Prediction about 2024 U.S. Presidential Election Outcome Using Bayesian Modeling*

Increasing Vote Percentage for Donald Trump over Time

Yunkai Gu Anqi Xu Yitong Wang

November 4, 2024

This study employs Bayesian modeling to predict the 2024 U.S. presidential election outcomes, and analyzing public support trends for Donald Trump. We using national and state level data to model how Trump's vote percentage evolves as the election approaches. The model captures temporal and regional patterns in support levels, which revealing a rise in Trump's favorability, particularly in the pivotal states such as Pennsylvania, Michigan, and Florida. These findings are valuable for campaign strategists and policymakers, providing a detailed understanding of voter motivations and helping anticipate shifts that could affect the election outcome.

Table of contents

*Code and data are available at: https://github.com/Kylie309/2024-US-election-prediction.

1 Introduction

The 2024 U.S. presidential election is set to be an important event in American politics, marked by changing voter demographics, shifting public opinions, and a more divided electorate. In this study, we utilize Bayesian modeling to project the percentage of support for Donald Trump in the 2024 U.S. Presidential Election. Recognizing that traditional individual polls may involve biases and inconsistencies, in order to solve it, we aggregate data from a range of high quality polls to produce a more stable and reliable forecast. Our model focuses on both national voting trends and state level voting trends, to capture more details and make the predictions more precisely. Some critical states, where subtle changes in support could heavily impact the final election outcome.

The main estimand of this analysis is the anticipated level of Trump's support across both national and state levels, we employing two Bayesian regression models, and incorporate variables such as polling firm, sample size, geographic region, and poll recency to capture nuanced voter dynamics. One model focusing on national trends and another capturing state level dynamics. The nationwide model uses temporal data with natural splines to track Trump's support trajectory as the election approaches. The dual-model approach, which includes natural splines for time adjustments and fixed effects for state-specific dynamics, provides a nuanced view of how both national and local factors influence voter preferences. This two model allows us to incorporate the latest data and dynamically update the forecast as new polling information becomes available.

The results of this study reveal an upward trend in percentage of support for Donald Trump as the 2024 U.S. presidential election approaches. The Bayesian model focusing on national voting trends indicates that Trump's vote percentage is expected to rise from approximately 45% six months prior to the election to nearly 50% as the election date nears. The other Bayesian model, which incorporates state-specific effects, highlights regional differences on the vote trend for Donald Trump. Specifically, there are states which show significantly higher support for Trump and steady increasing trend, such as Indiana, Misourri, Montana and Ohio. Both Bayesian models help to provide positive expectations about the final election outcome for Trump, making optimistic predictions.

Our study construct a advanced and diversified way to forcast the outcome than traditional poll. Individual poll often suffer from biases, inconsistencies, and a lack of adaptability as election day approaches, since its limitation of method and subjectivity of initiator. By using Bayesian modeling across both national and state levels, this study employs two models integrate data across multiple polls, which not only reduces biases associated with individual polls, but also reveals how support varies by state and changes over time. It provides a dynamic and precise forecast of voter preferences. This level of detail is essential for campaign teams and political analysts seeking to understand not just who leads, but where and when critical shifts in support may occur, allowing for crucial adjustments in messaging, resource distribution, and voter engagement efforts. Furthermore, it offers a rigorous framework for future election predictions, and emphasis the importance of adaptive modeling in capturing the true dynamics of voter's decision in real time, rather than relying solely on static opinion.

The remainder of this paper is structured as follows:

Section 2 introduces the overview of the data (Section 2.1), measurement (Section 2.2), data cleaning process (Section 2.3), as well as explanations, descriptions, table and graph summaries of outcome (Section 2.4) and predictor variables (Section 2.5) of the study. Section 3 explains the modeling process in detail, including procedure of model set-up (Section 3.1) and model justification (Section 3.2). Then, Section 4 presents the prediction outcome and results by plots, and Section 5 discusses the results and models in a broader context.

Appendix includes four parts: Appendix A presents methodology overview and evaluation of one certain pollster - Emerson; Appendix B provides detailed idealized methodology and survey design for the poll; Appendix C describes the data cleaning process in detail; Appendix D presents additional details during modeling process, including posterior predictive check Appendix D.1, comparison between posterior and prior distributions Appendix D.2, and diagnostics Appendix D.3.

2 Data

2.1 Overview

The analysis uses the dataset of national presidential general polls from FiveThirtyEight (FiveThirtyEight 2024). Following Alexander (2023), we consider to make predictions about the election outcome based on the data.

The analyses presented in this paper were conducted using R programming language (R Core Team 2023). The `tidyverse` packages (Wickham et al. 2019) were used in the process of data simulation, testing beforehand. After the original raw data was downloaded by using `tidyverse` package (Wickham et al. 2019), data cleaning process was done by using `tidyverse` package (Wickham et al. 2019), `lubridate` package (Grolemund and Wickham 2011), and `arrow` package (Richardson et al. 2024). We use `testthat` package (Wickham 2011) to develop the test for structure and format of simulation and analysis data. Then, models were constructed using `tidyverse` package (Wickham et al. 2019), `lubridate` package (Grolemund and Wickham 2011), `rstanarm` (Goodrich et al. 2022) package, and `splines` package (R Core Team 2024). The model results are then presented by `modelsummary` (Arel-Bundock 2022) package, and graphs were made with `ggplot2` package (Wickham 2016) and `RColorBrewer` package (Neuwirth 2022). Tables were constructed with `knitr` package (Xie 2021).

This analysis utilizes polling data from FiveThirtyEight (FiveThirtyEight 2024), a reputable source renowned for its aggregation and analysis of political polls. The primary dataset encompasses national presidential election polls for the 2024 election cycle, detailing voter preferences across various surveys conducted by authoritative organizations.

Although alternative datasets, such as those from Gallup and Ipsos, provide comparable national and state-level polling data, they were not included in this analysis for several reasons. The "poll-of-polls" methodology employed by FiveThirtyEight (FiveThirtyEight 2024) enhances its reliability, and the selected pollsters are well-respected, further bolstering the dataset's credibility. Additionally, FiveThirtyEight (FiveThirtyEight 2024) offers quality criteria variables that facilitate an evaluation of the data's reliability.

Consequently, focusing on the FiveThirtyEight dataset allows us to leverage the most robust available information to derive meaningful insights regarding predictions for the 2024 presidential election.

2.2 Measurement

To measure primary data, the raw polling data from FiveThirtyEight is structured to capture voter sentiment regarding the 2024 U.S. Presidential Election. Each dataset provides a snapshot of public opinion, transforming complex perceptions into quantifiable metrics, while compiling results from multiple polling organizations to present a comprehensive view of voter preferences.

The datasets from FiveThirtyEight utilize a "poll-of-polls" methodology, which aggregates results from various polls conducted by different organizations. This aggregation minimizes biases and enhances the predictive accuracy of the findings, establishing FiveThirtyEight as a more reliable source for forecasting election outcomes compared to datasets derived from singular polling organizations. Furthermore, FiveThirtyEight offers extensive metadata for each poll, including detailed methodology and sample demographics, which are critical for evaluating the quality and representativeness of the data. This level of transparency, such as numeric grading, is not consistently available in other datasets, rendering FiveThirtyEight's offerings more suitable for rigorous analysis.

Polling methodologies encompass specific techniques for data collection, typically employing random sampling to accurately reflect the broader electorate, often supplemented by stratified sampling to ensure diverse demographic representation. Surveys are conducted using various methods, including telephone interviews, online surveys, and face-to-face interactions, with the mode of administration influencing participant selection and response behavior, ultimately impacting the results. The specific poll methodology will be further elaborated upon in the appendix, focusing on one of the pollsters.

2.3 Data Cleaning

For the analysis in the paper, we derive two analysis datasets from the primary datasets which focus separately on national range and stage perspectives. Each entry in both datasets corresponds to a poll conducted by each pollster capturing the percentage of support for Donald Trump and number of days. To ensure the quality of data, we clean the data by only picking polls with numeric grades over 2.9. New variables are constructed for further analysis convenience. More details will be explored in appendix.

2.4 Outcome Variables

The outcome variable for this analysis is the support percentage for Donald Trump, represented as the pct variable in the datasets. This variable indicates the proportion of respondents who express support for Trump in various polls, providing a clear metric for evaluating his popularity over time as the 2024 U.S. Presidential Election approaches.

To better understand the distribution of the outcome variables, visualizations and summary statistics are realized. We will only use the national level data to plot and build tables since the distribution and summary statistics of it in national level data and state data level are approximately similar.

Figure 1 is plotted to present the support percentage of Donald Trump. The x-axis represents the national support percentage of Donald Trump ranging from 30% to 60%, while the y-axis indicates the counts of polls that fall within each support percentage range. The histogram

reveals a peak around 45% to 55%, suggesting a stable voter base, while the right-skewed pattern indicates a higher concentration of lower support percentages, with fewer polls reporting above 55%. Instances of outlier polls below 40% and above 55% reflect fluctuations in public opinion, potentially influenced by recent events and media coverage. The concentration of support suggests that Trump's backing may remain in the 45%-55% range as the election approaches; however, the presence of outlier polls indicates potential volatility, highlighting the need for ongoing analysis of polling data. Overall, the histogram provides critical insights into Trump's voter sentiment, emphasizing both the strength and variability of support leading into the election.

