Forecasting the 2024 U.S. Presidential Election: A Poll-of-Polls Approach for Predicting the Outcome*

Applying Aggregated Poll Data and Statistical Model to Reveal a Narrow Path to Victory in a Highly Competitive Race

Chendong Fei Xinze Wu Claire Ma

November 4, 2024

This paper develops a model to forecast the outcome of the 2024 U.S. presidential election by analyzing aggregated polling data, or "poll-of-polls," sourced from FiveThirtyEight. Using a generalized linear model, we assess national trends alongside key battleground state polls to predict each candidate's likelihood of victory. The findings indicate a closely contested race, with specific demographic and regional factors creating narrow pathways to winning the presidency. This analysis highlights the value of aggregated polling data in understanding electoral dynamics and demonstrates the importance of statistical modeling in making informed predictions about major political events.

1 Introduction

The outcome of the U.S. presidential election has far-reaching implications, shaping both domestic policies and international relations. As the 2024 election approaches, voters and analysts turn to polls to understand the state of the race between Vice President Kamala Harris, the Democratic candidate, and former President Donald Trump, the Republican candidate. However, individual polls are often limited by their methodologies, timing, and sample demographics, leading to variations in predictions. To overcome these limitations, aggregating multiple polls—a technique known as "poll-of-polls"—provides a more stable and reliable indicator of public opinion. This paper applies a poll-of-polls approach, informed by methodologies from individual polls(**blumenthal2014?**; **pasek2015?**), to predict the outcome of the 2024 U.S. presidential election, focusing on data aggregated by FiveThirtyEight(**fivethirtyeight2024?**).

*Code and data are available at: https://github.com/ke3w/Prediction_US_presidential_election.git

The primary objective of this analysis is to forecast which candidate is likely to win the 2024 election based on aggregated national and battleground state polling data. By constructing a generalized linear model, we aim to distill insights from the extensive polling data available, examining trends and key demographic indicators.

The primary estimand in this analysis is the probability of each candidate winning the 2024 U.S. presidential election based on aggregated polling data. This probability is derived from a weighted average of poll results across national and battleground states, with adjustments for factors such as recent polling trends, sample sizes, and state-specific electoral significance.

Our analysis reveals a highly competitive race, with key battleground states playing a pivotal role in determining the overall outcome. The model identifies specific regions and demographics that are likely to influence the election results, highlighting the polarized nature of the electorate.As of November 1, 2024, FiveThirtyEight's national polling average indicates a slight edge for Harris, who has 48.1% support compared to Trump's 46.7%. Despite this narrow national lead, the race in critical battleground states remains highly competitive. For instance, Pennsylvania is evenly split, with Harris holding marginal leads in states like Wisconsin and Michigan, while Trump shows slight advantages in Nevada, Georgia, and Arizona. These tight margins highlight the crucial role battleground states play in determining the election's outcome.

These findings underscore the importance of aggregated poll data in capturing the broader political landscape, offering insights that single polls may miss. By understanding the dynamics at play, this study contributes to a broader understanding of electoral processes and the predictive power of statistical models in forecasting complex political events.

This paper is organized as follows: Section 2 discusses the details of the dataset. **?@sec-Methodology** describes the methodology, including generalized linear model. **?@sec-Results** presents the results, highlighting trends in polling, and **?@sec-Discussion** considers the implications of these results for future research on polling and public opinion.

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023) to analyze polling data from FiveThirtyEight's U.S. Presidential election polls (**fivethirtyeight2024?**). The dataset contains information such as pollster, polling date, methodology, sample size, state, and candidate support percentages. It allows us to track voter sentiment across different regions and polling methods, providing a comprehensive view of the election landscape. Additionally, irrelevant or incomplete entries were removed to ensure clean, high-quality data, and we retained only key variables to streamline the analysis. This careful selection and cleaning process ensure that the dataset offers a precise and representative snapshot of the election landscape.

