Predicting Electoral Outcomes: Bayesian Modeling of Voter Support Trends for Kamala Harris and Donald Trump in the 2024 Presidential Election*

Xinqi Yue Yawen Tan Duanyi Su

November 4, 2024

This paper builds a predictive Bayesian model to forecast the 2024 US presidential election using a "poll-of-polls" approach, analyzing high-quality polls for Kamala Harris and Donald Trump. The model incorporates data on polling methodology, state trends, and candidate support. Our findings indicate that although no candidate is guaranteed to win a total of 270 electoral votes due to lack of polls in certain states, Kamala Harris is more likely to win the election with 222 electoral votes against Donald Trump's 109 electoral votes. This study contributes to an understanding of electoral forecasting, helping us interpret aggregated polling data to anticipate election results.

Table of contents

*Code and data are available at: https://github.com/xinqiyue/2024_US_Election_Prediction.

1 Introduction

The 2024 U.S. presidential election is generating significant public interest, with recent polling reflecting shifts in candidate support across various demographic and regional groups. Polling data, while insightful, presents challenges due to inherent biases, variability in polling methodology, and regional influences on voter behavior. This study aims to address these challenges using a "poll-of-polls" approach, aggregating high-quality national and state polls to forecast voter support for Kamala Harris and Donald Trump. By applying Bayesian modeling, this analysis incorporates both the temporal dynamics of candidate support and state-specific trends.

Our model predicts trends in voter support for the 2024 U.S. Presidential Election, focusing on Donald Trump and Kamala Harris. The results, visualized in Figure 5, indicate stable support for Harris at an estimated 52% nationally, with a confidence interval of approximately 49% to 55%. Trump's predicted support level centers around 46%, with a range of about 43% to 49%, showing a slight downward trend as the election date nears. State-specific polling results highlight variances, with Harris showing stronger support in traditionally Democratic areas like Maine CD-1, where predicted support reaches around 62%. In contrast, battleground states such as Georgia, Florida, and Pennsylvania exhibit closely contested support levels, reflecting competitive races in these regions. These predictions underscore the variability in voter sentiment across states, illustrating the importance of both national and regional factors shaping the electoral landscape.

In remainder of this paper, we structure our analysis as follows. Section 2 provides a detailed overview of the data used, including the data cleaning process, explanations of the predictor and outcome variables, as well as other variables that contribute to the model. We also present relevant data visualizations to illustrate initial trends and patterns. Section 3 focuses on the model setup and methodology, specifically detailing the Bayesian model used, its underlying assumptions, and software choices. Section 4 presents our main results, including both the model's outputs and the predictions of voter support for the key candidates. In Section 5, we discuss the findings, offer insights into the broader implications, and highlight the limitations and potential directions for future research. Finally, the appendices provide additional methodological analysis, an idealized survey methodology, and supporting details for the data and model diagnostics.

1.1 Estimand

Our primary estimand is the percentage of voter support for Donald Trump and Kamala Harris on election day. We estimate this support percentage both at the national level and by state, incorporating time and state-specific effects.

2 Data

2.1 Data Overview

Section 2 centers on a polling dataset for Donald Trump and Kamala Harris, filtered for quality by including only results from pollsters with a numeric grade above 2.5. Key variables such as state, end date, sample size, candidate name, and support percentage (pct) are identified. The data cleaning process standardizes column names, retains relevant fields, and addresses missing values to ensure high-quality data for analysis. The section includes a table presenting the first six entries of the cleaned analysis dataset, alongside multiple graphs: a bar graph comparing candidate support percentages by state, scatter plots illustrating trends in candidate support over time, and visualizations examining the relationship between sample size and support percentages.

The analysis utilizes FiveThirtyEight's dataset of national presidential general election polls (FiveThirtyEight 2024). Following the approach outlined in (Alexander 2023), we aim to predict the election outcome based on this polling data. The analyses were conducted in R (R Core Team 2023), utilizing several packages to streamline the process. The tidyverse packages (Wickham et al. 2019) were essential for data simulation and preliminary testing. Initially, the raw data was downloaded using the arrow package (Richardson et al. 2024), allowing for efficient handling of large datasets. The data cleaning process employed the dplyr (Wickham et al. 2023) and tidyr (Wickham, Vaughan, and Girlich 2024) packages to reshape and organize the data effectively.

Model construction involved the rstanarm (Goodrich et al. 2022) package for Bayesian regression modeling, complemented by the splines package (R Core Team 2024) to account for non-linear relationships in the data. The results were summarized using the modelsummary (Arel-Bundock 2022) package, while visualizations of the findings were created with the ggplot2 (Wickham 2016) package, ensuring clear communication of the model's outputs.

2.2 Data Clean

We clean the raw data to produce the analysis dataset using the tools listed in the Section 2.1, providing high-quality data for subsequent election analysis. Firstly, we standardize column names to lowercase and remove special characters. Then, we retain only the fields relevant to the election analysis, such as `state`, `end_date`, `sample_size`, `candidate_name`, and `pct` (percentage support), simplifying the data structure by removing unnecessary columns. Next, we remove rows with missing values and replace empty values in the `state` column with "National," indicating these records pertain to nationwide polls. Additionally, we convert the `end_date` column to date format, ensuring accuracy in date-based filtering. Then, we filter the data to retain only records that meet specific criteria: a `numeric_grade` of 2.5 or higher, a `candidate_name` of either "Kamala Harris" or "Donald Trump" (focusing on these two

candidates' support levels), an `end_date` on or after July 21, 2024 (focusing on recent polling when Biden withdrawal from the election), and a non-empty `end_date` (filtering out records with missing dates).

The cleaned data is displayed in Table 1, which contains a total of 5 variables:

Table 1: First 6 entries of Analysis Dataset

State	End date	Sample Size	Canadidate Name	PCT
National	2024-10-16	1000	Kamala Harris	49.5
National	2024-10-16	1000	Donald Trump	49.2
Arizona	2024-10-16	1435	Kamala Harris	48.0
Arizona	2024-10-16	1435	Donald Trump	51.0
National	2024-10-15	1457	Kamala Harris	48.0
National	2024-10-15	1457	Donald Trump	45.0

2.3 Predictors Explanation

- **state**: The U.S. state where the poll was conducted, which allows for state-by-state analysis and comparison of support levels.
- **end date**: The date the poll concluded, marking the end of data collection for that specific poll.
- **sample size**: The number of respondents in the poll, indicating the scope and potential statistical reliability of the results.
- **candidate_name**: The name of the candidate being polled, which in this data set focuses on Donald Trump and Kamala Harris.

2.4 Outcome Exaplanation

- **pct**: The support percentage each candidate received in the poll, which serves as the outcome variable for analysis.

pct is the support percentage each candidate received in the poll, which serves as the outcome variable for analysis. This variable is crucial for understanding the electoral dynamics, as it quantifies the level of public backing for each candidate at a given time and place. By analyzing the pct variable, researchers can assess trends in voter sentiment, compare the effectiveness of campaign strategies, and identify which demographic groups are more inclined to support one candidate over the other. Additionally, fluctuations in the pct can signal shifts in public opinion that may correlate with significant political events, media coverage, or changes in candidates' messaging. This outcome variable allows for a comprehensive analysis of the

electoral landscape, providing insights into how various factors influence voter behavior and preferences throughout the campaign.