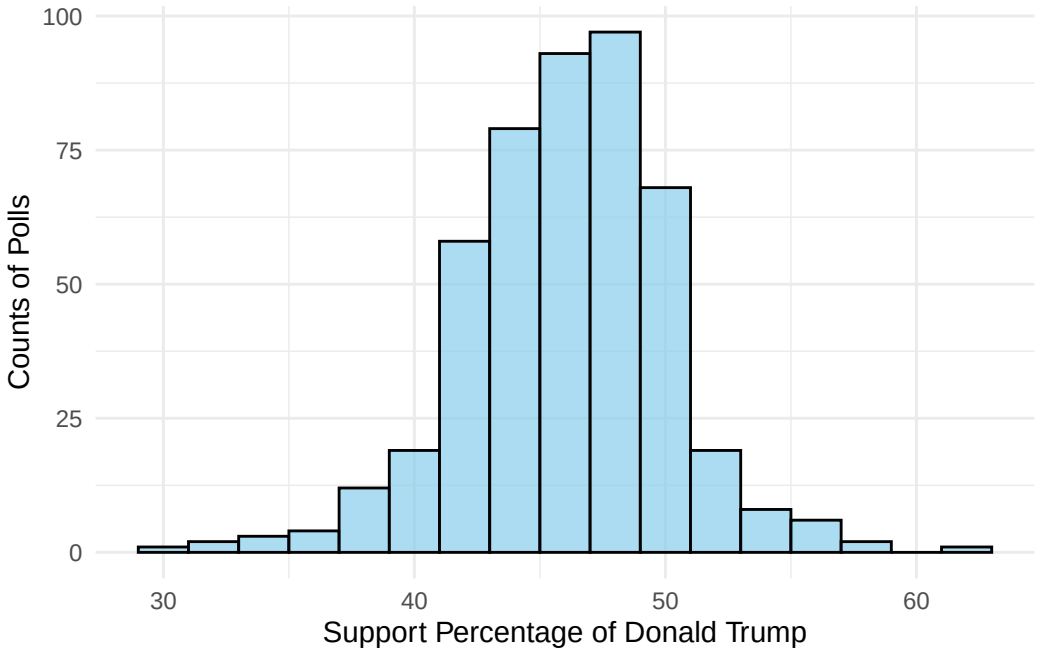


Figure 1: Support Percentage Distribution of Donald Trump

Table 1 presents summary statistics for Donald Trump's support percentage. The mean support percentage is approximately 46.4%, indicating that, on average, polls report a moderate level of backing for Trump. The median support percentage is 47%, suggesting that half of the polls report support below this value, reinforcing the stability of his voter base. The standard deviation of approximately 4.1% reflects a relatively small dispersion in support percentages across the dataset, indicating that most polls cluster closely around the mean. The minimum recorded support percentage is 29%, highlighting instances of lower confidence in Trump's support, while the maximum reaches 62%, illustrating moments of strong backing among certain voter segments. These statistics collectively suggest a generally stable but slightly variable support for Trump, which is critical for understanding the dynamics of his electoral appeal and for making informed predictions as the election date approaches.

Table 1: Summary Statistics for Trump's Support Percentage

Mean	Median	SD	Min	Max
46.39597	47	4.067369	29	62

2.5 Predictor Variables

In this analysis, two key predictor variables are examined: the number of days since the start of the polling period, represented by end_date_num, and the state of the respondents, indicated by the state variable in the dataset. Understanding these variables is essential for interpreting their impact on Donald Trump's support percentage and for making accurate predictions regarding the electoral outcome.

2.5.1 Number of Days

The variable end_date_num represents the number of days since a specified reference date, capturing the temporal aspect of polling data. This variable allows for analysis of how support for Trump may change as the election date approaches.

We will also use the national level data to plot and construct table for number of days since the distribution and summary statistics of this variable in national level data and state data level are approximately similar.

Figure 2 illustrates the distribution of the number of days since the start of the polling period, represented by the end_date_num variable. The x-axis shows the number of days, while the y-axis displays the count of polls conducted. The distribution reveals several peaks, especially at the start of the polling period, indicating a higher concentration of polling activity soon after it began. Spikes in polling counts are notable at days 0, 30, and 60, suggesting that these intervals may align with key events or shifts in the political landscape that increased polling efforts. Numerous low-count bars throughout the histogram suggest consistent polling, albeit unevenly distributed across days. The overall distribution indicates active polling throughout the period leading up to the election, with frequency fluctuations likely driven by campaign events, debates, and major news. This analysis underscores the importance of timing in polling data when evaluating shifts in support for Trump.

As Table 2 shown, the mean number of days is approximately 96.17, suggesting that, on average, polls were conducted about three months into the polling period. The median value is 99.5 days, indicating that half of the polls occurred after this point, reflecting a relatively consistent polling effort throughout the period. The standard deviation of approximately 48.41 days highlights some variability in the timing of the polls, with certain polls occurring significantly earlier or later than the average. The minimum value of 0 days indicates that some polls were conducted right at the start of the polling period, while the maximum value

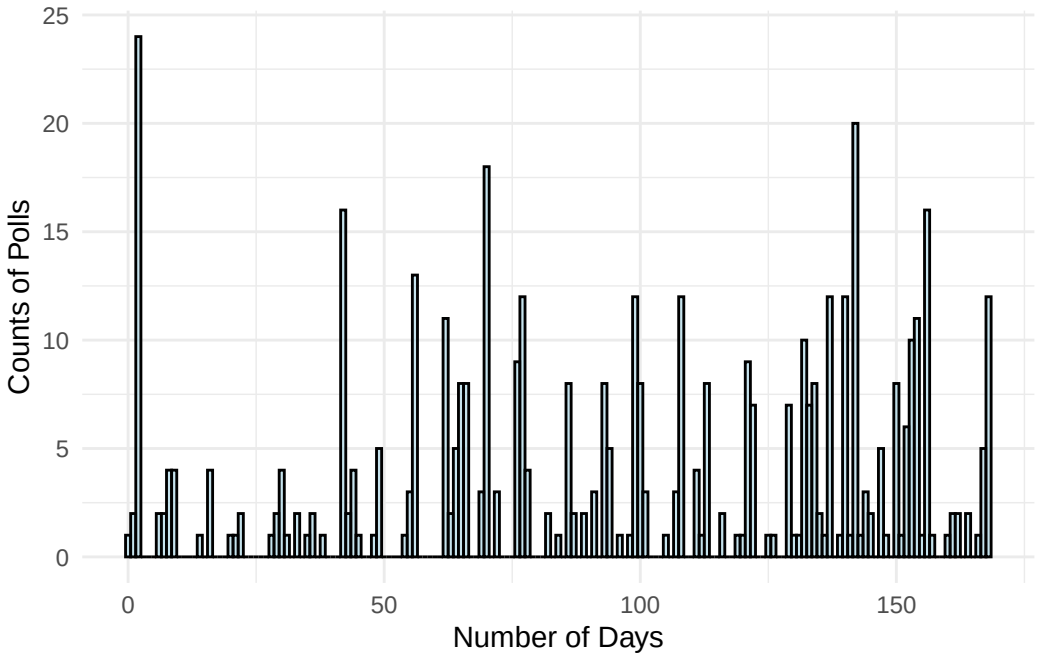


Figure 2: Distribution of Days Since Polling Period Started

of 168 days reveals that polling continued well into the latter stages, possibly capturing shifts in voter sentiment as the election date approached. Overall, these summary statistics suggest a sustained and varied polling effort over time, emphasizing the importance of considering the timing of each poll in the context of electoral events and public opinion trends as the election nears. This temporal analysis will be critical in understanding how support for Trump evolves leading up to the election.

Table 2: Summary Statistics of Days Since Polling Period Started

Mean	Median	SD	Min	Max
96.16737	99.5	48.40553	0	168

2.5.2 State

The state variable categorizes polling data by geographic location, providing a critical dimension for analysis. This variable allows for exploration of regional variations in support for Trump, which is crucial for understanding the electoral landscape.

Figure 3 illustrates the distribution of polling counts for Donald Trump across different states leading up to the 2024 U.S. Presidential Election in a bar chart. The x-axis represents individ-

ual states, while the y-axis shows the number of polls conducted per state. The chart indicates significant differences in polling activity, with states such as California, Florida, and Michigan having the highest counts, emphasizing their status as crucial battlegrounds with high voter interest and engagement. Notably, Michigan's high poll count points to a concerted effort to track voter sentiment in this key state. In contrast, states like Montana and Nebraska CD-2 show much lower polling counts, potentially due to less electoral competitiveness or limited polling resources. This distribution highlights geographic disparities in polling, showcasing where voter sentiment is closely monitored. Understanding these disparities is essential for analyzing Trump's support dynamics and anticipating electoral strategies as the election nears. The variation in polling frequency also implies that states with more polling might witness more pronounced shifts in voter sentiment, impacting campaign tactics and outreach initiatives.

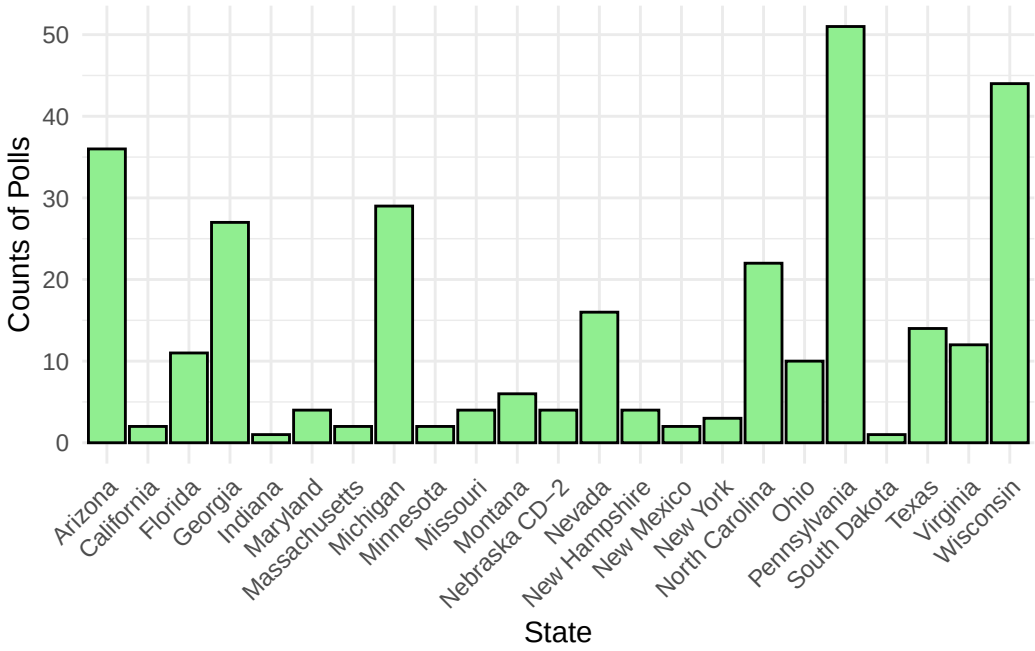


Figure 3: Counts of Polls by State

The summary statistics table, Table 3, outlines the counts of polling activity for Donald Trump across various states leading up to the 2024 U.S. Presidential Election. Pennsylvania, with 51 polls, leads in polling activity, signaling a strong focus on this key battleground. Wisconsin follows with 44 counts, further emphasizing its importance in the election landscape. Arizona (36 counts) and Michigan (29 counts) also show significant polling, underlining their competitive nature and strategic relevance. States like North Carolina (22 counts) and Nevada (16 counts) show moderate interest, suggesting they are important but not as heavily contested as leading states. On the other hand, states such as Montana, Maryland, Missouri, and New Hampshire have notably low polling activity, with only 4 to 6 counts each, indicating

potentially less competitive dynamics or fewer allocated resources. This table underscores geographic disparities in polling efforts, key for understanding where campaign strategies may focus and how voter sentiment might shift in critical states. The counts pinpoint battlegrounds that could decisively influence the election, highlighting the importance of targeted campaign activities in states with higher polling frequencies.

