Forecasting the 2024 U.S. Presidential Election*

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

Tim Chen Steven Li Tommy Fu

November 2, 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 aggregates and standardizes polling results from various organizations, applying quality metrics and methodological assessments to each poll. In this section, we detail our selected variables, discuss key measurements, outline important limitations of our data, and our data cleaning process.

Our analysis focuses on several key variables that directly influence polling accuracy and electoral predictions. Table 4 presents a sample of our dataset.

2.2 Poll Quality Variables

Table 5 summarizes three primary metrics that assess poll quality in our dataset:

- **Pollscore**: Represents "the error and bias attributable to a pollster, where negative numbers indicate better performance." Our data shows a mean pollscore of -1.06 (SD = 0.28), with values ranging from -1.50 to -0.50, indicating generally high-quality polling organizations in our sample.
- **Numeric Grade**: A rating given to pollsters to indicate their quality and reliability. With a mean of 2.89 (SD = 0.10) and a range of 2.70 to 3.00, our dataset maintains high standards. We established a minimum threshold of 2.7 to ensure methodological rigor.
- **Transparency Score**: A measure reflecting pollsters' transparency about their methodology, with 10 being the highest possible score. The data shows strong overall transparency with a mean of 8.59 (SD = 1.04), indicating that most polls in our sample maintain high standards of methodological disclosure.

2.3 Methodological Variables

- **Methodology**: Figure 3 shows the distribution of polling approaches used to conduct the polls. Live Phone methods and Online Panels are the predominant methods, with hybrid approaches combining multiple methods (such as IVR/Online Panel/Text-to-Web) representing evolving polling techniques adapting to changing communication patterns.
- **Sample Size**: Figure 6 illustrates the distribution of poll sample sizes, representing the total number of respondents participating in each poll. Our dataset shows considerable variation, with a mean of 1,114.76 respondents (SD = 583.72). The right-skewed distribution ranges from 450 to 4,253 respondents, with larger sample sizes typically associated with national polls.

2.4 Geographic and Temporal Variables

- **State Coverage**: Table 6 reveals the distribution of polls across different U.S. states. National polls lead with 261 surveys, followed by concentrated polling in key battleground states: Pennsylvania (86), Wisconsin (82), and North Carolina (64). This distribution reflects strategic focus on states where the electoral outcome is less certain.
- **End Date**: Figure 5 tracks polling frequency from July through October 2024. We observe increased activity during September and early October, with weekly poll counts peaking in late September. This pattern reflects intensified polling efforts as the election approaches.

2.5 Outcome Variable

- **Support Percentage**: Figure 4 displays our primary dependent variable: the percentage of respondents expressing support for each candidate. The distribution reveals a competitive race between Donald Trump and Kamala Harris, with support levels typically ranging between 40-50% for both candidates.

2.6 Measurement and Limitations

There are several measurement and limitation considerations for our dataset:

- **Poll Quality:** While our pollscore and numeric grade filters help ensure data quality, these metrics are based on historical performance and may not fully capture current methodological improvements or deterioration.
- **Temporal Dynamics:** Our dataset provides discrete snapshots of voter preferences rather than continuous measurement. This limitation is particularly relevant given the rapid evolution of political narratives and voter sentiment during presidential campaigns.
- **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.

2.7 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 Date* values to standardized date format
    ii. Calculated absolute supporter numbers from percentages and sample sizes
    iii. Created binary candidate indicators (Harris = 1, Trump = 0)

Other variable such as … were not selected for our analysis because …

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 investigate the factors influencing voter support for Kamala Harris and Donald Trump, we employed linear regression models using polling data. Our analysis aimed to assess how predictors such as sample size, pollster ratings, transparency scores, and state-level variables affect the reported percentage of support for each candidate.

We began with simple linear regression models to examine the relationship between sample size (`sample_size`) and the percentage of support (`pct`) for each candidate.

Figures Figure 7 and Figure 8 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

Recognizing the multifaceted nature of voter support, we extended our analysis using multiple linear regression (MLR) models. The initial model incorporated several predictors believed to influence polling outcomes:

$$\text{pct} = \beta_0 + \beta_1 \times \text{numeric\_grade} + \beta_2 \times \text{pollscore} + \beta_3 \times \text{transparency\_score} \\ + \beta_4 \times \text{sample\_size} + \beta_5 \times \text{state} + \beta_6 \times \text{methodology} + \epsilon$$

Variables were selected based on their relevance to polling accuracy and potential influence on voter preferences. Pollster ratings and transparency scores were included to account for differences in pollster reliability. State-level data were incorporated to capture regional variations in support. Sample size and methodology were included to control for their potential impact on polling results (see Figure 9 and Figure 10 as it illustrates residual plots and Q-Q plots).

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

To assess multicollinearity among predictors, we calculated the Variance Inflation Factor (VIF) for each variable(Table 1). High VIF values suggest multicollinearity, which can compromise the stability of coefficient estimates.

