Forecasting the 2024 U.S. Presidential Election*

Modeling State-Level Polling Data to Forecast the Electoral College Outcome

Tim Chen Steven Li Tommy Fu

October 31, 2024

We forecast the outcome of the 2024 U.S. Presidential Election between Kamala Harris and Donald Trump by developing multiple linear regression models based on comprehensive polling data collected throughout the election cycle. Incorporating variables such as state-level demographics, pollster reliability scores, transparency scores, and sample sizes, we applied the same statistical model to both candidates to ensure consistent comparison. Our analysis predicts that Kamala Harris will receive 216 electoral votes, while Donald Trump is projected to secure 147 electoral votes. Neither candidate achieves the 270 electoral votes required to win the presidency, highlighting the potential for a closely contested election. These findings underscore the significant impact of state-specific factors on voter support and suggest that neither candidate currently holds a decisive advantage. We recommend that future research includes dynamic modeling techniques and additional predictive variables, such as economic indicators and voter turnout rates, to enhance the accuracy of election forecasts. Our study emphasizes the complexities involved in electoral predictions and the necessity of balancing multiple factors in policy design and electoral analysis.

Table of contents

*Code and data are available at: https://github.com/timchen0326/US_presidential_election_forecast_2024.git.

1 Introduction

The 2024 United States presidential election presents unprecedented challenges for electoral forecasting. As the country navigates increasing political polarization and evolving voting patterns, the reliability of traditional polling methods has come under intense scrutiny (Viala-Gaudefroy 2024). The task of predicting voter behavior in America's diverse electorate is complicated by numerous factors, including shifting public opinion, rapidly changing political landscapes, and varying levels of voter engagement across different demographic groups.

Recent history has highlighted the complexities of election forecasting. The polling failures in 2016 and 2020—where polls significantly underestimated Republican support in key states—have prompted a fundamental reassessment of polling methodologies (Keeter 2024). These challenges are particularly acute in swing states, where margins of victory are often razor-thin and can determine the outcome of the entire election. The American Association for Public Opinion Research (AAPOR) identified several critical factors contributing to these polling errors, including the underrepresentation of Republican voters and difficulties in predicting voter turnout patterns (Viala-Gaudefroy 2024).

Survey methodology plays a crucial role in addressing these challenges. Well-designed surveys require careful consideration of sampling strategies, questionnaire design, and data collection methods to ensure representative results. As Keeter (2024) emphasizes, pollsters must now employ sophisticated weighting procedures and rigorous quality controls to overcome declining response rates and potential partisan non-response bias. Understanding the strengths and limitations of different polling approaches—from traditional probability sampling to newer online panels—is essential for accurate electoral forecasting.

This paper develops statistical models to forecast the outcome of the 2024 presidential election between Kamala Harris and Donald Trump. By leveraging multi-linear regression models, we predict the percentage of support for each candidate across different states, incorporating key variables such as pollster rating, sample size, and state-level demographics. Through aggregating these state-level predictions, we simulate Electoral College outcomes to provide insights into each candidate's probability of securing the required 270 electoral votes. Our analysis also includes a detailed examination of YouGov's polling methodology and proposes an idealized survey approach that could enhance the accuracy of election forecasting.

The remainder of this paper is structured as follows. Section 2 discusses the data used for this analysis, including key variables and sources, with particular attention to the quality metrics that affect polling accuracy. Section 3 outlines our modeling approach for each candidate, incorporating lessons learned from recent electoral cycles. Section 4 presents our Electoral College predictions based on the model outputs. Section 5 discusses the implications of our findings and suggests directions for future research. Finally, Section A evaluates YouGov's polling methodology, and our idealized survey methodology.

2 Data

2.1 Overview

Our study utilizes polling data from FiveThirtyEight's 2024 Presidential Election Forecast Database (FiveThirtyEight 2024), a comprehensive polling dataset maintained by ABC news. This database compiles and standardizes polling results from various organizations, applying quality metrics and assessments to each poll. Several key variables from this dataset are crucial to our analysis:

The dataset employed in this analysis encompasses polling data from various reputable sources, capturing critical variables related to polling organizations, sample sizes, candidate support levels, and polling coverage (state-level or national). The primary variables of interest are outlined as follows:

- **Poll ID**: A unique identifier for each poll entry, facilitating data tracking and management (e.g., 88590).
- **Pollster**: The organization conducting the poll, which serves as a primary indicator of the poll's methodological quality (e.g., YouGov, Ipsos).
- **Pollscore**: A quantitative measure of the pollster's reliability, where lower (negative) values suggest higher predictive accuracy (e.g., -1.1 for YouGov).
- **Sample Size**: The number of respondents included in each poll, influencing the statistical precision and margin of error (e.g., 1414).
- **Support Percentage (pct)**: The proportion of respondents expressing support for each candidate, serving as the dependent variable in subsequent analyses (e.g., 47% for Kamala Harris).
- **State**: Indicates the geographical focus of the poll, either at the state level or nationwide (e.g., National).
- **Candidate Name**: The candidate assessed in the poll, providing context for support levels and enabling candidate-specific analyses (e.g., Kamala Harris, Donald Trump).
- **End Date**: The date on which the poll concluded, offering temporal alignment for longitudinal analyses (e.g., 2024-10-07).

To ensure analytical rigor, the dataset underwent extensive preprocessing. This involved standardizing variable formats for compatibility with regression models, converting support percentages into numerical values, and systematically addressing missing data to minimize bias. These preprocessing steps ensured that the dataset adhered to the standards of a "tidy" dataset, enabling robust model development and enhancing interpretability of subsequent statistical analyses.

Table 3 shows a sample of the dataset.

