Forecasting the 2024 U.S. Presidential Election Using a Model-Based Approach*

Kamala Harris expects to win on 47.5%, leads Donald Trump by about 1% in 2024 US election

Diana Shen Jinyan Wei Jerry Yu

November 3, 2024

This study models the 2024 U.S. Presidential Election by analyzing polling data with adjustments for factors like recency and sample size, simulating an approach that captures up-to-date voter sentiment. Using recent 4 months aggregated poll data, the model identifies shifts in candidate support across key battleground states, providing a nuanced view of evolving preferences as election day approaches. By emphasizing recent, high-quality polls, the analysis improves predictive reliability, offering insights that could inform campaign strategies and public discourse. These findings highlight the value of weighting and time-sensitive adjustments in forecasting methods, balancing accuracy with responsiveness to the latest trends.

Table of contents

*Code and data are available at: https://github.com/DianaShen1224/Forecast-2024-US-election.

1 Introduction

Understanding voter sentiment is essential for both political campaigns and analysts, especially with the upcoming U.S. election on the horizon. Public opinion is highly dynamic and can change swiftly due to various influences, including media coverage, campaign tactics, and major events. This study aims to forecast the percentage of support for Kamala Harris, offering insights into the elements that shape voter preferences as the election nears. By examining data from multiple polling sources, we intend to pinpoint the key factors influencing support, such as the poll's end date, the polling organization, geographic location, and the poll's score.

Research by Gelman and King (1993) underscores that while polls can exhibit variability over a campaign period, they can provide reliable predictions when adjusted for temporal changes and fundamental factors. Additionally, Erikson and Wlezien (2008) highlight the importance of timing in polling, showing that data closer to election day tends to stabilize and therefore becomes more predictive. Our research seeks to fill a gap in existing literature that often fails to address the complexities of polling data, thereby enhancing our understanding of voter behavior within the electoral context.

The estimand of our analysis is the true percentage of voter support for Kamala Harris in the upcoming U.S. presidential election. The object of the estimation is forecasting the percentage of Kamala's vote share based on an aggregated polling data using a multiple linear regression model. Our goal is to track shifts in public opinion over time, offering a clearer pattern of voter's choice and potential election outcomes.

The main focus of our analysis is the percentage of support for Harris, which we will model using various predictor variables. We are particularly interested in how the end date, polling organization, state, and poll score affect voter support. Using a linear regression framework, we can quantify the relationships between these predictors and the support outcome, providing clarity on how each factor influences overall support for Harris. By estimating the coefficients for each predictor, we aim to drawn significant conclusions about their respective impacts on voter sentiment.

Our findings reveal a notable positive correlation between the end date and the percentage of support, indicating that as the election approaches, voter support tends to increase. We also observed considerable variability in support levels based on the polling organization and state, with certain pollsters consistently reporting higher support for Harris. The quality of the polls significantly affected results, with more reputable polls correlating with higher levels of support. These insights emphasize the necessity of considering both the timing of polls and the characteristics of different polling firms when analyzing public opinion.

This research is important because precise predictions of voter support are crucial for effective campaign strategies. By identifying the main factors influencing support for Harris, campaign teams can customize their outreach and messaging to resonate better with voters. Furthermore, recognizing the differences across various polling organizations and states can assist in resource allocation and strategic focus during the campaign. Given that elections can be decided by narrow margins, having trustworthy insights into voter preferences can substantially influence the final outcomes.

The structure of this paper is organized as follows: Section 2 provides details on the data sources and variables used in our analysis. Section 3 explains the modeling approach, including the assumptions and specifications of our linear regression framework. In Section 4, we present our findings, emphasizing the key predictors of Harris's support. Finally, Section 5 explores the implications of our results and suggests potential directions for future research. Section A provide external data detail, Section B provide model detail, Section C provide a exploration for pollster Siena College's methodology, and Section D provide a idealized methodology for polling the poll.

2 Data

2.1 Overview

We conduct our polling data analysis using the R programming language (Team 2023). Our dataset, obtained from FiveThirtyEight (FiveThirtyEight 2024),based on polling as of 2 November 2024, provides a detailed overview of public opinion in the lead-up to the election. Adhering to the guidelines presented in Rohan Alexander (2023), we explore various factors that influence voter support percentages, including the timing of the polls, the traits of polling organizations, and regional differences.

In this study, we utilized several R packages to enhance our data manipulation, modeling, and visualization capabilities. The tidyverse package offered a comprehensive set of tools for data wrangling and analysis, improving workflow efficiency (Wickham et al. 2019). The here package aided in managing file paths, allowing for easy access to our data files (Müller 2020). We relied on janitor to perform data cleaning, as it provides functionalities to identify and rectify quality issues within the dataset (Firke 2023). For handling date-related operations, the lubridate package proved invaluable, simplifying the manipulation of time variables (Grolemund and Wickham 2011). Arrow supported efficient data input and output in a performance-oriented format, essential for managing larger datasets (Richardson et al. 2024). We use ggplot2 (**citeggplot2?**) to visualize the analysis of data. In addition to the core packages, we employed several specialized libraries to facilitate specific analytical tasks. The modelsummary package enabled us to generate clean, interpretable model summaries, streamlining the reporting of regression and other statistical results (Arel-Bundock 2022). RColorBrewer provided color palettes for visualizations, enhancing clarity and visual appeal in our plots (Neuwirth 2022). The future package supported parallel processing, optimizing the runtime of intensive computations, especially beneficial in model training (Bengtsson 2021). Finally,for resampling and bootstrapping techniques, we utilized the boot package, which allowed us to assess model stability and reliability (Angelo Canty and B. D. Ripley 2024; A. C. Davison and D. V. Hinkley 1997). caret facilitated model training and validation, offering a unified interface for cross-validation and performance metrics (Kuhn and Max 2008). Our coding practices and file organization were informed by the structure outlined in Rohan Alexander (2023).

2.2 Measurement

The process of translating real-world events into our dataset requires a systematic approach to measurement and data gathering. In this research, we aim to assess public opinion regarding Kamala Harris as the upcoming U.S. presidential election approaches. Polling agencies formulate surveys featuring specific questions designed to capture voters' attitudes, including their likelihood of supporting Harris and their views on prevailing political issues.

Once the survey items are established, a representative sample is selected through stratified random sampling methods, ensuring a diverse demographic representation. Respondents are reached using various techniques, such as telephone interviews and online questionnaires.

After gathering the responses, the data is subjected to thorough cleaning and validation procedures to rectify inconsistencies and handle any missing information. This step is crucial for ensuring that the dataset accurately mirrors the electorate's sentiments. Following data cleaning, polling results are aggregated to smooth out individual poll biases and mitigate random error. This aggregation process leverages weighting based on poll quality, sample size, recency, and the historical accuracy of each polling agency, as seen in FiveThirtyEight's approach (Silver, 2018; Jackman, 2005). By combining multiple polls, the aggregation method produces a more reliable picture of public opinion trends over time.

Each entry in the finalized dataset reflects an individual's viewpoint at a given moment, enabling a detailed analysis of the factors that shape public opinion as the election nears. This structured methodology effectively converts subjective opinions into measurable data, providing valuable insights into voter behavior and preferences.

2.3 Variables

2.3.1 End Date

The end date is the final day of data collection for a poll, indicating when the survey period concluded. This date provides crucial context for the poll results, as public opinion may change over time in response to events, campaign actions, or other influencing factors. In our analysis, we only choose the polls with end date after July 21 2024, which Kamala Harris first declared in the election.

2.3.2 Recency

Recency is a newly created predictor variable in our dataset, specifically for use in our model. It is calculated as follows Recency = Election Date - End Date, where Election Date is set to November 5th. This variable quantifies the time interval between the poll's end date and the election date as a measure of the poll's relevance and potential impact on voter behavior. In general, polls conducted closer to the election date provide a more accurate picture of public opinion. The original dataset includes only dates, and by mutating this variable, we can more intuitively interpret the number of days before the election that each data point was collected. Converting this to discrete data enhances its usefulness in constructing predictive models. Figure 1 reveals a strong focus on recent polling data, with the highest frequency of polls occurring within the first few days and gradually declining as recency increases. Periodic peaks around 30, 60, and 90 days suggest some regularity in polling schedules, though the

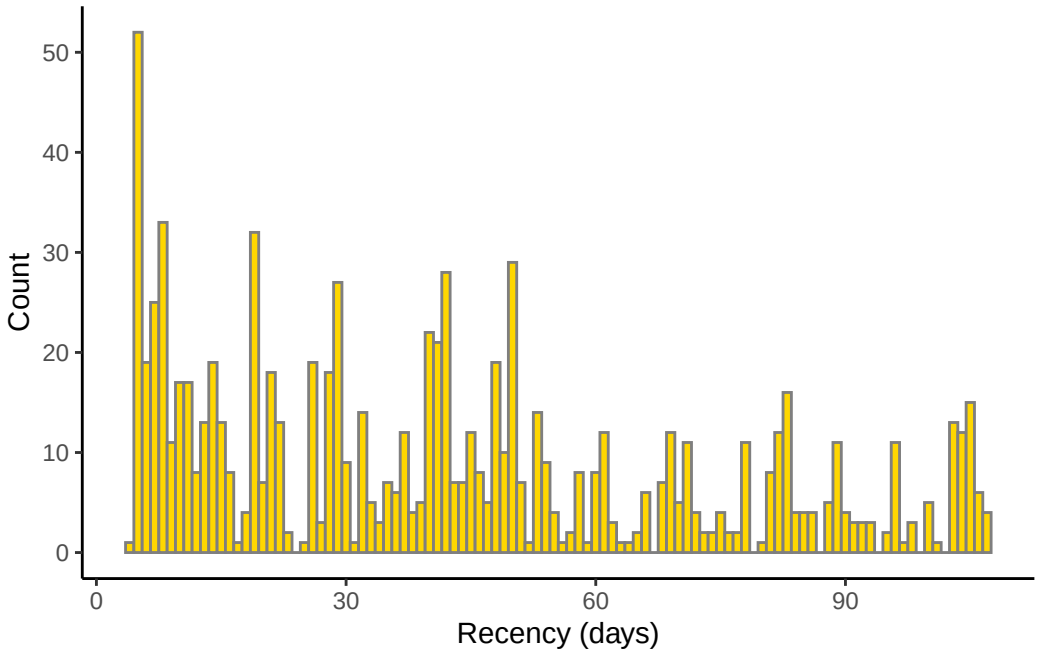


Figure 1: Distribution of the recency of the polls after July 21 2024

overall trend emphasizes newer data. This distribution highlights the prioritization of recent polls in the dataset.

2.4 Outcome variable

2.4.1 The support percentage of Kamala Harris.

Figure 2 illustrates the distribution of percentage support for Kamala Harris based on polling data, where support reflects the proportion of respondents favoring Harris in each survey. Most polls report support clustered around 50%, with the majority of values falling between 40% and 55%. This central peak suggests moderate, consistent support levels among respondents, with fewer instances of higher support levels above 55%. The right-skew in the distribution indicates occasional polls with elevated support, though these are less common. Overall, this visualization highlights the general sentiment and variability in support for Harris as captured across multiple polls.

