Demystifying Votes: A Hierarchical Bayesian Model for Predicting the 2024 US Presidential Election*

Decoding the dynamics and biases behind Trump's Approval Ratings with polling Data

Chenming Zhao

November 2, 2024

This study aims to use a hierarchical Bayesian model to predict whether Donald Trump will win the 2024 U.S. presidential election based on support rates. By analyzing data of 2024 US presidential general election polls across states, pollsters, and methodologies, we capture variations in support for Trump driven by these factors. Our results reveal significant differences in support between national and state-level polls, with fluctuations in key swing states potentially influencing the election outcome. There is a dynamic interaction between election results and public opinion trends, and this study provides insights into how various factors influence public sentiment, offering valuable predictions and understanding of voter dynamics in the U.S. election.

1 Introduction

The U.S. presidential election has consistently captured the attention of nearly everyone. The 2024 U.S. presidential election is taking place against a backdrop of intense political polarization and complex shifts in public opinion. With changes in the social and economic landscape post-pandemic, the rise of differing ideologies, and frequent political controversies, forecasting the outcome of the 2024 election has become more challenging than ever. Statistics show that the U.S. presidential election is one of the world's most expensive elections, involving massive amounts of money and manpower, and the 2024 election is expected to cost a record-breaking

$15.9 billion (see Josh Christenson and Glebova 2024). Both the Democratic and Republican parties have poured significant resources and funds into the campaign, employing different strategies to win over voters (see Ax 2024). The policy stances of different parties directly affect voters and national policies, and these differences will continue to influence the future direction of the electorate, the United States, and even the world. As the world's largest economy, the U.S. policy shifts directly impact global markets. Changes in U.S. trade policy often lead to shifts in international trade dynamics, affecting the economic development of other countries. Additionally, the U.S. foreign policy and military strategy hold a dominant position in global affairs. The diplomatic stance of the new president may also alter relations with other countries, impacting global security.

In this paper, we focus on the outcome of the 2024 presidential election. Using a dataset from FiveThirtyEight containing extensive polling results, we examine the influence of various factors on candidate support rates by accounting for differences in polling methods, state-level variation, and temporal changes. It is evident that polling methods often fail to fully capture the complexity of public opinion, and factors that introduce bias may intervene from the very design of polling methods. The chaos and complexity of the real world are also reflected in a large-scale political event like the 2024 presidential election, which involves a vast number of participants, wide-ranging groups, and massive investments and interests. Therefore, we employ a hierarchical Bayesian model as a predictive tool in our analysis. This model integrates polling data from FiveThirtyEight covering various states and polling organizations, while also considering the effects of regional and methodological influences. By incorporating variables such as days since Biden's withdrawal, polling methodology, and whether the poll is national or state-level, the model provides a dynamic view of Trump's support rate across time and place. With an R^2 of 0.641 and an RMSE of 2.31, the model effectively captures these patterns while maintaining predictive accuracy, indicating that it can provide meaningful insights into regional and national public opinion trends. Finally, we simulate data for Election Day on November 5 to provide a reliable basis for predicting the election outcome.

Our analysis reveals that differences in Trump's support rate across regions and polling methods are particularly pronounced. For example, support rates in traditionally conservative states like Texas differ significantly from those in traditionally liberal states like California, while trends in swing states such as Pennsylvania and Michigan are more unstable, with the potential to disrupt the election balance. This underscores the critical impact of geographic and methodological factors on public opinion, aligning well with the historical tendency of swing states to determine the election outcome. Additionally, our results suggest that polling methodologies introduce systematic differences in support rate estimation, with slight variations between results from online panels and phone surveys. In the predictive results, support for Trump at both the state and national levels hovers around 45% to 50%. Comparing this to the support rates of past U.S. presidents, we predict that Trump is unlikely to win in this election.

The remainder of this paper is structured as follows. Section 2 introduces the data sources, preprocessing steps, variable selection criteria, and the potential relationships between inde-

pendent and dependent variables. Section 3 provides a concise explanation of the hierarchical Bayesian model, including the design rationale and selection of priors and random effects, with more details on the model's posterior in Appendix B. Section 4 presents the main findings of the study, including a summary of the model output to support its accuracy and predictions of Trump's support rate across different regions on November 5, 2024. Section 5.1 in Section 5 discusses the implications of the results, Section 5.2 talks about limitations of the data and polls, complexity of conducting our real world into datas, and future directions.

2 Data

2.1 Overview

Polling data for the 2024 U.S. presidential election is provided by FiveThirtyEight (2024). This dataset compiles information from numerous polls conducted by various pollsters, detailing public support levels for presidential candidates. This data is periodically updated on FiveThirtyEight's website to reflect the latest polling information for the 2024 election. Each record includes information we need like state in which the poll was conducted (or conducted nationally), polling dates (start and end), the specific pollster, the sample size of each poll, the polling methodology, and support percentages for each candidate. This dataset serves as the foundation for analyzing trends and building predictive models for the winner of the upcoming US presidential election. Table tbl-pollingdata below previews this information for polls conducted in 2023 and 2024.

To simulate, test, analyze, and clean the polling data, the statistical programming language R was employed (R Core Team 2023). Key libraries that supported the data analysis include `tidyverse` (Wickham et al. 2019), `readxl` (Wickham and Bryan 2023), `dplyr` (Wickham et al. 2023), `knitr` (Xie 2023), `ggplot2` (Wickham 2016), `kableExtra` (Zhu 2024) and `patchwork` (Pedersen 2024). Additionally, the `rstanarm` (Goodrich et al. 2022) library was utilized to implement multilevel modeling for prediction purposes. Also this analysis inherits the folder structure from Professor Rohan Alexander from University of Toronto (Alexander 2024).

2.2 Measurement

The process of transforming real-world phenomena into structured data typically involves several critical steps, from designing surveys to final data entry. First, we identify what needs to be measured. In this case, we focus on public opinion regarding the 2024 presidential election, particularly support rates for various candidates. It is essential to define the data scope (e.g., national or state level) and key variables—such as polling information (e.g., polling organization, start and end dates), sample details (e.g., sample size and demographics), and candidate details (e.g., candidate name and party affiliation).

Next, polling organizations develop survey questions based on the desired data, considering sample size, sampling methods, and survey modes to capture respondents' preferences and voting intentions. Different methodologies impact the results, as they reflect the views of distinct groups, which may be influenced by factors like access channels, demographics, and technology usage. For instance, face-to-face surveys tend to capture opinions from offline-inclined groups, whereas voluntary online surveys exclude those less active online.

Once the survey is designed, polling begins, and responses are aggregated by the polling organization. Respondents' answers are recorded and categorized, converting their opinions into structured data points, though these lack the nuances of human complexity. To protect privacy, data are often anonymized and aggregated, retaining key insights. After collection, raw data undergo cleaning and processing to minimize sampling biases and errors. Yet, any single poll can have random and systematic error sources, so combining multiple poll results helps mitigate these errors (Blumenthal et al. 2010).