2.2 Measurement and Limitations

There are several measurement and limitation considerations for our dataset:

- **Poll Quality:** While we filter for high-quality polls using numeric grades, differences in polling methodologies may introduce systematic biases. The inclusion of pollster ratings helps account for historical accuracy but cannot completely eliminate these potential biases.
- **Temporal Dynamics:** Our dataset provides a snapshot of voter preferences during a specific timeframe. This static nature means we cannot capture the full dynamics of voter preference evolution over the campaign period.
- **Geographic Coverage:** Although we have national and state-level polling data, coverage varies by state. Battleground states typically have more frequent polling, while safer states may have sparse data, potentially affecting our state-level predictions.
- **Response Bias:** Despite careful methodology by pollsters, self-selection bias in survey participation and social desirability bias in responses remain potential concerns.

2.3 Outcome variables

Our primary outcome variable is the percentage of support (*pct*) for Kamala Harris and Donald Trump in each poll. This measurement represents the proportion of respondents who indicate they would vote for each candidate if the election were held on the day of the poll. The variable directly captures voter preferences and serves as the foundation for our electoral predictions.

Figure 1 below illustrates the distribution of support for both candidates, revealing several notable patterns. First, the support percentages cluster between 40% and 60%, reflecting the competitive nature of the race. Second, the distributions show slight differences between candidates, with Harris's support displaying more variation than Trump's. This pattern might reflect differences in voter certainty or polling methodology across different states and time periods.

2.4 Predictor variables

Our model incorporates several key predictor variables, each chosen for its theoretical importance in explaining polling variations and electoral outcomes:

- **Numeric Grade:** A composite measure (scale: 0-4) incorporating factors such as methodology rigor and historical accuracy
- **Pollscore:** A measure of historical polling accuracy (range: -4 to +4, where negative scores indicate better performance)
- **Transparency Score:** Quantifies the openness of polling methodology (scale: 0-10)
- **Sample Size:** Number of respondents, typically ranging from 500 to 3000

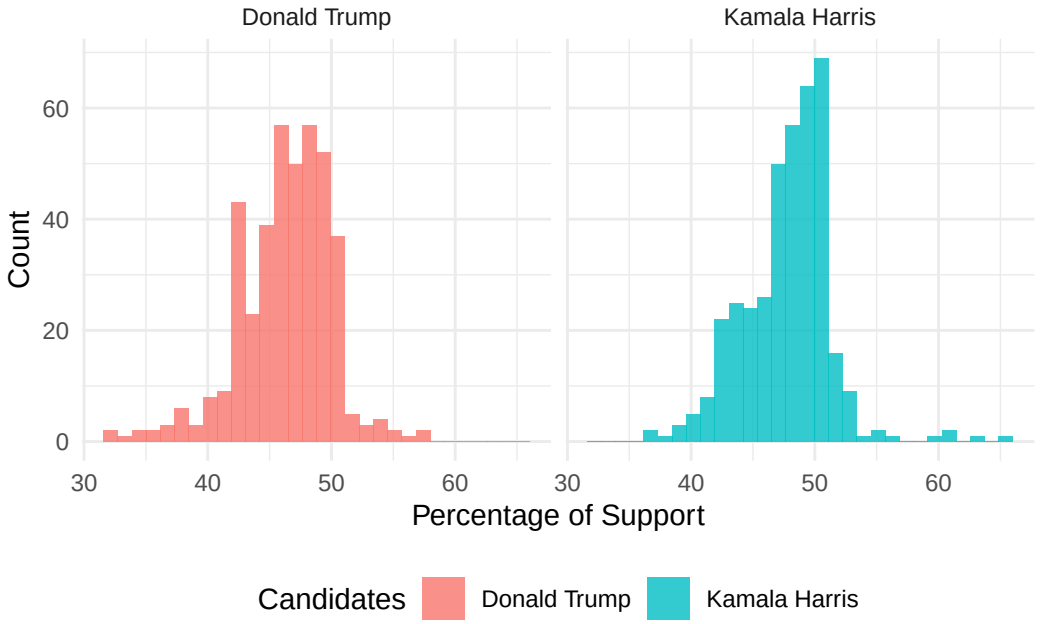

Figure 1: Distribution of Support for Kamala Harris and Donald Trump

- **State:** State-level indicators capturing regional political variations
- **Methodology:** Survey approach (e.g., online panel, phone interviews)

The relationship between these predictors and polling outcomes is complex and often interconnected. For example, while larger sample sizes generally reduce sampling error, this effect may be moderated by the poll's methodology and quality metrics. Similarly, geographic variations in polling accuracy suggest that the relationship between predictors and outcomes may vary systematically across states.

These variables were selected based on both theoretical foundations in electoral polling literature and practical considerations of data availability and quality. We focused on these core variables due to their consistent availability across polls and demonstrated importance in previous electoral forecasting efforts.

Edit note: Add a table in appendix to show sample with all variables used in model

2.5 Cleaning Process and Analysis

The data cleaning process employed R (R Core Team 2023) along with several specialized packages: tidyverse (Wickham et al. 2019) for data manipulation, dplyr (Wickham et al. 2023) for data transformation, janitor (Firke 2023) for consistent naming conventions, and lubridate (Grolemund and Wickham 2011) for date handling.

Our cleaning process followed several key steps:

1. Filtered for high-quality polls using a minimum threshold (*numeric_grade* ≥ 2.7)
2. Limited temporal coverage to post-campaign announcement period (after July 21, 2024)
3. Standardized geographic data:
    i. Coded missing state information as "National" polls
    ii. Verified state names for consistency
4. Transformed key variables:
    i. Converted end dates to standardized date format
    ii. Calculated absolute supporter numbers from percentages and sample sizes
    iii. Created binary candidate indicators (Harris = 1, Trump = 0)

Our variable selection process prioritized measures essential for polling accuracy and electoral prediction while eliminating redundant or non-informative fields. We retained key poll quality metrics including *pollscore* and *numeric_grade*, which provide crucial information about the reliability of each survey. Sample characteristics such as *sample size* and *methodology* were preserved to account for differences in polling precision. *State* and polling *end_date* information were maintained to capture regional variations and time-dependent patterns in voter preferences.

