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

cite R, packages, dataa The data used in this paper is source from FiveThirtyEight ("538presidentialpolls", 2024). T We use the statistical programming language R (R Core Team 2023)…. Our data…. Following (**tellingstories?**), 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 the two 2024 president nominates, 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, defined as those with a numeric grade of 2.0 or above and a transparency score of 4 or higher. We focused exclusively on polls where the *state* variable indicates national coverage, ensuring the poll represents nationwide opinions rather than a single state, which could introduce bias. Additionally, we filtered for polls conducted after July 21, 2024, the date when Kamala Harris announced her candidacy.

To prevent issues with missing data in our models, we excluded any rows where key variables (such as polling percentage or end date) were missing.

3.2 Model Specification

Two separate Bayesian linear regression models were fitted using the Brilleman et al. (2018) package. Both models specified the formula:

$$Y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \beta_0 + \beta_1 X_{i1} \tag{2}$$
$$\beta_0 \sim \text{Normal}(50, 5), \tag{3}$$
$$\beta_1 \sim \text{Normal}(0, 0.1) \tag{4}$$

Where Y_i is the support rate for Harris or Trump, separately, and X_{i1} is the number of days passed after the latest polling sample of Trump or Harris.

3.3 Model Justification and Limitations

We opted to use two separate models for Trump and Harris, as their supporter demographics differ significantly in ways that are challenging to capture within a single or even a small set of variables Center (2024). The polling dataset we obtained primarily reflects the characteristics and trends within individual polls rather than offering direct predictors of the presidential election outcome. To maximize the model's reliability, we refined the dataset by selecting polls with high credibility rather than attempting to predict outcomes directly from the available data.

Our model uses a simple linear approach with time as the sole predictor variable. While this may introduce potential bias by making a simplifying assumption about the trend over time, it also has the advantage of being straightforward to fit and interpret. This trade-off allows us to capture general trends while minimizing the risk of overfitting to limited or potentially inconsistent data sources.

4 Results

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

Table 1: Summary statistics of the Bayesian model for Harris

Term	Estimate	Std. Error	2.5% CI	97.5% CI
(Intercept)	40.466	1.988	37.287	43.462
days_after_earliest	0.010	0.003	0.005	0.014

Table 2: Summary statistics of the Bayesian model for Trump

Term	Estimate	Std. Error	2.5% CI	97.5% CI
(Intercept)	40.531	0.578	39.620	41.502
days_after_earliest	0.006	0.001	0.005	0.008

4.1 Overall trend

The intercept of both model is positive, with Harris at 40.47 and Trump slightly lower than Harris at 40.36. The trend of support rate for both candidates with respect of time is positive but Harris's increment (0.10) is larger than Trump's (0.06).

4.2 Model Diagnostics

The model diagnostics indicate a strong fit and reliable predictive performance for both candidates, as detailed in the plots in Appendix Section C. Posterior predictive checks (Figures **?@fig-pos_check_H** and **?@fig-pos_check_T**) show that the predicted values align well with the observed data, suggesting that the models effectively capture key data patterns. Furthermore, the comparison of posterior and prior variables indicates minimal deviation, affirming an appropriate choice of priors. The credible intervals for the predictors (Figures Figure 5 and Figure 6) encompass the observed support rates, demonstrating that the models adequately account for uncertainty in their predictions.

Additional robustness checks (Figures **?@fig-DI_H** and **?@fig-DI_T**) confirm the stability of these models. The R-hat values for all parameters are below 1.1, indicating good convergence of the MCMC chains and reliable parameter estimates. This convergence supports the conclusion that the sampling process reached a stable distribution, reinforcing the credibility of our inferences.

In conclusion, these diagnostics verify that the models are both stable and effective in reflecting trends in the data, confirming their reliability for further analysis.