2.5 Explanation of Other Variables Used

- **pollster**: The organization that conducted the poll, providing information on who gathered the data.
- **numeric grade**: A quality score assigned to the pollster, with higher values indicating greater reliability. Only pollsters with a score above 2.5 are included to enhance accuracy.
- **pollscore**: A specific rating of the individual poll, reflecting additional factors that impact the poll's quality and reliability.
- **methodology**: The method used by the pollster to conduct the poll, which may affect the reliability and interpretation of the results.
- **transparency score**: A measure of how openly the pollster reports their methodology and results. Higher scores suggest greater transparency and reliability.

2.6 Measurement

The goal of this measurement is to turn individual voter opinions into a reliable estimate of electoral college outcomes by analyzing polling data. This data, sourced from FiveThirtyEight (2024)—a trusted platform known for high standards—includes only polls that meet rigorous quality criteria, ensuring broad representation of likely U.S. voters. Polls provide essential details, such as the pollster's identity, survey dates, sample size, and methodology, covering data collection mode, demographic weighting, and adjustments to create a representative sample. These quality standards, while robust, are based on historical practices, which may overlook recent methodological shifts or emerging biases. Polling has inherent limitations. It provides snapshots of voter sentiment at specific times rather than ongoing updates, potentially missing rapid shifts in public opinion, especially near Election Day. Adjustments are made for recency, but participation and response biases—like self-selection and social desirability bias—can still impact accuracy by creating gaps between expressed opinions and actual voting behavior. Additionally, regional polling disparities, with battleground states polled more frequently than "safe" states, can lead to imbalances in representation, highlighting the challenge of achieving uniformly accurate forecasts across the nation.

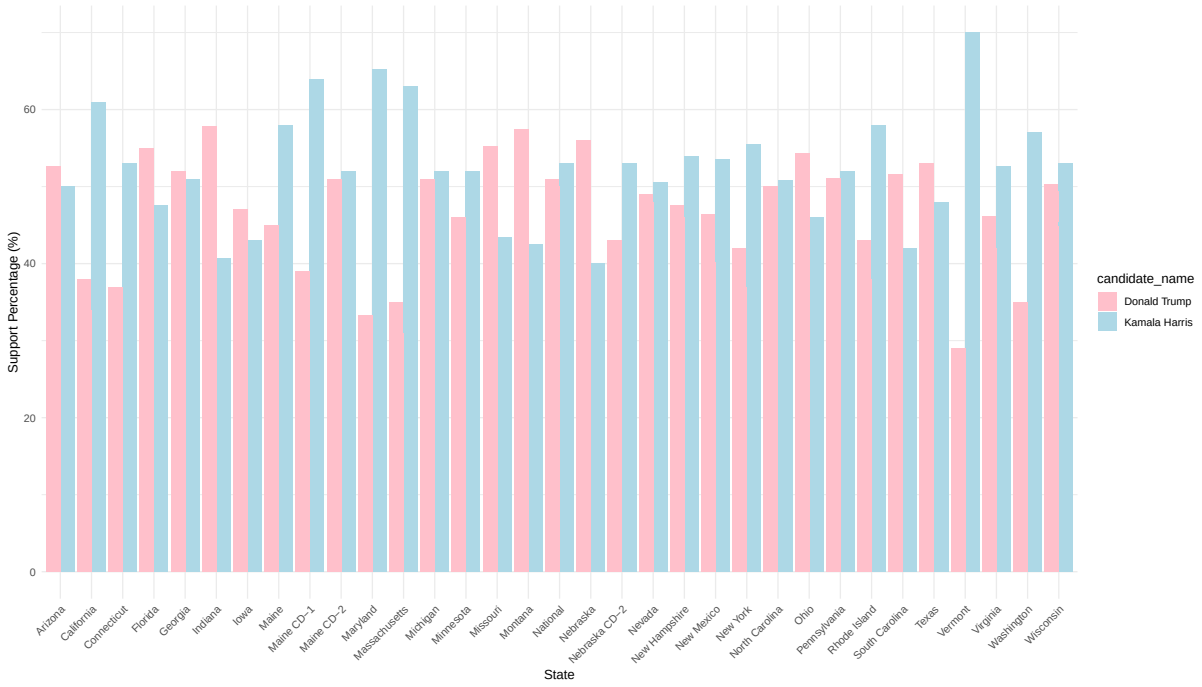

Figure 1: Support for Candidates by State

2.7 Data visualization

2.7.1 Exploring the Relationship Between State and PCT

Figure 1 compares support percentages for Donald Trump and Kamala Harris across U.S. states, using pink bars for Trump and light blue for Harris. The y-axis shows support percentages up to around 70%, while the x-axis lists states, each with two bars per candidate. Figure 1 reveals regional popularity variations. In some states, bars for both candidates are nearly equal in height, indicating competitive support. Other states show one candidate with a clear lead, highlights the varying levels of popularity for each candidate in different regions, which may reflect political, demographic, or regional factors influencing voter preferences. For instance, California leans toward Harris, with a taller blue bar, while Texas favors Trump, with a taller pink bar. These patterns reflect regional political tendencies, suggesting which states might be strongholds or battlegrounds. They highlight potential areas where each candidate could focus efforts to strengthen support or appeal to undecided voters.

2.7.2 Exploring the Relationship Between End Date and PCT

Figure 2 visualizes support percentages for Donald Trump and Kamala Harris over time, with each point representing a poll. The x-axis shows poll end dates, extending beyond July 21st,