We excluded several variables that offered little additional analytical value, such as *population_full*, which duplicated information available in other fields, and administrative data that did not directly influence predictions. This focused approach to variable selection allowed us to preserve all information necessary for accurate electoral forecasting.

3 Modeling Support for the Candidates

To analyze factors influencing candidate support, we use a linear regression model with sample size as the primary predictor. This model examines the relationship between sample size and the percentage of support for Kamala Harris and Donald Trump across different polls. By including sample size as a predictor, we can assess whether larger or smaller sample sizes impact reported support levels for each candidate.

For Kamala Harris, the linear regression results show that sample size has a statistically insignificant effect on her support percentage. This implies that other factors, such as pollster methodology or regional biases, may play a more substantial role in shaping her reported support.

In contrast, the results for Donald Trump indicate a weak but statistically significant negative effect of sample size on his support percentage. This suggests that, on average, larger sample sizes tend to show slightly lower support for Trump, though the effect size remains small.

Figures Figure 2 and Figure 3 illustrate the relationship between sample size and support percentage for Kamala Harris and Donald Trump, respectively. These plots provide a visual representation of the trends observed in the linear regression models for each candidate.

3.1 Multiple Linear Regression Models

To better capture the complexity of voter support, we constructed multiple linear regression (MLR) models for both Kamala Harris and Donald Trump. These models incorporate key predictors, including pollster rating (represented by pollscore), transparency score, sample size, and state-level data. By including these variables, the MLR models aim to control for multiple factors influencing voter support and to enhance the accuracy of our predictions.

The MLR model for Kamala Harris indicates that factors such as state and pollster rating significantly impact her predicted level of support across polls. The results, shown in Figure 4, suggest that Harris's support varies substantially by state and depends on the credibility of the pollster. Diagnostic checks, including residual and QQ plots, confirm that the model reasonably meets linear regression assumptions.

Similarly, the MLR model for Donald Trump highlights that state-level differences and pollster transparency play a significant role in explaining his support. Figure Figure 5 illustrates these findings, and diagnostic evaluations confirm that the model assumptions (linearity, independence, and homoscedasticity) are sufficiently met.

3.2 Multicollinearity Check Using Variance Inflation Factor (VIF)

To ensure that the predictors used in both models do not exhibit multi-collinearity, we checked the Variance Inflation Factor (VIF) for each predictor. High VIF values indicate multi-collinearity, which can affect the stability and reliability of the model coefficients.

Table 1: Harris MLR model Variance Inflation Factor (VIF)

	GVIF	Df	GVIF^(1/(2*Df))
numeric_grade	4.516	1	2.125
pollscore	3.050	1	1.746
transparency_score	7.164	1	2.677
sample_size	1.676	1	1.295
state	18.682	26	1.058
methodology	83.067	10	1.247

Based on the VIF results in Table 1, we refined the model by removing less significant predictors (such as methodology and transparency_score) to reduce multi-collinearity.This improves the model's accuracy and interpretability.

Table 2: Harris Refined MLR model Variance Inflation Factor (VIF)

	GVIF	Df	GVIF^(1/(2*Df))
numeric_grade	2.435	1	1.560
pollscore	2.353	1	1.534
sample_size	1.593	1	1.262
state	2.076	26	1.014

3.3 Stepwise Model Selection

To refine the multi-linear regression model, I applied a stepwise selection method to optimize the choice of predictor variables. Stepwise selection evaluates the model iteratively, adding or removing variables based on predefined criteria (typically AIC) to find a more parsimonious model. Using both forward and backward selection, stepwise method in the step() function assesses each predictor's contribution and determines if adding or removing it improves model fit. By allowing the function to examine variables in both directions, this process systematically removed non-significant variables that did not contribute meaningfully to the model's explanatory power, leaving only the most impactful predictors in the final refined model. The resulting plots, showing the "Residuals vs Fitted" and "Normal Q-Q" diagnostic views, allow for visual inspection of model assumptions and residual patterns, supporting the quality of the refined model.

3.4 National-Level Model Evaluation and Predictive Accuracy

After finalizing the model selection process, we evaluated the predictive performance of our models for Kamala Harris and Donald Trump using national-level polling data. Focusing on national polls, we split the dataset into training and test sets, maintaining consistency with a fixed random seed to ensure reproducibility. For each candidate, we developed a multiple linear regression model using pollscore and a log-transformed sample_size as predictors, capturing factors pertinent to national-level polling dynamics.

We then trained the models on the training subset and generated predictions for the test subset, evaluating model accuracy using the Root Mean Squared Error (RMSE). The RMSE metric quantifies the average prediction error in the test set, offering an indication of each model's reliability in predicting future national poll outcomes. The Harris model yielded an RMSE of 3.12, while the Trump model's RMSE was 2.40, indicating that both models provide reasonably accurate predictions, with the Trump model demonstrating slightly lower average error in predicting national polling support. These results reflect the models' robustness and highlight their utility for assessing national-level candidate support.

4 Electoral College Prediction

To forecast the winner of the 2024 U.S. Presidential election between Kamala Harris and Donald Trump, we performed the following series of steps, focusing on the distribution of electoral votes across states based on predicted polling percentages. This approach allows for an estimation of each candidate's electoral college support under the U.S. voting system.

