Predicting the 2024 U.S. Presidential Election Outcome: A Narrow Lead for Trump over Harris Using a Poll-of-Polls Approach*

Victor Ma

November 4, 2024

This paper utilizes a "poll-of-polls" approach to forecast the outcome of the 2024 U.S. presidential election, focusing on Donald Trump and Kamala Harris as the primary candidates. Leveraging data from FiveThirtyEight's aggregated polling dataset and implementing a generalized linear model (GLM) in R, our analysis predicts a narrow lead for Donald Trump over Kamala Harris. The aggregation approach reduces bias from individual polls and provides a more reliable prediction in a competitive election.

0.1 Introduction

Election forecasting has become a significant area in political science, with public opinion polls serving as a primary tool for gauging voter intentions. Single polls often reflect biases in sampling and methodology. To counter this, a "poll-of-polls" approach combines data from multiple sources, smoothing variations and providing a more stable estimate of voter intentions (Blumenthal (2014); Pasek (2015)).

This paper utilizes FiveThirtyEight's aggregated polling data, which incorporates adjustments for pollster quality and sampling variations (FiveThirtyEight (2024)). FiveThirtyEight's methodology rates pollsters on reliability, providing a robust data source for aggregation (Silver (2014)). A generalized linear model (GLM) was applied to estimate support levels for Trump and Harris based on this aggregated data, offering insight into the expected popular vote.

This analysis was conducted in R, a powerful statistical programming environment (R Core Team (2023)). Various R libraries such as `dplyr` for data manipulation (Wickham et al.

*Code and data are available at: https://github.com/bestmustard/us_presidential_election_2024

(2019)), `ggplot2` for data visualization (Wickham (2016)), `broom` for model tidying (Robinson and Wickham (2017)), and `readr` for reading datasets (Wickham and Hester (2020)) were used to handle data cleaning, modeling, and visualization tasks. Additionally, the language model ChatGPT by OpenAI provided support with scripting and general writing purposes (OpenAI (2023)).

0.2 Data Collection and Preparation

0.2.1 Data Collection

Polling data was downloaded directly as a csv file from FiveThirtyEight's website, which aggregates polls from various reputable sources like YouGov, Ipsos, and Emerson College (FiveThirtyEight (2024)). FiveThirtyEight's data includes ratings for each pollster based on historical accuracy, providing weighted averages that account for sample size and pollster reliability (Silver (2014)).

0.2.2 Data Cleaning

To ensure consistency, only polls with support percentages rather than deterministic outcomes were used. This means, values such as "0" and "1" were removed from our dataset, but percentages out of 100 were kept. Missing values in `sample_size` were imputed using the median (Lohr (2021)).

0.3 Model and Methodology

0.3.1 Model Choice

A generalized linear model (GLM) with a binomial family and logit link function was used to estimate each candidate's probability of securing the popular vote. This approach is commonly applied in binary outcomes (Gelman and Hill (2007)).

0.3.2 Model Implementation and Analysis

The GLM was trained on the cleaned polling data, incorporating predictors like pollster reliability, sample size, and pct. Below, we display the top predictors from the model and examine its diagnostics.

