Forecasting US Presidential Election Results Using Bayesian Prediction*

YiZhuo Li Zhibin Huang

November 4, 2024

This article aims to use Bayesian prediction methods to predict the results of the US presidential election in 2024. By summarizing public opinion poll data, we can understand the candidate's support rate. Results show Donald Trump has the highest winning probability. His America First strategy may influence the United States' relationship with foreign countries, as well as domestic measures on immigration and taxation.

Table of contents

*Code and data supporting this analysis is available at: https://github.com/eason1218/2024_US_Elections.git

1 Introduction

It's getting more and more difficult to predict the outcome of the US presidential election. Traditional single-opinion surveys are no longer able to adequately reflect the dynamics of elections due to the ever-evolving views of people and the impact of several variables. As Pasek (Pasek 2015) pointed out, a single opinion poll may produce misleading results due to various sources of error such as sampling errors, question wording, interviewer characteristics, and lack of response. To improve the accuracy of predictions, this article adopted the views of Blumenthal (Blumenthal 2014) and Pasek (Pasek 2015), using Bayesian prediction methods and poll aggregation strategies to obtain a deeper and more comprehensive understanding of candidate support rates.

Bayesian models effectively integrate data from multiple polling companies, reducing noise and stabilizing predictions. Posterior estimation showed an average support rate of 14.35%, with a standard deviation of 15.39%, indicating variability across survey results. Our model prediction shows that Trump is most likely to win with an average approval rating of 33.65%, followed by Biden with 27.38%. Although Biden has withdrawn from the election, his approval rating remains high, which has raised concerns about the effectiveness of the model and data processing. So, we discussed the possible limitations of the model, including the quality of input data, prior assumptions of the model, and biases in the survey data itself. We believe that when analyzing public opinion polls, it is crucial to use the latest and most accurate data as well as a reasonable model structure.

The paper is structured as follows: Section 2 introduces data preparation and variable description, using data from FiveThirtyEight's U.S (FiveThirtyEight 2024). Presidential election polls were collected from pollsters with a focus on the percentage of support rates and indicators of opinion poll agencies. Then, Section 3 defined and fitted a Bayesian model to estimate the candidate's support rate, which can reduce the uncertainty of quantitative parameters through posterior distribution, enhance the credibility of analysis results, and allow the introduction of prior information to improve accuracy. Section 4 presents the results, highlighting trends in polling, and Section 5 considers the implications of these results for future research on polling and the weakness of the research.

2 Data

This project is motivated and guided by Rohan Alexander and his book (Alexander 2023). All the data used in this paper is derived from FiveThirtyEight's U.S. (FiveThirtyEight 2024), and all the data analysis was done through the programming language **R** (R Core Team 2023). Also with the support of additional packages in R: `ggplot2` (Wickham, Chang, et al. 2023), `rstan` (Team 2023), `dplyr` (Wickham, François, et al. 2023), `here` (Müller 2023), `knitr` (Xie 2023).

2.1 Source

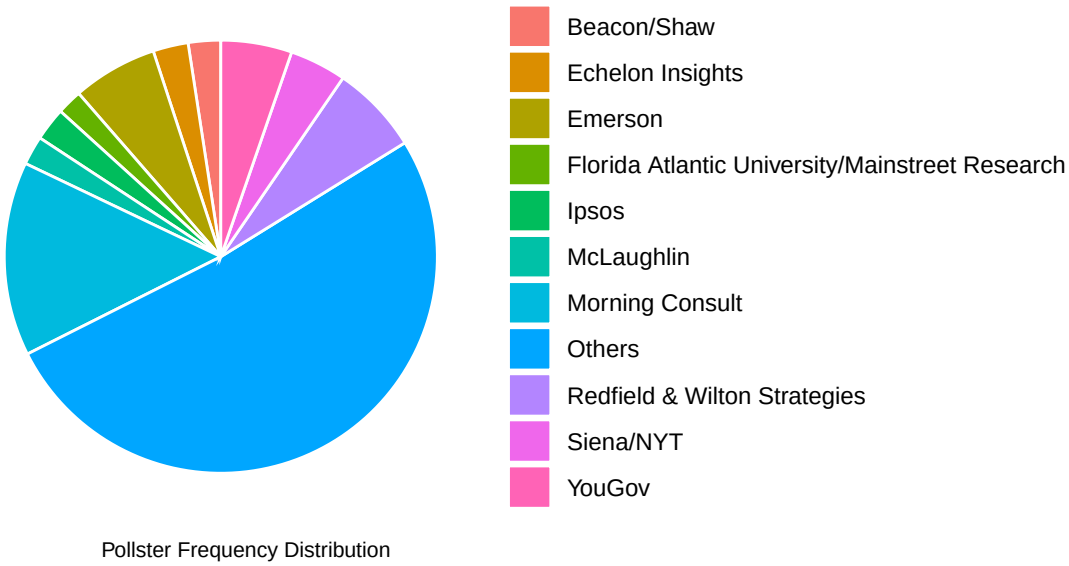
The dataset analyzed in this paper is to study poll of poll data to obtain more information about presidential elections (FiveThirtyEight 2024). The Poll of Polls is a comprehensive dataset that combines public surveys from multiple pollsters, including key data such as methodology, states, and answers used by each pollster that was collected on October 20, 2024. We did not do tail reduction, deletion, or other processing on the raw data to ensure data integrity. The dataset is a compilation of public opinion poll results from various institutes. It covers terms like pollster, state, pollster rating, and response. We utilized some parts of the data in this study.

