US President Prediction[*]

Predict the results of 2024 US President Election

Yun Chu Felix Li Wen Han Zhao

October 22, 2024

In this study, we aim to study the 2024 U.S. Presidential Election with the methodology of polls of polls across various states. With the insightful analyses on polls provided by Redfiled & Wilton Strategies, the conducted survey on the population highlights various key point for the prediction of US president. Using Generalized Linear Models (GLMs), this report analyzes the accuracy of prediction in president votes through several key variables from pollster. The model incorporates variables such as pollscore, methodology, transparency_score and hypothetical. The prediction finding highlights the significance of candicate trustworthiness among the voter decision, providing insights for swing voters. This analysis offers valuable information into the potential electoral outcomes driving the 2024 US president election.

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2....

[*] Code and data are available at: https://github.com/younazhao/US-President-Prediction/tree/main.

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023)…. Our data…. Following Alexander (2023), we consider…

Overview text

2.2 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

2.3 Outcome variables

Talk way more about it.

2.4 Predictor variables

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

3 Model

In this analysis, we aim to model the popularity trends for Kamala Harris and Donald Trump, based on high-quality polling data collected at the national level after Harris's declaration on July 21, 2024. We used Bayesian linear regression models with Gaussian error structures to estimate changes in polling percentages over time for each candidate.

3.1 Data Filtering and Preparation

The data were first filtered to retain only high-quality polls (with a numeric grade of 2.0 or above and a transparency score of 4 or above). We focused on polls where the `state` was either explicitly labeled as “National” or where the state information was missing (which we imputed as “National”). Additionally, we filtered polls collected after July 21, 2024, when Kamala Harris announced her candidacy.

- **Kamala Harris Data**: After filtering for high-quality polls and setting the timeframe, we calculated the number of Harris supporters per poll by taking the product of her polling percentage and the poll’s sample size.
- **Donald Trump Data**: We applied the same filtering and calculation methods to Trump’s data, ensuring comparability between both candidates.

To avoid issues with missing values in our models, we excluded any rows where key variables (e.g., polling percentage or end date) were missing.

3.2 Model Specification

Two separate Bayesian linear regression models were fitted using the `stan_glm` function. Both models specified the formula:

[pct end_date]

indicating that we are modeling the polling percentage as a function of time (end date of the polls).

3.2.1 Model for Kamala Harris

The model for Kamala Harris was fitted as follows:

- **Response Variable**: `pct` (polling percentage)
- **Predictor**: `end_date` (date of the poll)
- **Family**: Gaussian (normal distribution)
- **Priors**:
  - Normal prior for the slope: (Normal(0, 0.1))
  - Normal prior for the intercept: (Normal(50, 5))
  - Exponential prior for auxiliary parameters: (Exponential(1))

3.2.2 Model for Donald Trump

The model for Donald Trump was fitted using the same specifications as for Kamala Harris

4 Results

The results of the Bayesian linear regression models for Kamala Harris and Donald Trump are summarized in Table 1 and Table 2.

```
`modelsummary` 2.0.0 now uses `tinytable` as its default table-drawing
  backend. Learn more at: https://vincentarelbundock.github.io/tinytable/

Revert to `kableExtra` for one session:

  options(modelsummary_factory_default = 'kableExtra')
  options(modelsummary_factory_latex = 'kableExtra')
  options(modelsummary_factory_html = 'kableExtra')

Silence this message forever:

  config_modelsummary(startup_message = FALSE)
```

Table 1: Summary statistics of the Bayesian model for Harris

Table 1: Summary of Bayesian Model for Harris

Term	Estimate	Std. Error	2.5% CI	97.5% CI
(Intercept)	40.47	1.99	37.29	43.46
days_after_earliest	0.01	0.00	0.01	0.01

Table 2: Summary statistics of the Bayesian model for Harris

Table 2: Summary of Bayesian Model for Harris

Term	Estimate	Std. Error	2.5% CI	97.5% CI
(Intercept)	40.36	0.6	39.41	41.37
days_after_earliest	0.01	0.0	0.00	0.01

4.1 Overall trend

The intercepts for both models are significantly negative, indicating a downward trend in polling percentages over time for both candidates. The positive coefficients for end_date suggest a slight increase in polling percentages as the election date approaches. However, the magnitude of these increases is relatively small.

