Forecasting the 2024 U.S. Presidential Election*

Using Aggregated Polling Data to Predict the Outcome of the 2024 Presidential Race

Andy Jiang Onon Burentuvshin Rui Hu

November 4, 2024

The 2024 U.S. presidential election is shaping up to be a highly swing race, with polling data playing a crucial role in gauging public sentiment and forecasting the potential outcome. This paper aims to build a predictive model using a poll-of-polls approach, aggregating data from multiple pollsters to forecast the winner of the upcoming election. The analysis incorporates polling data from various sources, filtered to include only high-quality pollsters with a numeric grade above 3, to ensure the reliability of the predictions. Additionally, a detailed examination of one selected pollster's methodology is conducted, highlighting strengths, weaknesses, and biases in the survey process. The paper concludes with an idealized polling methodology, designed within a $100K budget, to enhance the accuracy of election forecasts. Results are presented in the context of understanding polling trends, potential biases, and the evolving landscape of voter preferences as the election approaches.

1 Introduction

The 2024 U.S. presidential election is shaping up to be one of the most closely watched and potentially pivotal elections in recent history. As public opinion shifts and political dynamics evolve, polling data has emerged as a critical tool for forecasting the outcome and understanding voter preferences. Polls provide a snapshot of where candidates stand at various points in time, reflecting the impact of campaigns, debates, and major events. However, the reliability and accuracy of individual polls can vary significantly based on methodology, sample size, and other factors, making it necessary to aggregate multiple sources of data for a more robust prediction.

*Code and data are available at: [https://github.com/AndyYanxunJiang/president-polls).

This paper aims to forecast the outcome of the 2024 U.S. presidential election by leveraging a "poll-of-polls" approach, which aggregates data from various pollsters to create a more comprehensive picture of the race. By focusing on high-quality polls and conducting a detailed analysis of the polling trends for the two main candidates, Donald Trump and Kamala Harris, we seek to uncover patterns in voter support and potential factors influencing shifts in public opinion. Additionally, the paper examines the methodologies of different pollsters to assess their strengths and weaknesses, which can help identify possible biases in the aggregated data.

The analysis will further explore the geographic distribution of support across key swing states, providing insights into the electoral landscape. Finally, the paper presents an idealized polling methodology with a $100,000 budget, designed to improve the accuracy and representativeness of election forecasts. By integrating multiple data sources and critically evaluating polling practices, this study aims to contribute to the ongoing discussion on the role of polling in democratic processes and election forecasting.

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023).

2.2 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

2.3 Outcome variables

Add graphs, tables and text. Use sub-sub-headings for each outcome variable or update the subheading to be singular.

Table 1: Sample of 10 Random Observations from the Ana

pollster	state	start_date	end_date	question_id	sample_size	population	party
Morning Consult	Pennsylvania	7/24/24	7/28/24	204799	804	rv	DEM
Morning Consult	Georgia	9/19/24	9/25/24	210918	989	rv	DEM
Morning Consult	Nevada	2/12/24	2/18/24	193849	445	rv	REP
Civiqs	Pennsylvania	7/13/24	7/16/24	203589	536	rv	DEM

Florida Atlantic University/Mainstreet Research	Nevada	5/19/24	5/21/24	199917	522	rv	REP
Siena/NYT	Arizona	10/22/23	11/3/23	184803	603	rv	REP
Siena/NYT	Wisconsin	4/28/24	5/9/24	198693	614	rv	REP
Franklin and Marshall College	Pennsylvania	8/9/23	8/20/23	180004	723	rv	REP
Morning Consult	Wisconsin	7/24/24	7/28/24	204800	700	rv	DEM
KAConsulting LLC	Wisconsin	5/15/24	5/19/24	203029	600	rv	REP

2.4 Predictor variables

3 Model

In this analysis, we utilize a logistic regression model to predict polling percentages for the 2024 U.S. presidential candidates over time. In order to focus on states with historical significance in elections. The model is specified as a Generalized Linear Model (GLM) with a binomial family and a logit link function, appropriate for modeling binary or proportion-based outcomes.

3.1 Model set-up

The dataset, model_data.csv, provides polling data with average_pct as the observed average polling percentage among 7 swing states. The end_date indicates the polling period's endpoint. To predict polling percentages beyond the available data, we created an additional sequence of 20 days beyond the dataset's last recorded date. Because we want to predict the result in November 5. This expansion allows for forecasting near-term polling trends. Also, capturing potential shifts in candidate support leading up to election day. The model was trained on existing data and applied to this extended date range using predict(model, newdata = ., type = "response").

