checking this box, I consent to this survey collecting personal information and information about my political views, my preferred 2024 presidential candidate, and who I voted for in 2020 for research purposes only.

- I consent.

1. Are you a registered voter in the United States ? Yes No
2. Which of the following describes your intention to vote or not in the upcoming presidential election? Yes, I will vote; No, I will not vote; I will probably vote; I probably will not vote; Undecided.

Demographics Questions

3. Would you consider yourself: White Black or African American Hispanic or Latino Asian American Indian or Alaskan Native Middle Eastern or North African Native Hawaiian or Pacific Islander Prefer not to say Other (specify)
4. What is your age? 18-29 20-44 45-64 65+ Prefer not to say
5. What sex were you assigned at birth, on your original birth certificate? “Female” “Male” Prefer not to say
6. How do you currently describe yourself (select all that apply)? “Female” “Male” “Transgender” Prefer not to say Other (Specify)
7. What is your household income? Less than $20,000 $20,000-59,999 $60,000-79,999 $80,000-99,999 $100,000 or more Prefer not to say
8. In which state do you currently reside? Dropdown list of states

Candidate Evaluation

9. If the 2024 presidential election were held today, who would you vote for?

Kamala Harris, Democrat Donald Trump, Republican Jill Stein, Green Party Write-in Don't know Prefer not to say Other (please specify)

(Optional) If you selected "Write-in" for the last question, please specify below:

10. Do you consider yourself a Democrat, a Republican, an Independent, or a member of another party? Democrat, Republican, Lean Democrat, Lean Republican, Independent Another Party Don't know Prefer not to say

Key Issues and Concerns

11. Rate these issues in order of importance to you (1 being most important and 7 being the least important): (use multiple choice grid with 7 rows and 7 columns) Abortion Immigration The state of democracy/corruption Foreign policy The economy Character Climate change?
12. If you had to assign a value from 1 to 5 to your level of optimism about the future of the United States, where 1 means highly pessimistic and 5 means highly optimistic, where would you place yourself? 1 (Highly pessimistic) 2 3 4 5 (Highly optimistic)
13. What would you say to someone who is undecided about voting in this election?

Confirmation Message

Thank you for your response. We appreciate the time, effort and honesty you put into answering this survey. Your answers have successfully been recorded and will be invaluable to our research!

Link to survey: https://forms.gle/h7MTA8k21ZbYxahT6

B Pollster methodology overview and evaluation

B.1 Overview

The Siena College/New York Times (Times/Siena) poll is a collaboration between the New York Times and the Siena College Research Institute that aims to capture voter sentiment in the battleground states and the nation overall (The New York Times 2024c). Its accuracy has made it a highly regarded poll for predicting U.S. presidential election results (Morris 2024b). The core concept of the Siena College/New York times survey methodology is obtaining a sample that is representative of the U.S. electorate as a whole (The New York Times 2024c).

We will evaluate the methodology of a New York Times/Siena College poll conducted in the seven battleground states (Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin) in late October 2024 (between October 24 and November 2, 2024) (The New York Times 2024b). We will refer to this poll as the Times/Siena late October 2024 battleground poll. In this poll, public opinion is transformed into data by contacting likely voters

in the seven battleground states from the L2 voter file via cell phone and using a structured survey with questions on candidate preference, key issues, and demographics (The New York Times 2024c). In this poll, public opinion is transformed into data by contacting likely voters from the L2 voter file via cell phone, using a structured survey with questions on candidate preference, key issues, and demographics.Each response is recorded as a data point. After collection, responses are adjusted through weighting to correct for non-response bias, ensuring the sample reflects the demographic makeup of the likely voter population (The New York Times 2024b). Once weighted, individual responses become entries in the dataset, which capture voter preferences and demographics for analysis. This structured dataset is then used to calculate support levels for each candidate and predict who will win the presidency.

B.2 Population and sampling frame

The target population, defined by Alexander (2024) is the population that Times/Siena polls aim to speak about, is registered voters and likely voters in the United States. Some polls conducted by Times/Siena have a target population of registered voters, while others use likely voters (The New York Times 2024a). The sampling frame, defined by Alexander (2024) as the individuals from the target population that Times/Siena can get data about is drawn from various voter registration databases, including the L2 voter database (The New York Times 2024c). The L2 database supplemented with phone number data from Marketing Systems Group to expand the survey's reach and contact as many likely voters as possible (The New York Times 2024b). In polls for the 2024 presidential election, Times/Siena has put an emphasis on gathering nationally-representative and state-specific samples for the seven battleground states that decided the 2020 presidential election and are expected to decide this election as well: Arizona, Nevada, Georgia, North Carolina, Michigan, Wisconsin, and Pennsylvania (The New York Times 2024c). A sampling frame that includes various voter registration databases helps ensure precise targeting and that the sample reflects the demographics of registered or likely voters in the upcoming U.S. election (The New York Times 2024c). The Times/Siena late October 2024 battleground poll is has a target population of likely voters in the seven battleground states and a sampling frame of voters in the L2 list of registered voters that live in those states (The New York Times 2024b).

