Forecasting the 2024 Election: Tight Margins and Swing State Uncertainty*

Harris Leads Trump by 0.8% as Voters Remain Divided%

Mariko Lee Karen Riani Cristina Su Lam

November 4, 2024

We forecast the 2024 U.S. presidential winner using state-level poll averages for Kamala Harris and Donald Trump, weighted by poll quality and sample size. Applying a logistic regression model, we find a slight 0.8% lead for Harris over Trump, using polls taken since the reelection bid on July 21, 2024. We weigh the polls with aims to increase the influence of the most reliable data in shaping the forecast—limiting biased results and showing a clearer voter sentiment amidst a particularly competitive election. These results show who may be leading the elections via polls, giving campaigns the insights they need and reinforcing public trust in the election process, especially as the outcome likely hinges on key swing states.

1 Introduction

Amid the chaotic, high-stakes 2024 U.S. presidential election, forecasting models offer crucial insights into voter sentiment as Americans face a tight race between top candidates Donald Trump and Kamala Harris. With the attempted assassination of Trump, Joe Biden's endorsement of Harris, and recent crises such as hurricanes disrupting voter access in swing states, public opinion remains volatile (ABC News 2024b). Key events like Trump's policy proposals on social security and public safety, Harris's advocacy for reproductive rights, and other rising debates on gun control have further polarized the electorate, making forecasting models essential in offering clarity amid this competitive landscape (CBS News 2024; Human Rights Watch 2022).

Using presidential polling data from FiveThirtyEight, our Bayesian model's estimand is the probability of support for Kamala Harris, weighted by pollster rating and sample size to improve the reliability of aggregated polling data. By refining the aggregation of polls through

[*] Code and data are available at: https://github.com/karenrni/Forecasting-the-2024-US-Presidential-Elections

logistic regression, weighted by sample size and pollster rating, this study contributes to the growing body of forecast models—including media outlets' own polls-of-polls and various Bayesian methods. Our model estimates only a slim 0.8% lead for Harris over Trump in the overall popular vote—a close margin that highlights the high uncertainty in public sentiment. In line with other forecasts that struggle to declare a clear frontrunner, this narrow lead reflects the volatility of the race and the importance of updated, representative polling data (The New York Times 2024).

This uncertainty emphasizes the significance of swing states in election forecasts—particularly states with nearly equal support for each party and a history of alternating between Democratic and Republican candidates in recent elections (Bloomberg 2024). Existing research identifies seven critical swing states, with Pennsylvania standing out as a tipping point in more than 1 in 5 simulations (Associated Press 2024). With 19 electoral votes, Pennsylvania could prove decisive in the event of a close race. The competitive nature of these swing states shows the need for continued data updates to maintain forecast accuracy in a closely contested race.

By prioritizing higher-quality polls and focusing on data collected after Harris's campaign announcement, the model seeks to mitigate biases and improve forecast reliability by using a polls-of-polls method over individual pollster data. This approach not only clarifies voter sentiment amid competing forecasts but also demonstrates the value of weighting variables such as pollster rating and sample size in enhancing prediction accuracy. While some states lack sufficient polling data—introducing potential limitations that could be improved with more data and model complexity—the overall findings provide strategic insights for decision-makers and reinforce public confidence in an election where no candidate holds a decisive lead. Appendix A provides a detailed assessment on The Washington Post, one of the polling methodologies used by prominent sources, including sampling approaches, weighting, and non-response handling strategies to ensure robust data quality.

The rest of this paper is structured as follows: Section 2 presents the data sources and methodology, Section 3 presents the forecasting model, then Section 4 followed by Section 5, which discusses the results and their implications. Section 6 concludes with a discussion of the limitations of the study and suggestions for future research, followed by **?@sec-appenx** with The Washington Post's pollster methodology and a budgeted idealized methodology.

2 Data

TODO: PROPER FORMATTING FOR DATA SEC, ADD FINAL DETAILS AND FIGURES

2.1 Overview

Our study uses state-level polling data from Donald Trump and Kamala Harris to predict the 2024 US presidential election. The polls used were sourced from FiveThirtyEight's 2024 Presidential Election Forecast Database (FiveThirtyEight 2024), which compiles polling data from multiple organizations and rates each poll's quality. These ratings assist in identifying and weighing the most reliable polls. The dataset used for this analysis spans polls starting from July 21, 2024, when Kamala Harris announced her reelection bid.

