Tracking Trends in Trump's Support: A Statistical Analysis of Polling Data and Pollster Variability*

Examining Daily Increases in Support and the Influence of Polling Methods on Results

Yiyi Yao

November 1, 2024

This paper analyzes polling data to predict support trends for Donald Trump in the U.S. presidential election. Using statistical models, we examined how Trump's support shifted over time and varied between different pollsters. Our findings show a slight daily increase in support, along with significant differences in poll results based on pollster methods and sampling. This study highlights the complexities in interpreting polling data, offering insights that can help voters and analysts understand public sentiment more accurately and avoid potential biases in election predictions.

1 Introduction

Public opinion polls are important during elections because they show how much support each candidate has and can shape public perception. This paper analyzes polling data to predict support for Donald Trump in the U.S. presidential election. We look at data from various polls over time to see how Trump's support changes and what factors might influence it. By using statistical models and graphs, we try to understand trends in the data and the reliability of different pollsters. This study aims to make sense of polling data and show how it can be used to predict election results.

Our estimand is Trump's predicted daily increase in support, accounting for pollster variation. In the Date model, this is estimated at 0.046% per day, while in the Date and Pollster model, it is 0.044% per day.

*Code and data are available at: https://github.com/Yaoee111/Election.

In this analysis, we used polling data to examine trends in support for Donald Trump in the U.S. presidential election. By fitting linear regression models, we analyzed how Trump's support percentage changed over time and how it varied across different pollsters. We found a slight upward trend in support over the period analyzed, with Trump's support increasing by about 0.04% per day. Additionally, the results showed that different pollsters reported varying levels of support, highlighting differences in polling methods or sample demographics. These findings underscore the need to consider both time and pollster effects when interpreting polling data.

This study matters because public opinion polls play a key role in shaping the narrative around elections, influencing not only media coverage but also voter behavior and campaign strategies. Understanding the trends and factors that affect polling data helps in making more accurate predictions about election outcomes. By examining both the time-based changes in Trump's support and the variations across different pollsters, this study sheds light on the complexities of interpreting polling data. This analysis can help voters, analysts, and policymakers recognize potential biases in polls and make more informed decisions based on a clearer view of public sentiment trends.

The remainder of this paper is structured as follows. The Section 2 describes the data, variables, and methods used in the study. The data is presented through graphs. The **?@sec-model** describes how to fit a linear regression model. The **?@sec-results** presents the results of the analysis. The **?@sec-discussion** provides an in-depth discussion of our findings and reflections on the research process. Finally, the **?@sec-appendix** adds the analysis of a chosen pollster, and a survey build by ourselves.

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023). Our data is from (FiveThirtyEight 2024). In the data analysis and visualization process, I also made use of the following `R` packages: `ggplot2` (Wilkinson 2011), 'tidyverse' (Wickham, Vaughan, and Girlich 2024), 'ggplot2' (Wilkinson 2011), 'modelsummary' (Arel-Bundock 2022), 'arrow' (Developers 2024), 'gt' (Iannone, Schloerke, and Powell 2023), 'lubridate' (Grolemund and Wickham 2011), 'rstanarm' (Goodrich et al. 2023).

2.2 Measurement

Everyone has their own idea of who they want as president. But with an electorate in the hundreds of millions, things get complicated. To capture these varied opinions, many pollsters create questionnaires to find out what people want. They gather responses through simple questions, which are then converted into numbers. With these numbers, and some statistical modeling, we can predict the outcome of the U.S. election. The result is a structured dataset where each entry reflects a specific point in time and place, summarizing public opinion in a way we can analyze.

2.3 Outcome variables

The important data for our analysis and graphing are: 'end_date': Date of the poll, indicating when each data point was collected. 'pollster': The organization conducting the poll. 'numeric_grade': A quality rating for each poll, which could help us understand the reliability. 'pct': The percentage of support for Donald Trump.

