Will Kamala Harris will win the 2024 Presidential election*

Analysing the polls using baysian modeling

Yan Mezhiborsky Cristina Burca Shreya Sakura Noskor

November 4, 2024

Abstract

We analyse the probability of Kamala Harris winning in the 7 swing states of USA using baysian modeling;the states we look at are Arizona, Michigan, Pennsylvania, Nevada, Georgia, Wisconsin and North Carolina. USA is a large country with its national economy affecting the global economic conditions, which is why predicting the future President of USA will help understand the future economic condiotion of the world. Through baysian modelling we found that North Carolina (47.26% support for Harris), Nevada(46.43% support for Harris), Wisconsin(48.38% support for Harris), Michigan(47.30% support for Harris), Pennsylvania(48.02% support for Harris) will vote Kamala Harris, while the rest of the swing states might vote Donald Trump, meaning that majority of the swing states will vote Kamala Harris. Provided this we predict that Kamala Harris will win the election. By creating a baysian model based on 'polls of polls' where we compare the results from different polls, we are able to make this prediction.

Table of contents

*Code and data are available at: https://github.com/Mezhi18/US_Election2024.git.

1 Introduction

The 2024 United States presidential election represents a pivotal moment in the country's political landscape. As in previous elections, swing states are projected to play a critical role in determining the outcome. Swing states, which are characterized by their shifting voting patterns and balanced support for both major political parties, have historically been the focus of intense campaign efforts and polling analyses. Understanding the dynamics and voter preferences in these states is crucial to gaining insight into the broader electoral trends that could shape the nation's future.

In this paper we aim to analyze polling data related to the 2024 election, with a particular emphasis on the swing states. By examining various polls and identifying patterns in voter sentiment, we seek to uncover the factors that may influence voter behavior in these highly contested regions. Our analysis will explore demographic shifts, the impact of key issues, and the level of voter engagement across different swing states. Through a comprehensive statistical approach, we aim to contribute to the understanding of the evolving electoral landscape and provide meaningful insights into the opinions shaping the 2024 presidential election. Using Bayesian regression modeling, we predict the percent of support for Kamala Harris in swing states based on pollsters and the number of days since she announced her candidacy. Details on our data, variables, and model are discussed further in the Data and Model sections.

Traditional election forecasts often rely on single polls or sources, but our approach combines results from multiple polls to provide a more accurate prediction. By focusing on swing states, which historically alternate between parties, we capture the areas most likely to influence the election outcome. The 2 major parties that are in competition are the Republicans and the Democrats. Some states exhibit a consistent pattern of voting for a particular party, regardless of the candidate. In contrast, swing states are more variable, providing key insights into who might win the election. Our goal is to predict the outcome of the U.S. presidential election by synthesizing data from multiple polls, focusing efficiently on the indecisive swing states.

To delve further into our approach, we introduce our estimated - the parameter of interest- which is the predicted support for Kamala Harris in the swing states (measured in %). We estimate this by creating Baysian model with 3 parameters; state, pollster, and days since Harris's candidacy announcement. We set priors using a normal distribution centered at 0, with a standard deviation of 2.5. Given that swing states tend to lack a dominant party preference, the expected support will be approximately 50%. We further explore these variables and their meanings other graphs in the (Section 2). From our analysis in (Section 3.2) we find that the average predicted support for Kamala Harris for the 7 swing states were as follows: Arizona at 46.26%, Nevada at 46.43%, Georgia at 46.59%, North Carolina at 47.26%, Michigan at 47.30%, Pennsylvania at 48.02%, Wisconsin at 48.38%. Based on this we predict that Arizona, and Georgia will vote Donald trump and the rest of the states will vote Kamala Harris.

Accurate election predictions are important as they help individuals, businesses, and governments anticipate potential shifts in policy, the economy, and international relations. A new president can affect the previously mentioned factors as well as address key issues like healthcare, education, and climate change. Knowing which candidate is likely to win enables groups to consider how their plans might align with anticipated policies, encouraging public engagement and participation in the democratic process. It also helps keep people informed and engaged in the political process, making them more likely to vote and participate in democracy in the future.

The paper is structured as follows: the Data section (Section 2) covers our methodology and data collection process, detailing the variables used in the analysis. The Model section (Section 3) introduces our Bayesian approach and includes related visualizations. The Results section (Section 3.2) provides a model summary and predicted support percentages for each swing state. In the Discussion section (Section 4), we interpret our findings, present final predictions, and discuss limitations, weaknesses and next steps. We also give a final word about who will win the upcoming election. We also have additional appendices (Section A), where we talk about the methodology of the polls and how they compare. The second appendix talks about a survey we designed with reasons and disadvantage. The third appendix shows the Posterior Predictive check of our model.

2 Data

2.1 Overview

As our paper is about the 2024 United States federal election and more specifically we are looking at the polling Data and the polling data comparing the two candidates, former President Donald Trump and Vice President Kamala Harris, for the upcoming election. Our original Data set had over 16,000 unique entries from different pollsters, the business or Organization that conducts the poll. Each poll has two entries, one giving the Data for the polling opinions of Donald Trump and the second for Kamala Harris. We have acquired our polling data from FiveThirtyEight (2024).

As there are over 50 variables many of which are redundant to our paper we will only discuss those that we have kept in our clean data as they are the only ones we use in our analysis.

- `pollster`: Shows name of the Pollster that conducted the poll.
- `sample_size`: The number of people that participated in the specific poll.
- `state`: This variable tells us in which States the poll was conducted.
- `candidate_name`: This is the full name of the selected candidate.
- `pct`: This tells us the percentage of participants that intend to vote for the selected candidate.
- `end_date`: The date the pollster finished conducting the poll.

The variables that we have create are:

- `num_harris`: The number of participants that intend to vote for Kamala Harris.

- `end_date_num`: The number of days since the first poll since Harris announced her candidacy, calculated using `end_date`.

Each pollster has a numeric grade from 1.0 to 3.0, which indicates the quality/ reliability of the respective pollster. Additionally, each pollster is also given a transparency score from 1.0 to 10.0 reflecting how 'transparent' the pollster is, or how much information is disclosed about its polls and methodology. It is important for pollsters to maintain high numeric grades and transparency scores because these metrics directly reflect the quality and reliability of their data. To ensure the highest level of accuracy in our predictions, we only include polls with a numeric grade of 1.5 or above and a transparency score of 6.0 or above.

We use the statistical programming language R (R Core Team 2023) to compile this paper. Our data comes from (FiveThirtyEight 2024) and was cleaned and visualized using the packages Tidyverse (Wickham et al. 2019), Lubridate (Grolemund and Wickham 2011), Janitor (Firke 2023), and Knitr (Xie 2023) and modeled using Rstanarm (Goodrich et al. 2022), ModelSummary (Arel-Bundock 2022) and BayesPlot (Gabry and Mahr 2023).

