Predicting voter support: Analyzing polling dynamics for Kamala Harris in the 2024 U.S. election*

Support for Kamala Harris predicted to rise as Election Day nears, with notable differences across states and pollsters

Xingjie Yao

November 3, 2024

This study analyzes polling data to predict Kamala Harris's percentage support in the upcoming U.S. election. Key predictors of voter sentiment include poll end date, pollster, state, and poll score. Our findings show that Harris's support tends to increase closer to Election Day and varies significantly by pollster and state. These insights are essential for shaping campaign strategies and understanding the role of polling in influencing voter behavior.

1 Introduction

In the context of the upcoming U.S. election, understanding voter sentiment is crucial for political campaigns and analysts alike. The dynamics of public opinion can shift rapidly, influenced by various factors such as media coverage, campaign strategies, and significant events. This study focuses on predicting the percentage support for Kamala Harris, with the objective of providing insights into the factors that influence voter support as the election approaches. By analyzing data from multiple polls, we aim to identify key predictors of support, including the end date of the polls, the pollster, the state, and the poll score. Filling the gap in the existing literature, which often overlooks the nuances of polling data, our research seeks to enhance the understanding of voter behavior in the context of the election.

The primary estimand of our analysis is the percentage support for Harris, which we model as a function of various predictors. Specifically, we are interested in how the end date, pollster, state, and poll score influence voter support. Our linear regression framework allows us to

*Code and data are available at: [https://github.com/Stella41603/2024-US-Election-Forecast.git].

quantify the relationships between these predictors and the outcome variable, providing a clear picture of how each factor contributes to the overall support for Harris. By estimating the coefficients associated with each predictor, we can draw meaningful conclusions about the direction and magnitude of their effects on voter sentiment.

The results of our analysis indicate a significant positive relationship between the end date and percentage support, suggesting that support increases as the election draws nearer. Additionally, we found substantial variability in support based on the pollster and state, with certain pollsters consistently reporting higher levels of support for Harris. The poll score also played a crucial role, as higher-quality polls were associated with increased percentage support. These findings underscore the importance of understanding both the temporal dynamics of polling and the characteristics of different polling organizations when interpreting public sentiment.

Accurate predictions of voter support are essential for effective campaign strategies. By identifying the key factors that influence support for Harris, political teams can tailor their outreach efforts and messaging to resonate more effectively with voters. Moreover, understanding the variability across different pollsters and states can guide resource allocation and focus during the campaign. As elections are determined by small margins, having reliable insights into voter preferences can make a significant difference in the final outcomes.

The remainder of this paper is structured as follows. In Section 2, we detail the data sources and variables used in our analysis. Section 3 outlines the modeling approach, including the assumptions and specifications of the linear regression framework. In Section 4, we present the findings of our models, highlighting the key predictors of percentage support for Harris. Finally, in Section 5, we discuss the implications of our results and potential avenues for future research.

2 Data

2.1 Overview

We use the statistical programming language `R` to conduct our analysis of polling data. Our data, sourced from FiveThirtyEight (R Core Team 2023; FiveThirtyEight 2024), provides a comprehensive view of public opinion leading up to the election. Following the guidelines established in Alexander (2023), we consider various factors that influence percentage support, such as the timing of the polls, the characteristics of the polling organizations, and regional variations.

In this analysis, several `R` packages were utilized to enhance data manipulation, modeling, and visualization. The `tidyverse` package provided a cohesive framework for data wrangling and analysis, streamlining workflows (Wickham et al. 2019). The `here` package simplified file path management, ensuring easy access to data files (Müller 2020). `Janitor` was essential for cleaning the dataset, offering tools to identify and rectify data quality issues (Firke 2023). The

`lubridate` package facilitated date manipulation, making it simpler to work with time-related variables (Grolemund and Wickham 2011). We used `testthat` to test the simulated and the analysis datasets (Wickham 2011). Lastly, `arrow` enabled efficient reading and writing of data in a performant format, which is crucial for handling larger datasets (Richardson et al. 2024). Coding and file structure were adopted from Alexander (2023).

2.2 Measurement

The transition from real-world phenomena to entries in our dataset involves a structured process of measurement and data collection. In our study, we focus on gauging public sentiment toward Kamala Harris as the U.S. presidential election approaches. Polling organizations design surveys with targeted questions that capture voter opinions, such as their likelihood of voting for Harris and their perceptions of current political issues.

Once the survey questions are crafted, a representative sample of the population is drawn using stratified random sampling to ensure diverse demographic representation. Respondents are contacted through various methods, including telephone interviews and online surveys.

After collecting responses, the data undergoes cleaning and validation to address inconsistencies and missing values. This ensures the dataset accurately reflects the opinions of the electorate. Each entry in the final dataset corresponds to an individual's opinion at a specific time, allowing for meaningful analysis of how various factors influence public sentiment leading up to the election. This systematic approach transforms personal opinions into quantifiable data, ultimately facilitating insights into voter behavior and preferences.

2.3 Outcome variable

2.3.1 The percentage support for Harris in the poll.

The percentage support for Harris in a poll indicates the proportion of respondents who expressed their support for Kamala Harris in a given survey. This value reflects the overall popularity or favorability of Harris within the specific group of respondents surveyed by a pollster. It is expressed as a percentage, with values ranging from 0 to 100, where a higher percentage suggests stronger support.

Figure 1 represents the distribution of percentage support for Harris across various polls. The data shows that most polls report support levels clustered around 50%, with a noticeable peak just above 50%. The distribution is roughly normal, though slightly right-skewed, as indicated by the few outlying values extending toward higher percentages (above 60%). There is a concentration of responses in the 45% to 55% range, indicating that a significant portion of polls reflect moderate levels of support for Harris. However, there are very few polls that report support below 40% or significantly above 60%, suggesting that extreme views (either

very low or very high support) are rare. This pattern could reflect Harris's relatively stable support base among the surveyed population.