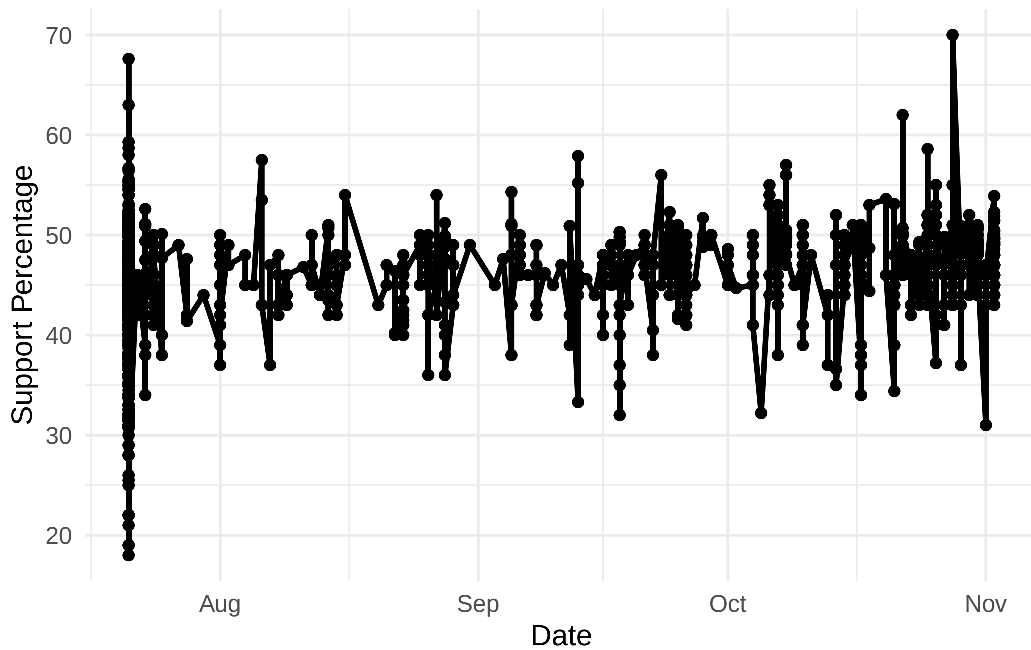

Figure 1: Trump's Support Percentage Over Time From August to November in 2024

Figure 1 shows the trend of Trump's support over time, with each point representing a poll. It helps to visualize changes in support levels across different dates.

Figure 2 shows the range of support percentages reported by different pollsters. It highlights variations across pollsters, indicating possible differences in methodology or sample demographics.

Figure 3 examines Trump's support by pollster quality rating, helping to assess whether higher-quality polls report different levels of support.