B.3 Respondent recruitment

Times/Siena selects the respondents for its polls at random, from a national list of registered voters gathered from various high-quality voter registration databases (The New York Times 2024c). The list of registered voters contains demographic information for each voter; this allows Times/Siena to reach out to the right number of voters from each party, state, region, and race (The New York Times 2024b). Once selected, respondents are called from one of the call centers based in New York, Florida, Virginia, Texas, or South Carolina (The New York Times 2024c). It often takes multiple attempts to reach respondents, and overall, around

2% of the calls from pollsters are answered (The New York Times 2024c). A response rate of 2% appears low at first, but it is very common for high-quality polls to have low single-digit response rates (Silver 2024). Times/Siena pollsters limit the calls to 15 minutes to keep respondents engaged (The New York Times 2024c). In the late October 2024 battleground poll, Times/Siena made around 850,000 calls to over 320,000 voters across the seven battleground states (The New York Times 2024b).

B.4 Sampling approach

The Times/Siena late October 2024 battleground poll uses a response-rate-adjusted stratified sampling approach. Stratified sampling is a probability-based approach that involves dividing the sampling frame of voters in the seven battleground states into distinct strata (Alexander 2024) based on statehouse district, party, race, gender, marital status, household size, turnout history, age and home ownership (The New York Times 2024b). Voters were then randomly selected from each stratum to get a sample that is representative of likely voters in the seven battleground states. The sample was adjusted for unequal probability of selection via stratum, to account for the probability that a respondent will actually vote in the 2024 election, and to match targets for the likely composition of the 2024 electorate (The New York Times 2024b). Assigning higher weightings to groups that were underrepresented in the sample was used to try to correct the non-response bias caused pollsters to underestimate support for Trump in 2016 and 2020 [The New York Times (2024b)].

The final sample included 7,879 likely voters from the seven battleground state [The New York Times (2024b)]. The breakdown is as follows: 1,025 from Arizona; 1,010 from Nevada; 1,004 from Georgia; 1,010 from North Carolina; 1,305 from Wisconsin; 998 from Michigan; 1,527 from Pennsylvania (The New York Times 2024b). The New York Times (2024b) oversampled and included a supplemental sample of unregistered voters from Wisconsin because the state has a history of significant election day voter registration.

B.5 Strengths and limitations

The Times/Siena late October 2024 poll's strengths include its focus on the key battleground states and its use of stratified sampling. Focusing on the seven key battleground states instead of the popular vote allows this poll to forecast the winner of the election. Effective use of stratified sampling can correct non-response bias and under-representation of Trump voters in the sample by weighting each strata according to the general population and increasing the weight of groups that were less likely to respond (Clinton 2024). Times/Siena polls do not use weighting by recalled vote to address non-response bias and correct polling misses from 2016 and 2020 (Cohn 2024). Times/Siena polls do not use weighting by recalled vote. In Section 5.1, we reported that weighting by recalled vote would have made polls from every election since 2004 less accurate and overestimated support for the losing candidate.

The Times/Siena late October 2024 poll's weaknesses and limitations include non-response bias, social desirability bias, coverage error, late shifts among undecided voters, and error in its estimation of the demographic composition of the electorate (The New York Times 2024b). Stratified sampling relies on the assumption that the views of the voters who participated in the survey align with the views of voters from similar demographics who did not participate in the survey (Clinton 2024). Even voters within the hardest-to-reach, least likely to respond demographics could be more likely to respond if they favor Harris than their counterparts who favor Trump, or vice versa. In a very close election cycle like this one, it is possible that the changes in public opinion in the seven battleground states that this survey presents are caused by the change in individuals who responded to this poll in late October and a similar poll at an earlier date (Gelman et al. 2016). Voting is socially desirable behavior, so there is a chance that this sample includes respondents who overestimated or overstated how likely they are to vote. This poll has a margin of error of plus or minus 3.6 percentage points in Arizona, 4.1 points in Georgia, 3.7 points in North Carolina, 3.9 points in Michigan, 3.8 points in Nevada, 3 points in Pennsylvania, and 3.7 points in Wisconsin in its sample of likely voters (The New York Times 2024b). This is consistent with the average polling misses of 3 to 4 percentage points reported by Silver (2024).

C Additional data details

C.1 Data cleaning

We used R (R Core Team 2023) and the dplyr (Wickham et al. 2023) and janitor (Firke 2023) packages to clean the raw presidential poll data that we obtained from FiveThirtyEight (2024). Our raw dataset contained polling data going back to 2021. We sub-setted it to only include polling data from May 5, 2024 and later (the six-month period leading up to the 2024 presidential election). However, President Biden ended his campaign on July 21, 2024 and Vice President Harris became the nominee, so we sub-setted it further to only include polling data from July 21, 2024 and later with a numeric grade of 3 that had a population of likely voters. We then filtered to only include polling data for Harris and Trump, since they are the two major party candidates. We did not use third-party candidate polling data.

We selected following variables (columns) to include in our analysis data:

- `poll_id`: the unique id of a poll.
- `pollster`: the pollster who conducted the poll.
- `numeric_grade`: the grade that the poll received. We only considered polls with a numeric grade of 3.0 or higher (Morris 2024b).
- `pollscore`: the Predictive Optimization of Latent skill Level in Surveys, Considering Overall Record of a poll (Morris 2024a).

- `state`: the state where the poll was conducted or NA if it is a national poll. We used the `mutate` function in base R (R Core Team 2023) to make state equivalent to "National" in rows where it was initially NA.
- `end_date`: the date that the poll was completed
- `population`: the narrowest category included in the poll's sample (likely voters < registered voters < all adults). We considered polls where the population is likely voters; this means that these polls included likely voters, but may have also included registered voters and all adults (FiveThirtyEight 2024).
- `party`: the party whose support that row of the dataset shows
- `candidate_name`: the candidate whose support that row of the dataset shows
- `pct`: the percentage of support that candidate with name `candidate_name` from `party` received in the poll with unique id `poll_id`

We constructed the `end_date_num` variable using the `as.Date()` function from base R (R Core Team 2023). `end_date_num` is equivalent to the number of days since President Biden ended his re-election campaign on July 21, 2024.