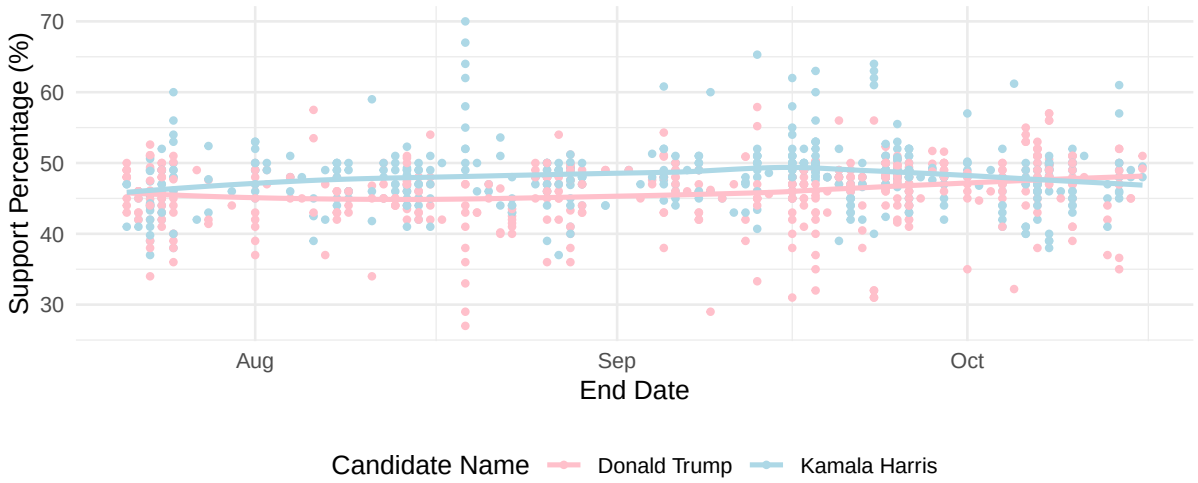

Figure 2: Trends in Candidate Support Over Time

while the y-axis reflects support percentages. Pink dots represent Trump, and light blue dots represent Harris.nTrend lines for each candidate show Harris consistently above Trump, indicating a slight but steady lead. Harris's trend remains around 50%, while Trump's is just below. Both lines are stable, with no major fluctuations in support over time, suggesting limited change in public opinion. The close clustering of points around each trend line indicates steady support levels for both candidates, with Harris maintaining a modest lead throughout the period.

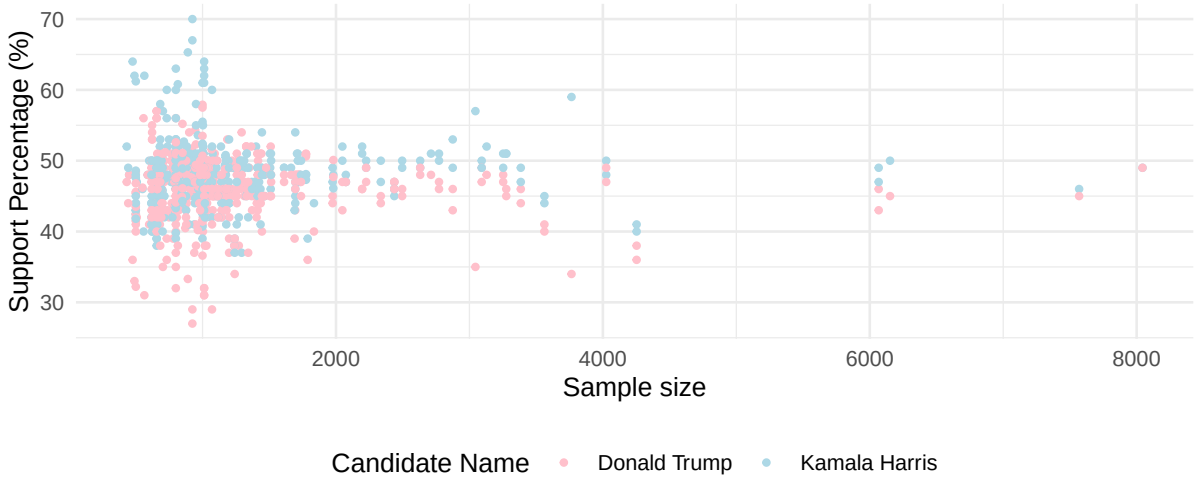

Figure 3: Support for Candidates by Sample Size

Figure 3 examines the relationship between poll sample size and support percentage for Donald Trump and Kamala Harris. The x-axis represents poll sample sizes, ranging from a few hundred to over 6,000, while the y-axis shows support percentages. Pink dots indicate Trump's support,

and light blue dots represent Harris. Most polls have sample sizes under 2,000, with support percentages clustering between 40% and 55% for both candidates. As sample sizes increase, support percentages stabilize around the 45%–55% range, though fewer large-sample polls exist. Smaller polls show more variability, with support percentages ranging from 30% to 60%, while larger polls yield more consistent results near 50%.Overall, the plot suggests that larger polls offer more stable estimates of support, while smaller polls tend to show greater fluctuation.

2.7.3 Exploring the Relationship Between End Date and State and PCT

Figure 4 comprises scatter plots illustrating the support rates for Donald Trump and Kamala Harris across various states, with the y-axis representing percentage support and the x-axis indicating the timeline. Red dots represent Trump's support, while blue dots indicate Harris's. The visualization shows fluctuating support levels for both candidates, with Harris generally receiving higher support in states like California and New York, whereas states like Florida and Indiana display a more competitive landscape. The national plot reveals trends in support, reflecting periods of gain or loss for both candidates, suggesting that voter sentiment is dynamic and influenced by external factors during the campaign. Overall, the figure highlights the changing electoral support for Trump and Harris across states and time.

3 Model

In this analysis, we use a Bayesian linear regression model to estimate the percentage of support (`pct`) for each candidate in the upcoming US presidential election. The model includes key predictors such as polling end date (`end_date`), candidate (`candidate_name`), sample size (`sample_size`), and state (`state`). Each predictor is selected based on its relevance to the variation in candidate support, as discussed in the data section.

3.1 Model set-up

In this Bayesian framework, support level are modeled as normally distributed and influenced by several predictors: end date, candidate name, sample size, and state. The variable y_i denotes the percentage of votes for a candidate in a specific poll, with β_i representing the candidate effect, γ_i reflecting the influence of sample size, and δ_i corresponding to the state effect. The intercept α indicates the baseline poll result, while each β_i coefficient quantifies the influence of its associated predictor on the vote percentage. Additionally, the variable state_j captures the effects attributed to different states, and end_date_i models temporal trends.

The model can be mathematically expressed as follows:

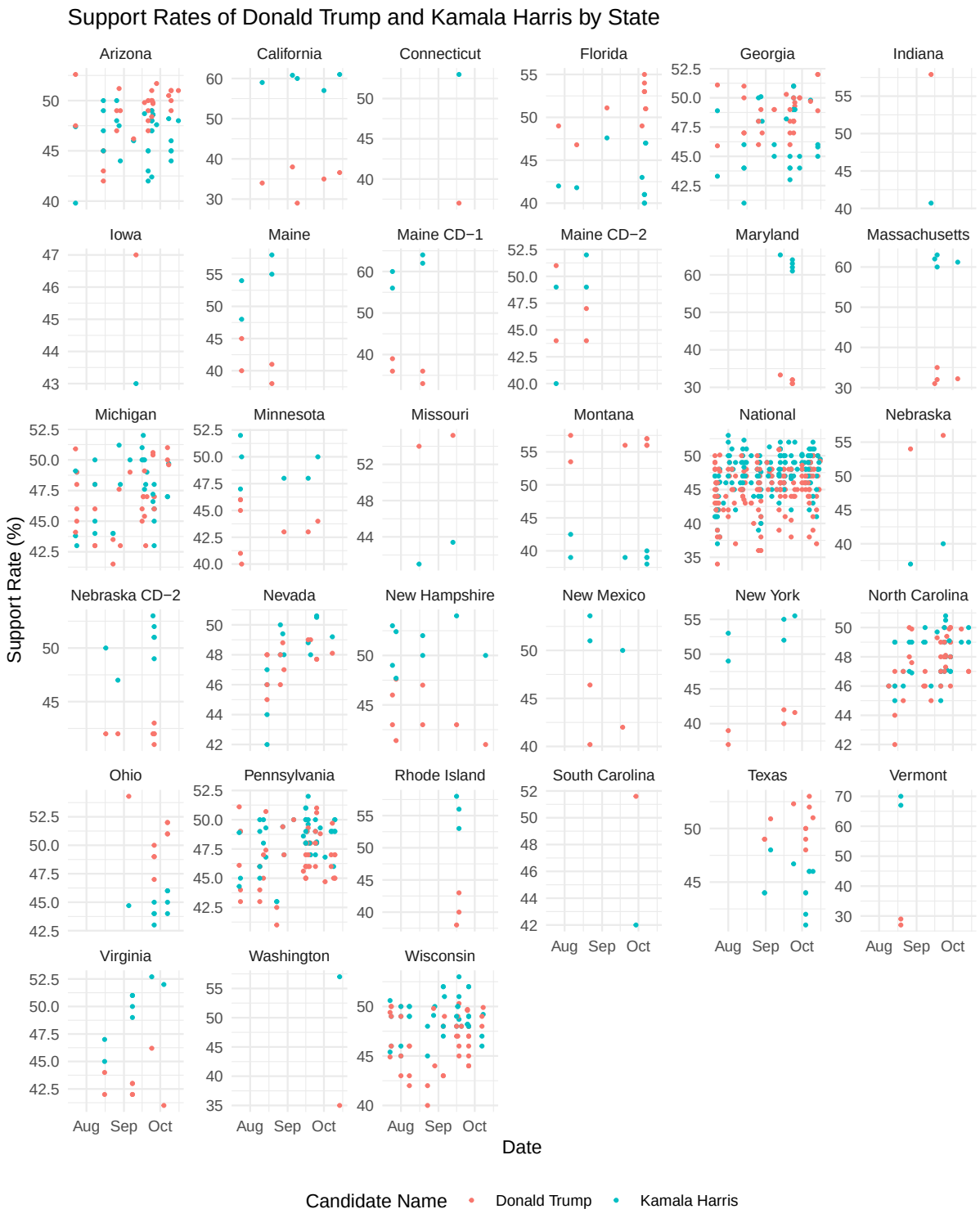

Figure 4: Support Rates of Donald Trump and Kamala Harris by State and end date

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$

$$\mu_i = \alpha + \beta_1 \cdot \text{end\_date}_i + \beta_2 \cdot \text{candidate\_name}_i + \beta_3 \cdot \text{sample\_size}_i + \sum_{j=1}^{M} \gamma_j \cdot \text{state}_j \tag{2}$$

$$\alpha \sim \text{Normal}(50, 10) \tag{3}$$

$$\beta_1, \beta_2, \beta_3 \sim \text{Normal}(0, 2.5) \tag{4}$$

$$\gamma_j \sim \text{Normal}(0, 2.5) \tag{5}$$

$$\sigma \sim \text{Exponential}(1) \tag{6}$$

Priors are defined as follows. For the intercept, a `Normal(50, 10)` distribution is utilized, reflecting a belief that the baseline support percentage is centered around 50% with moderate variability. This choice indicates a moderately informative prior that captures plausible ranges for the intercept. Each regression coefficient is assigned a `Normal(0, 2.5)` prior, which is weakly informative, allowing flexibility in parameter estimation while minimizing bias. These priors are selected to strike a balance between allowing the data to inform estimates and providing regularization to prevent extreme values and overfitting. The chosen priors are grounded in reasonable expectations regarding support ranges, aiming to enhance the robustness of the model.

Then, the complete model, summarizing the components, can be expressed as:

$$y_i \sim \text{Normal}\left(\alpha + \beta_1 \cdot \text{end\_date}_i + \beta_2 \cdot \text{candidate\_name}_i + \beta_3 \cdot \text{sample\_size}_i + \sum_{j=1}^{M} \gamma_j \cdot \text{state}_j, \sigma^2\right)$$

To further clarify the model:

- **Formula (1)** shows that the response variable y_i is normally distributed, with μ_i as the expected value and σ representing standard deviation.
- **Formula (2)** illustrates that μ_i is derived from the intercept α and the predictors.
- **Formula (3)** details the basis functions applied to the state variable, allowing for a nuanced relationship between states and vote percentages. The prior distributions are:
- **Formula (4)** specifies that α has a normal distribution with mean 50 and variance 10^2, indicating our expectation of average vote percentage.
- **Formula (5)** outlines the priors for γ_k, which are normal distributions centered at 0 with variance 2.5^2, reflecting the belief in potentially small effects from state variables.
- **Formula (6)** assigns an exponential distribution to the standard deviation σ, suggesting that smaller variance is more likely while allowing for broader variations.

3.2 Assumptions of the Bayesian Models

1. **Linearity**: The model assumes a linear relationship between independent variables (predictors) and the dependent variable (outcome, `pct`). The expected outcome is represented as a linear combination of the predictors. Nonlinear effects, if present, may not be fully captured, potentially affecting the model's fit.
2. **Normality of Errors**: It is assumed that the residuals (the differences between observed and predicted values) follow a normal distribution. This assumption is crucial for making valid inferences; violations can lead to incorrect conclusions and unreliable credible intervals. Diagnostic checks and posterior predictive checks are used to assess model fit and adherence to assumptions.
3. **Homoscedasticity**: The variance of the residuals is expected to be constant across all combinations of the predictors. If residuals display heteroscedasticity (non-constant variance), it can result in inefficient estimates and biased interpretations of the modeled relationships.
4. **Independence of Errors**: The model assumes that errors are independent of one another, meaning the error term for any single observation does not affect the error term for another observation. However, polling data often includes repeated measures from the same pollster, which may introduce dependence among observations.
5. **Additivity**: Predictors are assumed to additively influence `pct`, meaning interactions or non-additive effects are not considered. Future versions of the model might benefit from testing interactions, such as between state and candidate.
6. **Prior Distributions**: Bayesian regression requires the specification of prior distributions for the model parameters (coefficients). The selection of these priors is significant, as they can impact the resulting posterior distributions, particularly when data is limited.
7. **Parameter Estimation**: In Bayesian linear regression, parameters are estimated using a posterior distribution derived from both the likelihood of the observed data and the prior distributions. This framework allows for uncertainty quantification and more flexible inference.

3.3 Software and Validation

The model is implemented using the `rstanarm` package in R, which allows for efficient Bayesian inference with automatic convergence diagnostics. Model fit and convergence were monitored through trace plots and effective sample sizes. To further ensure robustness, we performed posterior predictive checks using `pp_check(model_bayes)` to evaluate model fit and identify any systematic deviations.