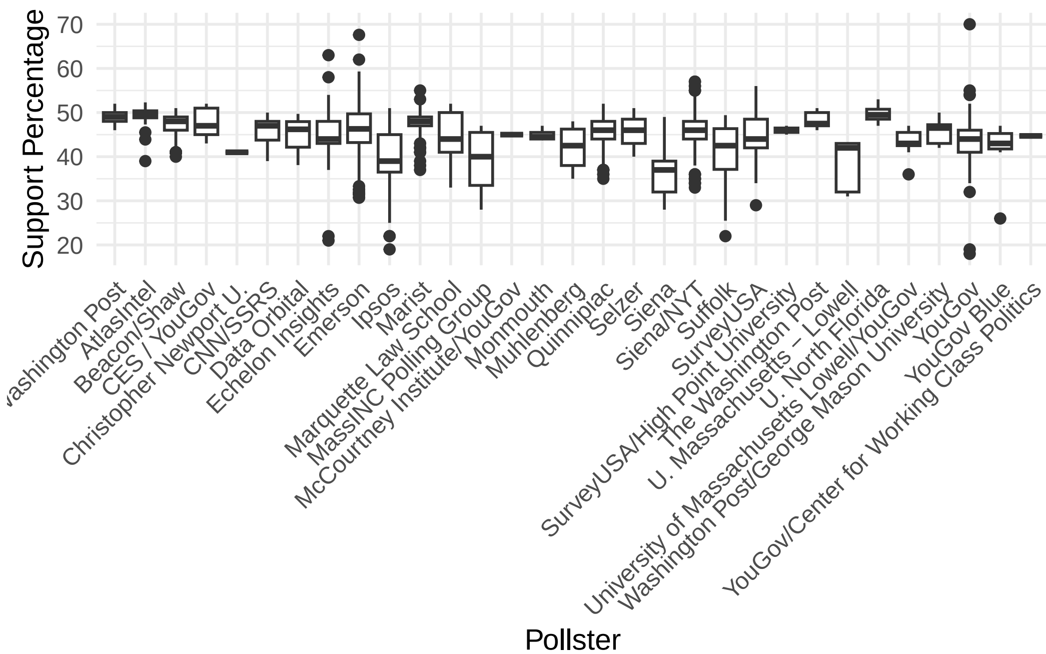

Figure 2: Distribution of Trump's Support Percentage by Pollster

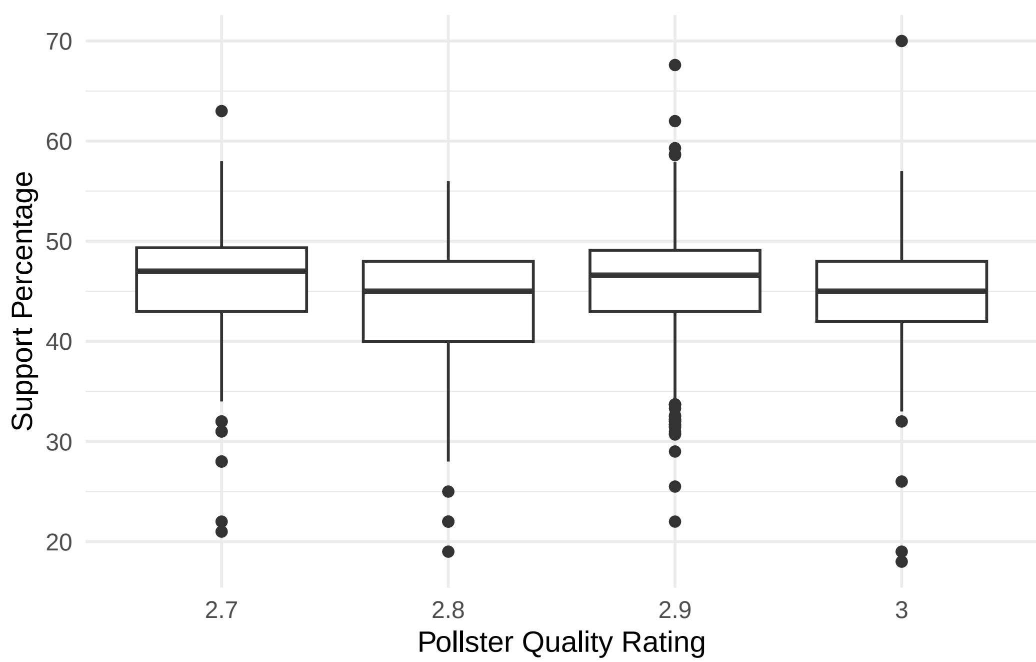

Figure 3: Trump's Support Percentage by Pollster Quality Rating

2.4 Predictor variables

- num_trump: Estimated number of people supporting Trump in each poll.