2.2 Methodology and Measurement

This paper employs a "polls-of-polls" approach to analyze and predict our outcome, combining data from multiple pollsters. By integrating results from various sources, this method captures a broader range of perspectives, reducing the influence of individual biases and offering a more balanced view. Unlike relying on a single poll, which may be shaped by its unique methodological biases, the "polls-of-polls" approach enhances reliability and increases the overall validity of our predictions.

Each pollster uses methodologies they believe are most effective for capturing an accurate snapshot of public opinion. Pollsters begin by selecting a sample of voters they consider representative of the broader population, factoring in various demographics such as age, region, and socio-economic background. However, polling methods differ significantly; some pollsters rely on phone interviews, while others use online surveys, text messaging, or a combination of methods (mixed-mode). After data collection, pollsters apply weighting adjustments to account for demographic imbalances, ensuring that underrepresented groups are adequately reflected in the results.

The diversity in methodologies helps reduce systematic bias compared to using a single poll source. Different methods may also be more effective in reaching certain demographics—such as older voters through phone surveys and younger voters through online platforms—broadening the inclusivity and representativeness of the sample.

In predictive modeling, accounting for pollster variability and methodology is essential to avoid overfitting to any single source. By including pollster as a predictor in the model, we allow it to recognize that poll results may vary not only due to real changes in public opinion but also because of differences in survey methodologies. This consideration enables the model to generate more accurate, generalized predictions, offering a reliable forecast that reflects a well-rounded view of public opinion.

2.3 Data Visualization

In this analysis, we examine polling trends for Kamala Harris across key swing states, focusing on recent polling data from Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin (Times 2024a). As shown in Figure 1, which displays the support percentages for both Harris and Donald Trump side-by-side by state, the data reveals a competitive race, with relatively close support levels for both candidates in most states. While Trump has a slight edge in Arizona, Georgia, and North Carolina, Harris maintains a modest lead in Michigan, Nevada, and Pennsylvania.

Moving to Figure 2, which isolates Harris's support percentages across states, we gain insight into the consistency and variability of her support. The box plot highlights states like Wisconsin and Pennsylvania,

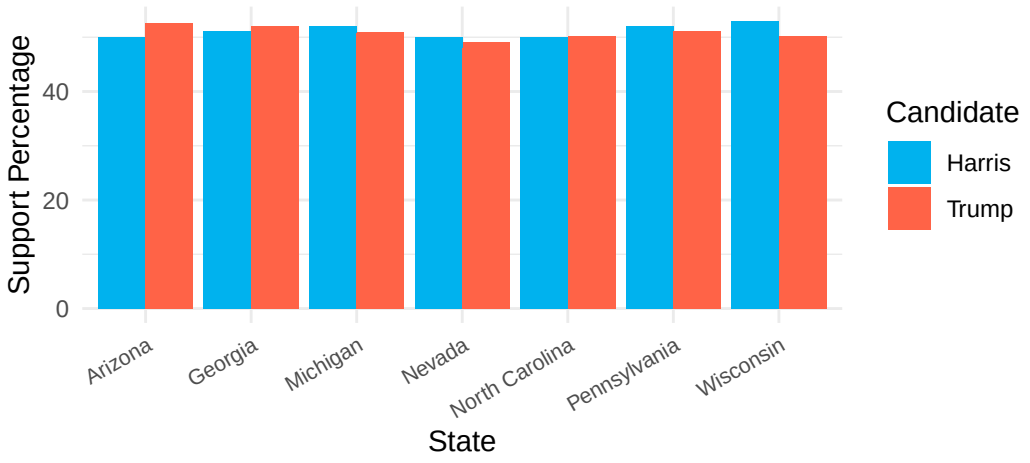

Figure 1: Distribution of Support Percentage by State for Harris vs. Trump

where the smaller interquartile ranges indicate less variability, suggesting that Harris's support is more stable in these regions. Conversely, states such as Nevada and Georgia show greater variability, as reflected in the larger spread of the boxes and whiskers, indicating that Harris's support fluctuates more in these areas.

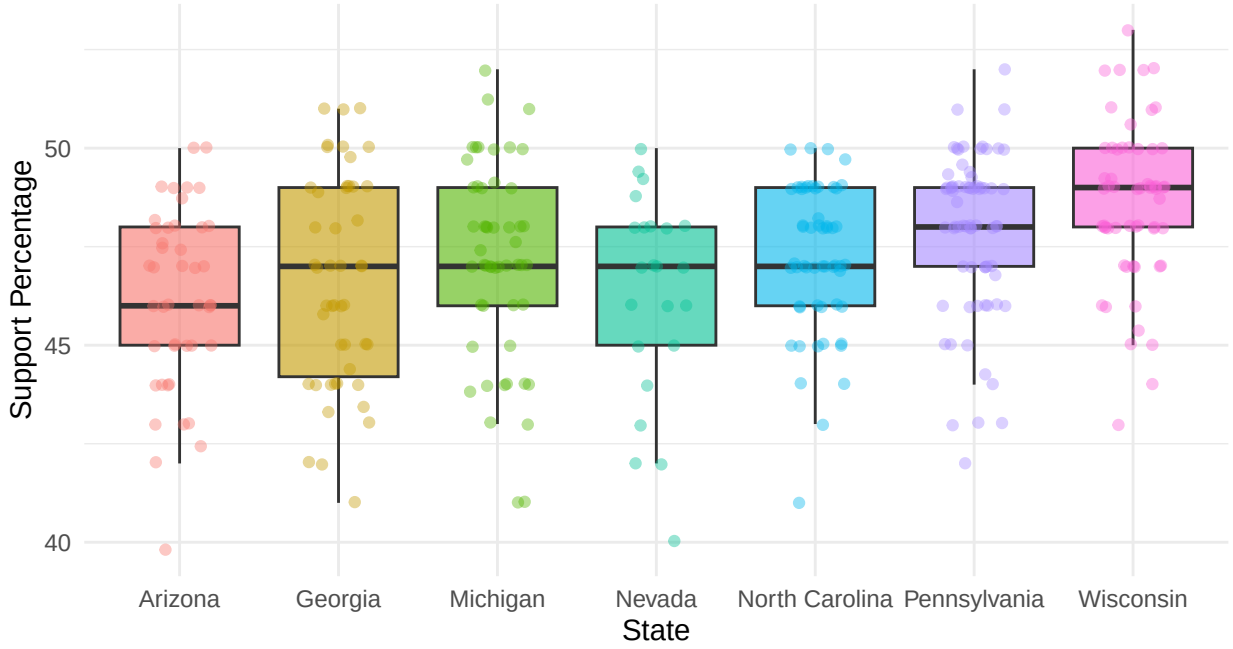

Figure 2: Support Percentage for Harris by State

To explore temporal trends, we introduce the variable `end_date_num`, representing the number of days since Harris declared her candidacy. The next figures examine support trends over time in individual states.