Our analysis focuses on the following variables: Polling rating (Pollscore): A numerical score reflecting the historical reliability of polling organizations used to weight polls by accuracy. Sample size: The number of respondents per poll, influencing the poll's margin of error and its weight in the forecast. Percentage support (pct): The proportion of respondents who express support for each candidate. State: The US state where the poll was conducted OR labeled "National" for nationwide polls. Pollster: The organization that conducted the poll, influencing the credibility and weight of our analysis.

Table X - Sample of the cleaned dataset, highlighting key variables crucial in the analysis. [Table]

2.1.1 Measurement and Limitations

We applied several measurement and limitations considerations to ensure the validity of our dataset: Poll Quality: Temporal Dynamics: Geographic Coverage: Response Bias:

2.1.2 Outcome Variables

The Percentage of support (pct) for Donald Trump and Kamala Harris is the main outcome variable. We forecast which candidate is leading by aggregating these results across states. Figure X represents the distribution of support percentages for the two candidates, showing clusters ranging from 40% to 60%, indicating the election's competitive nature.

[Figure X: Distribution of Candidate Support]

2.1.3 Predictor Variables

The following predictor variables were used to build the logistic regression model and refine our prediction: Pollscore: Historical reliability score for pollster (range:) Numeric Grade: (scale: 0-4) Transparency Score: (scale: 0-10) Sample Size: Number of respondents in each poll (typically between 500-3000) Methodology: Survey method (eg. phone interviews, online panels) State: State-level indicators that capture regional variations in support.

These selected variables are based on their relevance to polling accuracy and availability across the dataset. The analysis regarding their relationships with the outcome variables (percentage support) is further explored.

2.1.4 Cleaning Process and Analysis

The data cleaning process involved the following steps: Filtering for High-Quality Polls: We included only polls with a numeric grade of X or higher to reduce the impact of unreliable data. Standardizing Dates and Locations: Variable Transformation: Converted percentages into absolute numbers using sample size Created binary indicators for candidates (Trump = 0, Harris = 1)

Some of our data is of marriage license (**?@fig-marriage**), from (**TorontoOpenData?**)

3 Model

TODO: FINALIZE MODEL SECTION - VALIDATION - ALTERNATIVES - JUSTIFICATION - UPDATE MODEL BELOW

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 \times \text{Pollster}_i + \beta_2 \times \text{State}_i + \beta_3 \times \text{Sample Size}_i + \beta_4 \times \text{Pct}_i$$

- y_i is the dependent variable, representing the count of respondents who support Harris in a given poll, modeled as a binomial outcome. - β_0 is the intercept term, indicating the baseline log-odds of Harris support when all predictors are zero. - β_1, β_2, β_3, and β_4 are the coefficients for the predictor variables **Pollster**, **State**, **Sample Size**, and **Pct** (the percentage of support for Harris in a poll), respectively: - β_1 represents the adjustment in log-odds based on the specific pollster conducting the survey. - β_2 accounts for the impact of the state in which the poll is conducted. - β_3 adjusts for the influence of the poll's sample size. - β_4 represents the effect of the poll's percentage support for Harris on the log-odds. - The function $\text{logit}(\mu_i)$ transforms the linear combination of predictors into probabilities, providing the estimated probability of support for Harris.

4 Results

TODO: FINISH RESULTS INCLUDING GRAPHS, TABLES, SUMMARY STATS

Overall Percentage for Kamala Harris: 49.72895 % Overall Percentage for Donald Trump: 50.27105 % … state_winner n 1 Harris 19 2 Trump 23

5 Discussion

TO DO:

FINISH DISCUSSION, EDIT CURRENT BITS

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

What are some weaknesses of what was done? + implications

Environmental and Societal Factors in Surveys, contributing to the – bias

Recent envoronmental disasters have disruoted life in major swing states

Didnt account for partisian pollsters and the potential bias ,

Weaknesses in forecasting itself, How trump could lose despite pollster estimates

Lack of data, could introduce bias, not entirely representative, uncertainty due to declaration for Kamala, This can cause

6 Limitations

What is left to learn or how should we proceed in the future?