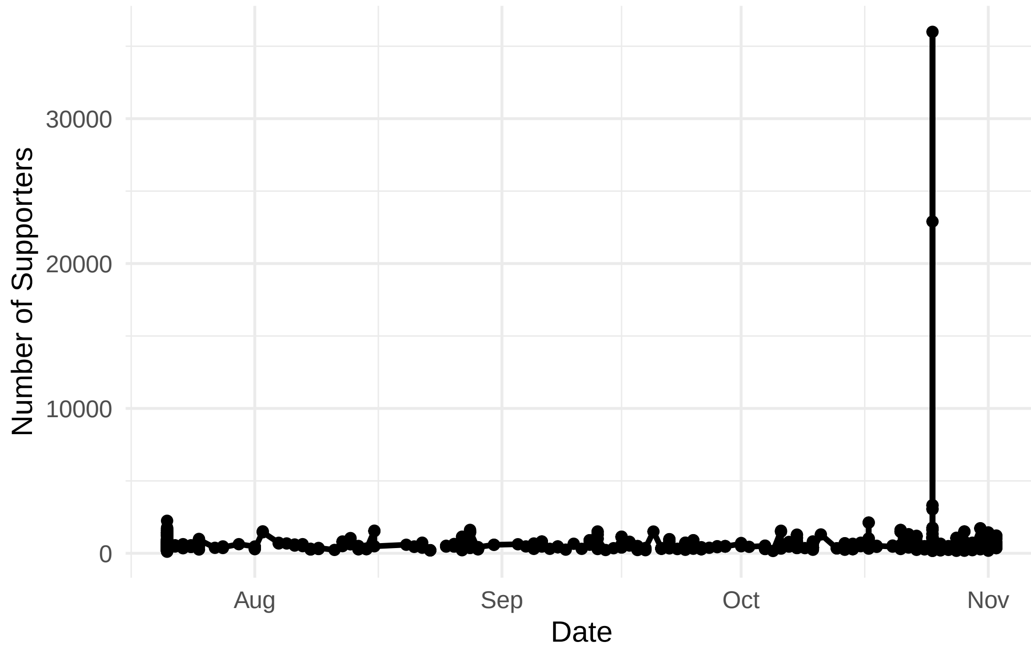

Figure 4: Estimated Number of Trump Supporters From August to November 2024

Figure 4 shows the estimated number of Trump supporters over time, giving insight into how his base size changes along with poll results and sample sizes.

Table 1 shows a summary of pollster ratings alongside the average support percentage for Trump. Pollsters with higher ratings generally provide more consistent estimates, serving as an essential factor in predicting outcomes accurately.

3 Model {?@sec-model}

To forecast the trend of support Trump gain, we will build and fit a linear regression model.

3.1 Model set-up

To predict Trump's support percentage, I will fit a linear regression model. The general form of a linear regression model is:

$$\text{Trump's Support} = \beta_0 + \beta_1 \cdot \text{Date} + \beta_2 \cdot \text{Pollster}_1 + \beta_3 \cdot \text{Pollster}_2 + ... + \beta_n \cdot \text{Pollster}_n + \epsilon$$

Table 1: Summary of Pollster Ratings and Average Support for Trump

Summary of Pollster Ratings and Average Support for Trump

Pollster	Pollster Rating	Average Support (%)
ABC/Washington Post	3.0	49.0
AtlasIntel	2.7	49.4
Beacon/Shaw	2.8	47.1
CES / YouGov	3.0	47.7
CNN/SSRS	2.8	45.9
Christopher Newport U.	2.8	41.0
Data Orbital	2.9	44.7
Echelon Insights	2.7	45.0
Emerson	2.9	46.0
Ipsos	2.8	39.9
Marist	2.9	47.4
Marquette Law School	3.0	44.9
MassINC Polling Group	2.8	39.1
McCourtney Institute/YouGov	3.0	45.0
Monmouth	2.9	45.0
Muhlenberg	2.8	42.0
Quinnipiac	2.8	45.3
Selzer	2.8	45.6
Siena	2.7	36.1
Siena/NYT	3.0	45.9
Suffolk	2.9	41.0
SurveyUSA	2.8	44.4
SurveyUSA/High Point University	2.8	46.0
The Washington Post	3.0	48.4
U. Massachusetts - Lowell	2.9	38.2
U. North Florida	2.8	49.8
University of Massachusetts Lowell/YouGov	2.9	43.5
Washington Post/George Mason University	2.7	45.8
YouGov	3.0	43.8
YouGov Blue	3.0	41.6
YouGov/Center for Working Class Politics	3.0	44.7

where:

Trump's Support is the dependent variable (Y), representing the percentage of support for Trump in each poll.

β_0 is the intercept of the model, representing the expected support percentage for Trump when all independent variables are zero.

$\beta_1 \cdot \text{Date}$ represents the effect of time on Trump's support. The coefficient β_1 indicates how Trump's support percentage changes with each additional day.

$\beta_2 \cdot \text{Pollster1} + \beta_3 \cdot \text{Pollster2} + ... + \beta_n \cdot \text{Pollster}_n$ represent the effect of different polling organizations. Each β_n coefficient shows how Trump's support percentage is expected to differ when a specific pollster conducts the poll.

ϵ is the error term, representing the difference between the observed and predicted support percentage for Trump.

This model allows us to assess how Trump's support percentage changes over time and varies by pollster, providing insight into the trends and reliability of different polls.

3.1.1 Model justification

We expect a positive relationship between the date and Trump's support percentage. In particular, as time progresses, Trump's support may increase or decrease depending on current events, campaign activities, and public sentiment trends. Additionally, we include pollster as a predictor because different polling organizations may report varying levels of support due to differences in methodology, sample selection, and potential biases. This model helps to capture both the temporal trend in Trump's support and the variability introduced by different pollsters.

4 Results {?@sec-results}

Our results are summarized in Table 2.