C.2 Polling averages from before President Joe Biden ended his re-election campaign

It is impossible to quantify or model the effects of Biden's unprecedented decision to end his re-election campaign in July 2024. There is no established methodology for analyzing this, so we made the decision to focus on Harris versus Trump polling from after July 21, 2024 because they are the two major presidential nominees. National and swing-state polling averages in the six months leading up to election day are shown in Figure 6 and Figure 7.

D Model details

The model summaries are shown in Table 6

D.1 Posterior predictive check

D.1.1 Harris Model

In Figure 8a we implement a posterior predictive check. This shows the comparison of the actual outcome variable, *pct_harris*, with simulations from the posterior distribution (Alexander 2024). Figure 8b shows a comparison of the posterior with the prior.

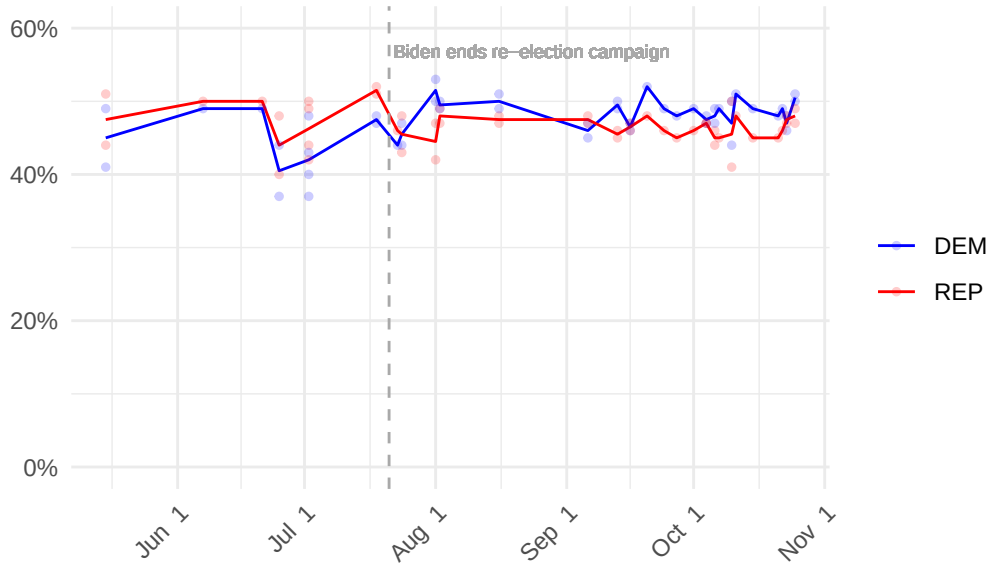


Figure 6: National popular vote averages for the Democratic and Republican presidential nominees since May 5, 2024 (six months before election day) show former President Trump leading before Biden ended his campaign, the gap narrowing after Harris became the Democratic nominee, Harris taking the lead in August, and a dead heat in September and October.

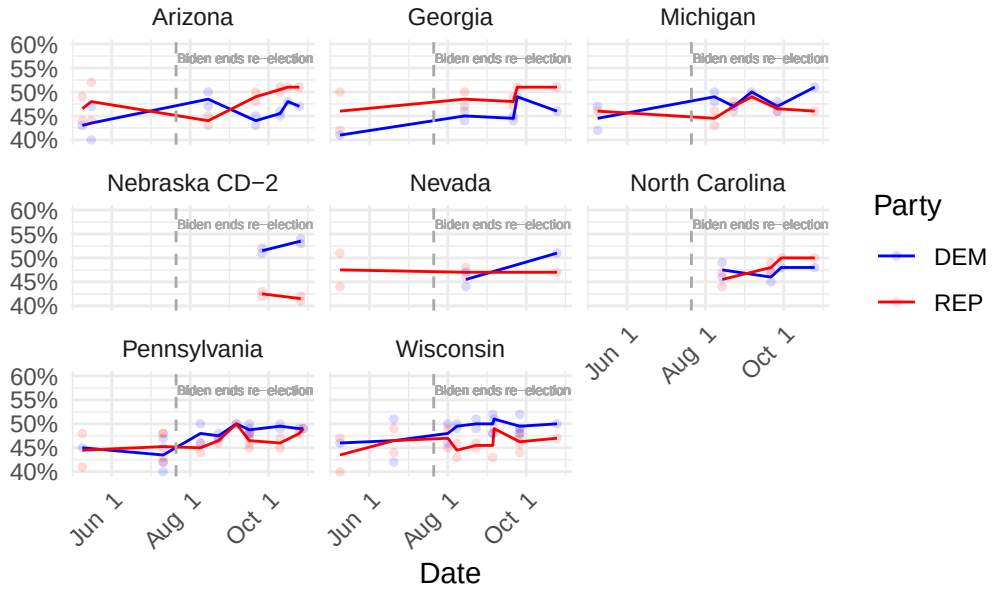
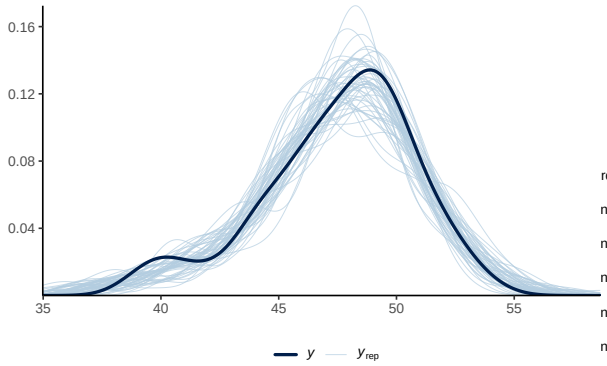