The initial VIF analysis revealed multicollinearity concerns, particularly with transparency score and methodology. To address this, we refined the model by removing less significant predictors with high VIF values, resulting in:

$$\text{pct} = \beta_0 + \beta_1 \times \text{numeric\_grade} + \beta_2 \times \text{pollscore} + \beta_3 \times \text{sample\_size} + \beta_4 \times \text{state}$$

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

This refinement reduced VIF values to acceptable levels, improving the model's reliability and interpretability.

3.3 Stepwise Model Selection

To optimize the model further, we applied stepwise selection using the Akaike Information Criterion (AIC) as the criterion for adding or removing predictors. The final model obtained is:

$$\text{pct} = \beta_0 + \beta_1 \times \text{pollscore} + \beta_2 \times \log(\text{sample\_size}) + \beta_3 \times \text{state} + \epsilon$$

The logarithmic transformation of sample size was introduced to account for diminishing returns in effect size with increasing sample sizes.

The final model was selected based on statistical significance, model diagnostics, and theoretical considerations. Pollster rating and state remained significant predictors, consistent with expectations that pollster credibility and regional factors influence polling results. The logarithmic transformation improved model fit and addressed heteroscedasticity associated with sample size.

All statistical analyses were conducted using (R Core Team 2023). The lm() function was used for regression modeling, and the car package (Fox and Weisberg 2019) was employed to calculate VIF values. Stepwise selection was performed using the step() function. Diagnostic plots were generated to assess model assumptions.

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

Table 3: Electoral Votes for Each Candidate

Candidate	Electoral Votes
Harris	306
Trump	232

To project the outcome of the 2024 U.S. Presidential election between Kamala Harris and Donald Trump, we conducted a statistical analysis that integrates predicted state-level polling percentages with the allocation of electoral votes. This methodology employs multiple linear regression models to estimate each candidate's support and translates these predictions into potential Electoral College results, adhering to the structure of the U.S. electoral system.

1. **Prediction of State-Level Polling Percentages**: We utilized finalized multiple linear regression models for both Kamala Harris and Donald Trump. These models were applied to their respective datasets to generate predicted polling percentages for each candidate across various states. The predictions reflect the expected levels of support based on historical polling data and relevant covariates included in the models.

2. **Aggregation of Predicted Support by State**: For each candidate, we computed the average predicted polling percentage within each state. This aggregation provides a concise state-level summary of anticipated support, facilitating direct comparisons between the candidates in each state (see Figure 1).

3. **Integration of Electoral Vote Allocations**: We established a dataset detailing the electoral vote distribution for each state, including the distinct allocations for Maine and Nebraska, which can split their electoral votes by congressional district. This dataset was merged with the state-level predicted support data to align electoral votes with the corresponding predicted support.