1. Predicting State-Level Polling Percentages: We used final multiple linear regression models for both candidates, Harris and Trump. These models generated predicted polling percentages for each candidate across different states, capturing their expected levels of support based on the poll data used.
2. Averaging Predicted Support by State: For both Harris and Trump, we calculated the average predicted polling percentage within each state. This step provides a state-level summary of each candidate's support based on our predictions, simplifying comparisons between them in each state.
3. Assigning Electoral Votes: We then merged our state-level predictions with the distribution of electoral votes, as assigned by state in the Electoral College. This dataset provided the electoral vote count for each state, including special allocations for districts in Maine and Nebraska, which can split their electoral votes by congressional district.
4. Determining State Winners: Using the averaged predictions, we identified the winner in each state by comparing the predicted support percentages. Harris or Trump was deemed the winner in a state if their predicted polling percentage exceeded that of their opponent.
5. Aggregating Electoral Votes: Finally, we summed the electoral votes for each candidate across all states where they were predicted to win. This yielded total electoral vote counts for both Harris and Trump, which are essential for determining the predicted winner under the U.S. system, where 270 electoral votes are required to secure the presidency.
6. Interpreting the Results: With Harris predicted to receive 216 electoral votes compared to Trump's 147, Harris demonstrates a stronger position in the Electoral College, though neither candidate reaches the 270-vote threshold required for a decisive victory. This outcome suggests that, while a clear electoral win is not forecasted, Harris holds a better chance of winning the election based on her lead in projected electoral votes. This lead places her closer to the threshold and may indicate a competitive advantage should undecided states or close races lean in her favor.

This electoral vote aggregation based on state-level predictions provides an interpretation of how polling data (Figure 8), analyzed through multiple linear regression models, translates into likely electoral support for each candidate. It effectively models potential election outcomes while respecting the structure of the Electoral College.

5 Discussion

5.1 Key Findings

This study set out to forecast the outcome of the 2024 U.S. Presidential Election between Kamala Harris and Donald Trump by developing a multiple linear regression (MLR) model based on comprehensive polling data. Importantly, the same MLR model was applied to both candidates, ensuring consistency in how predictors influenced the predicted support percentages for each.

The MLR model incorporated several significant predictors of voter support, including state-level factors, pollster ratings (pollscore), transparency scores, and sample size. By using the same model for both Harris and Trump, we were able to directly compare how these variables impacted each candidate's predicted support across different states.

For both Kamala Harris and Donald Trump, the model highlighted the substantial impact of state-level variations on their predicted support. Certain states showed higher predicted support for Harris, such as California and Maryland, which historically lean Democratic. Conversely, states like Indiana and Missouri exhibited higher predicted support for Trump, aligning with their Republican-leaning tendencies. The consistency of the model across both candidates allowed for a clear comparison of how state demographics and historical voting patterns influenced voter preferences.

Pollster reliability indicators, such as numeric grades and transparency scores, were significant predictors for both candidates. This suggests that polls conducted by more credible organizations tended to report more accurate and potentially higher support levels for each candidate. The sample size also played a role, although its impact varied slightly between the two candidates. By using the same model, we ensured that these predictors were weighted equally in the analysis of both Harris's and Trump's support levels.

When aggregating the state-level predictions to simulate the Electoral College outcomes, the findings suggest a competitive race. Harris is projected to receive 216 electoral votes, while Trump is predicted to secure 147 electoral votes. Neither candidate reaches the 270-vote threshold required to win the presidency, emphasizing the potential for a closely contested election. The remaining electoral votes are likely concentrated in battleground states, where voter preferences are more volatile and could ultimately determine the election outcome.

5.2 Model Strengths

A primary strength of this analysis lies in the consistent application of the same MLR model to both candidates. This approach ensures that the comparison between Harris and Trump is based on the same criteria and that any differences in predicted support are attributable to variations in the data rather than differences in the modeling approach.

By incorporating multiple predictors—including state-level data, pollster ratings, transparency scores, and sample size—the model provides a nuanced understanding of the factors influencing voter preferences. This multifaceted approach allows for more accurate predictions compared to models relying solely on national-level polling data or a single set of predictors.

The use of stepwise model selection and checks for multicollinearity enhanced the robustness of the MLR model. Stepwise selection optimized the predictor variables, ensuring that only those contributing significantly to the model were included. Checking for multicollinearity through the Variance Inflation Factor (VIF) helped prevent unreliable estimates due to correlated predictors, thereby improving the model's validity for both candidates.

Moreover, translating the predicted support percentages into Electoral College projections adds practical value to the analysis. Since the U.S. Presidential Election is determined by electoral votes rather than the popular vote, this approach offers a more realistic forecast of the election outcome. Using the same model for both candidates ensures that the Electoral College projections are directly comparable.

5.3 Limitations

Despite the strengths, several limitations need to be acknowledged:

- Reliance on Polling Data Quality: The accuracy of the model is inherently dependent on the quality of the polling data used. While the analysis accounted for pollster ratings and transparency scores, inherent biases and methodological differences between polls could still affect the results. Sampling errors, non-response biases, and underrepresentation of certain demographic groups are persistent issues in polling data that can skew predictions for both candidates.
- Static Snapshot of Voter Preferences: The model provides a static view based on the most recent polling data available. It does not account for the dynamic nature of election campaigns, where voter preferences can shift rapidly due to various factors such as political events, debates, or emerging issues. This limitation means the model may not accurately predict future changes in voter sentiment leading up to the election for either candidate.
- Exclusion of Other Influential Factors: The analysis did not incorporate other variables that can significantly impact election outcomes, such as economic indicators (e.g., unemployment rates, GDP growth), social movements, campaign strategies, or voter turnout efforts. The absence of these factors may limit the model's ability to fully capture the complexities of voter behavior for both Harris and Trump.
- Electoral College Simplifications: The Electoral College simulation assumes that the candidate with the higher predicted support in a state wins all its electoral votes (except for Maine and Nebraska). This winner-takes-all approach does not account for the

proportional allocation of electoral votes in those two states or potential variations in voter turnout that could affect the actual results.