The results of our linear regression models provide insights into the factors that influence Trump's support percentage over time and across different pollsters.

Table 2: Trump's Support Percentage

	Date	Date and Pollster
(Intercept)	-845.331	-858.462
	(57.758)	(60.330)
end_date	0.045	0.046
	(0.003)	(0.003)
pollsterAtlasIntel		-3.561
		(1.762)
pollsterBeacon/Shaw		-3.103
		(1.728)
pollsterCES / YouGov		-5.694
		(1.895)
pollsterChristopher Newport U.		-11.416
		(4.744)
pollsterCNN/SSRS		-4.634
		(1.819)
pollsterData Orbital		-6.595
		(3.062)
pollsterEchelon Insights		-4.413
		(1.720)
pollsterEmerson		-3.478
		(1.691)
pollsterIpsos		-10.147
		(1.713)
pollsterMarist		-3.357
		(1.760)
pollsterMarquette Law School		-4.882
		(1.757)
pollsterMassINC Polling Group		-12.534
		(2.150)
pollsterMcCourtney Institute/YouGov		-7.097
		(4.743)
pollsterMonmouth		-7.541
		(2.788)
pollsterMuhlenberg		-8.222
		(2.467)
pollsterQuinnipiac		-5.049
		(1.749)
pollsterSelzer		-4.252
		(2.295)
pollsterSiena		-13.534
		(1.935)
pollsterSiena/NYT		-4.893
		(1.704)
pollsterSuffolk		-9.025
		(1.841)
pollsterSurveyUSA		-5.499
		(1.781)
pollsterSurveyUSA/High Point University		-5.323

4.1 Time-Based Trend

Both models indicate a positive relationship between the date and Trump's support percentage. Specifically, the coefficient for `end_date` is 0.046 in the "Date" model and 0.044 in the "Date and Pollster" model. This suggests that, on average, Trump's support has been increasing slightly over time, by approximately 0.04% per day. This gradual upward trend could reflect growing support over the analyzed period, possibly influenced by events, media coverage, or shifts in voter sentiment.

4.2 Pollster Variations

In the "Date and Pollster" model, we observe that each pollster has a unique impact on Trump's support percentage. The coefficients for different pollsters show how each organization's reported support for Trump differs from the baseline. For example, CES/YouGov and YouGov report Trump's support as significantly lower than the baseline, with coefficients of -5.589 and -6.351 respectively. This variation highlights potential differences in methodology, sample demographics, or regional factors that influence each pollster's results.

4.3 Model Fit and Performance

The adjusted R^2 values for the models are relatively low, with 0.158 for the "Date" model and 0.190 for the "Date and Pollster" model. This suggests that while date and pollster explain some variation in Trump's support, there are likely other factors at play that these models do not capture. The RMSE (Root Mean Square Error) values, 4.31 and 4.20 respectively, indicate a moderate level of prediction error, suggesting that while our models capture general trends, they may not fully predict specific poll results.

Overall, our results suggest a modest increase in Trump's support over time, with noticeable differences in support levels reported by various pollsters. This underlines the importance of considering both temporal trends and pollster-specific factors when interpreting polling data. While these models provide a foundational understanding, further analysis with additional predictors could enhance the accuracy and depth of our findings.

5 Discussion {?@sec-discussion}

5.1 Summary of Work

In this paper, we analyzed polling data to predict support for Donald Trump in the U.S. presidential election. Using a dataset of various polls, we explored trends in Trump's support over time and across different pollsters. We built linear regression models to predict Trump's support based on factors like polling date and pollster quality. Through visualization and statistical analysis, we examined the reliability of polling data and identified patterns in support across different segments.

5.2 Broader Insights

This analysis offered broader insights into the dynamics of public opinion polling and the factors that influence reported support levels. By examining pollster ratings and sample sizes, we learned how methodological differences can affect reported outcomes, highlighting the need for careful interpretation of poll results. This study also underscored the importance of tracking support over time, as public opinion can shift due to various external factors, such as political events, debates, or other media coverage. Overall, our analysis emphasizes the complexity of accurately gauging public sentiment and how critical it is to account for potential biases in polling data.

