Forecasting the 2024 U.S. Presidential Election*

My subtitle if needed

Tim Chen   Steven Li   Tommy Fu

October 22, 2024

This paper presents a statistical model for forecasting the outcome of the 2024 U.S. Presidential Election. Using polling data from various sources, we develop multiple linear regression models to predict the percentage of support for the main candidates, Kamala Harris and Donald Trump, in different states. We aggregate these predictions to simulate the Electoral College vote and estimate the likelihood of either candidate winning. Our results suggest that neither candidate currently secures the required 270 electoral votes to win the presidency.

Table of contents

---

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

1 Introduction

Forecasting elections has long been one of the most challenging tasks for political scientists, statisticians, and analysts alike. The complexity of predicting voter behavior in large, diverse electorates like that of the United States stems from numerous factors, including ever-shifting public opinion, rapidly changing political climates, and varying levels of voter engagement. The 2024 U.S. Presidential Election is no exception. With polarized voter bases, unpredictable external factors, and the growing influence of non-traditional media, the race remains highly uncertain, making reliable forecasting models more important—and difficult—than ever.

Central to any electoral forecast is the use of polling data. Polls provide snapshots of voter intentions at specific points in time, shaping both public perceptions and expert expectations. Yet, polling comes with its own set of challenges. Pollsters employ different methodologies—ranging from online surveys to phone interviews—each with its own inherent biases. Sample populations may vary in size and representativeness, potentially skewing results in favor of certain demographic groups. Moreover, polling accuracy can be further complicated by factors such as non-response bias, the changing nature of the electorate, and the rise of unconventional voting patterns, particularly among younger and minority voters.

In this paper, we attempt to forecast the outcome of the 2024 U.S. Presidential Election by developing statistical models based on polling data from a variety of sources. By leveraging multiple linear regression models, we aim to predict the percentage of support for the main candidates—Kamala Harris and Donald Trump—across different states. Our analysis incorporates key variables such as pollster reliability, sample size, and state-level demographics, all of which contribute to shaping the electoral landscape. Ultimately, we aggregate these state-level predictions to simulate the outcome of the Electoral College, providing insights into the likelihood of either candidate securing the 270 votes required to win the presidency.

The rest of this paper is structured as follows: Section 2 discusses the data used for this analysis, including key variables and sources. Section 3 outlines the models developed for each candidate and presents the corresponding results. Section 4 discusses the aggregated Electoral College predictions based on the model outputs. Finally, Section 5 provides conclusions and offers suggestions for future research.

2 Data

2.1 Overview

The dataset used in this analysis draws from a variety of polling sources, providing comprehensive information about polling organizations, sample sizes, methodologies, and state-level data. Several key variables are crucial to the analysis:

- **Pollster rating:** 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:** 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.

To ensure accuracy and consistency, the data has been meticulously processed and cleaned. This involved aligning all variables for compatibility with regression analysis, converting relevant fields (e.g. support percentage) into numeric formats, and appropriately handling missing data to minimize potential biases. These reprocessing steps ensure the dataset is tidy and ready for model development.

Table 1 shows a sample of the dataset.

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

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

2.2 Measurement and Limitations

The primary limitations of this analysis stem from the quality and variability of polling data, which can introduce biases through factors like sample size and pollster methodology. Additionally, the model provides a static snapshot of voter preferences, without accounting for the dynamic nature of elections For instance, shifts in public opinion over time or the impact of campaign events. The exclusion of key external factors, such as economic conditions or voter turnout, also limits the model's ability to fully capture the complexities of election outcomes. Furthermore, the simplified Electoral College simulation assumes that polling data will accurately predict state-level results, which may not always be the case, especially in battleground states with volatile voter behavior.

2.3 Outcome variables

Our primary outcome variable is the percentage of support for Kamala Harris and Donald Trump in each poll. This variable is central to our analysis because it directly captures voter preference, which is the most relevant metric for forecasting election outcomes. By modeling the percentage of support, we can quantify how various factors—such as pollster quality, sample size, and state-level dynamics—influence the candidates' standings in the polls. This outcome variable also allows us to simulate the Electoral College results by aggregating predicted support across different states, which is crucial for determining the likelihood of either candidate winning the presidency.