Our model setup and variable selection are as follows:

$$\text{logit}(p_i) = \beta_0 + \beta_1(\text{Date}) + \beta_2(\text{Candidate})$$

where:

- p_i represents the probability of support for candidate i at a specific date.

- β_0 is the intercept, representing baseline support.
- β_1 captures changes in support over time (date effect).
- β_2 is a categorical variable accounting for each candidate.

3.1.1 Model justification

The choice of a logistic regression model was guided by the nature of polling data. Nature of polling data represents proportions that naturally bound between 0 and 1. The logistic function's S-shaped curve effectively models the relationship between time and polling percentage, accommodating non-linear trends that may arise as election day nears. Firstly, candidate names should be included as a categorical predictor in the model. It allows us to capture individual impact in support rates by candidate. In addition, a transformation from date to numeric is made on end_date in order to get a numerical format. It is called the end_date_numeric. It helps put date as a predictor in the model. By structuring the model this way, it can predict the percentage of the poll by both the impact of candidates and trends over time.

4 Results

The analysis of polling data across key swing states for the 2024 U.S. Presidential Election reveals insights into the levels of support for candidates Donald Trump and Kamala Harris. The findings are presented through three primary visualizations: the average polling percentage by state, the distribution of polling percentages, and the difference in average polling percentage between the two candidates.

4.1 Average Polling Percentage by State

From Figure 1, we can see that in the swing states of Arizona, Georgia, and North Carolina, Trump holds a narrow lead. In contrast, Harris leads in Pennsylvania, Wisconsin, Nevada, and Michigan. When looking at the overall average across these states, Harris slightly edges out Trump, indicating her overall advantage in these swing areas.

4.2 Polling Percentage Distribution

The box plot in Figure 2 demonstrates the range and consistency of support for both candidates across each state. Trump shows a wider range of support, reflecting variability, while Harris has relatively steady polling percentages across states, maintaining higher consistency. Harris's support is less variable and generally centers around higher polling percentages compared to Trump's.

4.3 Difference in Average Polling Percentage

Figure 3 highlights the difference in average polling percentages by state, with positive values indicating Trump's lead and negative values indicating Harris's. Trump's lead in the three states (Arizona, Georgia, and North Carolina) is relatively narrow, with margins of less than 1 percent. In contrast, Harris's lead in the states favoring her (Pennsylvania, Wisconsin, Nevada, and Michigan) ranges from 1 to 4 percent, indicating her stronger support in the states where she leads.