In Figure 3, Pennsylvania and Wisconsin display distinct trends. In Pennsylvania, support for Harris has been relatively consistent since early September, stabilizing around 48.5%. In contrast, Wisconsin shows a peak in support in mid-September, followed by a gradual decline that stabilizes in October. These observations suggest that while Pennsylvania remains a consistent stronghold for Harris, Wisconsin's voter sentiment may be more susceptible to fluctuations.

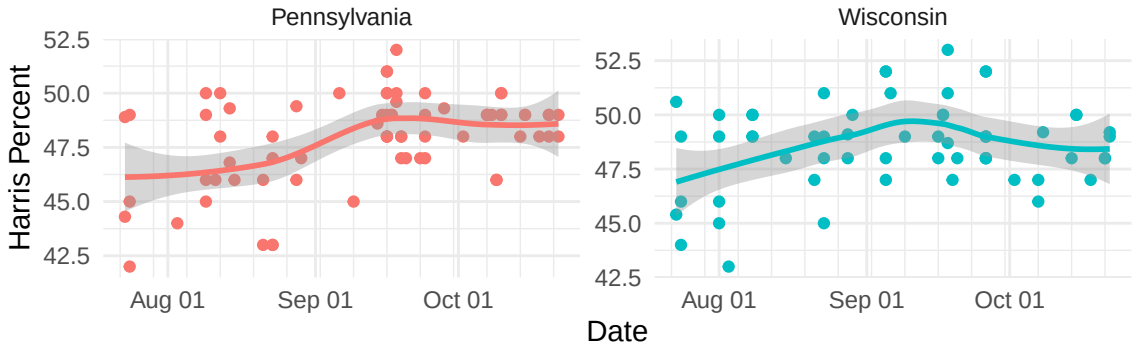

Figure 3: Predicted Support Percentage for Harris Over Time in Pennsylvania and Wisconsin Since Candidacy Declaration

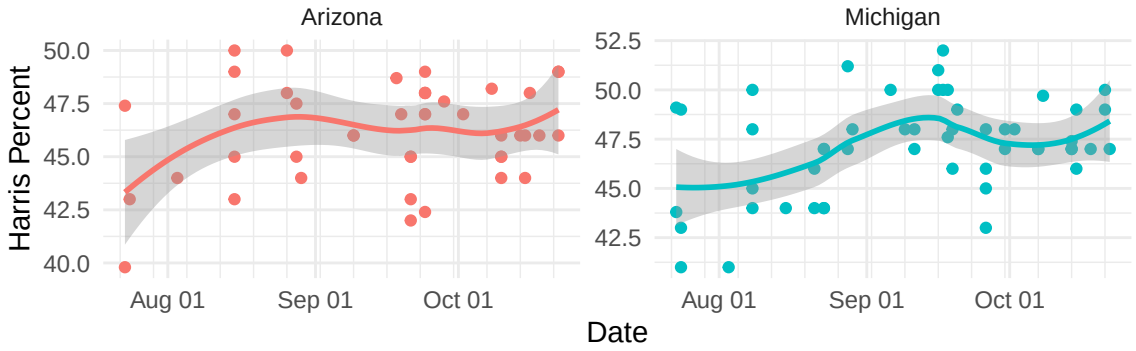

Figure 4: Predicted Support Percentage for Harris Over Time in Arizona and Michigan Since Candidacy Declaration

Figure 4 further examines Harris's support trends in Arizona and Michigan. Both states exhibit an upward trend in support starting in early October, which may suggest increasing favorability or strategic campaign efforts in these regions. This upward trajectory could indicate an opportunity for Harris to solidify her support in these competitive areas.

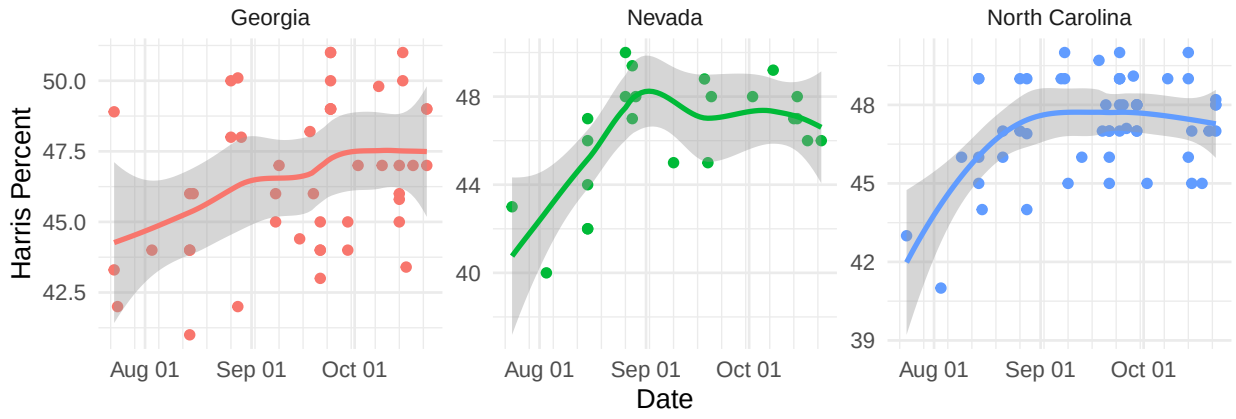

Figure 5: Predicted Support Percentage for Harris Over Time in Georiga, Nevada, and North Carolina Since Candidacy Declaration

In Figure 5, we analyze Georgia, Nevada, and North Carolina. Nevada demonstrates the highest variability in Harris's support, which could reflect the state's dynamic political response to recent campaign efforts. Georgia shows a steady upward trend, indicating growing support for Harris, while North Carolina reveals a slight decline since October, suggesting a possible shift in voter support.