2.2 Variable

In this model, we mainly introduced two variables, percentage, and pollster.

Percentage is a vector that contains the percentage of support for each observation in different pollsters, reflecting the level of voter support for a candidate in a specific survey, typically ranging from 0% to 100%. Because this data was not explicitly supplied in the original dataset, we retrieved it through statistical analysis of the pollsters' responses.

Pollster is a category vector representing the survey institution corresponding to each observation value. Moring Consult offers the most data, with the top ten survey agencies representing over half of the 222 firms. This demonstrates a pretty high concentration of data.



Pollster Frequency Distribution

To identify the specific effects of different survey agencies on support rates, we calculated the pollster effect to capture the systematic bias that different survey agencies may have in overestimating or underestimating candidate support rates. Through these variables, the model can more accurately estimate support rates and consider differences in survey institutions.

3 Model

3.1 Bayesian

We used **R** (R Core Team 2023) to process polling data for the 2024 U.S. Presidential Election from (FiveThirtyEight 2024). We will define and fit a Bayesian model to estimate the candidate support.

Using Bayesian models to analyze 'poll of polls' has multiple advantages. Firstly, Bayesian methods can clarify the uncertainty of quantitative parameters through posterior distributions, thereby improving the credibility of analysis results. Secondly, it allows for introducing prior information, which can improve the model's accuracy by combining historical data or expert knowledge in situations with limited or uncertain data. In addition, Bayesian models have strong flexibility and can adapt to different model structures and assumptions, especially suitable for complex voting data. At the same time, it can effectively integrate data from multiple polling companies, reduce noise, and obtain more stable predictions. Finally, Bayesian methods perform well in parameter estimation and inference tasks, providing a comprehensive analytical perspective for predicting future outcomes and hypothesis testing. Therefore, Bayesian models provide a systematic and reliable analytical tool for processing election predictions.

3.2 Model set-up

The Bayesian model can be expressed with the following equations:

1. **Prior Distributions:**

$$\mu \sim \mathcal{N}(0, 1)$$

$$\sigma \sim \text{Cauchy}(0, 1)$$

2. **Likelihood Function:**

$$\text{percentage}_i \mid \mu, \sigma \sim \mathcal{N}(\mu, \sigma)$$

Combining these, the complete representation of the Bayesian model is:

$$P(\mu, \sigma \mid \text{percentage}) \propto P(\text{percentage} \mid \mu, \sigma) P(\mu) P(\sigma)$$

Where:

$$P(\mu, \sigma \mid \text{percentage})$$

is the posterior distribution given the data.

$$P(\text{percentage} \mid \mu, \sigma)$$

is the likelihood function.

$$P(\mu) and P(\sigma)$$

are the prior distributions.