Overall, this Bayesian model provides a balanced approach, capturing key effects while maintaining interpretability and flexibility, making it suitable for predicting candidate support percentages in the election context.

3.3.1 Model justification

In this analysis, we utilize a Bayesian linear regression model to predict the percentage of voter support for Donald Trump and Kamala Harris in the 2024 U.S. Presidential Election. This choice of modeling framework is particularly suitable for our objectives, as it allows for the incorporation of prior knowledge regarding electoral dynamics while updating predictions based on newly observed data. The Bayesian approach provides a robust mechanism for estimating uncertainty in our predictions, which is crucial in the context of political polling where public sentiment can fluctuate significantly. The model structure includes key predictors: the polling end date, candidate name, sample size, and state. The use of these predictors is justified by their established relevance in previous electoral analyses, highlighting their roles in shaping voter preferences. The inclusion of `end_date` facilitates the examination of temporal trends in support, reflecting how candidate visibility and public sentiment evolve as the election date approaches. We model the outcome variable, support percentage (pct), as a normally distributed random variable influenced by these predictors. The mathematical formulation allows for the representation of the intercept, which denotes the baseline support level, alongside coefficients that quantify the influence of each predictor. The priors assigned to the model parameters are carefully chosen to reflect reasonable expectations based on historical voting patterns, thus enhancing the model's interpretability and effectiveness. Moreover, the Bayesian framework enables the modeling of potential interactions between state-specific effects and candidate support, allowing us to capture the regional variations that are often observed in electoral outcomes. By incorporating individual pollster effects as fixed coefficients, we can account for the inherent biases associated with different polling organizations, ensuring that our model provides a nuanced view of candidate support. This model's formulation strikes a balance between complexity and interpretability, allowing for a comprehensive analysis of the factors influencing voter sentiment while maintaining computational efficiency. The Bayesian approach not only enhances our understanding of electoral dynamics but also provides a statistically sound basis for forecasting potential election outcomes. Overall, the chosen model is well-aligned with the objectives of this study, facilitating a robust analysis of polling data to derive meaningful insights into the upcoming election.

4 Results

4.1 Model Result

Table 2 provide insights into the factors influencing voter support for Donald Trump and Kamala Harris, based on a total of 846 observations. The model intercept is estimated at -397.72,

Table 2: Model Results of Trump Percentage Vote based on Date and State

	Bayesian Model
(Intercept)	−397.72
end_date	0.02
candidate_nameKamala Harris	1.81
sample_size	0.00
stateCalifornia	−1.02
stateConnecticut	−1.11
stateFlorida	−0.53
stateGeorgia	0.35
stateIndiana	0.79
stateIowa	−0.97
stateMaine	−0.23
stateMaine CD-1	4.05
stateMaine CD-2	1.80
stateMaryland	−1.45
stateMassachusetts	−1.85
stateMichigan	−0.15
stateMinnesota	−0.76
stateMissouri	1.66
stateMontana	−1.02
stateNational	−1.18
stateNebraska	−0.40
stateNebraska CD-2	−1.24
stateNevada	0.18
stateNew Hampshire	0.41
stateNew Mexico	1.73
stateNew York	−1.92
stateNorth Carolina	0.32
stateOhio	−0.64
statePennsylvania	0.23
stateRhode Island	0.31
stateSouth Carolina	−0.51
stateTexas	0.02
stateVermont	−2.37
stateVirginia	−0.72
stateWashington	−0.85
stateWisconsin	0.43
Num.Obs.	846
R2	0.115
R2 Adj.	−0.056

serving as a baseline from which other coefficients are interpreted. Notably, the coefficient for the polling end date is 0.02, indicating a positive relationship between the date and support percentage, suggesting that as the election approaches, support for the candidates slightly increases, likely due to heightened visibility and engagement. The coefficient for Kamala Harris is 1.81, reflecting her higher expected support relative to Donald Trump, which indicates her appeal among voters during the polling period. Table 2 also includes state-specific coefficients that capture regional variations in support. For instance, states such as Maine CD-1 (4.05), Missouri (1.66), and New Mexico (1.73) show positive coefficients, indicating stronger support for the candidate in these regions. Conversely, negative coefficients for states like Vermont (-2.37), New York (-1.92), and Massachusetts (-1.85) suggest lower support levels, highlighting potential challenges in these areas. However, the model's adjusted R^2 value of -0.056 indicates that the predictors included do not explain a substantial amount of the variability in support percentages, suggesting that other unmeasured factors may influence voter preferences. The log-likelihood value of -2412.946, along with an effective sample size for the expected log pointwise predictive density (ELPD) of -2463.8 and a leave-one-out information criterion (LOOIC) of 4927.6, indicate moderate predictive accuracy. Additionally, the root mean square error (RMSE) of 4.18 suggests that the predicted percentages deviate from actual support levels by approximately 4.18 percentage points on average. Overall, these results underscore the importance of both temporal dynamics and regional differences in predicting voter support, while also indicating a need for further exploration of additional variables that may more effectively account for observed voter preferences.

Figure 5 produced predicted percentages of voter support for Donald Trump and Kamala Harris over time, as shown in the graph titled "Predicted Percentage Over Time." The y-axis represents the predicted support percentage, while the x-axis indicates polling end dates. Figure 5 reveals distinct trends in voter support for both candidates. The predictions for Kamala Harris (shown in teal) suggest a mean support level consistently around 52% during this period, with the confidence interval shaded in gray indicating a range from approximately 49% to 55%. This stability suggests that Harris maintains a competitive edge in voter support as the election date approaches.In contrast, the predicted percentages for Donald Trump (shown in red) indicate a mean support level around 46%, with a confidence interval ranging from about 43% to 49%. This pattern highlights a slight downward trend in Trump's predicted support, suggesting potential challenges in consolidating voter sentiment as the election draws near. The individual data points on the graph represent specific state polling results, with various shapes indicating different states. For example, Maine CD-1 stands out with a higher predicted support for Harris at approximately 62%, which is significantly above the national average, indicating a strong preference in this district. Conversely, states like Florida and North Carolina exhibit more competitive support levels, with both candidates' predicted percentages clustering around 47% to 50%.The presence of a broader range of predicted support percentages in battleground states such as Georgia and Pennsylvania reflects the volatility of voter sentiment in these regions, where Trump and Harris show closely contested support. The model indicates that Harris may benefit from higher support levels in traditionally Democratic states, while Trump faces a challenge in securing sufficient backing in key swing states. Over-