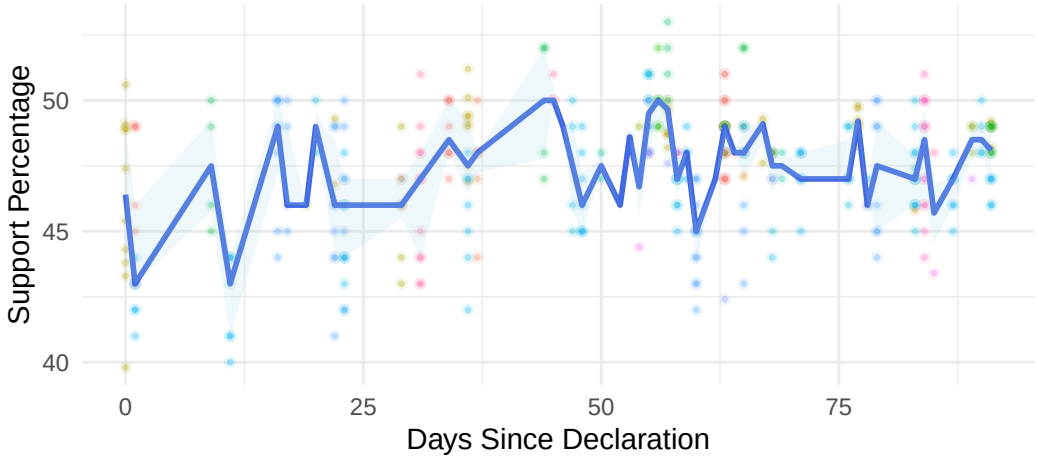

Figure 6: Variation in Harris's Support Percentage Across Pollsters Since Harris' Candidacy Declaration

In Figure 6, we examine the impact of pollster variability on Harris's support percentage over time. This visualization introduces the variable pollster, highlighting the differences in polling results that arise from varying methodologies. The chart reveals considerable fluctuations and occasional outliers in Harris' support, likely due to the distinct approaches each pollster employs in sampling and weighting responses.

Overall, these visualizations illustrate that Harris's voter support is not only state-dependent but also fluctuates significantly over time. In particular, many states exhibit notable shifts around mid-September, pointing to potential influences from external events or campaign dynamics during this period. Furthermore, the observed variability across pollsters emphasizes the necessity of a robust model that can account for differences

in pollster methodologies. By incorporating temporal patterns, state-specific characteristics, and pollster variability, our predictive model will aim to provide a more comprehensive and nuanced forecast of support for Harris in key regions, helping to isolate genuine shifts in public opinion from noise introduced by polling differences.

2.4 Outcome and Predictor variables

In this analysis, our primary outcome variable is the support percentage for Kamala Harris. This variable represents the proportion of respondents in each poll who indicated support for Harris. Given the competitive nature of the election, understanding how this support varies across states and over time is crucial for identifying trends and forecasting her overall performance.

In this analysis, our primary outcome variable is the support percentage for Kamala Harris. This variable represents the proportion of respondents in each poll who indicated support for Harris. Given the competitive nature of the election, understanding how this support varies across states and over time is crucial for identifying trends and forecasting her overall performance.

To explain the outcome, we consider several predictor variables:

1. **State**: Since support for candidates can vary significantly across different states due to regional demographics, political history, and local issues, the state variable allows us to capture these geographical differences. As shown in the earlier visualizations, some states demonstrate consistently higher or lower support for Harris, emphasizing the importance of including state-level distinctions in our analysis.
2. **Days Since Declaration**: Representing the number of days since Harris declared her candidacy, this temporal variable helps us capture shifts in voter sentiment over time. Our analysis reveals several critical points where support trends noticeably change, such as the upward or downward shifts in mid-September. This predictor allows the model to account for these temporal patterns and track how Harris's support evolves as the campaign progresses.
3. **Pollster**: Given that different pollsters employ varied methodologies, sampling techniques, and weighting schemes, the pollster variable captures the potential variability introduced by each polling organization. As illustrated in Figure 6, the results can vary considerably depending on the pollster, with some reporting notably higher or lower support percentages for Harris. By including pollster as a predictor, we aim to control for this source of variability, helping the model focus on underlying trends rather than discrepancies introduced by polling differences.

Each of these predictors provides a distinct layer of insight into Harris's support trends. Together, they form the foundation of our predictive model, which will use these variables to more accurately forecast shifts in voter support and clarify the impact of state, time, and pollster methodology on public opinion.

3 Model

3.1 Model Step- Up

The Bayesian model we employ leverages probabilistic reasoning to predict Kamala Harris's support across different states. Bayesian models offer a flexible framework for incorporating prior information alongside observed data to produce posterior distributions, which reflect updated opinions based on the data. This model was chosen for its ability to incorporate prior knowledge, handle uncertainty, and generate predictions that can adapt to the polling data, providing a more comprehensive forecast of Harris's support across swing states.

We used the `stan_glmer` function from the `rstanarm` package to create a Bayesian regression model with a binomial family and logit link. This model is designed to predict the proportion of respondents who support Kamala Harris based on several key factors.

The model specification is as follows:

$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (1)\\
\mu_i &= \beta_0 + \beta_1 \times \text{State}_i + \beta_2 \times \text{Pollster}_i + \beta_3 \times \text{Days since first poll}_i && (2)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (4)\\
\beta_2 &\sim \text{Normal}(0, 2.5) && (5)\\
\beta_3 &\sim \text{Normal}(0, 2.5) && (6)\\
\sigma &\sim \text{Exponential}(1) && (7)
\end{aligned}
$$

The outcome variable in this model is the proportion of respondents who support Harris in each poll, represented as y_i. This binomial outcome (successes out of the total sample size) is modeled as a function of several predictor variables that account for differences by state, pollster, and time.

- y_i is the dependent variable, representing the proportion of respondents who support Harris.
- β_0: The intercept term, representing the expected proportion of y_i when all other predictors are set to zero.
- **State** (β_1): Categorical variable that accounts for differences in support across seven key swing states.
- **Pollster** (β_2): Categorical variable capturing variability introduced by the polling organization.
- **Days since first poll** (β_3): Numeric variable representing the number of days since Harris declared her candidacy.

Each coefficient β follows the normal distribution with a mean of 0 and a standard deviation of 2.5, providing some regularization while allowing for a range of possible effects. The standard deviation σ is given an Exponential(1) prior, which constrains it to positive values and provides a realistic expectation of variability around the mean Each predictor captures a unique dimension of voter sentiment:

- **State** accounts for regional differences in support due to factors like demographics and political history.
- **Pollster** adjusts for methodological variations in polling, which we observed to influence support percentages.
- **Days Since Declaration** captures temporal trends, allowing the model to incorporate fluctuations in voter sentiment over the campaign period.

By combining these predictors, the model aims to forecast Harris's support by isolating genuine trends from random variability.

3.1.1 Model justification

Based on the data overview in Section 2.3, we observe that Harris's support varies across key swing states, and we expect these regional differences to influence our model's predictions. The state variable is anticipated to capture these geographic variations, with states like Arizona, Michigan, Georgia, and Pennsylvania showing relatively consistent high levels of support for Harris. In contrast, Nevada, North Carolina, and Wisconsin exhibit a declining trend in support. We expect the intercept to be around 50%, as this reflects the average level of support observed across all states and pollsters.