4.2 Model Diagnostics

The posterior predictive checks (Figures 1 and 2) and the credible intervals for predictors (Figures 3 and 4) indicate that the models fit the data well. The diagnostics for both Harris and Trump models show that the models are stable and the predictions are reliable.Further diagnostics (Figures 5 to 8) confirm the robustness of the models. The R-hat values are all below 1.1, indicating good convergence of the MCMC chains.

Appendix

A Additional data details

A.1 Model check

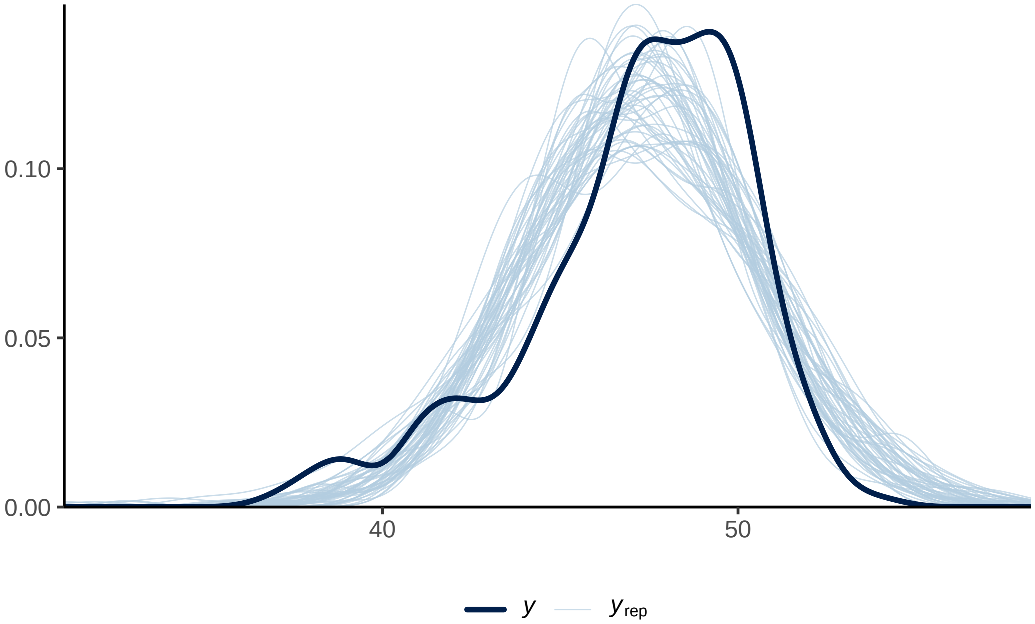

Figure 1: Posterior Predictive Check for All Models

A.2 pollster methodology analysis

This survey was conducted by Redfield & Wilton Strategies to assess the voting intentions of eligible voters in key U.S. swing states ahead of the 2024 Presidential Election. The primary goal of this poll is to provide an accurate and timely snapshot of public opinion in states where electoral outcomes are uncertain and could have a decisive impact on the overall result of the election. Swing states, due to their political volatility and diverse voter bases, are critical in determining the balance of power in the U.S. electoral system. Understanding voter preferences in these states is essential for political analysts, campaigns, and the general public.