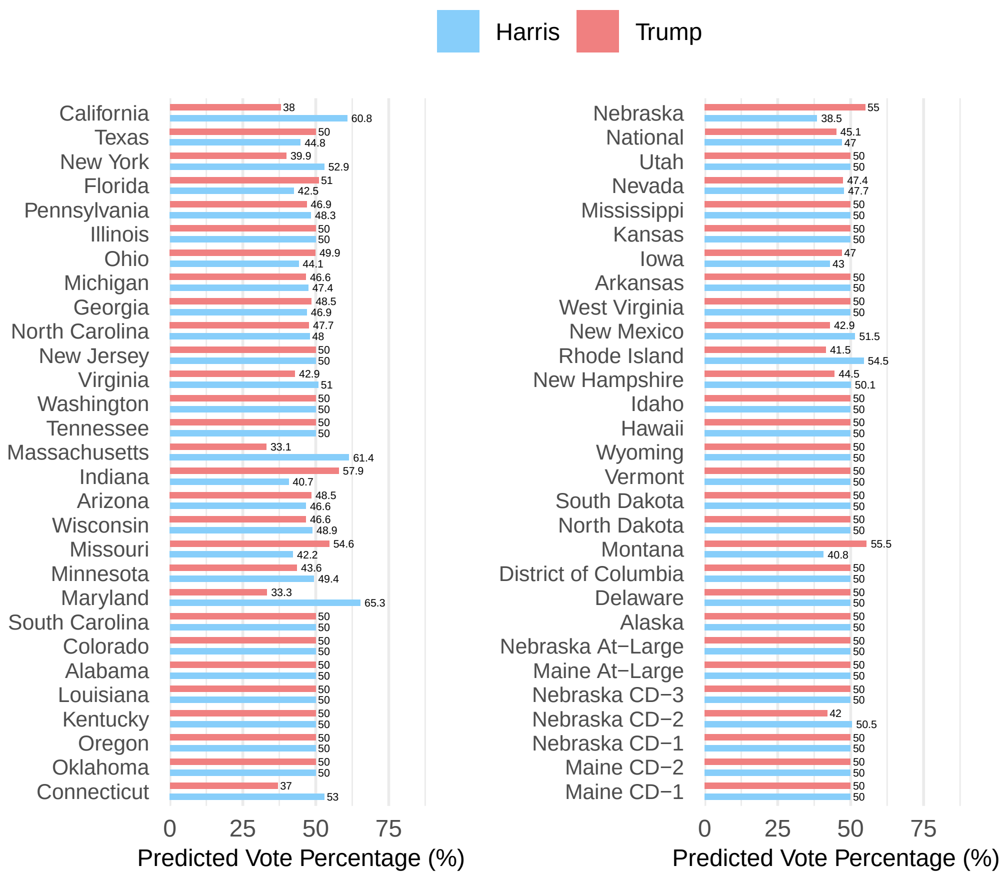

Figure 1: Predicted Vote Percentages by State

4. **Addressing Missing Data**: We identified states lacking predicted polling data and assigned a neutral predicted support of 50% to both candidates in these states. This assumption accounts for the absence of data while ensuring that all states are included in the analysis.

5. **Determination of State Winners**: Based on the aggregated predicted percentages, we determined the projected winner in each state. A candidate was designated as the winner if their predicted polling percentage exceeded that of their opponent. In cases where the predicted percentages were equal, we randomly assigned the winner using a stochastic approach to reflect the uncertainty inherent in tied predictions.

6. **Calculation of Total Electoral Votes**: We summed the electoral votes from all states won by each candidate to compute their total electoral vote counts. This calculation is critical, as securing at least 270 electoral votes is necessary to win the presidency under the U.S. electoral system.

7. **Interpretation of Results**: The analysis indicated that Kamala Harris is projected to receive a total of **306 electoral votes**, while Donald Trump is projected to receive **232 electoral votes**. Harris surpasses the 270-vote threshold required for victory, suggesting a strong position in the Electoral College based on our predictions. This outcome implies that Harris is likely to win the election, given her substantial lead in projected electoral votes (see Figure 2).

This statistical approach effectively connects state-level polling data to Electoral College projections. By employing multiple linear regression models to predict polling percentages and integrating these predictions with the electoral vote framework, we provide a detailed analysis of potential election results. This methodology reflects the structure of the U.S. electoral system and offers valuable insights into how predicted voter support may translate into electoral success.

Figure 2 displays a predicted U.S. Electoral College outcome in a hypothetical race between Harris and Trump, with blue and red indicating states projected to support each candidate, respectively. Harris is expected to win in traditional Democratic strongholds like the West Coast, Northeast, and parts of the Midwest, areas historically aligned with Democratic candidates due to factors such as urbanization, higher education levels, and diverse populations. In contrast, Trump's projected wins across the South and Great Plains reflect longstanding Republican preferences in these regions, shaped by conservative values and economic policies that resonate in rural and suburban areas. The distribution highlights both entrenched partisan divides and key battleground states, such as Florida and Pennsylvania, that could be decisive given their recent variability in party support.

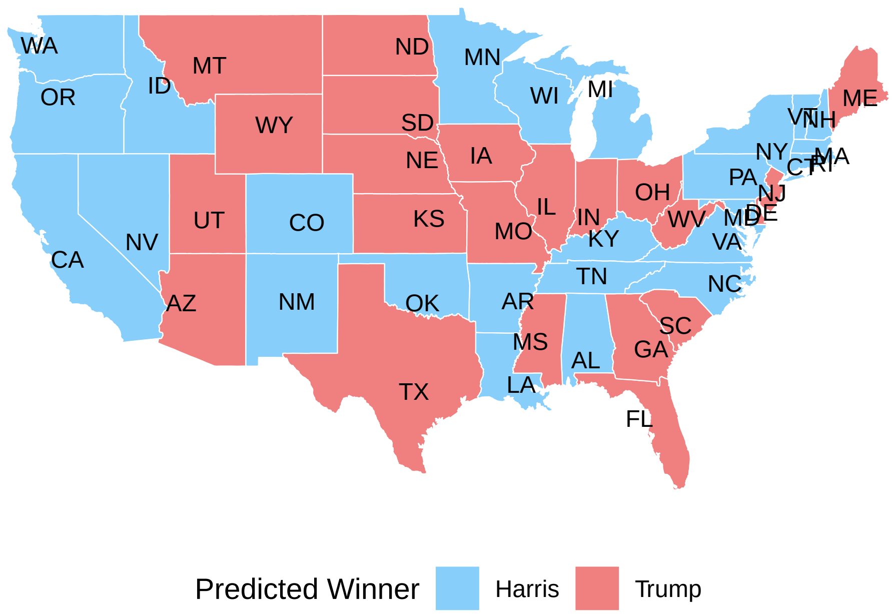

Figure 2: Electoral College Predition Map

5 Discussion

5.1 Summary of Findings

In this paper, we developed statistical models to forecast the outcome of the 2024 United States presidential election between Kamala Harris and Donald Trump. Utilizing polling data from FiveThirtyEight's 2024 Presidential Election Forecast Database, we analyzed factors influencing candidate support through linear and multiple linear regression models. We incorporated variables such as pollster rating, transparency score, sample size, and state-level data to predict the percentage of support for each candidate across different states. By aggregating these state-level predictions and integrating electoral vote allocations, we projected that Kamala Harris would receive 306 electoral votes, surpassing the 270 required to win the presidency.

