Forecasting the 2024 U.S. Presidential Election*

My subtitle if needed

Tim Chen Steven Li Tommy Fu

October 24, 2024

Do Abstract after finalizing everything

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 under representation 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.

Edit note: will need to expand/edit after getting final results

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. **?@sec-predict** 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

Steven will do data section

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.
- …other variables

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

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.

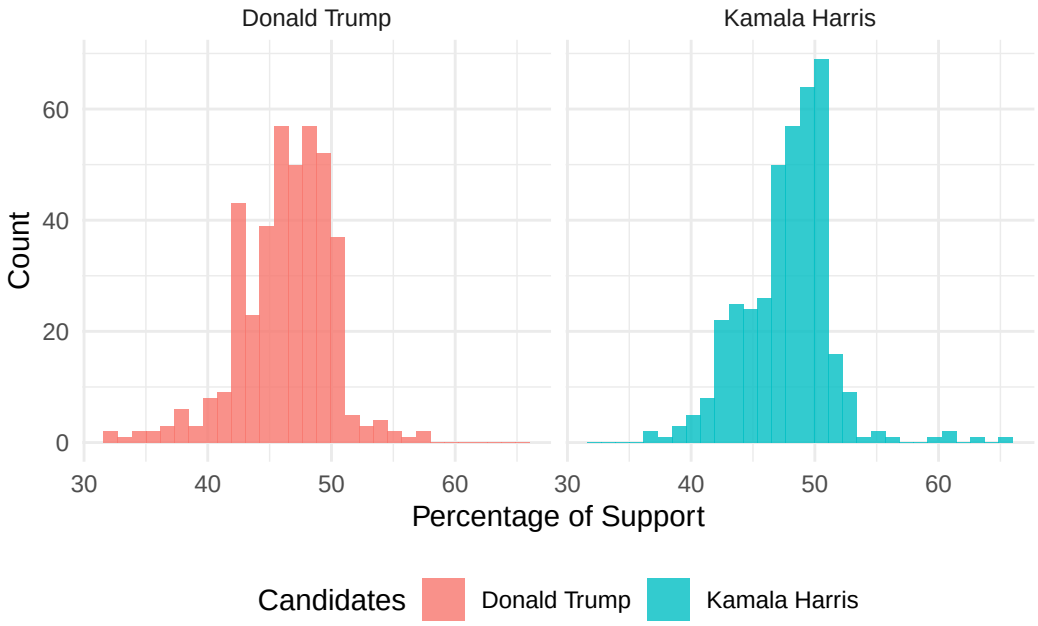

Figure 1: Distribution of Support for Kamala Harris and Donald Trump

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
- **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.

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 ($numeric\_grade \geq 2.7$)
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 Model

3.1 Model Overview and Theoretical Framework

- Justify the choice of multiple linear regression for election forecasting
- Discuss why this approach is appropriate given the data structure
- Explain how the model addresses the unique challenges of electoral prediction
- Define the mathematical notation for the model
- Explain each component of the model specification
- Discuss underlying assumptions and their implications

3.2 Baseline Model Development

- Present the initial model specification with all variables
- Explain the 80-20 train-test split methodology
- Detail the model validation approach
- Present mathematical notation:

Report key metrics such as, but not all required. U decide:

- R-squared values
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Out-of-sample testing accuracy

3.3 Model Refinement Using VIF Analysis

- Explain the importance of addressing multicollinearity
- Present VIF analysis methodology and thresholds used
- Show which variables were retained/removed and why
- Compare refined model performance to baseline
- Present updated mathematical notation
- Report comparative metrics showing improvement

3.4 Stepwise Model Selection

- Explain the stepwise selection methodology
- Justify the selection criteria (AIC/BIC)
- Present the final model specification
- Compare performance metrics across all three models
- Discuss why this represents the optimal model

3.5 Model Validation and Diagnostics

- Present residual analysis
- Discuss normality assumptions
- Address heteroscedasticity
- Present cross-validation results
- Discuss model limitations

4 Results

4.1 Model Performance Comparison

- Present comprehensive comparison table of all three models
- Include all relevant statistical measures
- Visualize performance metrics
- Discuss which model performs best and why

4.2 State-Level Predictions

- Present predictions for key battleground states
- Compare model predictions with recent polling averages
- Discuss uncertainty in predictions
- Present confidence intervals

4.3 Electoral College Projections

- Aggregate state-level predictions into electoral votes
- Present probability distributions for different outcomes
- Discuss implications for election outcome
- Address uncertainty in projections

4.4 Model Insights

- Discuss which variables are most influential
- Present effect sizes for key predictors
- Analyze regional variations in model performance
- Identify any surprising or counterintuitive findings

5 Discussion

5.1 Key Findings

- Summarize main results and their implications
- Compare findings with existing literature
- Discuss how results align with or differ from previous election modeling efforts
- Address practical significance of findings

5.2 Model Strengths

- Discuss advantages of the final model
- Compare with other approaches in literature
- Highlight unique contributions
- Explain practical applications

5.3 Limitations

- Address data limitations
- Discuss model assumptions and their impact
- Consider external validity
- Acknowledge potential sources of error

5.4 Future Research Directions

- Suggest methodological improvements
- Propose additional variables to consider
- Discuss potential extensions of the model
- Recommend practical applications

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 Sample Data

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

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

A.4 Multi-Linear Regression Models

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.