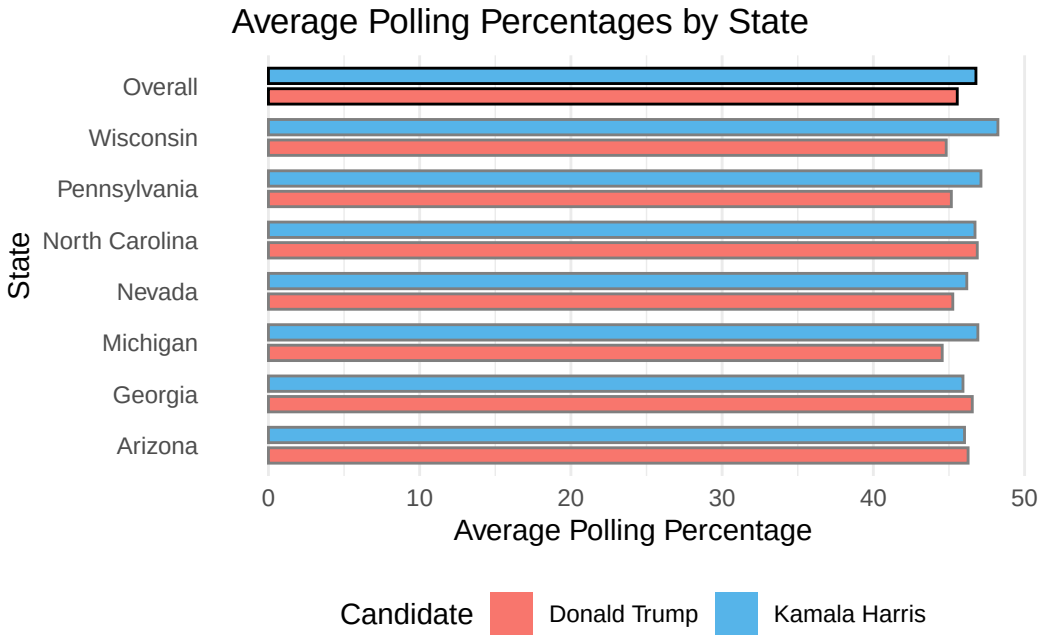


Figure 1: Average Polling Percentage by State

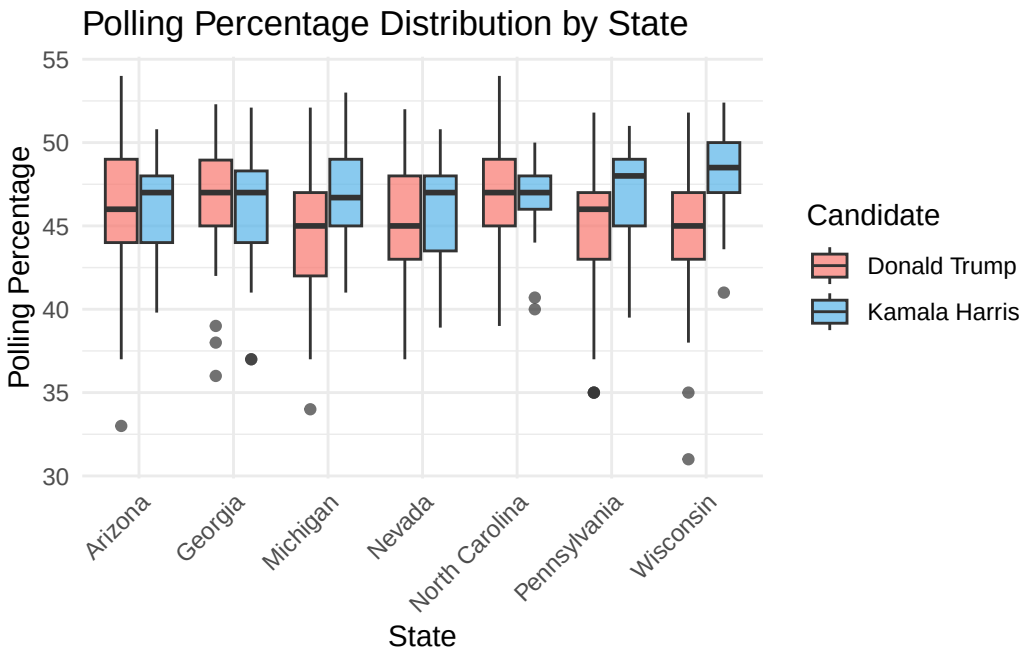


Figure 2: Polling Percentage Distribution

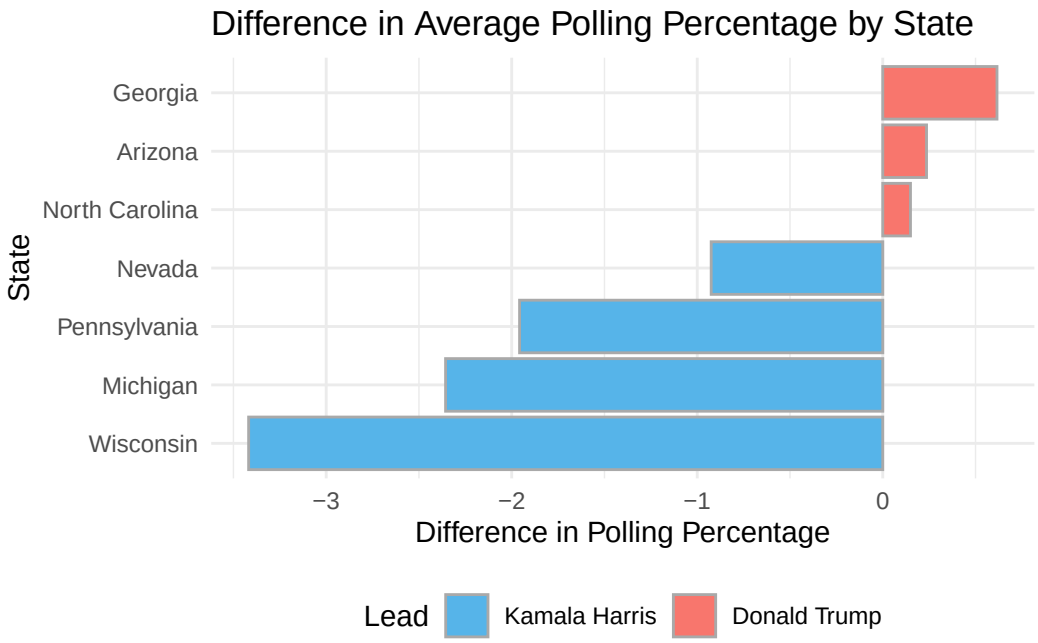


Figure 3: Difference in Average Polling Percentage

4.4 Swing State Polling Prediction Model

Figure 4 shows the trend of support rate changes for each candidate before the upcoming election, as well as the model's predicted values for the next 20 days.We predicts that Donald Trump may win based on our model.