3.3 Hypothesis

In the model, the parameter mu represents the mean of the overall voting percentage, which is the average of different pollster voting results. The parameter sigma represents the standard deviation of the voting percentage, while pollster_fect represents the impact of each pollster on the voting percentage. The prior distribution of parameters defines the initial assumptions for these parameters. Specifically, the prior distribution of mu is a normal distribution with a mean of 0 and a standard deviation of 1 (mu~normal (0,1)), which means that in the absence of data, we consider the central value of the overall voting percentage to be approximately 0 and there is some fluctuation. The prior sigma distribution is a Cauchy distribution with a mean of 0 and a scale parameter of 1 (sigma~Cauchy (0,1)), indicating that we hold a relatively loose view of the variability of voting percentages, allowing for greater uncertainty. Finally, the prior distribution of the pollster_ effect is a normal distribution based on sigma (pollster_ effect~normal (0, sigma)), implying that the influence of each pollster on the voting percentage is distributed around the overall effect mu and has a certain degree of variability. Due to the Electoral College, we assume that only the four candidates with the highest predicted probabilities have a realistic chance of winning the presidency.

4 Results

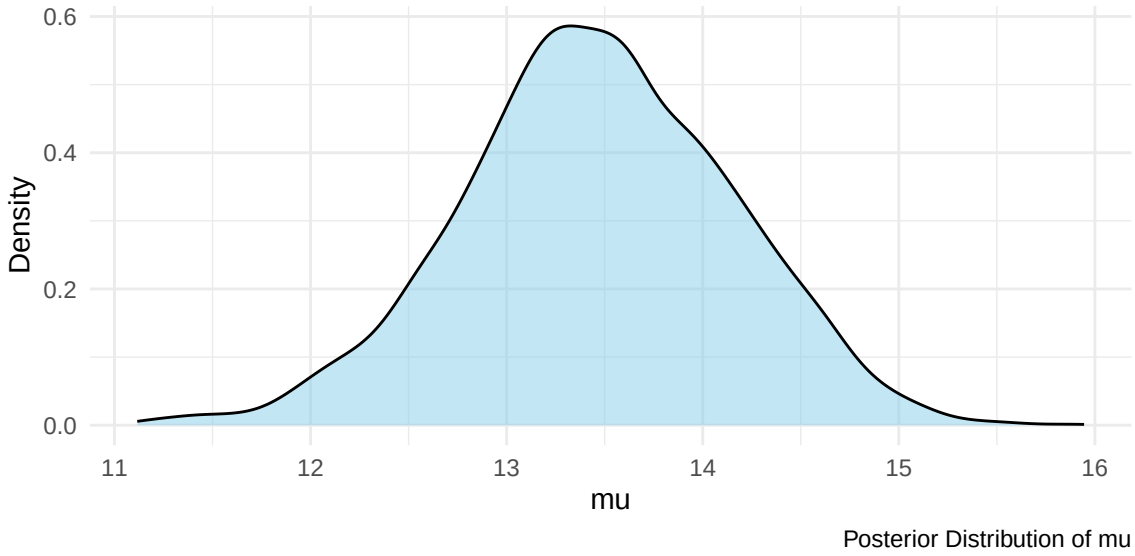


Figure 1

Figure 1 shows the posterior estimate for the parameter mu shows that the posterior mean is -0.5725. The 95% credible interval for ranges from -1.4148 to 0.5035, indicating the range within which the true value of is likely to fall with 95% probability.