A model validation is included in `scripts/05-exploratory_data_analysis.R`, and a Posterior Predictive Check and summary statistics for the model are included in Section A.3.

3.2 Model Results

The model predicts an average support percentage of approximately 47.33% for Kamala Harris across the selected states. This percentage reflects the central tendency of predicted support within the dataset, factoring in the influence of state, pollster, and the number of days since the first poll.

The results indicate variations in predicted support across states. For instance, Pennsylvania and Wisconsin show the highest average predicted support percentages (48.02% and 48.38%, respectively), while states like Nevada and Arizona display slightly lower support.

In states like Arizona, Georgia, Michigan, and North Carolina, the model predicts a steady increase in support for Harris over time, as shown in Figure 7. In states such as Nevada, Pennsylvania, and Wisconsin, the model predicts a decline in support over time, as seen in Figure 8.

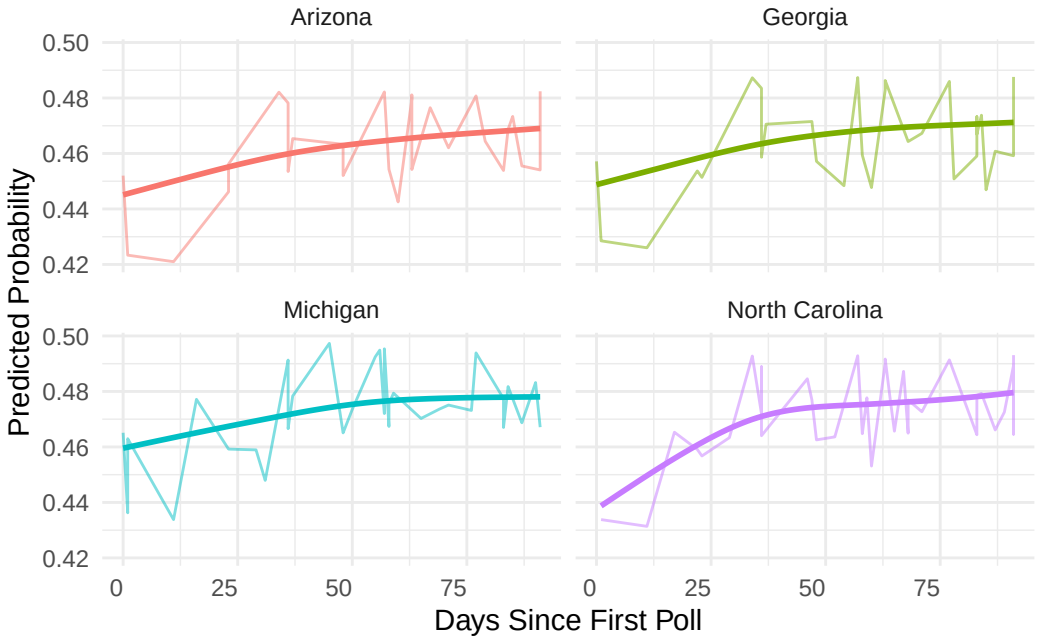

Figure 7: Predicted Probability of Support for Harris Over Time in Arizona, Georgia, Michigan, and North Carolina Since Harris' Candidacy Declaration

In Figure 8, we see states of Nevada, Pennsylvania and Wisconsin have a singular declining trend.

Table 1: Average Predicted Support for Harris by Swing State

State	Avg. Predicted Support for Harris
Arizona	46.26
Nevada	46.43
Georgia	46.59
North Carolina	47.26
Michigan	47.30
Pennsylvania	48.02
Wisconsin	48.38

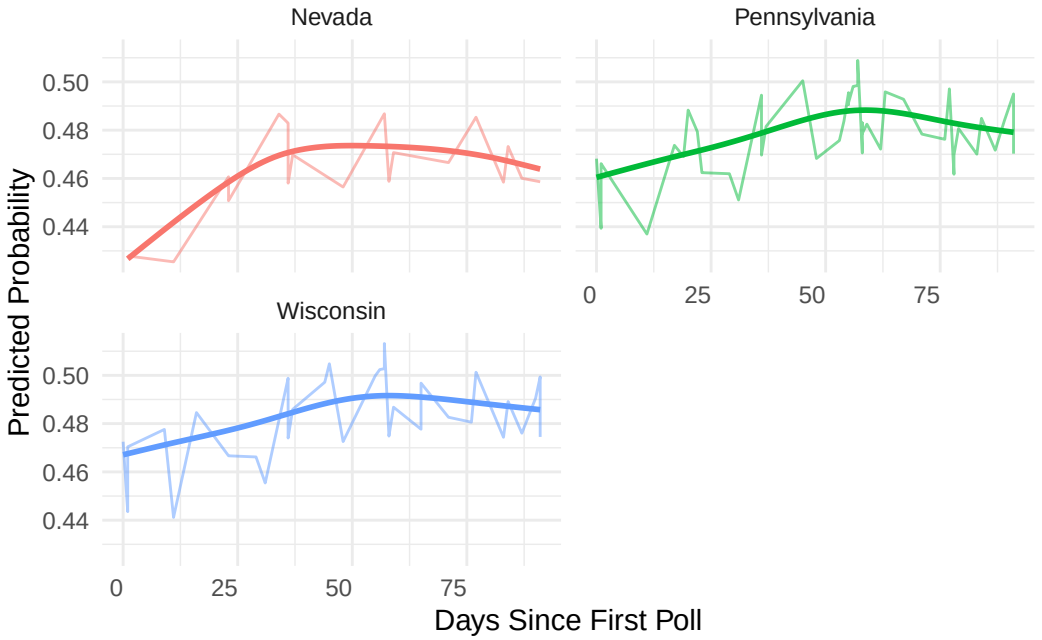

Figure 8: Predicted Probability of Support for Harris Over Time in Nevada, Pennsylvania, and Wisconsin Since Harris' Candidacy Declaration

4 Discussion

In this section, we will analyze each state individually, considering our collected data, generated models, and the voting history of each state in previous elections to project likely outcomes.

4.1 Arizona

Arizona is a very important swing state especially as the state has switches which party they voted for in in the last two elections with Trump winning 49% of the vote and Clinton with 45.5%, which considering Arizona is a swing state is a rather significant6 margin. While in the 2020 election Biden ended up winning Arizona with 49.4% of the vote and Trump 49.1%, which is a much smaller margin of victory and what makes Arizona so questionable in the 2024 election.

