My title*

My subtitle if needed

First author Another author

November 4, 2024

Abstract

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

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 forces shaping the 2024 presidential election.

Estimand paragraph

Results paragraph

Why it matters paragraph

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

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.

¿ does this go into data cleaning ?

We use the statistical programming language R (R Core Team 2023a). Our data comes from (FiveThirtyEight 2024) and was cleaned, modeled and graphed, using Robinson, Hayes, and Couch (2023), Wickham et al. (2019), Goodrich et al. (2022), Firke (2023), Grolemund and Wickham (2011), Robinson, Hayes, and Couch (2023), Arel-Bundock (2022), R Core Team (2023b).

2.2 Methodology

This paper uses 'polls-of-polls' method to analyze and predict our outcome, which combines polls from multiple sources (pollsters). By combining results from various sources, this approach incorporates diverse perspectives, which helps to minimize individual biases and provides a more balanced view. Unlike relying on a single poll, which can be influenced by its own biases, the 'polls-of-polls' method enhances reliability and increases the overall validity of the results.

… explain how pollsters survey people, what their different methodologies are, and why these things are important for an accurate prediction…

2.3 Measurement/ 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 [CITE NYT PAPER]. 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, 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.

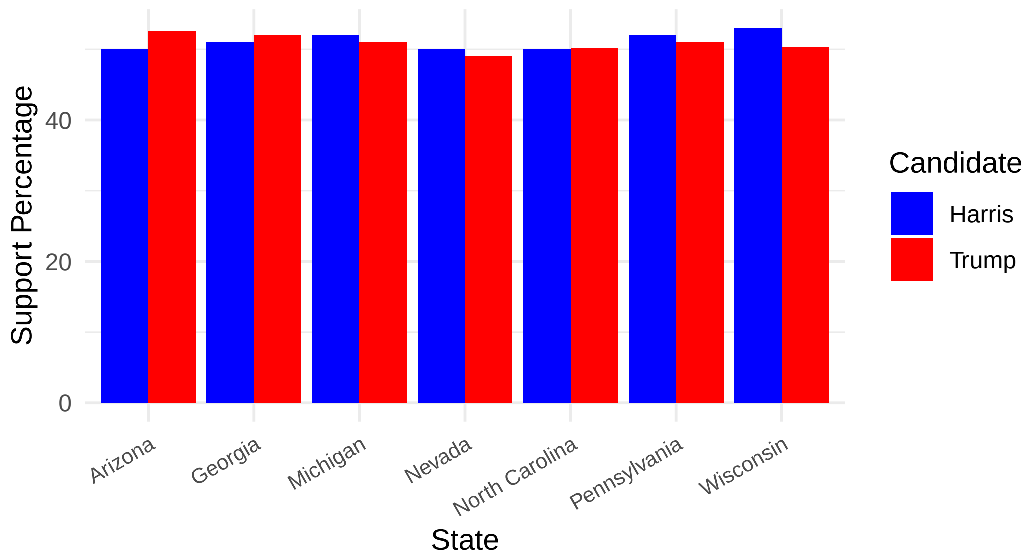


Figure 1: Distribution of Support Percentage by State

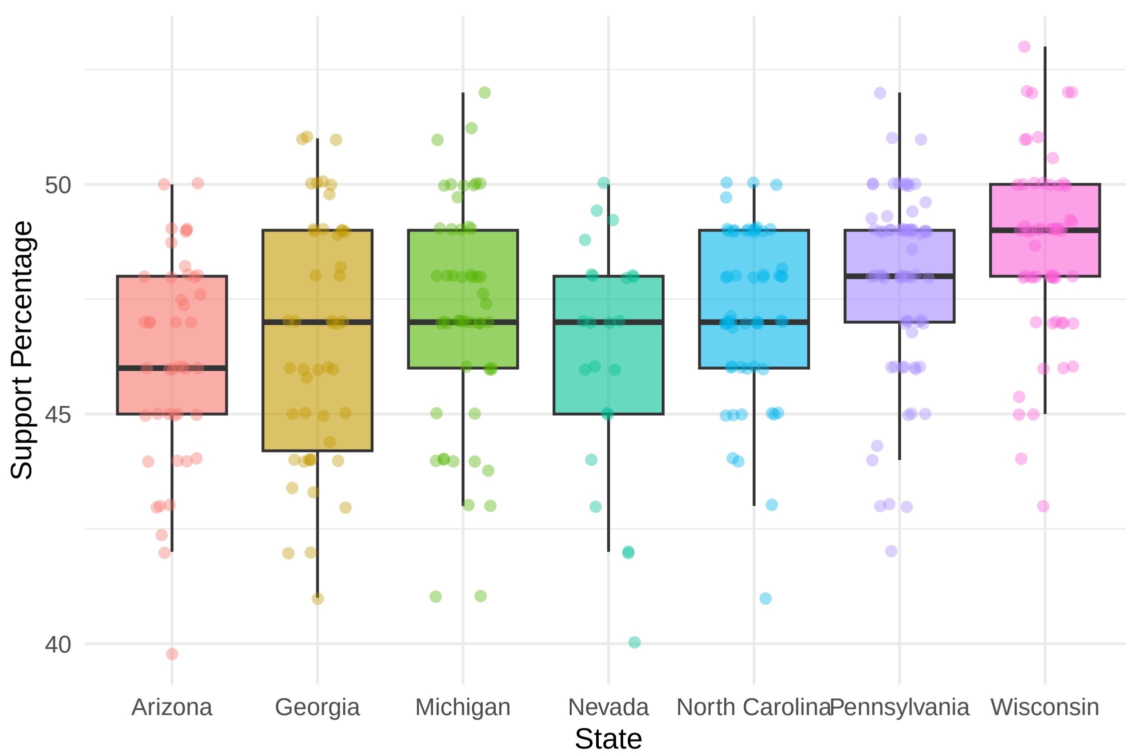


Figure 2: Support Percentage for Harris by State

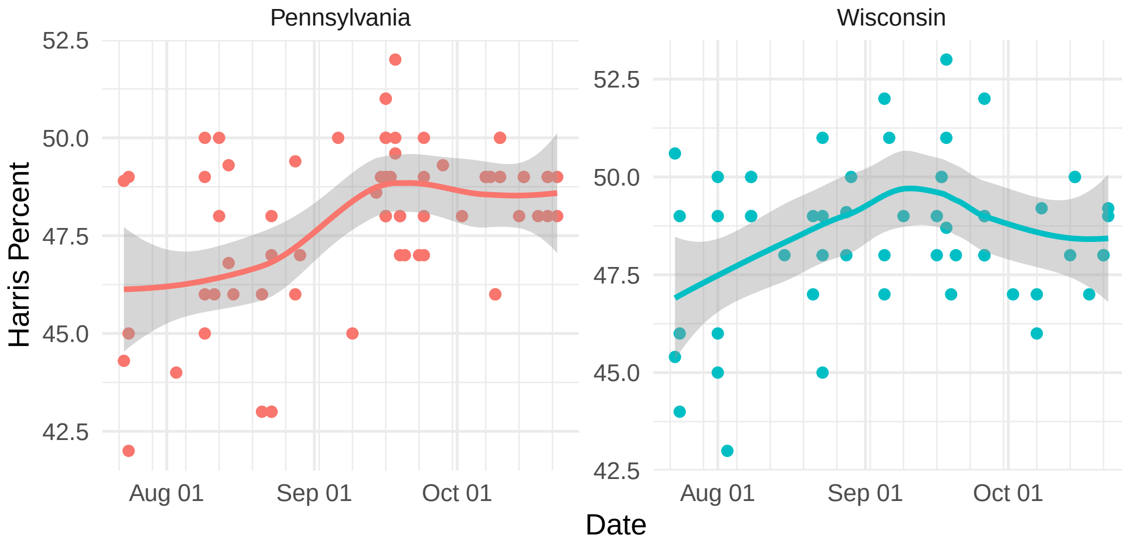


Figure 3: Support Percentage for Harris by State since Harris declared as Candidate for Wisconsin and Pennsylvania

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 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.

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.

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.