Processed data are entered into structured datasets. Ideally, well-formatted polling data are stored as individual records, typically including fields like polling organization, sample size, survey dates, margin of error, candidate support percentages, and methodology description. On platforms like FiveThirtyEight, these fields are standardized for ease of comparison and analysis. Finally, FiveThirtyEight and similar platforms aggregate these structured records and present them to the public or allow non-commercial research use, with data usually updated as new polls are released. This enables analysts and the public to track changes in candidate support, analyze trends, and interpret results. An interesting phenomenon, however, is that as election day approaches, poll results often "converge" or "follow" the average, reducing polling error, with evidence suggesting some "final adjustments" in methodology may occur (Blumenthal et al. 2010).

2.3 Analysis Data

This analysis is focused on Donald Trump, so we only collected observations where the `candidate_name` is Trump. At the same time, we prioritize the quality of the polls. In a political approval rate prediction model, which is often influenced by numerous complex factors, we encounter data that is typically complex and noisy. Low-quality poll results can introduce significant bias into our prediction model, further impacting the accuracy of inferences. Therefore, only polls with a `numeric_grade` above 2.7 were selected. Additionally, Trump's main opponent, Kamala Harris, only entered the race after Biden's withdrawal, a major political event that could have significant effects on the support rates for both sides. As such, we focused solely on polls conducted after Biden's withdrawal to enhance the model's prediction accuracy for the final outcome.

In researching the likelihood of Trump winning, we selected specific variables to focus on those most impactful for predicting approval rates. Here is an overview of data used for this analysis

4

Table 1. Following are selected outcome variables and predictor variables, and explanations respectively.

2.3.1 Outcome Variables

In this analysis of Donald Trump's support rate in the 2024 U.S. presidential election, the outcome variable is Trump's support rate (i.e., pct) because it is directly related to the election results. The primary research goal of this analysis is to predict Trump's chances of winning in the 2024 election, so the support rate is the main metric we aim to model and forecast. The support rate here represents the proportion of respondents in a specific poll who support Trump. We seek to understand how the support rate is influenced by different factors, such as polling methodology, time, or geographical variables. By modeling the fluctuations in Trump's support rate, we can infer his electoral performance, and assess the impact of various factors on his popularity. Understanding these dynamics provides insights into Trump's standing relative to his opponents and supports our overall election prediction.

2.3.2 Predictor Variables

In this analysis, the predictor variables are the factors that we believe can influence the support rate for Trump. These variables are chosen based on their potential to explain changes in support rate. Following are the predictor variables that will be included in our model, along with an explanation of their significance.

state/national: Political environments and voter preferences may vary significantly by state, and national polls cannot fully reflect conditions in individual states. The electoral college system in the U.S. makes state-level approval rates crucial for predicting the likelihood of winning.

pollster: Different polling organizations use varying survey methods, sampling strategies, and data-processing procedures, which impact the reliability and accuracy of the poll results. Knowing which polling organization published the data helps assess potential biases in the data.

methodology: We anticipate different survey methods significantly impact the representativeness of the results, as they directly influence the characteristics of the sampled population. Some methods may introduce biases toward specific demographic groups.

is_national: This variable helps us distinguish between national and state-level polls, ensuring the model can adapt and adjust across different geographic levels.

days_since_Biden_Withdrawal: This variable shows the difference between the start day of the poll and the number of days since Biden's withdrawal. Trump's main opponent, Kamala Harris, entered the race only after Biden withdrew. Major political events like Biden's

withdrawal can significantly impact voter approval rates, and tracking the time since this event helps assess both its short-term and long-term effects.

Other potential variables (such as voter demographics like gender, age, race, and income) could indeed influence voter approval rates, but we found that their correlation with presidential election outcomes was not very strong. Additionally, individual voter characteristics are not consistently available in most polls. Thus, we did not include these factors.

Table 1: The First Few Lines of Cleaned Analysis Data

start_date	state/national	pollster	methodology	Trump Support Rate	is_national	days_since_Biden_Withdrawal
2024-10-23	Texas	Siena/NYT	Live Phone	52	0	94
2024-10-23	Texas	Siena/NYT	Live Phone	50	0	94
2024-10-23	Texas	Siena/NYT	Live Phone	52	0	94
2024-10-23	Texas	Siena/NYT	Live Phone	51	0	94
2024-10-23	Nebraska	Siena/NYT	Live Phone	55	0	94
2024-10-23	Nebraska	Siena/NYT	Live Phone	53	0	94
2024-10-23	Nebraska	Siena/NYT	Live Phone	55	0	94
2024-10-23	Nebraska	Siena/NYT	Live Phone	53	0	94
2024-10-23	Nebraska	Siena/NYT	Live Phone	42	0	94
2024-10-23	Nebraska	Siena/NYT	Live Phone	41	0	94

2.3.3 Relations Between Variables

The following figures, Figure 1 and Figure 2, display various aspects of data related to Trump's support rate, helping us understand how support rates vary with different factors. In Figure 1, we have marked the occurrence count of each pollster or state on the x-axis, helping us identify potential sources of bias in the data. Low-frequency items (such as data points from certain pollsters or states) may reduce the representativeness of the model—low-frequency observations imply insufficient data for specific pollsters or states, which may not accurately reflect voter support rates in these areas. This incomplete data could lead to unreliable predictions for these pollsters or states, increasing prediction errors. Additionally, low-frequency data is more susceptible to extreme values. With fewer observations, certain outlier data points (e.g., a particularly high or low support rate in a single poll) can appear disproportionately significant, affecting the overall prediction. When fitting the model to low-frequency data, it may be more influenced by the "noise" in these data points, potentially introducing bias into the overall results. Systematic biases in low-frequency observations could spread throughout the model, impacting the overall accuracy of the predictions.

Although low-frequency observations may introduce some bias, directly removing them would result in information loss, especially in election predictions where the representativeness of different states and pollsters is crucial. Different pollsters and states have their own resources and priorities, leading to uneven data coverage. Some pollsters may prefer to publish data at specific times, while certain states, due to smaller populations or stable voter behavior, may receive less attention. Thus, it is challenging to completely balance the data quantity across different items. Although some observations have low frequency, they still contain valuable

information. For instance, even though smaller or remote states may have limited data, their voter preferences are still significant under the Electoral College system. Completely deleting these observations would diminish these states' representation in the overall analysis, weakening the model's accuracy in reflecting support rates. Additionally, we aim to cover as many pollsters and states as possible with the model, so we retain even low-frequency observations. Thus, we keep these observations and mitigate the bias they introduce by combining other variables to build a more comprehensive and robust predictive model.

The plot on the left side in Figure 1 shows the average support rates for the top and bottom ten pollsters, highlighting each pollster's most commonly used methodology. We can observe that the support rates between the top 5 and bottom 5 pollsters differ by about 10%, which helps us understand the bias introduced by certain pollsters' influence on support rates, as some pollsters may tend to show higher or lower support rates. This chart also examines the different methodologies used by these pollsters, as varying survey methods may affect the results by reaching different groups of voters. Focusing on these differences aids in assessing the reliability and bias of the data.

The plot on the left side in Figure 1 compares the average support rates across states with the national average, showing regional differences in voter preferences. Due to the Electoral College system in the U.S. presidential election, variations in state support rates can significantly impact the election outcome. This analysis highlights the differences between state and national polls, helping us understand Trump's relative support in each state. Paying attention to geographic disparities improves the model's prediction accuracy.