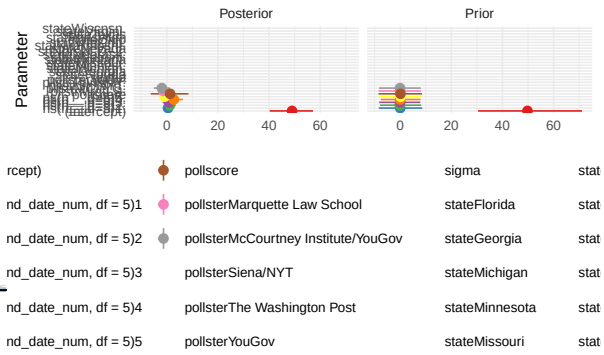
Figure 7: Democratic vs Republican presidential polling averages from May 5 to October 29, 2024, including both Biden and Harris as the Democratic nominee and Trump as the Republican nominee.

Table 6: Explanatory models of support for Harris and Trump based on end date (number of days since Biden ended re-election campaign on July 21, 2024), state, pollster, and pollscore

	Harris	Trump
(Intercept)	48.99	48.59
	(5.02)	(5.21)
ns(end_date_num, df = 5)1	0.49	0.96
	(1.10)	(1.04)
ns(end_date_num, df = 5)2	1.25	−0.58
	(1.41)	(1.44)
ns(end_date_num, df = 5)3	0.83	0.59
	(1.47)	(1.45)
ns(end_date_num, df = 5)4	2.74	−0.98
	(2.22)	(2.11)
ns(end_date_num, df = 5)5	−0.87	2.45
	(1.10)	(1.05)
stateFlorida	−4.13	4.45
	(1.19)	(1.21)
stateGeorgia	−0.81	0.61
	(0.93)	(0.96)
stateMichigan	1.57	−2.18
	(0.86)	(0.88)
stateMinnesota	4.10	−5.70
	(1.82)	(1.79)
stateMissouri	−5.98	4.82
	(1.82)	(1.79)
stateMontana	−5.88	7.63
	(1.41)	(1.37)
stateNational	1.66	−2.81
	(0.71)	(0.72)
stateNebraska	−5.55	4.29
	(1.46)	(1.55)
stateNebraska CD-2	6.75	−6.76
	(1.09)	(1.12)
stateNevada	0.42	−0.91
	(1.20)	(1.16)
stateNew Hampshire	3.28	−3.88
	(1.75)	(1.87)
stateNorth Carolina	0.74	−0.64
	(0.96)	(0.93)
stateOhio	−2.29	1.41
	(1.10)	(1.09)
statePennsylvania	2.06	−2.19
	(0.80)	(0.80)
stateTexas	−2.80	1.53
	(0.95)	(0.97)
stateVirginia	3.22	−4.82
	(1.82)	(1.91)
stateWisconsin	2.59	−2.29
	(0.85)	(0.85)
pollsterMarquette Law School	−1.34	−0.08
	(1.02)	(1.01)
pollsterMcCourtney Institute/YouGov	−1.93	−0.70
	(1.95)	(1.84)
pollsterSiena/NYT	−2.43	−0.12
	(1.98)	(1.95)
pollsterThe Washington Post	−1.71	1.22
	(1.28)	(1.26)
pollsterYouGov	−1.34	1.31
	(0.88)	(0.85)
pollscore	1.29	−0.01
	(4.44)	(4.54)
Num.Obs.	117	117
R2	0.715	0.697
R2 Adj.	0.629	0.603
Log.Lik.	−224.988	−224.011
ELPD	−249.5	−248.2
ELPD s.e.	8.2	7.6
LOOIC	499.0	496.3
LOOIC s.e.	16.3	15.2
WAIC	495.1	492.1
RMSE	1.57	1.56



(a) Posterior prediction check

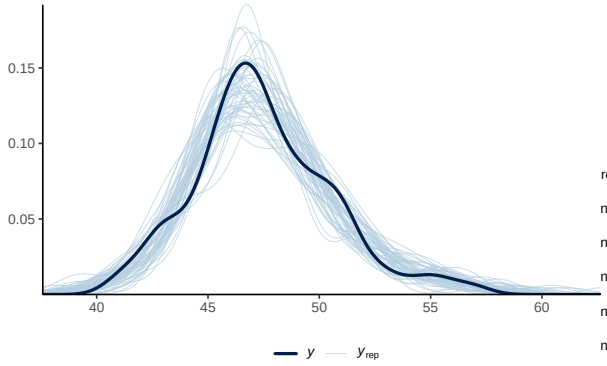


(b) Comparing the posterior with the prior

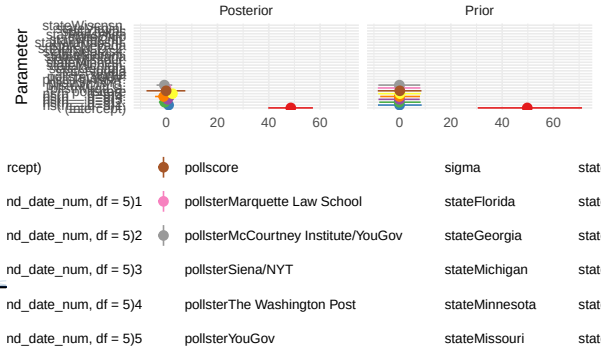
Figure 8: Examining how the model for support for Harris fits, and is affected by, the data

