Forecasting the 2024 U.S. Presidential Election*

My subtitle if needed

Tim Chen Steven Li Tommy Fu

October 24, 2024

Analysis of polling data from July to October 2024 reveals significant challenges in forecasting the U.S. presidential election between Kamala Harris and Donald Trump, with state-level variations and pollster methodology playing crucial roles in prediction accuracy. Using multi-linear regression models incorporating pollster ratings, sample sizes, and state-level data, we find that Harris is projected to receive 216 electoral votes while Trump is predicted to secure 147 electoral votes. State-level polling variations and methodological differences among pollsters significantly impact predicted outcomes, with pollster reliability metrics explaining substantial variance in candidate support levels. These findings highlight the continued importance of refining polling methodologies and expanding state-level coverage to improve electoral forecasting accuracy, particularly in an era of increasing political polarization.

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:

- **Pollster rating (Pollscore):** A numerical score reflecting the reliability and historical accuracy of each polling organization.
- **Sample size:** The number of respondents included in each poll, which influences the poll's margin of error.
- **Support percentage (pct):** The percentage of respondents expressing support for each candidate.
- **State:** The U.S. state where the poll was conducted, or in some cases, national-level polling data.
- …

Table 1 presents a sample of the processed dataset, showcasing the key variables necessary for the analysis.

Table 1: Sample Overview of Selected Variables in the Polling Dataset

Poll ID	Pollster	Pollscore	Numeric Grade	Transparency Score	Sample Size	pct	State	Methodology	Candidate	End Date
88071	YouGov	-1.1	3.0	9	1078	50	Pennsylvania	Online Panel	Kamala Harris	2024-09-06
88019	CNN/SSRS	-0.6	2.8	10	789	47	Pennsylvania	Probability Panel	Kamala Harris	2024-08-29
88380	Beacon/Shaw	-1.1	2.8	9	764	48	Arizona	Live Phone/Text-to-Web	Kamala Harris	2024-09-24
87918	YouGov	-1.1	3.0	9	1788	36	National	Online Panel	Donald Trump	2024-08-26
88402	Beacon/Shaw	-1.1	2.8	9	991	50	North Carolina	Live Phone/Text-to-Web	Kamala Harris	2024-09-24