We can see in Figure 4 that while the polling numbers for Harris started somewhat low around 42%, the have gone up significantly and have continued to trend upwards in the most recent polls. Although, in the Figure 1 trump still has larger support in the state, and the largest polling difference in Trump's favour out of all the swing states polled (270toWin (2024)).

We project Arizona will vote for: **Donald Trump**

4.2 Georgia

Georgia's last two elections outcomes were very alike Arizona's outcomes. In 2016, Trump secured the state with 50.8% of the vote to Clinton's 45.6%. However, Biden won the 2020 election with a slim victory of 49.5% to Trump's 49.3%, which emphasizes Arizona's last two election results with significant republican victory in 2016 and a rather marginal win in 2020.This marked the state's first Democratic victory since 1992. Georgia is a compelling state because most of the democratic vote comes from the Black community,

although Harris has seen relatively lower support among Black men. Given Biden's last narrow win, even a minor shift in voter preferences could significantly impact the 2024 outcome.

As shown in Figure 5, Harris's support rose slightly but plateaued around 47.5% and does not show any signs of increasing. Furthermore, Figure 1 shows that Trump currently holds a slight advantage, suggesting a likely outcome in his favor (270toWin (2024)).

We project Georgia will vote for: **Donald Trump**

4.3 Nevada

Nevada, primarily composed of desert and rural areas, sees its Democratic support concentrated in urban centers like Reno and Las Vegas compared to rural areas. The state has voted Democratic in the last four elections, with a 2% or greater margin in favor of Democratic candidates in the past two. Nevada has been a democratic state for the last 4 elections with at least a 2% difference in favor of the democrats in the last two elections (270toWin (2024)).

As seen in Figure 5, Harris experienced a spike in support in early September, peaking above 48%. Although her support gradually declined, stabilizing around 46%, Figure 1 still shows a marginal lead for Harris over Trump. Given Nevada's recent voting history and current polling, it is probable that Nevada will support Harris (270toWin (2024)).

We project Nevada will vote for: **Kamala Harris**

4.4 North Carolina

North Carolina has been an exclusively Republican state since 1980, with a singular exception in 2008 when the state was democratic and voted for Obama. Thus, North Carolina voted for Trump in the last two elections, where in 2016 there was a more significant win for Republicans with 49.8% of the vote, and 46.2% going to the Democrats. While in 2020, the Republicans maintained their support for the state but the margin of victory was slimmer with them receiving 49.9% of the vote and Biden getting 48.8% (270toWin (2024)).

Looking at Figure 5, Harris had a spike in early September for some states where she was polling around 48%, but afterwards had a minor but gradual decline. However, our model results in Figure 7 suggest a potential increase in Harris's support over time. North Carolina remains challenging to predict due to its Republican history and the neck-and-neck nature of the race shown in Figure 1, leaving the state's outcome highly uncertain.

We project North Carolina will vote for: **Donald Trump**

4.5 Michigan

Harris is currently leading the polls in Michigan with 49% of the vote compared to Trump's 47%. This is a slim margin, but it is the first of the states we've discussed where Harris is leading in the polls. In Figure 4, we see a significant jump in support for Harris, followed by a small dip in October, and then a gradual increase leading up to the election. Looking at the state's voting history in the last two elections, Michigan has aligned with Arizona and Georgia, voting for Biden in 2020 and for Trump in 2016, with vote percentages at 50.6% to 47.8% in 2020 and 47.5% to 47.3% in 2016, respectively. Unlike the previously mentioned states, however, 2016 was the first time since 1992 that Michigan voted for a Republican candidate. Considering Michigan is home to Detroit, the largest Black-majority city in the United States, and has a large population of union workers who often support the Democratic party due to its pro-union stance, it is likely that Harris will win the state (270toWin (2024)).

We project Michigan will vote for: **Kamala Harris**

4.6 Pennsylvania

Pennsylvania is an interesting case for several reasons. First, before Tim Walz was announced as Kamala Harris's running mate, one of the speculated candidates was Josh Shapiro, the current governor of Pennsylvania. Presidential candidates often consider the home states of potential running mates, as they can help with swing states. However, Harris ultimately did not choose Shapiro, so this is no longer a factor. Second, similar to Nevada and Michigan, Pennsylvania voted for Trump in 2016 with a narrow margin of 48.2% to 47.5%, and for Biden in 2020 with a slightly larger but still close victory of 50% to 48.8%. Notably, 2016 was the first time Pennsylvania voted for a Republican candidate since 1988 (270toWin (2024)). Lastly, if we look at Figure 3, we can see that Harris has been polling consistently at around 48% since mid-September, which is somewhat similar to the trend in Georgia. However, Pennsylvania appears to be leaning more Democratic than Republican compared to Georgia. In Figure 5, both states show a similar increase in support for Harris since polling began, reaching a plateau in mid-September at their current polling numbers.

We project Pennsylvania will vote for: **Kamala Harris**

4.7 Wisconsin

The last swing state we will be discussing is Wisconsin, which, as we can see in our prediction table (Table 1), has the highest support for Harris at 48.38% among all the swing states we examined. Looking at Figure 3, we see a small spike in early September for Harris, followed by a gradual stabilization at around 48% support, consistent with our model's projections. Additionally, in Figure 1, Wisconsin shows the largest difference in support for Harris compared to all the previously discussed swing states.

Finally, if we consider Wisconsin's voting history, the only time it voted Republican since 1988 was in 2016, when the state gave its electoral votes to Trump with 47.2% of the vote, compared to Clinton's 46.5%. In 2020, the state returned to blue, with Biden winning. Considering all factors, Harris is likely to win Wisconsin and reach the 270 electoral votes needed to win the election (270toWin (2024)).

We project Wisconsin will vote for: **Kamala Harris**

4.8 Electoral Vote Projection and Predicted Outcome

Without considering the swing states and assuming all other states will vote as expected, Kamala Harris has a slight lead of 226 electoral votes to Donald Trump's 219. Now, if we allocate the 43 combined votes of Arizona, Georgia, and North Carolina to Trump and assign the votes of Nevada, Michigan, Pennsylvania, and Wisconsin to Harris, which total 50 electoral votes, the result is a Democratic victory for Kamala Harris with 276 votes to Donald Trump's 262. We conclude that Kamala Harris is likely to win the upcoming United States election by 14 electoral votes. Below, we have included an image of our projected electoral map for the November 5th election (270toWin (2024)):