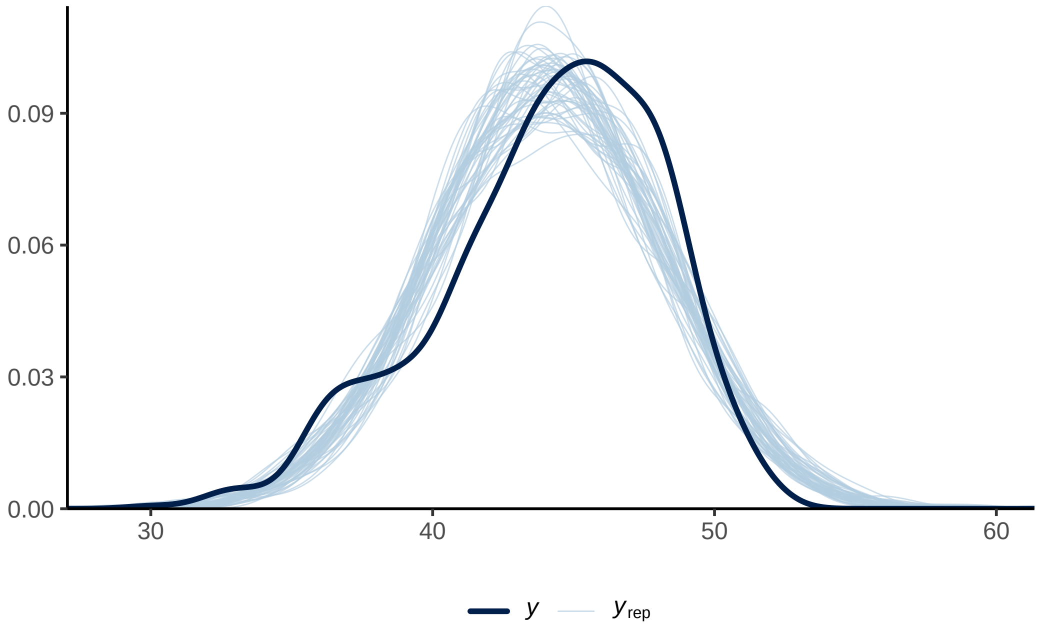

Figure 2: Posterior Predictive Check for All Models

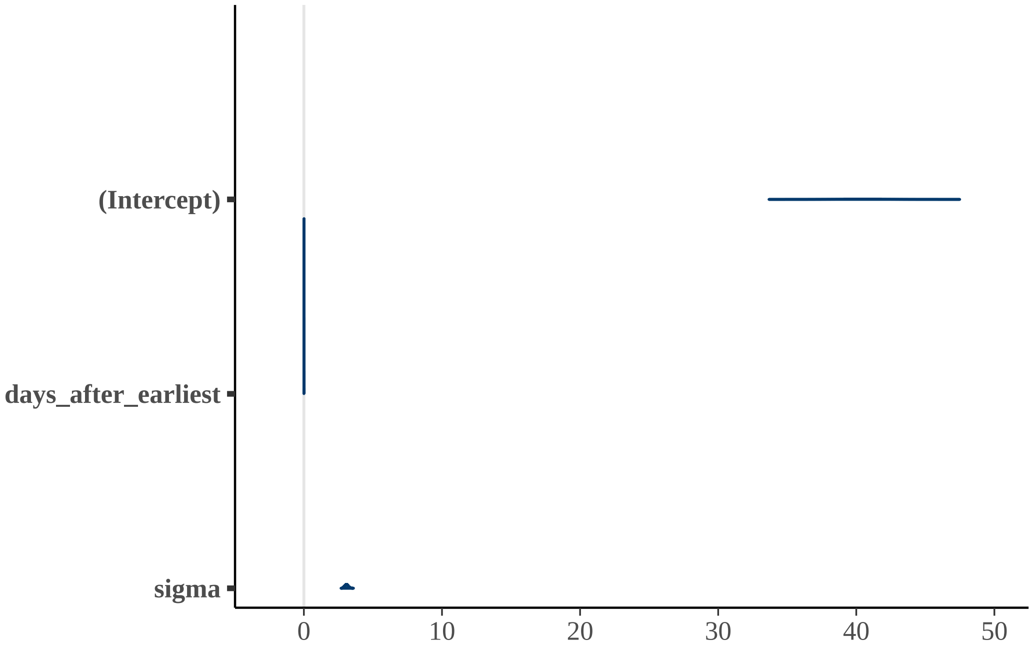

Figure 3: CI for predictors for Harris

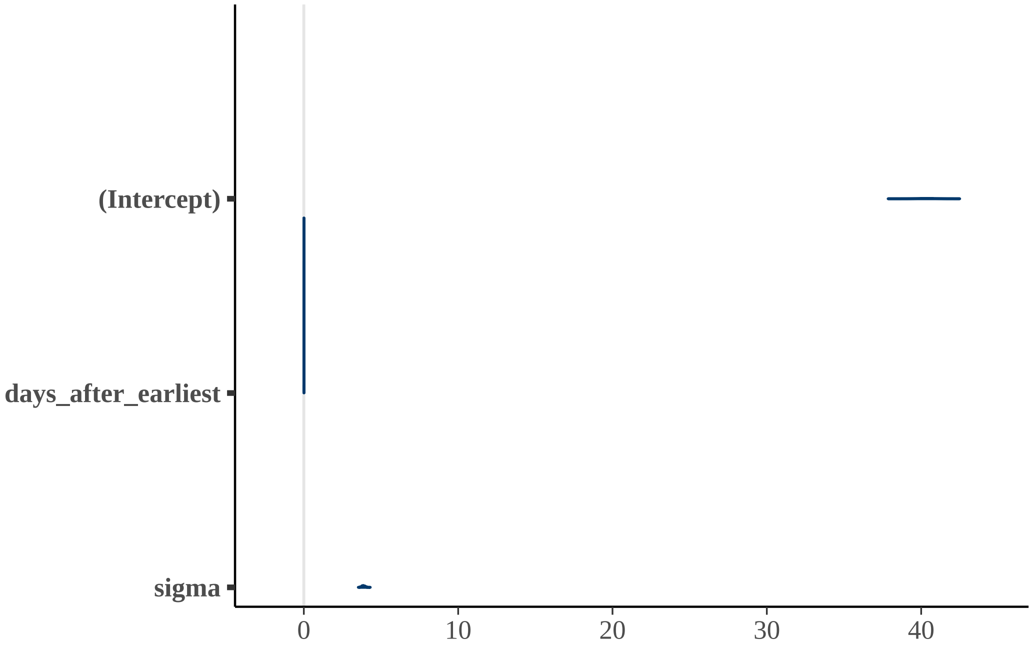

Figure 4: CI for predictors for Trump

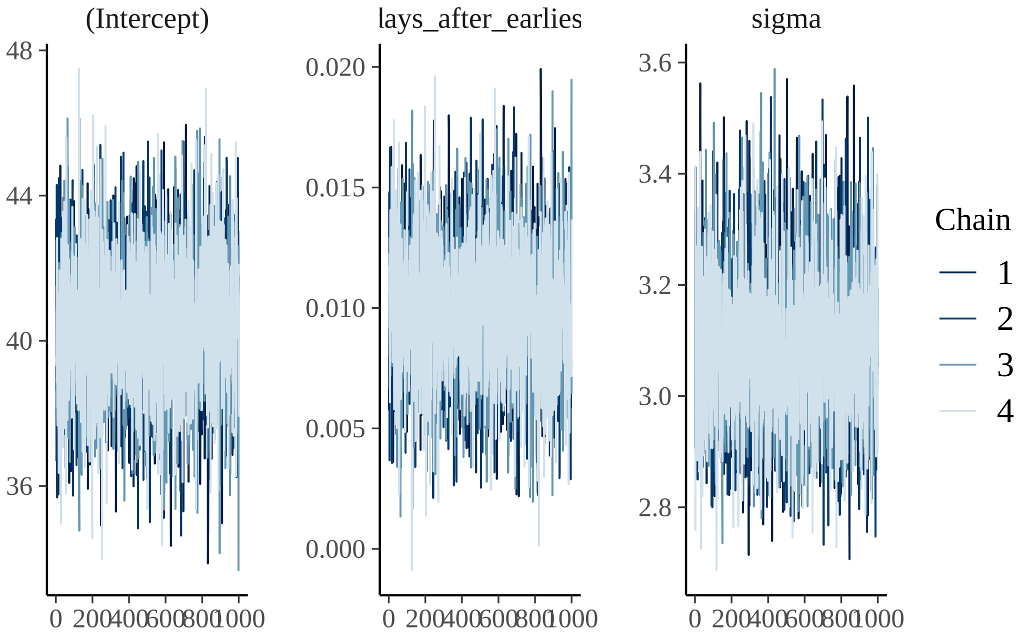

Figure 5: Model diagonostic for Harris

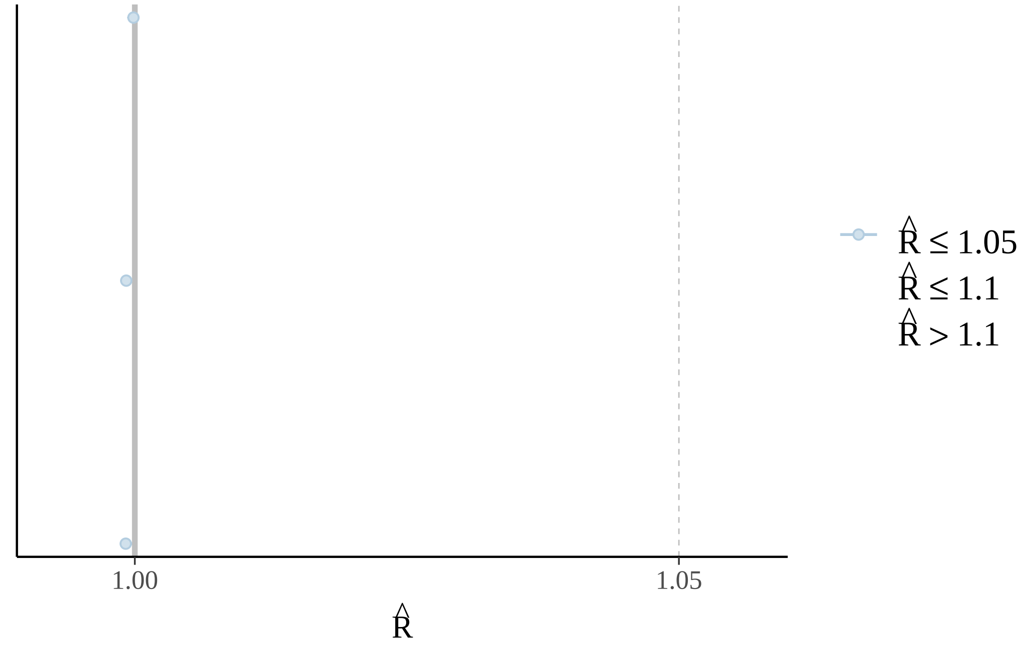

Figure 6: Model diagonostic for Harris

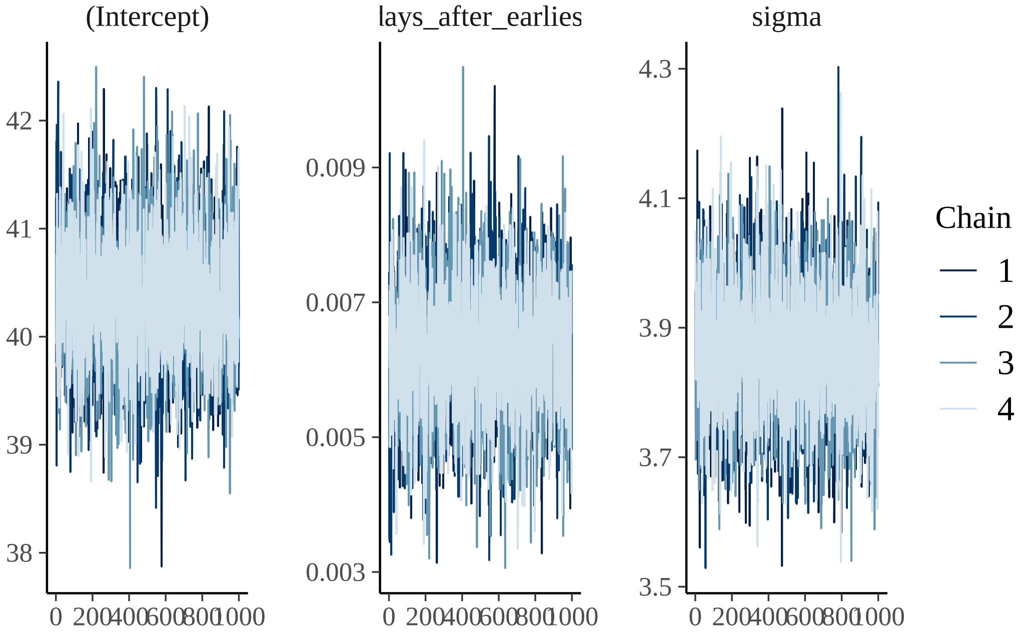

Figure 7: Model diagonostic for Trump