D.1.2 Trump Model

In Figure 9a we implement a posterior predictive check. It shows the comparison of the actual outcome variable, *pct_trump*, with simulations from the posterior distribution (Alexander 2024). Figure 9b shows a comparison of the posterior with the prior.



(a) Posterior prediction check



(b) Comparing the posterior with the prior

Figure 9: Examining how the model for support for Trump fits, and is affected by, the data

D.2 Distribution

Credibility intervals are the equivalent of confidence intervals when using a Bayesian model (Alexander 2024). We use Bayesian estimation to get a distribution for each coefficient. Figure 10 shows the 95% credibility intervals for the coefficients in the Harris model. Figure 11 shows the 95% credibility intervals for the coefficients in the Trump model. There is a 95% probability mass between the two ends of the 95% credibility interval (Alexander 2024).

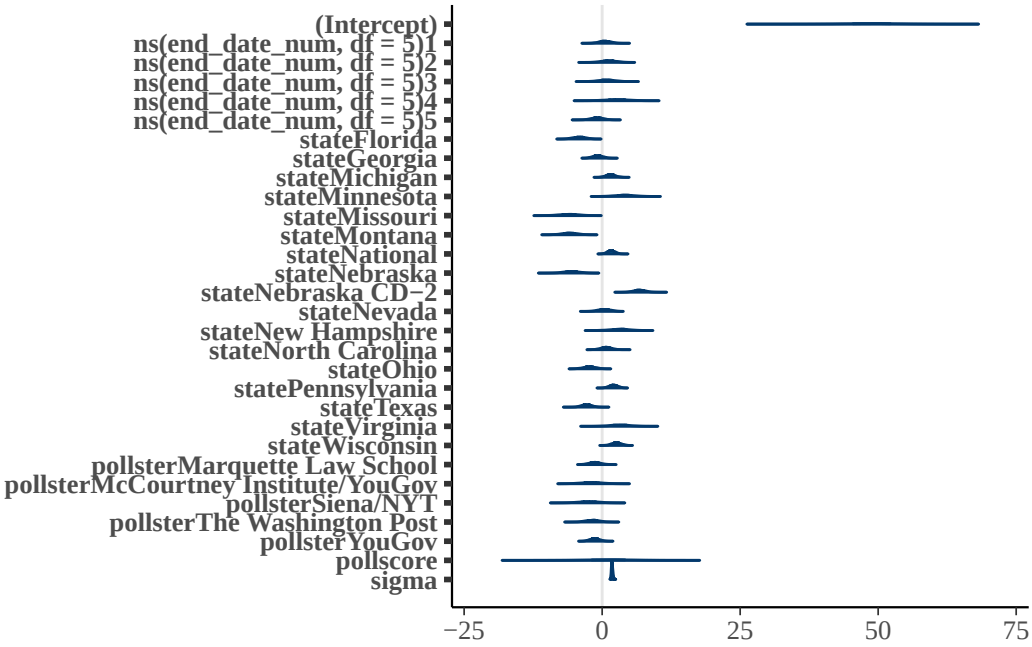


Figure 10: 95% credibility intervals for the Harris model

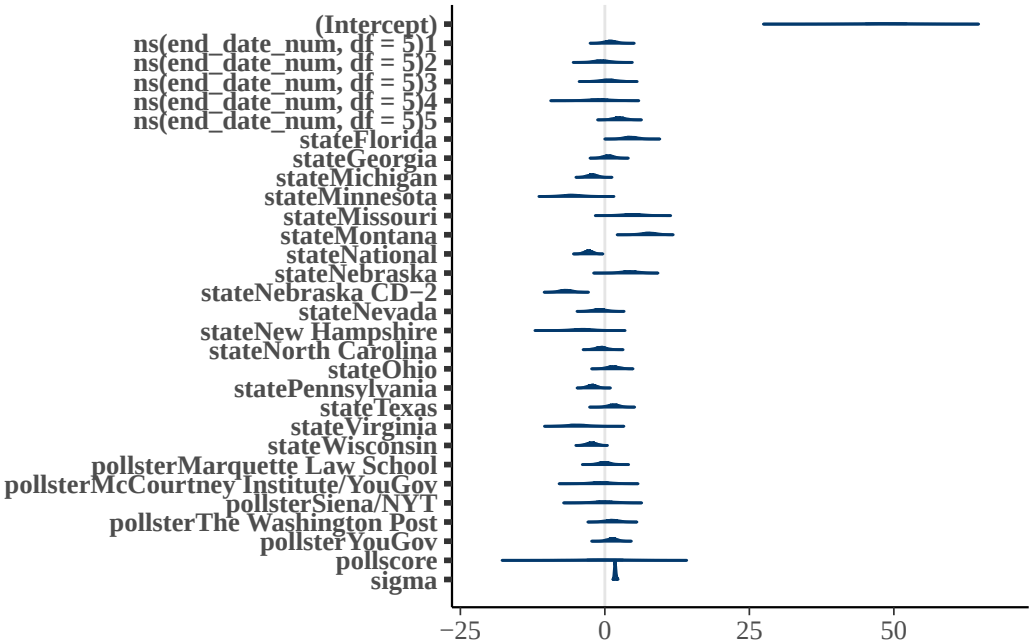


Figure 11: 95% credibility intervals for the Trump model

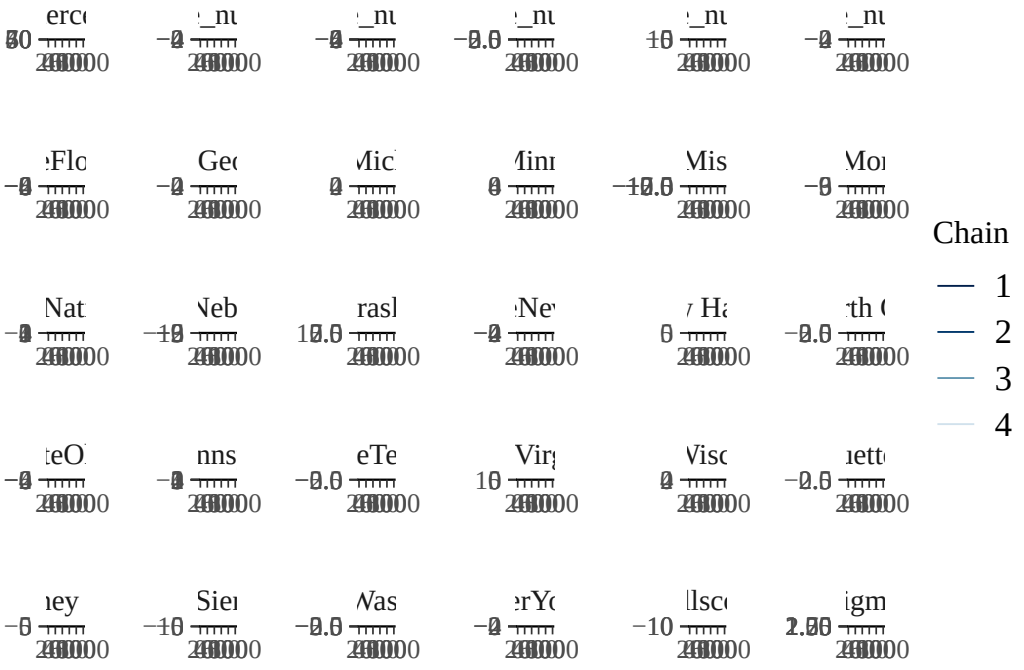
D.3 Model Diagnostics

D.3.1 Harris Model

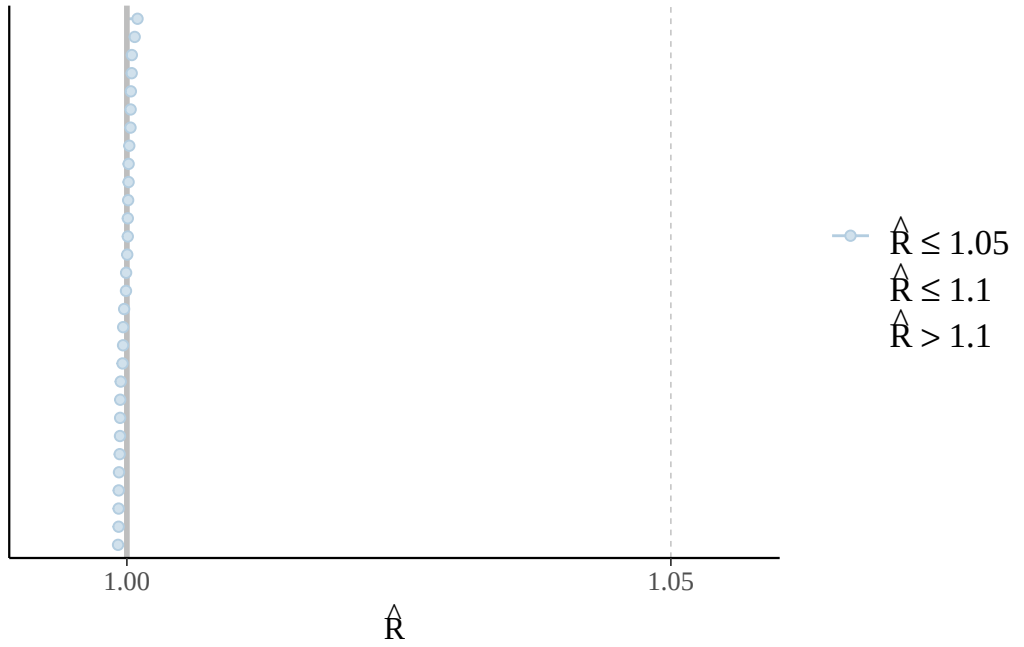
Figure 12a is a trace plot. It shows lines that bounce, are horizontal, and have some overlap between the chains. This suggests that there is nothing out of the ordinary with the Harris model (Alexander 2024). Figure 12b is a Rhat plot. It shows that everything is very close to 1. This suggests that there are no problems with our Harris model and we do not need to simplify it, remove predictors, modify predictors, use different priors, or re-run (Alexander 2024).