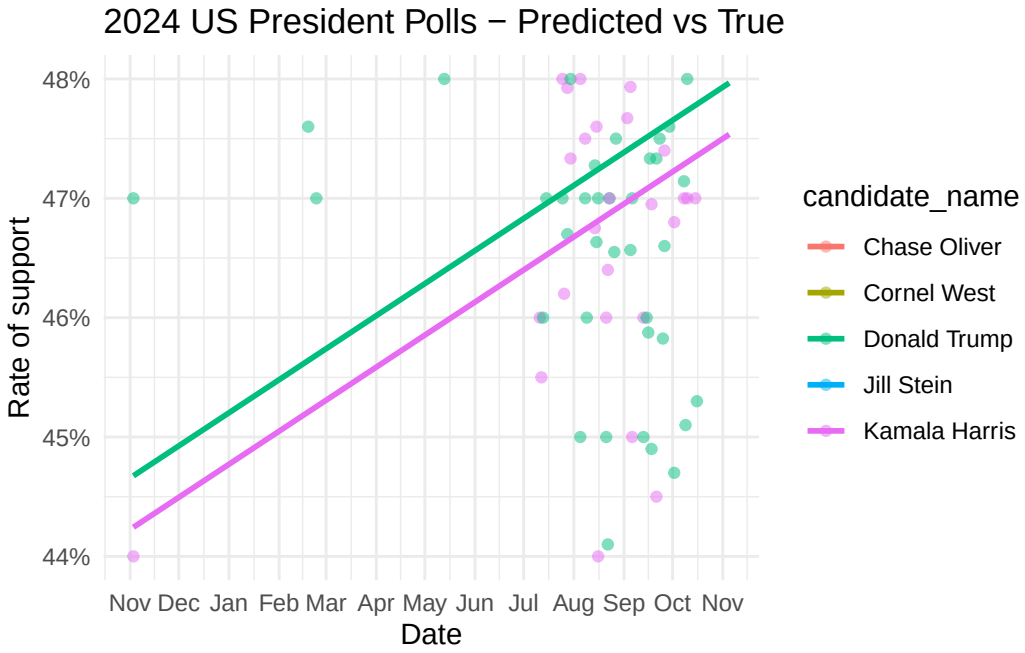


Figure 4: Swing State Polling Prediction Model

5 Discussion

5.1 Impact of Key Demographics and Policy Appeals

Harris's lead in several swing states may be influenced by her campaign's appeal to key demographic groups, particularly immigrants. The Democratic platform has emphasized social and economic benefits for immigrant communities, which may have strengthened her base in states with significant immigrant populations. Policies that promise support for healthcare, education, and other public services are likely contributing to Harris's appeal among these groups, solidifying her lead in states like Nevada and Michigan. This demographic support provides Harris with a strategic advantage, as these communities may play a crucial role in voter turnout and loyalty.

5.2 Influence of Endorsements and High-Profile Backing

Recently, high-profile figures like Elon Musk have publicly backed Trump, using their influence to mobilize support for his campaign. Musk's endorsement may energize Trump's voter base, especially among his followers and individuals interested in business and technology. This kind of visible support from influential figures could affect last-minute voter decisions, particularly in tightly contested states where Trump's lead is slim. If Musk's endorsement translates into increased campaign funding or social media influence, it could bolster Trump's standing in swing states where the margin remains close, potentially altering the final outcome on November 5th.

5.3 Limitations of the Analysis: Focus on Swing States

While the analysis focuses on key swing states, it does not consider the broader national picture, which includes states with clear, one-sided advantages for either candidate. This narrowed focus on swing states provides insights into swing areas but does not give a complete picture of the national landscape. Therefore, while Harris appears to have an edge in these swing states, it is unclear whether she is leading overall. The situation in solidly red or blue states could still shift the electoral balance, meaning Trump could potentially gain ground through strong performances in traditionally Republican states.