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

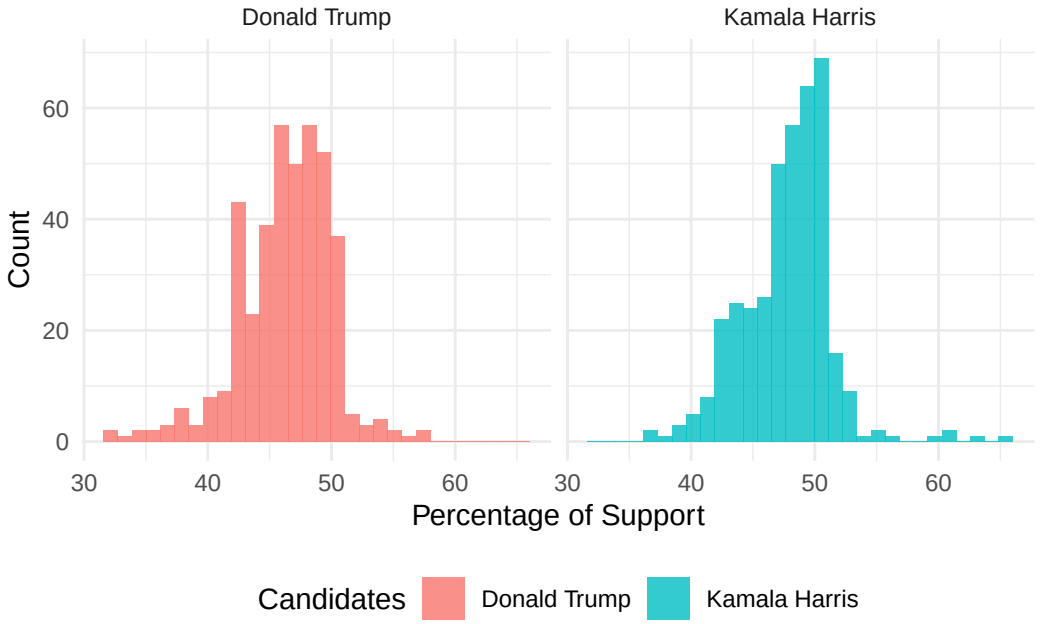

Figure 1: Distribution of Support for Kamala Harris and Donald Trump

- **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

3.1 Kamala Harris

We begin by modeling the percentage of support for Kamala Harris using a linear regression model. The predictors include the sample size, pollster ratings (e.g., pollscore and transparency score), and state. This model aims to quantify how these variables influence her support across different polls.

The results of the linear regression for Kamala Harris show that sample size has a statistically insignificant effect on her support. This suggests that factors other than sample size, such as pollster methodology or regional biases, may play a more significant role in determining the level of support she receives.

Figure 2 illustrates the relationship between the sample size and support percentage for Kamala Harris.

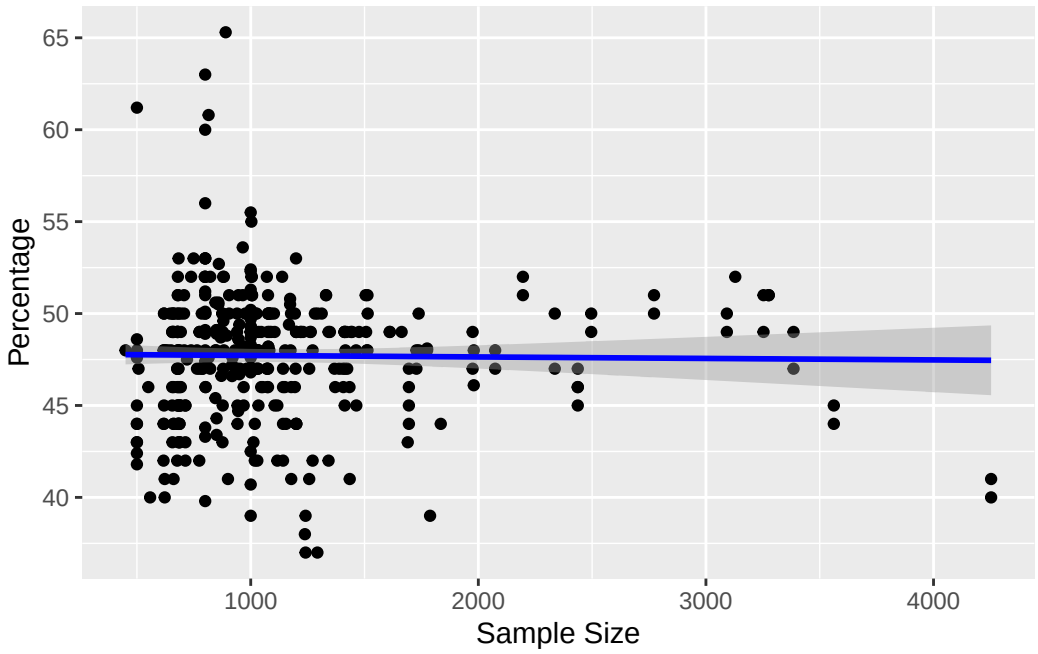

Figure 2: Linear Regression of Percentage vs Sample Size for Kamala Harris

3.2 Donald Trump

A similar linear regression model was applied to Donald Trump's data. The predictors remain the same, and the goal is to determine the factors that drive support for him.

The results in indicate that sample size has a weak but statistically significant negative effect on support for Trump. This suggests that larger polls tend to show slightly lower support for Trump, although the effect size is small.