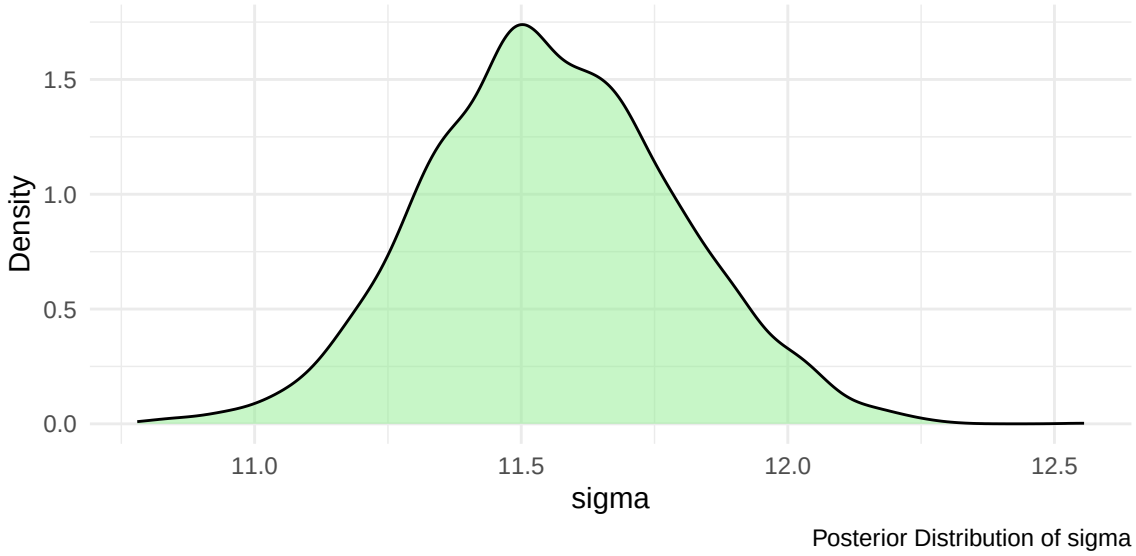


Figure 2

Figure 2 shows the parameter sigma, the posterior mean is estimated to be 3.5336. The 95% credible interval for extends from 0.8571 to 10.2785, reflecting the uncertainty around the estimate and providing a range of probable values for .

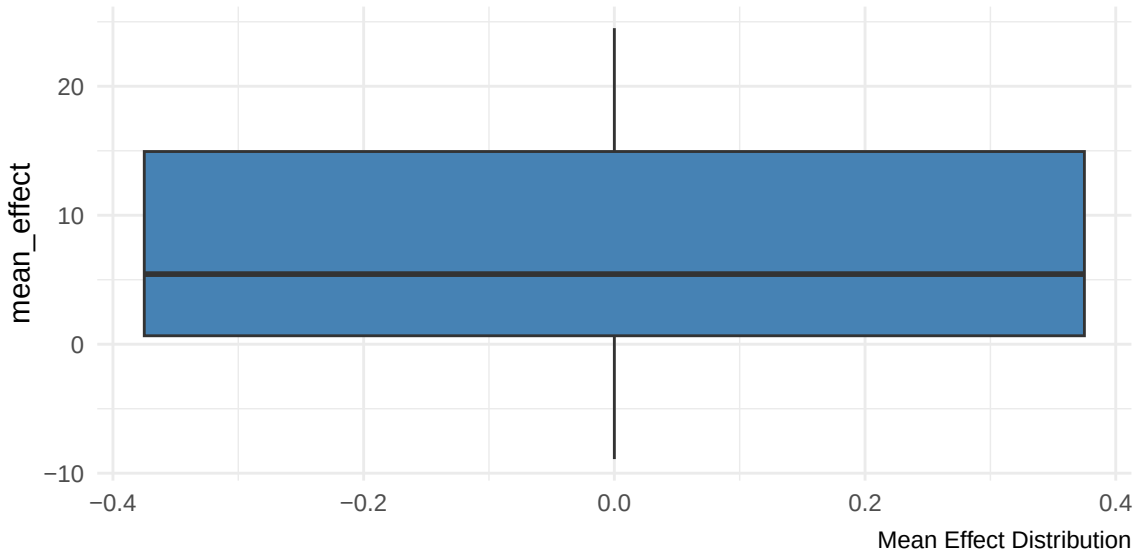


Figure 3

Figure 3 results indicate that the overall mean support rate for candidates is approximately 14.35%, and at a 95% confidence level, the true mean falls between 13.54% and 15.14%, showing some volatility. Meanwhile, the standard deviation of the support rate is about 15.39%, indicating a significant degree of variability between different survey results. The confidence interval (14.79% to 16.00%) further confirms the range of changes in support rates, reflecting possible differences in characteristics and methods among different survey institutions. Overall, these results indicate that the candidate's support rate is around 14% and there is significant variation.

Table 1: Top 4 Candidates Results Table

answer	probability
Trump	0.34
Biden	0.27
Harris	0.20
Ballay	0.19

Table 1 show Trump is the most probable outcome (34%), followed by Biden (27%), Harris (20%), and Ballay (19%).