Figure 1 below shows the distribution of support for both candidates across the polls included in the dataset.
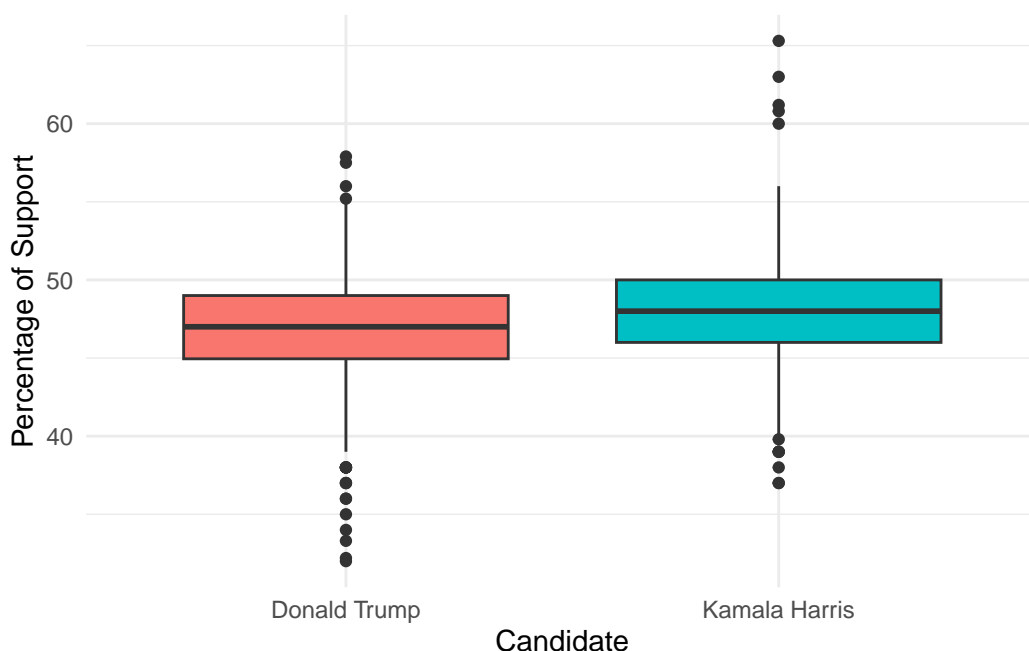

Figure 1: Distribution of Support for Kamala Harris and Donald Trump

2.4 Predictor variables

In our analysis, we utilized several key predictor variables to model voter support for Kamala Harris and Donald Trump in the 2024 U.S. Presidential Election. These predictor variables were selected based on their expected influence on polling results and their ability to account for variations in polling accuracy and voter behavior across different states.

One of the key predictors is numeric grade, which represents a composite measure of poll quality. This variable accounts for various aspects of a poll's reliability, such as the transparency of

its methodology and the pollster's historical accuracy. Pollscore, another important predictor, provides a similar measure of poll reliability but focuses more on the specific performance of the polling organization in recent elections.

Transparency score measures how openly polling organizations disclose their methodology and data, which can significantly impact the trustworthiness of the results. Polls with higher transparency scores are given more weight in the model, as they are generally considered more reliable. The sample size of each poll is also included, as larger sample sizes tend to reduce the margin of error and provide more precise estimates of voter support.

Geographic differences in voter preferences are captured through the state variable, which accounts for regional variations in political culture, demographics, and historical voting patterns. Lastly, methodology refers to the techniques and approaches used by the polling organizations, such as whether the poll was conducted online, by phone, or through other means.

All predictor variables—numeric grade, pollscore, transparency score, sample size, state, and methodology—are presented in Table 1.

2.5 Cleaning Process and Analysis

The data cleaning process uses R (R Core Team 2023) as well as other packages such as tidyverse (Wickham et al. 2019), dplyr (Wickham et al. 2023), janitor (Firke 2023) and lubridate (Grolemund and Wickham 2011). Begin by standardizing column name to ensure consistency and avoid errors during analysis. The dataset was filtered to focus solely on Kamala Harris and Donald Trump, with polls below a numeric grade of 2.7 excluded to maintain high-quality data. Missing state information was categorized as "National" polls, and the poll end dates were converted to a date format to include only polls conducted after the candidates officially declared their candidacies. This ensures the analysis reflects relevant, up-to-date voter preferences. Additionally, transformations such as calculating the number of supporters from the percentage support and encoding a binary variable for the candidates (Harris = 1, Trump = 0) were performed to enhance usability in the regression models.

Some variables were excluded for simplicity and relevance. For example, **'population_full'** and **'answer'** were dropped as they provided redundant or unnecessary information for this analysis. However, key variables like **'pollster'**, **'pollscore'**, **'numeric grade'**, and **'sample size'** were retained because they provide essential insights into the quality and precision of the polls, directly influencing the reliability of the models. These steps ensured the cleaned dataset was ready for analysis, with a focus on the most important predictors of voter support while minimizing extraneous information.