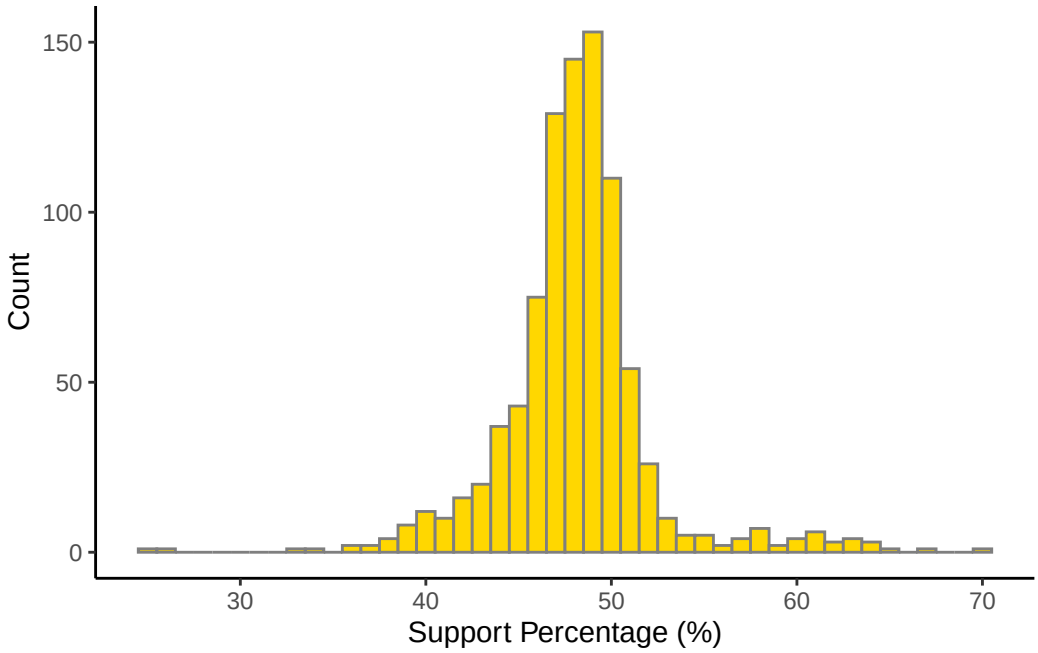


Figure 2: Distribution of support percentage of Kamala Harris

Table 1: Summary Statistics of categorical variable population and national_pull

		N	%
population	lv	620	68.3
	rv	288	31.7
national_poll	0	681	75.0
	1	227	25.0

2.5 Predictor variables

2.5.1 National Poll

National polls is another newly created variable in our dataset, represented by a binary indicator (1 for polls at national level, 0 otherwise). This distinction is important because not all polls are conducted nationally; some are state-specific and may reflect the region's preference for a particular presidential candidate. Ignoring this distinction may introduce significant bias into our model. By labeling each poll as national or non-national, we can give higher weight to national polls, which are more representative of overall public opinion, and thus improve the accuracy of our predictive model. Table 1 shows that for the national_poll variable, non-national polls (coded as 0) account for 681 cases (75.0%), whereas national polls (coded as 1) represent 227 cases (25.0%). This distribution highlights a focus on likely voters and a predominance of non-national polling data in the sample.

2.5.2 Population

The "population" variable in the dataset represents the status of each entry. it is the abbreviated description of the respondent group, categorizing participants as "lv" (likely voters) or "rv" (registered voters). They represent different levels of participation and potential voting behavior. Likely voters are individuals who are not only registered but are statistically more likely to vote based on historical data or survey responses. In contrast, registered voters include all individuals who are eligible to vote, regardless of their likelihood of participation. Table 1 shows that likely voters (lv) constitute the majority with 620 cases (68.3%), while registered voters (rv) make up 288 cases (31.7%).

2.5.3 State

Figure 3 displays the distribution of polls by state, showing how frequently polling organizations conducted surveys across various U.S. states and at the national level. The "National" category has the highest number of polls, indicating a strong emphasis on capturing overall U.S. sentiment. Certain states, such as Pennsylvania, Wisconsin, North Carolina, Arizona, Georgia, and Michigan, also show higher polling frequencies, likely because these are battleground states with the potential to influence the election outcome significantly. Toward the right side of the chart, states with minimal polling activity, including South Carolina, Iowa, and Washington, appear less frequently, possibly due to their historically predictable or less competitive nature. This distribution reflects the strategic focus of polling efforts, with organizations prioritizing both national sentiment and swing states where public opinion is more volatile. Overall, the chart provides insight into where polling resources are allocated as election day nears, emphasizing areas that could sway the final result.

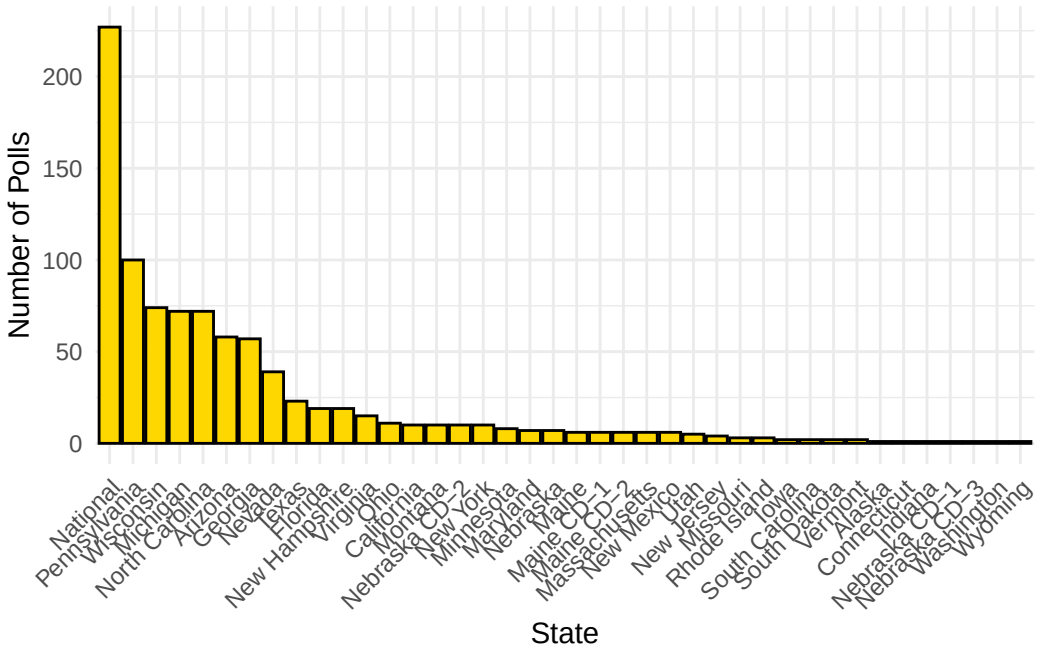


Figure 3: Count of polls by the US state

2.5.4 Pollster

Figure 4 displays the frequency of polls conducted by different pollsters. Each bar represents a pollster, with the height indicating the number of polls they conducted. Siena/NYT has the highest count, followed by YouGov and Emerson. Pollsters to the right have conducted significantly fewer polls, with some showing only one or two entries.

The pollster is the organization or firm that conducts the surveys, gathering and analyzing public opinion data on voter preferences. In this context, each pollster's count reflects its level of polling activity related to the election.

2.5.5 Numeric Grade

Figure 5 displays the distribution of numeric grades assigned to various pollsters, with the x-axis representing the numeric grade values and the y-axis indicating the frequency of each grade. The numeric grade is a metric that evaluates the quality or reliability of a pollster, taking into account factors such as methodology, historical accuracy, sample size, and pollster reputation. In this chart, we observe that the most common numeric grades are concentrated around 2.75 and 3.00, with a large spike at these values, suggesting that a significant number of polls are conducted by highly-rated pollsters. Fewer polls have grades below 2.5, indicating a relatively lower occurrence of polls from less-reliable pollsters in this dataset. This distribution

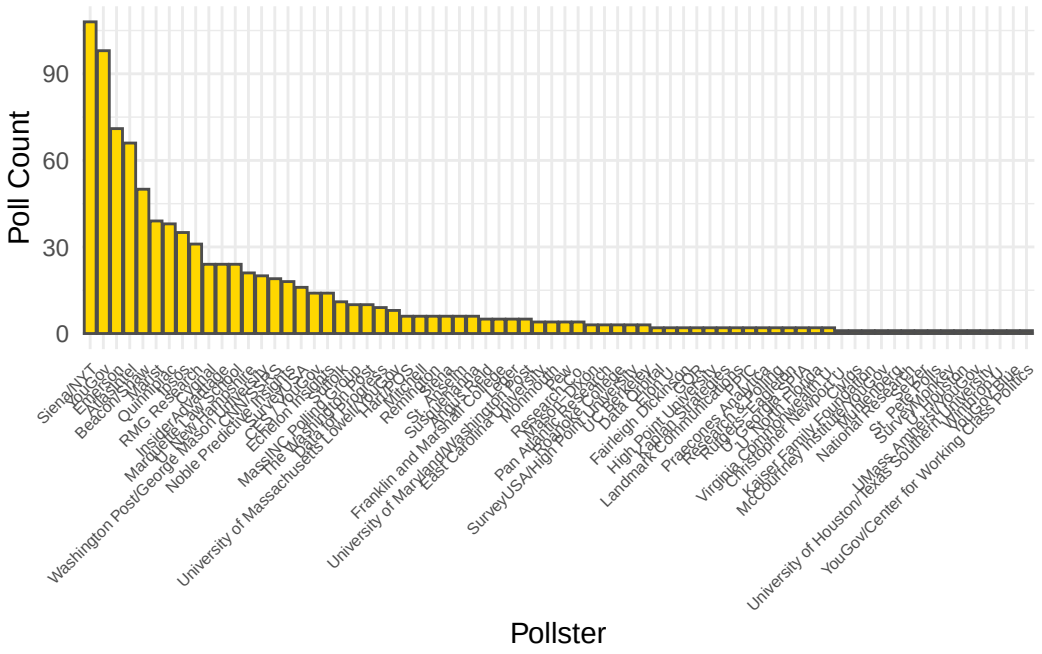


Figure 4: Frequency of Polls by Pollster

underscores the emphasis on high-quality pollsters within the dataset, ensuring a more reliable and consistent source of polling data for subsequent analysis.

2.6 Relationships between key varaibles

Figure 6 shows the trend of percentage support for Kamala Harris over time, with data points and smoothed trend lines for each pollster. The graph shows the trend of the percentage of respondents supporting Harris from August through October. Each color corresponds to a specific polling organization, with prominent pollsters such as Beacon/Shaw, Ipsos, Siena/NYT, Emerson, Quinnipiac, and YouGov. The smoothed lines reveal subtle trends over time, with some pollsters like Emerson and Beacon/Shaw showing a slight upward trend, while others like Siena/NYT display a downward trend. Figure 6 highlights the variability in poll results across different organizations, reflecting each pollster's methodology and sample.

Figure 7 illustrates Harris's support percentages over time, with each data point representing poll results colored by the numeric grade of the pollster. The smoothing lines for each numeric grade remain subtle, indicating minor variations in support trends over time across different pollster quality levels. Most points are clustered around the 50% support level, suggesting a generally stable voter sentiment for Harris, with only slight fluctuations across numeric grades. By limiting the maximum value of support at 80%, outliers and extreme variations are minimized, resulting in a clearer and more interpretable visualization. This approach