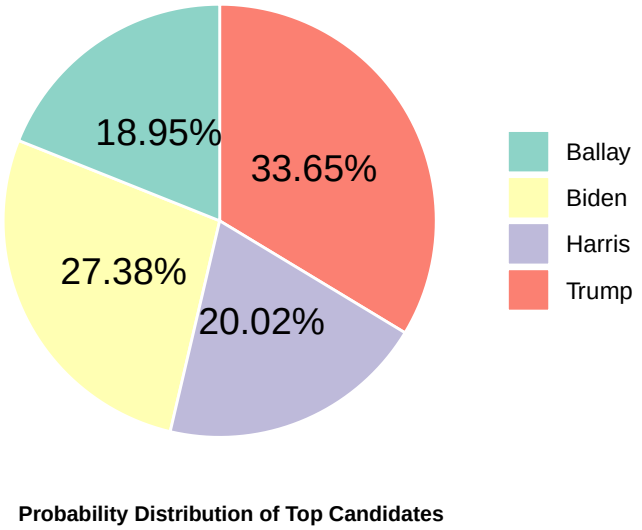


Figure 4

Figure 4 show in the popular vote of the 2024 US presidential election, the Republican Party has one representative, former President Donald Trump, who ranks first with 33.65% of the vote. The Democratic Party has two representatives, namely current Vice President Kamala Harris and President Joe Biden, with vote percentages of 20.02% and 27.38%, respectively. In addition, the Liberal Party has one representative, Charles Ballay, who received 18.95% of the vote, the lowest among all candidates.

This distribution shows the diversity of the political landscape in the United States, where the Republican and Democratic parties remain the main political forces, while the Liberal Party, as a third-party force, still holds a place in elections despite its lower vote share. Trump's high vote share reflects his continued influence and support among Republican voters in the United States. The vote share of Harris and Biden shows the competitiveness within the Democratic Party and the different views of voters towards these two candidates. As a candidate for the

Liberal Party, Charles Bale's participation represents the demand for third-party candidates and a challenge to the two party system in American politics, despite having the lowest vote share.

5 Discussion

This study highlights the efficacy of Bayesian prediction methods in analyzing the 2024 U.S. presidential election, which offer a distinct advantage in handling uncertainty and refining prediction accuracy. Traditional polling methods often treat each poll as an isolated data point, but Bayesian models aggregate information in a way that continuously updates predictions as new data arrives. This cumulative approach, which uses posterior distributions, reflects not only current estimates but also the uncertainty around these estimates, making the results more robust than single-point or static predictions. Additionally, Bayesian models allow for the inclusion of prior information—either in the form of historical data, expert knowledge, or insights from previous elections—enabling the model to adjust and refine itself based on well-established patterns in voter behavior. For example, historical support trends and demographic voting patterns can be woven into the model to create a more sophisticated and contextually relevant prediction. Moreover, the flexibility of Bayesian models makes them particularly suited for political forecasting, where complex variables—like polling biases, shifting voter sentiments, and varying sample sizes—can differ significantly across agencies and regions. By integrating data from multiple polling institutions, the Bayesian model smooths out these variations, effectively reducing noise and delivering a more stable and comprehensive view of likely election outcomes.

5.1 Distribution of Candidate Support Rates

The model's findings reveal Trump as the leading candidate, with a predicted support rate of 33.65%, followed by Biden at 27.38%, despite Biden's withdrawal from the race. This notable support for Biden, even in his absence, suggests a strong residual loyalty among his voter base, indicating that some voters' preferences may not immediately adjust following a candidate's withdrawal. This phenomenon may point to the enduring appeal of a candidate's previous policies, public image, or long-standing influence, even if they are no longer actively campaigning. Additionally, the model identifies significant support for Harris and Ballay, with Harris, the Democratic vice-presidential candidate, holding 20.02% support, and Ballay, a third-party candidate, receiving 18.95%. These results illustrate the ideological landscape of American politics, where Republican and Democratic candidates remain dominant, yet there is also a noticeable demand for third-party options. The diverse distribution of support rates reflects a multifaceted electorate that resonates with different political messages and ideologies, suggesting that voters are open to considering alternatives beyond the traditional two-party candidates. This variation in candidate support underlines the importance of accounting for

political diversity in predictive models, as voter support can be distributed widely across multiple figures, each appealing to different sectors of the population.