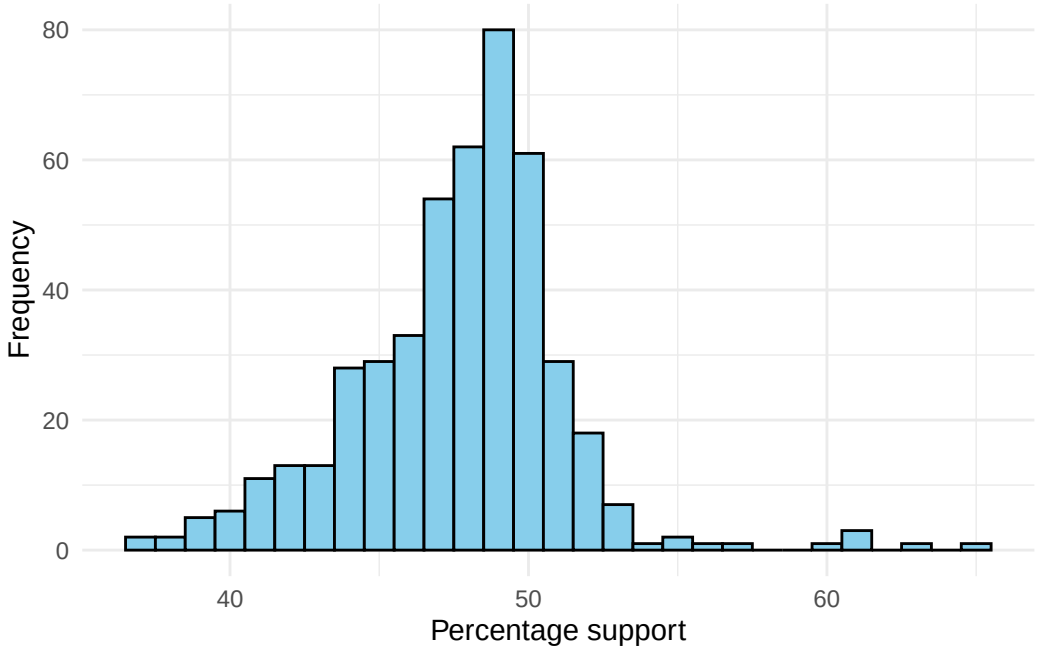


Figure 1: Percentage support for Kamala Harris in the 2024 U.S. election, showing a central peak around 50% with most values between 40% and 55%. The distribution indicates moderate support with few polls showing support below 40% or above 60%.

2.4 Predictor variables

2.4.1 Pollster

The pollster refers to the polling organization or source that conducted the poll. Polling organizations, such as Emerson, YouGov, or Quinnipiac, gather data from respondents to measure public opinion on a variety of topics, including political support. Each pollster may use different methodologies, sampling techniques, and geographic focuses, which can affect the results and the poll's reliability.

Figure 2 illustrates the number of polls conducted by different pollsters. Siena/NYT conducted the highest number of polls, with over 75, followed by YouGov and Emerson, both conducting more than 50 polls. Pollsters like Ipsos, Beacon/Shaw, and Quinnipiac also contributed significantly, each with 40-50 polls. After these top pollsters, the number of polls sharply declines, with organizations such as Marist, Marquette Law School, and AtlasIntel contributing fewer than 30 polls. A variety of other pollsters, including CNN/SSRS, SurveyUSA, and Echelon

Insights, have much smaller contributions. The least active pollsters, such as YouGov Blue and YouGov/Center for Working Class Politics, conducted only a handful of polls, showing that the majority of polling data comes from a small number of highly active pollsters.

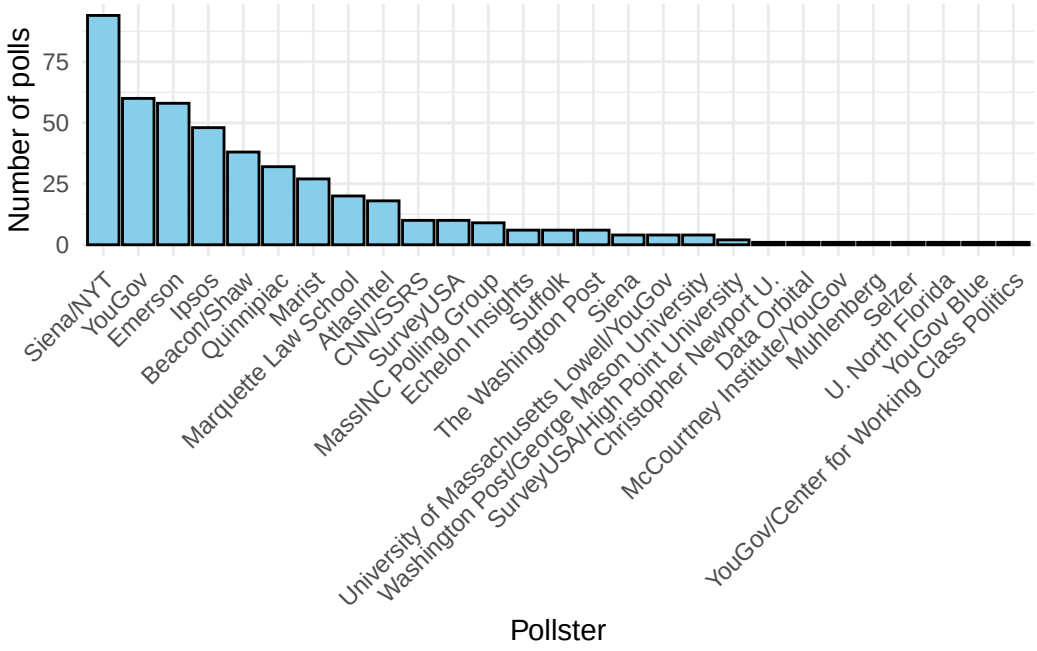


Figure 2: Showing the number of polls conducted by each pollster. Siena/NYT conducted the most polls, followed by YouGov, Emerson, and Ipsos. This distribution highlights the varying levels of polling activity across different organizations.

2.4.2 State

The state variable represents the geographic region where the poll was conducted. This can refer to a specific U.S. state, such as Arizona or California, or a broader national sample, indicated by "National." Polls conducted in individual states typically provide localized insights into voter preferences, which are crucial for understanding regional variations in support. National polls, on the other hand, aggregate opinions from across the country, offering a broader view of public sentiment.

Figure 3 shows the number of polls conducted across different states, with the states arranged in descending order based on the number of polls. The National polls dominate the dataset, with more than 150 polls, significantly higher than any individual state. Pennsylvania, Wisconsin, and North Carolina also show a relatively high number of polls, each with over 50. Other states such as Arizona, Georgia, and Michigan have moderate representation, while a large number of states, including California, Missouri, and Rhode Island, show much lower poll

counts, with fewer than 10 polls each. This pattern highlights the emphasis placed on national polling compared to state-level polling, though key battleground states receive more focus than others.