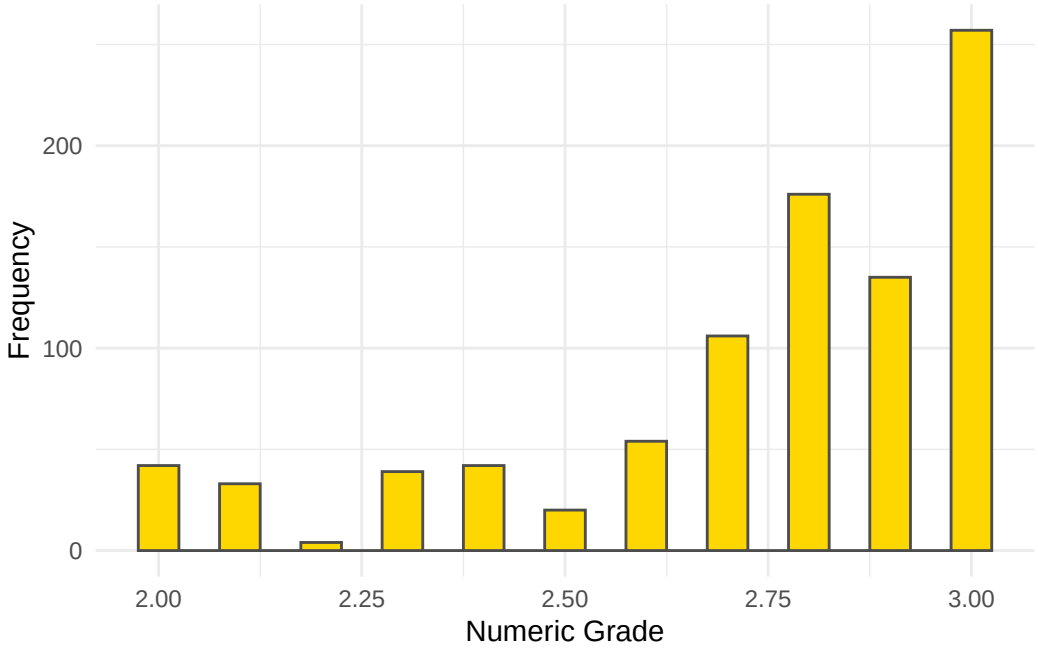


Figure 5: Distribution of numeric grade

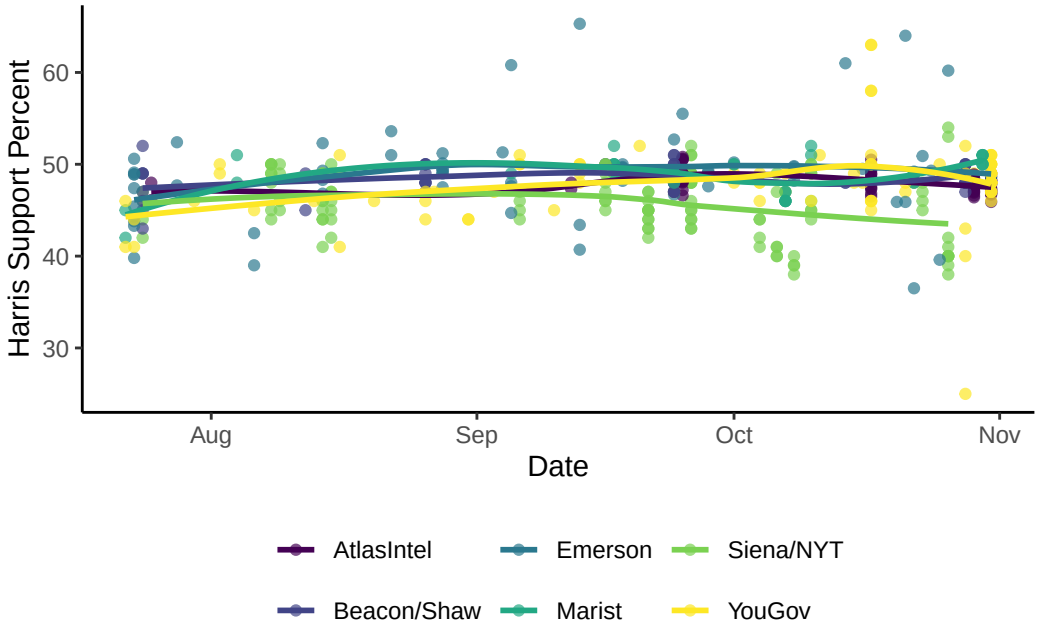


Figure 6: Harris Support Over Time by Pollster (Top 6 Pollsters)

emphasizes the central trend, allowing for clearer comparisons of support levels across pollsters of varying reliability without distraction from extreme values.

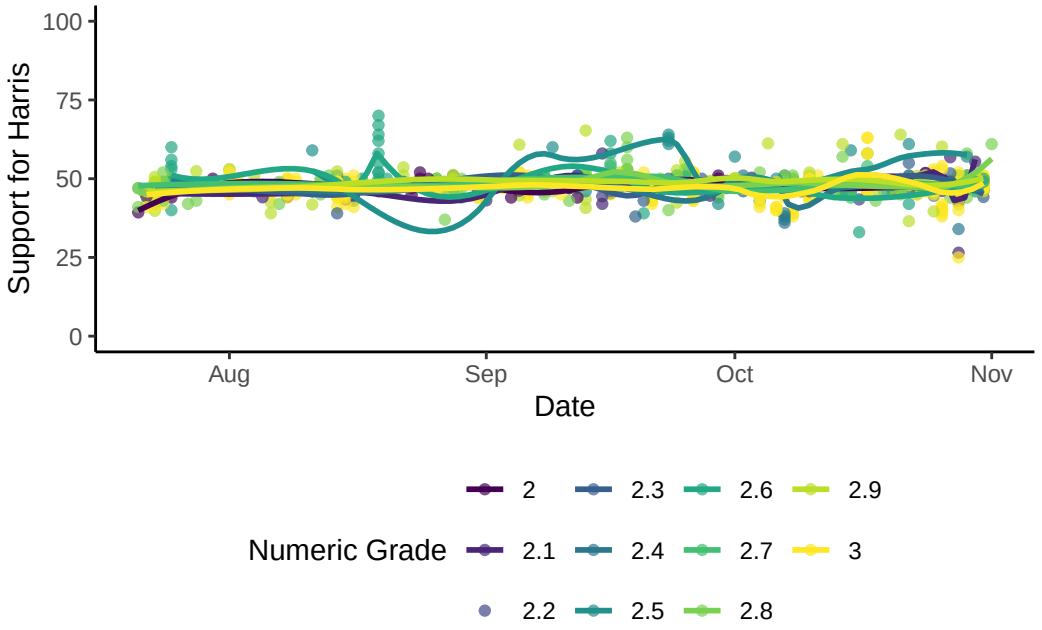


Figure 7: Harris Support Over Time by Numeric Grade

3 Model

Our modeling approach aims to quantify the relationship between various polling metrics and the percentage of support for each candidate, Kamala Harris and Donald Trump. For this analysis, we employ a linear model (LM) to examine how factors such as national polling trends, poll quality scores, and population size influence support percentages. The model is implemented using the `lm` function in R, with a Gaussian distribution to capture the variability in support rates.

In this analysis, we use predictors that capture both the methodological quality and structural aspects of each poll. Specifically, we include variables such as `national_poll`, which reflects national support trends; `pollster`, which represents differences in methodology and reliability across polling organizations; and `population`, which accounts for demographic and regional characteristics associated with each poll. Our combined weighting approach further integrates factors like recency and sample size to ensure that more recent, larger-sample polls have a greater impact on the model's predictions, providing a balanced and comprehensive assessment of candidate support.

The model assumes that the distribution of support percentage, given these polling characteristics, follows a normal distribution. This Gaussian assumption enables effective parameter estimation, a standard approach in linear regression. By applying moderate priors, we balance interpretability with the model's stability, allowing us to assess the impact of polling characteristics on candidate support with greater reliability.

3.1 Model Set-Up

The model predicts Harris's support percentage by constructing multiple linear regression model using the following predictor variables:

- **National Poll (`national_poll`)**: Represents the national trend in support for the candidate.
- **Pollster (`pollster`)**: Categorical variable identifying the organization conducting the poll, accounting for differences in methodology and reliability.
- **Population (`population`)**: Variable provides an abbreviated description of the respondent group, usually indicating their voting status (e.g., "lv" for likely voters or "rv" for registered voters).
- **State (`state`)**: The U.S. state where the poll was conducted or focused, if applicable.

We include Pollster and State variables as controls to account for potential biases introduced by different polling organizations and regional variations in voter preferences.

3.1.1 Unweighted Model

The unweighted model for Harris provides a baseline by treating all polls equally without adjustments for recency, sample size, or poll quality: The model takes the form:

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

$$\mu_i = \beta_0 + \beta_1 \cdot \text{National Poll}_i + \beta_2 \cdot \text{Pollster}_i + \beta_3 \cdot \text{Population}_i + \beta_4 \cdot \text{State}_i + \epsilon_i$$

$$\epsilon_i \sim \text{Normal}(0, \sigma^2)$$

Where:

- β_0 is the intercept term, representing the baseline level of support.
- $\beta_1, \beta_2, \beta_3, \beta_4$ are the coefficients for each predictor, indicating their influence on Harris's support percentage.
- σ^2 is the variance of the error term, representing unobserved variability in support.

The model is executed in R, using the `lm` function for a linear regression model. We apply weights based on `combined_weight` to account for the importance of each poll, incorporating factors such as recency and sample size. This approach helps ensure that more recent and reliable polls have a greater influence on the model's predictions.

3.1.2 Weighted Model Explanation

Our model uses a weighted approach to estimate candidate support more accurately by emphasizing higher-quality and more recent polls. The weighting scheme integrates factors like poll recency, sample size, pollster quality, and frequency, as inspired by the New York Times' methodology (The New York Times 2024a) for polling averages.

In this model, weights are applied directly to each poll's contribution in the **Weighted Least Squares (WLS)** estimation process, modifying the influence of each poll on the outcome. This results in the following expression for the estimated coefficients ($\hat{\beta}$):

$$\hat{\beta} = (X^T W X)^{-1} X^T W y$$

Where: W is the weight parameters calculated from the product of **sample size weight** and recency **weight**.

- **Sample Size Weight**: Calculated as the sample size divided by 2300, capped at 1. This emphasizes larger, more reliable polls while preventing extremely large samples from disproportionately impacting the analysis.
- **Recency Weight**: Applied through an exponential decay function, exp(-recency * 0.1), where "recency" represents the days since the poll was conducted. This ensures that more recent data carries more weight, while older polls have reduced influence.
- **Combined Weight**: The final weight is the product of all individual weights:

$$\text{Combined Weight} = \text{Recency Weight} \times \text{Sample Size Weight}$$

.

By incorporating these weights, we adjust the model's estimates to give more importance to recent, credible, and representative polls, enhancing the reliability of our predictions.

3.2 Model Justification

Existing research and political science theories suggest that factors such as sample size, recency, pollster reliability, and local demographics can significantly impact support percentages for candidates like Harris and Trump. Larger sample sizes are generally more reliable, while recent polls capture the latest shifts in public opinion. The methodology and timing of each poll also play a role: for instance, online surveys and telephone interviews may capture different segments of the population, and polling conducted during or near key political events often reflects more immediate public sentiment. Additionally, regional factors, such as local political dynamics or demographic characteristics, can influence support patterns.

Our model incorporates a combined weighting scheme inspired by the methodology outlined by the New York Times in their election polling averages. This weighting system integrates factors like recency and sample size, ensuring that more recent and larger-sample polls exert a greater influence on our model's predictions. By applying this weighting approach, we aim to achieve a balanced and representative dataset that reflects up-to-date trends while mitigating potential biases from older or smaller-sample polls.

A linear regression model was chosen to predict support percentages because the dependent variable is continuous and tends to approximate a normal distribution. Linear regression is an accessible and interpretable method for assessing how multiple factors contribute to an outcome, with each coefficient offering a straightforward interpretation of predictor influence. This model framework allows us to quantify the effects of various polling characteristics on candidate support.

Further justification for using this model stems from its alignment with the central limit theorem, which indicates that, given a sufficiently large sample, the distribution of support percentages should approximate normality. Moreover, the relationships between predictors like sample size and recency align with established theories in political behavior, providing a theoretical basis for the model. The linear regression approach also reduces the risk of overfitting, enhancing the generalizability of our results to a wider set of polling data.

Finally, to ensure the robustness and accuracy of our model, we conduct model validation and diagnostics, which are detailed in Appendix B. We use K-Fold Cross-Validation to validate the model, use Q-Q Plot and Residuals vs fitted diagram to diagnostic the model. These procedures help confirm that the model assumptions are met and that the predictions are reliable across different polling scenarios.

4 Result

This section examines the relationship between national polling trends, pollster effects, population(respondent group), and other factors with respect to the predicted support levels for Kamala Harris and Donald Trump in the 2024 U.S. Presidential Election. Using a dataset

comprising polling data from various states and organizations, we applied both unweighted and weighted linear regression models to identify the key predictors influencing each candidate's support. Below, we present the results of our models along with their implications.

Predictions were generated based on our test dataset. However, due to incomplete data for certain states, we were unable to make predictions across the entire U.S. This limitation arises from states with missing values in the dataset, which are consequently not represented in the model, affecting forecast accuracy in those regions.

4.1 Model Results

The linear regression models constructed for predicting support for Kamala Harris and Donald Trump in the 2024 U.S. Presidential Election utilized both unweighted and weighted configurations.The model summary of support for Harris is shown in Table 2, support for Trump is shown in Table 3. As shown in Table 2 and Table 3, the intercepts for Harris and Trump, estimated at 44.70% and 59.28% respectively in the unweighted models, while in weighted models, it's estimated at 45.2% and 54.8%, represent their baseline support levels when all predictors are at reference values. The high (R^2) values of 0.824 for Harris's unweighted model and 0.892 for her weighted model, along with similar values of 0.849 and 0.916 for Trump, indicate that a significant portion of the variation in support can be explained by the model's predictors. The relatively low Root Mean Square Error (RMSE) values further suggest that the models provide reasonably accurate predictions of support levels.