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.
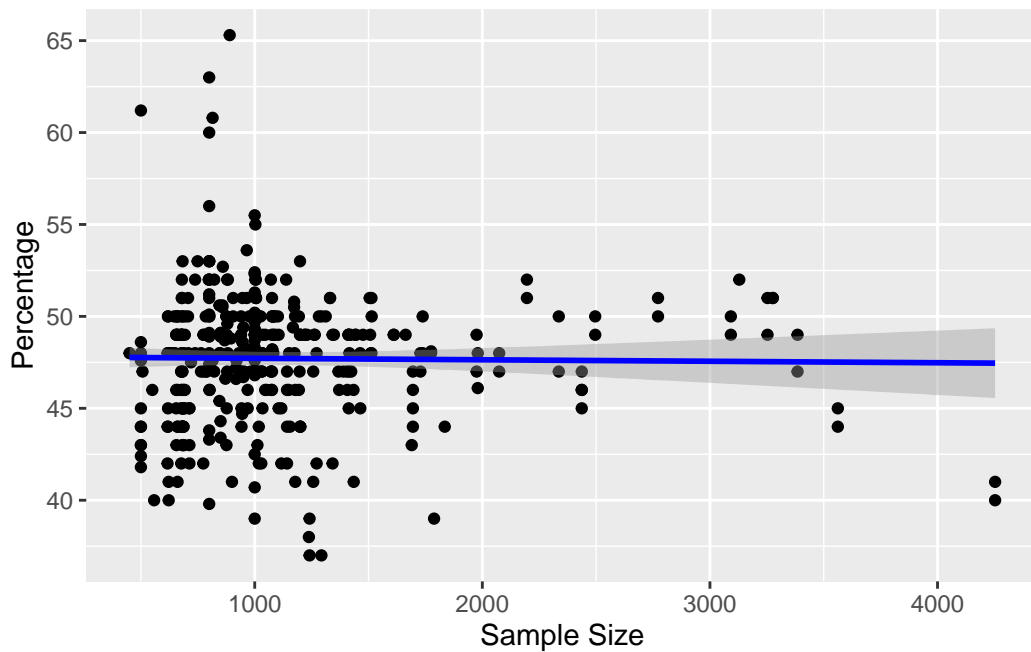

Figure 2: Linear Regression of Percentage vs Sample Size for Kamala Harris

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.
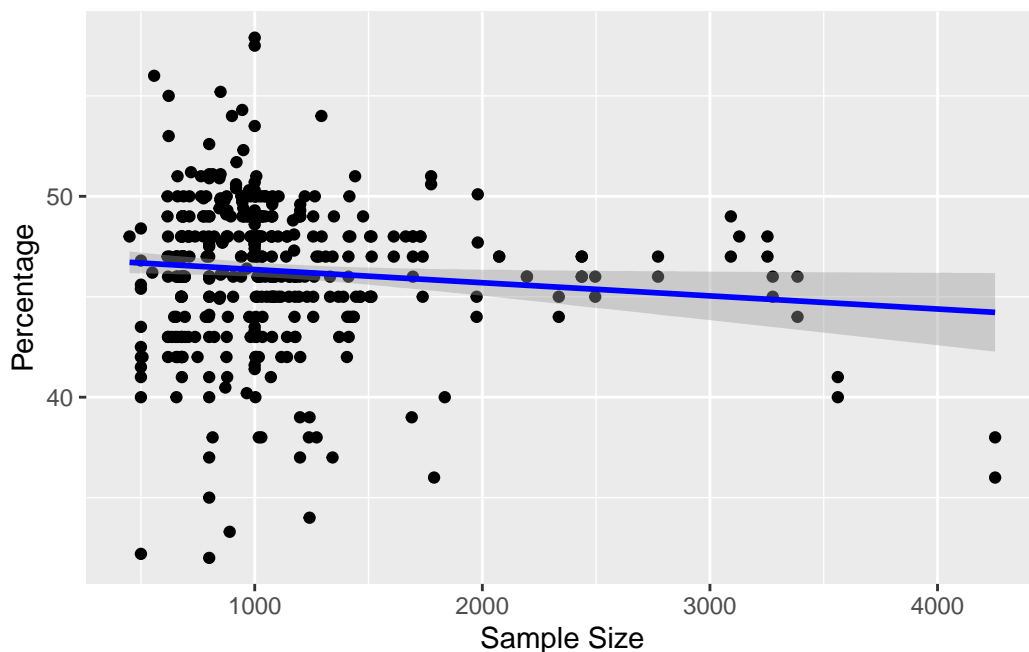

Figure 3: Linear Regression of Percentage vs Sample Size for Donald Trump

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.