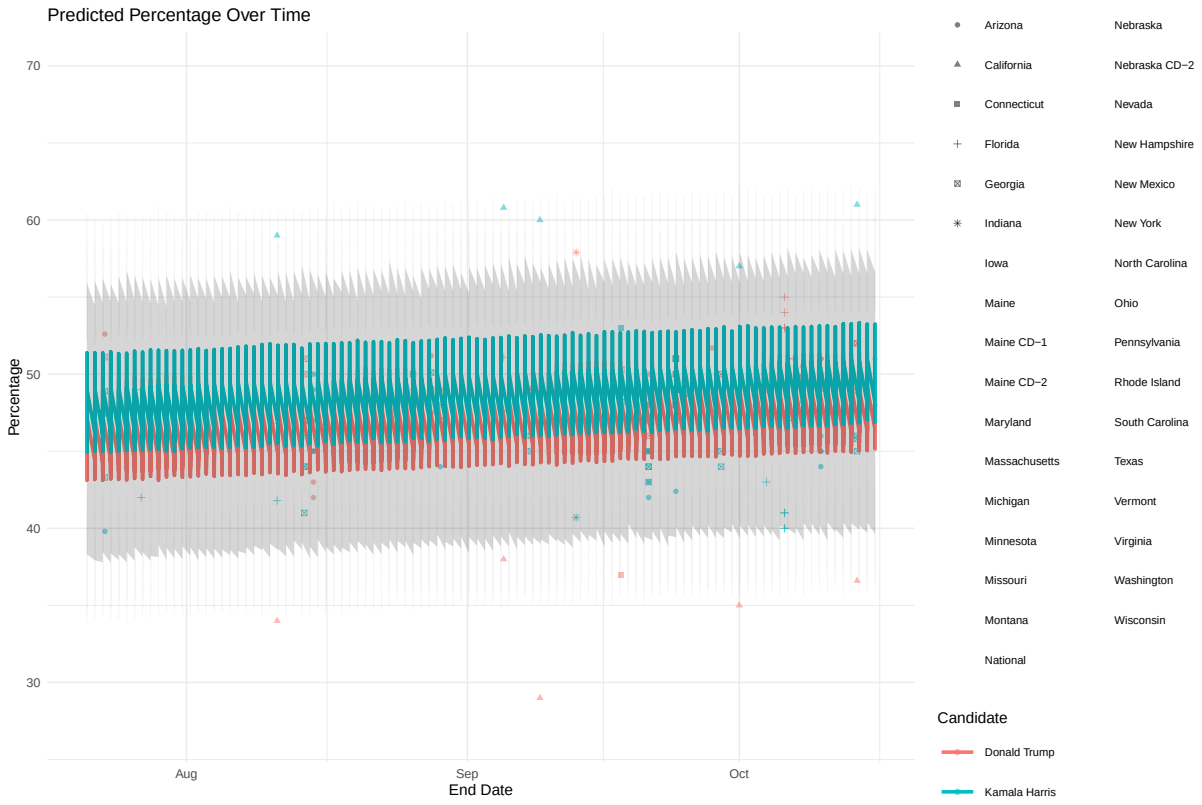

Figure 5: Predict Posterior Draws and Spline Fit for Vote Percentage

all, Figure 5 underscore the dynamic nature of voter preferences as the election approaches, with Harris appearing to lead in overall support while Trump must contend with fluctuating voter sentiments in critical states. These predictions provide a view of how various factors, including time and state-specific influences, shape the electoral landscape leading up to the 2024 U.S. Presidential Election.

4.2 Predict Result

Figure 6 produced a summary table of average support percentages for Donald Trump and Kamala Harris across various states, as well as Figure 7 predicted total vote counts based on these averages. Figure 6 highlights significant differences in voter support by state, revealing regional dynamics as the election approaches.

Figure 6 indicates that Kamala Harris tends to have higher support in several key states, such as California, where she leads with an average of 59.98% compared to Donald Trump's 34.29%. Similarly, Maryland shows strong backing for Harris at 62.90%, while Trump receives only 32.39% of the vote. In contrast, Trump demonstrates stronger support in traditionally Republican-leaning states like Indiana, where he averages 57.90%, and Florida, with 51.53% support. The data also highlights a competitive landscape in states like Georgia and Michigan, where the candidates are closely matched; in Georgia, Harris has 47.19% compared to Trump's 48.61%, while in Michigan, Harris garners 47.84% against Trump's 46.92%.

By combining the average support percentage of each state with the electoral votes in Figure 6, we calculated the predicted total votes of each candidate. Each state supports a large percentage of candidates to get the total number of votes in this state, and the total number of votes refers to 1Keydata (2024). According to these calculations, Figure 7 shows that Trump got a total of 109 votes, while Harris got 222 votes. Thus, although both of Donald Trump and Kamala Harris is not guaranteed to win 270 votes due to lack of data in certain states, Harris has obvious advantages in the election, especially with the strong support of several key states, and predicts that she may win this election.

5 Discussion

5.1 Overview of the Paper

This paper presents a comprehensive analysis of polling data to forecast support for candidates Kamala Harris and Donald Trump in the upcoming U.S. presidential election. By employing a Bayesian linear regression model, the study investigates how various factors such as polling date, candidate identity, sample size, and geographic location influence voter support. The findings contribute to understanding the dynamics of electoral support and the effectiveness of polling methodologies in predicting election outcomes.

State	PCT of Trump	PCT of Harris
Arizona	48.85625	46.49375
California	34.52000	59.56000
Connecticut	37.00000	53.00000
Florida	51.29000	43.04000
Georgia	48.80625	46.87812
Indiana	57.90000	40.70000
Iowa	47.00000	43.00000
Maine	41.00000	53.75000
Maine CD-1	36.00000	60.50000
Maine CD-2	46.50000	47.50000
Maryland	31.86000	63.06000
Massachusetts	32.55000	61.55000
Michigan	46.70938	47.54000
Minnesota	43.50000	49.16667
Missouri	54.60000	42.20000
Montana	56.14286	39.50000
National	45.37277	47.35494
Nebraska	55.00000	38.50000
Nebraska CD-2	42.00000	50.42857
Nevada	47.56154	47.80769
New Hampshire	44.00000	51.01250
New Mexico	42.86667	51.53333
New York	39.92000	52.90000
North Carolina	47.63500	48.00250
Ohio	50.58889	44.63333
Pennsylvania	46.80577	48.34200
Rhode Island	40.33333	55.66667
South Carolina	51.60000	42.00000
Texas	50.38182	44.70000
Vermont	28.00000	68.50000
Virginia	42.90000	49.71250
Washington	35.00000	57.00000
Wisconsin	46.71818	48.86667

Figure 6: Summary table of average pct for each candidate by state

Candidate	Total Votes
Trump	109
Harris	222

Figure 7: Summary total votes for each candidate

5.2 Insights About the World

One significant insight derived from this analysis is the importance of temporal dynamics in voter support. The model highlights that the timing of polling can substantially impact the reported support for candidates. This finding suggests that voters' preferences are not static; rather, they fluctuate based on current events, campaign strategies, and media coverage. Such insights underscore the need for political strategists and candidates to consider timing when conducting polls and planning campaign efforts.

Another critical takeaway is the varying levels of support for candidates across different states. The analysis reveals regional disparities in voter sentiment, indicating that geographic factors significantly influence political preferences. This highlights the importance of localized campaign strategies and targeted outreach, as what resonates in one state may not hold true in another. Understanding these regional differences can enhance candidates' ability to connect with voters and tailor their messaging effectively.