Figure 2 displays the trend in support rates over time for the top five pollsters with the most observations. We chose these pollsters because a sufficient number of observations is needed to accurately capture trend changes and reduce the impact of extreme values. Each subplot represents a pollster and shows the change in Trump's support rate since Biden's withdrawal. This helps us observe the dynamic change in support rates following major events, and we can also see how different pollsters respond to the same event. Analyzing time trends aids in understanding shifts in public opinion, providing more time-sensitive support rate predictions.

3 Model

Our model applies a hierarchical Bayesian approach, based on the framework proposed by Gelman, Hill, and Vehtari (2020). This method builds a layered structure that combines polling data collected by different organizations using various methodologies with regional and temporal trends to more accurately predict support rates on Election Day. Running within a Bayesian framework, our model employs a random walk prior to reflect the dynamically changing support rates following Biden's withdrawal in 2024. This prior distribution is based on historical election trends and the latest polling data, allowing the model to continuously update as new data becomes available.
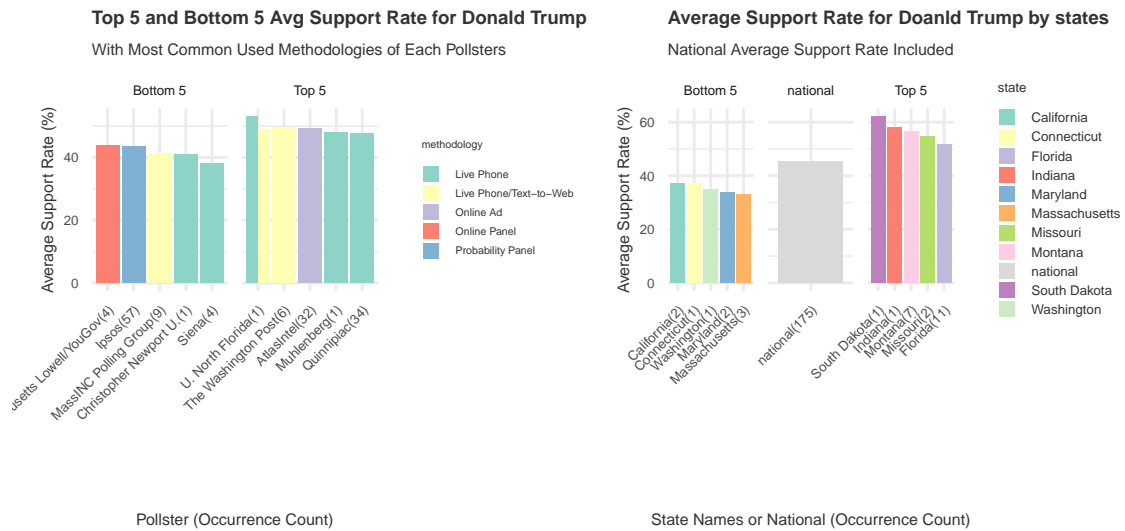
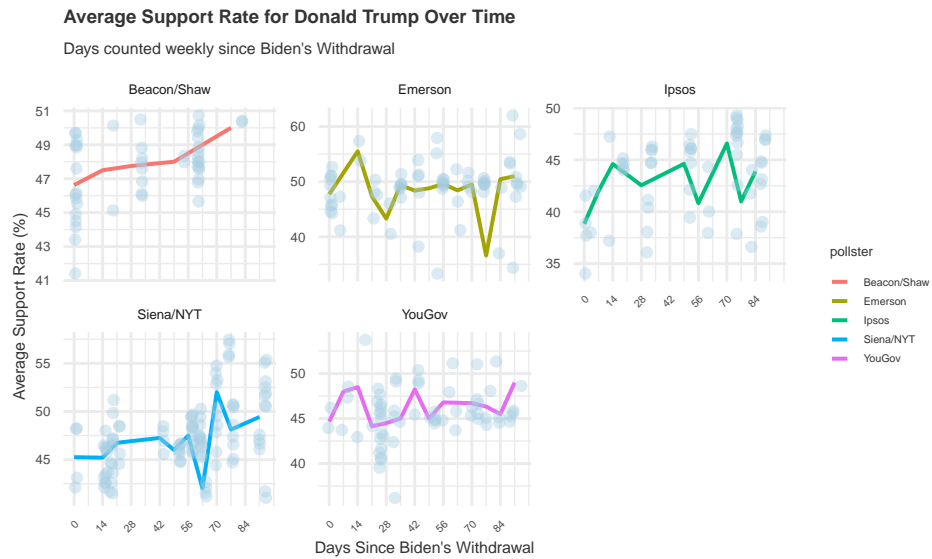Figure 1: Support rates for Donald Trump which showed in different variables plot 1

Figure 2: Support rates for Donald Trump which showed in different variables plot 2

The goal of our modelling strategy is to build a time series model to predict the percentage of support for Trump over time. This model incorporates multiple variables, including the polling methodology, the number of days since Biden's withdrawal from the race on 21 July 2024, whether the poll is at the national level, as well as random effects for each state and each pollster. These variables are chosen to capture the effects of time, polling characteristics, and geographic location on support rates, allowing us to obtain dynamic estimates of support for Trump and to predict the winner of 2024 U.S. presidential election.

Here we briefly describe the Bayesian analysis model used to investigate the factors influencing support for Trump in the polling data. The model incorporates both fixed effects, such as polling methodology and time since Biden's withdrawal, and random effects to account for variations across states and pollsters. Background details, model assumptions, and diagnostics are included in Appendix B.

3.1 Model set-up

This model is a generalized linear mixed model (GLMM), specifically:

$y_i \sim \mathcal{N}(\mu_i, \sigma^2)$

Define y_i as the percentage of support (pct) for Trump in poll in the i-th observation, assumed to follow a normal distribution with mean μ_i and variance σ^2. This model accounts for variations across different methodologies, time since Biden's withdrawal, and differences between national and state-level polls, with random intercepts for each state and random intercepts and slopes for each pollster.

The linear component of the model is defined as follows:

$$\mu_i = \alpha + \beta_1 \cdot \text{methodology}_i + \beta_2 \cdot \text{days\_since\_Biden\_Withdrawal}_i + \beta_3 \cdot \text{is\_national}_i + u_{\text{state}[i]} + (v_{\text{pollster}[i]} + \gamma_{\text{pollster}[i]}) \cdot \text{d}$$

where:

- y_i is the percentage of support (pct) for Trump in poll i.
- μ_i represents the predicted Trump support for poll i.
- $\alpha \sim \text{Normal}(45, 5)$ is the global intercept for Trump support.
- $\beta_1 \sim \text{Normal}(0, 5)$ represents the effect of polling methodology on Trump support.
- $\beta_2 \sim \text{Normal}(0, 5)$ represents the effect of days since Biden's withdrawal.
- $\beta_3 \sim \text{Normal}(0, 5)$ distinguishes national polls from state-level polls.

Random effects:

$u_{\text{state}} \sim \text{Normal}(0, \sigma_{\text{state}})$

$v_{\text{pollster}} \sim \text{Normal}(0, \sigma_{\text{pollster}})$

$\gamma_{\text{pollster}} \sim \text{Normal}(0, \sigma_{\text{pollster\_slope}})$

where:

- $u_{\text{state}[i]}$ is the random intercept for each state, allowing for state-level variations in Trump support.
- $v_{\text{pollster}[i]}$ is the random intercept for each pollster, capturing pollster-specific effects.
- $\gamma_{\text{pollster}[i]}$ is the random slope associated with the time variable (days_since_Biden_Withdrawal) for each pollster.