Future models could benefit from more complex models which account for other potentially influential factors such as poll methodology (i.e. online vs in-person surveying) and partisanship or respondents' political affiliations (Cambridge University Press 2024). Models using polls-of-polls approaches which use multi-level regression and post stratification tend to do particularly well in forecasting elections (REF tab 2), and increasing complexity to account for more influential factors like those mentioned above. They historically perform /…. Better estimation . and accounts for uncertainty propagation, which is particularly relevant in this quarter's election. (Reference in tab) Furthermore, using training data based on historical polls regarding democratic and republican election results and polling data would allow for a more robust model. This is especially because of Biden's late declaration for Harris, thus substituting fot prior state poll behaviour will help account for a lack of polling data. Although many forecasting models aim to estimate the national popular vote, Our model similarly could benefit from estimations for electoral vote

Appendix {#sec-appenx}

A Methodology Analysis of The Washington Post Polling

With an evaluation of sampling methodology, recruitment, handling non-response, and questionnaire design, this appendix analyzes the polling methodology used by The Washington Post in collaboration with ABC News. The objective is to assess the strengths and weaknesses of these approaches and understand their impact on polling accuracy.

A.1 Population, Frame, and Sample

The target population for The Washington Post and ABC News polls includes U.S. adults and registered voters. The sampling frame, or list from which the sample is drawn, includes landlines, mobile phones, and internet users, covering a broad demographic range. Using a combination of text-to-web polls and random digit dialing (RDD) for landlines and mobile phones, they aim to reach a probability sample where each individual in the population has a known chance of selection. This minimizes sampling bias and improves the representativeness of the sample (The Washington Post 2024a).

The Washington Post's polling averages use only national and state-level polls that comply with strict quality and transparency criteria. These surveys were chosen because they employ suitable stratification and weighting strategies in addition to random sample approaches (The Washington Post 2024b). To represent key demographics such as age, race, gender, and education, the samples are weighted (The Washington Post 2024a) (The Washington Post 2024b).

A.2 Sample Recruitment

Data collection includes live phone interviews and text-to-web surveys, ensuring diverse demographic representation. In a typical 2024 poll, sample collection consisted of 21% text-to-web invites, 15% landlines, and 64% mobile phones (The Washington Post 2024a). This technique facilitates access to younger and minority voters, who may not be well represented in conventional landline-based surveys.

ABC News also uses probability-based recruiting through the Ipos KnowledgePanel by using address-based sampling from the Delivery Sequence File of the US Postal Service. Since internet connections and equipment are offered at no cost, this guarantees that even households without internet connections or digital devices are involved (ABC News 2024a).

A.3 Sampling Approaches and Trade-offs

The Washington Post uses stratified random sampling to ensure that important demographics are represented proportionately to their voter base. Stratified sampling increases the likelihood of a balanced representation, while weighting addresses over- or under-representation of particular groups (The Washington Post 2024a) (The Washington Post 2024b).

When state-level polling data is scarce, The Washington Post incorporates historical voting records from the last two presidential elections (The Washington Post 2024b). This adjustment helps estimate current preferences but depends on past trends, which may miss recent shifts in voter sentiment (The Washington Post 2024a).

A.4 Non-response Handling

Non-response can introduce non-response bias, where the views of non-respondents may differ from respondents, potentially skewing results. The Washington Post addresses this issue using response weighting based on age, race, and education to align the final sample with the overall population distribution (**WashPost2023_Standard?**).

ABC News also addressed non-response bias by applying post-stratification adjustments and sending email reminders to non-respondents. In addition, The Washington Post and ABC News ensure that their samples are weighted to account for any anomalies in non-response (ABC News 2024a) (The Washington Post 2023).

Despite these initiatives, non-response bias is still a concern, especially for populations that are less inclined to take part in surveys, including younger or less politically active people (The Washington Post 2023).