5.2 Insights on Polling Methodologies and Voter Support

Our analysis indicates that pollster rating and state-level factors significantly influence reported support for both candidates. For Kamala Harris, the multiple linear regression model shows that higher pollster ratings and certain states are associated with increased support percentages. This suggests that polls conducted by more reliable pollsters may report higher support for Harris, possibly due to methodological rigor or sampling techniques that capture her voter base more accurately. For Donald Trump, pollster transparency and state differences play a notable role, indicating that how openly pollsters disclose their methodologies can affect the reported support for Trump.

5.3 Implications for Electoral Forecasting

The projection of Kamala Harris receiving 306 electoral votes underscores the importance of incorporating state-level analyses in electoral forecasting. Our findings highlight that national polling averages may not sufficiently capture the complexities of the Electoral College system. By focusing on state-specific predictions and accounting for variables that influence voter support in different regions, forecasters can achieve more accurate predictions. This approach acknowledges the heterogeneity of the electorate and the pivotal role of battleground states in determining election outcomes.

5.4 Limitations

Despite the strengths of our modeling approach, several limitations must be acknowledged. First, the reliance on historical polling data and pollster ratings assumes that past performance is indicative of future accuracy, which may not hold true if polling methodologies or voter behaviors change significantly. Second, the underrepresentation of certain states due to sparse polling data may affect the reliability of our state-level predictions. Assigning a default 50% support in states with missing data introduces uncertainty and may not reflect actual voter preferences. Additionally, the models may not fully account for dynamic factors such as late-breaking events, shifts in voter sentiment, or turnout variations that can influence election results.

5.5 Future Research and Recommendations

To enhance the accuracy of electoral forecasts, future research should consider incorporating real-time data and alternative data sources such as social media trends, economic indicators, and demographic shifts. Employing more sophisticated modeling techniques, such as hierarchical Bayesian models or machine learning algorithms, may capture complex interactions and non-linear relationships among variables. Furthermore, improving polling methodologies to address issues of response bias and sample representativeness is essential. Collaborations between pollsters, statisticians, and political scientists can foster the development of more robust predictive models that adapt to the evolving electoral landscape.

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 4: Sample Overview of Selected Variables in the Polling Dataset

pollster	Washington Post/George Mason University	CNN/SSRS	Siena/NYT	YouGov	CNN/SSRS
pollscore	-0.8	-0.6	-1.5	-1.1	-0.6
numeric_grade	2.7	2.8	3.0	3.0	2.8
transparency_score	9	10	9	9	10
sample_size	1005	931	1142	1033	708
methodology	Live Phone/Text-to-Web	Live Phone/Text-to-Web/Email/Mail-to-Web/Mail-to-Phone	Live Phone	Online Panel	Probability Panel
pct	42	48	48	47	48
state	Virginia	North Carolina	National	National	Michigan
candidate_name	Donald Trump	Donald Trump	Donald Trump	Donald Trump	Kamala Harris
end_date	2024-09-08	2024-09-25	2024-07-24	2024-10-04	2024-08-29

Table 5: Summary Statistics of Key Poll Metrics

Statistic	sample	pollscore	numeric	transparency
Mean	1114.76	-1.06	2.89	8.59
Median	1000.00	-1.10	2.90	9.00
SD	583.72	0.28	0.10	1.04
Min	450.00	-1.50	2.70	6.00
Max	4253.00	-0.50	3.00	10.00

