Forecasting the 2024 US Presidential Election: A Poll-of-Polls Approach to Predict the Outcome*

Utilizing Poll Aggregation and Statistical Modeling to Project a Clear Path for the Presidential Race

Ruiying Li Lorina Yang

October 31, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

Table of contents

1 Introduction

The ongoing U.S. presidential election is a major focus in the media, finance, and politics, making it essential to provide reliable forecasts and understand the factors influencing varying levels of candidate support. This study utilizes Bayesian linear and logistic regression models to analyze and predict support for Kamala Harris and Donald Trump, integrating poll-level data on support percentages, poll quality, sample size, and time trends.

Aim

Our primary objectives are to examine how poll characteristics—sample size and quality—influence reported support percentages and to track shifts in support over time for Harris and Trump. The linear regression model evaluates these poll factors, assessing their impact on support stability. Additionally, time-trend analysis identifies evolving voter sentiment, capturing a recent change in support rate for candidates. And second, to forecast aggregate support rates for Harris and Trump, facilitating comparison. We use the Bayesian linear regression model to estimate overall support rate aggregate by pollesters, a logistic model

*Code and data are available at: https://github.com/Lorina-Y/US-Presidential-election.git.

complements this by estimating each candidate's probability of success directly, adding a probabilistic perspective.

Estimand
The primary estimands include the impact of poll characteristics on support percentage and each candidate's probability of winning. The linear model provides insights into how polling factors influence reported support. For the Bayesian forecasting component, The key estimand in this analysis is the probability of support for Harris versus Trump for each poll, aggregated to estimate an overall probability of each candidate's support. This probability is informed by predictors including poll-specific support percentage (pct), poll quality (pollscore) and sample size, providing a detailed estimate of each candidate's likelihood of winning.

Results
Our results show that sample size and poll quality have minimal impact on support rates, suggesting robustness in reported voter sentiment. Time-trend analysis reveals that Trump's support was initially higher; however, recent data indicates a surge in support for Harris, overtaking Trump. The linear model's aggregate forecast shows nearly equal support rates (44.76% for Harris vs. 44.73% for Trump), while the logistic model offers a probabilistic view, indicating a slight edge for Harris. Together, these models provide a comprehensive understanding of the election dynamics.

Results To assess how specific poll characteristics influence reported support percentages (`pct`), we applied a Bayesian linear regression model focused on sample size and poll quality (`pollscore`). Results indicated a slight negative association between poll quality and support percentage, with support decreasing from around 46 to 44 as quality increased, suggesting limited impact. Additionally, a nearly flat trend line for sample size revealed no meaningful relationship with support levels. These findings suggest that poll characteristics such as quality and size have minimal effects on support, reinforcing the reliability of public sentiment assessments.

The results show that while the linear regression model predicts similar overall support rates for Harris and Trump (44.76% and 44.73%, respectively), the logistic model indicates a slight advantage for Harris, with a probability of support of 0.303, compared to Trump's 0.251. These results underscore the linear model's strength in examining predictor effects and aggregate trends, while the logistic model's probability outcomes provide additional insight into the competitive landscape, enriching the forecast with a candidate-to-candidate comparison. Together, these models offer a robust view of election dynamics, grounded in poll data characteristics and forecasted support probabilities.

1.1 Paper Structure

The remainder of this paper follows a structure Section **??**,Section **??** about MLR model,Section **??**,Section **??**,Section **??** about pollster methodology,Section **??** about our methodology on survey.

1.2 Software and packages used

We use the statistical programming language R (R Core Team 2023) and the following packages: tidyverse (Wickham et al. 2016), dplyr (Wickham 2023a), readr (Wickham 2023c), ggplot2 (Wickham 2023b), janitor (Baumer et al. 2023), lubridate (Wickham et al. 2023), broom (Robinson et al. 2023), modelsummary (P. 2023), rstanarm (Goodrich et al. 2022), here (**here?**).

2 Data

2.1 Overview

Our analysis, conducted in statistical programming language R (R Core Team 2023), utilizes polling data from FiveThirtyEight's repository on the 2024 U.S. Presidential election polls (FiveThirtyEight 2024). This dataset includes variables such as pollster, methodology, sample size, poll end_dates, and support percentages 'pct' for candidates, allowing us to examine how support fluctuates over time and in response to poll characteristics for each candidate. To align with our study's focus, we cleaned the dataset to remove inconsistencies and retained only relevant variables, with only focus on data of 2 Candidate "Trump" and "Harris", preparing the support rate for forecasting the possible winner and examine the trend.

2.2 Measurement appendix

This dataset reflects public sentiment on the 2024 Presidential race, gathered through various polling methods,which involves translating public opinion polling into measurable data for analysis.The polling methods including online surveys and phone interviews by organizations like Morning Consult and TIPP. Each poll captures respondent preferences for candidates, sampling strategies, and timing, along with poll related factors, such as sample_size and pollscore, forming structured data points. These variables represent the underlying polling practices and voter sentiment, after cleaning, the data enables us to analyze factors influencing support rates and model time trends in voter preferences, crucial for accurate election forecasting.