D.3.2 Trump Model

Figure 13a is a trace plot. It shows lines that bounce, are horizontal, and have some overlap between the chains. This suggests that there is nothing out of the ordinary with our Trump model (Alexander 2024). Figure 13b is a Rhat plot. It shows that everything is very close to 1. This suggests that there are no problems with our Trump model and we do not need to simplify it, remove predictors, modify predictors, use different priors, or re-run it (Alexander 2024).

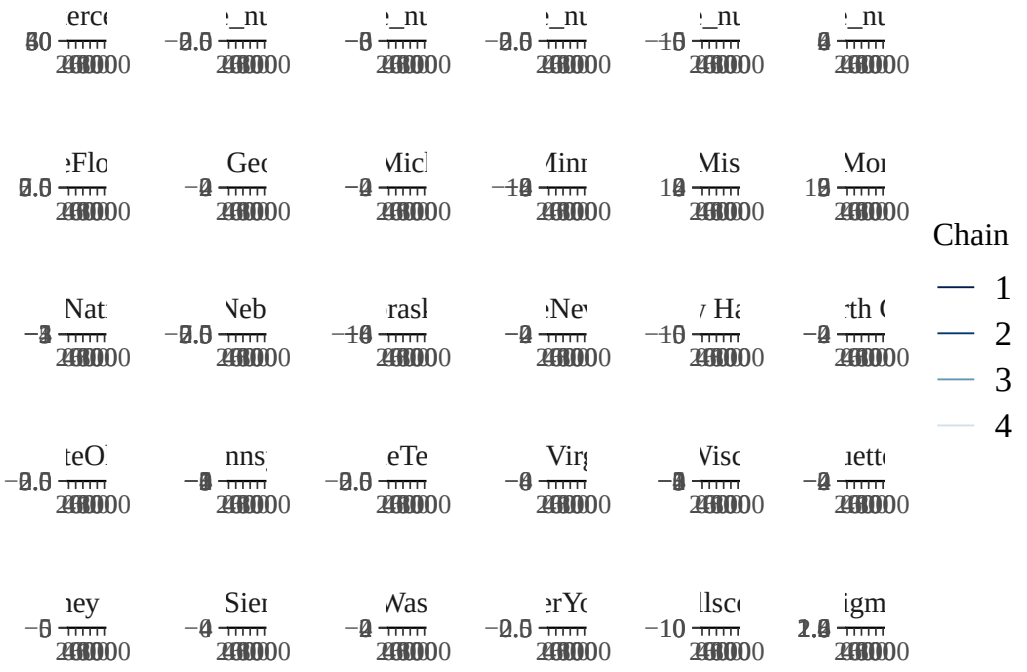


(a) Trace plot

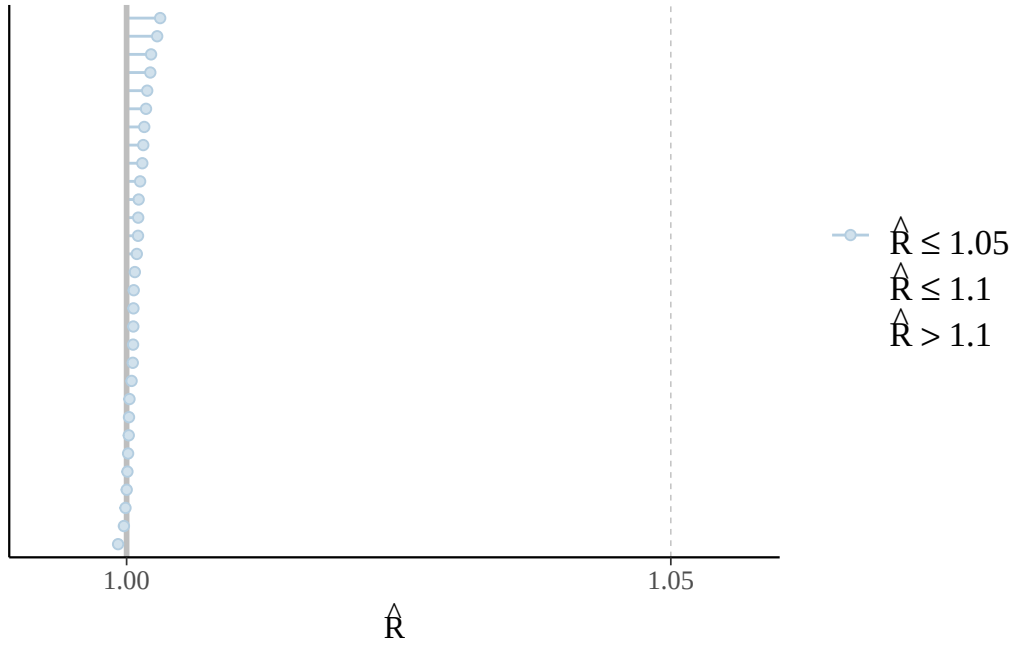


(b) Rhat plot

Figure 12: Checking the convergence of the MCMC algorithm for the Harris model



(a) Trace plot



(b) Rhat plot

Figure 13: Checking the convergence of the MCMC algorithm for the Harris model

References

270 to Win. 2024. "2024 Presidential Election Interactive Map." https://www.270towin.com/.

Alexander, Rohan. 2024. *Telling Stories with Data*. "University of Toronto". https://www.tellingstorieswithdata.com.