3.2 Model justification

3.2.1 Model Prior Selection, Assumptions, and Validation

We based our prior choices on historical polling data and previous research in electoral studies, aiming to provide informative priors without overly constraining the model. For instance, we set a prior on the intercept as Normal $\sim (45, 5)$, assuming Trump's support rate centers around 45%, with some uncertainty. Effect size priors (e.g., Normal $\sim (0, 5)$ for variables such as `methodology`, `days_since_Biden_Withdrawal` and `is_national`) reflect our expectation of moderate effects for these variables. This balance between informative priors and adaptability allows the model to accommodate data while providing a reasonable starting framework.

The model assumes that state-level differences and polling method effects are random and normally distributed. In reality, polling methods and state trends may exhibit non-normal behavior, especially during significant political events. Random effects for different polling agencies are introduced to account for biases associated with varying methodologies. However, we assume that each agency's methodology remains consistent over time, an assumption that may not hold if agencies adjust their sampling techniques. While the model effectively captures overall polling trends, it may not fully reflect subtle shifts within specific demographic subgroups (such as gender, race, socioeconomic status) since these details are not consistently available or may be obscured across poll results.

We implemented the model using the rstanarm package, leveraging MCMC sampling for Bayesian inference. To validate the model, we performed out-of-sample testing and calculated the RMSE (Root Mean Square Error), which resulted in 2.31. In the context of complex, socially influenced political surveys, this error level is considered acceptable. Convergence diagnostics in Figure 7 in Appendix B show that R-hat values are close to 1, and the Effective Sample Size (ESS) is generally large (often > 400), indicating good independence among samples. Sensitivity analyses on priors indicate that slight adjustments do not significantly alter the predictions, demonstrating model stability.

We also considered a simpler linear regression model which excluding hierarchical structures and a more complex model which incorporating additional economic data like state GDP and personal income. Through testing, we found that the simpler model could not adequately

capture differences across polling agencies and states, while the more complex model added computational cost with minimal gains in accuracy. Consequently, we chose the current approach to balance flexibility, interpretability, and computational efficiency.

3.2.2 Expectation of relationships between variables

We anticipate that different polling `methodology` may affect the estimation of support rates. While we cannot precisely determine which polling method might lean toward which demographic group, it is plausible to assume that methods like "online panels" or "phone surveys" might reach different segments of voters, thus introducing some bias into the support rate estimates. For instance, online surveys may engage more frequent internet users, while phone surveys may cover more traditional media users. This bias could lead to varying estimates of support for Trump.

For the variable `days_since_Biden_Withdrawal`, we hypothesize that, over time, Trump's approval ratings may fluctuate due to shifts in the election landscape, policy discussions, and changes in voter sentiment. This hypothesis is based on the dynamic nature of polling data: as elections near their end, poll results typically converge, reflecting more accurate forecasts for election day (see Blumenthal et al. 2010). We anticipate that the `days_since_Biden_Withdrawal` variable may exhibit a positive or negative trend, depending on the political climate and voter responses to Trump in particular periods.

Through the `is_national` variable, we distinguish between national and state-level polls, this variable enables the model to capture support rate differences arising from geographic and demographic differences. We expect that national polls, which typically have larger and more representative samples, are better suited to capturing overall trends, while state-level polls may reflect finer, localized biases and specific regional issues because national polls often encompass a broader and more balanced voter base, they provide a more stable forecast of overall voter trends. In contrast, due to geographic and demographic variations, state-level polls often exhibit greater fluctuations in error than national polls. These fluctuations stem from regional issues and diverse demographic characteristics, which may not align fully with national trends (see Kennedy et al. 2018).

Therefore, we included a random intercept (`1 | state`) for each state, with the intention that this random intercept would capture the differing impacts each state has on Trump's support rate. Similarly, by incorporating (`1 + days_since_Biden_Withdrawal | pollster`), we allow for variability in estimates by different pollsters over time. We anticipate that, due to differences in the stance or methodologies of these organizations, their estimates of Trump's support rate in the period following Biden's withdrawal will vary accordingly.

4 Results

4.1 Model Output

Our results are summarized in a table which is available here. This summary reveals the effects of different polling methods and organizations on our estimates. For instance, there is a small difference in estimates between data collected via online panels and phone surveys. The model adjusts for these differences by using a hierarchical structure, which accounts for inconsistencies caused by methodological and organizational biases. Specifically, variance parameters for different methods indicate the different impacts these methods have on sampling coverage and voter behavior, helping the model balance these differences.

The model's R^2 value is 0.641, indicating that it explains about 64% of the variation in support rates. This shows that the model effectively captures the main patterns in the data, although not every detail. Additionally, the Intraclass Correlation Coefficient (ICC) is 0.9, suggesting a high degree of clustering in the data, meaning that factors like state-specific and polling organization differences significantly contribute to overall variability. Other indicators, such as WAIC (2621.8) and RMSE (2.31), further confirm the model's good fit, showing that it maintains predictive accuracy without overfitting. Also, to balance precision and flexibility, we set a polling error rate of 3% based on historical data. This helps the model capture typical uncertainty without becoming overly confident.

4.2 Prediction Output

From Figure 3, the model predicts that nationwide support for Trump on Election Day, November 5, 2024, will center between 45% to 50% and shows moderate fluctuations. This variability is mainly attributed to national events and differences in polling methods. For instance, nationwide events such as economic shifts or unexpected political occurrences may influence voter sentiment, while differences in polling methods (such as phone surveys versus online panels) may introduce slight fluctuations in estimates due to differences in sample coverage.