A Appendix

B Additional data details

B.1 Model check

```
Drawing from prior...
```

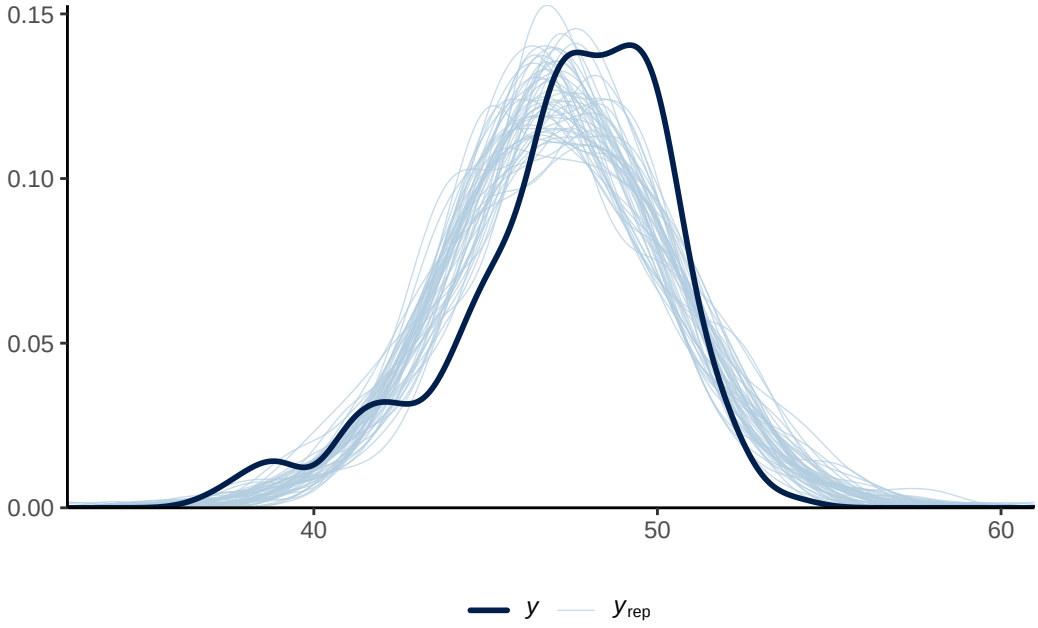

Figure 1: Posterior Predictive Check for All Models

```
Drawing from prior...
```