5.4 Potential Influence of Late Campaign Efforts

With a few days remaining before the election, both campaigns are likely to intensify their outreach efforts, particularly in swing states. Trump's recent increase in targeted advertisements and campaign rallies in swing states could energize his supporters, narrowing Harris's lead. On the other hand, Harris's campaign may focus on mobilizing key demographics, such as young and minority voters, to secure her advantage. Late-breaking events, public debates, or controversies could sway undecided voters and ultimately impact the election results in these states.

5.5 Limitations and Next Steps

This analysis is limited to examining current polling data in swing states and does not account for unexpected shifts in voter sentiment or turnout on election day. Additionally, focusing exclusively on swing states omits the influence of non-swing states that may still play a role in the electoral vote tally. Future analyses could incorporate national polling data to provide a more comprehensive perspective and examine demographic breakdowns within each state to understand which groups are driving support for each candidate. Integrating data on recent campaign activities, endorsements, and voter sentiment on pressing issues could also improve the predictive accuracy of election outcomes.

Appendix A

A Sampling Methodology Employed by New York Times and Siena College:

Our analysis and prediction utilizes data collected by a single pollster and it is important to scrutinize the sampling methodology behind the sampling methods used to determine the extent of the external validity of our predictions. Primarily, the data gathered by New York Times was done so with the collaboration with Siena College through live-interviewers reaching voters by telephone. The poll targets registered voters in the United States, with a particular focus on voters in swing states for presidential races. Consequently, the target population is registered voters of the United States of America and the sampling frame is the national list of registered voters from which NYC employs random sampling to create their pollsters. And from this sampling frame, respondents are randomly selected while balancing key demographic and political characteristics through both front-end sampling and post-survey weighting pursuant to appropriate representation of the entire target population. The poll relies on random digit dialing to reach voters, predominantly via cell phones. While landlines are also called, the dominance of cellphone usage aligns with contemporary communication trends. This results in their rendition of random sampling, which contributes to the external validity of data. However, though the sample was selected randomly, reaching voters by phone has become more difficult due to low response rates requiring many attempts to secure responses from certain individuals (response rates are as low as 2% at NYC). Despite this, the phone methodology is still seen as effective for its ability to quickly reach a diverse voter pool across various states as well as seen as the golden standard for polling. Although this can be seen as random sampling, the concern of low response rates and reliance on dialing to reach voters raises concerns of selection bias. There are questions as to whether a certain demographic will be more responsive to being reached through a cell phone. Moreover, voters that are reached over the phone may feel the urge to provide more 'socially desirable' responses that are not reflective of their true opinions. With a response rate of 2%, there is valid concern over non-response bias underrepresenting certain demographics. The poll addresses non-response through a combination of front-end sampling adjustments (calling harder-to-reach voters) and post-survey weighting. Weighting is applied to ensure that the final sample aligns with the demographic and political composition of the electorate, based on voter file data. Non-response bias is a persistent challenge, and even though weighting helps, it's impossible to be certain that respondents are fully representative of those who do not respond. Experiments with alternative methods, such as mail surveys with incentives of cash vouchers have shown to increase response rate up to 30% (New York Times).

B Sampling Approach: Stratified Sampling

B.1 Methodology

The Times/Siena Poll employs a stratified random sampling approach to ensure that the sample accurately represents the diversity within the target population of registered voters in the United States. This method involves dividing the population into distinct subgroups, or "strata," based on key characteristics such as state, age, gender, race, education level, and political affiliation. For example, in presidential elections, particular emphasis is placed on voters from swing states like Arizona, North Carolina, Georgia, and Pennsylvania. By focusing on these battleground states, the poll aims to capture the opinions of voters who are most likely to influence the election outcome. Once the strata are established, respondents are randomly selected within each group, ensuring that all demographic groups are appropriately represented in proportion to their actual distribution within the electorate.

B.2 Advantages

Stratified sampling offers the advantage of higher precision in representing the population's diversity, as each subgroup is proportionally represented. By breaking down the sample into distinct groups, the Times/Siena Poll can adjust for demographic variations between states or voter groups, which is particularly useful in a diverse electorate. This approach helps to minimize sampling error by ensuring that all significant voter subgroups are included, thus increasing the reliability of the poll's predictions.