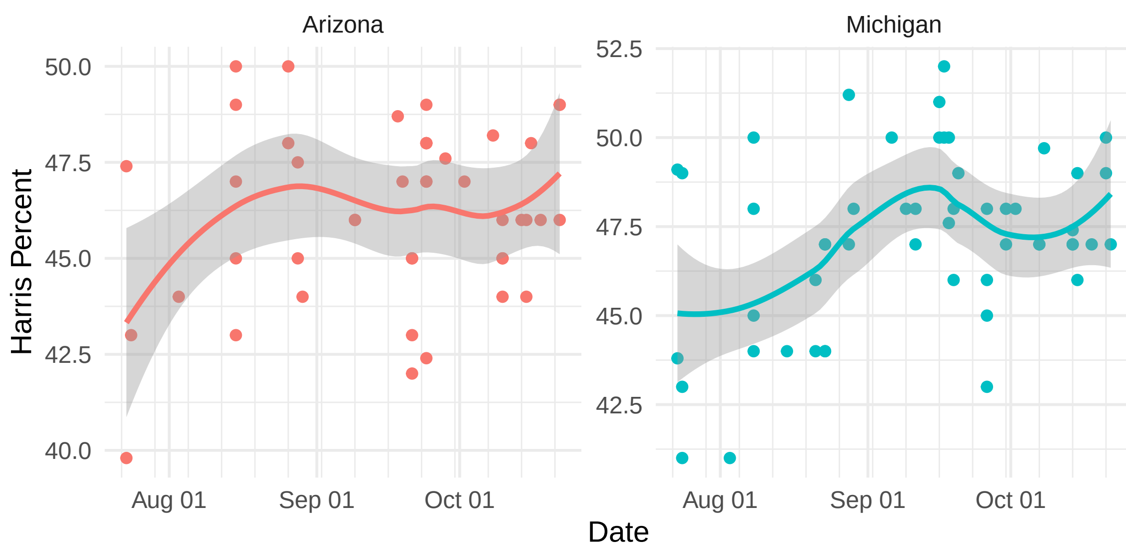


Figure 4: Support Percentage for Harris by State since Harris declared as Candidate

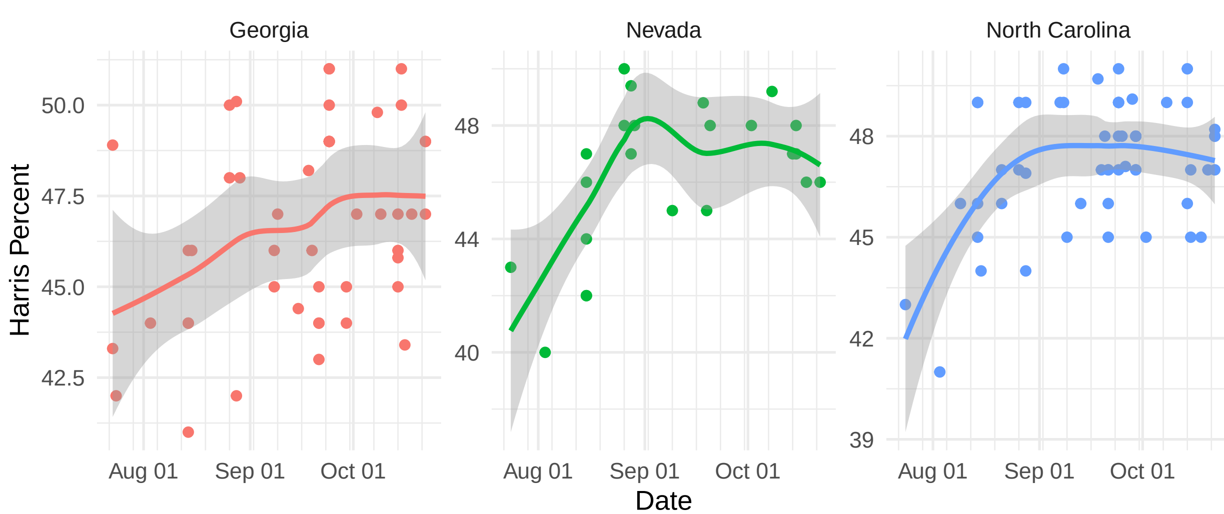


Figure 5: Support Percentage for Harris by State since Harris declared as Candidate