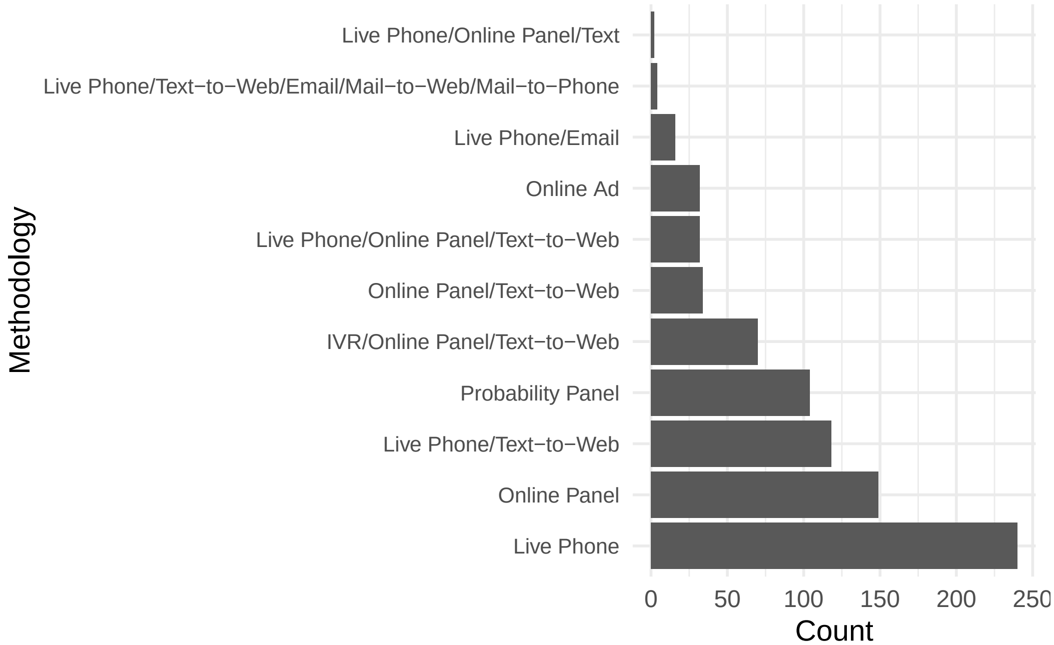

Figure 3: Distribution of Polling Methodologies

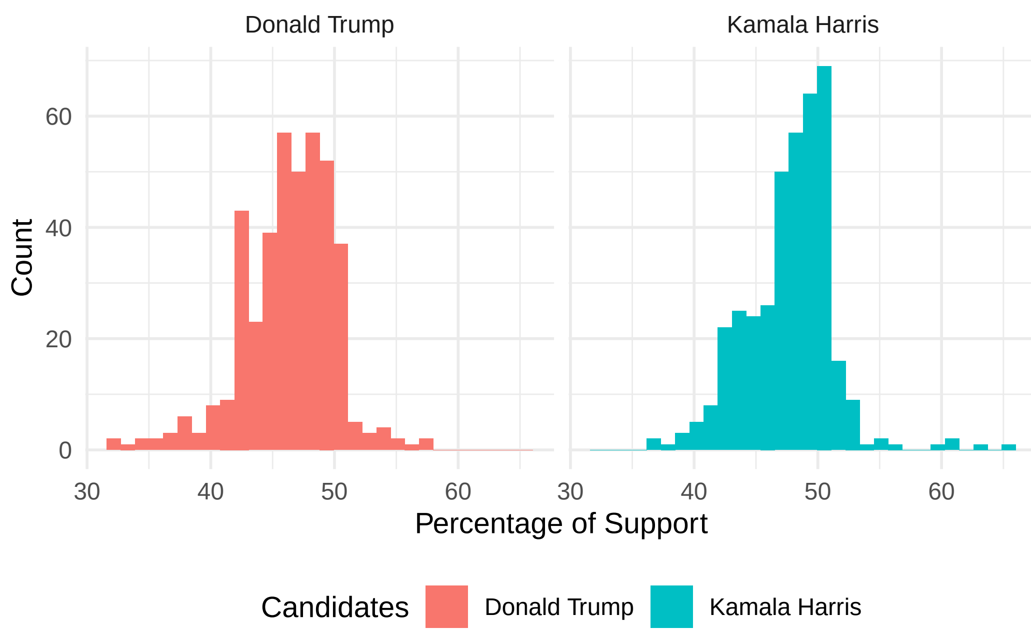

Figure 4: Distribution of Support Percentages by Candidate

Table 6: Polling Frequency by State

State	Number of Polls
National	261
Pennsylvania	86
Wisconsin	82
North Carolina	64
Michigan	58
Georgia	56
Arizona	52
Nevada	22
Minnesota	12
Nebraska CD-2	12
Texas	12
Virginia	12
Florida	10
New York	10
Ohio	10
Massachusetts	6
New Mexico	6
Missouri	4
Montana	4
Nebraska	4
New Hampshire	4
Rhode Island	4
California	2
Connecticut	2
Indiana	2
Iowa	2
Maryland	2