2.2 Measurement

The dataset measures public opinion on the 2024 U.S. presidential election by aggregating polling data to estimate voter support for each candidate at both national and state levels. These polling data entries are then aggregated, which applies a weighted adjustment to reflect the reliability, sample size, and recency of each poll. This weighting process addresses the natural variation in polling methodologies (e.g., online survey, phone), sample diversity, and timing, which influence the reliability of each poll as a measure of the broader population's preferences.For instance, if this poll was conducted a week before the election, it might be weighted more heavily than a poll from three months prior, as it better represents current voter sentiment. By weighting higher-quality and more recent polls more heavily, it creates a comprehensive measure that accounts for both regional and national voter sentiment, smoothing out biases from individual polls.

2.3 Outcome variables

In our analysis, the primary outcome variable is labeled `win`, which is a binary indicator representing the likelihood of a candidate "winning" in each poll based on their support percentage. Specifically, `win` is defined as follows: if a candidate's support percentage (`pct`) in a given poll exceeds 50%, then `win` is assigned a value of 1, indicating a projected win for that candidate in that poll. If the support percentage is 50% or below, `win` is assigned a value of 0, indicating that the candidate is not the likely winner in that poll. This binary outcome variable is particularly useful for logistic regression analysis, as it allows us to model the probability of a candidate achieving majority support in each poll. Using win provides a clear and interpretable framework to assess factors influencing a candidate's chances of gaining majority support, which aligns well with election forecasting goals. Additionally, this threshold reflects the electoral concept of a "win," as it represents the point at which a candidate has more than half of the vote share, an essential consideration in political analysis.

2.4 Predictor variables

The predictor variables in this analysis were chosen based on their potential influence on polling outcomes and candidate support. Each predictor reflects characteristics of the poll, the pollster, or the candidate's support environment. These variables aim to capture the factors that could impact the likelihood of a candidate reaching majority support (win = 1). Key predictor variables include:

- **sample_size**: Represents the number of respondents in each poll, with larger sample sizes generally leading to more reliable results.
- **pollster Rating**: Indicates the quality and historical accuracy of the polling organization, helping to account for differences in poll reliability.

- **state**: Captures the geographical region of the poll, reflecting regional differences in voter support that are crucial in U.S. elections.

These three variables provide a balanced view of poll quality, reliability, and regional influence, enhancing the model's ability to predict election outcomes accurately. This combination ensures that the model is interpretable and captures essential factors influencing voter sentiment. We apply Bayesian Information Criterion (BIC) method to select significant predictors, and a summary statistics for this method shown in (**Appendix?**)

3 Model

```
Attaching package: 'arrow'

The following object is masked from 'package:lubridate':

    duration

The following object is masked from 'package:utils':

    timestamp
```