5.4 Future Research

To enhance the predictive power and accuracy of election forecasting models, future research could consider the following approaches:

- Incorporate Time-Series Analysis: Introducing a time-series component would allow the model to account for trends and shifts in voter preferences over time. This dynamic approach could capture the impact of events such as debates, policy announcements, or external factors on voter sentiment for both candidates.
- Expand Predictive Variables: Including additional variables such as economic indicators, demographic data, and social media sentiment analysis could provide a more comprehensive view of the factors influencing voter behavior. Integrating data on unemployment rates or consumer confidence might reveal economic influences on candidate support.
- Voter Turnout Models: Developing models to predict voter turnout could significantly improve election forecasts. Turnout can vary widely between elections and is often influenced by voter enthusiasm, mobilization efforts, and barriers to voting. Incorporating turnout probabilities could adjust the predicted support percentages to reflect more realistic electoral scenarios.
- Advanced Modeling Techniques: Exploring advanced statistical methods or machine learning algorithms might capture complex nonlinear relationships between predictors and voter support. Techniques such as random forests, gradient boosting, or neural networks could uncover patterns not detectable through linear regression.
- Account for Electoral Nuances: With variations in how electoral votes are allocated and the potential impact of third-party candidates or ranked-choice voting in some states, future models should consider these electoral nuances. Simulating different voting systems could provide insights into how alternative methods might affect election outcomes.
- Cross-Validation with Alternative Data Sources: Validating the model against alternative data sources, such as exit polls or historical voting patterns, could test its robustness. Cross-validation helps ensure that the model is not overfitted to the polling data used and can generalize to different datasets.

A Appendix

A.1 YouGov Pollster Methodology Overview and Evaluation

YouGov conducts online surveys through their proprietary panel of U.S. adults, using non-probability sampling methods combined with sophisticated weighting procedures to achieve representative results. Their approach balances speed and cost-effectiveness with statistical rigor through careful sample selection and data quality controls.

A.1.1 Survey Population and Sampling

YouGov's target population typically comprises all U.S. adults or adult citizens, with their sampling frame consisting of their opt-in online panel covering approximately 95% of Americans. For general population surveys, they aim for 1,000-2,000 respondents, selected based on demographic and political characteristics to match the target population.

A.1.2 Panel Recruitment and Participation

Panel members are recruited through advertising and website partnerships, with surveys offered in multiple languages including Spanish to ensure broad representation. Participants receive points exchangeable for small monetary rewards, though many report being motivated by the desire to contribute to research.

A.1.3 Quality Control

YouGov employs several measures to maintain data quality:

- Verification of panelist identity through email and IP checks
- Response quality surveys to gauge reliability
- Monitoring of response times and patterns
- Removal of respondents who fail quality checks
- Question randomization to reduce bias

A.1.4 Non-response and Weighting

To address potential biases, YouGov applies statistical weighting based on demographics (age, gender, race, education) and political factors (voting behavior, party identification). Their weighting process considers multiple characteristics simultaneously to better reflect real-world demographic intersections.

A.1.5 Strengths and Limitations

The methodology's primary strengths include rapid data collection, cost-effectiveness, and the ability to track opinions over time. However, the nonprobability sampling approach may introduce biases, and the online-only format could underrepresent certain populations. While weighting helps address these limitations, it cannot fully account for all potential sources of bias.

A.2 Idealized Survey Methodology

This idealized survey methodology outlines a comprehensive plan for forecasting the US presidential election within a budget of $100,000. The approach is designed to be statistically sound, practical, and capable of accurately predicting election outcomes by considering both the popular vote and electoral college implications.

A.2.1 Sampling Strategy

The target population for this survey is eligible voters across the United States who are likely to participate in the upcoming presidential election. To achieve a representative sample:

- **Sampling Frame:** Utilize a combination of registered voter lists and demographic data from reputable sources such as the US Census Bureau.
- **Sampling Method:** Implement stratified random sampling to ensure representation across key demographics, including age, gender, race, education level, and geographic location.
- **Sample Size Calculation:** Aim for a sample size of approximately 10,000 respondents to achieve a margin of error of $\pm 1\%$ at a 95% confidence level.
- **Geographical Distribution:** Allocate samples proportionally across all 50 states and the District of Columbia, with oversampling in swing states to better predict electoral college outcomes.
- **Addressing Sampling Biases:** Apply weighting adjustments to account for underrepresented groups and ensure that the sample mirrors the overall voter population.

A.2.2 Recruitment Plan

To recruit respondents effectively, we will leverage online panels, social media advertising, and partnerships with community organizations to reach a diverse audience. Offering modest incentives, such as $5 digital gift cards, encourages participation while managing costs. Quota sampling within strata maintains demographic balance, and follow-up reminders along with mobile-friendly survey formats help reduce non-response bias. The data collection will occur over a two-week period to capture timely opinions without introducing temporal biases.

A.2.3 Survey Design Elements

The survey is crafted to elicit accurate and meaningful responses:

- **Question Types and Formats:** Use a mix of closed-ended questions and multiple-choice options for clarity and ease of analysis.

- **Response Options:** Include balanced and neutral response choices, with options for "Undecided" or "Prefer not to say."
- **Question Order and Flow:** Begin with general questions to build rapport, followed by more specific vote intention queries, and conclude with demographic questions.
- **Demographic Information:** Collect data on age, gender, race, education, income, and geographic location.
- **Political Affiliation and History:** Ask about party affiliation, past voting behavior, and political engagement.
- **Likely Voter Screens:** Include questions to gauge voting likelihood, such as past voting frequency and intention to vote in the upcoming election.
- **Vote Intention Questions:** Directly ask which candidate the respondent intends to vote for, ensuring confidentiality and anonymity.