Figure 3 shows the relationship between the sample size and support percentage for Donald Trump.

3.3 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 several predictors, including pollster rating (represented by pollscore), transparency score, sample size, and state-level data. By accounting for these factors, the MLR models allow us to control for more variables that influence voter support and provide more accurate predictions.

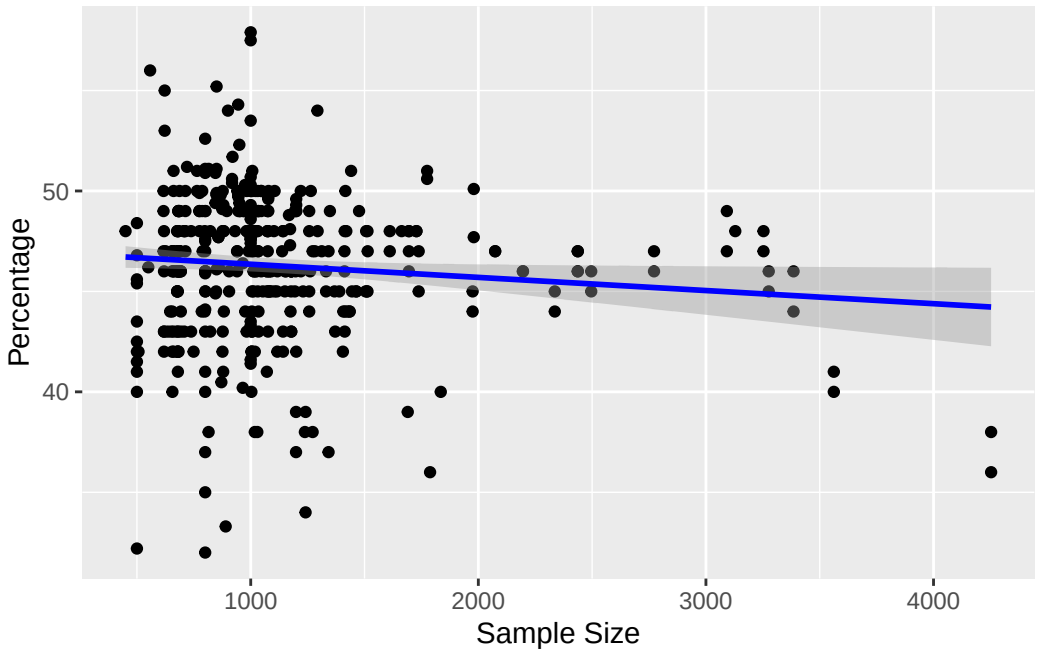

Figure 3: Linear Regression of Percentage vs Sample Size for Donald Trump

3.3.1 Kamala Harris Model

The MLR model for Kamala Harris takes into account multiple factors that may impact her support across different polls. This includes not just the sample size of the poll, but also how pollster reliability (numeric_grade), transparency of poll data (transparency_score), and state-level polling contribute to predicting voter preferences.

The results Figure 4 show that several predictors, including state and pollster rating, significantly impact her predicted percentage of voter support. These results suggest that voter preferences for Harris vary widely depending on the state and the credibility of the pollster.

The model diagnostics (such as residual plots, QQ plots) were evaluated to ensure the assumptions of linear regression hold. Figure X shows diagnostic plots, which indicate that the model performs reasonably well in terms of residual behavior and normality. Similarly, we build an MLR model for Donald Trump.

3.3.2 Donald Trump Model

Similarly, we constructed an MLR model for Donald Trump using the same set of predictors to assess the factors that influence his support across the country.

The results for Donald Trump in Figure 5 reveal that state-level variations and pollster transparency play a significant role in explaining his level of support. The model for Trump also passes diagnostic checks, as shown in Figure Y, indicating that the assumptions of linearity, independence, and homoscedasticity are reasonably met. Check Multi-collinearity using Variance Inflation Factor (VIF)

3.4 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 2: 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 2, 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 3: 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

3.5 Stepwise Model Selection

To further optimize the MLR models, we performed stepwise model selection, which systematically adds or removes predictors to minimize the Akaike Information Criterion (AIC) and improve model fit.