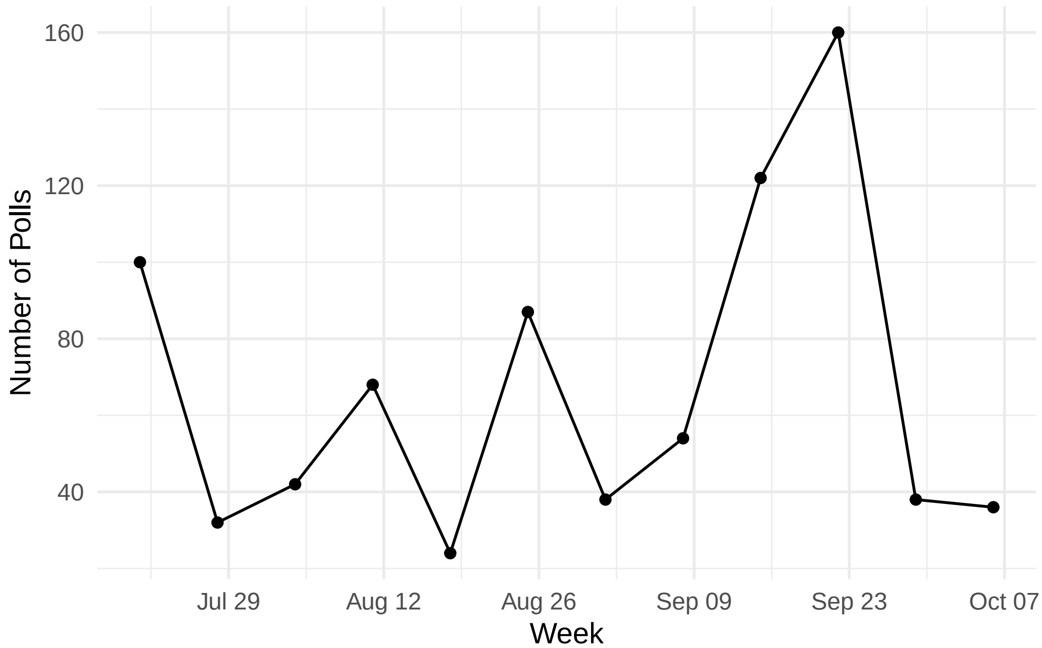

Figure 5: Number of Polls Over Time

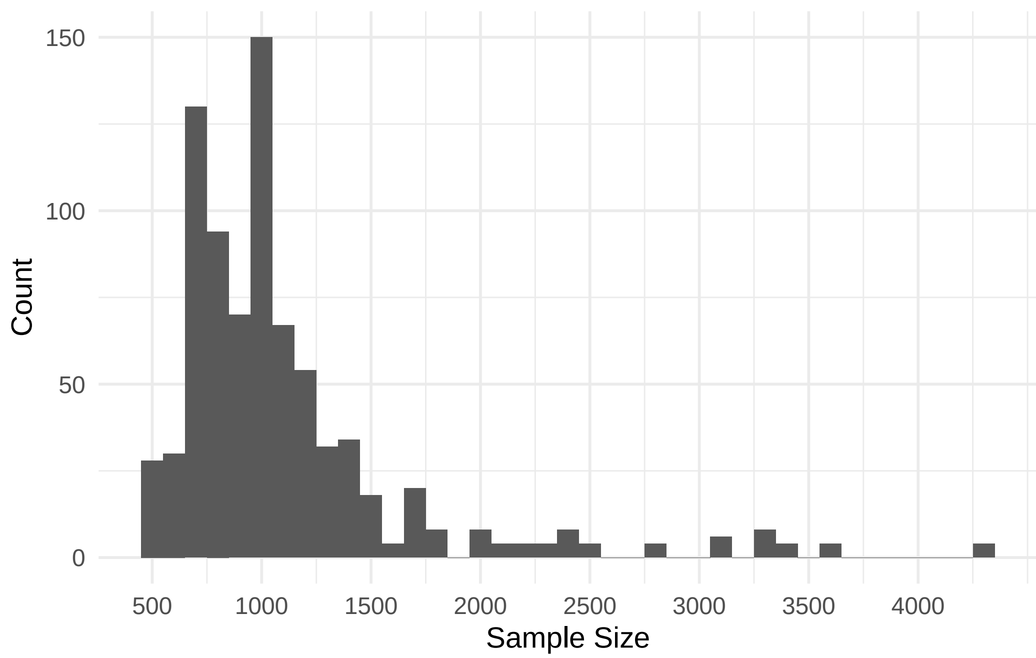

Figure 6: Distribution of Poll Sample Sizes

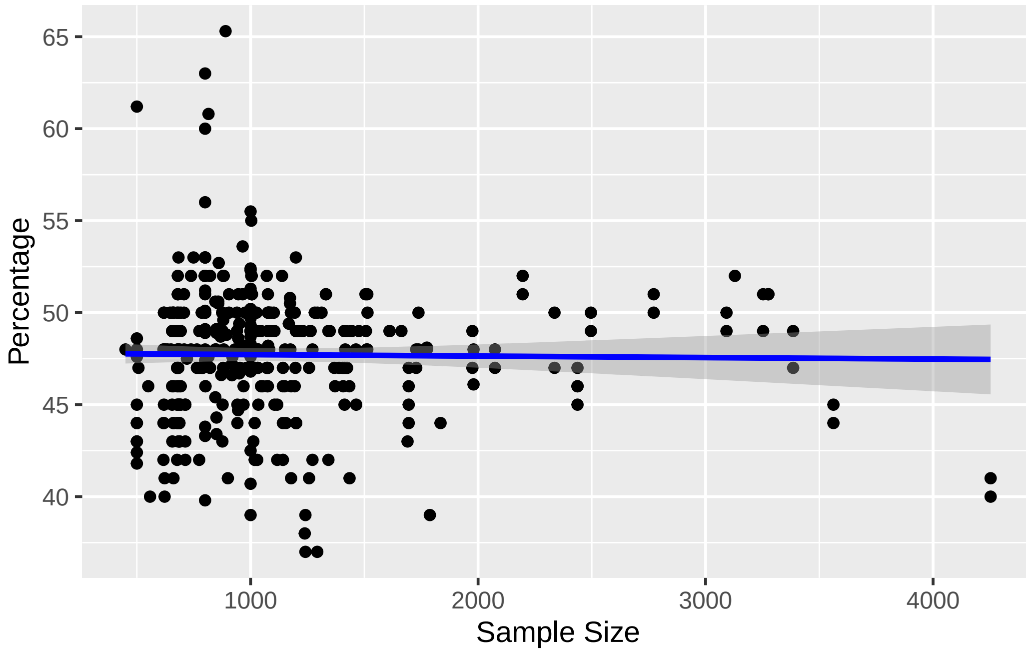

Figure 7: Linear Regression of Percentage vs Sample Size for Kamala Harris

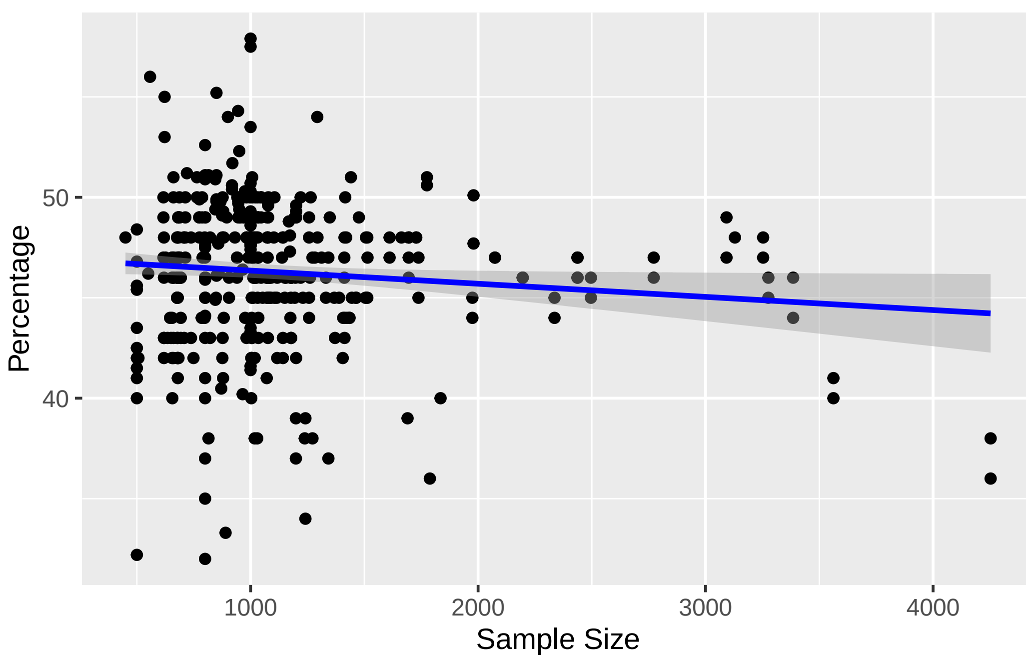