A.2.4 Quality Control

To maintain data integrity, we implement several quality control measures. Real-time validation checks within the survey prevent inconsistent or illogical responses. Attention-check questions identify disengaged respondents. We use unique survey links and track IP addresses to prevent duplicate submissions, while CAPTCHA verification deters automated responses. Incomplete or suspicious responses are excluded during data cleaning to ensure the final dataset is robust and reliable.

A.2.5 Data Processing

These data processing steps will be taken to ensure accurate analysis:

- **Weighting Methodology:** Adjust survey results using weighting factors based on demographic proportions in the voting population.
- **Handling Missing Data:** Employ imputation techniques or exclude cases with significant missing information.
- **Outlier Detection:** Identify and review outliers that may skew results, determining whether to retain or discard them.
- **Response Validation:** Cross-check responses for consistency and plausibility.
- **Poll Aggregation Approach:** Combine survey data with other reputable polls using meta-analytic techniques to enhance prediction accuracy.

A.2.6 Budget Allocation

A budget allocation of $100,000 ensures all aspects are adequately funded:

- **Recruitment Costs:** $40,000 for advertising and partnerships to reach potential respondents.
- **Incentive Payments:** $50,000 allocated for participant incentives ($5 x 10,000 respondents).
- **Survey Platform Fees:** $2,000 for premium features on a survey platform like Google Forms or an equivalent.
- **Data Analysis Tools:** $3,000 for statistical software licenses and data processing tools.
- **Quality Control Measures:** $3,000 for implementing validation systems and CAPTCHA services.
- **Administrative Costs:** $2,000 for project management and miscellaneous expenses.

A.2.7 Conclusion

This methodology presents a feasible and thorough plan to forecast the US presidential election within the specified budget. By adhering to best practices in survey design and execution, and by carefully considering both the popular vote and electoral college implications, the survey aims to provide accurate and reliable insights into voter intentions.

B Appendix

Table 3: Sample Overview of Selected Variables in the Polling Dataset

	1	2	3	4	5
poll_id	88590	88590	88590	88590	88558
pollster	YouGov	YouGov	YouGov	YouGov	Ipsos
pollscore	-1.1	-1.1	-1.1	-1.1	-0.9
sample_size	1414	1414	1230	1230	1272
pct	47	44	49	45	42
state	National	National	National	National	National
candidate_name	Kamala Harris	Donald Trump	Kamala Harris	Donald Trump	Kamala Harris
end_date	2024-10-07	2024-10-07	2024-10-07	2024-10-07	2024-10-07