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.
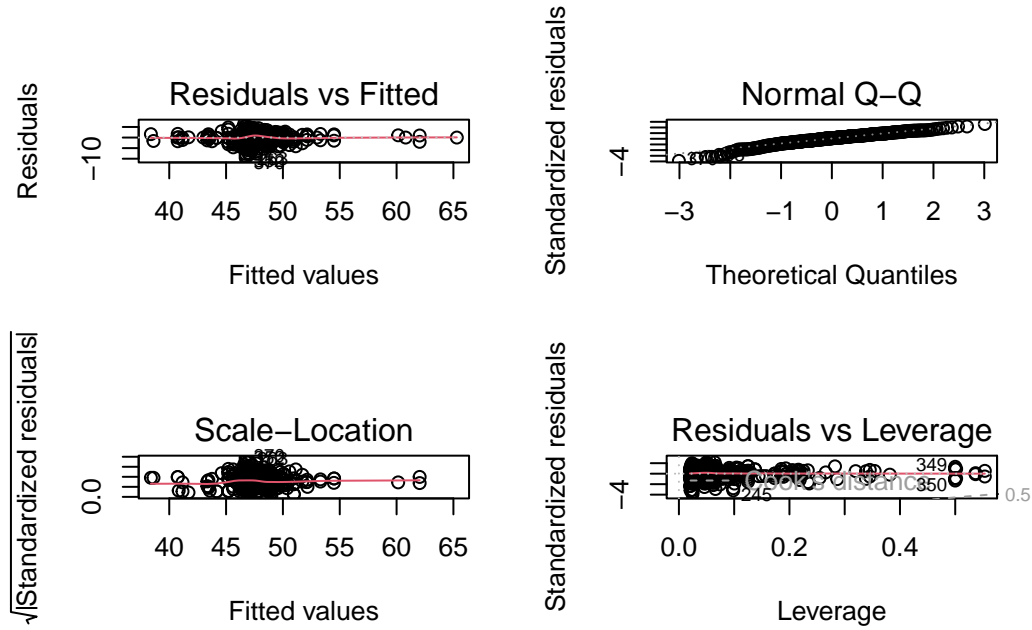
Figure 4: Multi-Linear Regression model for Kamala Harris

The results in 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)
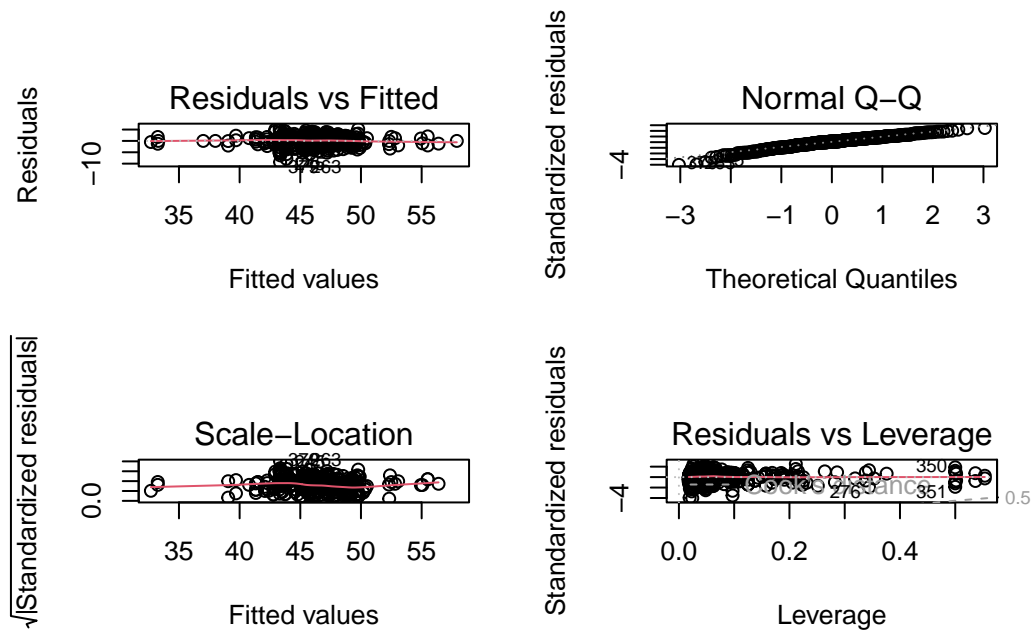
9