The results underscore the significance of **national_poll** as key predictors for both candidates' support, highlighting how national trends and pollster differences impact the reported support levels, as polls from the National level increase the support of Harris by 3.47%/4.27% , decrease the support for Trump by 7.61%/7.17%. The **population** predictor indicates that the type of respondent group contributes meaningfully to variations in support. With "likely voters" as the baseline level, the coefficient for "registered voters" (populationrv) suggests a lower level of support for both Harris and Trump in this group compared to likely voters. Specifically, the support for Harris of registered voters is 1.372 (unweighted model) or 0.529 (weighted model) lower than the likely voters. This implies that, on average, support among registered voters is lower than that among likely voters.

4.2 Predicted Voting Outcomes

Figure 8 illustrate the predicted support for Kamala Harris and Donald Trump from August 1, 2024, to November 2, 2024, using both unweighted and weighted linear regression models. In the Figure 8a unweighted model, we see that Trump's predicted support generally hovers between 40% and 50%, while Harris's support ranges from 45% to 50%. Both candidates' support levels appear relatively stable across the polling period, with minor fluctuations. This

Table 2: Summary of Multiple Linear Regression Model Predicting Harris Support Based on Polling Data

	Unweighted	Weighted
(Intercept)	44.70	45.2
	(2.07)	(19.1)
national_poll1	3.47	4.27
	(1.89)	(19.08)
populationrv	−1.372	−0.529
	(0.164)	(0.123)
Num.Obs.	908	908
R2	0.824	0.892
R2 Adj.	0.799	0.877
AIC	3789.6	5160.0
BIC	4333.3	5703.7
Log.Lik.	−1781.791	−2467.010
RMSE	1.72	1.89

Note. All models include a control for state and pollsters. The reference level of state is Alaska, the reference level of population is likely voter, and the reference level of pollsters is Alaska is Angus Reid.

Table 3: Summary of Multiple Linear Regression Model Predicting Trump Support Based on Polling Data

	Unweighted	Weighted
(Intercept)	51.87	54.8
	(2.02)	(18.0)
national_poll1	−7.61	−7.17
	(1.85)	(17.97)
populationrv	−1.060	−0.714
	(0.158)	(0.116)
Num.Obs.	930	930
R2	0.849	0.916
R2 Adj.	0.828	0.904
AIC	3833.7	5306.6
BIC	4380.1	5853.0
Log.Lik.	−1803.872	−2540.308
RMSE	1.68	1.90

Note. All models include a control for state and pollsters. The reference level of state is Alaska,the reference level of population is likely voter, and the reference level of pollsters is Alaska is ABC/Washington Post.

model, which treats all polls equally, suggests a consistent lead for Harris, albeit within a fairly close range.

The weighted model Figure 8b refines these predictions by applying weights based on recency and sample size, which reduces the influence of older or smaller polls. This approach results in smoother trend lines for both candidates, with less apparent variability compared to the unweighted model Figure 8a. Trump's support remains around the 40% to 50% range, while Harris's support stays in about 45% to 52% range. The weighting emphasizes more recent and larger polls, likely providing a more accurate reflection of current public opinion. Overall, both models consistently indicate a lead for Harris, with the weighted model slightly stabilizing the predictions. The average support for Harris and Trump are both consist for two models, which predicts that Harris will win on about 47%, leads Trump by about 1%.

4.3 Predicted Voting Outcomes for Recent Three Months by State

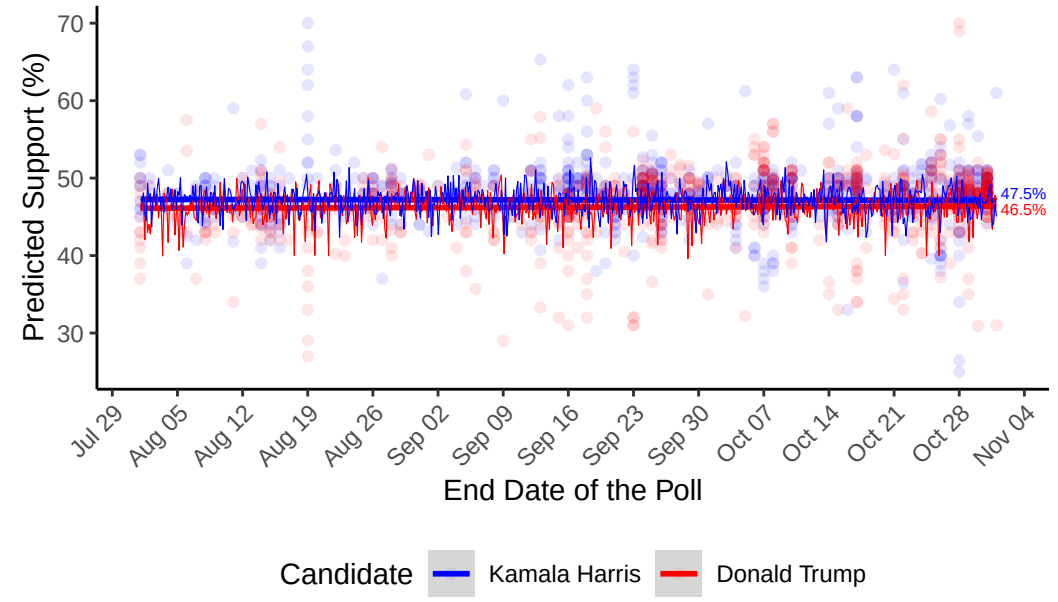
Figure 9 compare the predicted support for Kamala Harris and Donald Trump in key battleground states (Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin) for the 2024 U.S. election. Figure 9a uses an unweighted multiple linear regression model, while Figure 9b applies a weighted model, giving greater influence to polls with higher weights based on recency and sample size. In both figures, Trump's support, represented by red lines, and Harris's support, shown in blue, show close competition across most states, with slight variations in trends across the polling period from August to November 2024.

In the Figure 9a , support for Kamala Harris and Donald Trump fluctuates widely across battleground states, with Harris holding significant leads in Michigan, and Pennsylvania. Harris also hold advantage in Wisconsin but facing tight competition. Overall, without weighting, polls contribute equally, creating more volatility.

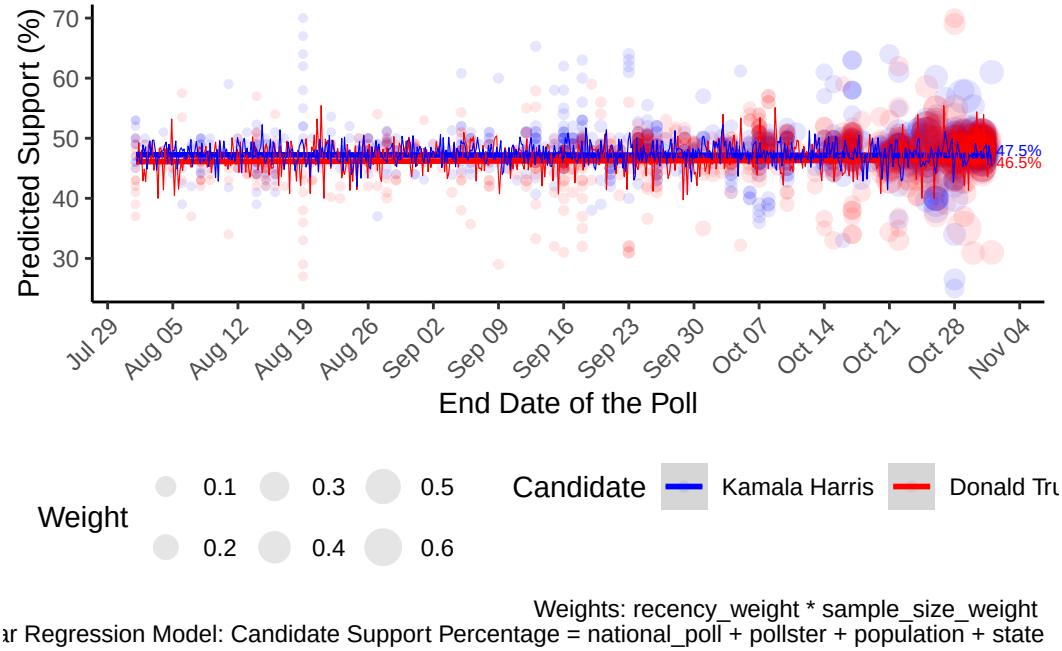
The Figure 9b stabilizes these trends by emphasizing recent, larger polls. Harris's support becomes more consistent, particularly in Pennsylvania, Michigan, and Wisconsin, where she shows a clearer lead. This weighted approach provides a more refined snapshot, highlighting Harris's likely advantage in these key states.

4.4 Predicted Support for Harris and Trump by Major Pollsters

Figure 10 depict the predicted support for Kamala Harris and Donald Trump in the 2024 U.S. election, from August 1, 2024, to November 2, 2024, using the unweighted and weighted multiple linear regression model. The red line shows average fluctuations in Trump's support, generally between 46% and 48%, with some noticeable peaks and dips, indicating variability over time. In contrast,the average Harris's support, represented by the blue line, remains more stable, consistently ranging around 47% to 48%, suggesting less volatility in voter sentiment for her.

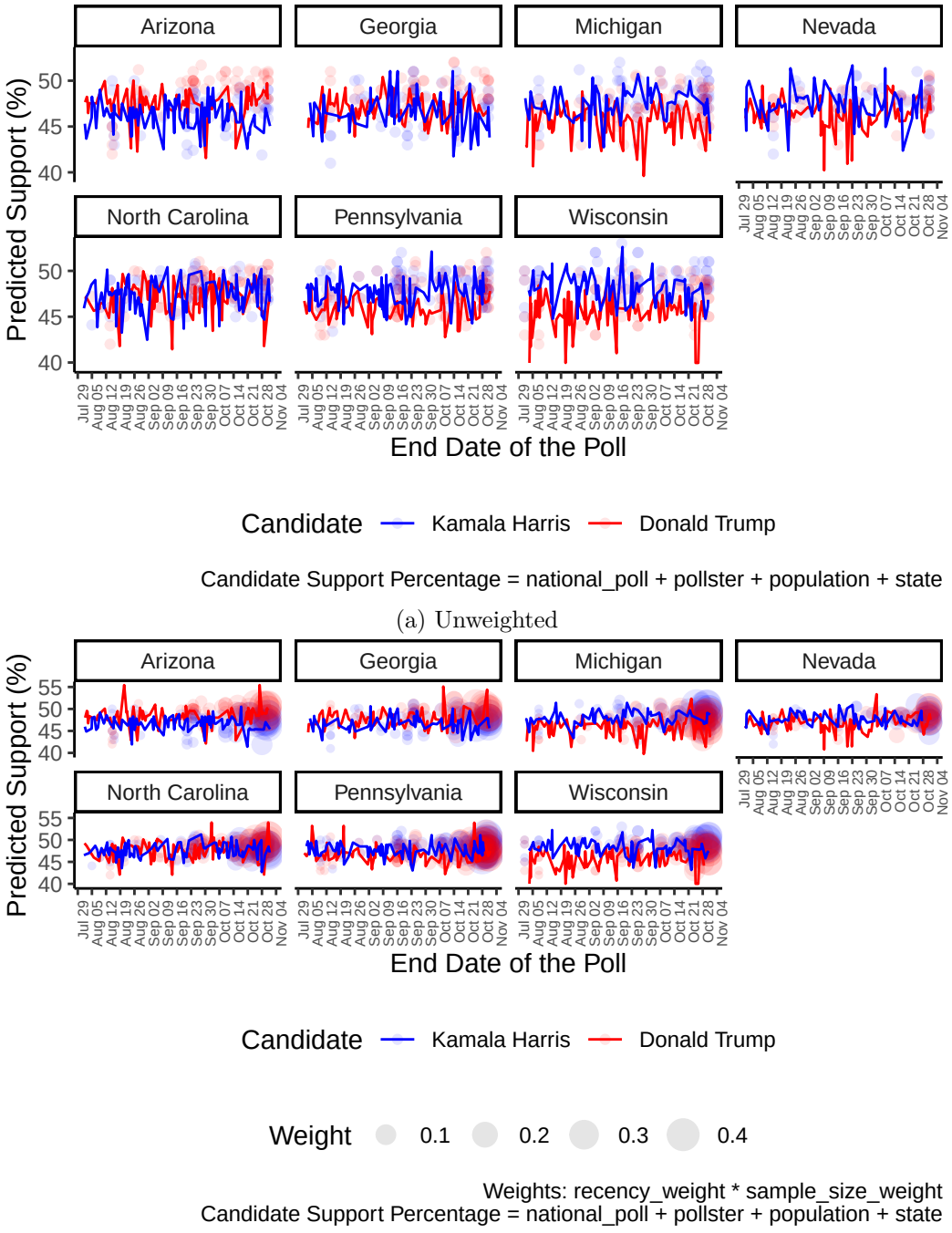


(a) Unweighted



(b) Weighted

Figure 8: Harris Kamala leads Donald Trump an average 2% in Predicted Support in the 2024 US Election (From August 1, 2024, to November 2, 2024)



(a) Unweighted

(b) Weighted

Figure 9: Harris is likely to lead Trump in Pennsylvania, Michigan, and Wisconsin by Model Prediction

Major Pollsters are represented by different colors, with circle sizes reflecting weights based on factors like recency and sample size. In the weighted model, larger circles correspond to higher-weighted polls, indicating they had a greater influence on the predictions. The model incorporates national poll data, adjustments for pollster effects, and population considerations, allowing for a refined prediction approach. This approach ensures that more recent and larger polls play a more significant role, capturing shifts in voter sentiment as the election approaches.