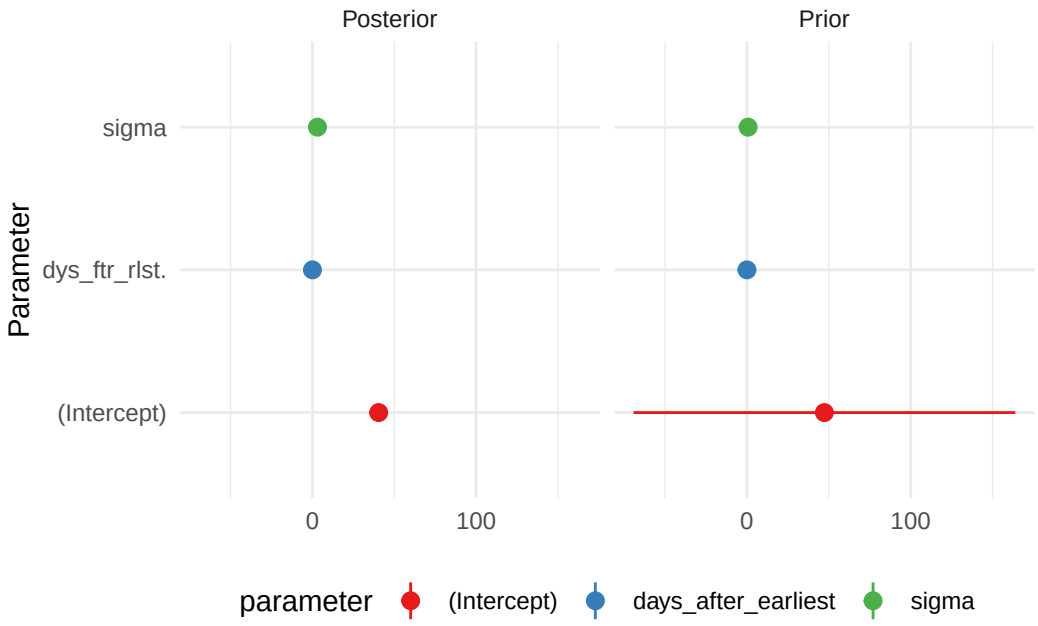

Figure 2: Posterior Predictive Check for All Models

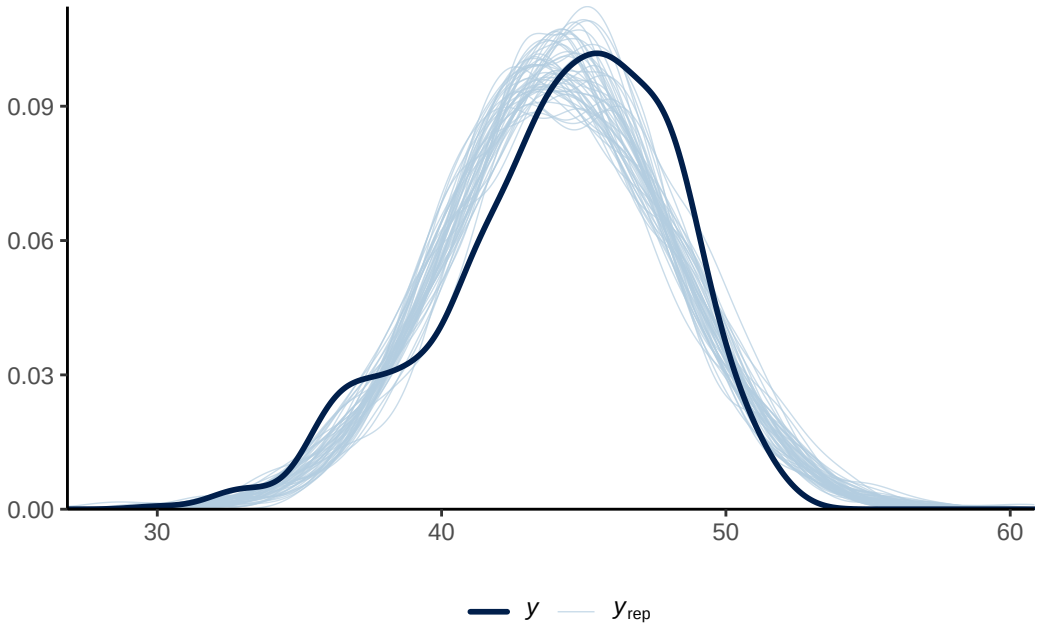

Figure 3: Posterior Predictive Check for All Models

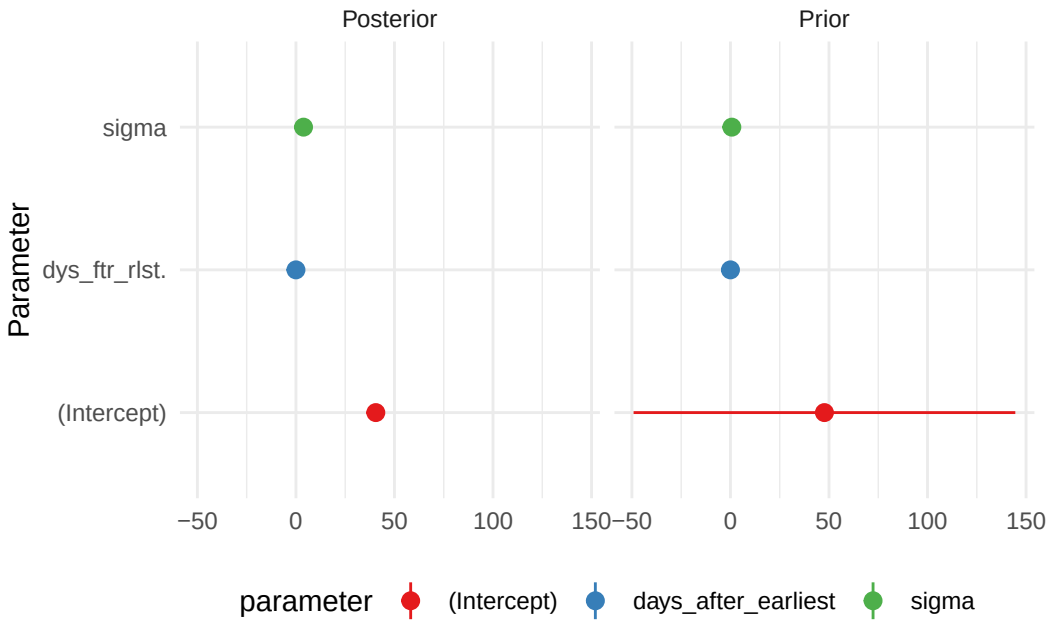

Figure 4: Posterior Predictive Check for All Models

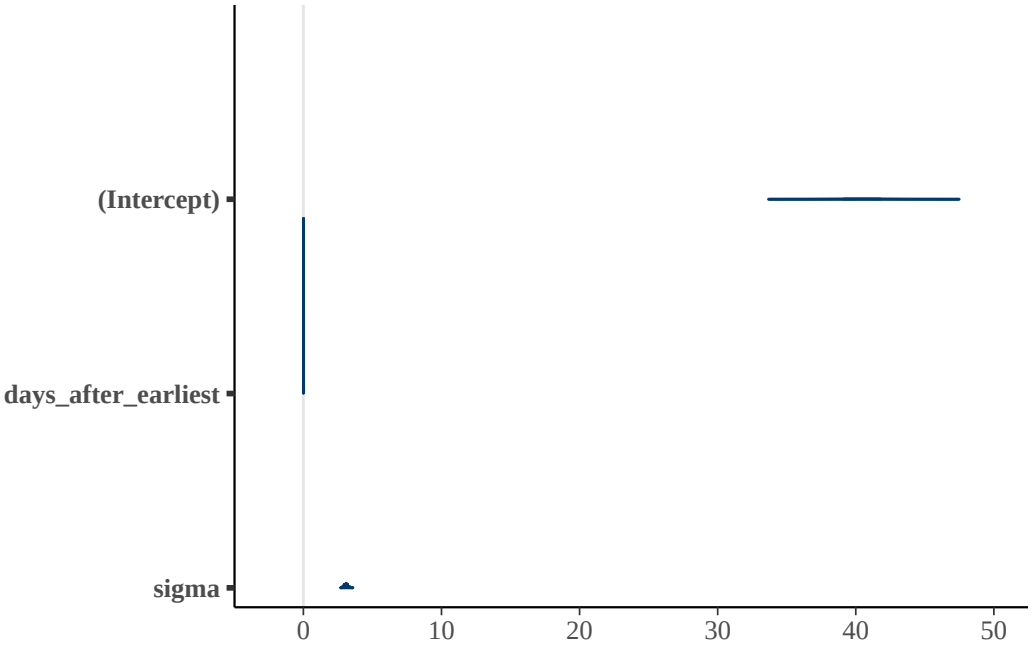

Figure 5: CI for predictors for Harris

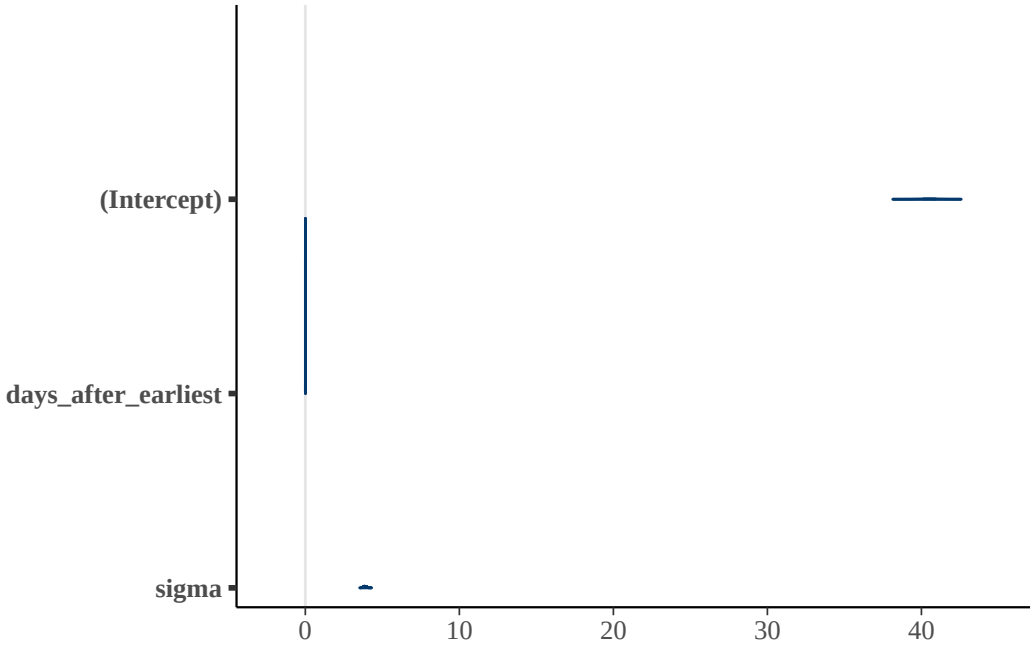

Figure 6: CI for predictors for Trump

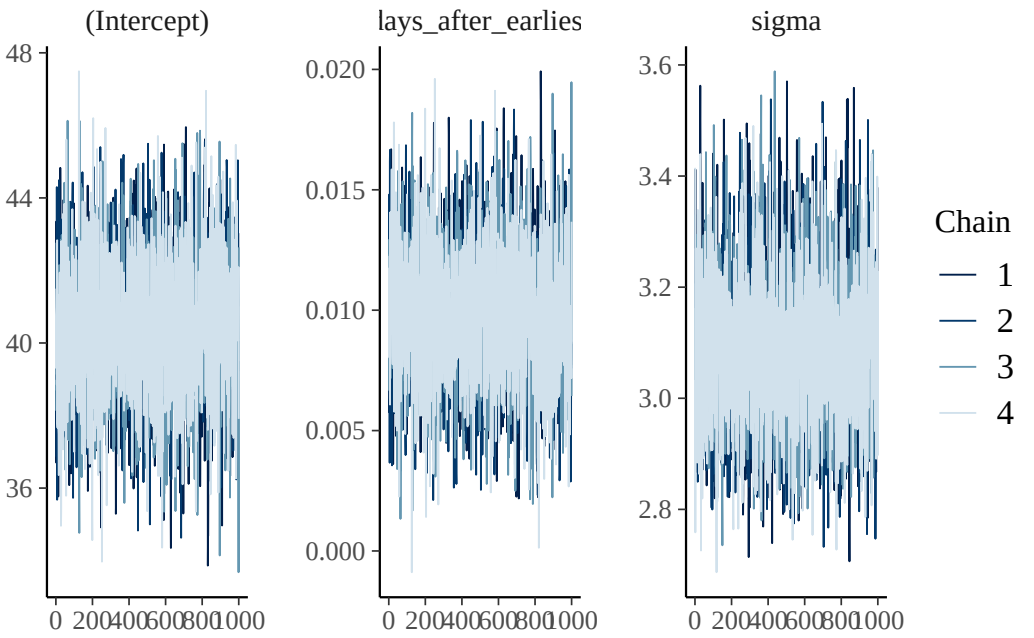

Figure 7: Model diagonostic for Harris

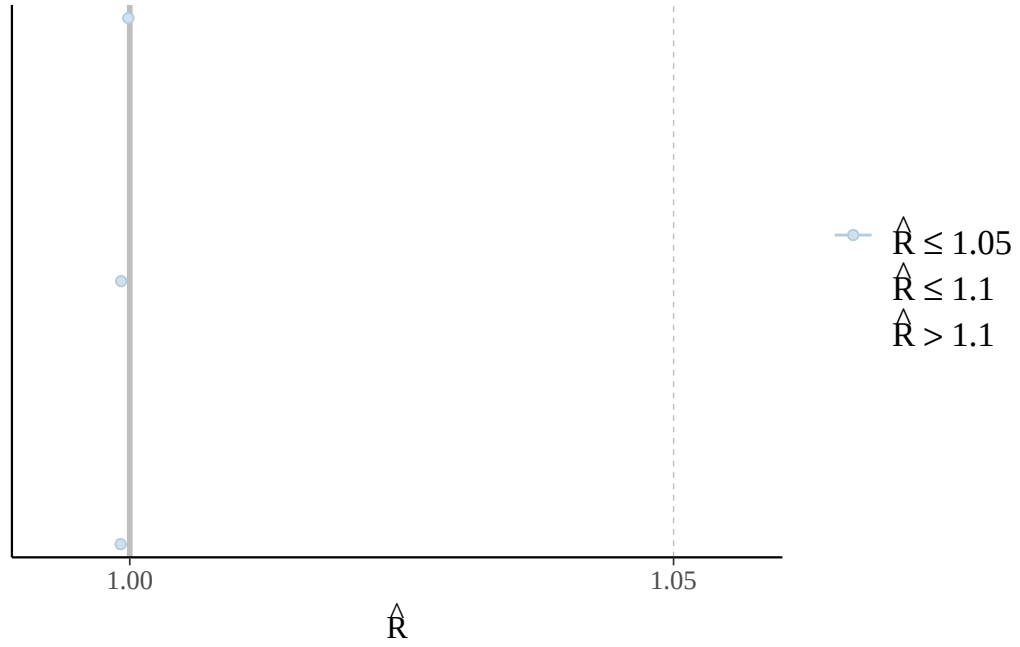

Figure 8: Model diagonostic for Harris

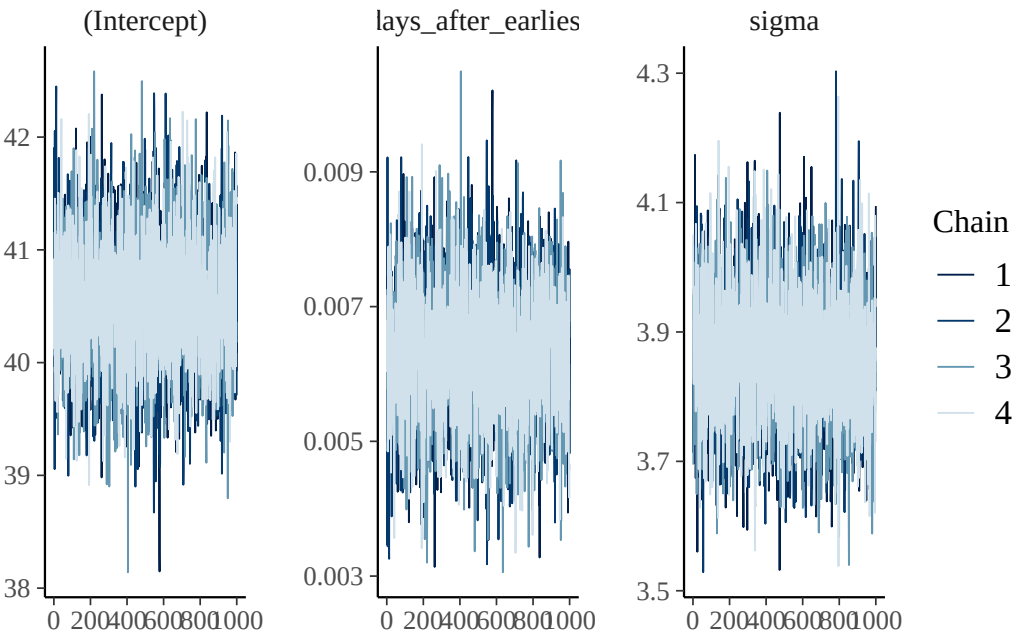

Figure 9: Model diagonostic for Trump