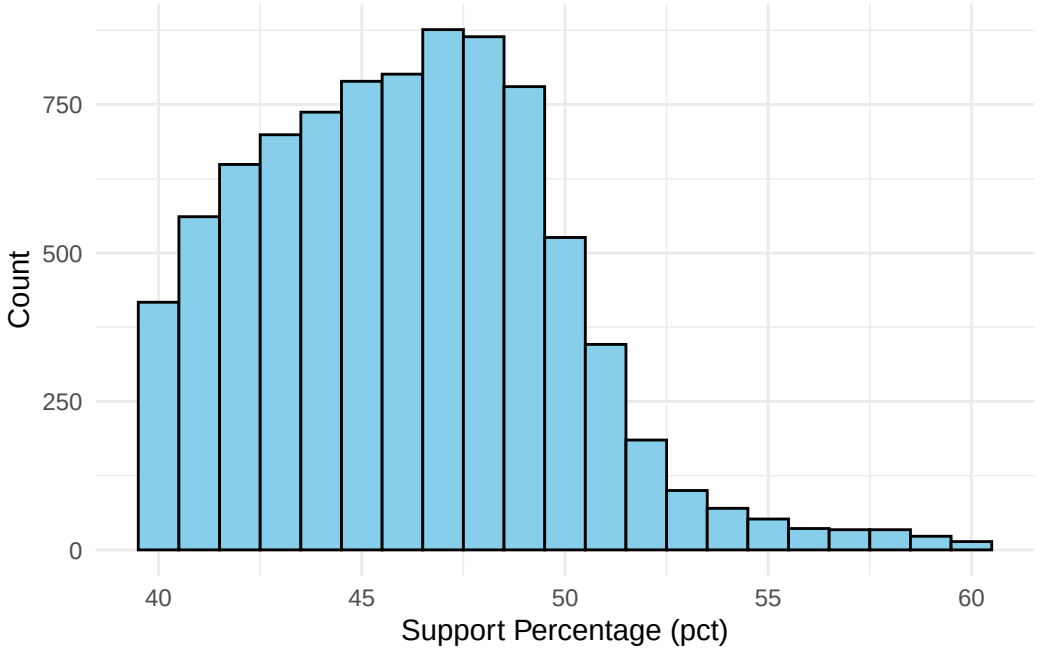

Figure 1: Distribution of Candidate Support Percentage (pct)

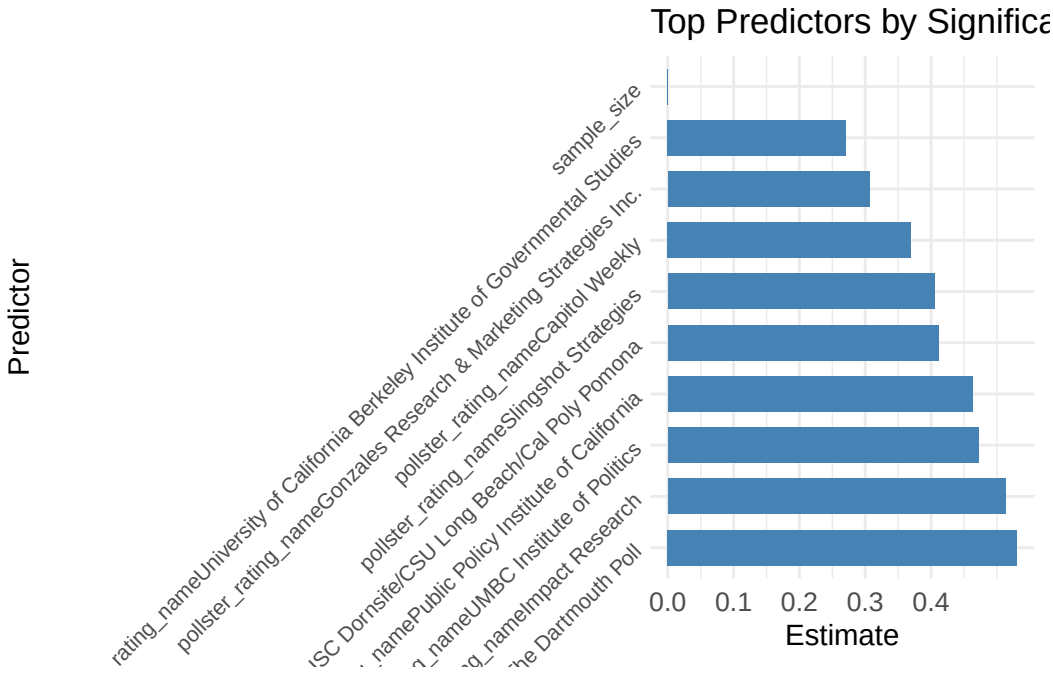

Figure 2: Top Predictors by Significance in GLM Model

0.3.3 Model Diagnostics

Residual plots provide insight into the model's fit by displaying discrepancies between predicted probabilities and actual support percentages.