Table 3: Summary Statistics of State

state	Counts
Pennsylvania	51
Wisconsin	44
Arizona	36
Michigan	29
Georgia	27
North Carolina	22
Nevada	16
Texas	14
Virginia	12
Florida	11
Ohio	10
Montana	6
Maryland	4
Missouri	4
Nebraska CD-2	4
New Hampshire	4
New York	3
California	2
Massachusetts	2
Minnesota	2
New Mexico	2
Indiana	1
South Dakota	1

3 Model

The models constructed in the paper are Bayesian linear regression models designed to predict the percentage of support for candidate Donald Trump (referred to as pct) based on the number of days since May 5, 2024, i.e. half year before the election takes place (referred to as end_date_num).

Two Bayesian models are constructed separately.

Firstly, a Bayesian model with spline using response variable as the percentage of support (pct) and predictor variable as the number of days (end_date_num) is constructed to make predictions for vote percentage for Trump nationwide.

Then, a similar Bayesian model with the same response and predictor variables, but with an additional predictor "state", which contains information of the state in which the polling occurred, is constructed to include "state" as a fixed effect in the model with spline.

The goal of our modeling strategy is twofold.

Firstly, we aim to understand the overall trend in the percentage of support for Trump over time from a national perspective. This is accomplished through the first Bayesian model, which uses the number of days (end_date_num) as the sole predictor. This model allows us to investigate how support changes as we approach the election, providing insights into temporal dynamics in voter preferences nationwide.

Secondly, we seek to explore the impact of state-level differences on voter support through the second Bayesian model, which incorporates state as an additional predictor alongside the number of days. By treating state as a categorical predictor variable, this model enables us to assess how the percentage of support varies across different states overtime. This approach helps to understand regional influences on voter behavior, potentially uncovering important variations that the first model may overlook.

Following sections define, explain and justify each model and the variables, as well as discuss underlying assumptions, potential limitations, software used to implement the model, and evidence of model validation and checking.

Background details and diagnostics are included in Appendix D.

3.1 Model Set-up

The Bayesian models were implemented using the R programming language (R Core Team 2023), specifically utilizing the `rstanarm` package of Goodrich et al. (2022). This package provides an interface for fitting Bayesian regression models using Stan, and the models are fit using the package.

3.1.1 Bayesian Model with Spline for Vote Percentage Nationwide

Define y_i as the percentage of the vote that candidate Donald Trump received in the poll for each observation i, which is the response variable. Then define n_i as the number of days since May 5, 2024, i.e. half year before the election takes place.

The Bayesian model could be defined by the following mathematical expressions:

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + f(n_i) \tag{2}$$
$$f(n_i) = \sum_{k=1}^{K} \gamma_k B_k(n_i) \tag{3}$$
$$\alpha \sim \text{Normal}(50, 10^2) \tag{4}$$
$$\gamma_k \sim \text{Normal}(0, 5^2) \quad \text{for } k = 1, 2, \ldots, K \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

Combining all the components, the complete model can be expressed as:

$$y_i \sim \text{Normal}\left(\beta_0 + \sum_{k=1}^{4} \gamma_k B_k(n_i), \sigma^2\right)$$

This Bayesian model involves the vote percentage of Trump (y_i, pct) as a function of the number of days since half year before the election (n_i, end_date_num), modeled using natural splines.

Formula (1) specifies that the response variable y_i is normally distributed, where μ_i is the predicted mean percentage of support for the i-th observation, and σ is the standard deviation of the normal distribution, representing the variability in the observed support percentages.

Formula (2) specifies that the mean μ_i is modeled as a linear combination of an intercept α and a function of n_i, represented by natural spline f. The intercept of α represents the expected support when all predictors are zero.

Formula (3) specifies that a natural spline function f is applied to variable n_i, which represents the non-linear effect of time, and allows the model to flexibly fit the relationship between time and support. In this function, K is the number of basis functions, $B_k(n_i)$ are the basis functions of natural spline, and γ_k are the coefficients for these basis function.

Then, formula (4), (5) and (6) presents the priors:

Formula (4) defines the prior normal distributions for intercept α, with mean equals 50 and variance equals 10^2. This prior for intercept suggests that the average vote percentage is expected to be around 50% with some uncertainty.

Formula (5) defines the prior normal distributions for parameter γ_k, with mean equals 0 and variance equals 5^2. This prior for splines indicates prior belief that the effects of number of days (n_i, end_data_num) are likely to be small but allowing for flexibility.

Formula (6) assigns a prior to the standard deviation σ, where 1 is the rate parameter of the exponential distribution. Using such prior for variance reflects the belief that smaller values are more likely but allows for larger variations if needed.

3.1.2 Bayesian Model with Spline for Vote Percentage and State as Fixed Effect

Similarly as before, define y_i as the percentage of the vote that candidate Donald Trump received in the poll for each observation i, which is the response variable. Then define n_i as the number of days since May 5, 2024, i.e. half year before the election takes place.

The Bayesian model could be defined by the following mathematical expressions:

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{7}$$
$$\mu_i = \alpha + f(n_i) + \beta_{\text{state}} \cdot \text{state}_i \tag{8}$$
$$f(n_i) = \sum_{k=1}^{K} \gamma_k B_k(n_i) \tag{9}$$
$$\alpha \sim \text{Normal}(50, 10^2) \tag{10}$$
$$\gamma_k \sim \text{Normal}(0, 5^2) \quad \text{for } k = 1, 2, \ldots, K \tag{11}$$
$$\sigma \sim \text{Exponential}(1) \tag{12}$$

Combining all the components, the complete model can be expressed as:

$$y_i \sim \text{Normal}\left(\beta_0 + \sum_{k=1}^{4} \gamma_k B_k(n_i) + +\beta_{\text{state}} \cdot \text{state}_i, \sigma^2\right)$$

This Bayesian model involves the vote percentage of Trump (y_i, pct) as a function of the number of days since half year before the election (n_i, end_date_num), modeled using natural splines. The mean is modeled as a linear combination of an intercept, a natural spline function of n_i, and a fixed effect for the state.

Formula (1) specifies that the response variable y_i is normally distributed, where μ_i is the predicted mean percentage of support for the i-th observation, and σ is the standard deviation of the normal distribution, representing the variability in the observed support percentages.

Formula (2) specifies the mean structure. The linear combination is composed of three parts: a parameter α; a function $f(n_i)$ which represents the non-linear effect of the predictor n_i, modeled using natural splines; and the coefficient β_{state} for the categorical variable state, which captures the average difference in percentage across the different states. Treating state effects as levels (or random effects) may also include weakly informative priors to avoid overfitting and encourage regularization.

Formula (3) specifies that a natural spline function f is applied to variable n_i, which represents the non-linear effect of time, and allows the model to flexibly fit the relationship between time and support. In this function, K is the number of basis functions, $B_k(n_i)$ are the basis functions of natural spline, and γ_k are the coefficients for these basis function.

Then, formula (4), (5) and (6) presents the priors:

Formula (4) defines the prior normal distributions for intercept α, with mean equals 50 and variance equals 10^2. This prior for intercept suggests that the average vote percentage is expected to be around 50% with some uncertainty.

Formula (5) defines the prior normal distributions for parameter γ_k, with mean equals 0 and variance equals 5^2. This prior for splines indicates prior belief that the effects of number of days (n_i, end_data_num) are likely to be small but allowing for flexibility.

Formula (6) assigns a prior to the standard deviation σ, where 1 is the rate parameter of the exponential distribution. Using such prior for variance reflects the belief that smaller values are more likely but allows for larger variations if needed.

3.1.3 Assumptions of the Bayesian Models

Underlying assumptions for two Bayesian models are discussed below:

1. Normality of residuals:

The models assume that the residuals of the model (the differences between observed and predicted values) follow a normal distribution. This assumption is important for the validity of inference, and violation of this assumption can lead to incorrect inferences and unreliable credible intervals.

2. Homoscedasticity:

The variance of residuals is assumed to be constant across all combinations of the predictors. If the residuals exhibit heteroscedasticity (i.e., non-constant variance), this may lead to inefficient estimates and biased conclusions about the relationships modeled.

3. Independence of observations:

It is assumed that the observations are independent of each other. If there are correlated observations (e.g., repeated measures from the same individual), this could lead to biased estimates.

4. Functional form:

The relationship between the response variable and the predictor variable is assumed to be modeled appropriately, using natural splines to allow for flexibility. If the functional form does not adequately capture the underlying relationship, this can lead to biased estimates. The choice of degrees of freedom in the spline affects how well the model fits the data.

5. Prior specification:

The prior distributions for the parameters (i.e. intercept and coefficients) are assumed specified based on prior beliefs about their likely values. The choice of priors can influence the posterior estimates, reflecting reasonable beliefs about the parameters.