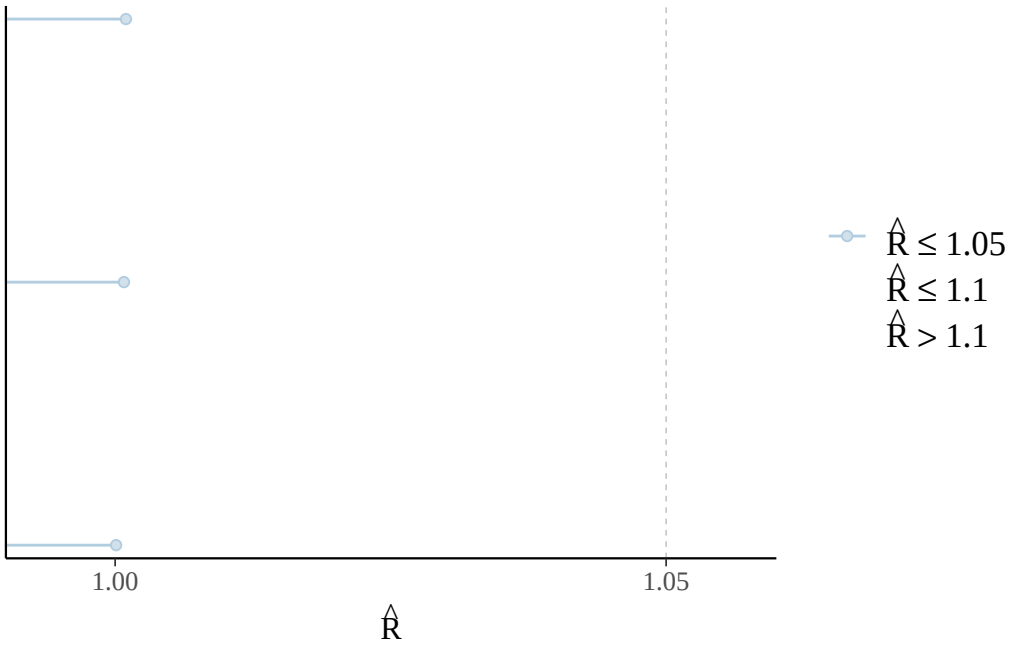

Figure 10: Model diagonostic for Trump

B.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.

B.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.

B.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.

B.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

B.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.

C Model details

D Appendix 2 - Idealized Methodology and Survey

D.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.

D.1.1 Sampling Approach

Cluster sampling is the sampling approach used. Specifically clusters are states. In our case, 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, units are selected based on postal