A.5 Questionnaire Design

The Washington Post creates its surveys with neutrality and clarity to avoid influencing responses. Questions are rotated to minimize question order bias, and multiple-choice options, including “No opinion,” are provided (The Washington Post 2024a). Randomizing questions prevents any unintentional influence on how respondents interpret and answer follow-up questions (The Washington Post 2023).

ABC News follows similar principles, offering surveys in both Spanish and English to ensure inclusivity. Leading questions are purposefully omitted from the questionnaires to better capture genuine public opinion across language barriers (ABC News 2024a).

A.6 Strengths and Weaknesses of the Methodology

Strength:
Sampling Method: The combination of RDD, text-to-web polls, and live phone interviews enable access to a wide demographic, including younger and harder-to-reach voters (The Washington Post 2024a) (The Washington Post 2023).
Post-stratification Weighting: Effective weighting adjusts for demographic imbalances, enhancing poll accuracy and representativeness (The Washington Post 2024a) (ABC News 2024a).
Transparent Approach: Using only high-quality polls and transparency in methodology increases the credibility of The Washington Post's polling data (The Washington Post 2023) (The Washington Post 2024b).

Weaknesses:
Non-response Bias: Despite response weighting, non-response remains an issue, particularly with groups less inclined to participate, such as younger individuals (The Washington Post 2023) (The Washington Post 2024a).
Dependency on Historical Data: In states with limited polling, depending on past election data may not accurately reflect recent shifts in voter preferences (The Washington Post 2024b).

A.7 Conclusion

The Washington Post and ABC News polling methodologies offer a framework for measuring public opinion in the 2024 U.S. presidential election. Employing various sampling strategies, stratification, and weighting improves sample representation. However, challenges such as non-response bias and reliance on historical data in under-polled states need continued adjustments to maintain polling reliability.

B Idealized Survey & Methodology - $100K Budget

B.1 Overview

Using a $100K budget, this appendix outlines a carefully designed survey methodology for predicting the 2024 US Presidential Election. The objective is to collect representative, high-quality data using recruiting, poll aggregation, and selective sample methods. Through rigorous validation, this approach ensures data accuracy and reduces common survey research errors.

B.2 Sampling Approach

We will implement stratified random sampling to ensure that key demographic and geographic subgroups are fairly represented. This approach reduces bias and offers more reliable insights into voter preference.

Stratification Criteria:
- Age Groups
- Gender
- Education Levels
- Geographic Representation
- Political Affiliation

Sample Size Goal: 10,000 respondents across states and demographics to achieve **high statistical power** with a margin of error below $\pm 1\%$

Trade-offs:
- Although stratified sampling increases representativeness, it necessitates accurate demographic information and may raise operating expenses.
- **Missing Data:** It's possible that some demographics (e.g. men) may be less likely to respond. Post-stratification weighting and data imputation will be used to address this issue.

B.3 Recruitment Strategy

Outline outreach and telephone surveys will be combined in our recruitment strategy to ensure widespread participation from various demographic groups.

Online Recruitment:
- Target ads on Google, Facebook, and Twitter to engage younger voters and urban populations.
- Budget Allocation: $25,000

Random-Digit Dialing (RDD):
- Phone outreach to reach older, rural voters with poor internet connection. - Budget Allocation: $30,000

Incentives:
- Participants are offered $5 gift cards to increase response rates. - Budget Allocation: $20,000

Non-Response Handling:
- Increase recruitment incentives for underrepresented groups and use numerous follow-up reminders.

B.4 Data Validation

To ensure the accuracy and reliability of responses, we will implement several data validation techniques:

Survey Logic Check:
- Recognize and flag responses contradicting one another (e.g., reporting under 18 but registered to vote).

Attention Check:
- Utilize questions to confirm respondents are actively engaged (e.g., “Select ‘Confirm’ to start questionnaire”)

Post-Stratification Weighting:
- Adjusting for over- and under-enumeration and weighting the sample to reflect the demographic of the US population.

Mode and Measurement Errors:
- We mitigate the impact of using mixed modes (online and telephones) by training enumerators and reducing enumerator bias. Misreporting will be reduced through straightforward questions.

B.5 Poll Aggregation Methodology

We will employ a poll-of-polls aggregation method to reduce bias and smooth fluctuation across individual polls.