3.2 Model Justification

The Bayesian model with natural splines for predicting election outcomes with respect to the date is the most appropriate model because of several reasons. Firstly, the electoral dynamics for vote percentage could be highly complex and therefore non-linear. Various factors occurring before the election date could largely affect the outcomes. By employing natural splines, the model allows for the flexibility to capture these non-linear trends, presenting how support for a candidate evolves as the election date approaches, thus providing predictions about vote percentage.

Moreover, the Bayesian framework inherently accommodates uncertainty in parameter estimates, which is particularly valuable in political polling where outcomes can be unpredictable. The ability to incorporate prior information and update beliefs based on observed data enhances the robustness of the model, making it be the most appropriate model.

The above final model decision is made after considering several alternative models and variants. One alternative is a simple linear regression model (SLR), which assumes a constant linear relationship between the predictor (date) and the response (percentage support). Although the model is extremely straightforward and easy to interpret, it is far too simple for representing trends of and conducting predictions about election outcomes. SLR also applies plenty of strict assumptions to the population, the predictor and response variables, and violation of these assumptions, which is highly possible under this scenario, could lead to biased results.

Another option was to use a generalized linear model (GLM). While GLMs can handle various response distributions, they still rely on a linear assumption for the predictors, which may not adequately reflect the evolving dynamics of voter sentiment over time. Moreover, GLMs do not provide the same level of flexibility in modeling non-linear relationships as splines do.

After setting up the Bayesian model with splines, we expect a positive relationship between the percentage vote of Donald Trump and time passed since half year before the election. In other words, we anticipate that as the election date approaches, Donald Trump's support would increase. This expectation arises from the increasing support from minority voters, particularly from black voters (New York Post 2024b). Also, all seven battleground states (Pennsylvania, Wisconsin, Michigan, Arizona, Nevada, Georgia, North Carolina) are expected to shift to support Trump (New York Post 2024c). The vice-presidential debates that took place could be the major reason, which explain Trump's appeal of 'Unabashed machismo vibe', motivating especially young voters to support him (New York Post 2024a).

4 Results

4.1 Overview

The results from two Bayesian models are summarized in Table 4.

Table 4 shows positive coefficients for natural splines for both two models (nationwide and state-specific). This suggests that as the election date approaches (i.e. more days passed since half year before election), the percentage vote increases in a non-linear relationship.

Model 2 additionally discusses state effects, including coefficients for each state respectively. FOr instance, state Massachusetts has the smallest, negative coefficient (-15.02), indicating that the response variable of vote percentage is significantly lower in California compared to the reference state. On the other hand, state South Dakota has the largest, positive coefficient (9.12), indicating a relatively higher response for the state.

4.2 Prediction and Spline Fit for Vote Percentage Nationwide

Figure 4 demonstrates the results of the first Bayesian model, including predictions and spline fit for vote percentage of Donald Trump in the perspective of the whole nation.

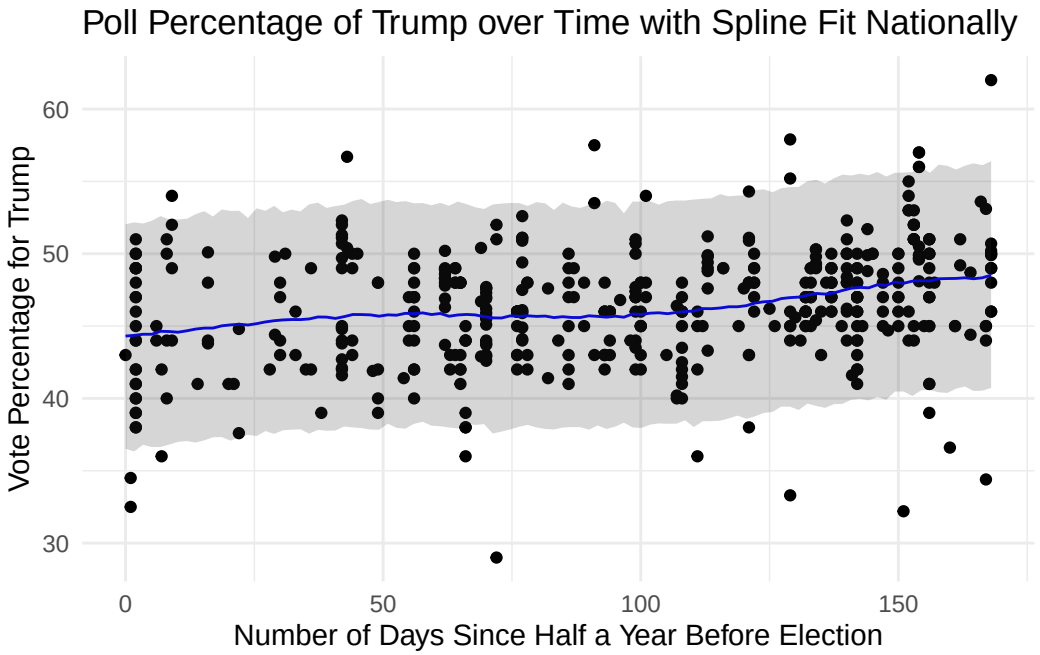


Figure 4: Predict Posterior Draws and Spline Fit for Vote Percentage Nationally

Table 4: Model Results of Trump Percentage Vote based on Date and State

	Model 1	Model 2
(Intercept)	44.27	45.59
ns(end_date_num, df = 5)1	1.19	0.21
ns(end_date_num, df = 5)2	1.60	2.92
ns(end_date_num, df = 5)3	3.10	1.85
ns(end_date_num, df = 5)4	5.58	6.28
ns(end_date_num, df = 5)5	3.17	2.94
stateCalifornia		−9.85
stateFlorida		3.62
stateGeorgia		0.24
stateIndiana		7.34
stateMaryland		−12.69
stateMassachusetts		−15.02
stateMichigan		−1.79
stateMinnesota		−1.91
stateMissouri		6.07
stateMontana		7.55
stateNebraska CD-2		−6.10
stateNevada		−0.36
stateNew Hampshire		−5.26
stateNew Mexico		−3.40
stateNew York		−5.42
stateNorth Carolina		−0.25
stateOhio		1.75
statePennsylvania		−1.42
stateSouth Dakota		9.12
stateTexas		1.24
stateVirginia		−4.38
stateWisconsin		−1.45
Num.Obs.	472	307
R2	0.089	0.603
R2 Adj.	0.068	0.557
Log.Lik.	−1310.219	−740.269
ELPD	−1316.8	−768.5
ELPD s.e.	21.8	11.9
LOOIC	2633.7	1537.0
LOOIC s.e.	43.6	23.9
WAIC	2633.7	1533.1
RMSE	3.87	2.67

The blue line drawn through the scatter points in Figure 4 indicates an upward trend in the data. It represents that as the time approaches the election (i.e. more day passed since half year before election), Trump's vote percentage is expected to be higher. Though the intercept is below 45%, indicating that half year ago Trump's support was lower than Harris, the percentage vote has climbed to nearly 50% with a steady trend over time. This phenomena notably illustrates a significant potential shift in voter support.

The trend aligns with the hypothesis that voter support typically intensifies as the election date nears, potentially due to increased campaign activities.

The black spots on the graph are individual observations. The shaded gray area represents the confidence interval for the predicted values, reflecting the uncertainty around the predicted values.

4.3 Prediction and Spline Fit for Vote Percentage by State

Figure 5 demonstrates the results of the second Bayesian model, including predictions and spline fit for vote percentage of Donald Trump in each state respectively.

The blue line drawn through the scatter points in each plot in Figure 5 presents the trend of Trump's vote percentage over time. For all 23 states listed, the blue lines indicate upward relationship between the response and predictor variables.

The individual plots for each state reveal diverse patterns of voter support, highlighting the complex electoral landscape of the United States. For some states such as Indiana, Misourri, Montana and Ohio, the vote percentage for Trump is higher than 50% for most of the time and is stably increasing. For other states, though the support is lower half year ago, the upward-sloping trend indicates final prediction about vote percentage to be potentially higher when the election date comes.

These state-specific analyses emphasize the importance of localized campaigning strategies, as voter preferences are influenced by state-specific issues.

5 Discussion

5.1 Overview of the Paper

This paper presented a analysis using Bayesian modeling to predict the outcome of the 2024 U.S. Presidential Election. This paper employed the "poll-of-polls" approach to make predictions, which aggregated multiple polls to minimize biases and enhance prediction accuracy.

The voting percentage of Donald Trump is the main response variable, and the number of days since half year before the election "end_date_num" and the regional location "state" are the predictors.

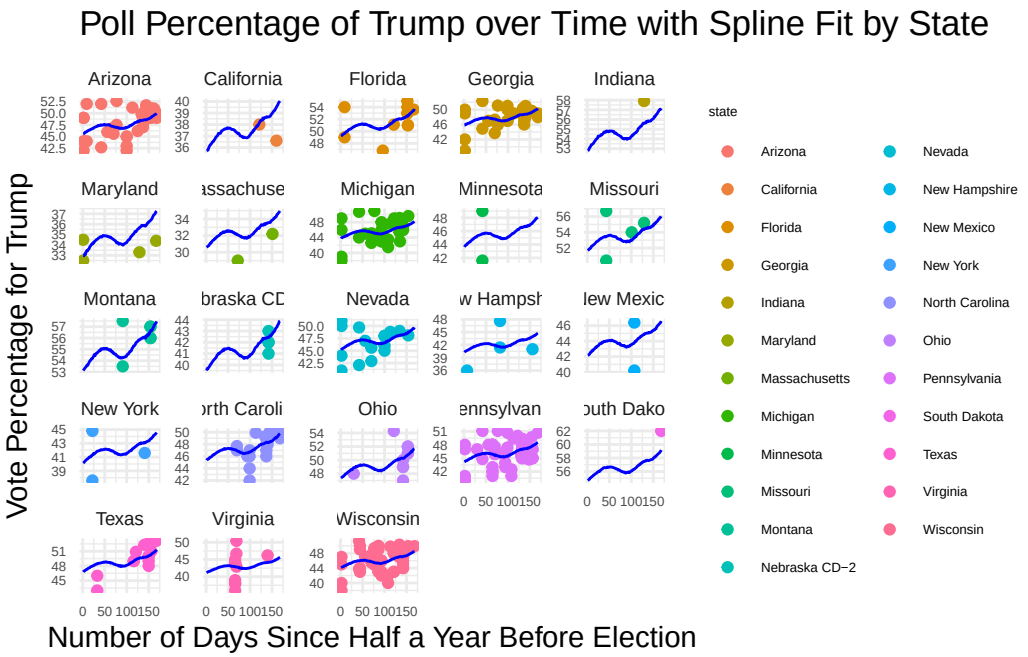


Figure 5: Predict Posterior Draws and Spline Fit for Vote Percentage in Each State

Two distinct Bayesian models were constructed: one that utilized nationwide data, and another that incorporated state-specific effects. The relationship between the response and predictor variables were captured by non-linear natural splines, providing trends of how vote percentage changes as time approaches the election date.