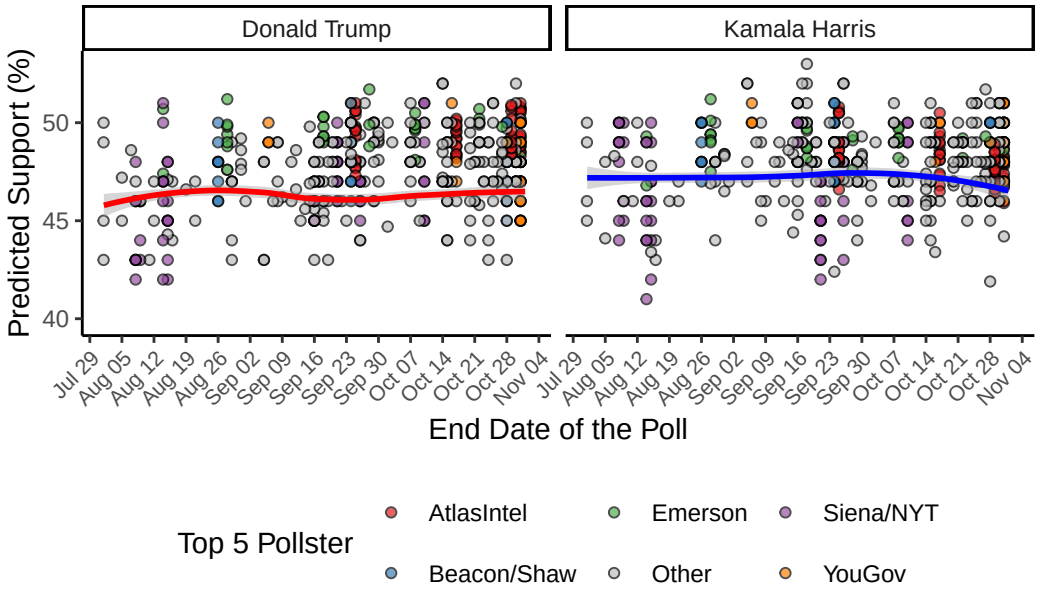
Siena/NYT and Quinnipiac consistently show stronger support for Harris, with Siena/NYT polls placing her around or above 50%. In contrast, Emerson polls display a closer race, often showing near-equal support for both candidates. AtlasIntel and Beacon/Shaw lean toward Trump, with AtlasIntel showing more volatility and Beacon/Shaw stabilizing near 45-50% for Trump. Polls from other organizations, in grey, introduce variability but hold less influence in the model. Overall, Siena/NYT and Quinnipiac favor Harris, while Emerson, AtlasIntel, and Beacon/Shaw present a more competitive picture.

5 Discussion

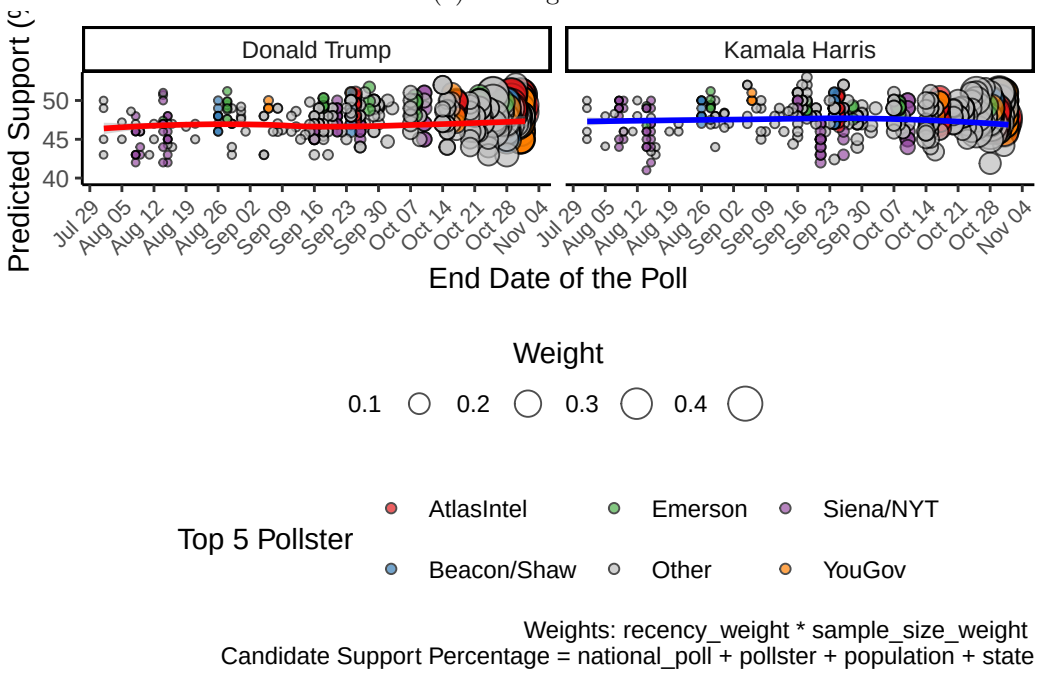
5.1 Key Findings and Real-World Implications

Our model reveals a portrait of a nation deeply divided, with Harris holding a narrow 49% edge over Trump's 47% in national support. However, as recent elections have shown, the national popular vote does not always translate into an Electoral College victory. The U.S. electoral system, which awards each state's electoral votes to the candidate with the majority of votes in that state, means that winning the popular vote nationally might still leave Harris short of the presidency. This structural quirk played a defining role in the 2016 election, where Hillary Clinton's popular vote win failed to deliver the electoral majority, paving Trump's path to the White House. Our findings underscore this enduring tension: while Harris may have a slight national advantage, the true battle will be fought state by state, in a handful of battlegrounds that could flip the election either way.

In states like Nevada and North Carolina, our model highlights just how close the race remains. North Carolina, for instance, shows both candidates locked in a near tie around 47% support—a statistical dead heat that brings the state's significance into sharp relief. Neveda, meanwhile, leans modestly toward Trump, a reflection of the unique demographic and political nuances shaping each battleground. These state-level insights illuminate a critical truth: while national polls offer a snapshot of overall sentiment, they risk obscuring the specific, local dynamics that will ultimately decide the outcome. The stakes are high; these are the states that could swing, the margins that will be watched closely on election night, as even slight changes in turnout or last-minute shifts in opinion could tip the balance. improve the number based on the graph results



(a) Unweighted



(b) Weighted

Figure 10: Major Pollsters Siena/NYT and Quinnipiac favor Harris, while Emerson, AtlasIntel, and Beacon/Shaw present a more competitive picture

Ultimately, these findings highlight the intricate dance between popular sentiment and the electoral mechanics of American democracy—a system where every vote counts, but some states matter just a bit more than others.

5.2 Influence of Weighting Methodology in Analyzing Voters' Preference

In our analysis and shown result, we find out that the use of a weighted model helps to reduce volatility in predicted support trends by assigning greater importance to polls that are both recent and based on larger sample sizes. This approach effectively smooths out abrupt fluctuations that might otherwise result from older or smaller polls, providing a more stable and coherent view of voter sentiment. hough our data filtered with recent polls after Harris's candidacy announcement, voter sentiment remains fluid, with ongoing debates and events likely to influence preferences. By emphasizing recent polls, the model is more responsive to the latest shifts in public opinion, which is especially relevant as election day approaches and voters' preferences may crystallize. Larger-sample polls, often conducted by established pollsters like Siena/NYT or Emerson, are assumed to be more reliable, and their increased weight in the model further reduces the likelihood of erratic changes driven by smaller, less consistent polls. This decreased volatility allows the model to present a clearer picture of the general support trend for each candidate, reducing noise and highlighting true shifts in sentiment. However, this smoothing effect may also mask some of the localized or short-term dynamics captured by less frequent or smaller-sample polls, which could provide unique insights into specific voter groups or state-level shifts.

5.3 Limitations of the Dataset and Model

Yet, as with all models, limitations persist. While our weighting system, with its emphasis on recent data, aims to capture a timely snapshot of voter sentiment, it may miss sudden changes sparked by campaign events or unexpected shifts in public discourse. The variability in our state projections—ranging from a low estimate of 363 electoral votes for Harris to a high of 471—reflects the inherent uncertainty in polling-based models, particularly given the constraints of our data sources. This wide range suggests that while data-driven insights can illuminate trends, they cannot fully account for the unpredictable nature of electoral outcomes. While the dataset provides an overall approval rating for Harris, it lacks a breakdown by factors such as age, gender, income, and education, which are critical to understanding the preferences of different groups of voters. Without them, our analysis may have overlooked significant changes in specific groups that could affect overall trends.

5.4 Next Steps

To enhance the predictive power of the model, historical voting data, economic indicators and major news events can be used as background variables. By analyzing the impact of similar

factors on voter sentiment in past elections, we can gain insight into patterns and trends that may apply to the current campaign. For example, high-profile events have historically influenced voter opinion, and understanding these influences can provide a more complete picture of the factors that drive support for each candidate. Given the underlying conditions that influence voter preferences over time, this additional context would allow the model to make more informed predictions.

Another possible improvement is the addition of detailed demographic information, such as age, gender, income level, and educational background, which would allow us to identify trends among specific groups of voters and thus greatly enhance our analysis. Different demographics often have unique responses to campaign issues and activities; for example, younger voters may prioritize different issues compared to older voters, and income level and educational attainment may influence political preferences in complex ways. By incorporating this level of granularity, future analyses can provide deeper insights into voter behavior and help develop more targeted campaign strategies.

Appendix

A Additional data details

A.1 Dataset and Graph Sketches

Sketches depicting both the desired dataset and the graphs generated in this analysis are available in the GitHub Repository `other/sketches`.

A.2 Data Cleaning

In this data-cleaning process, we focus on refining raw polling data for Kamala Harris and Donald Trump to enhance its quality and relevance for analysis. The process begins by loading the dataset and using the janitor package to standardize column names, ensuring consistent naming conventions throughout. We then filter the data to retain only essential columns and remove rows with missing values in key fields, including numeric_grade, pct, sample_size, and end_date.

For each candidate, we isolate polls specifically for Kamala Harris and Donald Trump, retaining only high-quality polls with a numeric_grade of 2 or higher—given that the average numeric_grade is approximately 2.175, with a median of 1.9. We also handle placeholder values in state information by setting entries marked as NA to National, and create a national_poll indicator, assigning a value of 1 for value in state equals to National and 0 for others. Dates are standardized using the lubridate package to facilitate accurate time-based analysis.

Recency are calculated based on the days elapsed since the poll's end date.Additionally,categorical variables, including pollster, state, candidate_name, population, and methodology, are converted to factors to prepare for analysis.