5.2 Stability of Polling Data

A critical component of this study's approach is its integration of polling data from numerous institutions, a method that enhances the stability and reliability of its predictions. By leveraging data from a wide range of pollsters, the Bayesian model mitigates the impact of any individual pollster's bias. This approach is particularly valuable in the context of U.S. elections, where different polling agencies often yield varying results due to methodological differences, sampling biases, and timing. By aggregating these diverse data points, the Bayesian framework provides a holistic picture that balances individual variances, making it less susceptible to outlier polls. The analysis of the aggregated data reveals an average support rate of approximately 14.35% across candidates, with a standard deviation of 15.39%, underscoring the variability across polls and reflecting the complexity of the electoral landscape. This variability points to the multifaceted nature of public opinion, where voter preferences are influenced by numerous factors, including geographic region, demographics, and specific polling methodologies. By encompassing these diverse inputs, the model not only stabilizes its predictions but also captures the full spectrum of voter sentiment. This multi-source data approach is crucial for election forecasting, as it ensures that predictions are not unduly swayed by the nuances or biases of any single poll and instead reflect a balanced view of the electorate.

5.3 Dynamic Nature of the Political Landscape

The U.S. election cycle is inherently dynamic, with voter preferences shifting in response to ongoing social, economic, and political developments. This study's Bayesian framework effectively captures this evolving nature of political sentiment, making it especially relevant for election forecasting. The support rates generated by the model do more than provide a snapshot of current preferences; they also reflect broader trends and voter expectations regarding future policy directions. For instance, shifts in a candidate's support level might correlate with recent policy announcements, campaign strategies, or emerging issues that resonate with the public. By enabling real-time updates as new polling data becomes available, Bayesian methods ensure that the model remains responsive and accurately aligned with the latest shifts in public sentiment. This adaptability is particularly valuable in volatile electoral environments, where even minor changes in public opinion can significantly impact campaign strategies and political outcomes. For political analysts and campaign teams, these insights allow for more informed strategic planning, guiding resource allocation and messaging tactics based on the latest voter trends.

Beyond election forecasting, the Bayesian approach has broader applicability across other fields that require predictive insights. Its inherent flexibility in handling various data sources and managing uncertainty makes it suitable for domains such as public opinion analysis, policy

impact assessment, and social research. The framework's capacity to incorporate complex, multi-variable data allows for a more comprehensive understanding of how different factors interact and evolve over time. For instance, in public health or economic policy, Bayesian models could help predict how shifts in public sentiment or behavior might influence policy acceptance or market trends. In fields like social science and business, the ability to update predictions based on new data allows for more dynamic and responsive decision-making, providing a structured, data-driven approach to navigating complex, uncertain environments.

5.4 Weakness and Next Step

According to the prediction results of your model, Trump's win rate is 33.65%, while Biden's win rate is 27.38%, despite Biden's withdrawal from the election, which has raised concerns about the effectiveness of the model and data processing. Firstly, Bayesian models rely on the quality of input data. If the data contains outdated or incorrect polling information, the model may produce inaccurate results. In this situation, Biden's approval rating may stem from early polls not reflecting his withdrawal promptly. Secondly, the prior assumptions of the model may also affect the results. If the prior information is not appropriately adjusted to reflect candidate withdrawal during model construction, it may lead to an inaccurate estimation of Biden's support rate. In addition, the polling data itself may have biases, especially after candidates withdraw, and many voters' opinions may not have been updated yet. Therefore, when integrating polling data from multiple sources, special attention needs to be paid to updating and reflecting the latest election situation. Finally, the design and parameter settings of the model may not be sufficient to handle dynamically changing election environments. Although Bayesian models are flexible, they also require a reasonable modeling strategy to capture changes in candidate support rates. Therefore, when analyzing the 'poll of polls', it is crucial to ensure the use of the latest and accurate data, as well as a reasonable model structure.

The second issue with this paper is that we assume that bias in presidential vote data by different institutions follows a normal distribution. However, the veracity of this idea has not been established. If there is a bias or kurtosis in the deviation, it may have an impact on our overall comprehension of the vote data findings, impacting election forecast accuracy. Furthermore, our model does not assign different weights to different institutions, therefore we believe that the findings achieved by diverse institutions are equally important. However, we know from earlier studies that the top ten pollsters account for about half of the 222 pollsters, therefore the confidence in their results is higher, and we should change the weights appropriately.