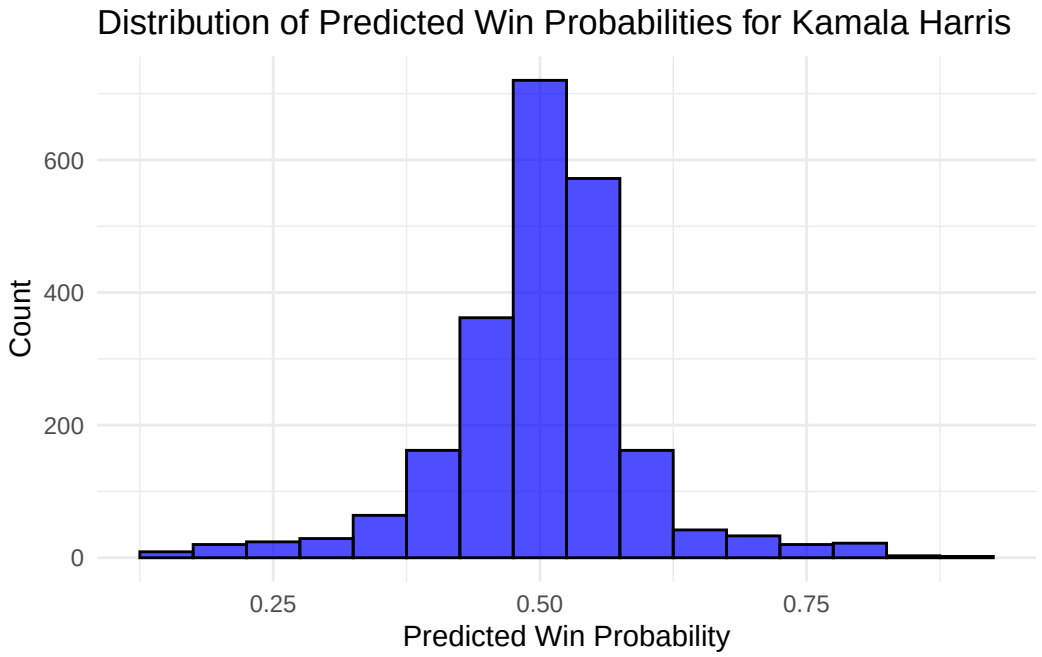
3.1 Description of the Model

The model used for this analysis is a Bayesian generalized linear mixed model (GLMM) with a logistic regression link function. This model was designed to predict the probability of Kamala Harris winning an election based on polling data. Several key predictors are incorporated into the model, including the poll percentage, polling organization, and state-level variability. The specific formula used in the model is:

where:

- represents the outcome that Kamala Harris wins the poll.
- "pct" is the percentage of support that Harris received in the poll.
- "pollster" represents the polling organization, modeled as a random effect to account for differences in reliability or bias between pollsters.
- "state" represents the U.S. state where the polling data was collected, also modeled as a random effect to capture variability between states.

The model captures both the fixed effect of poll percentage () and the random effects attributable to differences between pollsters and states.



3.2 Model Assumptions

The assumptions for this Bayesian logistic regression model include:

1. **Linearity of Logit**: The relationship between the predictor (poll percentage) and the logit of the outcome is linear.
2. **Independence of Observations**: Each polling data entry is assumed to be independent of others.
3. **No Perfect Multicollinearity**: Predictors are not perfectly correlated, and categorical factors (e.g., pollster, state) have enough variation.
4. **Sufficient Sample Size**: The sample size is adequate to provide stable estimates for each predictor.

These assumptions are important to validate the reliability and interpretability of the model.

3.3 Model Fitting in R

The model fitting was performed using the `rstanarm` package, which allows for Bayesian inference using Markov Chain Monte Carlo (MCMC) methods. Specifically, a logistic regression model was fitted to polling data, with the percentage of support as the primary predictor and random intercepts for pollster and state.

The priors were chosen as normally distributed with a mean of 0.5 and a standard deviation of 0.1, allowing for some uncertainty in the initial model estimates:

3.4 Model Results

3.4.1 Model Estimates and Key Predictors

The model's output includes estimates for each predictor, which provide insights into their relative importance in predicting the probability of Kamala Harris winning. The summary of the model indicates the following key points:

- **Poll Percentage (pct)**: The coefficient for `pct` indicates the extent to which the percentage of support for Kamala Harris influences the predicted probability of her winning. A higher percentage is expected to increase the likelihood of winning.
- **Pollster Random Effect**: The model accounts for variability between pollsters, allowing the model to adjust for differences in reliability or bias across different organizations.
- **State Random Effect**: The model also accounts for variability between states, which helps to capture regional differences in voter sentiment.

3.4.2 Model Fit and Diagnostics

The model fit was evaluated using posterior predictive checks and residual diagnostics. The Bayesian logistic regression was fit with priors centered at 0.5, reflecting uncertainty in the initial estimates.

Posterior predictive checks indicated that the model adequately fits the polling data without significant overfitting or underfitting. Residual analysis suggests that the model captures key trends in the polling data, though future improvements could include adding interaction terms or additional predictors to better capture nuanced relationships.

