Predicting the 2024 US Presidential Election with a Model-Based Forecast*

Kamala Harris Predicted to Lead with 341 Electoral Votes Based on a Poll-Based Prediction

Yuanyi (Leo) Liu Dezhen Chen Ziyuan Shen

November 3, 2024

In this paper, we develop a linear model to predict the outcome of the 2024 U.S. presidential election using a "poll-of-polls" approach that aggregates data from multiple well-known pollsters. By analyzing state-level polling trends, our model predicts that Kamala Harris is likely to secure key states, leading to an estimated electoral vote count of 341 for Harris and 197 for Donald Trump. This approach, which combines multiple polls, reduces bias and improves reliability, providing a more accurate picture of public sentiment leading up to the election.

Table of contents

*Code and data are available at: Forecasting the 2024 US Presidential Election.

1 Introduction

The results of presidential elections influence national and international policies and determine governance and economic priorities. An accurate election forecast is a valuable tool for political strategists, the media, and the public. The 2024 U.S. presidential election is expected to be highly competitive, with Donald Trump and Kamala Harris as the leading candidates. Despite the abundance of polling data, individual polls are often subject to bias and short-term fluctuations, making it difficult to predict election outcomes with confidence. This paper addresses this issue by using aggregate polling data to provide more reliable predictions of election outcomes.

In this study, our estimand is Donald Trump's percentage of polling in each state ahead of the 2024 U.S. presidential election. The object of the estimation is Trump's percentage of polling data in each state based on aggregated information from various sources. By using a linear regression model, we aim to capture changes in public opinion over time and provide a clearer understanding of voter preferences and the likely election outcome.

Our findings suggest that while Trump has a competitive chance, Harris is projected to win key states and receive a majority of electoral votes. Harris is estimated to receive 341 electoral votes to Trump's 197, and the uncertainty around these estimates is represented by confidence intervals, which take into account polling changes and model assumptions. This is important because accurate election forecasts help reduce uncertainty, which allows political stakeholders to allocate resources efficiently and gives the public a better understanding of likely outcomes. Aggregate polling reduces uncertainty and provides a clearer picture of potential outcomes.

The structure of the paper is organized as follows: following Section 1, Section 2 presents the data collection and cleaning process, along with an overview of the variables used in the analysis. Section 3 introduces the forecasting models, explaining why the selected models are suitable for predicting election outcomes based on aggregated polling data. Then, Section 4 provides the results, highlighting key trends and predictions. Eventually, Section 5 concludes with a discussion of the findings, evaluating the reliability of the forecasts and identifying potential limitations of the models.

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023) to process and analyze polling data for the 2024 U.S. Presidential election. The dataset used for this analysis was obtained from the FiveThirtyEight 2024 U.S. Presidential Election Polls (Best and Bycoffe 2024). It consists of polling data for the 2024 general election, covering various polling organizations and methodologies. Following methodologies discussed by "Telling Stories with Data" (Alexander 2023), we forecast election outcomes using the "poll of polls" method, which aggregates results from multiple polls to reduce bias and provide a more accurate representation of voter sentiment. For key operations, please refer to Appendix C.

The dataset includes 15,891 rows and 52 columns, covering various pollster attributes such as pollster name, state, methodology, and polling results. To ensure the reliability of our analysis, we filter the data to include only polls with a numeric grade of 2.5 or higher, which represents high-quality, reputable pollsters. This filtering allows us to focus on polls that follow rigorous standards and have demonstrated transparency and accuracy.

Additionally, we limit the dataset to polls conducted after July 15, 2024, when Donald Trump officially announced his campaign, ensuring that the data reflects the most current public sentiment. Other similar datasets from prior elections, such as data from previous general election cycles, could have been used for comparison. However, given that this analysis focuses on the upcoming 2024 election and the shift in public opinion, the most relevant data is specific to the current cycle.

2.2 Measurement

Polling data is a measurement of public sentiment captured through various survey methods. Polling organizations collect responses from individuals representing different segments of the population and ask them about their voting preferences. These responses are then weighted to reflect a more accurate representation of the electorate, based on factors like age, gender, race, and geographic region.

For this dataset, the poll results reflect the percentage of respondents who support a particular candidate. Different polling methods—such as live phone interviews, online panels, and mixed methodologies—capture this information. Each polling organization applies its own methodology, which can influence the results. For example, polls that use live interviews may experience different respondent behavior than those that use online surveys. More details on polling methodologies can be found in the ABC News methodology explanation.

The key measurement process involves converting real-world voter preferences into a dataset of percentages that reflect support for a candidate at a specific point in time. Polling organizations typically provide this data at a state or national level, which allows for both localized and broad interpretations of voter sentiment.

2.3 Outcome variables

2.3.1 Percent support for Donald Trump

The primary outcome variable is the percentage (`pct`) of respondents who indicated support for Donald Trump in the 2024 general election. This percentage is calculated based on the total number of respondents in each poll divide by `sample_size`. The data includes observations from various states, providing both national and state-specific measures of Trump's support.

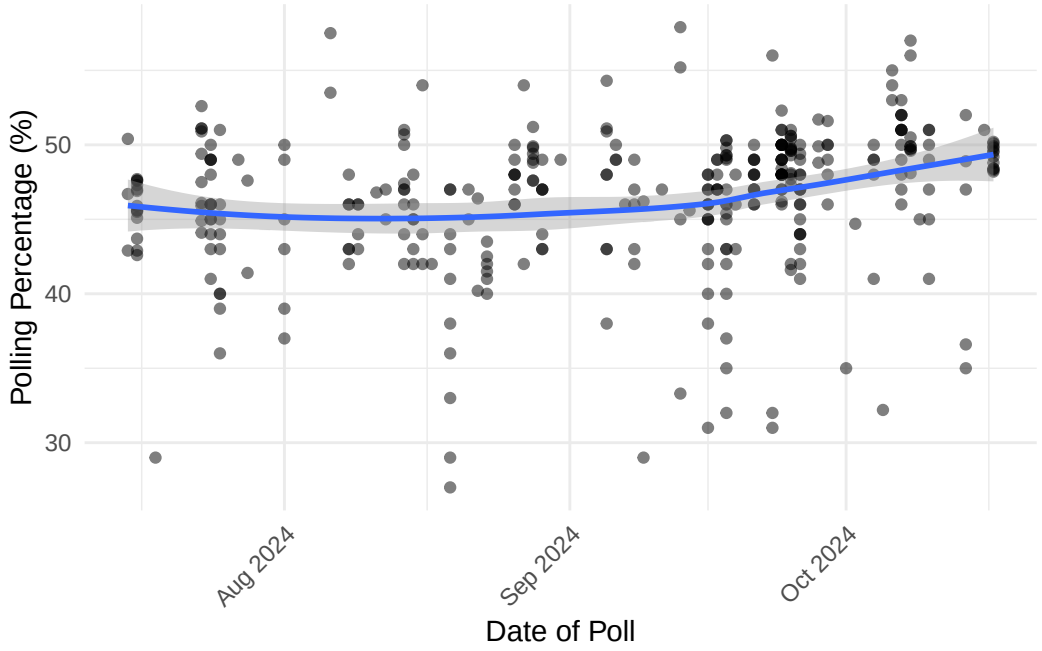

Figure 1: Donald Trump's polling percentages from July to October 2024. Each dot indicates the polling percentage for Trump on a specific date, while the blue line shows the trend in Trump's support.