At the state level, Figure 3 shows predictions indicate significant variations between states. For example, support in Texas is considerably higher than in California, with a gap of over 10 percentage points, while swing states like Pennsylvania and Michigan display greater volatility. This result further validates the model's robustness; the introduction of a random intercept for each state reflects systematic differences across states. This design aligns with data trends, suggesting that states with diverse geographic and demographic profiles exhibit unique political tendencies.
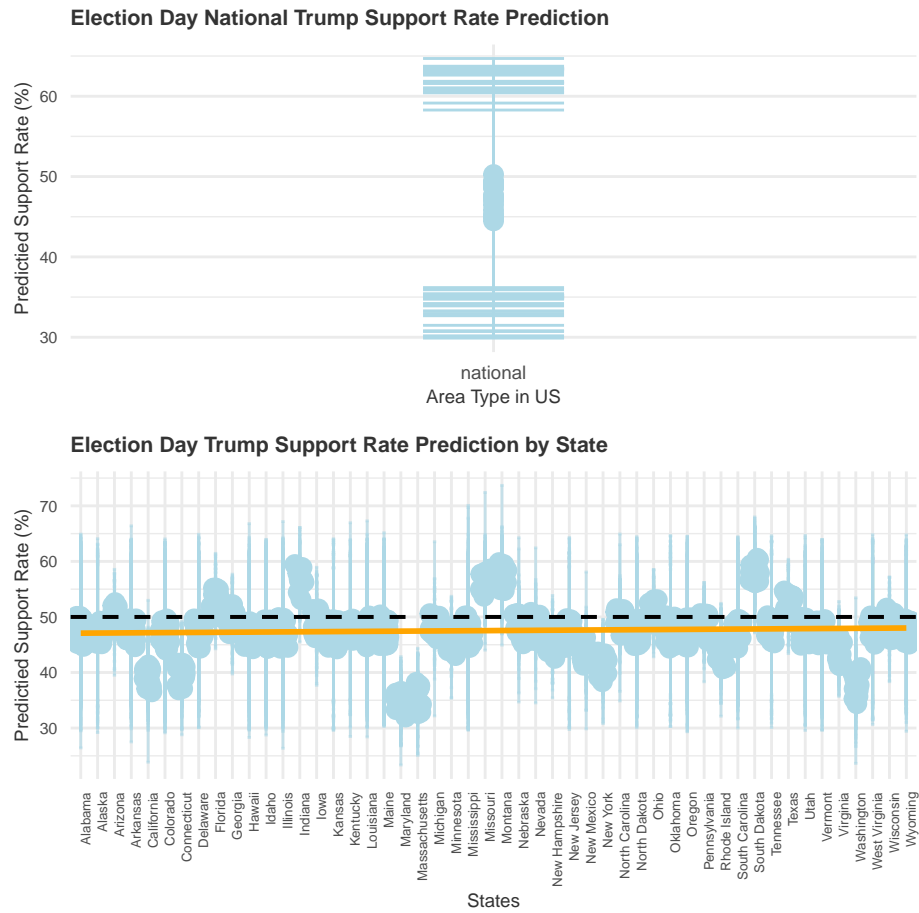
Figure 3: Prediction of average support for Donald Trump on 2024-11-05

5 Discussion

5.1 Winner Prediction

Based on the support rate predictions for Trump at both the national and state levels displayed in Figure 3 in Section 4, it appears challenging for Trump to emerge victorious in the upcoming U.S. presidential election on November 5, 2024, against his primary opponent, Harris. The first chart shows the national support rate prediction for Trump, centering around 45% to 50%. Historically, any candidate needs to reach or exceed the 50% threshold to secure a decisive victory. The current predictions reveal some level of uncertainty, although minor shifts in voter sentiment might occur leading up to Election Day, leaving Trump with some possibility of winning.

The Figure 3 also illustrates the support rates at the state level, revealing significant differences between red and blue states. For instance, Trump's predicted support rate is higher in traditionally conservative states such as Texas, while in strong Democratic states like California, his support rate is lower. In swing states like Pennsylvania and Michigan, the support rate hovers around 50%, aligning with historical trends where these states tend to have close contests.

To draw final conclusions from these predictions, it is essential to consider historical polling data and past election results. In previous U.S. presidential elections, shifts in sentiment in swing states have played a crucial role in determining the outcome. For example, the 2016 election highlighted the importance of swing states like Pennsylvania, Michigan, and Wisconsin, where polls underestimated Trump's support. A study by the Pew Research Center noted that last-minute sentiment shifts and polling errors led to an unexpected advantage for Trump (Pew Research Center 2021). Studies by *The Economist* emphasize the critical impact of swing states on election outcomes, also noting that national polls do not always reflect the dynamics of the Electoral College (see The Economist 2020). The Electoral College system can result in a candidate winning the popular vote but losing the election, or vice versa. This system has affected U.S. elections multiple times, demonstrating the importance of state-level results relative to national polling trends.

Given the current support rate predictions from our model and the historical performance of swing states, it is projected that Trump cannot secure a clear victory. Furthermore, the volatility in swing states suggests that if these states lean Democratic as they did in 2020, Harris would likely win. However, the final outcome will depend on voter turnout, the last stages of the campaign, and any major events leading up to Election Day.

5.2 Weaknesses and next steps

In discussing the limitations of polls and our model, we aim to evaluate its robustness and the challenges it faces in capturing the complexities of voter behavior. Each step, from data

collection to model construction, involves certain assumptions and potential biases. Below is a detailed discussion of these aspects:

5.2.1 Defects of Existing Data

Firstly, a major challenge for the model lies in estimating a large number of parameters with limited polling data. This scarcity of data can make the model highly sensitive to choices in priors and interstate covariance structures. For example, relying on historical patterns from states with limited data may fail to reflect the latest shifts in voter sentiment. This issue is also seen across different polling organizations and survey methods. As Gelman, Hill, and Vehtari (2020) noted, the model can fluctuate between overconfidence and unreasonable uncertainty. To address this, we could consider incorporating alternative data sources, such as social media sentiment or historical population voting trends, to supplement polling data in underrepresented areas. However, these sources carry their own issues, as they may lack reliability and rigor.

Then, in handling systematic bias introduced by different polling methods, the model accounts for variance by assigning specific variance parameters for each method (e.g., phone interviews vs. online panels). Each method may systematically overestimate or underestimate support for a particular candidate. For instance, online panels may attract more politically engaged individuals, leading to biases compared to random phone interviews. To reduce these biases, the model includes variance adjustments specific to each method. Further improvements could involve re-weighting polls based on historical accuracy or introducing dynamic adjustments for newer polling methods. However, as pointed out in Blumenthal et al. (2010), historical data often include responses to unique events that are not repeatable, making it difficult to assess the accuracy of historical information as a reference. Similarly, the model uses a random walk prior to accommodate fluctuations in public support over time, yet it may struggle to capture abrupt changes driven by political events, such as a debate or controversial statement. The inherent randomness of world events makes it challenging to fully address this problem.

Geographic factors are another focus, with special attention to swing states, which have historically close election results and volatile voter groups. Swing states pose unique challenges for predictive models, as their fluctuations are often inadequately reflected through historical patterns or limited recent polling. Additionally, smaller sample sizes in these states may lead the model to rely too heavily on prior assumptions. Potential solutions include increasing polling frequency in swing states, integrating voter registration data and demographic shifts, or introducing specific polling error adjustments for swing states to enhance model accuracy where small shifts could significantly impact election outcomes.

Furthermore, the polling data we focus on captures voter intent rather than voting probability. Factors such as likelihood of turnout or eligibility are concealed beneath the support rates, and changes in turnout due to major public events may affect election day outcomes. Because these details are not included in the polls, the model cannot fully capture these changes. If the

model could dynamically adjust voting behavior estimates based on actual turnout indicators, such as mail ballot requests or early voting numbers, it would better reflect changes in voter behavior and strengthen predictions of the presidential election outcome.

5.2.2 The Influence of Reality

5.2.2.1 Bias and Transparency

The influence of polling agencies' biases and methodological transparency on election predictions has long been a research focus. Every polling agency's choices—whether intentional or unintentional—can introduce biases into results. In highly polarized political environments, factors like question wording, sampling techniques, and demographic weighting can amplify potential biases. For example, research from the Pew Research Center and Stanford University shows that differences in methods, such as educational weighting or nonresponse adjustments, can lead to significant polling biases, especially for politically sensitive figures like Donald Trump Stanford News Service (2020).

Another complex issue is "partisan nonresponse bias," where certain political groups are less likely to participate in polls. This trend has become more prominent in recent years, notably affecting conservative respondents. Pollsters must take extra care in sampling to accurately represent all population segments, which presents particular challenges in practice and makes it hard to eliminate these biases entirely (see Pew Research Center 2024b).