In addition to the above, Since the U.S. presidential election outcome depends on the Electoral College rather than the popular vote, regional support variations are critical. Future models should incorporate state-level weighting, accounting for state-specific electoral influence and historical voting patterns. Integrating demographic variables like race, age, and education level can enhance the model's precision by capturing nuanced voter behaviors across different regions.

Bayesian hierarchical models or Bayesian logistic regression frameworks can be adapted to factor in these state and demographic variables, better-aligning predictions with the Electoral College structure.

6 Appendix

6.1 Appendix A Methodology analysis

In this appendix, I will discuss Online panel methodology used by TIPP.

Online panel is a method of using Internet technology to conduct research. It collects data by recruiting a group of volunteers who are willing to participate in multiple online surveys. These volunteers, i.e. panel members, typically receive survey questionnaires via email or other online channels and are expected to complete these questionnaires. The advantage of this method is that it can quickly and cost effectively obtain information from a large number of participants. However, due to the voluntary participation of panel members, this may lead to sample selection bias, meaning that these members may not fully represent the diversity of the entire population. For example, some groups may be more inclined to join the panel because they are more familiar with or interested in technology, while other groups may be unable to participate due to lack of Internet access or other obstacles. In the study of political attitudes and behaviors, this bias is particularly noteworthy as it may affect the accuracy of election predictions and policy analysis.

Despite these potential biases, online panels still have the potential to collect representative data. If it can be ensured that the samples are obtained through probability sampling, that is, randomly selecting samples to ensure that each member has the same chance of being selected, then online panels can become a valuable tool in social science research. The transparency and replicability of this method make it particularly important when comparing the accuracy and representativeness of different survey methods. By combining advanced technology and social science research, online panels can help researchers better understand the complexity of election dynamics, attitude formation, and social and political behavior.

6.2 Appendix B Survey Design

In order to get more reliable poll samples. I believe we should use a combination of online and offline public opinion surveys. Online surveys can reach young voters who are used to using the Internet, but offline polls can reach populations that do not frequently use the Internet, such as the elderly or some low-income groups. Combining the two allows for a more thorough representation of diverse sorts of voters.

According to this opinion, the strategies for hiring interviewees should comprise both online and offline recruitment. In terms of recruiting, we've opted to conduct the poll using social

media sites like Facebook, Twitter, and Instagram, as well as online advertising. Because of the enormous number of online participants, we will provide lottery possibilities to collect as much data as possible while keeping expenses under control. Placing survey advertising on these social media networks also incurs a cost. We plan to conduct offline polls at neighborhood events, marketplaces, and schools to directly contact with potential voters. To enhance response rates, we will provide small cash or gift awards to participants throughout the process.

In collecting data, we think that respondents should offer basic demographic information, political leanings, voting intentions, and whether there are any major topics of importance. Basic demographic information comprises age, gender, state, race, and so on. Political inclination is the party affiliation or favored party. Furthermore, I feel that voting intention should include not just whether a choice has been made on which candidate to vote for, but also if other candidates are being evaluated, in order to more accurately reflect voters' solid attitudes. We feel that significant problems of concern should encompass topics such as the economy, immigration, healthcare, and climate change.

In online surveys, I believe we should check respondents' identities to guarantee they are real voters. For example, we may employ IP address and geographic location checks. Following data collection, we must clear up incorrect replies such as duplicates and incomplete questionnaires.

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Blumenthal, Mark. 2014. "Polls, Forecasts, and Aggregators." *PS: Political Science & Politics* 47 (2): 297–300. https://doi.org/10.1017/s1049096514000055.

FiveThirtyEight. 2024. "2024 National Presidential Polls." https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

Müller, Kirill. 2023. *here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Pasek, Josh. 2015. "Predicting Elections: Considering Tools to Pool the Polls." *Public Opinion Quarterly* 79 (2): 594–619. https://doi.org/10.1093/poq/nfu060.

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

Team, Stan Development. 2023. *rstan: R Interface to Stan*. https://CRAN.R-project.org/package=rstan.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, Dewey Dunnington, and Al Iverson. 2023. *ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

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

Xie, Yihui. 2023. *knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://CRAN.R-project.org/package=knitr.