5.3 Limitations of the Study

Despite its contributions, the study is not without weaknesses. One notable limitation is the reliance on existing polling data, which may introduce biases if the data collection methods are not standardized across pollsters. Variations in methodology, sample demographics, and question phrasing can lead to discrepancies in reported support levels. Additionally, the model assumes a linear relationship between predictors and outcomes, which may oversimplify the complexities of voter behavior. A more nuanced approach that incorporates non-linear effects or interactions between variables could provide deeper insights into voter dynamics.

5.4 Future Directions

Looking ahead, several avenues for further research emerge from this study. First, it would be valuable to explore the impact of specific events, such as debates or major news stories, on voter sentiment. Incorporating real-time data and sentiment analysis from social media could enrich the understanding of how public opinion shifts in response to external influences. Furthermore,

expanding the model to include demographic factors—such as age, income, and education—could yield a more comprehensive picture of voter support and the diverse motivations behind electoral choices.

In conclusion, while this paper offers valuable insights into electoral dynamics, it also highlights the complexities of voter behavior and the need for continued exploration in this field. Future research should strive to incorporate a broader range of variables and methodologies to enhance the predictive accuracy of electoral forecasts and better inform political strategy.

Appendix

A Methodology Analysis of Texas Politics Project Poll

A.1 Introduction

This section provides a examination of the methodology utilized by YouGov for the October 2024 Texas Statewide Survey (Blank (2024)), stating the reliability of the organization and its methods, along with an analysis of the survey's strengths, weaknesses, and potential improvements.

A.2 Population, Frame, and Sample

- **Population**: The survey aimed to represent Texas registered voters, defined as individuals who meet the eligibility requirements to vote in Texas, including age, residency, and registration status. This demographic is critical given the state's political landscape and upcoming elections.
- **Frame**: The sampling frame is the specific population from which a sample is drawn, constructed from diverse data sources to ensure validity. Key components included:
  - **American Community Survey (ACS)**: This provides demographic information that reflects the socio-economic characteristics of Texas residents, essential for establishing a representative sampling frame.
  - **Public voter registration records**: These ensure that the survey captures a valid sample of registered voters, allowing for accurate political representation.
  - **2020 Current Population Survey (CPS)** and **National Election Pool (NEP) exit polls**: These sources offer insights into voting behaviors and preferences, enriching the sampling frame with relevant context and historical data.
- **Sample Size**: The initial sample comprised 1,338 respondents, which was refined to 1,200 for the final dataset through a process known as matching. This process enhances the demographic representativeness of the sample across key characteristics such as gender, age, race, and education.

A.3 Sample Recruitment

- **Recruitment Methods**: YouGov employs a proprietary opt-in survey panel that includes approximately 1.5 million U.S. residents. Key recruitment methods include:

  - **Web Advertising Campaigns**: Targeted ads based on keyword searches increase the likelihood of engaging relevant respondents, leveraging digital platforms for broad reach.
  - **Permission-Based Email Campaigns**: This method allows for direct outreach to potential participants who have expressed interest in surveys, ensuring a more willing and engaged respondent pool.
  - **Telephone-to-Web and Mail-to-Web Recruitment**: These methods broaden the outreach, ensuring access to respondents who may not engage online, thus increasing the overall sample diversity.

- **Diversity and Representation**: The multifaceted recruitment strategy is designed to mitigate biases and enhance the demographic diversity of the panel. This is essential for capturing a representative sample in a politically diverse state like Texas.

A.4 Sampling Approach and Trade-offs

- **Two-Stage Sampling Approach**:

  1. **Target Sample Selection**: A probability sample is drawn from the defined target population, ensuring it accurately reflects the demographic composition of Texas. Probability sampling means that each individual in the population has a known, non-zero chance of being selected.
  2. **Matched Sample Selection**: Each target sample member is matched to one or more respondents from the YouGov panel, utilizing a set of variables available in consumer and voter databases. Matching involves finding respondents in the panel who share key demographic and behavioral traits with the target sample.

- **Strengths**: This methodology allows for the creation of a sample that closely mirrors the target population's characteristics, theoretically increasing the external validity of the survey results—external validity refers to the extent to which findings can be generalized to a broader context beyond the study sample.

- **Weaknesses**:

  - **Non-Random Selection**: The reliance on an opt-in panel introduces potential biases associated with self-selection. Respondents who choose to participate may differ significantly from those who do not, potentially affecting the survey's internal validity (the degree to which the survey accurately measures what it intends to measure).
  - **Matching Limitations**: While proximity matching aims to find the closest respondents, it may not account for all variables that could influence responses, leading to unobserved biases. Unobserved biases occur when factors influencing survey responses are not measured or controlled for in the analysis.

A.5 Handling Non-Response

- **Weighting Strategy**: To address non-response bias, YouGov implemented a robust weighting methodology:
  - **Propensity Score Adjustment**: This involves estimating the likelihood of response based on demographic characteristics, which are then used to weight the matched cases back to the target population. Propensity score weighting is a statistical technique that adjusts for differences in observed characteristics between respondents and non-respondents.
- **Challenges**: The success of this strategy hinges on accurate assumptions regarding non-response characteristics, which can be difficult to ascertain. If the reasons for non-response are systematically related to the survey topics or demographics, this could undermine the validity of the results, highlighting the importance of addressing potential non-response bias.

A.6 Questionnaire Evaluation

- **Question Design**: The YouGov questionnaire included a series of structured questions aimed at eliciting clear responses regarding voter preferences and opinions on key issues. Structured questions are those that offer fixed response options, facilitating easier analysis.
- **Pros**:
  - **Clarity and Relevance**: Questions were designed to be straightforward, addressing current political concerns likely to engage respondents effectively.
  - **Comprehensive Coverage**: The questionnaire covered a wide array of topics pertinent to the election, providing rich data for analysis and ensuring relevant insights into voter sentiment.
- **Cons**:
  - **Length and Complexity**: The extensive nature of the questionnaire could lead to respondent fatigue, increasing the likelihood of incomplete or rushed responses, which may affect data quality. Respondent fatigue occurs when participants become tired or disengaged from a survey, leading to lower-quality data.
  - **Framing and Bias**: The manner in which questions are framed can introduce bias. Questions that lack neutral wording may lead to skewed responses based on how they are interpreted by participants, potentially distorting the data collected.

A.7 Recommendations for Improvement

1. **Enhanced Recruitment Strategies**: Consider integrating more diverse recruitment channels, such as partnerships with community organizations, to reach underrepresented groups who may not be captured through standard methods.
2. **Improved Matching Techniques**: Exploring advanced statistical techniques or machine learning approaches for matching respondents could enhance the accuracy of the matched sample.
3. **Shortening the Questionnaire**: Streamlining the questionnaire to focus on the most critical issues could reduce respondent fatigue and improve engagement, ensuring higher quality data collection.
4. **Pilot Testing**: Implementing pilot surveys to test question clarity and structure could help identify potential biases or misunderstandings before full deployment, ensuring that the questions effectively capture the desired information.
5. **Transparency in Weighting Methods**: Clearly communicating the assumptions and methods used in weighting would enhance transparency and allow for a more critical evaluation of the findings, fostering trust in the results.