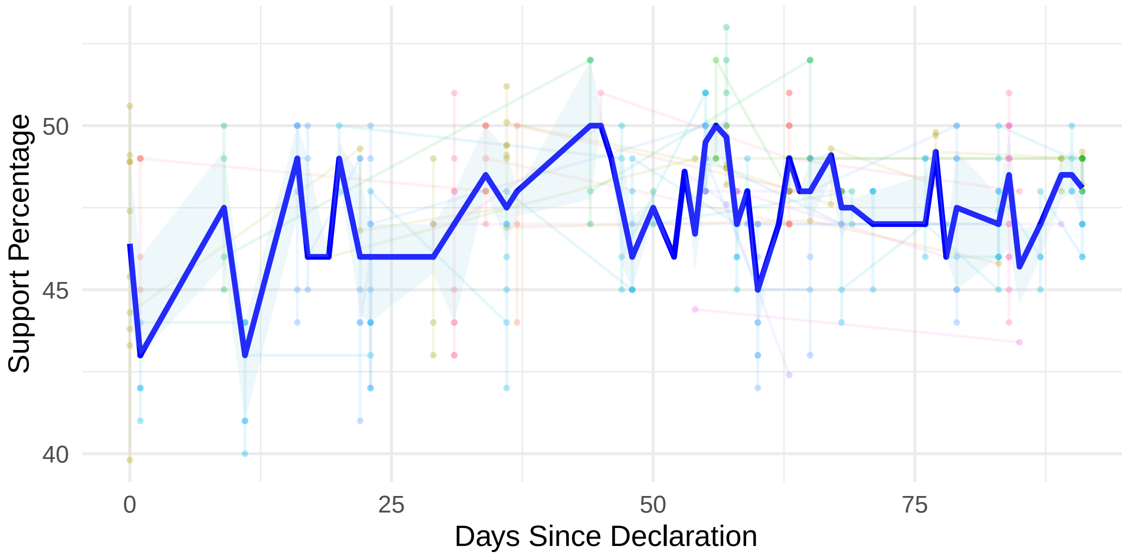


Figure 6: Variation in Harris's Support Percentage Across Pollsters Over Time

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.

To introduce the outcome and predictor variables effectively, you'll want to set the stage for your model by clearly defining what you are trying to predict (the outcome variable) and the factors that you believe influence it (the predictor variables). This section should provide context on why each variable is included and how it might impact the results, without yet delving into the specific mechanics of the model itself.

Here's an example of how to structure this paragraph:

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

We define our model as:

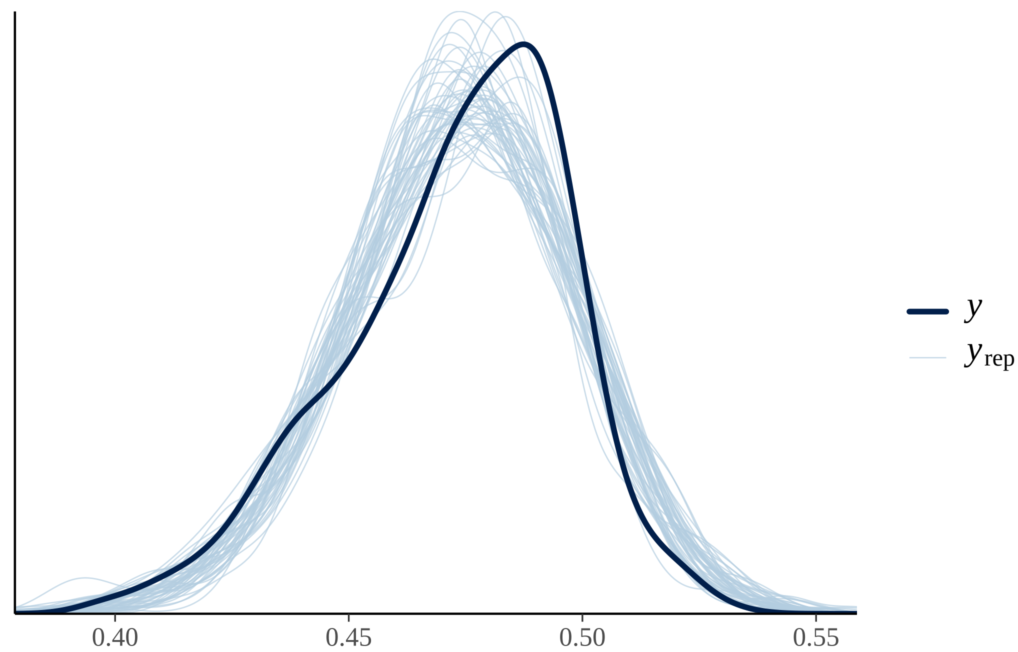
$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (1)\\
\mu_i &= \beta_0 + \beta_1 \times \text{State}_i + \beta_2 \times \text{Pollster}_i + \beta_3 \times \text{Days since first poll}_i && (2)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5)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