Clinton, Josh. 2024. "Poll Results Depend on Pollster Choices as Much as Voters' Decisions." Good Authority. https://goodauthority.org/news/election-poll-vote2024-data-pollster-choices-weighting/.

CNN. 2016b. "Presidential Results." https://www.cnn.com/election/2016/results/president.

———. 2016a. "Presidential Results." https://www.cnn.com/election/2020/results/president.

Cohn, Nate. 2024. "How One Polling Decision Is Leading to Two Distinct Stories of the Election." https://www.nytimes.com/2024/10/06/upshot/polling-methods-election.html.

Durand, Claire, Courtney Kennedy, Mark Blumenthal, Scott Clement, Joshua Clinton, Charles Franklin, Kyley McGeeney, and Lee Miringoff. 2017. "An Evaluation of 2016 Election Polls in the United States." https://www.researchgate.net/publication/317170048_AN_EVALUATION_OF_2016_ELECTION_POLLS_IN_THE_U_NITED_STATES.

Firke, Sam. 2023. *janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

FiveThirtyEight. 2024. "National Presidential Polls, 2024." https://projects.fivethirtyeight.com/polls/national/.

Gelman, Andrew, Sharad Goel, Douglas Rivers, and David Rothschild. 2016. "The Mythical Swing Voter." https://doi.org/10.1561/100.00015031.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm/.

Kennedy, Courtney, Jesse Lopez, Scott Keeter, Arnold Lau, Nick Hatley, and Nick Bertoni. 2021. "Confronting 2016 and 2020 Polling Limitations." https://www.pewresearch.org/methods/2021/04/08/confronting-2016-and-2020-polling-limitations/.

Kolenikov, Stas. 2016. "Post-Stratification or Non-Response Adjustment?" https://doi.org/10.29115/SP-2016-0014.

L2. 2024. "DataMapping: Winning Strategies Start with the Best Voter Records." L2 Better Data Better Decisions. https://www.l2-data.com/datamapping/.

Mercer, Andrew, and Arnold Lau. 2023. "Comparing Two Types of Online Survey Samples." https://www.pewresearch.org/methods/2023/09/07/comparing-two-types-of-online-survey-samples/.

Morris, G. Elliott. 2024a. "How 538's Pollster Ratings Work." ABC News. https://abcnews.go.com/538/538s-pollster-ratings-work/story?id=105398138.

———. 2024b. "What Are the Best Pollsters in America." ABC News. https://abcnews.go.com/538/best-pollsters-america/story?id=105563951.

Neyman, Jerzy. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97 (4): 558–625. http://www.jstor.org/stable/2342192.

Office of Institutional Reseaerch. 2024. "How Do i Analyze My Survey Results." https://ir.wfu.edu/large-survey/how-do-i-analyze-my-survey-results/.

PASEK, JOSH. 2015. "Predicting Elections: Considering Tools to Pool the Polls." *The Public Opinion Quarterly* 79 (2): 594–619. http://www.jstor.org/stable/24546379.

Pew Research Centre. 2024. "Harris, Trump Voters Differ over Election Security, Vote Counts and Hacking Concerns." Pew Research Centre Election 2024. https://www.pewresearch.org/politics/2024/10/24/harris-trump-voters-differ-over-election-security-vote-counts-and-hacking-concerns/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Raifman, Sarah, Michelle DeVost, Jean Digitale, Yea-Hung Chen, and Megan Morris. 2021. "Respondent-Driven Sampling: A Sampling Method for Hard-to-Reach Populations and Beyond." *Springer Nature Link* 9. https://doi.org/https://doi.org/10.1007/s40471-022-00287-8.

Silver, Nate. 2024. "Nate Silver: Here's What My Gut Says about the Election, but Don't Trust Anyone's Gut, Even Mine." https://www.nytimes.com/2024/10/23/opinion/election-polls-results-trump-harris.html?unlocked_article_code=1.UU4.pFkQ.F2hD-woxmiEj&smid=url-share.

Smith, Michael, Maryam Witte, Sarah Rocha, and Mathias Basner. 2019. "Effectiveness of Incentives and Follow-up on Increasing Survey Response Rates and Participation in Field Studies," no. 230. https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-019-0868-8.

Stantcheva, Stefanie. 2023. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." *Annual Review of Economics* 15: 205–34. https://www.annualreviews.org/content/journals/10.1146/annurev-economics-091622-010157.

The New York Times. 2024a. *Cross-Tabs: Late October 2024 Times/Siena Poll of the Likely Electorate*. The New York Times. https://www.nytimes.com/interactive/2024/10/26/us/elections/times-siena-nyc-poll-registered-voter-crosstabs.html.

———. 2024b. "Cross-Tabs: October 2024 Times/Siena Poll of the Likely Electorate in the 2024 Battlegrounds." https://www.nytimes.com/interactive/2024/11/03/us/elections/times-siena-battleground-poll-crosstabs.html.

———. 2024c. "You Ask, We Answer: How the Times/Siena Poll Is Conducted." The New York Times. https://www.nytimes.com/article/times-siena-poll-methodology.html#link-2fc0eb29.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Zhang, Ziyi, Shoufei Zhu, Jaron Mink, Aiping Xong, Linhai Song, and Gang Wang. 2019. "Beyond Bot Detection: Combatting Fraudulent Online Survey Takers." https://gangw.cs.illinois.edu/www22-bot.pdf.