code. Using simple random sampling, 100 distinct postal codes are randomly selected. People whose living address have 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.

D.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.

D.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.

D.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)}.

D.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. If the election were held today, who would you most likely vote for?

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

- Not Decided Yet
- Prefer not to say

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 many characters are present in the word “President”?

- 6
- 8
- 10
- None of the Above

6. What age group are you in?

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

7. What is your sex?

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

8. What is your Ethnity/Race?

- White
- Black or African American
- Hispanic or Latino
- Asian
- Native American or Alaska Native
- Native Hawaiian or Other Pacific Islander
- Middle Eastern or North African
- Prefer not to say
- Other (please specify) [optional text box]

9. What is your Education Level?

- Less than high school
- High school diploma or equivalent
- Some college, no degree
- Associate's degree
- Bachelor's degree
- Master's degree
- Professional or doctoral degree (JD, MD, PhD, etc.)
- Prefer not to say

10.How do you plan to vote?

- In-person on Election Day
- In-person early voting
- Mail-in/Absentee ballot
- Undecided

11. Is there anything else you'd like to share about your voting decision or concerns?

D.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

D.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

Brilleman, SL, MJ Crowther, M Moreno-Betancur, J Buros Novik, and R Wolfe. 2018. "Joint Longitudinal and Time-to-Event Models via Stan." https://github.com/stan-dev/stancon_talks/.

Center, Pew Research. 2024. "The Harris-Trump Matchup." https://www.pewresearch.org/politics/2024/10/10/the-harris-trump-matchup/.

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.