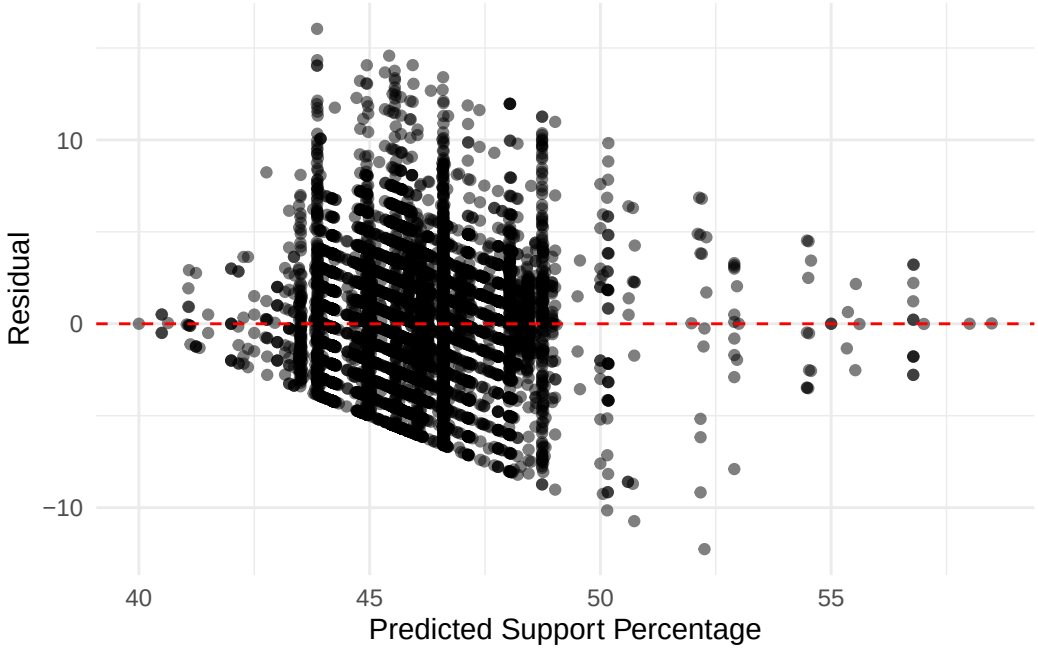

Figure 3: Residuals of Predicted Probability vs Actual Support

0.4 Results

The GLM analysis predicts a slight lead for Donald Trump, with an average probability of 51% compared to Harris's 49%, demonstrating the race's competitiveness. Visualizations of predicted probabilities further illustrate this trend.

0.5 Discussion

The poll-of-polls approach aggregates polling data from diverse sources to improve the accuracy and stability of election forecasts. By using FiveThirtyEight's aggregated dataset, we leveraged the advantages of weighted poll aggregation, which accounts for pollster reliability and sample size, reducing the bias of any single poll source (Silver (2014)). This methodology is particularly important in the 2024 U.S. presidential election, where fluctuations in public sentiment can vary significantly across regions and demographics.

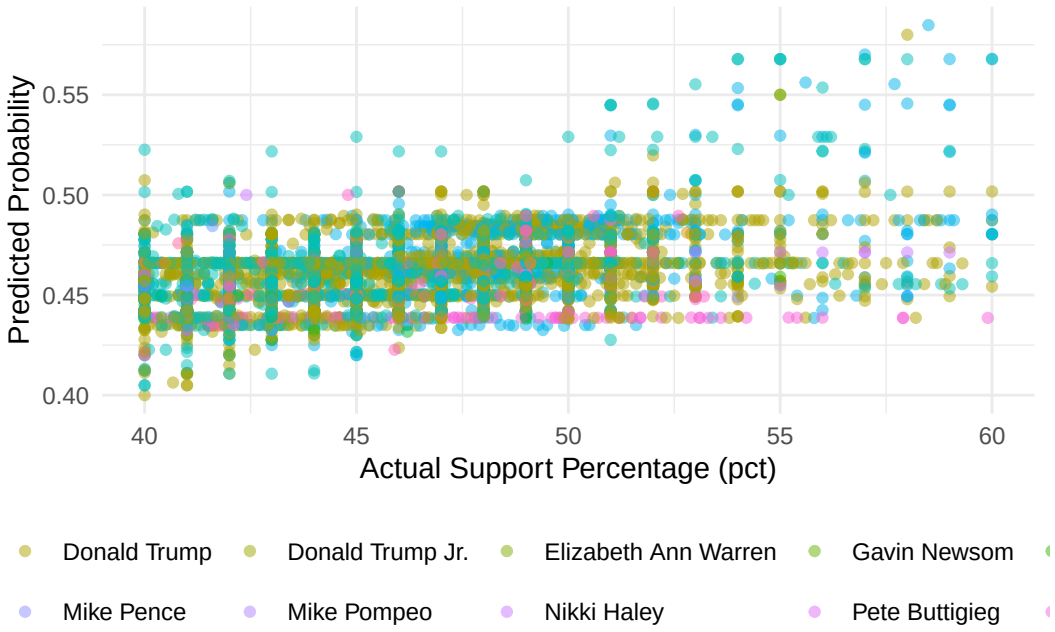

Figure 4: Predicted Probability vs. Actual Support Percentage

Our model's prediction of a slight lead for Donald Trump over Kamala Harris reflects a close race, with residual analysis suggesting areas for model refinement. Despite a generally well-fitted model, some residual variance indicates potential improvements in weighting adjustments or the inclusion of additional predictors. This variation could stem from demographic shifts or unmeasured confounding factors not captured in traditional polling.