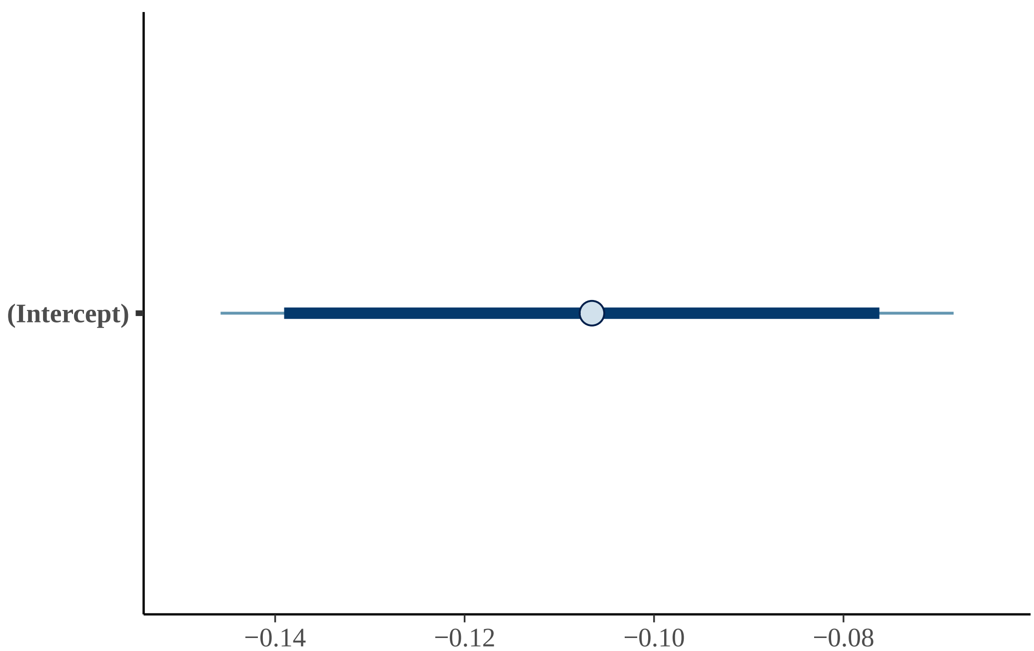
We used the `stan_glmer` function from `rstanarm` package to create a Bayesian regression model with a Normal distribution. The dependent variable in the proportion of respondent who support Kamala Harris, and our model aims to predict Harris' support based on several important factors, modeled by:

where:

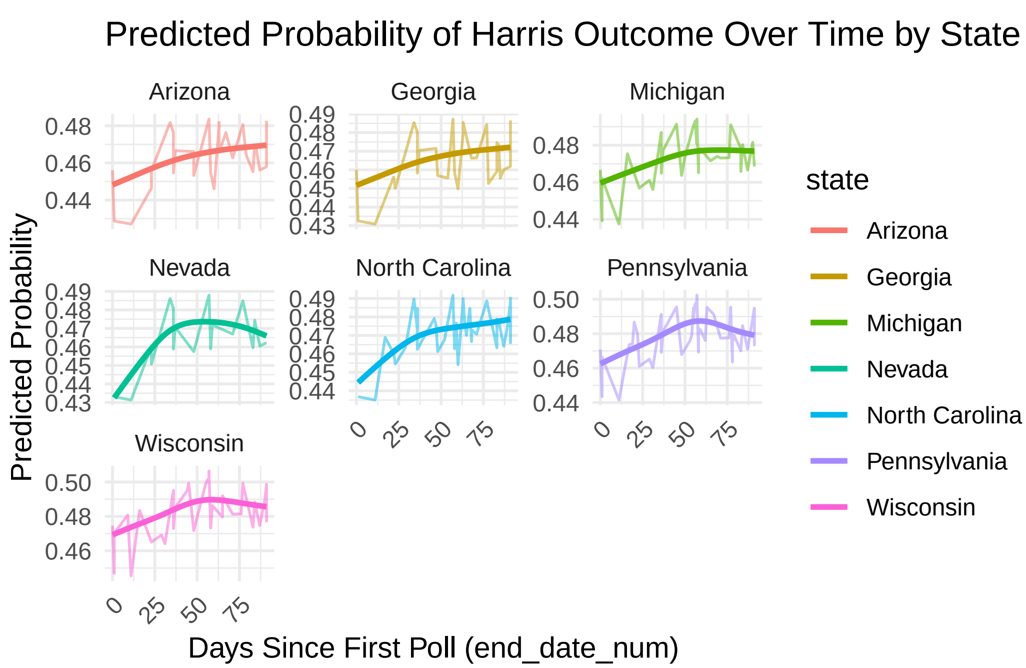
- y_i is the dependent variable, represneting the proportion of respondents who support Harris
- β_0 is the intercept term, representing the expected proportion of y_i when all other predictors are set to zero.
- β_1 corresponds to the `state` choosing one of seven states,
- β_2 corresponds to the `pollster`,
- β_3 is the value representing number of days since the first poll.