The stepwise model selection improved the model by retaining the most significant predictors and eliminating those with little explanatory power.

3.6 Final Predictions for Electoral College Votes

Using the final models for both candidates, we predicted the percentage of support in each state and aggregated the results to simulate the Electoral College outcome. The predicted percentage of support for each candidate is used to determine the likely winner in each state.

4 Electoral College Prediction

To forecast the winner of the 2024 election, we aggregate the predicted percentages of support from our models for each state and calculate the Electoral College votes. The candidate with 270 or more Electoral College votes is predicted to win the election.

```
[1] "Harris Electoral Votes: 216"
```

```
[1] "Trump Electoral Votes: 147"
```

```
[1] "No candidate reached 270 electoral votes."
```

According to our models, Kamala Harris is projected to receive 216 electoral votes, while Donald Trump is predicted to win 147 electoral votes.

5 Discussion

5.1 Key Findings

The multiple linear regression (MLR) models developed in this paper provide valuable insights into the factors that influence voter support for Kamala Harris and Donald Trump. Through the analysis of polling data, several key predictors were identified as significant, including state, pollster rating (pollscore), and transparency score.

The MLR model for Kamala Harris highlights the importance of state-level factors in determining her support. The model shows that voter preferences vary significantly across different states, with certain states (e.g., California and Maryland) showing higher support levels, while others (e.g., Indiana and Missouri) show lower support. Pollster reliability, captured through numeric grades and transparency scores, also plays an important role, suggesting that voters may respond differently based on the credibility of the pollster.

Donald Trump Similarly, the MLR model for Donald Trump reveals that state-level variations are critical to understanding his level of support. The model indicates that Trump's support is more stable across certain states, but there are also notable outliers where his support fluctuates. Pollster characteristics, such as pollscore and transparency score, significantly impact Trump's predicted support, reflecting the importance of poll quality in predicting election outcomes.

Electoral College Forecast Aggregating state-level predictions into Electoral College votes demonstrates the potential competitiveness of the 2024 U.S. Presidential Election. Harris is projected to receive 216 electoral votes, while Trump is predicted to receive 147 electoral votes. This forecast suggests that the election remains highly uncertain, with several key battleground states likely determining the final outcome.

5.2 Model Strengths

One of the main strengths of this analysis is the incorporation of multiple predictors that allow us to account for various factors influencing voter support. By considering state-level data, pollster ratings, and transparency scores, the models provide a more nuanced prediction of voter behavior compared to models that rely solely on national-level polling.

The use of stepwise model selection and multicollinearity checks further enhanced the robustness of the models by optimizing the choice of predictors and ensuring that the models do not suffer from unstable estimates caused by correlated predictors.

Additionally, the aggregation of state-level predictions into Electoral College outcomes presents a more realistic forecast of the election, as the U.S. Presidential election is ultimately decided by electoral votes, not popular votes.

5.3 Limitations

Despite the strengths of the models, there are several limitations that should be addressed:

- **Poll Reliability and Sampling Bias:** The accuracy of the predictions depends heavily on the quality of the polling data. Although the model accounts for pollster ratings, polling methodologies can still introduce biases, particularly in states with limited polling data. Sampling errors and non-response biases could skew the results, especially in smaller states or regions with inconsistent polling coverage.
- **Static Prediction:** The model provides a static snapshot of voter support based on current polling data, which may not capture the dynamic nature of voter preferences over time. As election day approaches, voter preferences may shift due to campaign events, debates, or other external factors. Without time-series data, the model may fail to account for these trends.