Figure 8: Linear Regression of Percentage vs Sample Size for Donald Trump

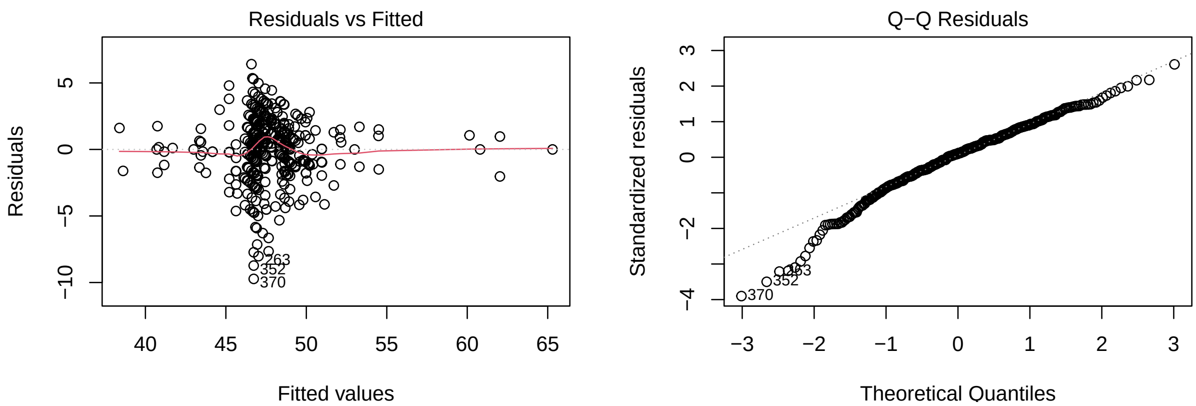

Figure 9: Multi-Linear Regression model for Kamala Harris

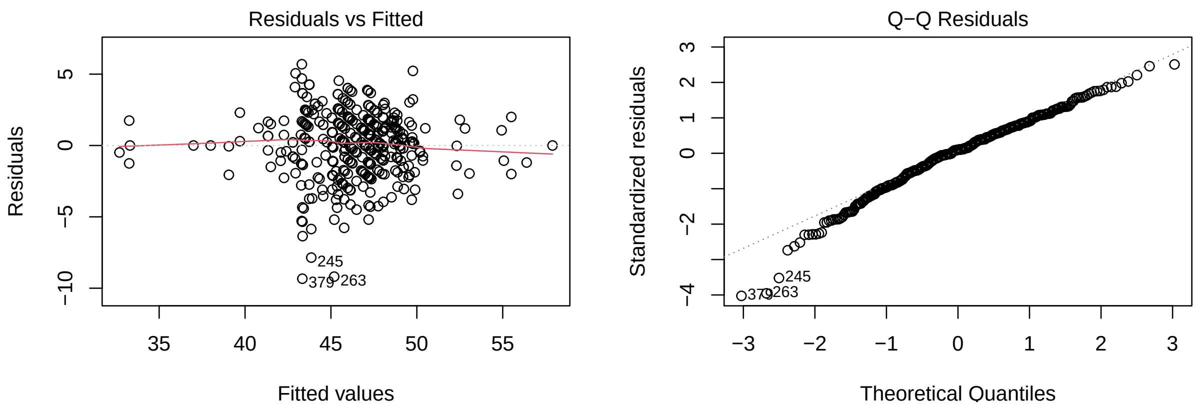

Figure 10: Multi-Linear Regression model for Donald Trump

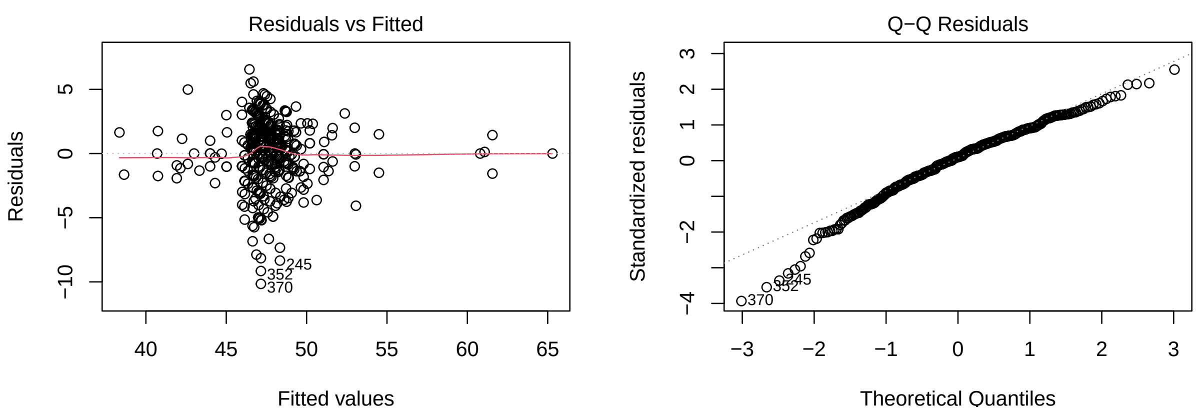

Figure 11: Harris Refined Multi-Linear Regression Model