The cleaned datasets for both candidates are then saved as Parquet files for efficient storage and access in further modeling and analysis. This structured approach ensures that the data is accurate, complete, and optimized for insightful statistical analysis.

A.3 Attribution Statement

B Model details

B.1 Model validation: K-Fold Cross-Validation

```
For the Harris model, the RMSE is 2.43, R-squared is 0.63, and MAE is 1.60.

For the Trump model, the RMSE is 2.57, R-squared is 0.66, and MAE is 1.62.
```

We use a 10-fold cross-validation on two linear regression models—one for Harris and one for Trump. The models use three predictors: national_poll, pollster, and population. The output provides key metrics, which breaks down here:

RMSE (Root Mean Square Error): Measures the average magnitude of prediction errors (lower is better).

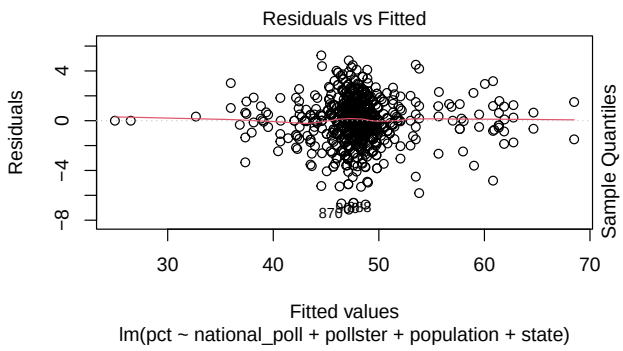
Harris model: RMSE of 2.43, indicating an average prediction error of around 2.43 percentage points.
Trump model: RMSE of 2.57, showing a slightly average prediction error on average.

R-squared: Represents the proportion of the variance in the response variable explained by the model (higher is better).
Harris model: R-squared of 0.63, meaning the model explains about 63% of the variance in Harris's polling data.
Trump model: R-squared of 0.66, meaning the model explains about 66% of the variance in Trump's polling data.

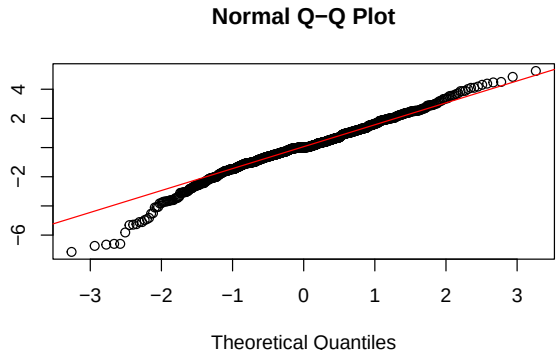
MAE (Mean Absolute Error): Shows the average absolute difference between observed and predicted values (lower is better).
Harris model: MAE of 1.6, meaning that, on average, the predictions are off by 1.62 percentage points.
Trump model: MAE of 1.62, indicating slightly higher precise predictions compared to the Harris model.

Interpretation Summary Predictive Accuracy: The Harris model has slightly better predictive accuracy than the Trump model, as reflected by its lower RMSE and MAE values.
Model Fit: Harris' model explain roughly 63% of the variance in their respective datasets. Trump's models explain roughly 66% of the variance in their respective datasets.This suggests that other factors not included in the model may play a significant role in explaining the remaining variance.
This summary indicates the models are moderately predictive, with room for improvement in accuracy and fit, potentially by adding more predictors or adjusting model specifications.

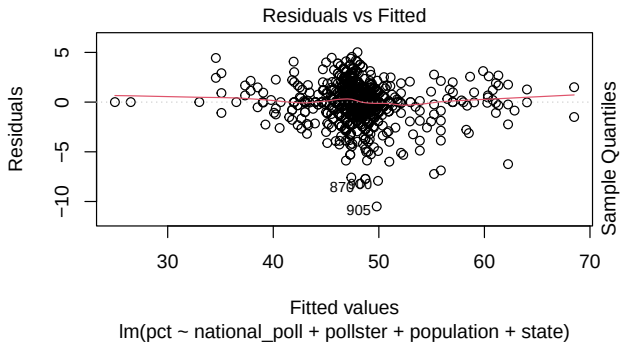
B.2 Diagnostics



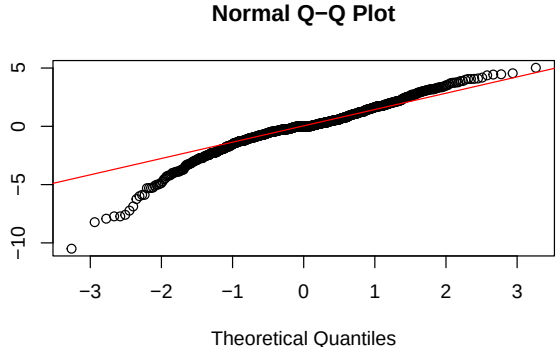
(a) Residual Plot for Unweighted Model



(b) Q-Q Plot for Unweighted Model



(c) Residual Plot for Weighted Model



(d) Q-Q Plot for Weighted Model

Figure 11: Diagnostics of model using residual vs fitted plot and norm Q-Q plot -Support for Harris

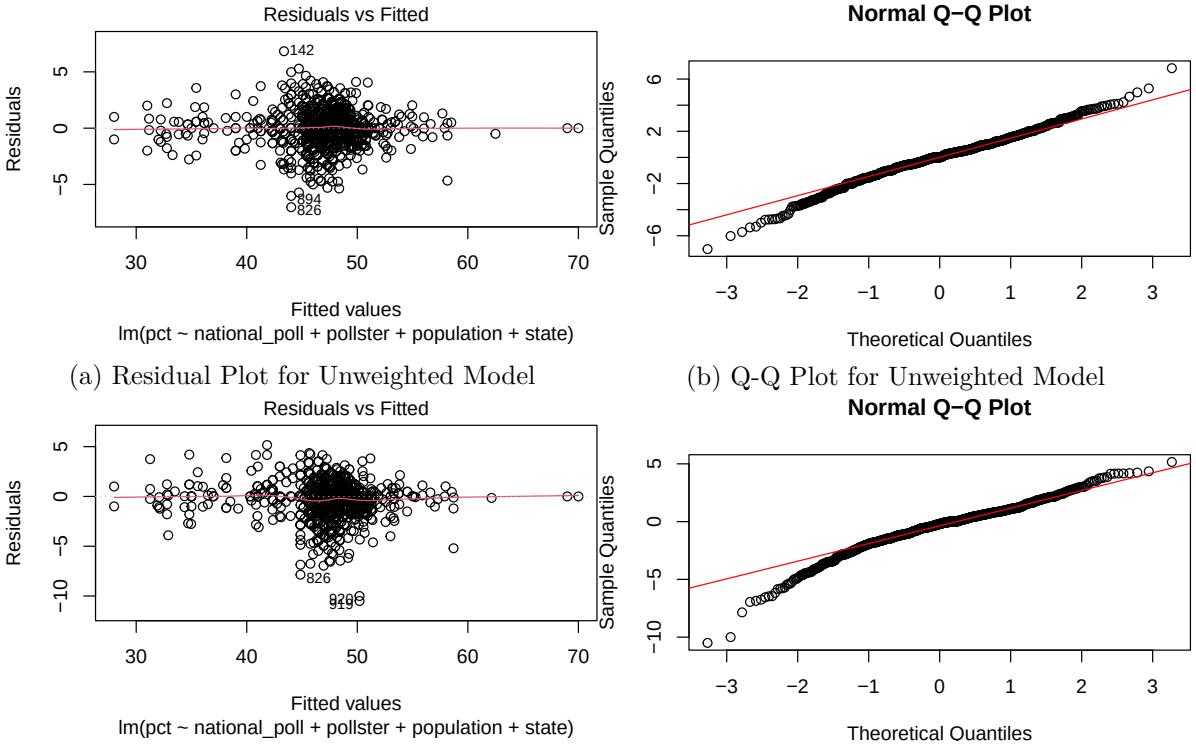
Generally, we use Residual vs Fitted plot and Q-Q Plot diagnostic our model. Residual vs Fitted plot aare Residuals (differences between observed and predicted values) plotted against fitted values. Ideally, these residuals should be randomly scattered around the zero line to indicate that the model does not have systematic errors. The Q-Q plot for the unweighted model shows how the residuals align with a theoretical normal distribution. Ideally, residuals should follow a straight line in this plot if they are normally distributed, which is an assumption of linear regression.

Figure 11a and Figure 11c are residual plots of un-weighted model for Harris support. It showsthe residuals are generally spread around zero, with no clear pattern. This suggests the model is relatively well-specified.

Figure 11b and Figure 11d is are Q-Q plots of un-weighted and weighted model plot for Harris support. It shows most residuals fall along the line, especially in the middle range. This suggests that our model satisfy the normality assumption. However, some points at the head and tail deviate, indicating potential outliers or non-normality in the extreme residual values.

This slight deviation at the ends suggests the model might have some issues with extreme predictions but performs reasonably well overall.

In summary, both models show a reasonably good fit. Both models exhibit minor deviations from normality and a few notable outliers, which may warrant further model adjustments for improved prediction accuracy.



(a) Residual Plot for Unweighted Model

(b) Q-Q Plot for Unweighted Model

(c) Residual Plot for Weighted Model

(d) Q-Q Plot for Weighted Model

Figure 12: Diagnostics of model using residual vs fitted plot and norm Q-Q plot -Support for Trump

Figure 12a and Figure 12c shows the residuals plotted against the fitted values for the unweighted model for Trump's Support. It shows that the residuals are generally spread around zero with no clear pattern for both model, suggesting that the model is well-specified.

Figure 12b and Figure 12d shows the Q-Q plot for the unweighted and weighted model for Trump's Support, showing that most residuals fall along the line, especially in the middle range, suggesting that the model satisfies the normality assumption. However, some points at the head and tail deviate, indicating potential outliers or non-normality in extreme residual values.

Summary Both models exhibit a reasonably good fit, with the weighted model offering slight improvements in managing non-linearity and extreme values. Despite this, both models show

minor deviations from normality and some notable outliers, suggesting that further model adjustments may be beneficial for improved prediction accuracy.

C The New York Times/Siena College Polling Methodology

This appendix outlines the methodology used by the Siena College Polling Institute for conducting its surveys. Known for its methodological rigor, Siena College focuses on accurately capturing voter sentiment during elections and has conducted polls in key states such as Michigan, Wisconsin, and Ohio.

In this section, we examine the main components of Siena's polling methodology, including the target population, sampling frame, recruitment processes, and sampling strategies. Additionally, we discuss how Siena addresses non-response and review the strengths and weaknesses of its questionnaire design. By detailing these elements, this appendix clarifies how Siena College ensures the reliability and validity of its polling results, offering valuable insights into voter behavior and election dynamics.

C.1 Pollster Overview

Siena College Polling Institute is a prominent pollster known for its comprehensive and methodologically rigorous surveys. It specializes in political polling and is particularly recognized for its work in understanding voter sentiment during elections. Established in 1980 at Siena College in New York's Capital District, the institute carries out both expert and public opinion polls. (Siena College Research Institute 2024).

C.2 Population, Frame and sample

Refer from Rohan Alexander (2023), we defined three key terms as:
Target population : The collection of all items about which we would like to speak/ the entire group about which we want to draw. conclusions
Sampling frame : A list of all the items from the target population that we could get data about.
Sample : The items from the sampling frame that we get data about.

The target population for Siena's polls includes registered voters eligible to vote in Michigan, Wisconsin, and Ohio.

The sampling frame is a comprehensive list of registered voters, which includes demographic information for each voter. This enables the pollsters to ensure an appropriate representation of voters across various parties, races, and regions (The New York Times 2024c).

The sample of registered voters sourced from the voter file maintained by L2, a nonpartisan vendor, and supplemented with additional cellular phone numbers matched from Marketing Systems Group. The sample for the poll totals 2,055 likely voters, with 688 from Michigan, 687 from Ohio, and 680 from Wisconsin, surveyed from September 21 to 26, 2024.

C.3 Sample Recruitment

Siena use phone poll to recruit sample. Telephone polling is a way to gather public opinion by contacting individuals via landlines and mobile phones, using live interviewers to improve data quality and capture nuanced responses. Through random digit dialing or voter registration databases, researchers achieve a representative sample across demographics.

