My title*

My subtitle if needed

First author Another author

November 3, 2024

Abstract

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

Table of contents

1 Introduction **1**

2 Data **1**

2.1 Overview . . . 1

2.2 Methodology . . . 2

2.3 Measurement/ Data Visualization . . . 2

2.4 Outcome variables . . . 4

2.5 Predictor variables . . . 5

3 Model **5**

3.1 Model set-up . . . 9

3.1.1 Model justification . . . 9

4 Results **9**

5 Discussion **9**

5.1 Arizona . . . 9

5.2 Georgia . . . 9

5.3 Michigan . . . 9

5.4 Nevada . . . 9

5.5 North Carolina . . . 9

5.6 Michigan . . . 9

5.7 Pennsylvania . . . 9

5.8 Wisconsin . . . 9

5.9 Weaknesses and next steps . . . 9

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

Appendix **10**

A Additional data details **10**

B Model details **10**

B.1 Posterior predictive check 10

B.2 Diagnostics 10

C Appendix 1 **11**

D Appendix 2 **11**

References **12**

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 polls 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.
- `start_date`: The date the pollster began conducting the poll.
- `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.

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

We focus on the swing states of the U.S, according to [cite NYT PAPER]

These are the statistics of each swing state by percetnage by state.

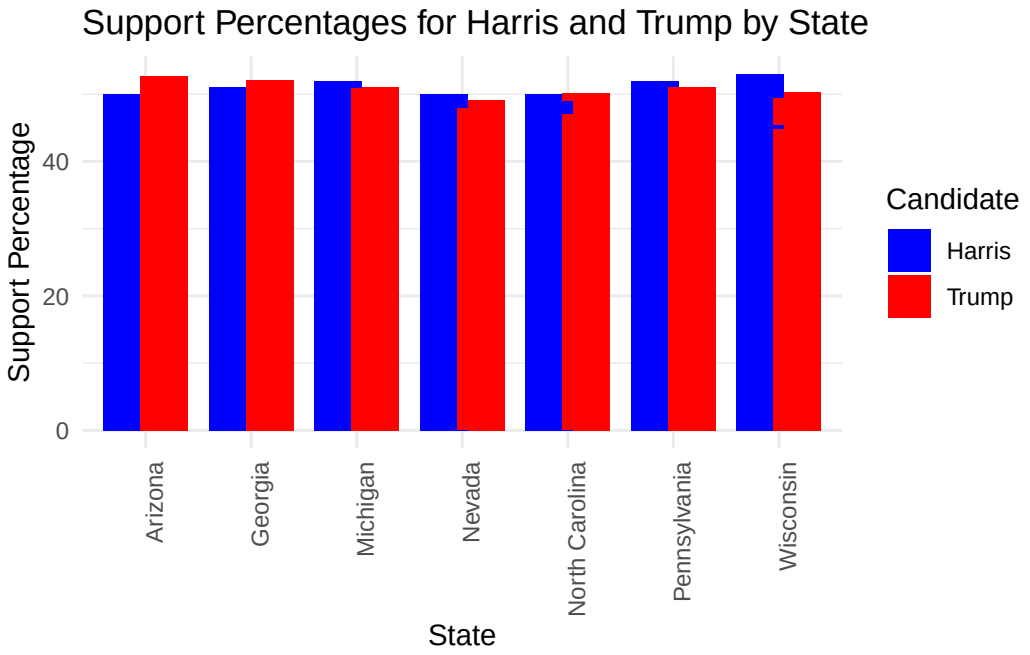


Figure 1: Distribution of Support Percentage by State

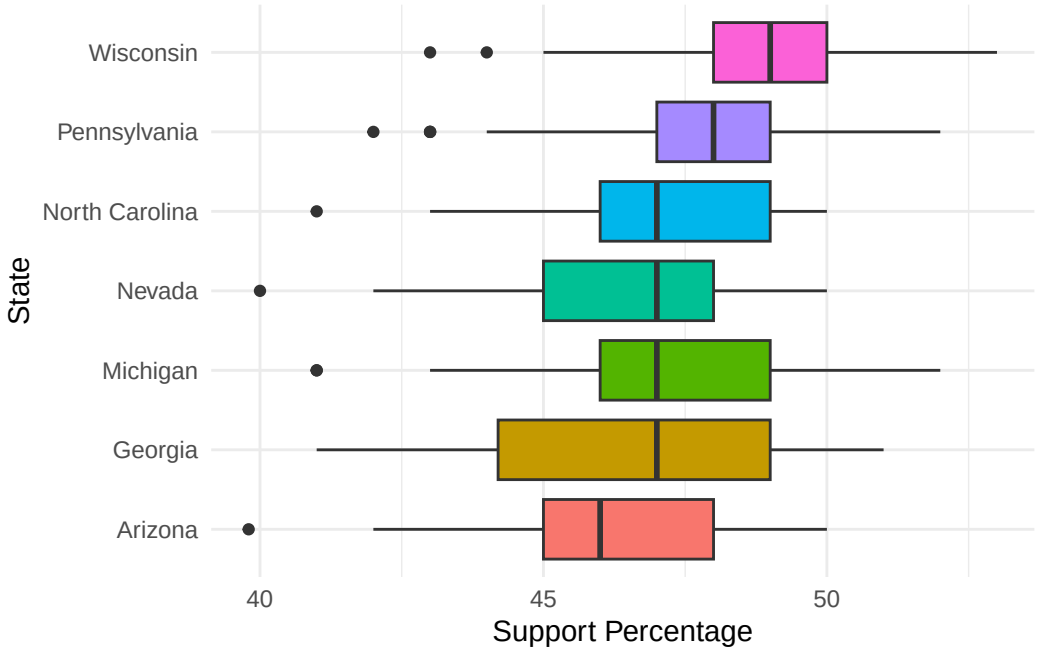
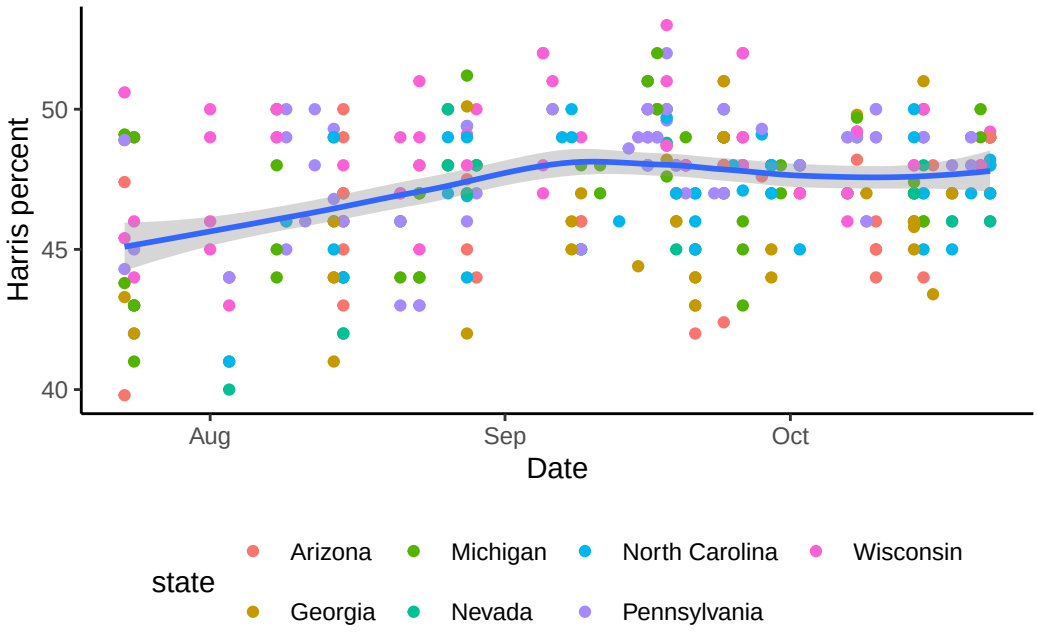
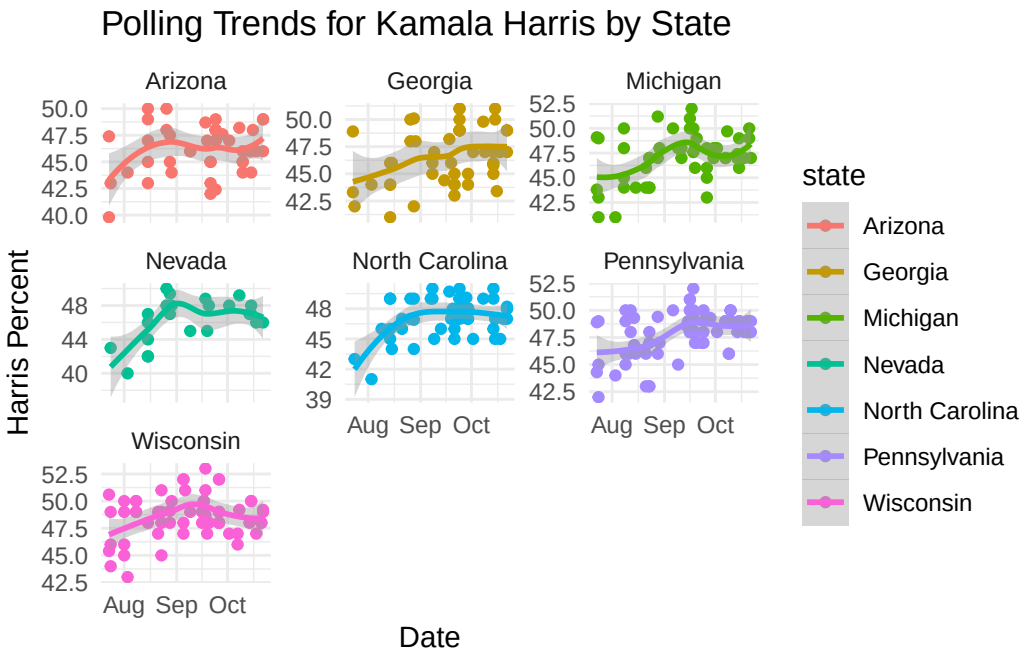