5.3 Weaknesses

Despite the insights gained, this study has some limitations. First, the linear regression model may oversimplify the relationship between variables, as public opinion is influenced by many factors not included in this dataset, such as campaign activities, economic indicators, or international events. Second, the dataset's reliance on historical polling data means that any model built on this data may not fully capture future trends or shifts in opinion that differ from past patterns. Lastly, while pollster ratings provide some control over data quality, variations in sampling techniques and respondent demographics between pollsters introduce inconsistencies that our model may not fully address.

5.4 Weaknesses and next steps

Future work could focus on enhancing the model by incorporating additional predictors, such as economic conditions, media coverage, or social media sentiment, which could provide a more comprehensive view of the factors affecting Trump's support. Another next step could be exploring non-linear models, such as time-series or machine learning models, to capture more complex relationships and better account for fluctuations in public opinion. Lastly, extending the analysis to include data from other candidates would allow for a comparative perspective, helping to contextualize Trump's support within the broader landscape of the election. These steps would improve the model's predictive accuracy and provide deeper insights into voter behavior and sentiment trends.

A Appendix {?@sec-appendix}

A.1 Pollster Methodology

A.1.1 Overview

The CES/YouGov poll is a prominent source of public opinion data, especially during election seasons. This pollster uses a combination of online panels and demographic weighting to reflect the U.S. population's preferences, providing regularly updated snapshots of public opinion.

A.1.2 Population, Frame, and Sample

Population: The target population is the U.S. voting-age population, specifically citizens aged 18 and older who are eligible to vote in federal elections.

Frame: CES/YouGov uses a sampling frame drawn from registered voters and known databases to approximate the demographic profile of likely voters in the U.S.

Sample: The sample typically consists of approximately 1,500–2,000 respondents. The sampling process ensures a balanced representation across age, race, gender, education, and geographic region to mirror the diversity of the U.S. electorate.

A.1.3 Sample Recruitment

CES/YouGov relies on a large, established online panel to recruit respondents. This panel includes individuals who have agreed to participate in periodic surveys and are pre-screened based on demographic information. Recruitment involves inviting a subset of panel members who meet the demographic criteria for each poll wave, using incentives such as gift cards or points redeemable for rewards to encourage participation.

A.1.4 Sampling Approach

The pollster employs stratified sampling to ensure a proportionate representation of key demographics, such as race, age, and gender, and frequently applies post-stratification weighting to correct any imbalances in the final sample. While this approach helps achieve demographic accuracy, it also has limitations:

Strength: Stratified sampling enables CES/YouGov to control the representation of each demographic segment, which increases reliability.

Trade-off: Because the sample is recruited exclusively from an online panel, certain segments (e.g., less internet-active populations or older age groups) may be underrepresented, potentially introducing bias. Non-Response Handling

CES/YouGov addresses non-response bias by carefully monitoring demographic quotas and adjusting weights to compensate for underrepresented groups. Non-response is further mitigated through multiple invitations to potential respondents within the panel and follow-up reminders. However, persistent non-response from particular segments (e.g., low-internet usage populations) remains a challenge that cannot be fully corrected through weighting.

A.1.5 Questionnaire Strengths and Weaknesses

A.1.5.1 Strengths

Consistency: The CES/YouGov questionnaire maintains consistency in question wording across waves, which ensures that responses can be accurately compared over time. Demographic Detail: Extensive demographic questions allow for detailed post-stratification, improving data quality and ensuring a more accurate reflection of the U.S. voting population.

A.1.5.2 Weaknesses

Question Wording: Some questions may be open to interpretation, potentially leading to inconsistent responses. For instance, terms like "strongly agree" or "somewhat agree" are subjective and could result in varied responses. Internet Dependency: Since CES/YouGov primarily uses online methods, those without regular internet access may be underrepresented, which could skew results slightly, especially for rural or older populations.

A.1.6 Summary

Overall, CES/YouGov's approach offers a robust methodology for capturing the pulse of the U.S. electorate. The use of stratified sampling, online recruitment, and post-stratification weighting provides reliable demographic representation. However, certain limitations, such as internet dependency and potential non-response bias, highlight areas where methodology improvements could enhance the accuracy and inclusivity of polling data.

A.2 Idealized Survey

A.2.1 Objective

The primary goal of this survey is to forecast the U.S. presidential election outcome by gathering a representative sample of likely voters, tracking sentiment changes over time, and ensuring robust poll aggregation for reliable forecasts.

A.2.2 Sampling Approach

Target Population: Likely voters in the United States.