Additionally, there is a feedback loop between published poll results and voter behavior: trends shown in polls can influence undecided voters or shift the enthusiasm levels of a candidate's supporters. The presence of "shy" or undecided voters further complicates the challenge of accurate polling. Despite methodological advancements, polling remains prone to errors that are often underestimated compared to the actual unpredictability of elections.

5.2.2.2 Financial Constraints

Financial limitations are a major barrier to achieving ideal polling coverage, particularly when trying to gather representative polling data across all U.S. states. Polling each state comprehensively is costly, especially in low-population states or states deemed "safe" for one candidate. In these cases, data scarcity means the model largely relies on data from more populous or swing states, which may not fully represent the diversity of the nationwide electorate, introducing potential biases.

According to a Pew Research Center report, polling biases often stem from financial constraints and methodological choices, such as the use of cost-effective online panels, which may not effectively capture perspectives from older, rural, or technologically less-adept groups. These groups often require different recruitment strategies and data collection methods, increasing polling costs (see Pew Research Center 2021). Another Pew report highlights that participation in extreme partisan regions is often low, which makes balanced representation more challenging

and increases the financial and logistical burden of implementing corrective weighting and sampling techniques (see Pew Research Center 2024a).

In theory, expanding polling coverage could improve model representativeness and accuracy. However, due to the high cost of polling and the statistical law of diminishing marginal returns, this expansion is often constrained by budget and logistical support. The concept of diminishing marginal returns means that as more resources (like time, money, or personnel) are invested in an activity, the additional benefit from each extra unit of investment gradually decreases. For instance, the first slice of pizza might be very satisfying, but by the third or fourth slice, enjoyment decreases. For polling, investing resources to cover more regions initially increases data representativeness significantly, but as major groups are covered, further investments yield smaller improvements, inevitably leaving certain groups underrepresented in the results.

5.2.2.3 The Balance Between Model Complexity and Interpretability

While our model integrates multiple parameters and hierarchical structures to address the variability across states and polling methodologies for a more accurate election support forecast, this complexity also introduces significant challenges in interpretability. For non-expert users, understanding how random intercepts or variance parameters influence the final prediction requires a certain level of statistical knowledge. Even though we attempt to make the model's main drivers (e.g., state-level effects or specific methodological impacts) more understandable through simplified summaries and visualizations, several issues remain.

Visualizations often struggle to fully capture the complex structure of our model. The model includes data from 50 U.S. states, multiple polling organizations, and various survey methodologies. If we attempted to display every detail within a chart, the visualizations would become overwhelmingly complex, making it difficult for readers to extract key insights. Conversely, if we limit the display to a subset, such as only highlighting key states or selected polling methods, this approach risks losing the completeness of information. These simplified visualizations fail to convey all sources of variation within the model, potentially leading readers to overlook significant factors and misunderstand the impact of omitted influences. Additionally, this partial representation can foster a feedback loop where oversimplified summaries influence potential voters' perceptions, impacting the trends in unpredictable ways.

5.2.2.4 The Randomness of the Real World

Voter preferences can shift rapidly due to events like debates, economic changes, or international incidents, challenging election predictions. Regular polling may fail to capture these shifts if not updated frequently, leading to outdated and potentially misleading forecasts. The Harvard Data Science Review highlighted this uncertainty in the 2020 election, where state-level opinion volatility was influenced by economic and educational factors, necessitating models that incorporate these dynamic interactions (Gelman, Hill, and Vehtari 2020).

With elections occurring only every four years, opportunities to validate prediction models are limited. Standard techniques like cross-validation are challenging to implement due to data scarcity, pushing models to rely heavily on historical data, which risks overfitting to past trends. Harvard Data Science Review notes that this reliance can result in models exhibiting excessive certainty or unreasonable uncertainty, especially when recent polling data is lacking.

Appendix

A Additional data details

B Model details

R package `bayesplot` (Gabry J 2024) is used for posteriors.

Due to space limitations, we have chosen to display only the posterior and prior comparison for one polling methodology variable in the posterior analysis. If we were to display all methodology variables, the chart would become overly complex and difficult to interpret. Although the effects of multiple methodology variables differ, displaying only one variable may make the model's diversity and sensitivity to biases across different polling methods appear insufficient. Here, we have selected a representative methodology, which allows us to avoid being distracted by excessive data. The output results for other methodologies are similar to IVR/Online Panel/Text-to-Web or hover around zero due to limited data.

B.1 Posterior predictive check

Figure 4 presents a comparative display of the prior and posterior distributions for selected parameters in our model. These parameters include the model's intercept (Intercept), a polling method-related variable ("mIVR/OP/T–W." representing `IVR/Online Panel/Text-to-Web`), a flag indicating whether the poll is national (is_national), the number of days since Biden's withdrawal (dys_snc_Bd_W.), and the model's error term (sigma). The prior distributions represent our initial expectations for these parameters before incorporating the data, while the posterior distributions reflect the influence of actual data on these estimates. By comparing the priors and posteriors, we can observe how the data has refined the estimates for each parameter. For instance, the prior for the IVR/Online Panel/Text-to-Web variable is relatively broad, indicating initial uncertainty about its impact, whereas the posterior distribution is much narrower, suggesting that the data supports a certain level of effect from this polling method.

The intercept here represents the baseline estimate of Trump's support rate when all other variables are zero. The posterior distribution for the intercept is centered around 50, suggesting that the model's baseline estimate for Trump's support rate is approximately 50%, which is relatively high. Additionally, the posterior distributions for `is_national` and `days_since_Biden_Withdrawal` are close to zero, indicating that these variables have minimal or uncertain impact on the support rate. The error term (sigma) represents the model's estimation of uncertainty in the data. A higher posterior mean for sigma suggests that the model detects a moderate level of variability in the data, possibly due to fluctuations in public opinion across states or over time.

We observe that the posterior distribution for `is_national` is concentrated near zero, indicating that there is no significant difference in support rates between national and state-level polls. Similarly, `dys_snc_Bd_W.` is a variable capturing the time elapsed since Biden's withdrawal, intended to reflect any time-related influence. The posterior distribution near zero suggests that this variable has no substantial effect on Trump's support rate, possibly indicating that public opinion on Trump has remained relatively stable over time. The `mIVR/OP/T--W.` parameter represents a specific polling method ("IVR/Online Panel/Text-to-Web"), which is included to capture the impact of different polling methods on support rates. The posterior indicates that while the influence of this method is relatively limited, it still contributes some variability to the model.
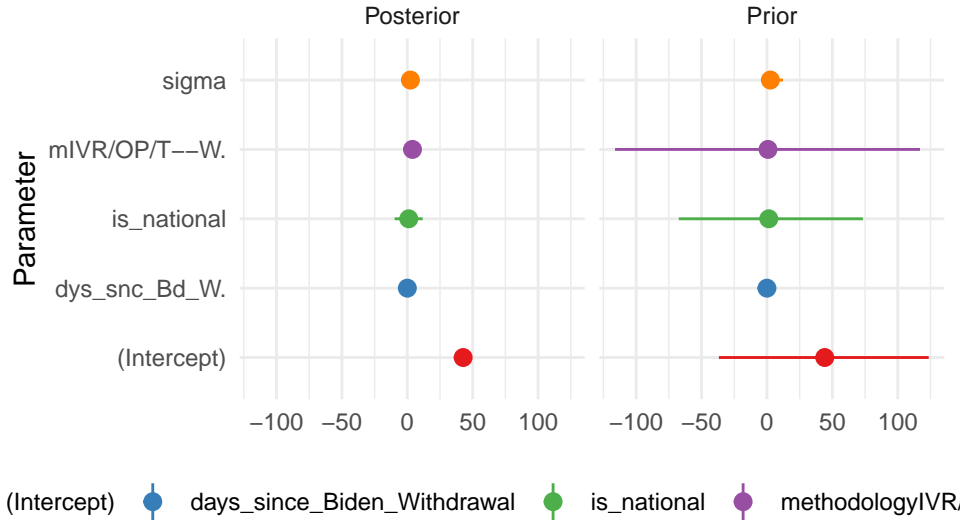