According to The New York Times (2024b), the polls are conducted in both English and Spanish by live interviewers at call centers located in Florida, New York, South Carolina, Texas, and Virginia. The respondents are randomly selected from a national database of registered voters and are contacted via both landlines and cellphones.

C.4 Sampling Approach

Siena employs a response-rate-adjusted stratified sampling of registered voters sourced from the voter file maintained by L2, a nonpartisan vendor, and supplemented with additional cellular phone numbers matched from Marketing Systems Group. The New York Times selected the sample in multiple stages to address differences in telephone coverage, nonresponses, and notable variations in telephone number productivity by state.

Stratified sampling is typically utilized to ensure all strata of the population are represented. When considering our population, it typically consists of various groupings. These can range from a country being divided into states, provinces, counties, or statistical districts to a university comprising faculties and departments or even demographic characteristics groups among individuals. A stratified structure allows us to categorize the population into mutually exclusive and collectively exhaustive sub-populations known as "strata"(Rohan Alexander 2023).

In this scenario, we want to collect the polls from all strata of our target population to balance our poll result. The sample was stratified by political party, race, and region, and screened by M.S.G. to ensure that the cellular phone numbers were active.

C.4.1 Strength and Weakness

Stratified sampling enhances sample representativeness by ensuring that smaller subgroups are adequately included, allowing researchers to allocate resources more efficiently and gain deeper insights into specific groups. However, this method can lead to **higher costs** due to the

extensive data collection and analysis needed, especially when sampling large regions. Stratified sampling also introduces **complexity in data analysis**, requiring advanced techniques to accurately interpret subgroup data and appropriate weighting for each stratum. Additionally, poorly defined strata or imbalanced sampling can lead to sampling bias. While stratified sampling provides strong representation and analytical depth, it also brings challenges related to cost, complexity, and potential bias if not executed with care.

C.5 Non-response Bias

An interview was deemed complete for inclusion in the voting preference questions if the respondent stayed engaged in the survey after answering the two self-reported variables used for weighting—age and education—and provided responses to at least one question concerning age, education, or the presidential election candidate reference. If these conditions were not met, the interview was recorded as a non-response.

To handle the non-response bias, Siena choose to use weighting adjustments. Weighting is like balancing a scale to make sure each group in the survey counts the right amount. It changes the importance of each answer depending on how likely people are to skip the survey (Kinga Edwards 2024).

Siena use several steps to address nonresponse bias and ensure the reliability of the results. The weighting process was conducted by The New York Times using the R survey package and involved multiple adjustments. Siena's weighting process involved adjusting samples for unequal selection probabilities and turnout likelihood, based on 2020 data. Further adjustments aligned the sample with likely electorate targets from the L2 voter file. The final weight combined modeled turnout (80%) and self-reported intentions (20%), mitigating nonresponse bias and ensuring the sample accurately reflected the characteristics and behaviors of likely voters, thereby enhancing result validity.

C.6 Questionnaire Design

C.6.1 Response bias defination

In the design of the questionnaire, there will be some common bias that may occur when running the questionnaire.

Stantcheva, Stefanie (2023) define these bias as:

- Moderacy response bias is the tendency to respond to each question by choosing a category in the middle of the scale.
- Extreme response bias is the tendency to respond with extreme values on the rating scale.

- Response order bias occurs when the order of response options in a list or a rating scale influences the response chosen. The primacy effect occurs when respondents are more likely to select one of the first alternatives provided, and it is more common in written surveys. This tendency can be due to satisficing, whereby a respondent uses the first acceptable response alternative without paying particular attention to the other options. The recency effect occurs when respondents choose one of the last items presented to them (more common in face-to-face or orally presented surveys).
- Social desirability bias typically stems from the desire of respondents to avoid embarrassment and project a favorable image to others, resulting in respondents not revealing their actual attitudes. The prevalence of this bias will depend on the topic, questions, respondent, mode of the survey, and the social context. For instance, in some circles, anti-immigrant views are not tolerated, and those who hold them may try to hide them. In other settings, people express such views more freely.
- Acquiescence is the tendency to answer items in a positive way regardless of their content, for instance, systematically selecting categories such as "agree," "true," or "yes".

C.6.2 Strengths and Weakness

Strengths:

The questionnaire is concise and straightforward, reducing respondent fatigue and enhancing clarity, which is crucial for maintaining engagement. By incorporating both closed- and open-ended questions, it allows for both quantifiable data and rich qualitative insights. Clear response categories help reduce moderacy bias, encouraging participants to choose decisively rather than defaulting to neutral answers. Additionally, varied question types help mitigate acquiescence bias by encouraging honest responses and avoiding leading language.

Weaknesses:

However, the questionnaire has some limitations. Its reliance on agree-disagree and yes-no formats may increase acquiescence bias, as respondents may lean toward favorable answers. Furthermore, some demographic nuances may be inadequately addressed, potentially leading to nonresponse bias from underrepresented groups. The risk of response order bias is also present, especially if randomization of options is not implemented, increasing the chance of recency effects in verbally-administered surveys.

Additionally, the absence of assured anonymity could lead to social desirability bias, where respondents alter answers to project a favorable image. Lastly, with over 50 questions, the length of the survey may increase dropout rates, especially in time-intensive formats like telephone surveys, thereby raising nonresponse bias.

In summary, while the questionnaire is clear and well-structured, it faces challenges from potential biases including acquiescence, nonresponse, social desirability, and order effects. Future

improvements should focus on diversifying question types, ensuring demographic inclusivity, and refining question phrasing to reduce bias and enhance validity.

D Idealized Methodology for US Presidential Election Forecast

This appendix details the methodology and design for conducting a U.S. presidential election forecast survey with a budget of $\$100,000$. The objective is to generate an accurate and reliable prediction of the election outcome while ensuring data quality through meticulous sampling, recruitment, validation, and aggregation of results.

D.1 Sampling Approach

To ensure a representative sample of likely voters, I will employ a Composite Measure sampling method based on past ballots cast data from the 2020 U.S. elections. After determining the sample size for each state, I will use stratified sampling based on demographics, dividing the population into subgroups and taking random samples from each subgroup. This Composite Measure sampling approach, as referenced in Clark Letterman (2021), enhances our chances of selecting respondents from states or regions that have historically exhibited higher voter engagement compared to the general population distribution. While some states may have larger populations, we aim to adjust the sampling to reflect higher anticipation rates.

To illustrate this Composite Measure of size, consider two states with similar populations. For instance, although State A and State B both have 1 million eligible voters, State B consistently shows a higher ballots cast in past elections. Therefore, we will increase the proportion of polls conducted in State B. In this scenario, State A has a historical turnout rate of 50%, while State B has a turnout rate of 70%. In a purely population-based sampling approach, both states would have an equal chance of being selected for polls: 50% for State A and 50% for State B. However, by incorporating ballots cast, we modify these probabilities to increase the likelihood of selecting State B due to its higher historical turnout.

In the subsequent steps, we will detail how to utilize **ballots cast** as a crucial factor in creating a **composite measure of size** for sampling U.S. election polls. Rather than relying solely on population size, we will adjust the sample allocation based on historical voter turnout, ensuring that regions with higher engagement are more prominently represented in our polling data.

D.1.1 Step 1: Define the Sampling Data

We begin by gathering data on the **eligible voter population** and **historical ballots cast** for different states. For simplicity, we'll focus on two states: **State A** and **State B**.

State	Eligible Voters	Ballots Cast
State A	1,000,000	500,000
State B	1,000,000	700,000

D.1.2 Step 2: Calculate Total Ballots Cast

The **total ballots cast** across both states is the sum of ballots cast in each state:

Total Ballots Cast = $BallotsCast_A + BallotsCast_B$

Total Ballots Cast= 500,000 + 700,000 = 1,200,000

D.1.3 Step 3: Calculate Composite Measure of Size

Using the total ballots cast across both states, we calculate the **proportion** of each state's ballots relative to the total. This proportion serves as the composite measure of size, which will guide our sample allocation.

For State A:

Sampling Proportion A = $\frac{500,000}{1,200,000} \approx 0.417$

For State B:

Sampling Proportion B = $\frac{700,000}{1,200,000} \approx 0.583$

D.1.4 Step 4: Allocate Sample Based on Ballots Cast

Finally, we allocate the sample size according to these calculated sampling proportions. For example, if conducting 1,000 surveys, the allocation would be:

- **Polls for State A**:

Polls for State A = $1,000 \times 0.417 \approx 417$

- **Polls for State B**:

Polls for State B = $1,000 \times 0.583 \approx 583$

By using the number of ballots cast, we ensure that the sample allocation reflects historical voting participation, giving each state an influence proportional to its voter turnout in previous elections.

By using historical ballots cast to adjust our polling sample, we ensure that regions with higher voter engagement have a greater influence on the polling results. Consequently, we can

produce more accurate and representative poll outcomes that account for the varying levels of voter participation across the country.

D.1.5 Stratification Variables

After determining the sample size for each region, we will use stratified sampling across key demographic categories, such as age, gender, race/ethnicity, and education level. This method ensures that our final sample proportionally represents each subgroup within every region, accurately reflecting the diversity of the U.S. voting population. To achieve this, the sample will be stratified according to critical demographic and geographic variables, with strata information sourced from U.S. Census data available through IPUMS USA (Ruggles et al. 2024).

D.2 Target Population

Our target population is all U.S. citizens eligible to vote in the 2024 U.S. presidential election (age>=18).

D.3 Sample frame

Based on the recruitment method we discussed later, our sampling frame could be all registered voters in online panels like Qualtrics and YouGov and the millions of U.S. voters who are reachable via social media platforms like Facebook and Instagram.

D.4 Sample

We plan to survey 400 respondents. To optimize the limited sample size and enhance the stratified sampling approach, we will group states into four regions: Midwest, Northeast, South, and West. Sample sizes for each region will be allocated according to the proportion of total ballots cast in each region during the 2020 election. This regional allocation ensures representative sampling while aligning with past voting patterns, as outlined in Table 5.

Table 5: Regional Grouping of States in the USA

Region	**States**
Midwest	Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin

Region	States
Northeast	Connecticut, Delaware, District of Columbia, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont
South	Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, West Virginia
West	Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, Wyoming

Table 6: Regional Voting Data and Sample Size Allocation calculated using Composite Measure Sampling Proportion based on 2020 US Election regional ballots cast

Region	Total Ballots Cast	VEP	Composite Measure Sampling Proportion	Sample Size
Midwest	35,134,960	50,932,439	0.214805579	86
Northeast	32,262,303	47,473,317	0.200216867	80
South	54,746,770	84,563,831	0.356644666	143
West	37,594,304	54,139,892	0.228332888	91

Table 6 shows the regional breakdown of the 2020 election data, sourced from (Wikipedia contributors 2024), including total ballots cast, and Voting Eligible Population (VEP). The sample size for each region shown in Table 6 was determined based on the **Composite Measure Sampling Proportion**.

D.5 Recruitment of Respondents

To maximize our budget, we will concentrate on online recruitment methods, which provide a cost-effective and efficient means of reaching a diverse and representative sample of voters nationwide.

Online Recruitment Strategy: Aiming for a total of 400 respondents, we will focus our resources on survey implementation and high-quality data collection, developing the survey in-house and recruiting participants through online survey platforms.

Online Panel Providers (Qualtrics, YouGov): We will recruit 200 respondents via reputable panel providers like Qualtrics and YouGov. These platforms ensure high-quality samples by

verifying voter registration, offering a solid foundation of reliable data due to their strict participant verification protocols.

Social Media Recruitment: An additional 400 respondents will be recruited through targeted ads on platforms like Facebook and Instagram. Given the typically lower data quality from social media sources, we anticipate that about 50% of responses may be invalid, and we will oversample accordingly to secure a sufficient number of valid responses. To incentivize participation, respondents will receive a small monetary reward, such as a gift card, upon survey completion. We will also use targeted ads and eligibility screening (e.g., age and U.S. voter registration status) to ensure that respondents are likely to be eligible voters.