Using package `ggplot2` (Wickham 2016), Figure 1 shows Donald Trump's polling percentages over time from July to October 2024. Each black dot represents an individual poll conducted on a specific date, indicating the percentage of voters who supported Trump in that poll. The blue line represents a trendline, smoothing out the individual poll results to show the general trajectory of Trump's support over time.

Initially, Trump's polling percentage appears stable with small fluctuations around 45%. However, as the campaign progresses into October, there is a slight upward trend in Trump's polling percentage, suggesting that his support increased as the election neared.

The shaded region around the blue line represents the confidence interval, indicating the range within which the true polling percentage is likely to fall, accounting for sampling variation.

2.4 Predictor variables

2.4.1 Polling methodology

The `methodology` variable describes the method each pollster used to collect data. It includes approaches such as live phone interviews, online surveys, and mixed methods. For polls that used multiple collection methods (e.g., phone and online), we labeled them as "Mixed Voting" for simplicity. The methodology is important because different approaches can introduce varying degrees of bias, influencing the final reported percentages.

2.4.2 State

The `state` variable identifies whether the poll is state-specific or national. In state-specific polls, the data reflects localized voter preferences, while national polls aggregate opinions across the entire country. For this analysis, we ignore national polls and focus entirely on state polls because electoral outcomes are determined on a state-by-state basis in the U.S. election system.

Table 1: The electoral votes allocated to each U.S. state, listed in two sets for better readability. Each state's electoral vote count reflects its representation in the Electoral College.

States	Electoral Votes	States	Electoral Votes
Alabama	9	Nebraska	5
Alaska	3	Nevada	6
Arizona	11	New Hampshire	4
Arkansas	6	New Jersey	14
California	54	New Mexico	5
Colorado	10	New York	28
Connecticut	7	North Carolina	16
Delaware	3	North Dakota	3
Florida	30	Ohio	17
Georgia	16	Oklahoma	7
Hawaii	4	Oregon	8
Idaho	4	Pennsylvania	19

States	Electoral Votes	States	Electoral Votes
Illinois	19	Rhode Island	4
Indiana	11	South Carolina	9
Iowa	6	South Dakota	3
Kansas	6	Tennessee	11
Kentucky	8	Texas	40
Louisiana	8	Utah	6
Maine	4	Vermont	3
Maryland	10	Virginia	13
Massachusetts	11	Washington	12
Michigan	15	West Virginia	4
Minnesota	10	Wisconsin	10
Mississippi	6	Wyoming	3
Missouri	10	District of Columbia	3
Montana	4		

Table 1 illustrates the number of electoral votes allocated to each U.S. state, based on data from the official government website (Nolan 2008). This information will be used to map the predicted results for each state later, allowing us to calculate the total number of electoral votes that Donald Trump is projected to receive.

2.4.3 Polling score and transparency

The `numeric_grade` variable is a rating assigned to each polling organization, indicating the reliability of the poll. Pollsters with higher numeric grades are considered more reliable and consistent in their methodology (e.g. higher than 2.5). Additionally, the `transparency_score` measures how openly a pollster shares their methodology and data, which further affects the trustworthiness of their results.

Figure 2 presents two bar charts illustrating the distribution of Polling Scores (Numeric Grades) and Transparency Scores across the polling organizations in the analysis:

- **Polling scores**: Most organizations have numeric grades between 2.8 and 3.0, indicating that the majority of pollsters are considered highly reliable. Very few have grades below 2.7, which would suggest lower reliability.
- **Transparency scores**: Transparency scores show more variability, with many pollsters achieving scores above 9, reflecting a high level of openness in their methodology. Fewer organizations have transparency scores below 6, indicating that while some pollsters are less open, the majority strive for transparency in their practices.

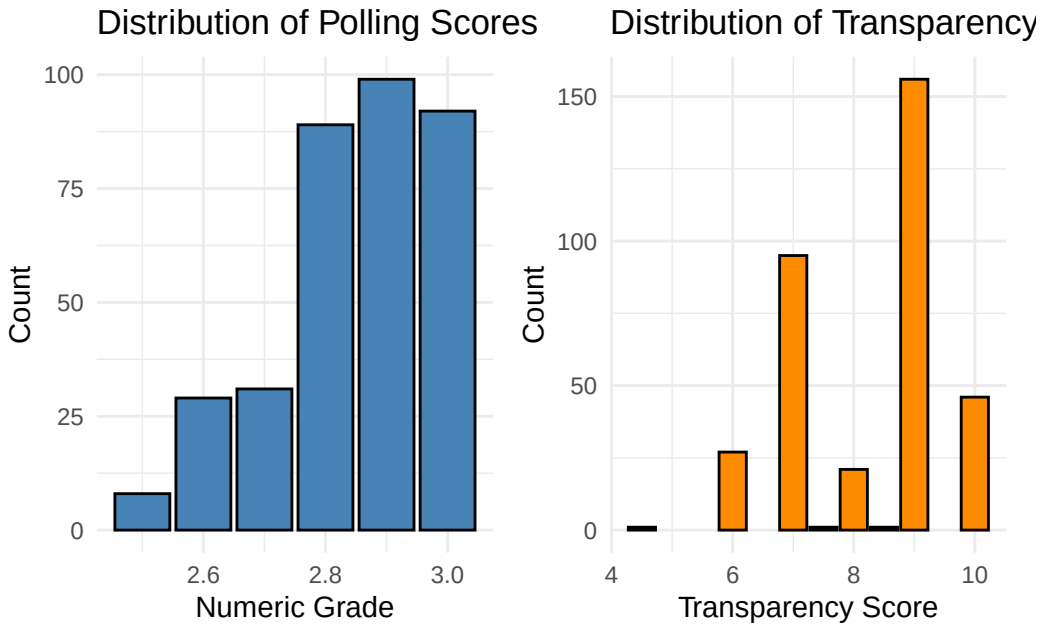

Figure 2: Distribution of Polling Scores and Transparency Scores of Polling Organizations in the 2024 U.S. Presidential Election

2.4.4 Sample size

The `sample_size` variable represents the total number of respondents surveyed in each poll. A larger sample size generally leads to more reliable and precise estimates of voter sentiment, as it reduces the margin of error. In contrast, smaller sample sizes can result in greater variability and less confidence in the results.

2.4.5 Poll duration

The `duration` variable is a constructed variable that represents how long Donald Trump has been in the 2024 presidential campaign, measured as the number of days from his official campaign announcement on July 15, 2024 until the `end_date` of each poll. This variable helps quantify how much time has passed since Trump officially entered the race and allows us to examine how his support has evolved over the course of his campaign.