B.3 Challenges and Trade-offs

Although stratified sampling improves representativeness, it presents certain challenges in execution. First, identifying and reaching a sufficient number of respondents from harder-to-reach groups (such as young voters, minority populations, and those without reliable phone access) can be difficult, especially given the 2% response rate. To counterbalance the low response rates, the poll requires multiple call attempts and careful post-survey weighting to adjust for any potential underrepresentation. Additionally, while phone polling offers broad geographic reach and a standardized mode of data collection, it limits the sample to voters with phone access who are willing to participate in phone interviews, potentially introducing selection bias if certain demographics are less reachable by phone.

C Idealized Methodology

Pursuant to accurately predicting the winner of the presidential election with the electoral college system, there are assumptions that need to be assessed. The United States of America's electoral college system creates a special environment where votes in some states are disproportionately more influential than that of others and candidates are not elected based on popular vote, but on the majority of electoral votes they get.

C.1 Assumption 1:

Nation-level Popular Vote is irrelevant in our methodology since president electors are chosen based on electoral votes.

C.2 Assumption 2:

Based on historical and polling data, some states' electoral votes are predetermined as either Democrat or Republican. Therefore, it is more prudent to shift focus over to swing states where outcomes are uncertain. This assumption is accompanied with the challenge of determining what states qualify as "swing states". However, on this issue, there is consensus on which states qualify as swing states. Pollsters such as New York Times have their metric for determining which states are 'battleground states'. As of 2024, our methodology should direct all our resources to acquiring voting data on the following states (New York Times). This assumption results in each presidential candidate starting off with a set number of electoral votes.

C.3 Swing States

According to 2024 polling data conducted by New York Times, the swing states are determined to be the following:

Pennsylvania Georgia North Carolina Michigan Arizona Wisconsin Nevada

C.4 Feasibility of Reaching Voters: Justification of Online Panels

According to the New York Times, with a sample of 1000 individual votes, there is a margin of error of around three to four percentages. With $100,000, there are limitations on the labor and resources that is required to poll 1000 voters while balancing out demographics through dialing. Furthermore, there is additional filtering of our sample frame to determine likely voters. Assuming New York Times' 2% voter response rate, to sample 1000 voters, at least 50,000 voters will need to be called. It is likely that our team will be dialing way more than 50,000 times since it is an unreasonable assumption that every voter we call would pick up or

be available to provide a response. There is also the fact that our team does not carry the same credibility and history of acquiring poll data as the New York Times does, resulting in more reluctance for voters to share information with us. Therefore, it is not feasible for our methodology to employ telephone dialing as our method to reach voters. Instead, we focus on online panels and surveys with incentives for completing the survey correctly. However, online panels and surveys come with their own set of challenges and weaknesses that must be addressed in our methodology. Primarily, the problem of self-selection in online panels disrupts the random selection that allows the external validity of our predictions. Moreover, there is apossibility that we acquire responses from respondents that are not likely to vote, gaining data from outside our target population. Even if we distribute the panels randomly among swing states, responses are still corrupted by selection bias that occurs from responses that are provided from demographics that feel more passionately about the election or are more likely to respond to online panels. Unlike telephone dialing, we have less control over the specific demographic we want to target, resulting in a less stratified data that is representative of the target population overall.

C.5 Remedies for Online Panels

C.5.1 Weighting Adjustments

Although online panels offer an imperfect solution to the limited resources our team will work with, there are remedies that can address underrepresentation brought by selection bias. As NYC and Siena College does, post-survey weighting can help us represent our target population more accurately. This is an effective way to adjust our sample based on demographic data and historical turnout patterns within each state. By weighting responses to mirror actual voter demographics from a particular state, the survey results can approximate a more realistic projection of the electorate. These weighting adjustments help to maintain the integrity of our sample, closely resembling the target population of a state.

C.5.2 Screening for Likely Voters

Although weighting adjustments helps us represent our target population more closely, there is nothing to stop respondents from outside our target population to participate in our panels. To produce accurate predictions of election results, our sample needs to be acquired from a target population that will participate in the election. Therefore, it becomes necessary to include screening questions designed to filter for likely voters, based on voting history or engagement levels. Our survey includes questions that pertain to voting history, interest and engagement, voting intention, and eligibility. Responses to these questions will be used to produce an aggregate "likely voter" score that filters for the target population enabling for more accurate predictions of the presidential election.

C.6 Elements of Our Survey

To gather accurate responses from a diverse and representative set of likely voters, our survey design carefully adheres to best practices in methodology. Given our goal to avoid attrition bias, we've restricted the estimated survey completion time to under ten minutes. This constraint is stricter than the fifteen-minute benchmark typically used by NYC and Siena College for telephone surveys, as online surveys require shorter durations to maintain respondent engagement. Our survey is divided into three main sections: Preamble, Eligibility, and Party Affiliation. The survey we constructed can be found in Appendix B.