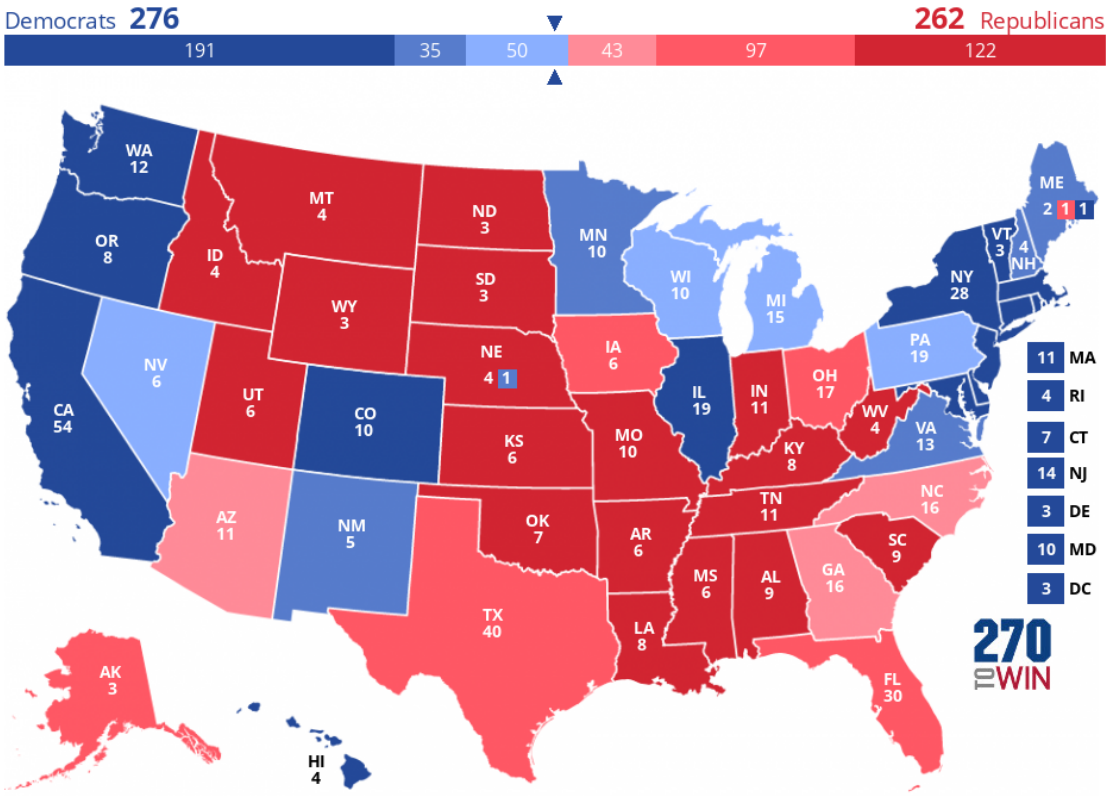

Figure 9: Predicted Electoral Votes for the 2024 US Election

4.9 Limitations and Weaknesses

There are several limitations and weaknesses that could have affected our research and statistical analysis. First, the variable `numeric_grade`, which weighs pollster results for accuracy, could have been better utilized to determine how we evaluate each pollster. By ensuring that we selected the most accurate pollsters available and incorporating this into our model, we might have increased the accuracy of our predictions.

Second, since polling data often changes, especially leading up to an election, it is challenging to make real-time adjustments as new data becomes available. Additionally, because Harris only announced her candidacy at the end of July, we did not have a substantial time frame to analyze her polling data. Typically, presidential candidates start campaigning much earlier, which gives a more comprehensive data set for analysis.

Moreover, since we only created a model for Harris and not for Trump, we had limited comparisons. While we believe our results would not have significantly changed, having a model based on Trump could have provided a more balanced analysis. We also could have benefited from examining recent voting histories in each state for members of Congress, both in the Senate and the House of Representatives, as well as the current political party in power, such as the party affiliation of each state's governor.

4.10 Next steps

As the election quickly approaches, there are not many additional factors we can consider at this stage. However, there are still a few areas for future improvement. First, while we did take into account the historical voting patterns of each state, we did not perform a rigorous analysis of that data. This is something we could explore more deeply in future studies. Additionally, because poll results are inherently uncertain and are not absolute predictors of the election outcome, we could use confidence intervals in the future to better illustrate this uncertainty.

Next, we could incorporate more variables into our analysis to create a stronger and more accurate model, which would help us make more reliable predictions. Lastly, we should consider intangibles more thoroughly, such as recent events both locally and globally, the socioeconomic status of voters, and the voting behavior of different ethnic groups. It would also be beneficial to analyze how population and demographic shifts since the last election might impact voting trends, as well as general public sentiment toward the political parties and their presidential candidates.

A Appendix

A.1 Pollster Methodology

The New York Times/Siena College Poll is a respected polling partnership known for providing data-driven insights into U.S. public opinion on political, social, and economic issues. Recognized for its rigorous methodology, the poll focuses on capturing the views of registered and likely voters across the country, with particular emphasis on key battleground states during election cycles. Frequently cited for its accuracy and transparency, it is widely regarded as a reliable source by the public, political analysts, and the media.

The poll conducts live telephone interviews with a national sample of registered voters, reaching over 90% of respondents via cellphones. The target population consists of U.S. registered voters, with the sampling frame drawn from a comprehensive voter file covering 200 million individuals. This file contains demographic and geographic information to ensure sampling is representative across age, race, political affiliation, and region. The poll places particular focus on swing states like Arizona, Georgia, Michigan, Nevada, Pennsylvania, and Wisconsin, where small shifts in voter opinion can determine the electoral outcome. Respondents are recruited randomly from the voter file to capture a diverse range of political perspectives, demographics, and regional backgrounds. To mitigate non-response bias, which is significant given a response rate of around 2%, the poll applies demographic weighting to underrepresented groups, such as young adults and non-college graduates (Times (2024b)).

The methodology of telephone polling provides benefits, such as enabling random sampling across a broad population. However, low response rates and the shift away from landlines to cellphones mean certain groups, particularly younger adults, may be underrepresented despite demographic weighting adjustments. While the questionnaire is crafted to be clear, neutral, and balanced across political views, the brief format of phone surveys—designed to improve response rates—limits depth. Additionally, phone surveys may not reach segments of the population more responsive to digital surveys or in-person interviews, potentially missing out on specific voter perspectives. Each respondent's answer is then adjusted for weighting to ensure proportionality, and results are aggregated to reflect a balanced, representative outcome of public opinion, rather than raw, unadjusted numbers.

Overall, while the New York Times/Siena College Poll is highly regarded for its thorough approach, methodological constraints, such as low response rates and limited outreach to digitally-focused populations, introduce challenges to representativeness.