The Bayesian model with natural splines was chosen for analysis because it is appropriate for capturing complex electoral dynamics, allowing for flexibility, evaluating the trend across time, thus providing predictions about the election outcome.

The results investigated by the paper basically met our anticipation. As time approaches the election date, the predicted support for Trump increased stably, increasing from 45% half year ago to nearly 50% recently. Specifically speaking from state perspective, steady increase trend still persists in predictions of supports from people in different states, and several states have high vote percentage lead.

In conclusion, we predict that the voting and support inclination will continue to favor Donald Trump until the day of the election.

5.2 Fluid Nature of Voter Intensions

What we learned from this study is the understanding of how temporal dynamics influence voter preferences in electoral contexts.

Besides providing predictions about the upcoming 2024 election, the increasing trend in vote percentage for Donald Trump as the election approaches found by the paper suggests that voters are highly responsive to the political incidence. The model's ability to capture these shifts provides valuable insights for political strategists and analysts, emphasizing the necessity to continuously monitor voter sentiments in the lead-up to elections. As mentioned before, frequent presidential debates, promises made by politicians at the last minute, or a sudden outbreak of an event would incite more people to vote for candidate, or change their mind and switch to support the other.

This finding underscores the fluid nature of voter intentions, and the critical role of time in shaping electoral outcomes. For political strategists, the paper emphasizes the importance of imposing more provocative measures as the election date is nearer. For voters, this paper could help them understand the dynamic flow of political elections clearly, and help to maintain distinct, calm overview during election activities.

5.3 State as an Important Factor

Moreover, the second Bayesian model in the paper included state as a fixed effect, providing information on not only the electoral behaviors of the voters from a broad national perspective, but also on the regional variations in support trends.

By considering regional issues, the predictions and analysis about the vote tendency with respect to time could be conducted more completely, and enable people to view the reasons behind the formation of such preferences from the perspective of regional differences. For instance, the model identifies states like South Dakota, where Trump's support is relatively high, versus states like Massachusetts, which shows a comparatively low, negative coefficient.

The study reveals the disparities in voter support across different locations, highlighting the regional variations in electoral behavior, and reinforcing the idea that a "one-size-fits-all" approach to campaigning is highly inadequate. The finding encourages politicians to encourage politicians to "suit the remedy to the case", and illustrates that electoral strategies must be tailored to address the unique characteristics and concerns of voters in different regions in order to gain the trust and support of voters.

The findings also potentially indicate that other demographic factors such as age, race, and educational background could play a crucial role in shaping voter preferences. In particular, younger voters may exhibit different support patterns compared to older demographics, reflecting generational shifts in political engagement and priorities. The paper highlights the need for political campaigns to engage with diverse voter segments, recognizing that their motivations and preferences may differ significantly.

5.4 Weaknesses

Potential limitations of the Bayesian models are described below:

1. Overfitting with complex models:

Due to the complexity of the models, using splines with high degrees of freedom may lead to overfitting, where the model captures noise rather than the underlying trend. Overfitting reduces the model's generalizability to new data and diminishes the predictive validity for the models.

2. Missing data:

Some data included to make predictions nationwide using the first Bayesian model have missing values in the predictor "state", and was cleaned out during the process of making predictions about how the vote percentage varies across different states with the second Bayesian model. These missing values in key variables could lead to biases in the model. For instance, if significant segments of the population are underrepresented due to missing data, the conclusions drawn from the analysis may not accurately reflect the true electoral trend.

3. Assumptions violation:

If the assumptions and priors of the models are violated, situations where the models may not be appropriate could occur. For instance, if the population of the response variable (pct) exhibits a skewed distribution instead of the normal distribution assumed, the models would be inappropriate and provide biased results.

In reality, voter preferences may also exhibit clustering effects, particularly within communities or social networks. This could lead to correlated observations that the current models do not adequately account for, potentially skewing the results.

5.5 More to Be Done

Though applying complete statistical modeling and analysis, this study only provided predictions about the election outcome, and trends of vote support to one certain candidate over time. Further researches and investigations are essential to refine the predictive modeling included in this paper.

Firstly, more models could be considered to be used in the predicting procedure. One model type may seem to be too simple and have too many assumptions and limitations, making it difficult to provide an overview of such electoral context. More model validation and sensitivity analyses could be done to improve the modeling process, so that the robustness of the models against different assumptions and priors could be assessed to make the model more fit.

Secondly, data collection methodologies could be improved as well. In this study only one raw dataset was used, arising potential issues in true representativeness of the polling data. Finding and using more data would lead to more accurate and relaible predictions in future electoral analysis.

Last but not least, evaluations on the possible reasons and external factors behind the trends of the vote percentage could be investigated. For instance, significant media coverage of Trump's campaign activities or strategic appearances in battleground states could sway public opinion and drive changes in voter support.

Specifically, since the second Bayesian model included in this paper considered state as a fixed effect, further researches could analysis why such regional variations exist. Is this because people in each region have their own unique sets of values and beliefs, or is it because people in a certain region have experienced the same events? Future studies could explore the integration of additional predictors, and consider various factors such as socio-economic reasons, historical issues, and ethic identity.

Appendix

A Pollster methodology overview and evaluation: Emerson

A.1 Background of Emerson College Polling

Emerson College Polling is a famous U.S. polling organization as known for its precise and timely predictions, particularly in political forecasting such as the U.S, election. Emerson uses multiple methods approach, combining both traditional and modern techniques to capture a representative outcome of the election (Emerson 2024a). Emerson College uses a mix of landline calls, text-to-web links, and online panel surveys to reach a broad range of people. This approach helps them collect decisions from a diverse group of likely voters. In mid-October 2024, Emerson College Polling surveyed 1,000 likely U.S. voters to assess voter preferences for the upcoming presidential election. The poll results showed a competitive race, with Kamala Harris at 49% and Donald Trump closely following at 48%, reflecting a near split among voters (Emerson 2024c). Following appendix provides a detailed breakdown of ECP's methodology for this poll.

A.2 Population, Frame, and Sample

In polling, population, frame, and sample are basic concepts that determine the accuracy and reliability of the outcomes. An accurately defined population ensures the poll's focus, a robust frame prevents undercoverage, and a well-selected sample reduces bias, providing a clear representation of public thinking (ScienceDirect 2023). The more closely a sample resembles the actual population, the more reliable the conclusions drawn from it will be (Illinois 2023).

1. Population: The population in a poll is the complete group whose opinions the poll seeks to understand (Illinois 2023). For election polling, this group often includes all eligible or likely voters in a region or demographic segment. Emerson's target population consists of likely U.S. voters, determined based on their past voting records, registration status, and stated intention to vote (Emerson 2024c). For their 2024 surveys, they specifically target people who most likely to participate in upcoming elections, making adjustments to closely model the anticipated voter's tendency

2. Frame: The sampling frame is the actual list or database used to draw a sample from the population. It should ideally include all individuals in the population so each has a chance of selection (AAPOR 2022). A well-designed frame minimizes "coverage error," which occurs when parts of the population are excluded from the frame, thereby reducing bias. Emerson's sampling frame consists of U.S. Census data, voter registration files, and pre-validated online panels. This frame is carefully stratified by demographics like age, race, gender, education, and region to ensure that all key

groups are included, especially hard to reach populations such as rural or younger voters https://emersoncollegepolling.com/october-2024-national-poll-harris-50-trump-48/

3. Sample: The sample is the group of individuals selected from the frame to participate in the poll (ScienceDirect 2023). It represents the population in a smaller form, allowing pollsters to draw conclusions without surveying every person (AAPOR 2022). While large sample sizes can increase a poll's statistical precision, accuracy can still be compromised by biases from poorly crafted questions, flawed data collection, or improper analysis. Beyond roughly 1,000 respondents, the sample's quality and selection methods are often more important than sheer size (SciLine 2023). Emerson uses stratified sampling to set quotas that match the demographic composition of the U.S. electorate (Emerson 2024b), which helps ensure that the sample reflects key population characteristics. They gather responses through a combination of probability and non-probability sampling methods, including online panels, text-to-web (MMS-to-web) surveys, and landline IVR calls. For national polls, Emerson typically samples about 1,000 respondents and then weights the data to further align with U.S. demographic distributions, reducing sampling error and improving accuracy.

A.3 Sample Recruitment Methods

1. Cell Phone MSS-to-Web: Emerson applies a text-to-web approach (Emerson 2024a), sending potential respondents a survey link through MMS, this method is especially effective in reaching younger or mobile-first voters, they are more likely to response by one click rather than receiving an half-hour long call.

2. Landline IVR: IVR calls are used to reach older demographics or rural population (Emerson 2024a) who may be less engaged online or not familiar with new generation technology. The traditional method such as the landline calls are easier to accepted by them, and increases response rates among older adults and those in rural areas, which ensuring balanced representation across age groups.

3. Online Panels: Emerson also utilizes verified online panels (Emerson 2024a), where respondents' eligibility is cross-checked with voter file data, in order to exclude not potential voter's response, this step reinforces sample accuracy.