3 Model

Our modeling approach aims to quantify the relationship between various polling metrics and the percentage support for Donald Trump. For this analysis, we use a linear model to examine

how factors such as poll score, methodology, poll duration, sample size, transparency score, and state influence support percentages. The model is implemented using the `stan_lm` function, with a Gaussian distribution to capture the variability in support rates.

In this analysis, we use predictors that reflect both the methodological quality and structural characteristics of each poll. Specifically, we include variables such as `numeric_grade`, which quantifies the poll's quality, `methodology`, `duration` (the length of the election period), `sample_size`, `transparency_score`, and `state` (the regional factor).

The model assumes that the distribution of support percentage, given these polling characteristics, follows a normal distribution. This Gaussian assumption facilitates parameter estimation, which is a standard approach in linear regression. To prevent overfitting and ensure interpretability, we assume moderate priors, maintaining balanced uncertainty across our predictors. This approach enables us to assess the impact of polling characteristics on candidate support while ensuring stability in our findings. Background details and diagnostics are included in Appendix D.

3.1 Alternative model

Initially, we attempted to categorize the `state` variable into two levels: state-specific polls and national polls. However, after modeling, we found that this approach was less effective for predicting election outcomes on a state level. Since the election is determined by the state-by-state electoral college, it became clear that using a state-specific variable would more accurately capture the variation needed to predict outcomes by state. This refinement allowed the model to better align with the electoral structure, resulting in a more accurate reflection of support distribution across states.

3.2 Model set-up

The model predicts Trump's support percentage using the following predictor variables:

- Poll score (`numeric_grade`): Represents the quality rating of the poll.
- Methodology (`methodology`): Different methods used to conduct the poll, such as live phone, mixed voting, or online ads.
- Poll duration (`duration`): The the length of the election period was conducted.
- Sample size (`sample_size`): The number of respondents in the poll.
- Transparency score (`transparency_score`): A measure of how transparent the polling data and methodology are.
- State (`state`): A categorical variable representing the U.S. state where the poll was conducted.

The model takes the form:

$$
\begin{aligned}
y_i \mid \mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \beta_0 + \beta_1 \cdot \text{Poll score}_i + \beta_2 \cdot \text{Methodology}_i \\
&\quad + \beta_3 \cdot \text{Poll duration}_i + \beta_4 \cdot \text{Sample size}_i \\
&\quad + \beta_5 \cdot \text{Transparency score}_i + \beta_6 \cdot \text{State}_i + \epsilon_i \\
\epsilon_i &\sim \text{Normal}(0, \sigma^2)
\end{aligned}
$$

Where:

- β_0 is the intercept term.
- $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$ are the coefficients for each predictor.
- σ^2 is the variance of the error term.

The model is executed in `R` (R Core Team 2023) using the `rstanarm` package (Goodrich et al. 2022). Default priors from `rstanarm` (Goodrich et al. 2022) are used, with the priors set to have a mean of zero and a moderate standard deviation to ensure a reasonable level of regularization.

3.2.1 Model justification

Existing research and political science theories suggest that factors such as poll quality, sample size, transparency, methodology, poll duration, and the state where the poll takes place can influence Trump's support percentage. Higher quality polls and larger sample sizes are believed to yield more reliable estimates, while transparency scores can affect levels of public trust, potentially altering support. Moreover, different polling methodologies (e.g., online surveys, telephone interviews) may affect how representative the sample is, while longer polling duration can reflect shifts in public sentiment over time. State-specific factors, such as local political events or population characteristics, can also contribute to the variation in support. The timing of the poll, particularly in relation to key political events, might further influence public opinion.

A linear regression model was chosen to predict Trump's support percentage because the dependent variable is continuous and tends to follow a normal distribution. Linear regression is a straightforward method to analyze how multiple factors contribute to an outcome, and it offers easy-to-interpret coefficients for each predictor. The model helps us determine the extent to which different aspects of polling influence the level of public support for Trump.

Further justification for using this model comes from its alignment with the central limit theorem, which suggests that with sufficient sample size, the distribution of poll percentages

should approximate normality. Additionally, the relationships between the predictors (e.g., poll quality, methodology type) and support align well with established theories in political behavior, giving a solid theoretical underpinning to our model. The simplicity of a linear regression approach also helps mitigate overfitting, ensuring our results are generalizable to a broader set of polling data.

4 Results

Section 4 examines the relationship between polling score, methodology, poll duration, transparency score, sample size, and state, with respect to Donald Trump's polling performance in the 2024 U.S. Presidential Election. Using a dataset that captures polling data across various states and polling organizations, we apply a linear regression model to identify key factors that influence Trump's support. Below, we present the results of our model and its implications.

We predicted the election outcomes based on our test dataset. However, due to missing data for certain states in our original dataset, predictions could not be generated for all states. To address this gap, we reviewed historical trends from recent elections involving Joe Biden; if a state consistently leaned toward the same candidate, we assumed this trend would continue and used it to estimate outcomes for those missing states. This limitation arises because the states with NA values in the dataset are not represented in the model, preventing accurate forecasts for those regions. Further discussion on this issue is provided in Section 5.

4.1 Model results and interpretation

The linear regression model built by our training dataset, using 242 data points, estimated the factors influencing Trump's support levels. For brevity, Table 2 only shows the first ten rows of the model's coefficients, with the full model summary available in Appendix D. The intercept is estimated at 44.224, representing the baseline support level when all predictors are at their reference levels. The model achieved an R^2 value of 0.803, indicating that 80.3% of the variance in polling outcomes is explained by the predictors included in the model. The adjusted R^2 value of 0.753 confirms that the model captures the relationships between variables without overfitting.

Table 2: Summary of key coefficients from the linear regression model predicting polling outcomes

	coefficient
(Intercept)	44.224
numeric_grade	0.779
as.factor(methodology)Mixed Voting	1.692
as.factor(methodology)Online Ad	0.960

	coefficient
as.factor(methodology)Online Panel	-1.037
as.factor(methodology)Probability Panel	-1.035
duration	0.028
sample_size	0.002
transparency_score	-0.220
stateCalifornia	-15.426

Certain predictors have a more obvious impact than others. For example, `numeric_grade` (0.779) shows that a one-unit increase in grade corresponds to a 0.779 percentage point increase in predicted polling outcomes. The type of methodology also plays a important role: "Mixed Voting" (1.692) and "Online Ad" (0.960) methods are associated with positive effects, while "Online Panel" (-1.037) and "Probability Panel" (-1.035) methods have negative effects relative to the baseline. Additionally, the state-specific coefficient shows a substantial effect. This underlines the importance of geographic context in polling predictions, especially for politically distinct areas.

Other variables, like `transparency_score` (-0.220) and `duration` (0.028), have smaller effects. The negative coefficient for `transparency_score` suggests a slight decrease in predicted support as transparency increases, potentially reflecting a sensitivity among the public to the transparency level in polling methods. Meanwhile, `sample_size` (0.002) shows a small positive effect, indicating that larger sample sizes modestly increase predicted support. A complete overview of all predictors is provided in Appendix D.

4.2 Predicted electoral outcomes

We employed a regression model to estimate the percentage of votes Trump is likely to receive in each state. Using these predicted results and the electoral vote allocations, we identified the winner for each state and calculated the total electoral votes for both Trump and Harris.