A.8 Conclusion

In summary, YouGov's methodology for the October 2024 Texas Statewide Survey demonstrates several strengths, particularly in its diverse recruitment strategies and thorough weighting processes. However, inherent challenges associated with opt-in panels, matching techniques, and questionnaire design warrant careful consideration. By addressing these weaknesses and implementing the suggested improvements, the reliability and validity of future surveys can be enhanced, ultimately providing richer insights into voter behavior and preferences.

B Idealized Methodology and Survey

B.1 Objective and Overview

This survey methodology aims to forecast the U.S. presidential election outcome by gathering representative data from a diverse cross-section of American voters. The target population includes all eligible U.S. voters, spanning a wide array of demographic backgrounds. With a budget of $100,000, this methodology employs probability sampling techniques, effective recruitment, and thorough data validation protocols to ensure data accuracy and minimize bias. By applying probability sampling, each individual has a known chance of selection,

allowing the results to be statistically valid and generalizable. This design accounts for demographic, geographic, and political influences, enhancing prediction reliability (**fivethirty?**; **pewresearch?**).

B.1.1 Core Objectives

- **Representative Sampling**: Achieve a sample reflecting U.S. voting demographics, ensuring insights are applicable to the entire electorate.
- **Data Quality**: Implement validation processes, including cross-verification, to ensure accuracy.
- **Statistical Modeling and Aggregation**: Apply statistical models and aggregate polling data to enhance forecast accuracy, as demonstrated in prior election studies.

B.2 Sampling Strategy

The sampling strategy combines **stratified random sampling** and **quota sampling** for comprehensive representation across key demographics. Stratified random sampling ensures each subgroup within the population is represented by dividing the population into strata (e.g., age, gender, race) and sampling randomly within each stratum. Quota sampling supplements this by guaranteeing minimum representation across essential demographics.

B.2.1 Stratification Variables

- **Age Groups**: 18-29, 30-44, 45-64, 65+
  - **Gender**: Male, Female, Non-binary/Other
- **Race/Ethnicity**: White, Black, Hispanic/Latino, Asian, Indigenous, Other
- **Education Level**: No high school, High school graduate, College graduate, Post-graduate
- **Income Bracket**: <$30,000, $30,000-$60,000, $60,000-$100,000, >$100,000
- **Geographic Region**: Northeast, Midwest, South, West

This structure aligns with survey sampling best practices, improving accuracy by ensuring proportional representation (**taherdoost2016sampling?**; **pewresearch?**). However, balancing stratification variables with cost considerations remains a challenge, as over-stratification can be cost-inefficient.

B.2.2 Sample Size

To achieve high accuracy, we will survey **10,000 respondents**, which provides a $\pm 1\%$ margin of error at a 95% confidence level. This sample size enables detailed subgroup analyses, such as state or demographic-specific trends, increasing the forecast's reliability.

B.2.3 Weighting

Post-stratification weighting will adjust for any underrepresented or oversampled demographic groups, ensuring the final sample accurately mirrors the voting population (**kalton2003weighting?**). For example, younger voters or minorities will be weighted according to their population proportions.

B.3 Recruitment Strategy

Multi-channel outreach (digital ads, email, and civic organization partnerships) will mitigate non-response bias by ensuring diverse respondent participation. This strategy helps correct response imbalances that could distort survey accuracy (**dillman2014internet?**).

- **Digital Advertisements**: Targeted ads on platforms like Facebook and Instagram to attract diverse demographic groups.
- **Email Outreach**: Invitations to registered voters via available databases.
- **Civic Partnerships**: Collaborations with non-profits to enhance diversity and reduce non-response bias.
- **Incentives**: Participants can enter a lottery for a $100 gift card as motivation.

B.4 Data Validation and Quality Assurance

Ensuring data integrity is essential for an accurate election forecast. Validation protocols will include:

- **Captcha Verification**: To prevent bot entries.
- **Contact Verification**: Confirm email or phone authenticity.
- **Response Time Monitoring**: Flag unusually quick completions for review.
- **Voter Registration Check**: If feasible, confirm registration to verify eligibility.
- **Audits**: Random follow-ups to verify responses.

These measures align with data integrity standards in electoral polling (**pew_survey_methodology?**).

B.5 Poll Aggregation and Data Analysis

B.5.1 Poll Aggregation

Using a poll-of-polls approach, this survey's data will combine with data from reputable polling firms (e.g., YouGov, Ipsos), creating a balanced forecast.

- **Weighting by Methodology and Recency**: Polls are weighted based on methodological rigor and recency, favoring recent data.
- **Bias and Variability Adjustments**: Aggregated data will adjust for any bias among individual polls.

B.5.2 Modeling Approach

Bayesian hierarchical models will address variability across states, demographics, and regions, enabling both popular vote and Electoral College predictions.

B.6 Budget Allocation

- **Recruitment and Outreach**: $70,000
- **Incentives**: $10,000
- **Survey Platform and Administration**: $5,000
- **Validation Tools**: $5,000
- **Analysis Software**: $10,000

B.7 Survey Implementation

The survey will be conducted through **Google Forms**, providing an accessible, cost-effective platform. Access the survey here: Google Form Survey.

B.7.1 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 duanyi.su@mail.utoronto.ca (Xinqi Yue, Yawen Tan, Duanyi Su).

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

Section 1: About Eligibility

Are you a U.S. citizen?

- Yes
- No (Please end Survey)

Do you meet your state's residency requirements?

- Yes
- No (Please end Survey)

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

- Yes
- No (Please end Survey)

Are you registered to vote by the voter registration deadline. (North Dakota does not require voter registration.)

- Yes
- No
- Wish to register later

Section 2: About Demographics

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

Which age group do you belong to?

- 18-29

- 30-44
- 45-64
- 65 or older

What is your gender? - Male

- Female
- Prefer not to say
- Other

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 live 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.7.2 Design Considerations

- **Question Wording**: Structured to be clear, neutral, and direct.
- **Pilot Testing**: A pre-deployment pilot will refine the question flow, ensuring clarity and ease for respondents.

C Additional data details

D Model details

D.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

D.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

References

1Keydata. 2024. "State Electoral Votes." https://state.1keydata.com/state-electoral-votes.php.

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

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.

Blank, B. J. H. 2024. "With Voting about to Start in Texas, Trump and Cruz Maintain Single-Digit Leads in New University of Texas/Texas Politics Project Poll." The Texas Politics Project. https://texaspolitics.utexas.edu/blog/voting-about-start-texas-trump-and-cruz-maintain-single-digit-leads-new-university-texastexas.

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/.

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.

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://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2024. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.