Weighting Criteria:
Sample Size: larger samples receive more weight to mirror greater reliability.

Recency: More recent polls are given higher weight to capture modern voter sentiment.

Pollster Rating: Polls from highly rated pollsters receive higher weights to reduce the impact of bias.

B.6 Survey Implementation

Google Forms was used to create and implement the survey, allowing for efficient data collection and safe storage. The main section and sample questions are listed below.

Access the survey: Google Form

Survey Overview
Title: 2024 US Presidential Election Poll
Purpose: To gather public sentiment and predict election outcomes.
Estimated Time: Less than 5 minutes
Confidentiality: All responses are anonymous and used only for research purposes.

1. What is your age?

- Under 18
- 18-24
- 25-34
- 35-44
- 45-54
- 55-64
- 65+

2. What is your gender?

- Male
- Female
- Non-binary / Prefer not to say

3. What is the highest level of education you have completed?

- Less than high school
- High school diploma or GED
- Some college
- Bachelor's degree
- Master's degree or higher

4. Are you registered to vote?

- Yes
- No

5. Who do you intend to vote for in the upcoming presidential election?

- Donald Trump (Republican)
- Kamala Harris (Democrat)
- Other
- Undecided

6. How likely are you to vote in the upcoming election?

- Very Likely
- Somewhat Likely
- Not Likely

B.7 Budget Breakdown

- **Online Recruitment** : $25,000
- **RDD Recruitment**: $30,000
- **Incentives for Participants**: $20,000
- **Data Processing & Validation**: $15,000
- **Miscellaneous Expenses**: $10,000

Total: $100,000

B.8 Conclusion

This survey methodology uses stratified sampling, multi-channel recruitment, and rigorous data validation procedures to ensure accurate forecasting of the 2024 US Presidential Election. We provide a more stable and reliable prediction through poll-polls aggregation, smoothing out fluctuations across polls. This design balances accuracy, inclusivity, and efficiency with a carefully considered $100K budget, ensuring the poll gathers meaningful insights into voter sentiment and behavior.

References

ABC News. 2024a. "ABC News Polling Methodology Standards." https://abcnews.go.com/US/PollVault/abc-news-polling-methodology-standards/story?id=145373.

———. 2024b. "Hurricanes Helene and Milton Impact on 2024 Election." https://abcnews.go.com/538/hurricanes-helene-milton-affect-2024-election/story?id=114783517.

Associated Press. 2024. "Election Polls and Forecasts." https://apnews.com/article/election-polls-kamala-harris-donald-trump-b497bd1015d35ed563e085c8db075802.

Bloomberg. 2024. "Trump and Harris Tied in Key States." https://www.bloomberg.com/news/features/2024-10-23/new-poll-has-trump-harris-tied-in-key-states-just-12-days-to-election.

Cambridge University Press. 2024. "Information, Incentives, and Goals in Election Forecasts." *Judgment and Decision Making*. https://www.cambridge.org/core/journals/judgment-and-decision-making/article/information-incentives-and-goals-in-election-forecasts/CDED59D35F6599BF9D31348A79B196F9.

CBS News. 2024. "Trump-Harris Campaign Promises in the 2024 Election." https://www.cbsnews.com/news/trump-harris-campaign-promises-2024-election/.

Human Rights Watch. 2022. "U.s. Supreme Court Overturns Roe v. Wade." https://www.hrw.org/news/2022/06/24/us-supreme-court-topples-roe-v-wade-blow-rights.

The New York Times. 2024. "Trump-Harris Polls and the Tight 2024 Election." https://www.nytimes.com/2024/10/21/upshot/trump-harris-polls-election.html.

The Washington Post. 2023. "Polling Methods and Standards (2023)." https://www.washingtonpost.com/documents/f409a864-d58c-430a-8adc-ee6a8a344854.pdf?itid=lk_inline_manual_6.

———. 2024a. "ABC News/the Washington Post Polling Methodology (2024)." https://www.washingtonpost.com/documents/b4c28200-18a7-412e-a4da-4e8cf9b91fd5.pdf.

———. 2024b. "Presidential Polling Averages Methodology." https://www.washingtonpost.com/elections/2024/06/26/presidential-polling-averages-methodology/.