These recruitment methods help Emerson reach a wide audience, though self-selection bias may arise, particularly with online panel participants, as individuals joining these panels are often more politically engaged.

A.4 Sampling Approach's Trade-offs

Emerson's sampling approach uses stratification to set quotas based on demographic factors (Emerson 2024e), such as age, race, gender, education, and location. Post-stratification weighting further adjusts these quotas to align the sample with Census and voter registration data.

A.4.1 Advantages

1. Inclusivity: By applying a combination of recruitment methods, Emerson reaches a range of voter groups across demographics, which avoid the incomplete context representation bias.
2. Budget Saving: Digital methods such as text-to-web and online survey are generally less expensive than the traditional methods (landline call).
3. Effectiveness: Using text-to-web and online panels allows a quick receiving of the answers, for rapid adjustments and data collection.

A.4.2 Disadvantages

1. Non-Probability Limitations: Non-probability sampling means not all individuals have an equal chance of being selected, in some of Emerson's methods such as landline call, also known as robopolls or interactive voice response calls, are cost-effective but typically only reach landline users, the focus on landlines limits the diversity of respondents, impacting the sample's representativeness (SciLine 2023).
2. Access Bias: Voters without internet or mobile access may be underrepresented, particularly among lower-income or rural demographics, which can increase the incomplete representation bias.

A.5 Handling Non-Response

To address non-response bias, Emerson applies demographic weighting to underrepresented groups (Emerson 2024c), such as younger voters or certain racial demographics, they put more weight for these groups. Emerson occasionally provides incentives to respondents, which encourages survey completion. This is especially effective in boosting participation rates for online panel members, who may be less likely to complete surveys without additional motivation. However, people with lower political engagement levels may still be less likely to respond, which could impact representativeness.

A.6 Questionnaire Design's Advantages and Disadvantages

A.6.1 Advantages

1. Clarity and Style of Question: Emerson focuses on straightforward, unbiased questions to ensure respondents understand each item consistently (Emerson 2024d), closed-ended questions make voters are easy to give the answer without too much thinking, which decrease the drops-out.
2. Relevance and Adaptability: Questions are updated regularly to reflect current political events, enhancing the survey's relevance and accuracy.

A.6.2 Disadvantages

1. Simplicity: Online surveys typically rely on multiple-choice or straightforward questions (Emerson 2024d), which may limit deeper insights into complex opinions or voter's tendency.
2. Question Order Bias: The order of questions presented can unintentionally influence voter responses (SciLine 2023). For example, asking participants to rate their favorability towards Kamala Harris before asking about other candidates (Emerson 2024d) could mislead respondents, subtly framing how they perceive subsequent individuals. This can lead to "order effects," in polls, order effects can create a bias, especially if the initial question involves a polarizing candidate, it may cause strong emotions that influence answers to following questions.
3. Potential Response Fatigue: Frequent participation in polls may lead to rushed responses, impacting quality.

A.7 Conclusion

Emerson College's approach combines traditional and digital methodologies to reach a comprehensive and representative sample of U.S. likely voters. Even the inclusion of diverse communication methods enhances accessibility and speed, challenges such as self-selection bias and non-probability sampling limitations still appear. Despite these, Emerson's effective stratification and weighting strategies have made its poll significant in prediction of U.S. elections.

B Idealized Methodology and Survey

B.1 Overview

The idealized survey methodology is designed to forecast the outcome of the upcoming U.S. presidential election accurately. This approach will focus on a robust sampling strategy, targeted recruitment, comprehensive data validation, and effective poll aggregation. A budget of $100,000 will be allocated to ensure that each aspect of the methodology is fully supported and that the resulting data is accurate and representative.

B.2 Sampling Approach

B.2.1 Target Population

The target population consists of U.S. registered voters eligible to vote in the upcoming presidential election. This population includes diverse demographic groups, varying by age, race, gender, geography, and political affiliation, all of which are critical to accurate representation.

B.2.2 Sampling Frame

To build a representative sampling frame, recent national voter registration data from reliable sources, such as the U.S. Election Assistance Commission, will serve as the foundation. This data will be enhanced by voter profile databases from third-party aggregators like Catalist, which consolidate demographic, geographic, and political affiliation data. 'Survey Methodology' (Groves et al. (2011)) discusses how stratified sampling can increase the efficiency of survey estimates by dividing the population into subgroups and sampling each group independently. Thus, to ensure the sample reflects the diversity of the U.S. electorate, quotas will be set based on demographic data from the U.S. Census and Bureau of Labor Statistics, including age, race, gender, income, and urban-rural distribution.

B.2.3 Sample Size

We will take a sample of at least 3,000 respondents to balance statistical accuracy with budget constraints.

B.2.4 Stratified Random Sampling

Stratified random sampling enhances the accuracy of survey estimates because it ensures proportional representation of different demographic groups. This minimizes the risk of sampling bias, where certain groups may be overrepresented or underrepresented, skewing the results.

Hence, to achieve a representative sample, stratified random sampling will be employed. This method involves dividing the population into homogeneous subgroups based on key demographic characteristics, then randomly selecting samples from each stratum proportionate to its size in the population. The strata will include:

- Age: Use age brackets (18-29, 30-44, 45-64, 65+) based on voter turnout trends, as younger age groups are often underrepresented.
- Gender: Stratify by gender to capture any potential disparities in voting behavior.
- Race and Ethnicity: Reflect population proportions of racial and ethnic groups, as provided by census data.
- Geography: Stratify by state and, within each state, by rural, suburban, and urban classifications to capture regional voting patterns.
- Political Affiliation: Where available, include political party affiliation, such as Democrat, Republican and Independent, to account for differing levels of partisan engagement.

Within the division, we will randomly select respondents within each stratum to match the proportions of these demographic categories in the general population. Our stratified random sampling works by first dividing the entire population into subgroups, or strata, based on specific characteristics relevant to the research, such as age, gender, or political affiliation. Once these strata are defined, random samples are drawn independently from each subgroup, ensuring that each is proportionally represented according to its size in the overall population or adjusted for adequate representation of smaller yet important groups. These samples are then combined to form the complete sample, ensuring that all significant subgroups are included. This approach reduces sampling variability, improves representation, and allows for more precise estimates compared to simple random sampling, as each subgroup's characteristics are accurately reflected in the final sample.

B.2.5 Trade-offs

This approach minimizes bias but can be costly due to increased recruitment needs in underrepresented groups as mentioned in 'The Total Survey Error Approach' (Weisberg (2009)). Random sampling may also limit the ability to reach specific subsets of voters.

B.3 Data Validation

As data quality is imperative to reach a reliable prediction, we will apply the following examinations to assure the validation:

- Verification of Responses: Employ CAPTCHA for online responses to limit bot entries and cross-check voter registration data to confirm respondent eligibility.
- Consistency Checks: Highlight and review responses with inconsistencies, such as duplicate IP addresses or extreme response times, and validate through follow-up when feasible.
- Data Cleaning: Remove incomplete or invalid responses and use imputation techniques for minor missing data if the omission rate is low.

B.4 Poll Aggregation Methodology

We will pick the poll-of-polls approach aggregates results from multiple polling sources to produce a more stable and representative forecast of the U.S. presidential election. 'Polling and the Public' (Asher (2016)) outlines the benefits of weighted poll aggregation and the importance of outlier detection. Outlier detection ensures that any poll results significantly different from others are reviewed to determine if they reflect genuine public opinion or anomalies due to errors or biased sampling. This process prevents these outliers from disproportionately influencing the overall forecast. Accordingly, this method helps to mitigate individual poll biases, smooth out fluctuations, and capture a broader picture of voter sentiment. Polls will be weighted by the following factors:

- Sample Size: Larger sample size polls will receive greater weight to reflect their higher statistical reliability.
- Recency Emphasis: The most recent polls are given greater influence in the aggregation to reflect real-time changes in voter preferences. Time-weighting is applied to each poll within the last month, with those closest to the current date contributing most to the overall estimate.
- Outlier Identification: Identify outlier polls by analyzing their deviation from the rolling average of other polls. Polls with extreme deviation from the average will be assessed individually to determine if they represent unique insights or sampling anomalies.

B.5 Survey

This survey will be conducted using Google Forms, which is an effective platform for data collection. The survey can be accessed by the link, Google Form Survey.

B.5.1 Survey Structure

Title:

2024 U.S. Presidential Election Forecast Survey

Introduction:

We appreciate your participation in this survey, which aims to forecast the outcome of the 2024 U.S. Presidential election. Your responses are essential for our research.

Please note:

- Your answers will be treated with complete confidentiality.
- Participation in this survey is voluntary.
- We encourage you to provide honest and thoughtful responses.
- The survey is estimated to take about 5 minutes to complete.
- If you have any questions or concerns, feel free to reach out to our research team at anjojoo.xu@mail.utoronto.ca (Angel Xu, Yunkai Gu, Yitong Wang).

Thank you for your valuable contribution! As a token of our appreciation, each participant will receive $5 upon completion of the survey.

Section 1: Eligibility

Are you a U.S. citizen?

- Yes
- No (End Survey)

Do you meet your state's residency requirements?

- Yes
- No (End Survey)

Will you be 18 years old or elder by the Election Day?

- Yes
- No (End Survey)

Are you registered to vote by your state's voter registration deadline. (North Dakota does not require voter registration.)

- Yes

- No
- Plan to register later
- Maybe

Section 2: Demographics

The following questions will help us understand the background characteristics of our respondents.

Which age group do you belong to?

- 18-29 years
- 30-44 years
- 45-64 years
- 65 years or older

What is your gender? - Male

- Female
- Other
- Prefer not to say

What is your race or ethnicity? Please select all that apply. - White

- Black or African American
- Hispanic or Latino
- Asian
- Native American or Alaska Native
- Native Hawaiian or Other Pacific Islander
- Prefer not to say
- Other

Which U.S. state do you currently reside in?

[answer box]

What region do you live in within your state?

- Rural
- Suburban

- Urban
- Prefer not to say

What is your political affiliation?