2.3 Response variables

In our analysis, we use two response variables to capture different aspects of candidate support. For the linear model, the response variable is `pct`, representing the percentage of respondents indicating support for each candidate in individual polls. This continuous variable allows us to explore how factors like poll quality (`pollscore`) and sample size influence reported support percentages. For the logistic model, we use `candidate_dummy`, a binary variable coded as `1` for Harris and `0` for Trump. This binary outcome enables us to estimate the probability of

support for each candidate, providing a probabilistic forecast of each candidate's chance of winning based on aggregated poll characteristics.

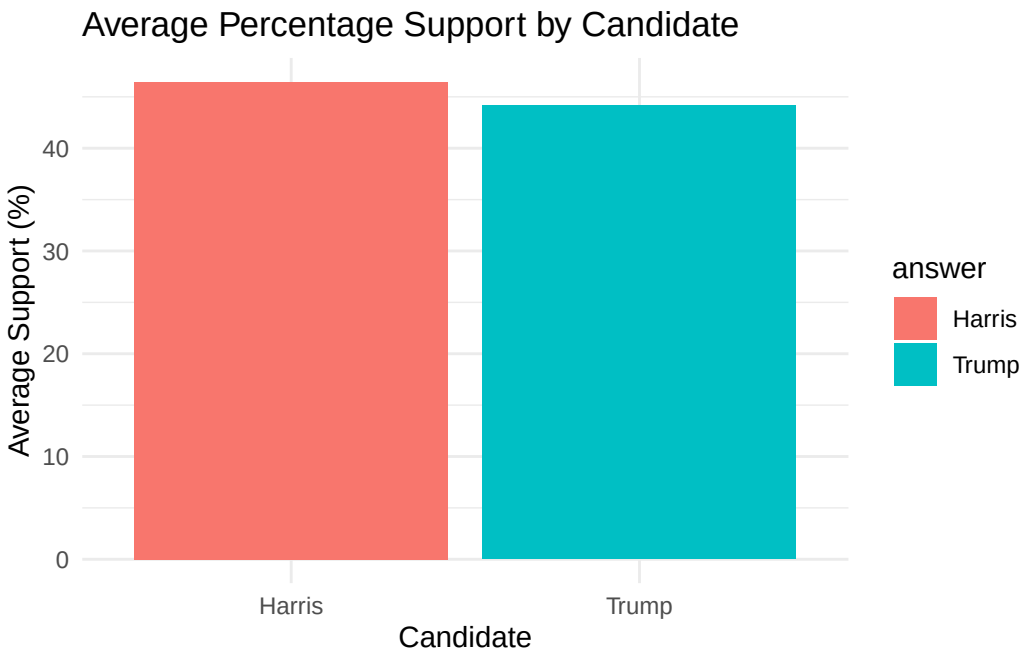

Figure 1: Average Percentage Support(pct) for Harris and Trump

2.4 Predictor variables

The model incorporates a set of key predictor variables aimed at refining forecast accuracy by capturing unique attributes of each poll. **Average Support (`pct`)** is a predictor in the second Bayesian logistic model, it provides the support percentage reported by each poll for the candidate in question, serving as a direct indicator of voter preference. **Poll Quality (`pollscore`)** reflects the reliability of each poll based on historical accuracy and transparency, with higher scores indicating greater reliability. **Sample Size (`sample_size`)** accounts for the number of respondents per poll, controlling for the variability that smaller samples may introduce. Together, these predictors enable a nuanced analysis of support that accounts for each poll's quality, methodology, and respondent base, leading to a robust and interpretable model of voter preference.

3 Model

3.1 Model Overview

This analysis employs a Bayesian linear regression model and a Bayesian logistic regression model to predict support for each candidate, Harris and Trump, in the 2024 U.S. presidential election. By using a "poll-of-polls" approach, we aggregate data from multiple polls, creating a comprehensive view of voter support. The linear model estimates how poll characteristics like sample size and poll quality influence reported support percentages (pct), capturing trends over time. Meanwhile, the logistic model provides a probabilistic forecast, estimating each candidate's likelihood of winning based on aggregated support metrics. Both models incorporate Bayesian inference, allowing us to account for variability and uncertainty in polling data and offering a robust election forecast.

3.2 Model selection

We selected the Bayesian linear and logistic models based on their capacity to integrate prior information and address different aspects of our research objectives. The linear model was chosen for its ability to assess predictor effects on support percentages (pct), making it ideal for understanding poll characteristics' impact on support rates. The Bayesian logistic model was selected for its suitability in estimating support probabilities, capturing the competition between candidates. Model fit and Bayesian methods was evaluated use posterior predictive checks, also allow for rigorous model diagnostics, validating the models against observed polling data and enhancing forecast reliability.

3.3 Model set-up

3.3.1 Bayesian linear model

In the Bayesian linear regression model, the response variable 'pct' represents the percentage of respondents supporting each candidate in individual polls. This model can be express as below, In the Bayesian linear regression model, the outcome variable, `pct`, represents the percentage of respondents supporting each candidate in individual polls. This model is expressed as below, where (_0) is the intercept, (_1) and (_2) are coefficients for sample size and poll quality (pollscore), respectively, and (_i) represents the error term, capturing random variability in support percentages.

$$\text{pct}_i = \beta_0 + \beta_1 \cdot \text{sample\_size}_i + \beta_2 \cdot \text{pollscore}_i + \epsilon_i$$