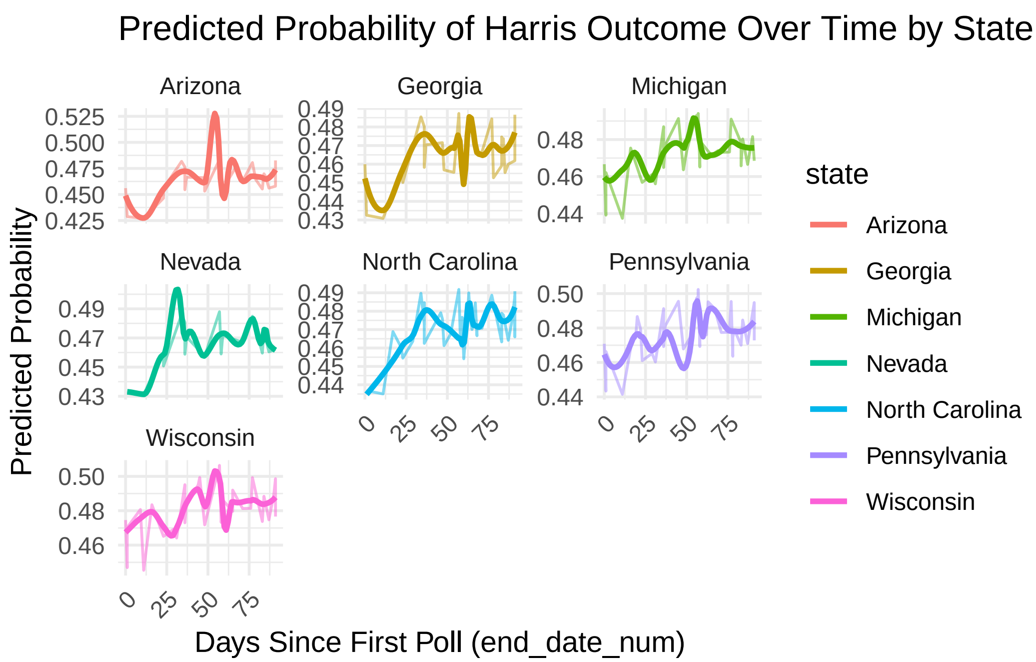
All variables follow the normal distribution with a mean of 0 and a standard deviation of 2.5.


y
y rep
0.40
0.45
0.50
0.55


(Intercept)
−0.14
−0.12
−0.10
−0.08

	First model
(Intercept)	-0.11
	(0.02)
Sigma[end_date_num × (Intercept),(Intercept)]	0.00
	(0.00)
Sigma[pollster × (Intercept),(Intercept)]	0.00
	(0.00)
Sigma[state × (Intercept),(Intercept)]	0.00
	(0.00)
Num.Obs.	345
ICC	0.6
Log.Lik.	-1370.772
ELPD	-1414.7
ELPD s.e.	12.1
LOOIC	2829.4
LOOIC s.e.	24.3
WAIC	2827.1
RMSE	0.02






```
expected_probs <- posterior_epred(bayesian_model, newdata = harrisdata)

# Calculate the mean predicted probability across all observations (overall predicted support percentag
overall_predicted_support <- mean(rowMeans(expected_probs)) * 100
overall_predicted_support
```

```
[1] 47.34767
```

```
expected_probs <- posterior_epred(bayesian_model, newdata = harrisdata)

# Calculate the median predicted probability for each observation
median_probs <- apply(expected_probs, 2, median)

# Add median predicted probabilities to `harrisdata`
harrisdata <- harrisdata %>%
  mutate(median_prob = median_probs)

# Calculate the average predicted support by state
predicted_support_table <- harrisdata %>%
  group_by(state) %>%
  summarize(average_predicted_support = round(mean(median_prob) * 100,2))

# Print the table
predicted_support_table
```

```
# A tibble: 7 x 2
  state          average_predicted_support
  <fct>                              <dbl>
1 Arizona                             46.4
2 Georgia                             46.7
3 Michigan                            47.3
```

```
4 Nevada                             46.5
5 North Carolina                     47.2
6 Pennsylvania                       48.0
7 Wisconsin                          48.4
```