C.6.1 Preamble

The Preamble serves as an introductory section that aims to establish transparency and motivate respondents to complete the survey accurately. This introduction begins with an explanation of the lottery-style incentive offered for survey completion. Rather than small individual payments, participants have a chance to win a larger prize if they complete the survey, a choice that aligns with budget constraints while enhancing participation rates. To encourage honest responses, respondents are assured of the anonymity of their answers, as well as the fact that their responses will be used solely for research purposes. This transparency fosters trust and can reduce hesitation when sharing political preferences. Information about the survey's purpose is provided sparingly to minimize any potential response bias, as we communicate that the survey concerns voting and political engagement while omitting specific details that could influence responses, such as particular issues or candidate names. Additionally, respondents have the option to input a referral code if invited by a friend or family member, which can increase their chances of winning the lottery and encourage participation through word-of-mouth. We also request information on the relationship with the referrer to monitor potential demographic clustering in cases where social or familial ties might influence responses.

C.6.2 Eligibility

The Eligibility section focuses on ensuring that respondents meet the criteria necessary for inclusion in our target population, specifically targeting likely voters within designated swing states to improve the relevance and accuracy of our sample for election predictions. This section begins by asking for respondents' state of residence, age, race, and other preliminary demographic information. To further narrow our sample to those most likely to vote, we employ several screening questions related to voting history, political engagement, and voting intentions. These questions help exclude individuals who are unlikely to participate in the election, thus refining our sample to reflect the likely voting population.

C.6.3 Party Affiliation

Party Affiliation addresses respondents' political leanings and past voting behavior. This section explores their current preference for presidential candidates and parties and asks how they voted in previous elections. By differentiating between party and candidate preference, we can identify any variance in their voting behavior and assess whether respondents' decisions align more strongly with party or individual candidate considerations. This section also explores whether respondents are informed about their choices, asking questions to determine whether they are aware of both candidates or have researched their policies. By gathering these insights, we can create a profile of the respondent's political orientation, which allows for a nuanced interpretation of voting intentions that goes beyond simple party preference.

C.7 Distribution of Online Panels

The online panel method provides valuable flexibility in how surveys are distributed, enabling targeted outreach in regions crucial to our election prediction model. Since the distribution strategy will absorb a significant portion of the budget, this section outlines our plan for cost-effective, targeted, and controlled distribution, aimed at achieving a sample that represents voters in key swing states while addressing biases inherent to online panel surveys.

C.7.1 Swing State Representation

To capture a meaningful sample for predicting election results, we target respondents exclusively from key swing states where voting outcomes remain uncertain. By focusing our resources on swing states, we enhance the relevance of our sample to the Electoral College system, which relies heavily on swing state outcomes. Achieving this requires not only careful selection of digital platforms but also strategic advertising in locally relevant spaces. The lottery-style incentive, with a $5,000 voucher offered to one lucky respondent, serves as an appealing motivator designed to increase participation. We will advertise the survey on websites, forums, and marketplaces that have high traffic within our target states. Examples include social media channels, local news sites, and state-specific online communities. Additionally, we may utilize geo-targeted ads on platforms like Google and Facebook to directly reach residents of these swing states, optimizing our budget by focusing on the digital spaces our target demographic frequents. A significant portion of the budget will thus be allocated to these state-specific online advertisements, which not only promote the survey but also help ensure that respondents fit our target profile.

C.7.2 Referral Mechanics

To further broaden our reach without exponentially increasing advertising costs, we incorporate a referral system that rewards respondents for inviting others to complete the survey. Each

referral allows the original respondent to gain an additional entry in the lottery, increasing their chances of winning. This referral incentive taps into snowball sampling, where respondents refer people within their network to participate, expanding our pool of respondents organically. However, using snowball sampling in an election survey introduces potential biases, particularly the risk of overrepresentation of individuals with similar political views, as respondents are likely to refer to close contacts who may share similar ideologies. To mitigate this, we will place specific restrictions on the referral program. Each respondent will be limited to referring only one other individual, minimizing the potential for overly clustered referrals. Additionally, we will require respondents to identify their relationship with each referral, helping us monitor homogeneity in our responses. By balancing the benefits of snowball sampling with these restrictions, we aim to expand the reach of our survey into difficult to reach demographics without compromising the representativeness of the sample.