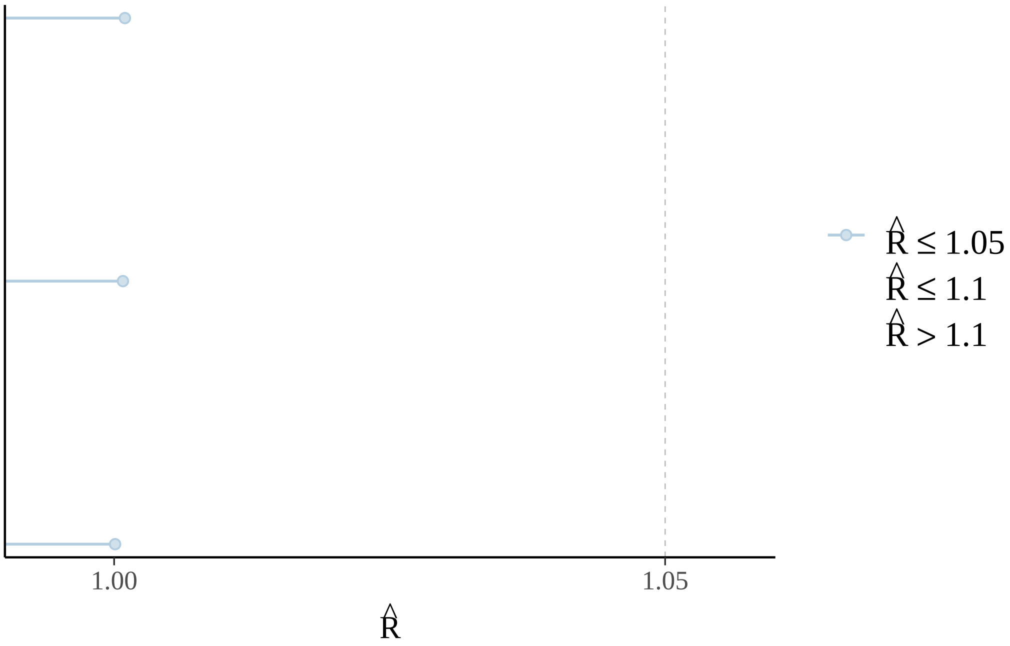

Figure 8: Model diagonostic for Trump

A.2.1 Population of Interest

The population of interest for this survey consists of all eligible voters residing in major U.S. swing states, specifically Arizona, Florida, Georgia, Michigan, North Carolina, and Pennsylvania. These states are known for their fluctuating political alignments and are expected to play a crucial role in the upcoming election.

A.2.2 Sampling Frame

The population sampled includes eligible voters from Arizona, Florida, Georgia, Michigan, North Carolina, and Pennsylvania. Participants were selected via an online panel.

A.2.3 Sample

The sample sizes for each state were as follows:

Arizona: 750 respondents

Florida: 1,350 respondents

Georgia: 927 respondents

Michigan: 970 respondents

North Carolina: 880 respondents

Pennsylvania: 1,070 respondents

A.3 Weakness & Strength of the methodology

In terms of strengths, Redfield & Wilton has a great reputation for producing reliable polling data. They have employ a mix of online and telephone survey, which add in various resources of collecting their data. This approach can help reduce bias. In addition, Redfield & Wilton often target swing states, which makes their polls results important to the US president election.

The weakness of their methodology would incorporate a certain potential bias based on their political leaning of their clients or media. This could cause a certain neutrality in their dataset.