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

3.1 Model set-up

We run the model in R (R Core Team 2023a) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

What do we expect… We expect a predictions in the high 40% based on our current data.

4 Results

Our results are summarized in **?@tbl-modelresults**.

5 Discussion

5.1 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 **?@fig-pollingharris** 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.

We believe Arizona will vote: **Donald Trump**

5.2 Georgia

Considering the last two elections Georgia is very similar to Arizona, as in 2016 Trump won the state with 50.8% to Clinton's 45.6%. While in the 2020 election Biden won with a slim victory of 49.5% to Trump's 49.3%, which again is very similar to Arizona's last two election results with significant republican victory in 2016 and a rather marginal win in 2020, which was the first democratic victory for the state since 1992. Georgia is in an interesting situation as lot of the democratic vote in the state comes from the black community, although Harris a has less support from male black population, as the last victory for Biden was as a small margin, even a small percent of the population switching their vote could heavily impact the vote in 2024.

In **?@fig-pollingharris** at the beginning of the polling data the support for Harris was relatively low, while it did start to grow the polling percentage somewhat plateaued at around 47.5% and does not show any signs of growing. Furthermore, in Figure 1 we can see that Trump has the edge over Harris considering all of these factors there is a very real possibility that Trump does win Georgia at the end of the day.

We believe Georgia will vote: **Donald Trump**

5.3 Michigan

Harris is currently leading the polls in Michigan with 49% of the vote and Trump with 47%, this is a rather slim margin, but the first in the discussed states where Harris is leading in the polls. In **?@fig-pollingharris** we can see that there was a significant jump for Harris and a small dip in October followed by a gradual increase leading up to the elections. When we look at the voting history of the state in the last two elections, the have voted for the same candidates as Arizona and Georgia, where they voted for Biden in 2020 and for Trump in 2016, with the voting percentages at 50.6% to 47.8% in 2020 and 47.5% to 47.3 in 2016 with respect to the winner. Although unlike the previously mentioned states 2016 was the first time the state voted for the republican candidate since 1992. Considering Michigan is home to the largest black majority city in the United states, Detroit, and a large majority of union workers who often vote democratic as it is seen as the more union-friendly party between the two it is rather like that Harris will win the state.

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

5.4 Nevada

5.5 North Carolina

5.6 Michigan

5.7 Pennsylvania

5.8 Wisconsin

5.9 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

Why we choose NYT? - `numeric_grade` is 3 - `pollscore` is -1.5, a score of reliability called "Predictive Optimization of Latent skill Level in Surveys, Considering Overall Record, Empirically.", where negative numbers are better - `transparency_score` is 9, reflects pollsters transparency about their methodology (calculated based on how much information it discloses about its polls and weighted by recency) - `population_full` is 'rv', respondents are registered voters

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...

```
pp_check(first_model) +
  theme_classic() +
  theme(legend.position = "bottom")

posterior_vs_prior(first_model) +
  theme_minimal() +
  scale_color_brewer(palette = "Set1") +
  theme(legend.position = "bottom") +
  coord_flip()
```

Figure 7: Examining how the model fits, and is affected by, the data

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows... This suggests...

?@fig-stanareyouokay-2 is a Rhat plot. It shows... This suggests...

Checking the convergence of the MCMC algorithm

C Appendix 1

D Appendix 2

Here we will be talking about how we would conduct a $100,000 surevey 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 respondants. 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 reserach (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 reassurence that the data that is being collected will not be used against them, and so we teel 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.

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 necessairly mean that people use facebook more than the other platforms. So we we decided to combine startified 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.

References

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/.

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

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

Robinson, David, Alex Hayes, and Simon Couch. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

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/.

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.