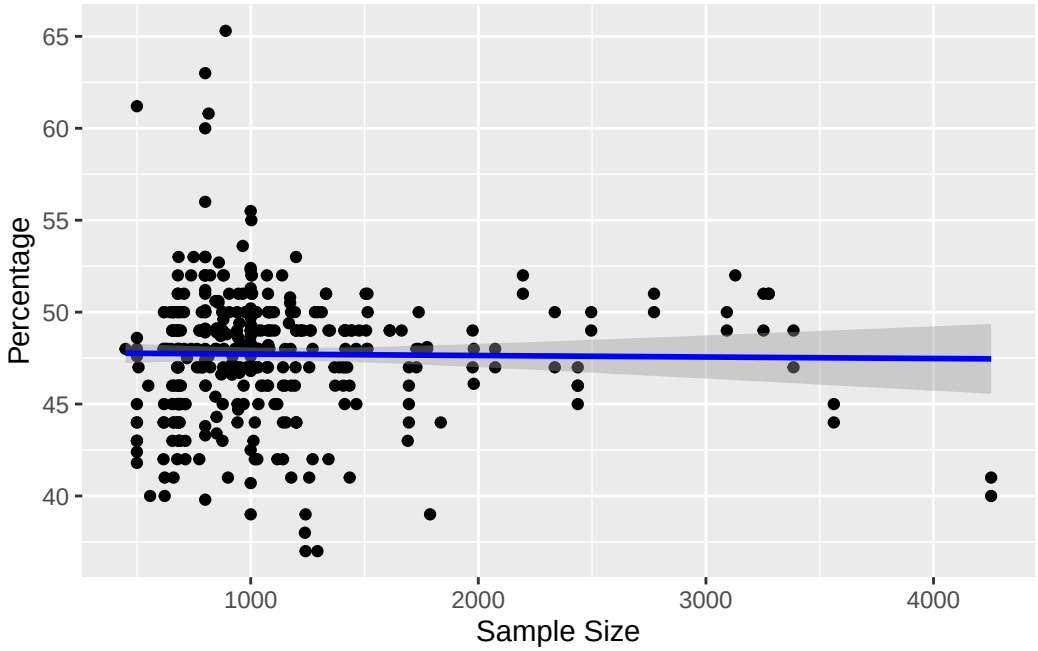

Figure 2: Linear Regression of Percentage vs Sample Size for Kamala Harris

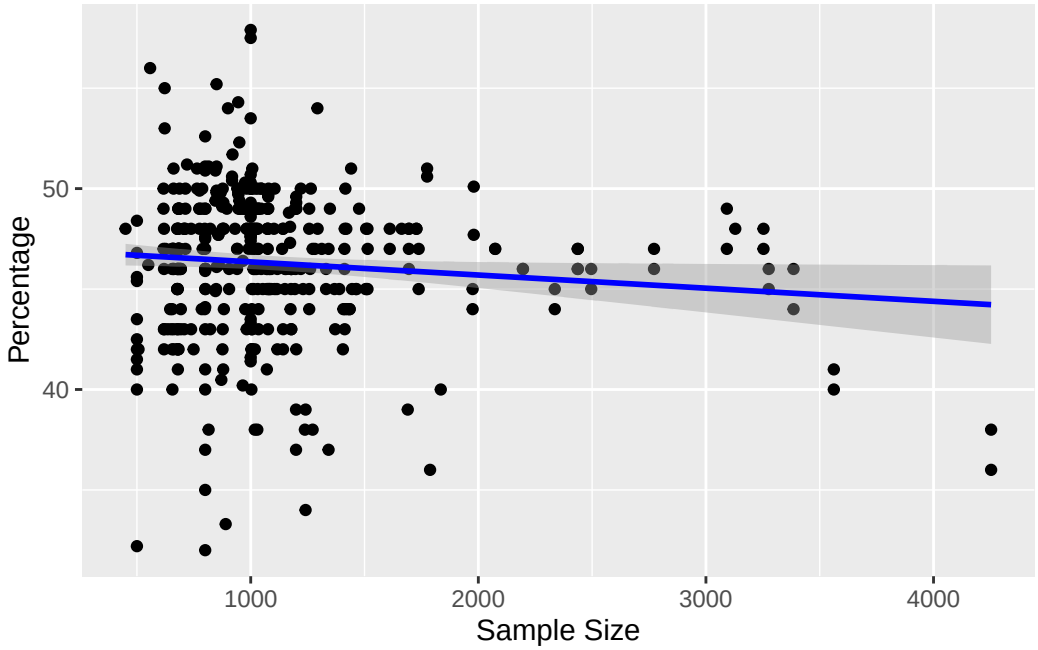

Figure 3: Linear Regression of Percentage vs Sample Size for Donald Trump

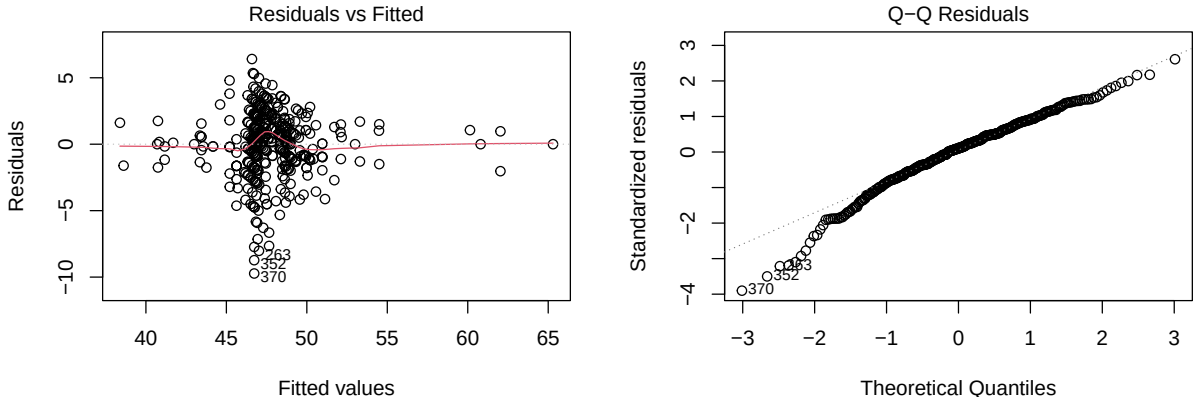

Figure 4: Multi-Linear Regression model for Kamala Harris

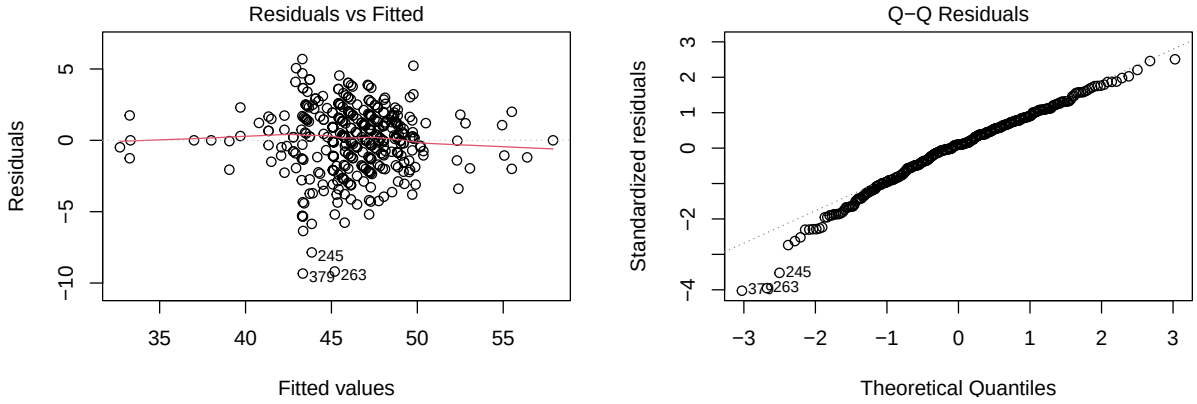

Figure 5: Multi-Linear Regression model for Donald Trump

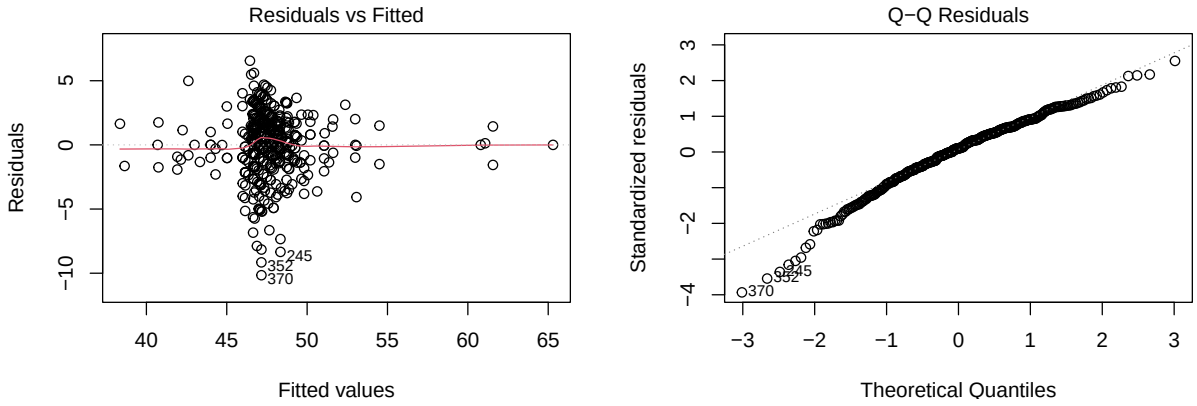

Figure 6: Harris Refined Multi-Linear Regression Model

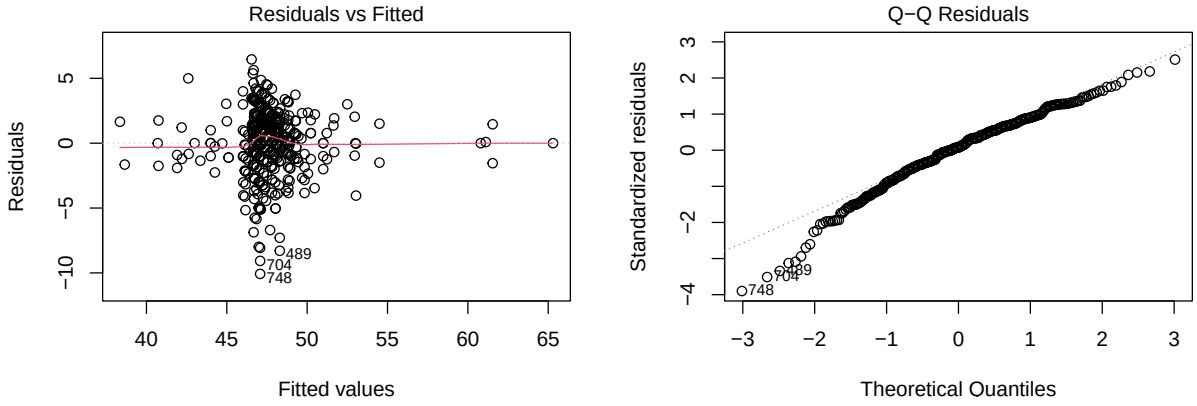

Figure 7: Harris Final Multi-Linear Regression Model

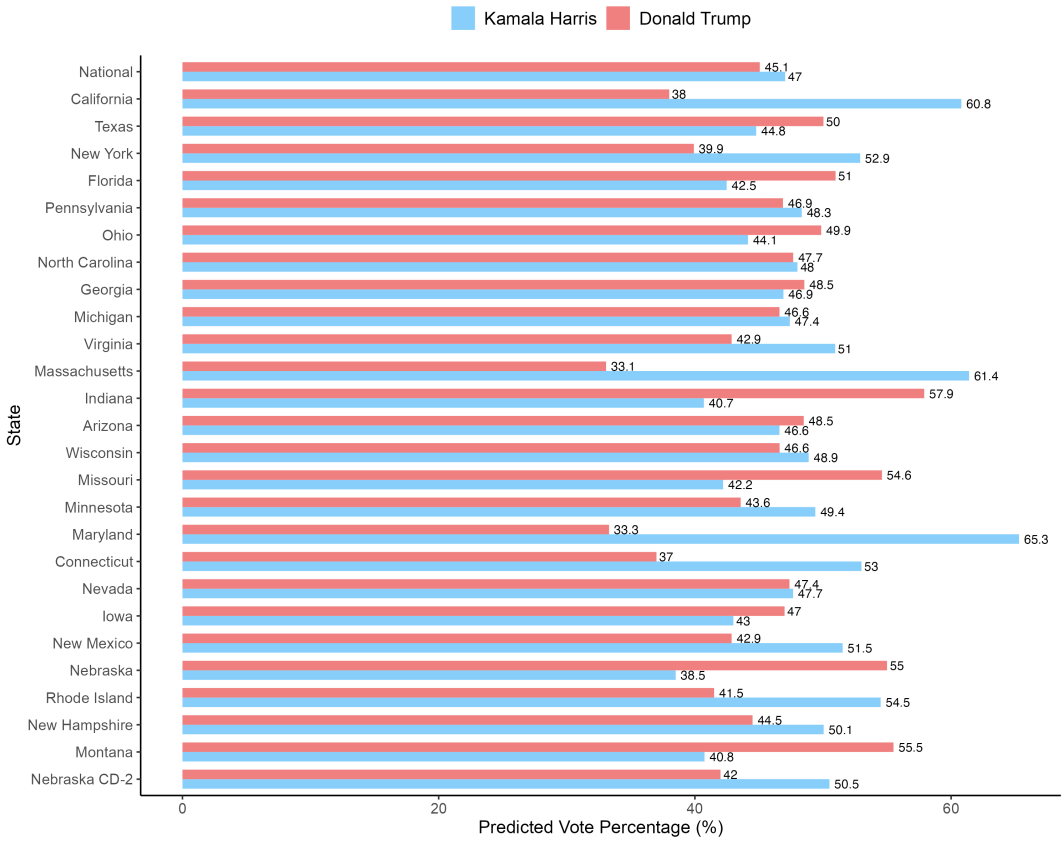

Figure 8: Predicted Vote Percentages by State

References

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

FiveThirtyEight. 2024. "2024 Election Polls." FiveThirtyEight. https://projects.fivethirtyeight.com/polls/.

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/.

Keeter, Scott. 2024. "Key Things to Know about U.S. Election Polling in 2024." https://www.pewresearch.org/short-reads/2024/08/28/key-things-to-know-about-us-election-polling-in-2024/.

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

Viala-Gaudefroy, Jérôme. 2024. "2024 US Presidential Election: Can We Believe the Polls?" *The Conversation*. https://theconversation.com/2024-us-presidential-election-can-we-believe-the-polls-240834.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.