D.6 Handling Non-response bias

Nonresponse bias occurs when participants are unwilling or unable to answer certain questions or complete the survey. Handling non-response bias matters because it can skew survey results, leading to inaccurate conclusions that do not accurately represent the entire population's views or characteristics. Given that our survey takes approximately 15 minutes to complete, there is a risk of nonresponse bias. To mitigate this, as highlighted by Survey Monkey (2024), we set clear expectations about the survey's purpose and estimated completion time, encouraging participants to stay engaged. Additionally, we apply post-stratification, adjusting survey weights to ensure that our respondent group aligns closely with the actual population characteristics, reducing bias, and improving representativeness (Kinga Edwards 2024).

D.7 Respondent Validation

To ensure high data quality and accuracy, we will conduct rigorous respondent validation through several verification steps, ensuring only eligible and relevant participants are included. Respondents will confirm their voter registration status, with a portion cross-referenced against voter databases or verified via reliable panel providers like Qualtrics and YouGov, thus focusing the survey on registered and likely voters. Eligibility will also be confirmed through screening questions verifying age (18+) and U.S. citizenship, allowing only those who meet these criteria to proceed. Attention-check questions embedded throughout the survey will filter out inattentive participants, with non-compliant responses excluded from the final sample. Additionally, unique identifiers such as IP and email addresses will be tracked to prevent duplicate submissions, ensuring each response is unique. Finally, post-survey data cleaning will address any inconsistent or incomplete responses to maintain dataset reliability. These validation steps are essential for producing data that accurately reflects the opinions of eligible, registered, and attentive respondents, thereby enhancing the survey's overall validity.

D.8 Poll Aggregation

After collecting survey responses, we will aggregate data from our two recruitment sources: online panel providers and social media platforms. To weight these panels accurately, we'll first identify key demographic variables (e.g., age, gender, race/ethnicity, education level) and establish population benchmarks using U.S. Census data (Ruggles et al. 2024). For the online panel (200 respondents) and social media panel (400 respondents), we will calculate weights by comparing each panel's demographic distribution to these benchmarks, adjusting for under- and over-represented groups.

Once individual weights are calculated for each panel, we will integrate them into a single weighting scheme that aligns with the target population's demographic makeup. These weights will then be applied during data analysis, allowing underrepresented groups to have an appropriate influence on the results. Finally, we'll conduct post-stratification adjustments to ensure that the combined sample accurately reflects the U.S. voting population, yielding reliable insights into voter preferences and behaviors.

D.9 Survey Design

The survey is designed to capture essential insights into voting intentions, candidate favorability, and the issues influencing voter decisions. It will be concise and straightforward, taking no longer than 15 minutes to complete.

Survey Link
The survey has been implemented using Google Forms. You can access it here: Survey Link.

In our survey, several questions are adapted from the Emerson College Polling data (Emerson College Polling 2024). We apply insights from Stantcheva, Stefanie (2023) to minimize response biases. Common response biases identified in survey design include moderacy bias, extreme response bias, ordering bias, acquiescence bias, experimenter demand effect (EDE), and social desirability bias (SDB). Our survey primarily focuses on strategies to reduce moderacy bias, extreme response bias, ordering bias, SDB, and acquiescence bias.

D.9.1 Definition of the response bias

We have defined the bias we want to solve in Section 2.5.4.

D.9.2 Solution to the response bias in our survey

To mitigate bias, we enhance our survey in the following ways, drawing on recommendations from Stantcheva, Stefanie (2023):

Addressing Extreme/Moderacy Bias: We use a minimum of five response options for scale questions to provide more nuanced choices, reducing the likelihood of respondents defaulting to extreme or middle answers.

Mitigating Response Order Bias: For nominal questions, we randomize response options, and for ordinal questions, we vary the order. Open-ended formats and pauses (e.g., "Who would you vote for? [pause] Candidate A or Candidate B?") further minimize order effects.

Minimizing Social Desirability Bias (SDB): Our online survey format and minimal introductory information (only stating it's for academic research in Statistics) reduce SDB. We guarantee respondent anonymity on the survey landing page and before sensitive questions, reminding participants that all answers are confidential. Additionally, a feedback section at the end allows respondents to express any concerns.

Reducing Acquiescence Bias: We avoid agree-disagree formats, instead using direct scales (e.g., "very unfavorable" to "very favorable") and item-specific options (e.g., "Approve, Disapprove, Neutral") to capture a full range of views.

D.10 Budget Breakdown

Budget Breakdown With a total budget of $\$100,000$, the allocation for various components of the survey implementation and data collection is as follows:

Survey Design and Development: $\$2,000$ Covers question formatting, testing, and online integration (e.g., Qualtrics) to ensure user-friendly and relevant survey design.

Online Panel Providers (Qualtrics, YouGov): $\$80,000$ 200 respondents recruited at \$400 each, leveraging verified voter panels for high-quality data.

Social Media Recruitment (Facebook, Instagram): $\$12,000$ 400 respondents recruited via targeted ads (\$30 each). Anticipating a 50% invalid rate, allowing for 200 valid responses after validation.

Data Validation and Quality Control: $\$6,000$ Includes voter registration checks, attention filters, and post-collection quality review, especially for social media responses, to ensure data integrity.

D.11 Copy of U.S. Presidential Election Polls Survey

Welcome to our 2024 U.S. Presidential Election Polls Survey. Your participation in this survey is vital in helping us understand voters' preferences and opinions on key issues. Rest assured that your responses are anonymous and will only be used for statistical analysis.

This survey is for academic research in Statistics. It consists of 26 carefully designed questions and should take approximately 12-15 minutes to complete.

For any questions or concerns regarding this survey, please contact:
Email: diana.shen@mail.utoronto.ca; jinyan.wei@mail.utoronto.ca; huayan.yu@mail.utoronto.ca

---

Privacy Notice for Respondents

Your privacy is our priority. In this survey, your responses are completely anonymous, ensuring that no one can link your answers back to you. We encourage you to share your true opinions, as this survey is conducted by a neutral, nonpartisan entity. Your data will only be used for research purposes, and you will not be identified individually. If you have concerns, we ask for your feedback at the end of the survey to ensure transparency and trust.

Section 1: Survey Questions

1. **What is your party registration or affiliation?**

   - Democrat
   - Republican
   - Independent/Other
   - Prefer not to say
   - Other: _____

2. **If the Presidential Election were held today, would you vote for Kamala Harris or Donald Trump?**

   - Kamala Harris
   - Donald Trump
   - Someone else
   - Undecided
   - Prefer not to say

3. **Although you are undecided, which candidate do you lean toward?** (Only answer if you chose "Undecided" in Question 2)

   - Kamala Harris
   - Donald Trump

4. **How favorable are you towards the following candidates?**

	Very unfavorable	Unfavorable	Moderate	Favorable	Very favorable
Kamala Harris					
Donald Trump					

5. **How likely are you going to vote in the 2024 election?** (1 = Definitely not to vote, 10 = Definitely will vote)

    - 1
    - 2
    - 3
    - 4
    - 5
    - 6
    - 7
    - 8
    - 9
    - 10

6. **How do you plan to cast your vote?**

    - In-person on election day
    - Early voting in-person
    - By mail
    - Unsure

7. **Did you vote in the 2020 U.S. presidential election?**

    - Yes
    - No
    - Prefer not to say

8. **If you voted in 2020, who did you vote for?**

    - Joe Biden
    - Donald Trump
    - Other
    - Prefer not to say

9. **Imagine the following candidates: Candidate A favors cutting taxes but has a weak stance on climate change, and Candidate B focuses on healthcare but supports increased military spending. Who would you vote for?**

- Candidate A
- Candidate B

10. **What do you think is the most important issue facing the United States?** [Select at most 3]

    - Economy
    - Healthcare
    - Climate Change
    - Immigration
    - National Security
    - Education
    - Social Security
    - Other: _____

11. **Select option 3 from the list below:**

    - Option 1
    - Option 2
    - Option 3
    - Option 4

12. **Which social media platforms do you use to get political news?** (Select all that apply)

    - Facebook
    - Twitter
    - Instagram
    - YouTube
    - None

Section 2: Demographic Information

Privacy Notice for Demographic Information Collection
Your demographic information is collected anonymously and will be used for statistical purposes only, helping us analyze trends across different groups. We ensure that your individual responses cannot be traced back to you, maintaining full confidentiality. Your privacy and honest participation are important to us.

1. **What is your age group?**

    - 18-24
    - 25-34

- 35-44
- 45-54
- 55-64
- 65+

2. **Region:**
   - Northeast
   - South
   - Midwest
   - West
3. **For statistical purposes only, can you please tell me your ethnicity?**
   - Hispanic or Latino of any race
   - White or Caucasian
   - Black or African American
   - Asian
   - Other or multiple races
4. **Can you please tell me your gender?**
   - Men
   - Women
   - Other
   - Prefer not to say
5. **What is the highest level of education you have attained?**
   - High school or less
   - Some college
   - Bachelor's degree
   - Graduate degree

Section 3: Feedback

1. **Do you have any concerns or feedback regarding the survey, surveyor, or entity?**
   Your feedback is important to us and will help ensure transparency and trust in the research process.

Thank You

Thank you for taking the time to complete this survey. Your honest feedback is invaluable and will contribute greatly to our research. We appreciate your participation!

References

A. C. Davison, and D. V. Hinkley. 1997. *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511802843.

Angelo Canty, and B. D. Ripley. 2024. *Boot: Bootstrap r (s-Plus) Functions*.

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.

Bengtsson, Henrik. 2021. "A Unifying Framework for Parallel and Distributed Processing in r Using Futures." *The R Journal* 13 (2): 208–27. https://doi.org/10.32614/RJ-2021-048.

Clark Letterman. 2021. "Q&A: How Pew Research Center surveyed nearly 30,000 people in India." https://medium.com/pew-research-center-decoded/q-a-how-pew-research-center-surveyed-nearly-30-000-people-in-india-7c778f6d650e.

Emerson College Polling. 2024. https://emersoncollegepolling.com/october-2024-national-poll-harris-50-trump-48/.

Erikson, Robert S., and Christopher Wlezien. 2008. "Are Political Markets Really Superior to Polls as Election Predictors?" *Public Opinion Quarterly* 72 (2): 190–215. https://doi.org/10.1093/poq/nfn010.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

FiveThirtyEight. 2024. "Dataset: 2024 u.s. Presidential Election Polls." https://projects.fivethirtyeight.com/polls/data/president_polls.csv.

Gelman, Andrew, and Gary King. 1993. "Why Are American Presidential Election Campaign Polls so Variable When Votes Are so Predictable?" *British Journal of Political Science* 23: 409451.

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

Kinga Edwards. 2024. "What is Non-Response Bias and Why It Matters." https://www.surveylab.com/blog/what-is-non-response-bias/.

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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

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.

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

Ruggles, Steven, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rodgers, and Megan Schouweiler. 2024. "IPUMS USA: Version

15.0." Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V15.0.

Siena College Research Institute. 2024. "Cross-Tabs: September 2024 Times/Siena Poll of the Michigan Likely Electorate." https://www.nytimes.com/interactive/2024/09/28/us/elections/times-siena-michigan-crosstabs.html.

Stantcheva, Stefanie. 2023. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." Journal Article. *Annual Review of Economics* 15 (Volume 15, 2023): 205–34. https://doi.org/https://doi.org/10.1146/annurev-economics-091622-010157.

Survey Monkey. 2024. "What is nonresponse bias and how to avoid errors." https://www.surveymonkey.com/mp/nonresponse-bias-what-it-is-and-how-to-avoid-its-errors/.

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

The New York Times. 2024a. "How We Conduct Our Election Polls." *The New York Times*. https://www.nytimes.com/article/election-polling-averages-methodology.html?unlocked_article_code=1.UE4.dFWl.2m54BRRnn8BR&smid=url-share.

———. 2024b. "You Ask, We Answer: How The Times/Siena Poll Is Conducted." https://www.nytimes.com/article/times-siena-poll-methodology.html.

———. 2024c. "Cross-Tabs: September 2024 Times/Siena Polls in Michigan, Ohio and Wisconsin." *The New York Times*, September 28, 2024. https://www.nytimes.com/interactive/2024/09/28/us/elections/times-siena-rust-belt-crosstabs.html.

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.

Wikipedia contributors. 2024. "Voter Turnout in United States Presidential Elections — Wikipedia, the Free Encyclopedia." https://en.wikipedia.org/w/index.php?title=Voter_turnout_in_United_States_presidential_elections&oldid=1252529917.