Forecasting the United States 2024 Presidential Election*

Kamala Harris wins 60th presidency with 279 Electoral College Votes

Robert Ford Michelle Ji Cher Ning-Li

November 4, 2024

This report analyzes compiled polling surveys for the 2024 U.S. presidential election using FiveThirtyEight data (FiveThirtyEight 2024b) and the 'pool of polls' method. From the analysis, bayesian and generalized linear models were built to forecast the outcome of the presidential election based on popular vote and the electoral college system. The models forecast Kamala Harris as the winner of election, winning 279 electoral college votes versus the 258 votes Donald Trump acquires. Election forecasting is crucial for informing voters, guiding campaign strategies, and engaging the public in the electoral process. Additionally, it helps predict potential policy shifts and holds political institutions accountable, contributing to a more transparent and informed democracy.

1 Introduction

Every four years, the United States presidential election occurs and is one of the most significant political events internationally. The complex process of the election involves both the popular and electoral college system. The candidate that receives the popular vote by the voters wins the electoral votes for that state (Weber 2024). After the popular vote is accounted for, the candidate who receives the majority of the electoral votes from all states wins the presidential election. The number of electoral votes varies by state and is proportional to each state's respective population (Weber 2024).

Specifically for the 2024 presidential election, the top candidates include Donald Trump and Kamala Harris, both with differing opinions on key topics such as immigration, technology, abortion, and transatlantic affairs. For example, regarding technology, both parties are concerned with the rapid progression of artificial intelligence's capabilities but differ in the response and regulation of it. Republicans believe that the current practices are unfair and disproportionately target their voices, whereas democrats generally are in favor of tighter regulation (NBC 2024). Another key issue important to voters and the candidates is abortion. As former president, he supported restrictions on abortion access and advocated for policies that align with the conservative movement's opposition to abortion rights (McKenzie Beard and Abbie Cheeseman and Justine McDaniel 2024). On the other hand, Harris has been a key supporter in the pro-choice movement and has been an advocate for protecting and expanding access to abortion services, as well as female autonomy (Chad de Guzman and Koh Ewe 2024). The outcome of this election will have profound consequences for various legislative policies and the future direction of the United States.

*Code and data are available at: https://github.com/Ford-Robert/us-presidential-election.

To forecast the 2024 United States presidential election, a generalized linear model and Bayesian model were built. These models will help us predict who will win the election based on the electoral college system. A key finding from our model is that we predict Kamala Harris to win the presidential election with 265 electoral college votes. This would be a historic win, as she would be the first woman to become US president, a significant milestone in American politics. One drawback of our model stems from representativeness of the dataset, as it does not include all 50 states in the polls, which may impact the electoral college votes.

This paper is broken down into various sections, including Data, Modeling, Results, Discussion, Conclusion, and Appendices. This paper uses data made available by FiveThirtyEight (FiveThirtyEight 2024b) which compiles individual polling surveys based on state and methodology. Section 2 explores the data, highlighting key aspects that may be useful to future policymakers or campaign strategists, as well as details the variables present in the dataset. Section 3 presents the models that were built and used to forecast the election. Section 4 details the conclusions of our model and Section 5 explores possible implications and insights of our conclusions. Section A include an idealized methodology and survey that we could hypothetically run, with the task of forecasting the US presidential election. This section also includes an in-depth analysis of one specific pollster's methodology found within the larger dataset.

2 Data

The dataset used was obtained through FiveThirtyEight Interactives (FiveThirtyEight 2024b) and is focused on the United States presidential general polls for 2024. The polling data extracted utilize the "pool of polls" method, which combines multiple polling results into a single estimate. By aggregating data from several polls, this method aims to create a more stable and reliable measure of public opinion by reducing the impact of outliers, sampling error, and individual poll biases. This dataset is crucial to help us build a model to forecast the 2024 US Presidential election and can help us analyze key features in voter habits for future political analysis. Note that polling data collection stopped on October 19. Our model's results are based solely on data available as of that date and do not incorporate any subsequent polling updates.

There were two other potential datasets that could have been used, titled "Presidential Polling Averages" and "President Primary Polls", both found through FiveThirtyEight (FiveThirtyEight 2024a). The "Presidential Polling Averages" dataset only included data from the Biden versus Trump election in 2020 and was not updated to include data for the upcoming 2024 election and therefore inadequate. The "Presidential Primary Polls" dataset included very similar information as our chosen dataset, however, it included candidates that have since dropped out or were not major running candidates, such as Nikki Haley and Michelle Obama. It includes observations that are unnecessary and creates noise, therefore we did not choose this dataset either.