- Democratic
- Republican
- Independent
- Libertarian
- Green Party
- Prefer not to say
- Other

Section 3: Voting Behavior and Intentions

These questions focus on voting registration and plans for the upcoming election.

How likely is it that you will vote in the 2024 U.S. Presidential Election?

- Very likely
- Somewhat likely
- Somewhat unlikely
- Very unlikely

If the election were held today, which candidate would you most likely vote for?

- Donald Trump
- Kamala Harris
- Not sure
- Prefer not to say
- Other

How confident are you that your choice would remain the same by Election Day?

- Not at all confident
- Slightly confident
- Moderately confident

- Very confident
- Completely confident

Section 4: Engagement with the Election

This section aims to understand how actively our respondents follow and discuss the election.

How closely do you follow news and updates related to the 2024 U.S. Presidential Election?

- Very closely
- Somewhat closely
- Not very closely
- Not at all

How often do you discuss the 2024 U.S. Presidential Election with friends, family, or colleagues?

- Daily
- Weekly
- Occasionally
- Rarely
- Never

Section 5: Additional Insights

We appreciate any additional thoughts our respondents may have on the upcoming election.

Do you have any further comments or insights about the factors that might affect the 2024 U.S. Presidential Election? (optional)

[Answer box]

End Message:

Thank You for Completing the Survey!

We appreciate your time and thoughtful responses. Your participation is invaluable in helping us gather insights for our research on the 2024 U.S. Presidential Election forecast. Your answers will contribute to a more comprehensive understanding of voter trends and factors influencing this election.

If you have any further questions or would like to know more about this study, please feel free to reach out to our research team at anjojoo.xu@mail.utoronto.ca.

As a thank you, each participant will receive a $5 reward shortly after survey completion.

B.6 Survey Budget Allocation

- Sampling & Recruitment: $50,000. To ensure a large and diverse sample, the majority of the budget are distributed for online recruitment and sampling.
- Incentives: $15,000. Fund small incentives for respondent participation.
- Data Validation and Cleaning: $10,000. Employ validation mechanisms to ensure data quality.
- Poll Aggregation and Analysis: $15,000. Use advanced analytical methods for accurate aggregation.
- Miscellaneous Costs: $10,000. Cover unforeseen costs in recruitment, data cleaning, or analysis.

B.7 Survey Design Considerations

The survey was designed with a focus on clarity, respondent comfort, and data reliability. Questions flow logically, from demographic information to voting intentions, perceptions, and engagement, reducing cognitive load. Wording is conversational, neutral, and direct, minimizing confusion and potential bias. Multiple-choice questions provide efficiency in response and analysis, while open-ended questions are selectively used for unique insights. A pilot test will be conducted to refine clarity and flow, ensuring that all questions are relevant, easy to answer, and respectful of the respondent's experience.

B.8 Methodology Strength and Weakness

B.8.1 Strengths

This methodology leverages stratified random sampling and poll-of-polls aggregation, providing a balanced, representative forecast by combining demographic weighting and time-based adjustments. Outlier detection and confidence intervals further enhance reliability, making the forecast adaptable to real-time shifts in voter sentiment.

B.8.2 Weaknesses

Limitations include potential sampling and non-response bias, as underrepresented groups and undecided voters may be difficult to capture. Additionally, reliance on self-reported data and aggregated poll results introduces variability, while temporal limitations may impact accuracy close to Election Day. The forecast provides a likely outcome range but cannot fully account for unexpected events or behavioral shifts.

C Additional Data Cleaning details

The data cleaning process involves several critical steps to ensure that the dataset is suitable for analysis and accurately reflects voter sentiment regarding the 2024 U.S. Presidential Election. Below is a detailed description of the cleaning operations performed on the raw polling data.

The dataset is filtered to retain only those polls conducted by high-quality pollsters, defined by a numeric grade of 2.9 or higher. Additionally, only polls for the target candidate, Donald Trump or Donald Trump Jr., are included. The `end_date` variable, which represents when each poll was completed, is converted from character format to date format using the `mdy` function from the `lubridate` package. To focus on relevant data, the dataset is further filtered to include only those polls conducted within six months before the election date (after May 5, 2024). Afterwards, only the necessary columns for analysis are retained, including `end_date`, `sample_size`, `state`, and `pct`. Then, a new variable, `end_date_num`, is created to represent the number of days since the minimum end date in the dataset. This numeric format allows for easier modeling. For certain models, it is necessary to convert the percentage of votes (`pct`) into actual vote counts. A new variable, `num_vote`, is calculated by multiplying the percentage by the sample size and rounding to the nearest integer. Finally, any remaining missing values are removed from the dataset to ensure that the data is clean and complete for subsequent analysis.

These cleaning steps ensure that the dataset is robust, consistent, and ready for analysis, ultimately providing a solid foundation for modeling voter preferences in the 2024 U.S. Presidential Election.

D Additional Model details

D.1 Posterior Predictive Check

During the modeling process, posterior predictive checks (PPC) is used.

PPC is the comparison between what the fitted model predicts and the actual observed data, which validates whether the fitted model is compatible with the observed data. The aim is to detect if the model is inadequate to describe the data (Andrés López-Sepulcre ORCID iD 2024).

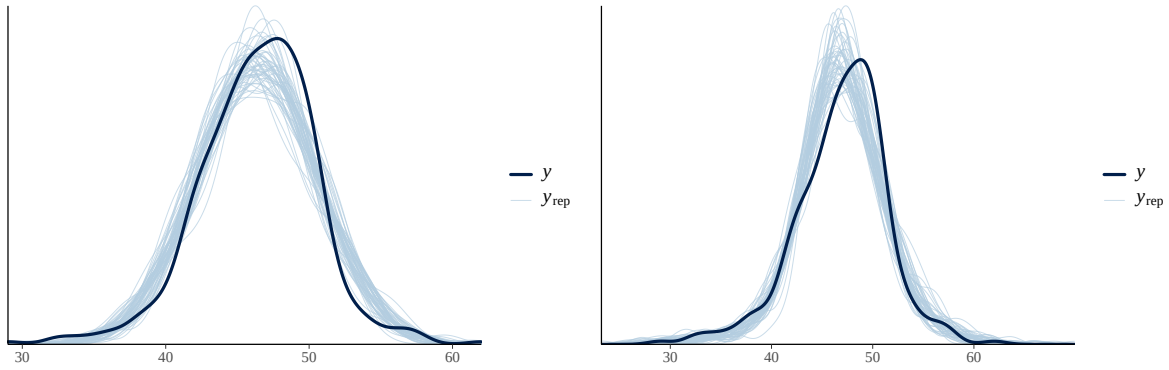
In the PPC plots, the x-axis represents the values of the response variable (in this case, the vote percentage of Donald Trump), and the y-axis represents the density or probability of those values.

The dark black line presents the density of the observed data (y), which is a smoothed estimate of the distribution of actual concentrated data points. On the other hand, the light blue

lines present the density of the replicated data (y_{rep}), which is generated from the posterior predictive distribution of the model.

Whether the black line matches well with the blue lines indicates how well the model fits the observed data. If the black line aligns well with the blue lines, it suggests that the model is performing well in capturing the actual data distribution, and vice versa.

Figure 6 shows the posterior predictive checks for both Bayesian models.



(a) Posterior Predictive Check of Bayesian Model 1 (National)

(b) Posterior Predictive Check of Bayesian Model 2 (by State)

Figure 6: Posterior Predictive Check of Two Bayesian Models

In Figure 6a, the black line basically follows the trend of the light blue lines, indicating the observed data consistently falling within the range of simulated data. This suggests that the model provides a good fit, effectively predicting the actual trend of the data.

Similarly, Figure 6b shows the posterior predictive checks for the second model, which predicts Trump's percentage vote over time by state. Although the black line has minor deviation away from the light blue lines, and a part of the black line falls outside the range of simulated lines, the majority part of the black line matches with the light blue lines.

This suggests that the model basically provides a good fit, effectively predicting the actual trend of the data. But the fit is not that well compared to model 1, since there is a larger deviation in this plot than the previous PPC plot for model 1.

D.2 Posterior vs. Prior

Comparing the posterior with the prior is also necessary for model validation. It examines how the model fits and is affected by the data.

Prior distribution represents the beliefs about the parameter values before observing any data, which is constructed by our initial assumptions. Posterior distribution reflects updated beliefs about the parameters after observing the data.

In the plot, the posterior distribution is demonstrated on the left side, and the prior distribution is on the right side. Each colored dot on the graph represents a parameter, and the horizontal line indicates the uncertainty.

By comparing the posterior with the prior, whether data has a significant impact on the parameter estimates and whether the model is well-fit could be validated.

Figure 7 shows the posterior vs. prior plot for the Bayesian model in the national perspective. We can see that in the posterior distribution, parameters have a smaller spread compared to the prior, indicating less uncertainty. There are also some differences between the prior and posterior distributions, suggesting that the model is well-fit and captures certain patterns in the data.