Table 3 presents a summary of the predictions, including Trump's estimated percentage, the number of electoral votes in each state, and the projected winner. For instance:

- In Arizona, Trump is expected to receive 48.94% of the vote, resulting in a win for Harris with 11 electoral votes.
- In Texas, Trump is projected to receive 51.83%, thereby winning all 40 electoral votes for that state.

Table 3: Prediction for Trump and Harris by Electoral College

State	Trump Predicted %	Electoral Votes	Winner
Arizona	49.50	11	Harris
California	33.92	54	Harris
Florida	51.57	30	Trump
Georgia	47.61	16	Harris
Maine	42.67	4	Harris
Maryland	31.59	10	Harris
Massachusetts	34.03	11	Harris
Michigan	46.96	15	Harris
Minnesota	42.47	10	Harris
Nebraska	44.85	5	Harris
Nevada	47.58	6	Harris
New Hampshire	44.51	4	Harris
New Mexico	40.42	5	Harris
New York	38.56	28	Harris
North Carolina	47.75	16	Harris
Ohio	49.29	17	Harris
Pennsylvania	47.39	19	Harris
Rhode Island	38.13	4	Harris
Texas	51.62	40	Trump
Virginia	43.55	13	Harris
Wisconsin	46.18	10	Harris

Since we are missing data for some states in our dataset, we turned to historical trends from recent elections involving Joe Biden to fill these blanks. For each missing state, we assessed whether it consistently leaned toward a particular candidate in past elections. If so, we assumed this trend would persist and used it as a basis for estimating outcomes in the current election.

Table 4 summarizes the outcomes for the missing states based on this approach. It shows the predicted winner for each state based on historical trend. For instance, Alabama, Alaska, Arkansas, Idaho, Indiana, and Iowa are expected to go to Trump, contributing their respective electoral votes to his total. Conversely, states like Colorado, Connecticut, Delaware, Hawaii, and Illinois are projected to favor Harris, thus allocating their electoral votes to her.

Table 4: Estimated Electoral Outcomes for Select U.S. States Based on Historical Voting Trends

State	Electoral Votes	Winner
Alabama	9	Trump
Alaska	3	Trump

State	Electoral Votes	Winner
Arkansas	6	Trump
Colorado	10	Harris
Connecticut	7	Harris
Delaware	3	Harris
Hawaii	4	Harris
Idaho	4	Trump
Illinois	19	Harris
Indiana	11	Trump
Iowa	6	Trump
Kansas	6	Trump
Kentucky	8	Trump
Louisiana	8	Trump
Mississippi	6	Trump
Missouri	10	Trump
Montana	4	Trump
New Jersey	14	Harris
North Dakota	3	Trump
Oklahoma	7	Trump
Oregon	8	Harris
South Carolina	9	Trump
South Dakota	3	Trump
Tennessee	11	Trump
Utah	6	Trump
Vermont	3	Harris
Washington	12	Harris
West Virginia	4	Trump
Wyoming	3	Trump
District of Columbia	3	Harris

4.3 Final electoral outcomes

Based on our analysis, Table 5 summarizes the total predicted electoral votes for Trump and Harris in the 2024 U.S. Presidential Election. In the United States, a candidate must secure at least 270 electoral votes to win the presidency. According to these predictions, Kamala Harris is expected to secure the majority with 341 electoral votes, while Donald Trump is projected to receive 197. This suggests a clear victory for Harris.

Table 5: Predicted Final Electoral Votes for the 2024 U.S. Presidential Election: Trump vs. Harris

Candidate	Final Votes
Trump	**197**
Harris	**341**

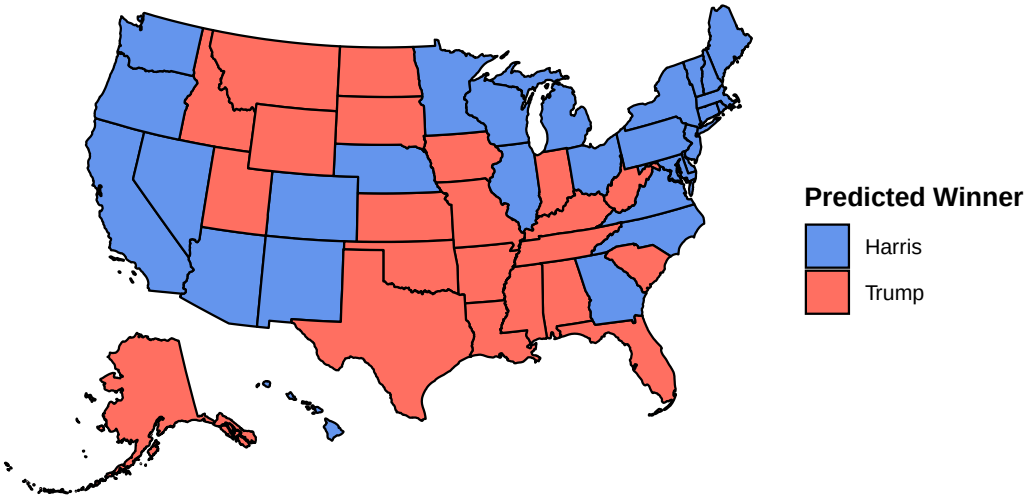

Figure 3: The predicted winner for each U.S. state in the 2024 presidential election based on our model. States shaded in blue are projected to favor Harris, while those in red are projected to favor Trump.

By using package `usmap` (Di Lorenzo 2024), Figure 3 displays the predicted outcome for each state in the 2024 U.S. Presidential Election, with red representing states where Trump is expected to win and blue for those favoring Harris.

It shows Trump gaining support across central and southern states, including Texas and Missouri, which have traditionally leaned Republican. Harris, meanwhile, is projected to lead in states along the West Coast and in the Northeast, such as California and New York, regions with a history of Democratic support. This visual representation complements the previous electoral vote projection, where Harris is anticipated to surpass the 270-vote threshold needed to win the presidency.

5 Discussion

5.1 Interpretation of polling data and electoral implications

This paper examines the factors affecting Donald Trump's polling performance in the 2024 U.S. Presidential Election. Our linear regression model shows that numeric grade, polling

methodology, and state-specific factors are key predictors of Trump's support. Higher numeric grades, representing perceived poll reliability, are positively associated with Trump's polling numbers, suggesting that well-regarded polls may capture a more favorable sentiment. Polling methodology also plays an important role, with different methods (e.g., "Mixed Voting" or "Online Panel") showing varied effects on the results, likely reflecting differences in respondent engagement or accessibility. State-specific factors underscore the importance of regional dynamics. The Electoral College system makes state-level support essential, and unique local issues and demographics in each state contribute to Trump's polling outcomes.

Transparency score and sample size also affect polling reliability. Polls with higher transparency and larger sample sizes provide more stable projections. While these factors improve data quality, they have less direct influence on voter preferences than state and methodology variables.