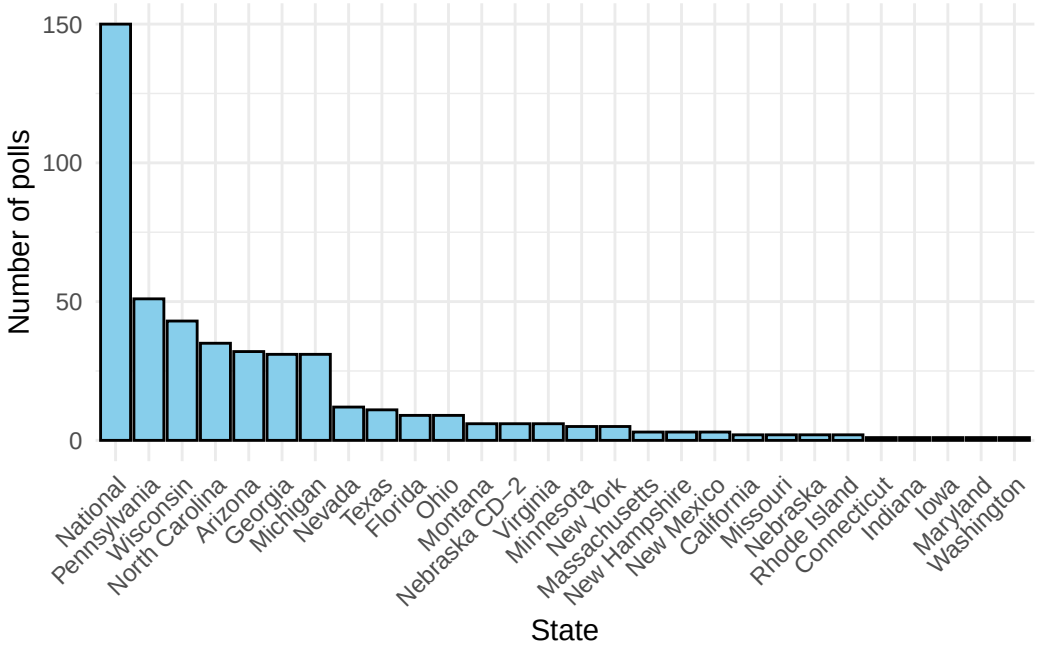


Figure 3: Number of polls conducted per state, with a large number focused on national polling. Pennsylvania, Wisconsin, North Carolina, and Arizona are among the states with the highest polling frequency, reflecting their importance as swing states in the election. Fewer polls were conducted in other states, indicating lower polling focus in those areas.

2.4.3 End Date

The end date refers to the date when the poll was completed. This is the final day when respondents' data was collected, marking the conclusion of the survey period. The end date is important for understanding the context of the poll, as public opinion can shift over time due to current events, campaign developments, or other external factors.

2.4.4 Poll score

The poll score is a numerical rating that reflects the overall quality or reliability of the poll. It may be based on factors such as the pollster's track record, transparency, sampling methodology, and adherence to best practices in polling. A higher pollscore typically suggests a more

reliable poll with accurate representation, while a lower score may indicate potential issues with the poll's quality, such as bias, small sample sizes, or methodological flaws.

Figure 4 shows the distribution of poll scores across the dataset. Most of the poll scores fall between -1.4 and -1.2, with the highest concentration at -1.2, indicating that many polls received similar ratings in this range. There is another noticeable cluster around -0.8 to -0.6, showing a smaller group of polls with higher scores. The lower end of the distribution, around -1.6, also has a significant number of polls. This pattern suggests that most polls tend to receive lower scores, with relatively few polls achieving scores closer to -0.4, which could indicate better performance or quality. The negative scores overall suggest that the poll ratings system is weighted towards lower values, possibly reflecting the stringency of the scoring criteria.

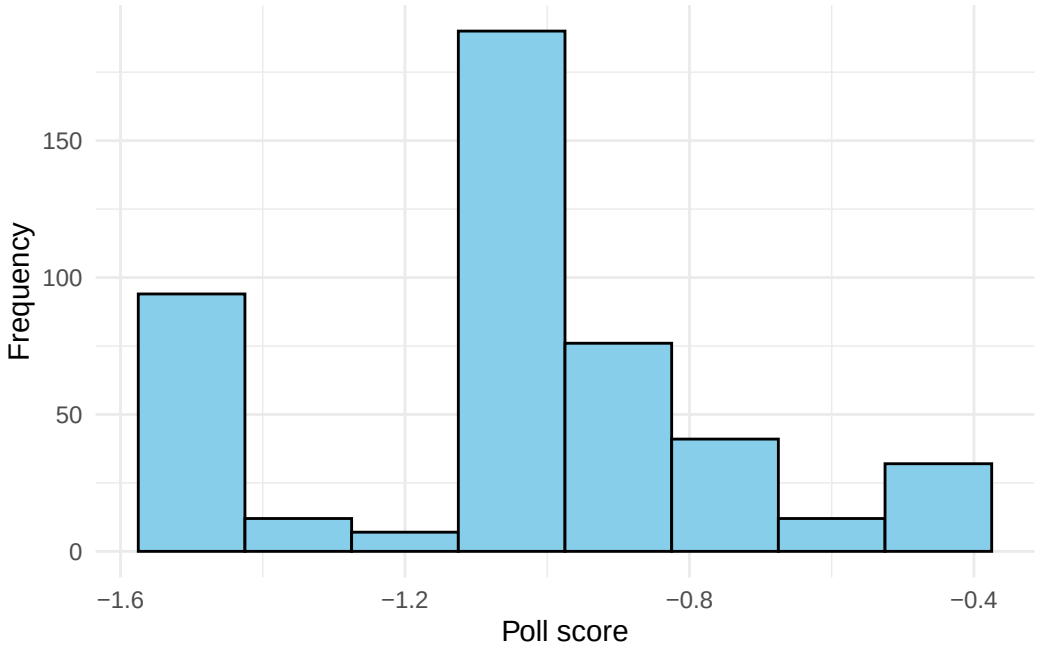


Figure 4: The distribution of poll scores across various ranges. The majority of poll scores are concentrated around -1.2, with fewer polls in the higher and lower score ranges.

2.5 Relationships between key varaibles

Figure 5 illustrates the distribution of percentage support for Harris across different states. Each box represents the interquartile range (IQR), which contains the middle 50% of the data, while the line inside the box indicates the median percentage support in that state. The whiskers extend to the smallest and largest values within 1.5 times the IQR, and the dots outside the whiskers represent outliers.

Key observations:

- Maryland, Massachusetts, and California show the highest percentage support, with median values around or above 60%.
- Nebraska, Montana, and Indiana show the lowest percentage support, with median values below 45%.
- National polls, as well as many battleground states like Georgia, Arizona, and Pennsylvania, have median support levels around 50%, but the data shows substantial variability within these states, as indicated by wider IQRs.
- Some states, like Nebraska and Florida, exhibit more outliers, suggesting that there are some polls reporting significantly different support levels compared to the majority.
- This plot highlights regional differences in Harris's support, with some states showing much stronger or weaker support than others, and variations within states due to differing poll results.