Overall, Redfield & Wilton has been considered as a competent, reputable and reliable pollster with his variaty and methodology on polls. Even though the pollster contained a potential bias, it has been a reliable resource in prediction of US president election.

B Model details

C Appendix 2 - Idealized Methodology and Survey

C.1 Overview

This section introduces an idealized methodology and survey with $100K budget to predict the 2024 US presidential election. The goal is to maximize the accuracy of the prediction under the budget. The subsections that describe the details of the idealized methodology and survey include sampling approach, respondents recruiting method, data validation, poll aggregation and the survey questions.

C.1.1 Sampling Approach

Cluster sampling is the sampling approach used. Specifically clusters are the swing states as they are the states that would affect the election result, i.e. Arizona, Florida, Gegorgia, Iowa, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin. In each swing state/clusters, units are selected based on postal code. Using simple random sampling, 100 distinct postal codes are randomly selected. People whose living address has the postal codes selected are the target respondents. For the postal codes that lie in the non-residential area, the postal codes would be ignored.

C.1.2 Recruitment Strategy

The strategy to be implemented to recruit respondents is a combination of physical and online recruitment.

If the building corresponds to the postal code is a condo building, a paper with QR code to the survey is going to be put in the lobby or elevator, or sent to residents through the property management. If the building is a residential house, then a letter with the QR code would be put in the mailbox.

All respondents are rewarded with a $5 deposit to their bank account or a gift card of their choice. Each IP address is limited to answer the survey once.

C.1.3 Data Validation

To improve accuracy of the prediction results, the following data validation approaches will be done.

- Postal Code Validation

  - All the postal codes got from respondents are going to be validated to check if it is one of the randomly selected postal codes. If not, the response would not be considered when the prediction is done.

- Just-for-Rewards Prevention
  - To identify the responses that are not seriously answered, the last question of the survey is set test if the respondents are serious and careful when answering the questions.

- Age Validity
  - Responses with age under 18 but answered "Yes" to "Are you registered to Vote" are discarded.

C.1.4 Poll Aggregation

For each of the swing states, the result would be calculated based on the votes to Trump versus Harris because they are the candidates with the greatest chances of winning. Based on the decisiveness of the respondents and the likelihood of voting from the responses, the individual votes will be weighted when calculating the results for each swing states.

To aggregate the polls for all states, the result of each states is multiplied by its electoral votes. After summing the electoral votes for Trump or Harris, the estimated electoral votes that Trump or Harris get is available. The candidate that has more than 270 electoral votes would be predicted to win the election {U.S. National Archives and Records Administration (2024)}.

C.1.5 Survey

Google form of the survey is available here: https://docs.google.com/forms/d/1obaebX3zqw7WJEd1lHrF-DVBXdZt5tbjHofjsWj9zf8/prefill

1. What is the postal code of where you live?
2. Are you registered to vote?

- Yes
- No

3. Who would you vote for?

- Donald Trump - Republican
- Kamala Harris - Democrat
- Other Candidates

- Not Decided Yet

4. How decisive are you to vote for the option you chose in the last question?

- Very Decisive
- Pretty Decisive
- A Bit Indecisive
- Very Indecisive

5. How likely are you to vote for the election?

- Very Likely
- A Bit Likely
- Not Likely

6. What age group are you in?

- 0 - 18
- 19 - 30
- 31 - 50
- 51 - 70
- 71+

7.
8.
9.
10. How many characters are present in the word “President”?

- 6
- 8
- 10
- None of the Above

C.1.6 Budget Specification

- Physical and Online Recruitment: $30,000
- Rewards for Respondents: $50,000
- Data Collection & Validation: $10,000
- Other: $10,000

Total: $100,000

C.1.7 Conclusion

The survey uses cluster sampling to sample the swing states based on postal codes. After data collection and validation, weighting based on the likelihood of voting and decisiveness of the respondents, the total votes for Donald Trump and Kamala Harris are calulated. The candidate with more than 270 votes are predicted to win the election{U.S. National Archives and Records Administration (2024)}.

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

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

U.S. National Archives and Records Administration. 2024. "About the Electoral College." 2024. https://www.archives.gov/electoral-college/about.