In contrast, poll duration has minimal impact in our model, suggesting that short-term and long-term polls yield similar support levels in this context. This finding implies that voter sentiment is more influenced by geographic and methodological context than by the length of polling periods. Overall, this model underscores the importance of state-level factors and methodology over specific polling practices in understanding Trump's support trajectory in this election.

5.2 Influence of state policies on voter preferences

Beyond the mechanics of polling, our analysis suggests that state-specific policies and regional cultural factors could influence voter preferences and eventually shape electoral outcomes, although they are not directly measured in our model. For example, state policies on recreational marijuana legalization may reflect broader ideological values that align with particular candidates or parties. To examine this further, we compared our predictions map with a visualization of states where recreational marijuana is legal.

Figure 4 shows states where recreational marijuana is legal. Compared with Figure 3, many states that have legalized marijuana, such as California, Washington, and New York, also appear in our prediction as likely to support Harris, suggesting a correlation between progressive state policies and support for more liberal candidates. This policy environment may align with Democratic platforms on social issues, which could explain stronger support for Harris in these regions.

In contrast, states that prohibit recreational marijuana use—many of which are predicted to lean toward Trump—might favor more conservative social policies, possibly influencing their political alignment. This could reflect a cultural environment that aligns more closely with Republican stances, contributing to stronger support for Trump in these areas.

While the connection between marijuana legalization and voter preferences is not absolute, the alignment of state policies with broader political values suggests that cultural factors may

influence voter preferences. Additionally, policies related to economic investment, healthcare, or education could shape these patterns, offering further influences into state-level political dynamics.

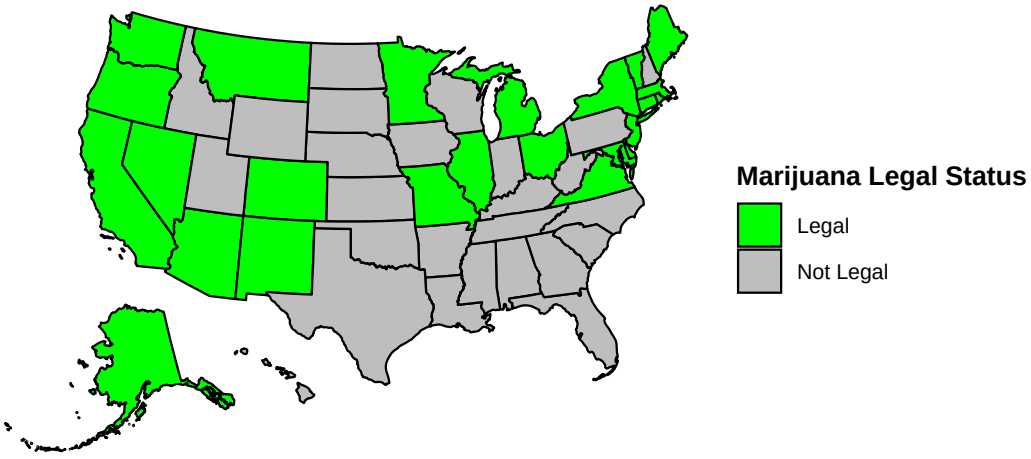

Figure 4: The legal status of recreational marijuana use in each U.S. state as of 2024. States shaded in green have legalized recreational marijuana, while those in gray have not.

5.3 Limitations of the data and model

This paper has several limitations that affect the accuracy and breadth of its predictions. First, the dataset is incomplete, with some states lacking polling data. These omissions prevent a fully representative analysis, as the absence of data from certain regions leaves out their unique political dynamics. Furthermore, focusing on state-level data may obscure more localized trends, particularly in diverse districts with distinct political identities. For instance, urban and rural areas within the same state often show different voting patterns that a state-level analysis may overlook.

Another limitation lies in the model's emphasis on certain predictors while excluding others. By prioritizing variables like poll duration and sample size, we may miss factors that could better explain variations in support, such as demographic and economic data. Although the model performs well on a broad level, these limitations underscore the need for a more inclusive approach to fully understand the range of influences on voter behavior.

5.4 Future research directions

Future studies could address these limitations by including demographic variables, such as age, income, and party affiliation, to provide a clearer view of voter preferences. Additionally, incorporating a wider range of state-level policies, such as healthcare, education, and economic

initiatives—could deepen the analysis of how local issues shape voter choices. Expanding the dataset in this way would allow a more detailed understanding of how social and economic factors interact with political preferences.

Increasing the model's geographic specificity to include smaller regions, such as congressional districts or precincts, could lead to more accurate forecasts, especially in closely contested states. This finer level of detail could capture unique district-level trends that a state-level model misses. Furthermore, as polling methods evolve, combining machine learning techniques with traditional regression could allow the model to adapt to shifts in public opinion, resulting in more responsive and timely predictions. This approach would provide a valuable foundation for tracking and forecasting voter preferences as political landscapes evolve.

Appendix

A Pollster methodology overview and evaluation

A.1 Overview of Siena/NYT

Siena College Research Institute (SCRI) and The New York Times launched a collaborative political poll in July 2013. Siena/NYT was known for its rigorous methodology and accuracy, which provided its first real-time midterm election poll in 2018 (Nagourney and Igielnik 2024). In 2024, FiveThirtyEight named it the most accurate pollster in the United States. The polling methodology emphasizes live phone interviews and representative sampling and focuses on transparency and statistical accuracy to provide a deep understanding of voter sentiment.

A.2 Target population, frame, and sample

- Target Population: Registered voters or likely voters in the United States, especially voters in major battleground states during an election in 2024.
- Sample Frame: U.S. registered voters from a national voter file maintained by L2, a nonpartisan provider of detailed voter data, including demographics and voting history. The sample includes voters with matching cell phone numbers.
- Sample Size: Sample sizes vary across different polls. The poll in late October 2024 surveyed over 2,500 voters, and 2,097 of them completed the full survey. The margin of error for likely voters is about ± 2.2 percentage points.

A.3 Sample recruitment

The sample was recruited by telephone. Registered voters were selected from the voter files held by L2 and contacted via their listed telephone numbers including landlines and mobile phone numbers. More than 260,000 calls were made to more than 80,000 voters, with over 98% of respondents using mobile phones. Interviews were conducted by live telephone interviews in both English and Spanish.

A.4 Sampling approach and trade-offs

Siena/NYT used a stratified sampling method (Nagourney and Igielnik 2024). Polls were stratified by demographic characteristics such as political party, race and region. This improves accuracy but adds cost and complexity. It also makes it challenging to reach certain populations such as young voters.

Advantages of stratified sampling:

- Generates a representative sample that captures the diverse opinions of different voter populations.
- Helps to ensure that under-represented groups are given appropriate weighting, and reduces potential bias.
- Provides a more accurate reflection of the wider population by stratifying on key characteristics.

Disadvantages of stratified sampling:

- Increased cost and time to complete polls, especially for live telephone interviews.
- There are challenges in reaching specific demographic groups, such as younger voters who may be less likely to participate in the poll.
- Reliance on weighting adjustments to correct for response discrepancies, which does not fully eliminate residual bias.