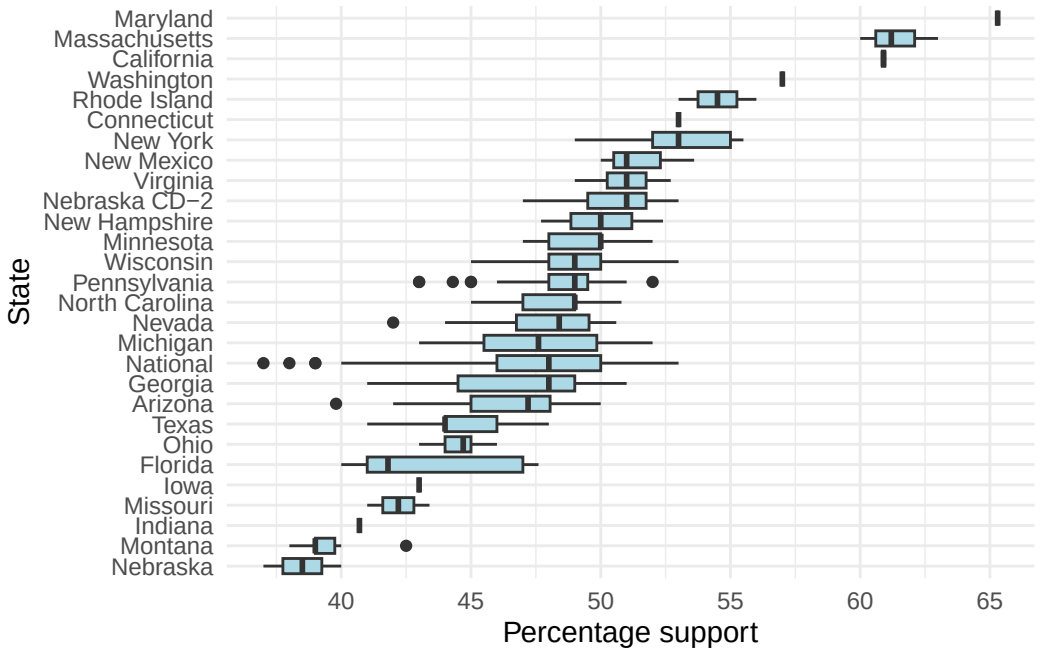


Figure 5: Percentage support for Kamala Harris by state. The distribution of support varies significantly across states, with some states like Nebraska showing lower median support, while others like Maryland and Massachusetts display higher levels.

Figure 6 illustrates the relationship between percentage support for Harris and the poll score, with each point representing an individual poll. The red trend line indicates a slight positive correlation, suggesting that as poll scores improve (become less negative), the percentage support for Harris increases modestly. Despite this trend, the relationship is relatively weak, as there is considerable variation in support across polls, regardless of their scores. Most polls

cluster around lower scores, and the percentage support shows moderate fluctuations within that range.

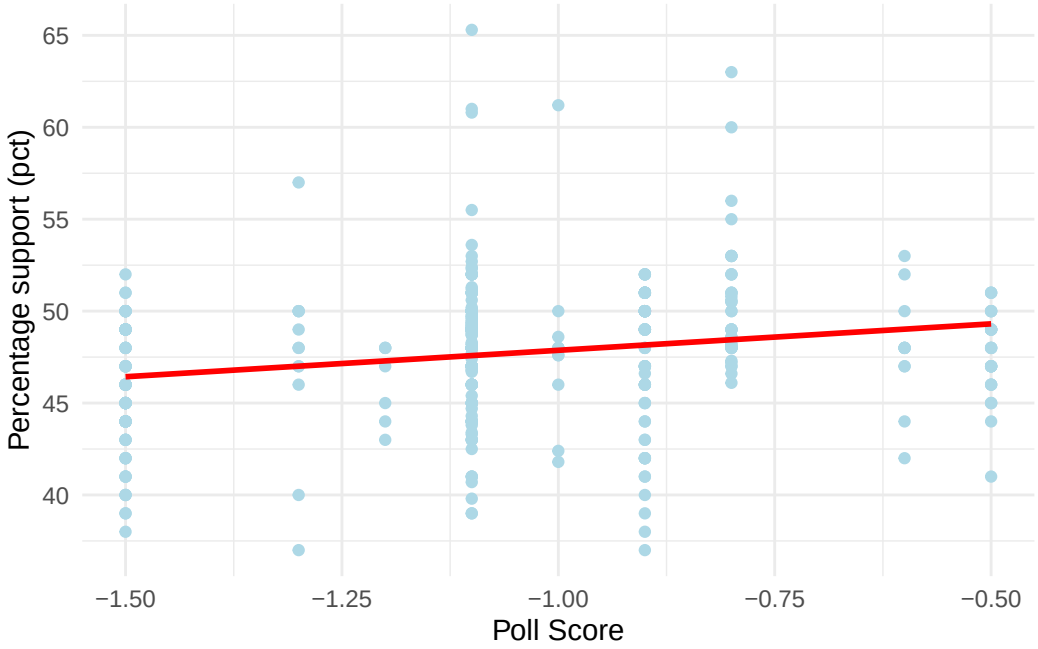


Figure 6: Percentage support for Kamala Harris versus poll score, with a fitted trend line. The positive slope of the trend line suggests a slight increase in support as poll scores improve, though slightly.

3 Model

The goal of our modeling strategy is twofold. Firstly, we aim to quantify the relationship between key predictor variables, such as the end date of the poll, the pollster, the state, and the poll score, and the percentage support for Harris. Secondly, we seek to assess the predictive power of these variables in explaining the observed variation in percentage support. Here, we briefly describe the linear regression model used to investigate these relationships. The model includes end date, pollster, state, and poll score as predictors, allowing us to evaluate their individual contributions to percentage support. By estimating the coefficients for each variable, we can infer the direction and magnitude of their effects. Background details on model specification, assumptions, and diagnostics are provided in Appendix .1 and Appendix .2. Model validation is presented in Appendix .3.

The modeling decisions align with the data section by treating end date as continuous to capture the linear effect of time on support, preserving detail. Pollster and state are treated

as categorical variables to account for fixed differences between groups without imposing an order. Poll score is modeled as continuous to retain its granularity and reflect the effect of poll quality on support. These choices ensure that key characteristics of the data are represented without loss of information.