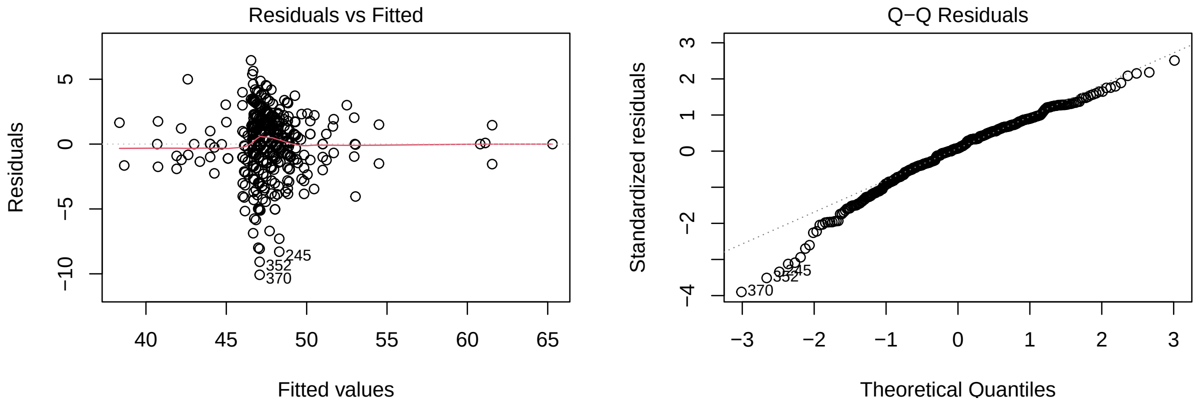

Figure 12: Harris Final Multi-Linear Regression Model

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/.

Fox, John, and Sanford Weisberg. 2019. *An r Companion to Applied Regression*. https://cran.r-project.org/package=car.

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.