A.5 Non-response handling

Weighting adjustment: The Times weighted the sample using the survey software package in R and adjusted for unequal probability of selection in each stratum. The sample was also split by political party and weighted according to parameters of registered voter characteristics from the voter file. This weighting step ensures that respondents from underrepresented demographic groups such as those without a college degree are given more weight so that the results reflect the voting population overall.

Non-Response Model: To adjust for non-response bias, the L2 voter file was stratified by 'state precinct', 'political party', 'race', 'gender', 'marital status', and 'voting history'. The average expected response rate is based on a single-place non-response rate model from previous Siena/NYT surveys, which can be adjusted appropriately to account for different response rates in different groups.

Multiple calls and language adaptation: Interviewers made more than 260,000 calls to more than 80,000 voters, resulting in a sample of 2,516 respondents. Interviewers conducted the interviews in English and Spanish, and bilingual interviewers adapted the language of the interviews to respondent preferences, which helped to increase the participation of Spanish-speaking respondents. Re-contacts by interviewers reduced the non-response rate.

A.6 Questionnaire evaluation

Strengths:

- Minimize bias through the use of straightforward and non-leading questions.
- Transparency was maintained by publishing the questions verbatim to allow for public scrutiny.

- Bilingual interviewers ensured that respondents could participate in their preferred language (English or Spanish), which increased the participation of Spanish-speaking voters and improved the quality of respondents' answers.

Weaknesses:

- The length of the poll can lead to respondent fatigue, which affects the quality of the data and increases dropout rates.
- It can be challenging to limit calls to less than 15 minutes, which can lead to incomplete or rushed responses, thus affecting the overall reliability of the data.

A.7 Conclusion

The Siena/NYT poll uses stratified sampling and live phone interviews to reach a representative sample of voters. While the use of composite weighting and bilingual interviews provides advantages such as transparency and reduced bias, challenges remain including high costs and the potential for bias due to non-response. Despite these drawbacks, Siena/NYT's commitment to transparency and rigorous methodology makes it a reliable source for understanding voter sentiment and opinion.

B Idealized methodology and survey

B.1 Overview

The objective of this survey is to predict voter sentiment in the upcoming U.S. presidential election, by using a budget of $100,000 to collect accurate and varied data on voting intentions, candidate favourability, and key issues from a representative sample of registered or likely voters across the United States. The methodology is designed to maximize accuracy, minimize bias and take into account a variety of demographic, geographic and political factors that influence voting behaviour.

B.2 Sampling approach

The sampling approach is designed to increase coverage and ensure that a representative voting population is covered. This will be achieved by using a combination of stratified random sampling and quota sampling of key demographics.

B.2.1 Stratification variables

- Age: 18-29, 30-44, 45-64, 65+
- Gender: Male, Female, Non-binary
- Race: White, Black/African American, Hispanic/Latino, Asian, Native, Other
- Income Level: <$30,000, $30,000-$59,999, $60,000-$99,999, >$100,000
- Education level: High School, Some College, Bachelor's Degree, Graduate Degree, Other
- Geographic region: Northeast, Midwest, South, West
- Political Party: Democratic, Republican, Independent, Other

B.2.2 Sample size

The target sample size was 5,000 respondents, which are stratified by the Stratification Variables described above. This stratification ensures that all major voter groups are proportionately represented. 5,000 samples have a margin of error of about ±1% and a confidence level of 95%, which is appropriate for predicting elections.

B.3 Recruitment strategy

The recruitment strategy was separated into two main approaches to ensure a diverse and representative sample:

Online survey panels:

- Work with established survey panels such as Lucid or YouGov.
- These panels have access to large representative databases of voters.
- Recruiting through panels ensures quality and reliability, as they have built-in validation and quality control processes.

Social media advertising:

- Use social media platforms such as Facebook, Instagram and LinkedIn to reach potential respondents.
- Target groups that are underrepresented in traditional panels, including young voters and minority groups.
- Social media ads will be customized with demographic information to increase reach and participation.

Incentives: Offer gift card lottery prizes to all participants to encourage participation and ensure a high response rate.

B.4 Data validation and quality control

To maintain the integrity of the data, we will take the following measures:

- **Attention checks:** At least two 'attention check' questions will be included in the survey to identify inattention or robots.
- **Duplicate detection:** Use email or IP-based verification to prevent duplicate responses. Ensure that each respondent provides only one response set to keep the dataset unique.
- **Response time monitoring:** Track the time it takes respondents to complete surveys. Responses with apparently faster-than-average completion times are labelled as potentially low-quality.
- **Data consistency check:** Analyses the consistency of responses, especially between related questions. Inconsistent answers will be labelled for review.
- **Panel partner validation:** For respondents recruited through an online survey panel, rely on the panel provider's in-built validation systems. These systems use various quality checks, such as identity verification and response consistency, which ensure high-quality data from panel participants.

B.5 Poll aggregation and forecasting

B.5.1 Poll aggregation

- 'Poll of polls' approach: Survey data will be combined with data from some reliable polling sources such as Lucid, YouGov and others. This approach aggregates the results of multiple reputable polls to improve the forecast reliability by reducing bias and individual polling errors.
- Weighting adjustments: Polls are weighted according to sample quality, size, repeatability, and demographic match. Demographic mismatches are adjusted using raking to ensure that the aggregated data accurately represents the U.S. population.

B.5.2 Forecasting method

- Bayesian modelling: Aggregate polling data is combined with historical electoral trends and demographic information using Bayesian statistical models. This approach allows for dynamic updating as new data becomes available, which improves accuracy.
- Uncertainty Quantification: Includes confidence intervals to express uncertainty in forecasts for a more detailed interpretation.

B.6 Budget allocation

- Survey Panel Recruitment: $50,000
- Social Media Advertising: $50,000
- Incentives(lottery prizes): $15,000
- Data Cleaning and Validation: $5,000
- Poll Aggregation and Modeling: $10,000

B.7 Survey implementation

The survey will be available via Google Forms to ensure broad accessibility and easy data collection. The link to the survey will be distributed via email invitations to panellists and targeted social media adverts. To ensure maximum reach and diversity, multiple channels will be used, including Facebook, Instagram, and other social media platforms. The link to the survey is here.

Distribution plan:

- **Panel invitations:** Members of established survey panels (e.g. Lucid, YouGov) will receive email invitations to participate in the survey.
- **Social media activities:** Targeted adverts will be placed on social media platforms to reach under-represented groups, especially young voters and national minority groups.
- **Follow-up reminders:** An automated reminder email will be sent to participants who have not completed the survey to maximize response rates.
- **Survey Monitoring:** Response rates will be closely monitored throughout the data collection period and sampling methods will be adjusted if certain population quotas are not reached.
- **Accessibility:** The survey will be mobile-friendly, which will allow respondents to complete the survey on any device.

B.8 Survey structure

Survey introduction: Welcome! We are surveying to better understand voter preferences and priorities for the upcoming U.S. presidential election. Your participation is important and your answers will help us accurately predict the outcome of the election. It will take approximately 10 minutes to complete the survey. All responses are confidential and will be used for research purposes only.