Figure 2: Distribution of Support Percentage by State





2.4 Outcome variables

Add graphs, tables and text. Use sub-sub-headings for each outcome variable or update the subheading to be singular.

Talk more about it.

And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

Talk way more about it.

2.5 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

We define our model as:

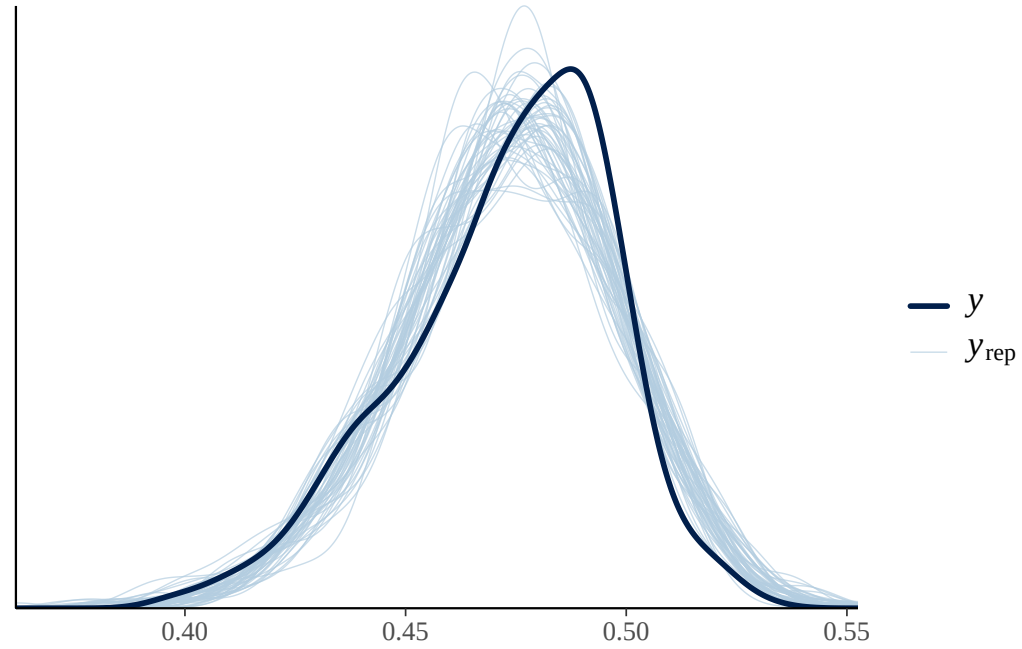
$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \beta_0 + \beta_1 \times \text{State}_i + \beta_2 \times \text{Pollster}_i + \beta_3 \times \text{Days since first poll}_i \tag{2}$$
$$\beta_0 \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta_1 \sim \text{Normal}(0, 2.5) \tag{4}$$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