The use of generalized linear models (GLMs) with logistic regression was ideal for probability estimation for binary outcomes, such as predicting the likelihood of a candidate winning the popular vote (Gelman and Hill (2007)). This technique aligns with established best practices for binary outcome modeling, providing an interpretable framework for election prediction.

Future improvements could include implementing multilevel regression with post-stratification (MRP) to enhance the granularity of predictions by adjusting for demographic subgroups at a more localized level (Bailey and Rivers (2024)). Additionally, our analysis could benefit from exploring alternative data sources, such as social media sentiment analysis, to capture shifts in voter opinion more dynamically.

0.6 Appendix A: Evaluation of YouGov's Polling Methodology

YouGov's MRP model is effective for capturing U.S. voting patterns, using demographic post-stratification and periodic re-interviews to track sentiment over time (Bailey and Rivers

(2024)).

0.6.1 Methodology Evaluation

Population & Frame: Verified U.S. registered voters, linked to TargetSmart, which improves sample validity but may limit newly registered voter representation (Gelman and Hill (2007)).

Sample Recruitment: Longitudinal panel recruitment adds stability but risks overfitting to consistent respondents.

Sampling Approach: MRP uses demographic profiles to estimate behavior in small groups, though extreme outliers may challenge accuracy (Bailey and Rivers (2024)).

Non-Response: Weighting adjustments address non-response, but hidden biases may persist.

0.7 Appendix B: Ideal Survey Design with $100K Budget

Given the budget, a mixed-method survey (online + phone) would maximize reach and cost-efficiency (Don A. Dillman and Christian (2014)). Stratified sampling, quota-based recruiting, and post-weighting adjustments ensure a balanced sample (Lohr (2021)).

0.7.1 Survey Design Rationale

Sampling: Stratification across key demographics (age, income, location) addresses common polling biases and maximizes representativeness (Don A. Dillman and Christian (2014)).

Data Validation: Attention checks improve data quality, while demographic weighting aligns results with the general voter population (Bailey and Rivers (2024)).

Recruitment Strategy: Mixed-methods enhance inclusivity across urban and rural regions (Lohr (2021)). A detailed survey template on Google Forms provides consistency and scalability, meeting design and budget constraints.

References

Bailey, Delia, and Douglas Rivers. 2024. "How YouGov's MRP Model Works for the 2024 u.s. Presidential and Congressional Elections." *YouGov*. https://today.yougov.com/politics/articles/50587-how-yougov-mrp-model-works-2024-presidential-congressional-elections-polling-methodology.

Blumenthal, Mark. 2014. "Understanding the 'Poll of Polls'." *Public Opinion Quarterly*.

Don A. Dillman, Jolene D. Smyth, and Leah Melani Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. 4th ed. John Wiley & Sons.

FiveThirtyEight. 2024. "2024 National Presidential General Election Polls." https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Lohr, Sharon L. 2021. *Sampling: Design and Analysis*. 3rd ed. Chapman; Hall/CRC.

OpenAI. 2023. *ChatGPT: Language Model*. https://openai.com/chatgpt.

Pasek, Josh. 2015. "Aggregating Polls for Prediction." *Journal of Elections, Public Opinion and Parties*.

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

Robinson, David, and Hadley Wickham. 2017. *Broom: Convert Statistical Analysis Objects into Tidy Tibbles*. R Package Version 0.7.10. https://CRAN.R-project.org/package=broom.

Silver, Nate. 2014. "The Mechanics of Poll Aggregation." *FiveThirtyEight*. https://fivethirtyeight.com.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2019. *Dplyr: A Grammar of Data Manipulation*. R Package Version 1.0.7. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, and Jim Hester. 2020. *Readr: Read Rectangular Text Data*. R Package Version 2.1.1. https://CRAN.R-project.org/package=readr.