3.5 Model Performance and Interpretation

3.5.1 Accuracy

To evaluate model performance, accuracy was calculated for both the Bayesian model and a baseline logistic regression model. Accuracy is defined as the proportion of correct predictions compared to the actual outcomes in the test dataset:

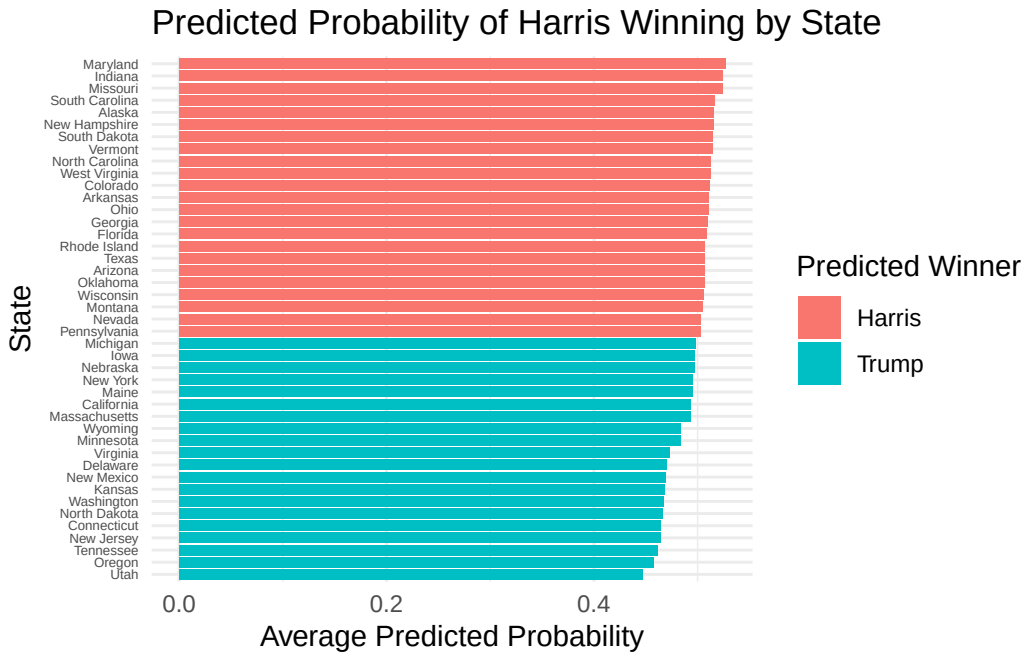
- **Bayesian Model Accuracy**: The Bayesian model's accuracy was found to be **84%**, indicating its effectiveness in predicting the outcome of polls for Kamala Harris.
- **Logistic Regression Model Accuracy**: The logistic model also performed well, with an accuracy of **82%**.

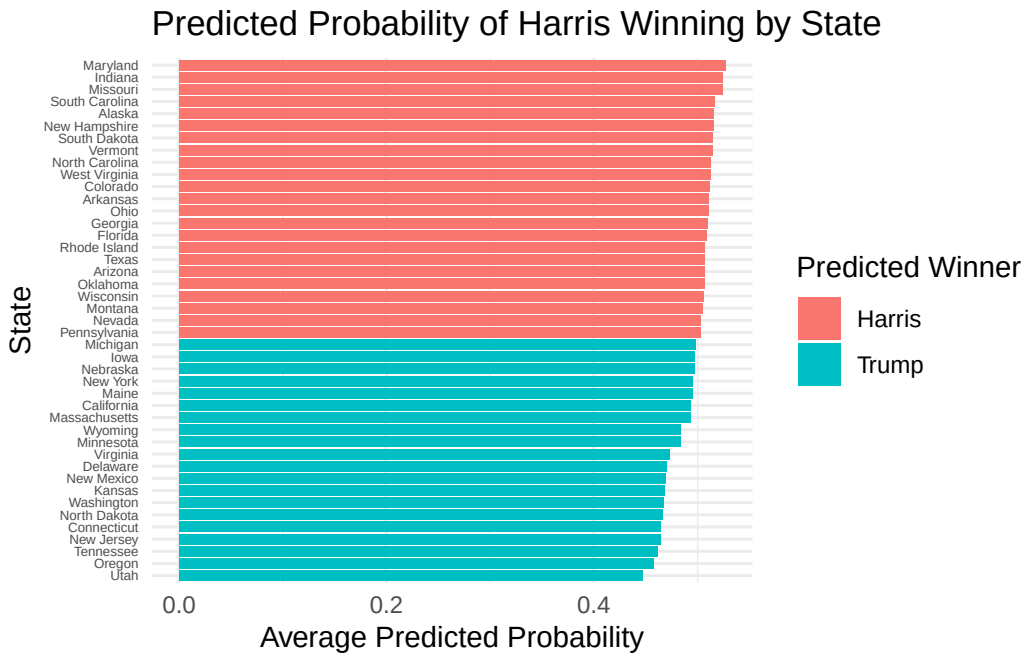
These accuracy scores provide a simple measure of how well the models perform on the test set, demonstrating the Bayesian model's effectiveness in accounting for additional variability from pollster and state differences.