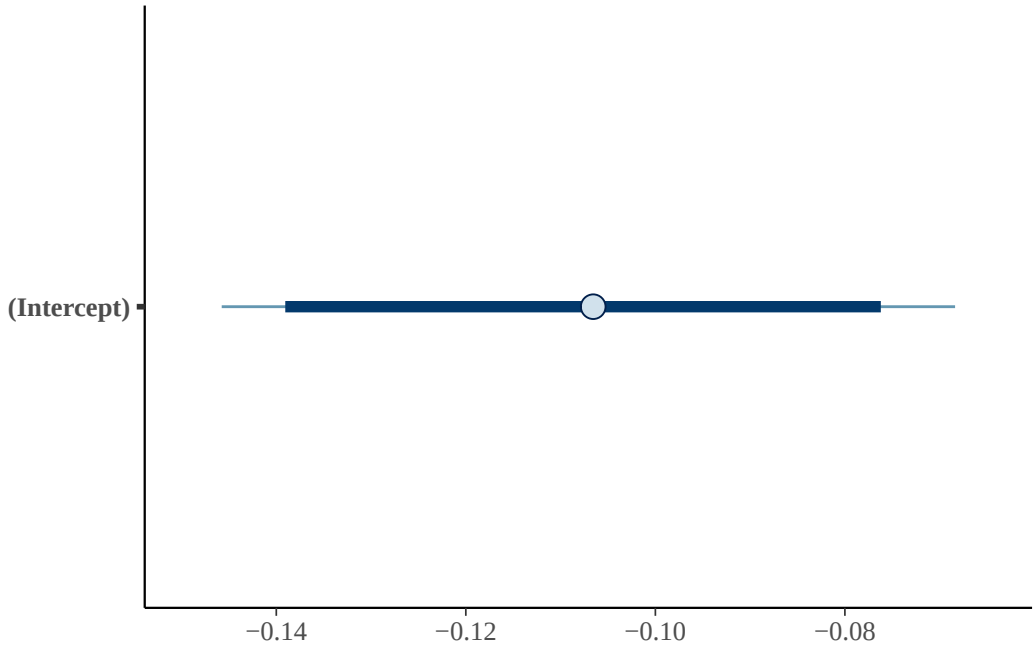
We used the `stan_glmer` function from `rstanarm` package to create a Bayesian regression model with a Normal distribution. The dependent variable in the proportion of respondent who support Kamala Harris, and our model aims to predict Harris' support based on several importnat factors, modeled by:

where:

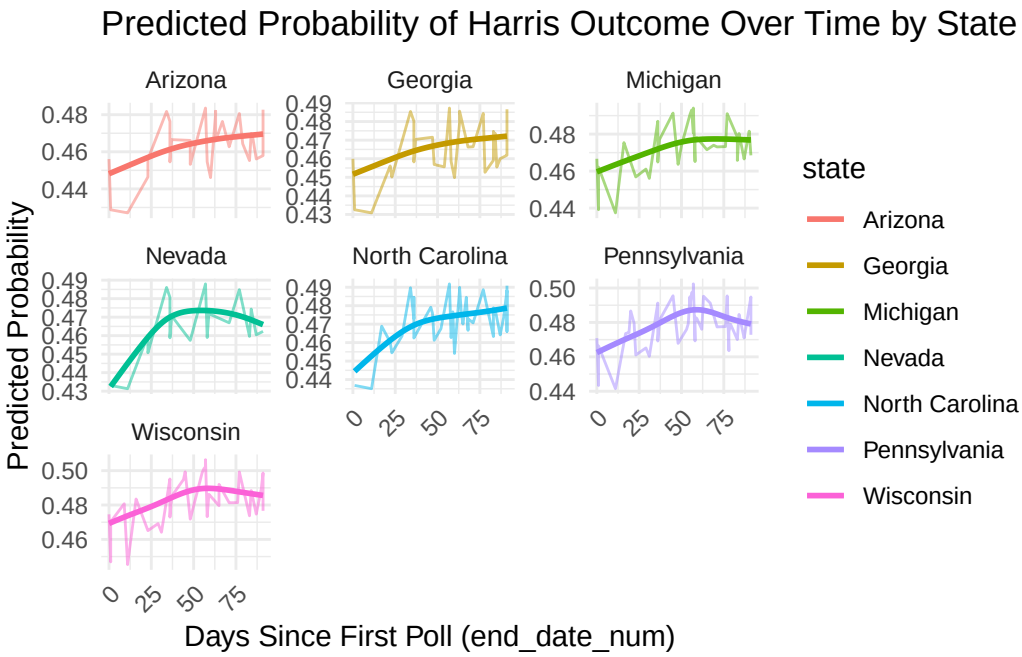
- y_i is the dependent variable, represneting the proportion of respondents who support Harris
- β_0 is the intercept term, representing the expected proportion of y_i when all other predictors are set to zero.
- β_1 corresponds to the `state` choosing one of seven states,
- β_2 corresponds to the `pollster`,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

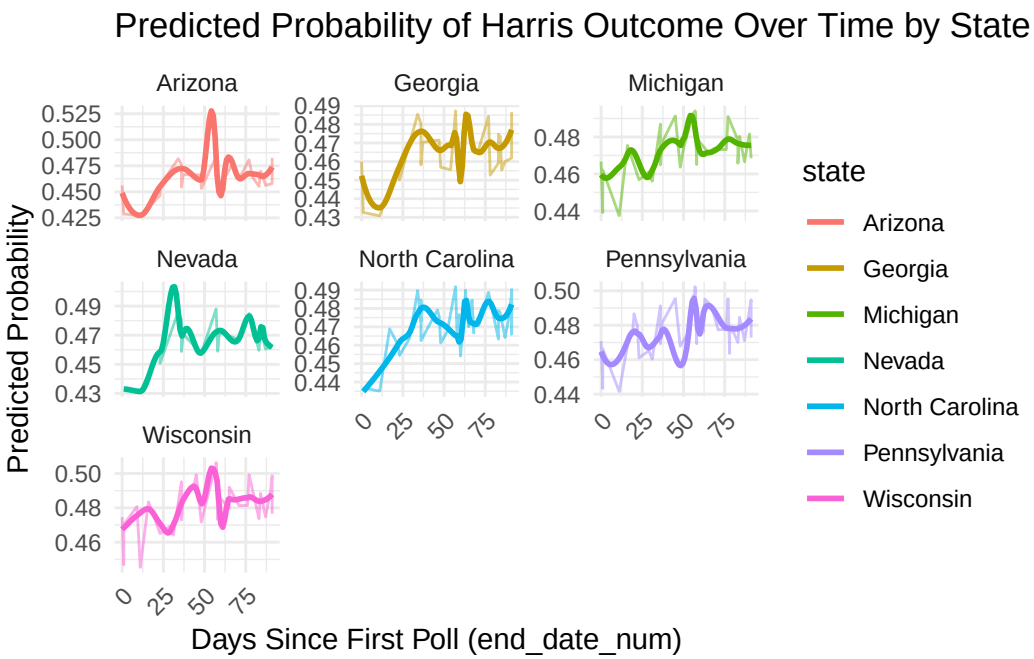
All variables follow the normal distribution with a mean of 0 and a standard deviation of 2.5.


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 percentage)
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 = mean(median_prob) * 100)

# 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                         48.3
```

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

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

5.2 Georgia

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Michigan

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 3: 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

======= »»»> eda40e1517683f0ad12e0bcd901797a2bb2abd7d

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.

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.