Figure 4: Comparison of Prior and Posterior Distributions for Selected Model Parameters

Figure 5 shows a comparison between the distribution of observed data y and the distribution of simulated data y_{rep} generated by our model. The dark solid line represents the distribution of the actual observed data, while the multiple light lines represent the distribution of multiple simulated datasets generated by the model. This comparison helps assess the model's fit quality that if the distribution of simulated data y_{rep} closely matches that of the observed data y it then suggests that the model effectively captures the characteristics of the observed data, demonstrating good predictive capability and a reasonable posterior distribution.

In Figure 5, we see that the distribution of the simulated data (light lines) aligns closely with the real data distribution (dark line), especially in terms of the peak and shape of the distribution. This indicates that the data generated by our model is consistent with the actual data, suggesting that the model can adequately replicate the distributional characteristics of the real data. This comparison validates the model's predictive ability and reliability, giving us greater confidence in the model's subsequent predictions.

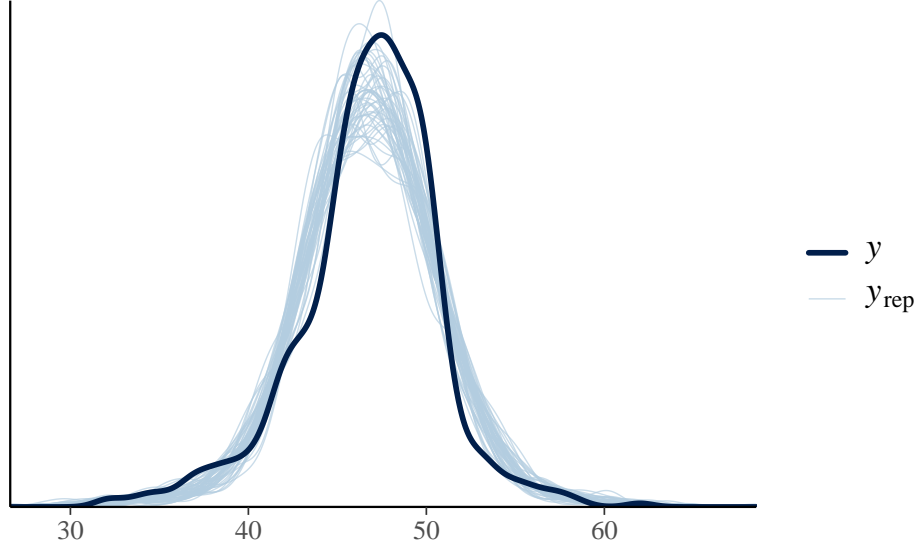Figure 6 shows the MCMC sampling trace plots for four key parameters in our model, including

Figure 5: PP Check for Model

the `IVR/Online Panel/Text-to-Web` method variable, the `days_since_Biden_Withdrawal` variable, the error term `sigma`, and the intercept term. From Chain 1 to Chain 4, with each chain representing an independent MCMC sampling process. Our expectation is that each sampling chain will stabilize and fluctuate around a central value without showing a clear trend (i.e., the mean of the samples converges over time), and there will be no significant differences between the chains. If these conditions are met, it indicates that the model has reached convergence, and the posterior distribution is reliable. From Figure 6, we observe that each parameter's sampling chains display a certain level of stability, with the chains fluctuating around a central value after reaching a steady state, without showing upward or downward trends. The distributions across chains are similar, indicating consistency in sampling results and supporting model convergence. This suggests that our model has successfully achieved a stable state within the posterior distribution, and the sampling results are reliable.

This plot shows the $\hat{R}$ diagnostic statistic for our model, which is commonly used to assess the convergence of MCMC sampling. Ideally, the $\hat{R}$ value should be close to 1, indicating consistent variance across different MCMC chains, signifying that the model has reached a stable state. The horizontal axis represents the $\hat{R}$ values, where values closer to 1 indicate better convergence. Our expectation is that all parameters have $\hat{R}$ values below 1.05, indicating minimal variance differences between chains and consistency across them. This plot shows that nearly all $\hat{R}$ values are concentrated around 1, with all falling below the threshold of 1.05 (indicated by the vertical dashed line on the right), suggesting that the sampling chains have converged well, providing high reliability for the model results.

Figure 8 shows the posterior distributions of certain parameters in our model, including the `(Intercept)`, a methodology variable `methodologyIVR/Online Panel/Text-to-Web`, a time
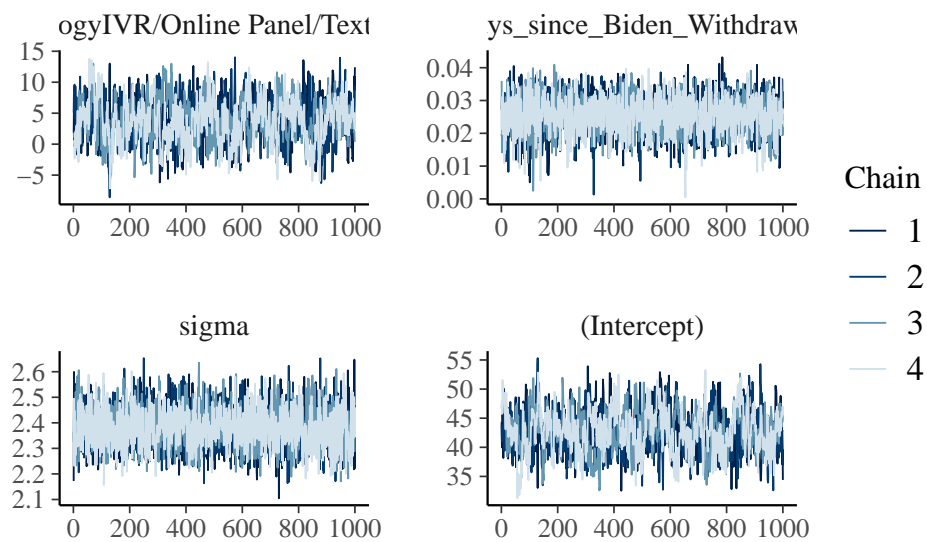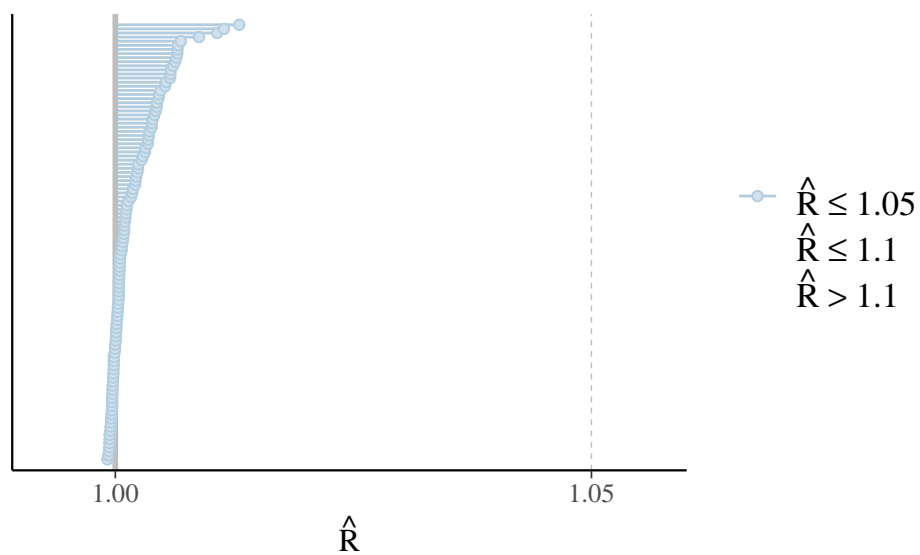
Figure 6: MCMC sampling trace plots for model