3.1 Model set-up

3.1.1 Model 1: Percentage support as a function of end date

$$y_i = \beta_0 + \beta_1 \cdot x_{1i} + \epsilon_i$$
$$\epsilon_i \sim \text{Normal}\left(0, \sigma^2\right)$$

Where:

- y_i is the percentage support for Harris,
- x_{1i} is the end date of the poll,
- β_0 is the intercept,
- β_1 is the coefficient for the end date,
- ϵ_i is the error term, assumed to follow a Normal distribution with mean 0 and variance σ^2.

3.1.2 Model 2: Percentage support as a function of end date, pollster, state, and poll score

$$y_i = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \beta_3 \cdot x_{3i} + \beta_4 \cdot x_{4i} + \epsilon_i$$
$$\epsilon_i \sim \text{Normal}\left(0, \sigma^2\right)$$

Where:

- y_i is the percentage support for Harris,
- x_{1i} is the end date of the poll,
- x_{2i} is the pollster,
- x_{3i} is the state,
- x_{4i} is the poll score,
- β_0 is the intercept,
- $\beta_1, \beta_2, \beta_3, \beta_4$ are the coefficients for the respective predictor variables,
- ϵ_i is the error term, assumed to follow a Normal distribution with mean 0 and variance σ^2.

We run the models in R (R Core Team 2023).

3.1.3 Model justification

We expect a positive relationship between the timing of the poll, the pollster, the state, and the percentage support for Harris. Specifically, polls conducted closer to election day may show higher support due to increased campaign visibility, while certain pollsters and states may have systematic effects on the results. Additionally, higher poll scores likely reflect better poll quality, which could lead to more accurate measurements of support. By including these variables, the model aims to predict percentage support based on key factors that influence polling outcomes. The linear regression framework provides a clear interpretation of how each predictor contributes to the percentage support.

4 Results

Our results are summarized in Table 1.

In the context of the models analyzed, several predictors significantly influence the percentage support for Harris.

- End Date: The coefficient for the end date predictor indicates that as the polling date approaches the election, the percentage support for Harris tends to increase. This effect is consistent with the notion that heightened campaign visibility and voter engagement closer to election day can lead to more favorable polling results.
- Pollster: The coefficients for different pollsters reveal that some organizations report higher levels of support for Harris compared to others. For instance, pollsters like Mass-INC Polling Group and Siena/NYT show positive coefficients, suggesting that polls conducted by these organizations tend to yield higher support for Harris, possibly due to differences in methodology, sample selection, or political leanings of the pollsters.
- State: The state variable also has a notable impact on the percentage support. Certain states, such as Maryland and California, exhibit higher coefficients, indicating that voters in these regions are more likely to support Harris compared to others. This reflects regional political dynamics, demographics, and historical voting patterns that influence public opinion.
- Poll Score: The poll score is another critical predictor. A higher poll score, which reflects better quality and reliability of the poll, is associated with increased percentage support for Harris. This implies that more credible polls are likely to provide a more accurate reflection of public sentiment.

Overall, the combination of these predictors provides a multifaceted understanding of the factors influencing the percentage support for Harris. Each predictor contributes uniquely to the model, allowing for a nuanced interpretation of how different elements impact voter sentiment leading up to the election.

Table 1: Model results

	Model 1	Model 2
(Intercept)	−416.971	−564.687
	(132.353)	(97.427)
end_date	0.023	0.031
	(0.007)	(0.005)
pollsterBeacon/Shaw		0.293
		(0.686)
pollsterChristopher Newport U.		−0.362
		(3.401)
pollsterCNN/SSRS		−0.490
		(0.937)
pollsterData Orbital		−0.905
		(2.460)
pollsterEchelon Insights		0.561
		(1.144)
pollsterEmerson		0.644
		(0.662)
pollsterIpsos		−1.342
		(0.702)
pollsterMarist		0.807
		(0.732)
pollsterMarquette Law School		0.197
		(0.807)
pollsterMassINC Polling Group		1.667
		(1.253)
pollsterMcCourtney Institute/YouGov		−0.812
		(2.439)
pollsterMuhlenberg		−1.299
		(2.448)
pollsterQuinnipiac		−0.952
		(0.703)
pollsterSelzer		−3.966
		(2.460)
pollsterSiena		−1.626
		(2.730)
pollsterSiena/NYT		−1.793
		(0.618)
pollsterSuffolk		−1.233
		(1.162)
pollsterSurveyUSA		−1.173

5 Discussion

5.1 Understanding the Predictive Power of Key Variables

The primary goal of our analysis is to accurately predict the percentage support for Harris in the upcoming U.S. election. The results from our models reveal significant insights into how various predictors impact voter sentiment. For instance, the end date has a positive effect on support, indicating that as the election approaches, public support for Harris increases. This finding aligns with expectations that heightened campaign activity and media attention lead to greater voter engagement.

The inclusion of the pollster variable highlights the disparities in polling outcomes based on the organization conducting the survey. Certain pollsters consistently report higher levels of support for Harris, suggesting that methodological differences—such as sampling techniques and demographic targeting—play a crucial role in shaping poll results. This insight emphasizes the importance of critically evaluating polling data sources when interpreting public sentiment.

Furthermore, the state variable reveals regional variations in support, with states like Maryland and California showing significantly higher coefficients. These differences can be attributed to local political climates, demographic factors, and historical voting patterns. Understanding these nuances is vital for strategizing campaign efforts and resource allocation in key battleground states.

5.2 The Importance of Quality Indicators

Another critical aspect of our analysis is the influence of the poll score on percentage support. The negative correlation between lower-quality polls and support underscores the importance of relying on credible and well-designed surveys to gauge public opinion accurately. Higher poll scores are associated with more reliable predictions, emphasizing the need for voters and analysts alike to scrutinize the methodology behind polling data.

This finding has broader implications for the electoral process, as inaccurate polling can lead to misguided campaign strategies. As such, understanding which pollsters produce more reliable data can aid campaigns in making informed decisions about where to focus their efforts in the lead-up to the election. By highlighting these factors, our models not only provide insights into current voter sentiment but also underscore the importance of quality polling in electoral forecasting.

5.3 Implications for Campaign Strategy

The results of our analysis offer valuable insights for campaign strategy. The positive impact of the end date on percentage support suggests that campaigns should intensify their outreach

efforts as the election draws closer. Engaging voters through targeted messaging and increased visibility can potentially sway undecided voters and solidify support among existing backers.

Additionally, the regional variations in support indicate the need for tailored messaging that resonates with specific demographics in different states. Campaigns should focus on understanding local issues and concerns to effectively address voters' needs, thereby enhancing their overall appeal.

Moreover, by recognizing the variability introduced by different pollsters, campaign teams can choose their data sources wisely to inform their strategies and avoid overreliance on potentially biased or low-quality polling data.