A.2 Survey

Here we will be talking about how we would conduct a $100,000 surevy to gather data about this upcoming election to ensure there is minimal error in the data we are collecting and, our data represents the facts we want them to represent. Out of the $100,000 we will allocate $15000 for the necessary development and administration of the data. This means we will be spending this money to create the survey, and have a infrastructure in place to hold the large amount of data as well as cover any security fees to keep this data private and safe. Then we allocate $50000 to advertise the survey. We will need a large sample so we will be spending most of our budget for this. We will have a different urls for each advertisement we have so that we can incorporate this in our data to see which demographic is accessing the survey through which platform. The platforms to advertise this is through spotify, facebook, instagram, and some news networks. If possible we will try to use this money to get it endorsed by branches of government to show our reliability. Reliability is also what citizens respond to so this will get a lot of respondents. We will spend $25,000 for modeling our data as they tend to be expensive with data this size. And the rest of the $15,000 will be leftover cost for anything that we don't foresee. If there is any leftover we can add a survey participation price such as having a free 3 month trial for the platforms we mentioned above (that is if they have a subscription based membership).

We will not be using telephone surveys because according to research (Survey 2024) we find the most people, don't pick up calls from unknown numbers and even the people that pick it up they are less inclined to

answer the questions of the survey.

Next to look at the actual contents of the survey. They can be accessed through this link: Sample survey question. There are 3 things that we were careful of when we were creating this survey. First thing we considered was transperency. People need the reassurance that the data that is being collected will not be used against them, and so we feel them what data we do collect and how the data we collect cannot be used to identify a person. The second thing we focus on is readability of the questions. We tried to make them as simple as possible using accessibile language, and tried to keep it short as well (Tourangeau, Rips, and Rasinski (2000)). We also prioritised the size of the survey. We kept it to a short 11 questions that will tell us their political standing in the past and present. We know what current issues are important to them as well as their age group and the state they are from. This will help us gather data without inconveniencing the person. There are disadvanatages to the way we structured the survey. For example, we were keeping time and security in mind, so we restricted the survey to 11 questions, however, this limits the amount of answers we get from the reader. We may be missing some crucial information that helps us figure out which group of people may vote for whom, such as ethnicity and wages (Stantcheva (2023)).

Finally we take a look at how we sample from the gathered data. We thought about random sampling but we were worried about having unequal proportions of the demographic we would advertise to. So if more people that go to facebook responded, that doesn't necessarily mean that people use facebook more than the other platforms. So we we decided to combine stratified and cluster sampling(Stantcheva (2023)). The idea is that we will choose based on the different platforms we advertise to first to have a cluster sampling method. And then we will sample the data by dividing the people by the certain aspects of the surevy, like if they are in different age groups or if they are registered voters. This is called Stratified Sampling (Stantcheva (2023)). Then we take a union of all the samples and we gather a group that we are able to build models based off of.

There are obvious disadvantages as well. For example, we have fewer questions meaning when we sample based on the question we might have cases where, some groups are over or under represented. This can happen for example when most of the people who answered might be from age 18-50 because those are the age groups with access to a phone and are frequently visiting the sites we advertise our survey on. Also the different apps may skew the data as Facebook has a older audience as opposed to Instagram and Tiktok. Another disadvantage is the fact that we have 2 different sampling methods that we use and we combine them both, meaning that this can be an expensive process. Doing so may exceed the allocated budget however, it will be worth it because we minimize the sampling bias we talk about earlier where we may have a group that has more access to the survey. The last disadvantage we also foresee is the fact that younger audience may not be as interested in filling out a survey about politics as they feel, it might not directly affect them. We try to minimize this by offering compensation for their time when they fill out the survey, i.e. free membership services for 3 months on Spotify.

Although there are both advantages and disadvantages, we feel that this survey is well though out and doesn't allow for ambiguity in the answers of the participants.

A.3 Posterior Predictive Check

In Figure 10, we check the Posterior predictive check of model of our model. Since the raw data aligns closely with our predictions, our model is a good prediction of the data.

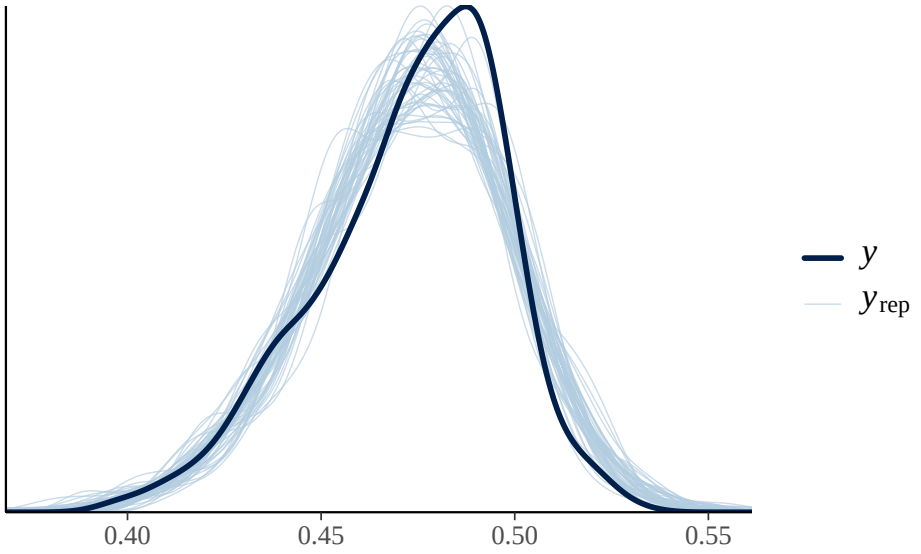

Figure 10: Posterior Predictive Check of Model

References

270toWin. 2024. "Historical Presidential Election Information by State." https://www.270towin.com/states/.

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

FiveThirtyEight. 2024. "2024 National Presidential Polls." https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

Gabry, Jonah, and Tristan Mahr. 2023. *Bayesplot: Plotting for Bayesian Models*. https://mc-stan.org/bayesplot/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

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

Stantcheva, Stefanie. 2023. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." Journal Article. *Annual Review of Economics* 15 (Volume 15, 2023): 205–34. https://doi.org/https://doi.org/10.1146/annurev-economics-091622-010157.

Survey, ID. 2024. "Phone Survey." https://www.idsurvey.com/en/phone-survey/.

Times, New York. 2024a. "Tracking the Swing States for Harris and Trump." https://www.nytimes.com/interactive/2024/us/elections/presidential-election-swing-states.html.

———. 2024b. "You Ask, We Answer: How the Times/Siena Poll Is Conducted." https://www.nytimes.com/article/times-siena-poll-methodology.html.

Tourangeau, Roger, Lance J. Rips, and Kenneth Rasinski. 2000. "Respondents' Understanding of Survey Questions." In *The Psychology of Survey Response*, 23–61. Cambridge University Press.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.

org/knitr/.