If you have any questions or concerns, please feel free to contact our research team at winniekeai23@gmail.com (Ziyuan Shen, Yuanyi (Leo) Liu, Dezhen Chen).

As a thank you for your participation, you will be entered into a raffle to win a grand prize of a gift card. We are deeply grateful for your participation and appreciate the time and effort that you put in.

Survey question:

Part 1: Screening for eligibility

1.Are you an eligible U.S. citizen?
* Yes / No
2.Are you currently 18 years of age or older?
* Yes / No
3.Have you completed your registration to become a voter?
* Yes / No / Plan to register

Part 2: Demographics

4.What is your age?
* 18-24 / 25-34 / 35-44 / 45-54 / 55+
5.What is your gender?
* Male / Female / Non-binary / Prefer not to say
6.Which of the following best describes your race or ethnicity? * White / Black / African American / Hispanic or Latino / Asian / Native / Other
7.Which of the following best describes your education level?
* High School / Some College / Bachelor's Degree / Graduate Degree / Prefer not to say/ Other
8.What was your family/individual income last year? * <$30,000 / $30,000-$59,999 / $60,000-$99,999 / >$100,000 / Prefer not to say
9.Which region of the U.S. do you currently reside in?
* Northeast / Midwest / South / West
10.Do you generally identify with a political party?
* Democratic / Republican / Independent / None / Prefer not to say

Part 3: Voting intentions

11.How likely are you to participate in the upcoming presidential election?
* Very Likely / Somewhat Likely / Not Sure / Unlikely / Very Unlikely
12.If the presidential election were held today, for which candidate would you vote?
* Kamala Harris (Democrat) / Donald Trump (Republican) / Undecided / Prefer not to say
13.What are the top three issues that will influence your vote?
* Economy / Healthcare / Education / Gun Policy / Taxes / Foreign Policy / Social Security / Civil Rights / Technology / Privacy / Other[text]

Part 4: Information sources and engagement

14.Which of the following sources do you rely on most for political news and information?
* Television News / Newspapers / Online News Websites / Social Media / Radio / Friends

and Family / Other
15.How often do you engage in discussions about political topics with friends or family?
* Daily / Weekly / Monthly / Rarely / Never

Part 5: Verify

16.Please select 'Agree' to verify that you are paying attention.
* Agree / Disagree
17.Do you agree to participate in this survey? Your responses will be kept confidential and used only for research purposes.
* Yes / No

End Part:

Thank you for your participation!

Your responses are valuable to us in anticipating the upcoming election and understanding voters' priorities. If you have any questions or would like more information about our research, please feel free to contact us at winniekeai23@gmail.com(Ziyuan Shen, Yuanyi (Leo) Liu, Dezhen Chen).

C Data manipulation and cleaning

In the data cleaning process, we prepared the raw election data for analysis by applying transformations, filtering, and restructuring using several R packages, including `dplyr` (Wickham et al. 2023), `janitor` (Firke 2024), `tidyverse` (Wickham et al. 2019), `arrow` (Richardson et al. 2024), and `rsample` (Frick et al. 2024).

1. **Standardizing column names**: Using the `clean_names()` function, we converted column names to a consistent lowercase format, making them easier to reference and manipulate.
2. **Filtering polls by quality and candidate**: With the `dplyr` package from `tidyverse`, we filtered the dataset to include only higher-quality polls with a `numeric_grade` of 2.5 or above, ensuring that only credible polls were considered. We further filtered to focus on polls related specifically to Donald Trump (`candidate_name == "Donald Trump"`), aligning the dataset with our analysis objective.
3. **Date conversion and filtering**: We converted the `end_date` column to a Date format using `mdy()` from `lubridate` (part of `tidyverse`), ensuring date consistency. We then filtered the data to include only polls conducted after July 15, 2024—the date of Trump's campaign announcement—allowing us to focus on his active campaign period.
4. **Creating duration and support columns**:

- **Duration**: We created a `duration` variable using the `difftime()` function from base `R`, which calculates the number of days between each poll's end date and Trump's campaign start date. This allows us to observe how polling data shifts over time.
- **Number of Trump supporters**: We calculated `num_trump`, the estimated number of respondents supporting Trump in each poll, by applying the support percentage (`pct`) to the sample size and rounding the result with `dplyr` functions.

5. **Methodology update**: We used `str_detect()` from `stringr` (in `tidyverse`) to identify polls that listed multiple polling methods (e.g., those containing "/") and standardized them to "Mixed Voting" to clarify the `methodology` column.

6. **Selecting and renaming columns**: We used `dplyr` functions to select the relevant columns (`poll_id`, `numeric_grade`, `methodology`, `transparency_score`, `end_date`, `sample_size`, `pollster_name` (renamed from `display_name`), `state`, `pct`, `duration`, and `num_trump`), streamlining the dataset for analysis.

7. **Removing duplicates and missing values**: With `dplyr`, we removed duplicate rows using `distinct()` and dropped rows with missing values using `drop_na()`, ensuring that only complete and unique data entries were retained.

8. **Splitting the data**: We utilized `rsample` to perform a stratified split on the cleaned dataset, using the `state` variable to ensure balanced representation across states. The data was divided into training (70%) and testing (30%) sets, preparing it for modeling.

Finally, we saved the cleaned dataset and the training/testing splits in CSV and Parquet formats using `arrow`, facilitating efficient storage and future accessibility.

D Model details

D.1 Model summary

Table 6 presents the coefficients from our model analyzing factors that influence polling outcomes. The intercept value, 44.178, represents the baseline level of support when all other predictors are at their reference or baseline levels. Key variables include "numeric_grade," with a positive coefficient indicating that higher grades correlate with increased support. Different polling methodologies show varying effects; for example, "Mixed Voting" has a positive coefficient of 1.697, while "Online Panel" and "Probability Panel" methods have negative coefficients, suggesting that these methods might be associated with lower support levels in polling outcomes.

Table 6: Coefficients from a regression model examining factors influencing polling outcomes, including polling methodology, duration, transparency, and state-level indicators.

	coefficient
(Intercept)	44.178
numeric_grade	0.794
as.factor(methodology)Mixed Voting	1.697
as.factor(methodology)Online Ad	0.957
as.factor(methodology)Online Panel	−1.034
as.factor(methodology)Probability Panel	−1.042
duration	0.028
sample_size	0.002
transparency_score	−0.219
stateCalifornia	−15.440
stateConnecticut	−12.525
stateFlorida	3.034
stateGeorgia	−0.343
stateIndiana	8.087
stateMaine	−4.212
stateMaryland	−18.070
stateMassachusetts	−15.594
stateMichigan	−1.031
stateMinnesota	−4.158
stateMissouri	6.835
stateMontana	7.622
stateNebraska	−3.796
stateNevada	−1.660
stateNew Hampshire	−4.135
stateNew Mexico	−5.961
stateNew York	−8.545
stateNorth Carolina	−0.751
stateOhio	1.606
statePennsylvania	−1.395
stateRhode Island	−9.639
stateSouth Carolina	1.865
stateTexas	1.876
stateVermont	−17.375
stateVirginia	−5.150
stateWashington	−12.253
stateWisconsin	−1.668
Num.Obs.	242
R2	0.803
R2 Adj.	0.753
Log.Lik.	−526.328
ELPD	−563.9
ELPD s.e.	16.5
LOOIC	1127.7
LOOIC s.e.	33.1
WAIC	1121.8
RMSE	2.07

D.2 Posterior predictive check

In Figure 5, we implement a posterior predictive check, which displays the overlap between the observed data (denoted by y) and the replicated data generated from the model (denoted by y_{rep}). In a well-fitting model, the distribution of these replicated data should resemble the distribution of the observed data. Here, the replicated lines closely follow the shape and central tendency of the observed data's density, especially around the peak. This similarity indicates that the model is capturing the overall pattern of the data reasonably well, suggesting a good fit. Deviations between the replicated and observed lines would indicate areas where the model might not be accurately capturing the data structure.