- **Unaccounted Variables:** Although the model includes important predictors like state, pollscore, and transparency score, other potentially influential factors, such as economic conditions, campaign spending, and voter turnout, are not included in the analysis.These unaccounted variables may introduce inaccuracies in the final predictions.

5.4 Future Research

Directions Future work could improve upon this analysis by addressing some of the limitations mentioned above. For instance, incorporating time-series data could allow the model to capture how voter preferences evolve in response to external factors such as campaign events, economic developments, and political endorsements. A dynamic forecasting model that updates predictions as new polls are released would provide more timely and accurate forecasts.

Moreover, integrating other influential variables, such as economic indicators (e.g., unemployment rates, inflation), voter turnout models, and campaign spending data, could enhance the predictive power of the model. Including demographic data (e.g., age, education, income) could also improve the granularity of predictions, especially in battleground states where demographic shifts are critical to election outcomes.

Finally, expanding the model to account for ranked-choice voting in certain states could provide a more accurate forecast in scenarios where third-party candidates or run-off elections play a significant role.

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.

A.3 Multi-Linear Regression Models

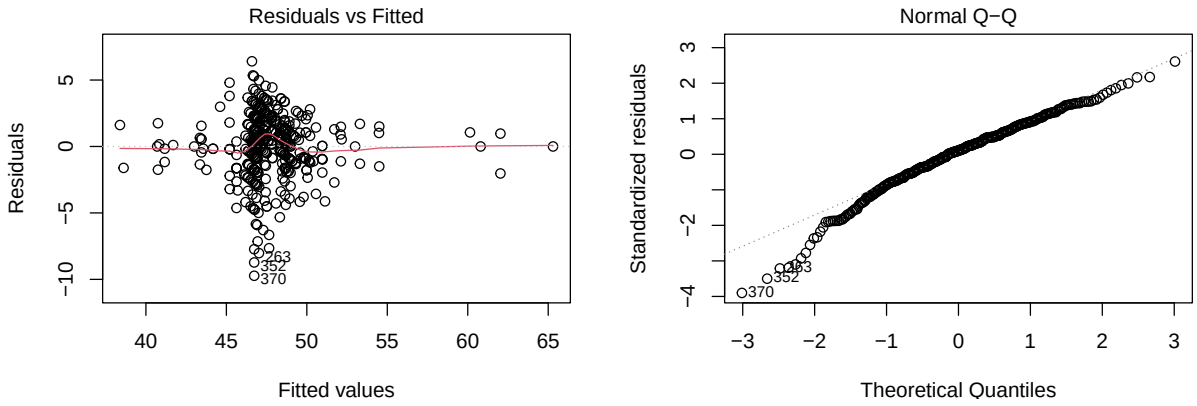

Figure 4: Multi-Linear Regression model for Kamala Harris

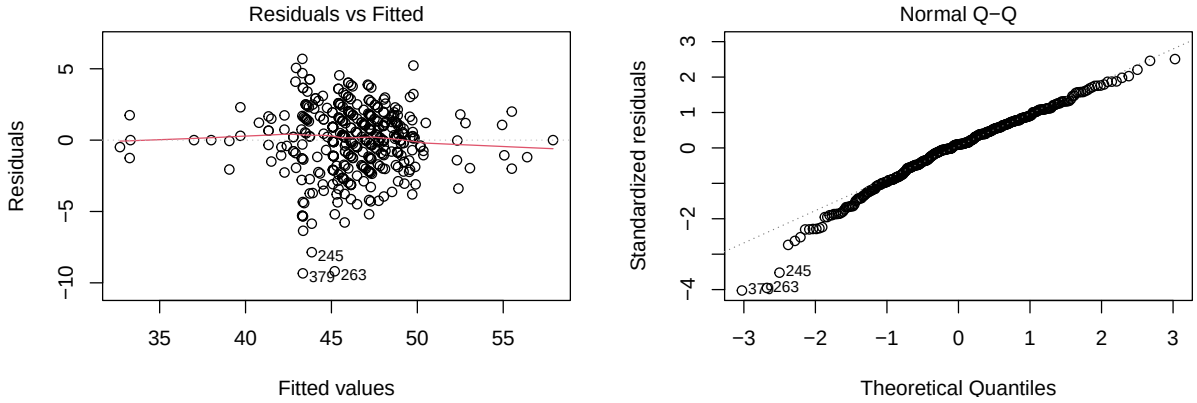

Figure 5: Multi-Linear Regression model for Donald Trump

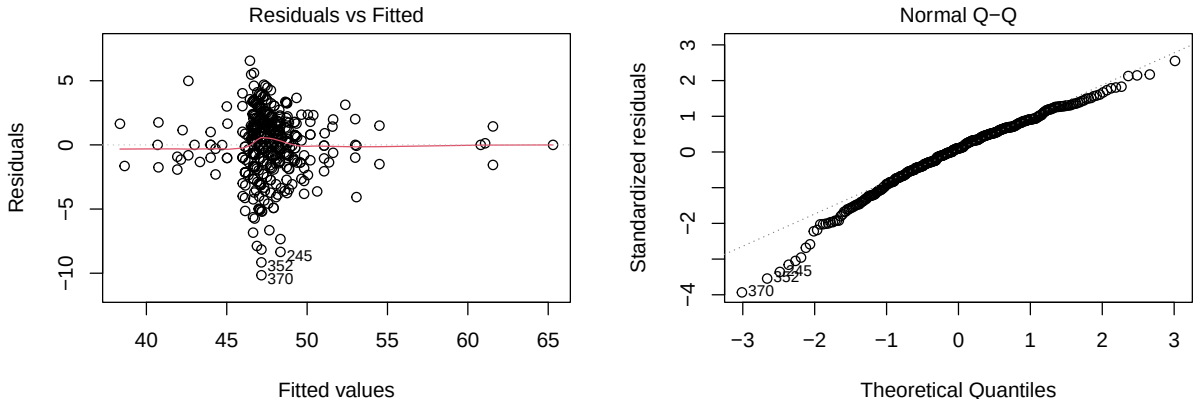

Figure 6: Harris Refined Multi-Linear Regression Model

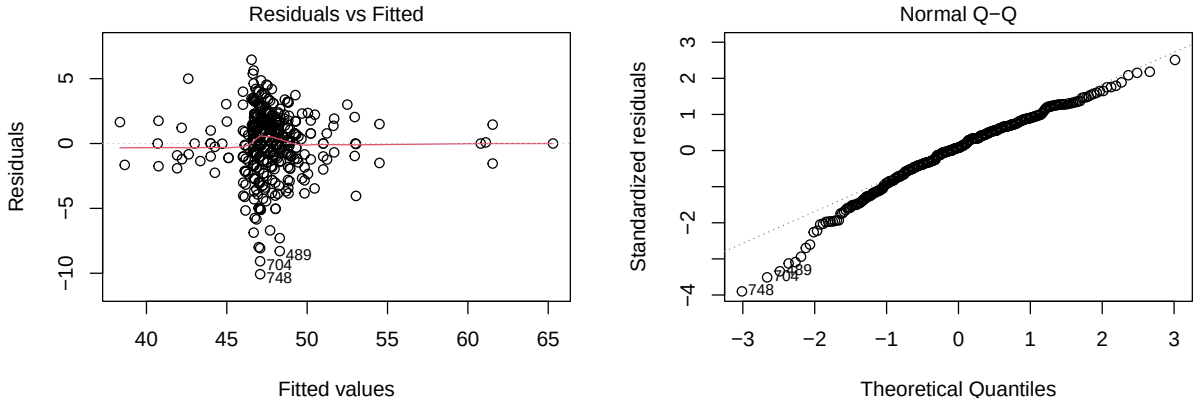

Figure 7: Harris Final Multi-Linear Regression Model

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.