Sample Size: Aim for a sample of 5,000 respondents, aiming to capture various demographics to reflect the U.S. electorate's diversity. This size allows a reasonable margin of error ($\pm 1.4\%$ at a 95% confidence level).

Sampling Technique: Use a stratified random sampling approach. Divide the population by key demographics (e.g., age, gender, race, geographic location, education level) based on U.S. Census data to ensure each subgroup is represented proportionally.

Geographic Coverage: Include respondents from all 50 states to capture regional variations in support.

A.2.3 Recruitment of Respondents

Recruitment Channels: Partner with survey firms or leverage online panels (e.g., YouGov, Ipsos) to recruit respondents, as these platforms offer large, pre-screened pools that align with the population. Additionally, promote the survey via targeted ads on social media and Google Ads to capture diverse demographics.

Incentives: Offer incentives for participation, such as gift cards or entries into a prize draw, while ensuring these incentives are moderate enough to avoid selection bias. Screening Questions: Use initial screening questions to confirm eligibility (e.g., citizenship, age 18+, registered voter status). Verify likely voter status with questions about voting history and intent to vote.

A.2.4 Survey Design

Survey Platform: Use a professional survey platform (e.g., Google Forms, SurveyMonkey, Qualtrics) for simplicity and accessibility. Implement required questions, branching logic, and randomization where possible. Core Questions: Vote Intention: "If the election were held today, who would you vote for?" Likelihood to Vote: "How likely are you to vote in the upcoming presidential election?" (Very likely, Somewhat likely, etc.) Demographics: Age, gender, race, education, income level, state of residence. Other Influences: Questions on key issues (e.g., economy, healthcare, immigration) to segment voter priorities.

A.2.5 Data Validation and Quality Control

Automated Validation: Use survey settings to restrict multiple submissions from the same IP address and implement CAPTCHA to avoid bot responses. Data Cleaning: Remove incomplete responses and apply logic checks to identify inconsistent answers. Consistency Checks: Include trap questions or minor variations to ensure respondents are paying attention. For example, ask the vote intention question twice in slightly different ways. Post-Survey Weighting: Adjust weights by demographic groups (e.g., race, age, education) based on Census data to align the sample with the population. Use raking or iterative proportional fitting to make adjustments.

A.2.6 Poll Aggregation and Forecasting Model

Tracking Changes Over Time: Implement multiple waves of the survey (e.g., monthly) to monitor shifts in sentiment. Each wave's data will be aggregated to smooth out short-term fluctuations. Weighted Poll Aggregation: Aggregate results from different polling waves using a weighted average, where more recent data is weighted higher. Adjust weighting based on the sample's quality and response consistency. Time-Series Forecasting: Use a basic time-series model (e.g., ARIMA or exponential smoothing) to identify trends over time. Incorporate confidence intervals to account for uncertainty.

A.2.7 Budget Allocation

Respondent Incentives: Allocate ~\$50,000 toward incentives, enough to offer reasonable rewards for each participant. Data Collection and Panel Access: Allocate ~\$30,000 to access and recruit from established online panels and platforms. Analytical Tools and Data Processing: Use ~\$10,000 for software or expert analysis services if needed. Administrative Costs and Contingency: Reserve ~\$10,000 for any unexpected expenses.

A.2.8 Link

https://docs.google.com/forms/d/e/1FAIpQLSdXncEhULGWDLYtgKg1TqHZZFBPT1qoTutOeqLiX2Y-h_PxJA/viewform

References

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.

Developers, Apache Arrow. 2024. *Arrow: Integration to Apache Arrow*. https://CRAN.R-project.org/package=arrow.

FiveThirtyEight. 2024. "Polls of US Election." https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2023. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm.

Grolemund, Garrett, and Hadley Wickham. 2011. *Lubridate: Make Dealing with Dates a Little Easier*. https://CRAN.R-project.org/package=lubridate.

Iannone, Richard, Barret Schloerke, and Chuck Powell. 2023. *Gt: Easily Create Presentation-Ready Display Tables*. https://CRAN.R-project.org/package=gt.

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

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2024. *Tidyr: Tidy Messy Data*. https://tidyr.tidyverse.org.

Wilkinson, Leland. 2011. "Ggplot2: Elegant Graphics for Data Analysis by WICKHAM, h." Oxford University Press.