Figure 7: R hat Diagnostic statistic for model

variable `days_since_Biden_Withdrawal`, and the error term `sigma`. From the shape and location of the posterior distributions, we can see the degree to which the observed data supports the estimates of these parameters and the range of uncertainty around them. The `(Intercept)` distribution is centered around 30% to 55%, suggesting that the model estimates a 45% baseline support rate, which aligns with our expectations for Trump's support rate. The methodology variable shows a relatively wide distribution, indicating that the data does not provide strong directional support for its influence, suggesting that it may have a minimal or uncertain impact on the support rate. Also, the time variable has a posterior distribution close to zero, indicating that the number of days since Biden's withdrawal does not significantly impact the support rate. The error term `sigma` is at a low level, suggesting that the model views the data as having low volatility and being relatively concentrated. Thus, we can conclude that the model effectively captures the significant impact of the baseline support rate but but due to data uncertainty, it struggles to reveal the influence of factors like methodology and time. This phenomenon may reflect inconsistencies in the data across different poll methods or limitations in the the sample.
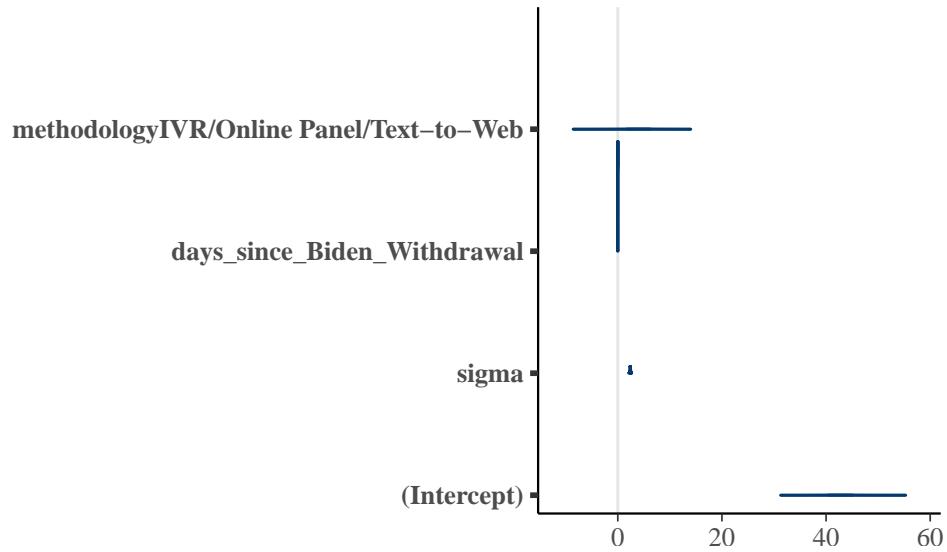

Figure 8: posterior distributions of four parameters in model

B.2 Diagnostics

References

Alexander, Rohan. 2024. "Starter Folder." https://github.com/RohanAlexander/starter_folder.git.

Ax, Joseph. 2024. "U.s. Democrats Deploying Funds to Boost down-Ballot Candidates in Non-Battleground States." 2024. https://www.reuters.com/world/us/us-democrats-deploying-funds-boost-down-ballot-candidates-non-battleground-2024-06-10/.

Blumenthal, Mark, Charles Franklin, Nate Silver, Scott Keeter, and Andrew Gelman. 2010. "Symposium: Polls, Forecasts, and Aggregators." *Public Opinion Quarterly* 74 (5): 845–65. https://doi.org/10.1093/poq/nfq076.

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

Gabry J, Mahr T. 2024. *Bayesplot: Plotting for Bayesian Models.* https://mc-stan.org/bayesplot/.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2020. *Regression and Other Stories.* Cambridge: Cambridge University Press. https://doi.org/10.1017/9781139161879.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Josh Christenson, Ryan King, and Diana Glebova. 2024. "Harris and Trump Hit \$2.5B in Campaign, PAC Fundraising but Don't Break 2020 Records." 2024. https://nypost.com/2024/10/28/us-news/harris-and-trump-hit-2-5b-in-campaign-pac-fundraising-but-dont-break-2020-records/.

Kennedy, Courtney, Andrew Mercer, Stephen Ansolabehere, Claude Berrebi, Matthew DeBell, Scott Keeter, Douglas Rivers, and Christopher Wlezien. 2018. "Disentangling Bias and Variance in Election Polls: Toward an Optimized Use of Polling Data for Forecasting." *International Journal of Public Opinion Research* 30 (4): 542–64. https://doi.org/10.1093/ijpor/edx021.

Pedersen, Thomas Lin. 2024. *Patchwork: The Composer of Plots.* https://patchwork.data-imaginist.com.

Pew Research Center. 2021. "Confronting 2016 and 2020 Polling Limitations." 2021. https://www.pewresearch.org/methods/2021/04/08/confronting-2016-and-2020-polling-limitations/.

———. 2024a. "Issues and the 2024 Election." https://www.pewresearch.org/politics/2024/09/09/issues-and-the-2024-election/.

———. 2024b. "Key Things to Know about U.S. Election Polling in 2024." https://www.pewresearch.org/short-reads/2024/08/28/key-things-to-know-about-us-election-polling-in-2024/.

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

Stanford News Service. 2020. "9 Things to Know about Election Polling and Data." Stanford University. https://news.stanford.edu/stories/2020/10/9-things-know-election-polling-data.

The Economist. 2020. "How This Forecast Works: The Economist's 2020 US Presidential Election Forecast." https://projects.economist.com/us-2020-forecast/president/how-this-works.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and Jennifer Bryan. 2023. *Readxl: Read Excel Files.* https://readxl.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation.* https://dplyr.tidyverse.org.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r.* https://cran.r-project.org/web/packages/knitr/index.html.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax.* http://haozhu233.github.io/kableExtra/.