5.4 Weaknesses and next steps

Despite the insights gained, our analysis has some weaknesses. The model may not account for interaction effects between predictors, such as how the impact of the end date might vary across different states or pollsters. Additionally, there could be unobserved variables that influence voter sentiment but are not included in the model, leading to omitted variable bias.

Moving forward, future research should aim to explore these interaction effects and incorporate a broader range of predictors, such as voter demographics and sentiment analysis from social media. Implementing a mixed-methods approach could also enhance the robustness of our findings by integrating qualitative data on voter attitudes.

Furthermore, as the election approaches, it will be essential to continually update the models with new polling data to ensure that predictions remain relevant. By addressing these weaknesses and adapting our modeling strategy, we can improve our understanding of the factors influencing percentage support for Harris in the dynamic landscape of the upcoming election.

Appendix

.1 Model details

The linear regression model assumes a linear relationship between these predictors and percentage support, independent observations, and homoscedasticity (constant variance of errors). Additionally, the residuals are expected to follow a normal distribution. If these assumptions are violated—such as non-linear relationships or heteroscedasticity—the model's estimates may be biased. Potential limitations include ignoring interaction effects between the predictors, and the risk of omitted variable bias if important factors are not included. The model may not be appropriate in cases where non-linearity or outliers dominate the data, necessitating alternative methods, such as using the raw counts instead of percentage support for Harris.

.2 Model diagnostics

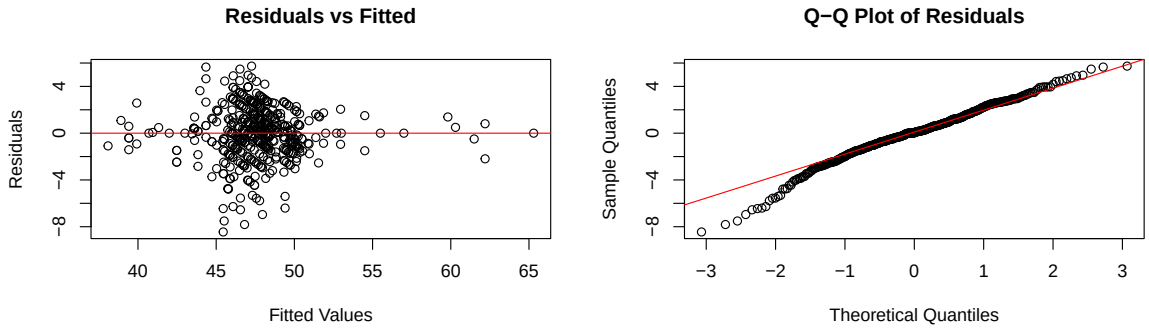
Figure 7a is a residuals versus fitted values plot. It shows no pattern in the residuals, indicating that the assumption of linearity holds true. This suggests that the linear regression model is appropriately specified and that the relationship between the predictors and the outcome variable is adequately captured without systematic bias. The absence of a discernible pattern also implies that the model does not suffer from issues such as heteroscedasticity, where the variance of the errors would change with fitted values.

Figure 7b is a Normal Q-Q plot of residuals. It shows slight deviation from the line, particularly in the tails of the distribution. This suggests that while the residuals are generally normally distributed, there may be minor departures from normality, especially for extreme values. These deviations could indicate the presence of outliers or non-normality in the residuals, which may affect the validity of hypothesis tests and confidence intervals derived from the model. It may be worth exploring further to ensure that the assumptions of normality are sufficiently met for accurate inference.

.3 Model validation

Root Mean Square Error (RMSE) is a standard way to measure the error of a model in predicting quantitative data. It represents the square root of the average squared differences between predicted and observed values. A lower RMSE value indicates a better fit of the model to the data, with the unit of measurement the same as the dependent variable (in this case, percentage support). In the models provided, Model 1 has an RMSE of 3.52, while Model 2 has a significantly lower RMSE of 2.22, indicating that Model 2 provides more accurate predictions of percentage support.

Adjusted R^2 adjusts the R^2 value for the number of predictors in the model, providing a more accurate measure of model performance, particularly when multiple predictors are involved.



(a) Residuals vs. fitted values

(b) Normal Q-Q plot of residuals

Figure 7: Assessing model assumptions, Model 2

It penalizes the addition of non-significant predictors that do not improve the model. Model 1 has an adjusted R^2 of 0.024 , suggesting that it explains very little of the variance in the data. In contrast, Model 2 has a much higher adjusted R^2 of 0.564 , indicating that a substantial portion of the variance in percentage support is explained by the predictors included in the model.

When comparing the two models, Model 1 includes only the end date as a predictor and performs poorly, as indicated by its low R^2, adjusted R^2, and higher RMSE. Conversely, Model 2 incorporates multiple predictors, including pollster, state, and poll score, which results in a significantly better fit, as shown by its high R^2 (0.614) and lower RMSE (2.22). The lower AIC (2166.2) and BIC (2393.8) values for Model 2 further reinforce its superiority by indicating a more parsimonious model with better predictive accuracy. Overall, the enhanced performance of Model 2 suggests that the additional variables contribute meaningfully to understanding the factors influencing percentage support for Harris.

.4 Methodology of Suffolk University Political Research Center

The Suffolk University Political Research Center (SUPRC) is a well-regarded pollster known for conducting national and regional surveys to measure public opinion on a wide range of political issues (Center 2024). With a reputation for methodological rigor, SUPRC has become a notable source of insights into voter sentiment, election forecasting, and public attitudes on pressing political topics. This report provides a comprehensive examination of SUPRC's methodology, highlighting key features of its sampling approach, questionnaire design, and data collection techniques, as well as the strengths and potential limitations of their polling process. Such an analysis is essential for understanding the reliability and validity of the results SUPRC reports, especially as polling data plays an influential role in shaping media coverage, public perceptions, and political strategies.

Target Population and Sampling Frame

SUPRC targets the general adult population of the United States, with a specific focus on registered voters as the primary population for national polls. For state-level or regional surveys, SUPRC may adjust the sample to reflect the voting-eligible population in those specific areas. The sampling frame used by SUPRC is derived from lists of registered voters, which are obtained from state voter registration databases and supplemented by other public and private sources. These databases often provide demographic information that allows SUPRC to stratify their sample by key demographic variables, including age, gender, race, and geographic location. By incorporating these factors, SUPRC can design samples that are broadly representative of the electorate, ensuring that the results reflect a comprehensive cross-section of public opinion.