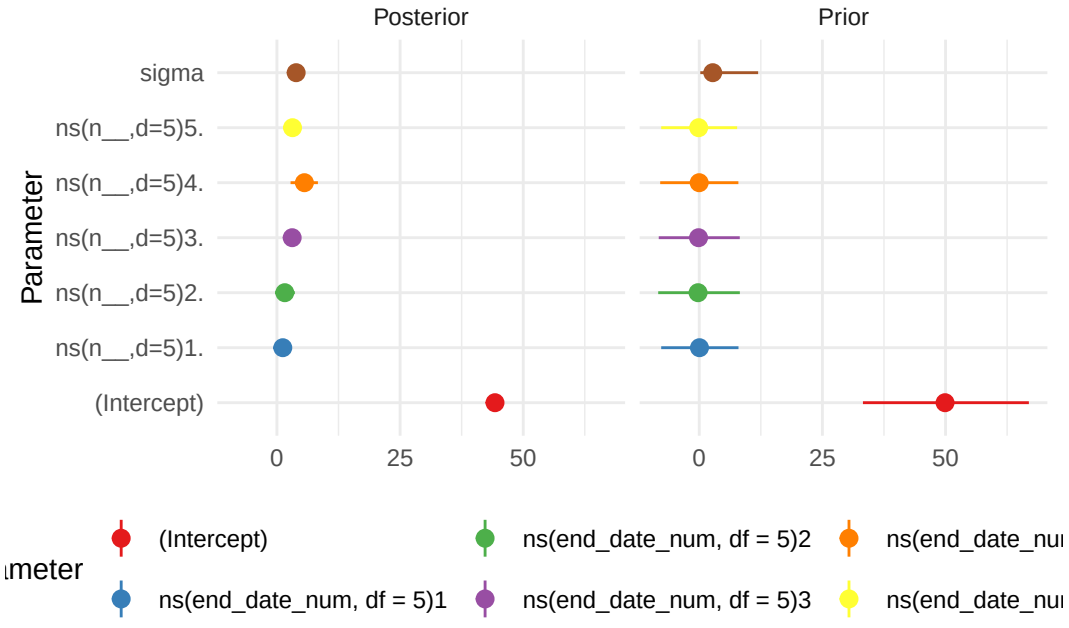


Figure 7: Posterior vs Prior Plot for Model 1 (National)

D.3 Diagnostics

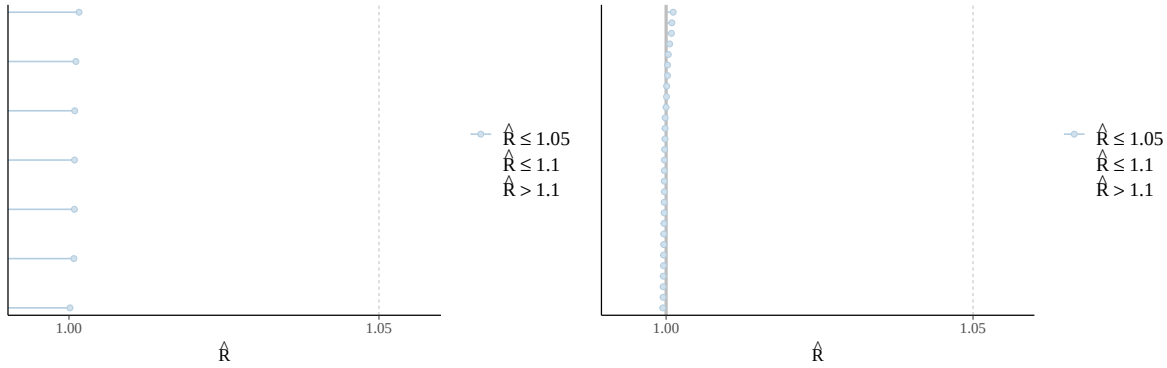
D.3.1 R-hat Plots

$\hat{R}$ is a diagnostic measure that compares the variance within each chain to the variance between chains.

R-hat plot shows the values of the $\hat{R}$ statistic on the x-axis, which is a measure of convergence for Markov Chain Monte Carlo (MCMC) simulations in Bayesian analysis, and shows different parameters or chains on the y-axis.

The plot categorizes the parameters based on their $\hat{R}$ values. It is a crucial diagnostic tool in assessing the validity of the model. If most parameter have values $\hat{R} \leq 1.05$, the parameters have good convergence. This indicates that these parameters have likely reached their stationary distribution, and the chains have converged to similar distribution, suggesting that the sampling process has mixed well.

Figure 8a and Figure 8b presents R-hat plots for the two Bayesian models constructed in the paper respectively. Both plots show that all parameters in two models have values $\hat{R} < 1.05$ and around 1, indicating good convergence and a well-fitted model.



(a) R-hat Plot of Bayesian Model 1 (National) (b) R-hat Plot of Bayesian Model 2 (by State)

Figure 8: R-hat Plots of Two Bayesian Models

D.3.2 Trace Plots

Trace plot shows samples from all the chains. Each chain is represented by a different color or line style.

If all chains appear to mix well and oscillate around the same mean, and overlap significantly, it suggests that the parameter estimates are stable.

Figure 9 presents a trace plot for the Bayesian model which is in the national perspective. We can see that there is no clear separation between the chains, suggesting that there is lack of convergence issue and well-fit model. However, relatively large fluctuations in the plot indicate some uncertainty, and less effective and precise outcome. This could be probably fixed by adding more data into the modeling procedure.

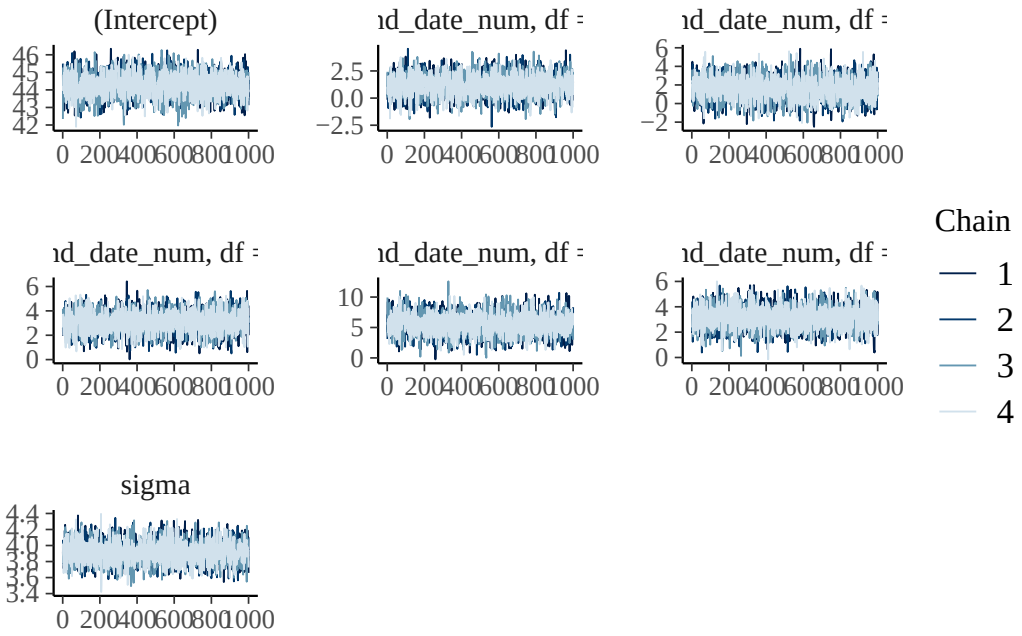


Figure 9: Trace Plot for Model 1 (National)

References

AAPOR. 2022. "Sampling Methods for Political Polling." https://aapor.org/wp-content/uploads/2022/12/Sampling-Methods-for-Political-Polling-508.pdf.

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

Andrés López-Sepulcre ORCID iD, Matthieu Bruneaux. 2024. "Posterior Predictive Checks." https://cran.r-project.org/web/packages/isotracer/vignettes/tutorial-100-posterior-predictive-checks.html.

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Asher, Herb. 2016. *Polling and the Public: What Every Citizen Should Know*. Cq Press.

Emerson. 2024a. "Emerson College Redefines Political Polling Landscape." https://today.emerson.edu/2024/10/16/emerson-college-redefines-political-polling-landscape/.

———. 2024b. "October 2024 National Poll: Harris 50." https://emersoncollegepolling.com/october-2024-national-poll-harris-50-trump-48/.

———. 2024c. "October 2024 Tracking National Poll: Harris 49." https://emersoncollegepolling.com/october-2024-tracking-national-poll-harris-49-trump-48/.

———. 2024d. "Polling Data - Google Spreadsheet." https://docs.google.com/spreadsheets/d/130z_dUtgMB3_yta2v8knWxjNv9Z2HymD/edit?gid=1544401297#gid=1544401297.

———. 2024e. "September 2024 Swing State Polls: Trump and Harris Locked in Tight Presidential Race." https://emersoncollegepolling.com/september-2024-swing-state-polls-trump-and-harris-locked-in-tight-presidential-race/.

FiveThirtyEight. 2024. "Presidential General Election Polls (Current Cycle)." https://projects.fivethirtyeight.com/polls/data/president_polls.csv.

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

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Groves, Robert M, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau. 2011. *Survey Methodology*. John Wiley & Sons.

Illinois, University of. 2023. "Polling and Sampling: Confidence Intervals and Hypothesis Testing." https://discovery.cs.illinois.edu/learn/Polling-Confidence-Intervals-and-Hypothesis-Testing/Polling-and-Sampling/.

Neuwirth, Erich. 2022. *RColorBrewer: ColorBrewer Palettes*. https://CRAN.R-project.org/package=RColorBrewer.

New York Post. 2024a. "Male Voters Under 30 Voting MAGA in 2024 Explain Trump's Appeal: 'Unabashed Machismo Vibe'." https://nypost.com/2024/10/28/us-news/male-voters-under-30-voting-maga-in-2024-explain-trumps-appeal/.

———. 2024b. "Trump Leading Harris in All but One Swing State Thanks to Strong Black Support: Polls." https://nypost.com/2024/10/31/us-news/trump-takes-all-but-one-swing-state-thanks-to-strong-black-support-polls/.

———. 2024c. "Why Donald Trump May Still Have the Upper Hand as Polls Show Him Deadlocked with Kamala Harris." https://nypost.com/2024/10/27/us-news/why-donald-trump-may-still-have-the-upper-hand-as-polls-show-him-deadlocked-with-kamala-harris/.

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

———. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

ScienceDirect. 2023. "Sampling Frame - Overview, Applications and Limitations." https://www.sciencedirect.com/topics/mathematics/sampling-frame#:~:text=The%20sampling%20frame%20is%20the,once%20in%20a%20sampling%20frame.

SciLine. 2023. "Surveys and Polling: What Reporters Need to Know." https://www.sciline.org/elections/surveys-polling/.

Weisberg, Herbert F. 2009. *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. University of Chicago Press.

Wickham, Hadley. 2011. *Testthat: Get Started with Testing. The R Journal*. Vol. 3. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

———. 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.

Xie, Yihui. 2021. *Knitr: A Comprehensive Tool for Dynamic Report Generation in r*. https://CRAN.R-project.org/package=knitr.