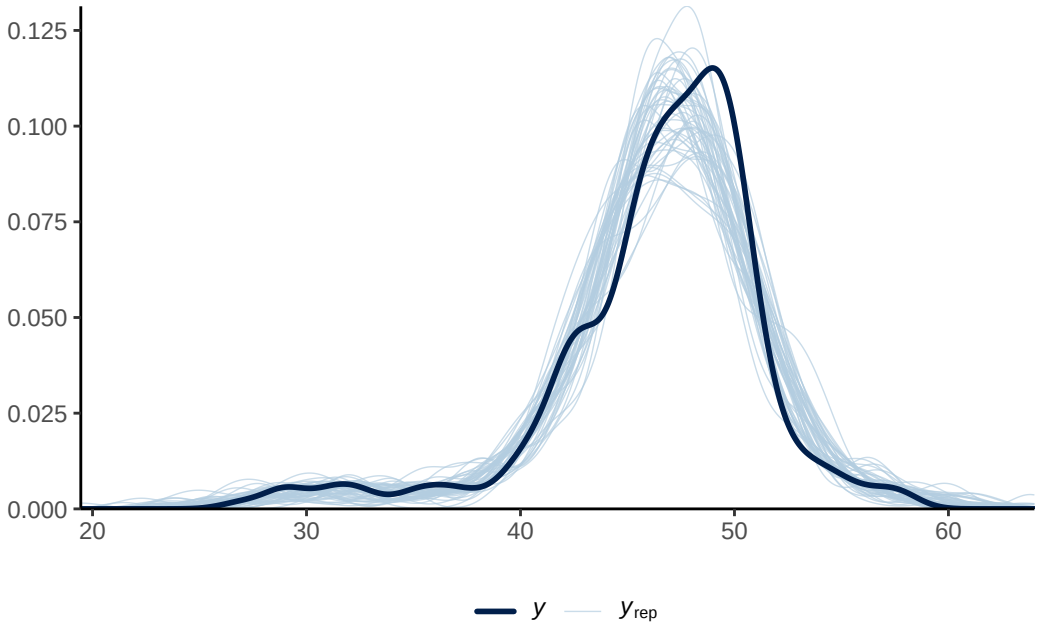

Figure 5: Comparison of Observed Data Density (dark line) and Model-Generated Replications (light lines) shows close alignment, suggesting the model captures the main distributional characteristics of the data.

D.3 Diagnostics

Figure 6 is a visualization of the Gelman-Rubin diagnostic, often denoted as R-hat, which assesses the convergence of a Bayesian model's Markov Chain Monte Carlo (MCMC) samples. It shows the values of for various parameters in the model. An value close to 1 indicates good convergence, meaning the chains for each parameter have mixed well and are sampling from the same posterior distribution. In Figure 6, all parameters have an value at or below 1.05, suggesting that the model has achieved adequate convergence and the estimates are reliable.

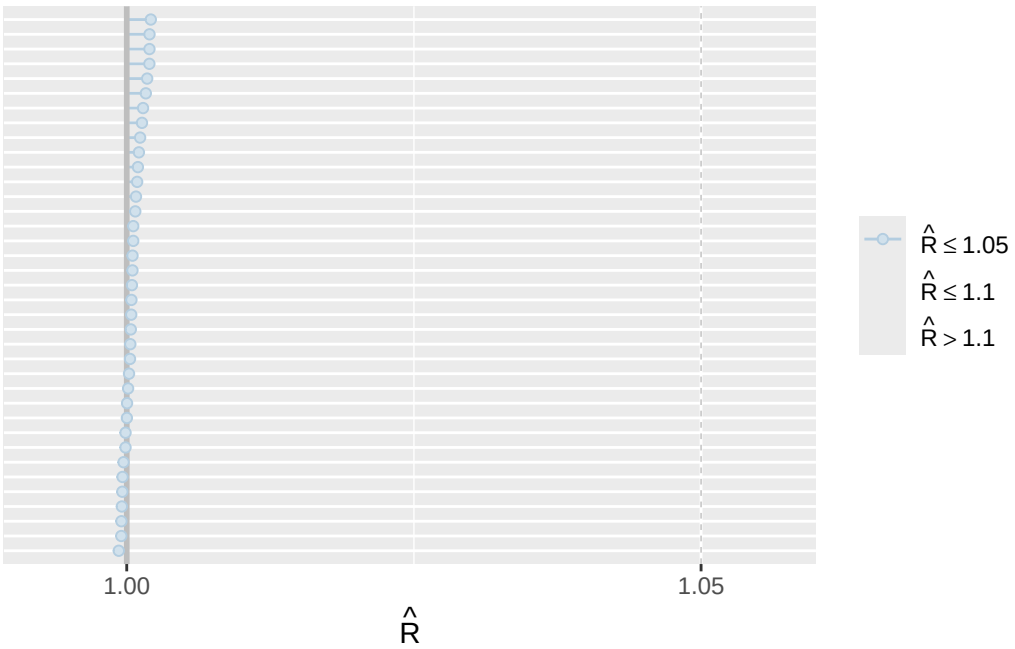

Figure 6: All parameters have R-hat values at or below 1.05, indicating strong convergence of the MCMC chains and reliable parameter estimates.

References

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

Best, Ryan, and Aaron Bycoffe. 2024. "National: President: General election: 2024 polls." *FiveThirtyEight*. https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

Di Lorenzo, Paolo. 2024. *usmap: US Maps Including Alaska and Hawaii*. https://usmap.dev.

Firke, Sam. 2024. *janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Frick, Hannah, Fanny Chow, Max Kuhn, Michael Mahoney, Julia Silge, and Hadley Wickham. 2024. *rsample: General Resampling Infrastructure*. https://rsample.tidymodels.org.

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

Nagourney, Adam, and Ruth Igielnik. 2024. "Harris and Trump deadlocked to the end, final times/Siena National Poll finds." *The New York Times*. The New York Times. https://www.nytimes.com/2024/10/25/us/politics/poll-harris-trump-times-siena.html.

Nolan, Jacqueline V. 2008. "United States electoral college, votes by State." *Library of Congress*. https://www.loc.gov/resource/g3701f.cp000001/.

R Core Team. 2023. *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://github.com/apache/arrow/.

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