Appendix B

D Idealized Questionnaire and Survey

https://docs.google.com/forms/d/117wC8f1rBW_bd5glEdiR-7823-02A3HiD5EzQuT1GVw/edit

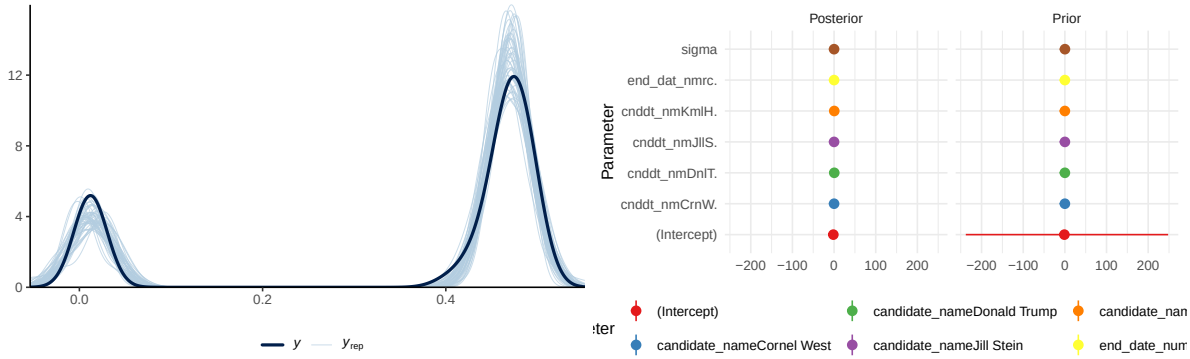
E Additional data details

F Model details

F.1 Posterior predictive check

In Figure 5a we implement a posterior predictive check. This shows that the posterior predictive density closely follows the observed data density. Discrepancies between the two suggest that the model may be missing some aspect of the data structure or variability, which could guide potential model adjustments or refinement.

In Figure 5b we compare the posterior with the prior. This shows the prior for each parameter is centered around zero, but the posterior shows more precise estimates for each candidate and the date variable, reflecting the data's influence. Also, the posterior narrowing suggests that the model successfully extracted information from the data, reducing uncertainty around parameter estimates compared to the prior. model_data <- read_parquet(here::here("data/02-analysis_data/model_data.parquet"))



(a) Posterior prediction check

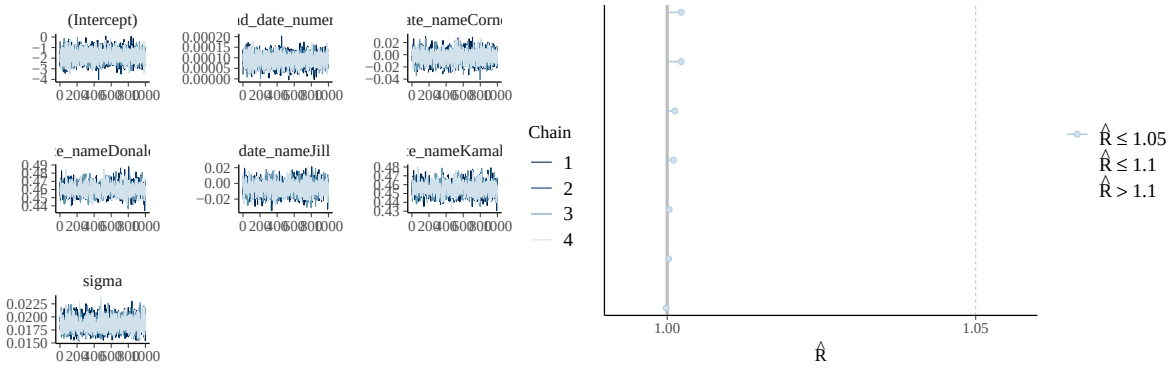
(b) Comparing the posterior with the prior

Figure 5: Examining how the model fits, and is affected by, the data

F.2 Diagnostics

Figure 6a is a trace plot. It shows chain mixing, stationarity and no drifting. This suggests that the model's MCMC sampling achieved convergence for each parameter. The chains are well-mixed and stationary, providing confidence in the posterior estimates derived from this sampling process.

Figure 6b is a Rhat plot. It shows the Rhat values for the parameters estimated in the model. Values close to 1 indicate that the chains have mixed well and the posterior distribution has been adequately explored. This suggests that the model has converged successfully and the estimates are reliable for inference.



(a) Trace plot (b) Rhat plot

Figure 6: Checking the convergence of the MCMC algorithm

References

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