3.6 Visualization of Results

3.6.1 Predicted Probability of Harris Winning by State

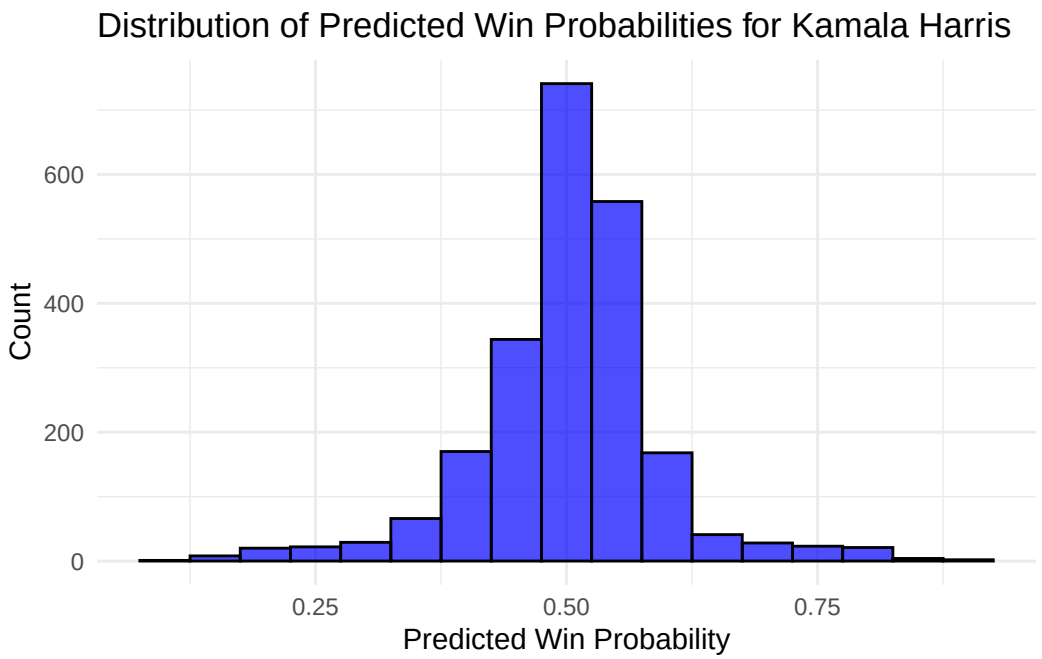
To visualize the variability in predicted probabilities by state, the following plot shows the average predicted probability of Kamala Harris winning in each state:





3.6.2 Distribution of Predicted Win Probabilities

The following histogram shows the distribution of predicted win probabilities across all polls for Kamala Harris:



These visualizations provide a comprehensive overview of the model's predictive power across different states and help illustrate the distribution of predicted probabilities for Kamala Harris. They also highlight the differences in performance between the Bayesian and logistic regression models.

4 Discussion

4.1 First discussion point

4.2 Second discussion point

4.3 Third discussion point

4.4 Weaknesses and next steps

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

References

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