To further improve representativeness, SUPRC employs demographic quotas during the sampling process. This involves setting target numbers for respondents from different demographic groups based on their prevalence in the broader population. By reaching specific targets for underrepresented groups, SUPRC can reduce the risk of skewed results and enhance the reliability of their findings. For example, if a state has a high percentage of younger or Hispanic voters, SUPRC ensures that these groups are adequately represented in their sample, which helps to capture a more accurate picture of local sentiment.

Mixed-Mode Sampling Approach

SUPRC utilizes a mixed-mode sampling approach, combining both landline and cell phone surveys to reach respondents. This dual-mode strategy is particularly important in an era where mobile phones have become the primary communication tool for many, especially younger and urban voters. Landline-only surveys can introduce biases by excluding segments of the population that are less likely to have landlines, such as young adults and low-income households. By including cell phones in their sampling frame, SUPRC can mitigate this bias and enhance the inclusiveness of their sample.

However, this approach comes with trade-offs. Although telephone interviews generally yield higher response rates compared to online or mail surveys, they are also vulnerable to non-response bias if certain groups are less likely to answer calls from unknown numbers. Additionally, the timing of calls can impact who is available to participate; for instance, daytime calls may disproportionately reach retirees or those not working traditional hours. SUPRC attempts to balance these factors by scheduling calls at different times of day and on different days of the week to increase the likelihood of reaching a diverse respondent pool.

Moreover, reliance on self-reported data in phone surveys can introduce potential inaccuracies, such as social desirability bias, where respondents may tailor their answers to appear more socially acceptable. For instance, respondents might be hesitant to share controversial opinions on hot-button issues like immigration or law enforcement. SUPRC acknowledges these limitations and seeks to minimize bias through neutral question phrasing and assurance of respondent anonymity, encouraging more honest responses.

Addressing Non-Response and Weighting

Non-response is a challenge for most pollsters, and SUPRC employs several strategies to address it. To improve response rates, SUPRC makes multiple attempts to reach each selected respondent, calling at various times of day and using follow-up calls when initial attempts are unsuccessful. If particular demographic groups are underrepresented in the completed sample, SUPRC may conduct targeted outreach to ensure adequate representation. For instance, if early survey results show that young voters are underrepresented, SUPRC may prioritize reaching additional respondents within that age group.

Once data collection is complete, SUPRC applies weighting techniques to align the sample with known population parameters. This involves adjusting the results to correct for any over- or under-representation of demographic characteristics such as age, gender, race, and education. Weighting ensures that the sample is more representative of the target population and reduces the impact of potential sampling bias. For example, if the final sample contains a disproportionately high number of college-educated respondents, SUPRC would adjust the weights to better match the actual educational composition of the broader population.

However, while weighting can help correct imbalances, it cannot fully eliminate non-response bias if certain groups remain underrepresented or unreachable. For instance, lower-income respondents who may not have stable contact information might still be missing from the sample. As a result, SUPRC acknowledges that even with weighting, some residual bias may persist, particularly in hard-to-reach populations.

Questionnaire Design and Question Framing

SUPRC's questionnaires are carefully crafted to capture a wide range of political opinions and concerns, covering topics such as candidate favorability, issue importance, and voter priorities. The questions are designed to be clear and straightforward, reducing the likelihood of respondent confusion and improving data accuracy. SUPRC frequently updates their questions to

reflect current events, ensuring that their surveys remain relevant and timely. For example, during election years, the questionnaire may include additional questions on candidate performance and policy positions.

The framing of questions is critical, as even subtle wording differences can influence responses. SUPRC takes measures to ensure neutrality in question wording to avoid leading respondents toward particular answers. For instance, instead of asking, "Do you think Candidate X is doing a good job?", SUPRC might ask, "How would you rate the performance of Candidate X?" This phrasing minimizes the potential for bias and allows respondents to answer based on their own interpretation.

SUPRC also incorporates open-ended questions that allow respondents to elaborate on their views in their own words. This approach provides richer qualitative data, revealing insights that may not emerge from closed-ended questions. For example, instead of simply asking whether a respondent supports a policy, an open-ended question might capture their reasons, revealing underlying values or concerns. However, analyzing open-ended responses is labor-intensive and can be challenging to quantify, making it difficult to incorporate into the broader dataset.

Strengths and Limitations

SUPRC's methodological rigor has several strengths. Their representative sampling and mixed-mode approach allow them to reach a broad spectrum of the population, while weighting adjustments further enhance accuracy. The inclusion of open-ended questions provides depth, offering a more nuanced understanding of public opinion. Additionally, SUPRC's commitment to updating question content ensures their surveys reflect current political and social contexts.

However, like all polling methodologies, SUPRC's approach has inherent limitations. Telephone-based surveys are subject to non-response and social desirability biases, which may affect data accuracy. The mixed-mode strategy, while enhancing reach, cannot entirely eliminate these issues. The reliance on self-reported data also introduces potential inaccuracies, as respondents may inadvertently misstate their opinions or behaviors. Furthermore, despite rigorous weighting, some level of non-response bias may remain if certain demographic groups are systematically underrepresented.

In conclusion, the Suffolk University Political Research Center employs a robust methodology aimed at capturing a representative snapshot of public opinion across the United States. While their approach addresses many of the challenges inherent in polling, understanding the nuances of their methodology is essential for interpreting SUPRC's findings accurately. Recognizing both the strengths and limitations of their methods enables a more informed assessment of their polls' reliability, contributing to a broader understanding of public sentiment and its implications for political discourse and election outcomes.

.5 Idealized Methodology for Forecasting the U.S. Presidential Election

.5.1 Overview of Methodology

With a budget of $100,000 allocated for forecasting the U.S. presidential election, the primary goal is to develop a methodology that captures a representative snapshot of the electorate's preferences, accounting for demographic and geographic variations. This methodology involves a systematic sampling approach, multi-channel respondent recruitment, rigorous data validation, carefully crafted survey design, and sophisticated poll aggregation. This robust approach aims to ensure that the results accurately reflect public opinion across diverse groups, providing valuable insights into voter sentiment and election dynamics.

.5.2 Sampling Approach

A well-constructed sampling approach is critical to ensure that the survey results accurately represent the voting population. The ideal approach is stratified random sampling, which segments the population into distinct subgroups (or strata) based on key demographic variables. This method enables proportional representation of each subgroup in the sample, reducing sampling bias and enhancing the precision of the estimates.

1. Target Population:

The target population is registered voters in the United States, as these individuals are most likely to participate in the election. While this excludes non-registered adults, focusing on registered voters provides a more accurate reflection of the likely electorate.

2. Sampling Frame:

Constructing a reliable sampling frame involves obtaining voter registration databases from state election offices. These databases are supplemented with demographic data from sources like the U.S. Census Bureau to include additional variables such as age, gender, race, and income level. This enhances the comprehensiveness of the frame, ensuring it covers all relevant segments of the population.