Figure 5: Multi-Linear Regression model for Donald Trump

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.
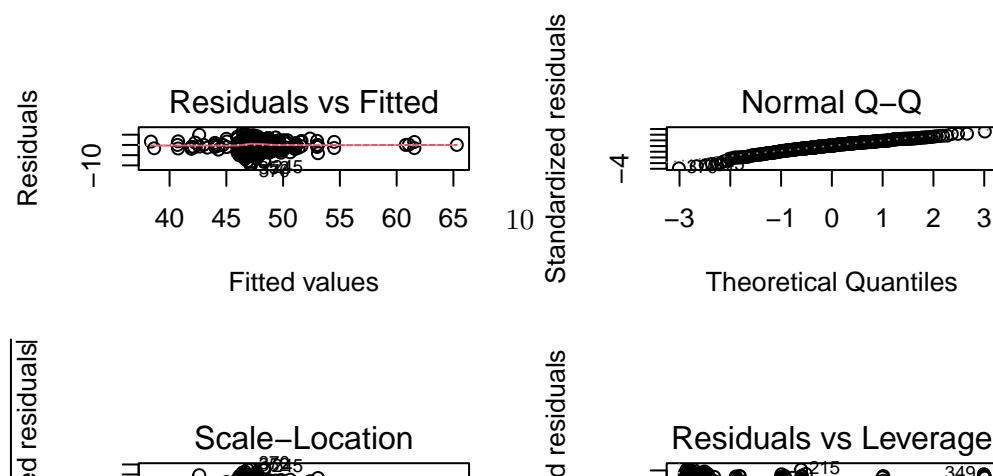
Table 3: Harris Refined MLR model Variance Inflation Factor (VIF)

	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.

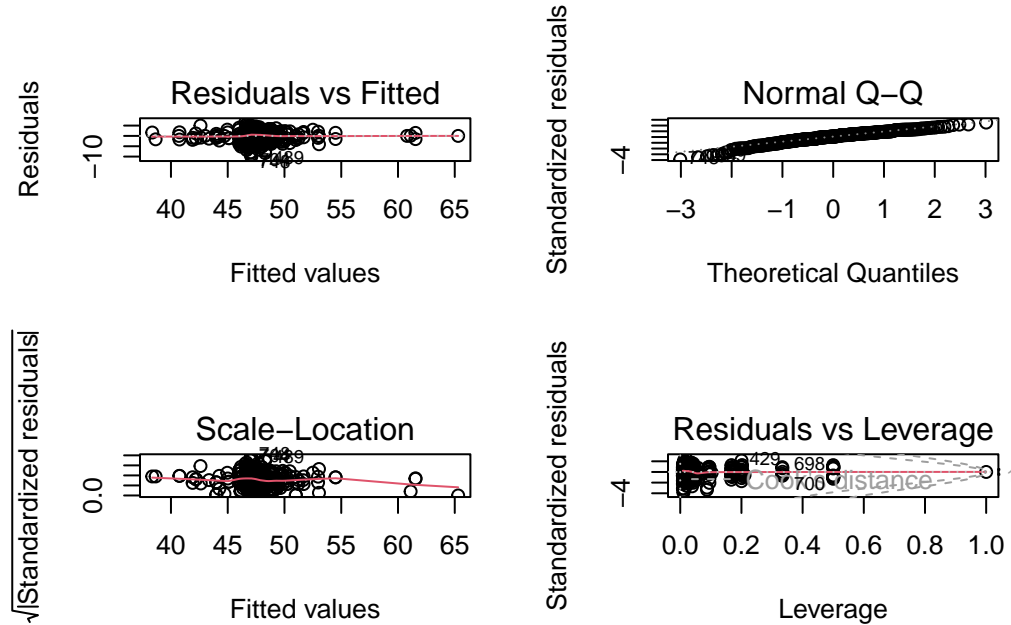
Figure 7: bbb

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, neither candidate secures the 270 Electoral College votes needed to win. Kamala Harris is projected to receive 216 electoral votes, while Donald Trump is predicted to win 147 electoral votes. However, due to the lack of 270 electoral votes for either candidate, the election remains highly competitive, and further developments may shift the balance.

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. According to the predictions, neither Kamala Harris nor Donald Trump currently secures the necessary 270 electoral votes to win. 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.

References

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

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/.

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

Wickham, Hadley, 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.