3. Strata Variables:

Key demographic variables for stratification include:

- Age: Age groups, e.g., 18-29, 30-44, 45-59, 60+, to capture generational differences.
- Gender: Male, Female, and non-binary options to reflect gender diversity.
- Race/Ethnicity: Categories based on census data, e.g., White, Black, Hispanic, Asian, etc.
- Income Level: Income brackets to represent socio-economic diversity.
- Geographic Location: Urban, suburban, and rural regions, as well as states or regions (e.g., Northeast, South), to capture geographic diversity.

4. Sample Size:

A sample size of approximately 10,000 respondents is targeted, providing a margin of error around ±1% at a 95% confidence level. This sample size allows for a granular breakdown of demographic and geographic subgroups, enhancing the reliability of state-level predictions and providing robust insights into demographic trends.

5. Budget Allocation for Sampling:

With the $100,000 budget, costs are allocated for:

- Data acquisition (e.g., voter registration databases)
- Survey design and distribution
- Data cleaning and analysis
- Compensation for respondents (if required, to boost response rates)

.5.3 Respondent Recruitment

Ensuring a diverse and representative sample requires recruiting respondents through multiple channels. This multi-channel approach reaches underrepresented groups and mitigates the limitations of any single recruitment method.

1. Online Panels:

Partnering with established survey firms that maintain large, demographically diverse online panels is a cost-effective way to recruit respondents. These firms can target specific demographic characteristics, enabling oversampling of hard-to-reach groups if necessary.

2. Social Media Campaigns:

Platforms like Facebook, Instagram, and Twitter offer advanced targeting options that can help reach particular demographics, such as younger voters or those in rural areas. Paid ads can target users based on age, location, and interests, improving recruitment efficiency and reach.

3.Community Outreach:

Collaborating with community organizations (e.g., local nonprofits, cultural centers) can increase participation among underrepresented groups, such as minorities or low-income populations. These organizations can help promote the survey, fostering trust and encouraging engagement among community members.

4. Telephone Recruitment:

Although more costly, telephone recruitment can be effective for reaching populations that may be less active online, such as older adults. Combining landline and cell phone contacts helps capture a broader audience, especially for those less likely to engage with online surveys.

.5.4 Data Validation

To ensure data quality and accuracy, multiple validation techniques will be implemented during and after data collection.

1. Response Validation:

The survey will include logic checks to flag inconsistencies. For instance, if a respondent indicates they are a registered voter but also states they did not vote in the last presidential election, their response may be flagged. Such discrepancies will be reviewed to maintain data integrity.

2. Demographic Verification:

After data collection, responses will be cross-checked against demographic benchmarks (e.g., census data) to verify that the sample aligns with the broader population on critical variables. Adjustments can be made if necessary, such as weighting underrepresented groups to improve representativeness.

3. Follow-Up Surveys:

Conducting follow-up surveys with a subset of respondents can provide additional validation by confirming initial responses and checking for consistency over time. This also offers an opportunity to assess potential biases introduced by attrition.

4. Duplicate Detection and Fraud Prevention:

To prevent fraudulent or duplicate entries, unique identifiers (e.g., IP addresses, cookies) will be used to track responses and avoid multiple entries from the same individual. Anti-fraud measures ensure that the data represents unique individuals, enhancing reliability.

.5.5 Survey Structure

The survey is designed with a user-friendly layout to encourage completion and minimize dropout rates. Key sections include:

1. Introduction:

A brief introduction explains the purpose of the survey, provides assurance of anonymity, and outlines data privacy protections. This section builds trust and increases participant comfort.

2. Demographic Questions:

Basic demographic questions (age, gender, education, income, etc.) provide essential information for analysis and weighting, ensuring that responses reflect the intended sample characteristics.

3. Voting Behavior Questions:

Questions about past voting behavior and future voting intentions (e.g., "Did you vote in the last presidential election?", "How likely are you to vote in the upcoming election?") help gauge voter engagement and likelihood of turnout.

4. Candidate Preference Questions:

Core questions focus on voter preferences for candidates, such as "If the election were held today, who would you vote for?" These questions allow for straightforward measurement of current support levels.

5. Opinion Questions on Key Issues:

These questions address critical issues like the economy, healthcare, and climate change. Responses to these issues provide context for understanding the factors driving candidate support.

6. Conclusion:

A closing message thanks respondents for their participation and outlines how the results will be used, fostering transparency and trust.

.5.6 Poll Aggregation

To enhance the predictive power of the survey results, a poll aggregation strategy will be used, combining our survey data with existing polling data from reputable sources. Poll aggregation allows us to incorporate data from multiple sources, creating a more comprehensive and stable forecast.

1. Data Weighting and Reliability Scoring:

Each poll is weighted based on reliability factors such as sample size, methodology, and historical accuracy of the pollster. More reliable sources receive higher weights, ensuring that our predictions prioritize high-quality data.

2. Statistical Techniques for Aggregation:

Methods like Bayesian modeling or weighted averaging are employed to integrate data from various polls. Bayesian modeling allows us to account for prior information and adjust predictions as new data arrives. Weighted averages provide a straightforward method of combining polls, balancing recent trends with overall stability.

3. Trend Analysis and Model Calibration:

To detect shifts in public opinion, trend analysis is applied to the aggregated data. The model is periodically calibrated to ensure it reflects the latest polling data, maintaining its predictive accuracy.

4. Uncertainty Estimation:

Given the inherent uncertainty in polling, the final forecast includes confidence intervals, which reflect the range within which the true support level is likely to fall. This provides a realistic view of the forecast's precision and accounts for sampling variability.

.5.7 Summary

In summary, this detailed methodology for a $100,000 election forecast project combines rigorous sampling, multi-channel recruitment, robust validation, and advanced poll aggregation to produce a reliable and nuanced understanding of voter preferences. The approach addresses potential sources of bias and enhances the representativeness of the findings, providing valuable insights into the electoral landscape. This comprehensive framework supports evidence-based analysis, informing public discourse and campaign strategies alike.

.5.8 Implementation

The survey was created using Google Forms. Survey link: https://forms.gle/8GWgVWEtrrnL59mT7

References

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

Center, Suffolk University Political Research. 2024. "National Issues Poll with USA TODAY." https://www.suffolk.edu/academics/research-at-suffolk/political-research-center/polls/national#collapse-October-21-2024-National-Issues-Poll-with-USA-TODAY.

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

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

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/.

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

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

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

Wickham, Hadley. 2011. "testthat: Get Started with Testing." *The R Journal* 3: 5–10. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

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.