The variables used in the dataset include pollster name, pollster weight, method, state, sample size, pollscore, collection/end date, transparency score, candidate, support percentage, election date, days to election, initial weight, weight, and poll region. Pollster Name indicates which pollster the individual poll came from and the weight is assigned by FiveThirtyEight depending on sample size and if they have multiple polls out in a short time to adjust for any bias. Method is the way the poll was deployed, including but not limited to live phone, online panel, and text-to-web. State indicates what state the poll was conducted and sample size is the size of the poll. The pollscore is the rating FiveThirtyEight gives each indidivual poll, on factors such on ethical and methodological elements.The candidate variable indicates the candidate a certain percentage of people in the poll voted for and "support percentage" represents that specific number. The election date and days to election variables indicate how far out the poll was conducted with respect with the election date of November 5, 2024. We also reorganized the dataset to include a poll region, which divides the US into geographical regions such as rust belt, northeast, etc. The table in Section A.3 details the first few observations of the cleaned dataset, which we used to build our model.

For further analysis, the cleaned dataset was separated into two different datasets by beach name, for simplicity purposes. Data was analyzed through the R programming software (R Core Team 2023) and packages such as tidyverse (Wickham et al. 2019), ggplot2 (Wickham 2016), knitr (Xie 2014), and dplyr (Wickham et al. 2023) were used to help download, clean, simulate, analyze, and test the data.

Some pollsters are included in the dataset more than others based on the frequency of their conducted poll surveys and if they align with FiveThirtyEight's guidelines, which can be seen in Figure 1. Figure 1 highlights the distribution of the top 25 pollsters and how frequently they appear in the dataset, with Morning Consult being the most popular pollster with almost 500 appearances. Noting which pollsters are more trustworthy and provide reliable data is important for evaluating the quality of future analyses.

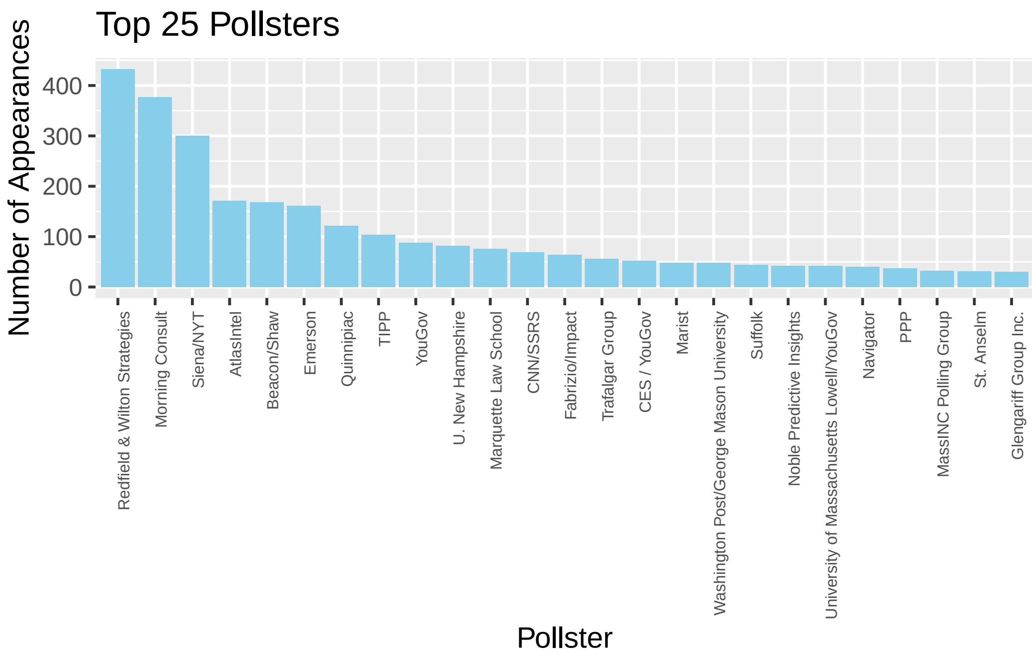

Figure 1: Top 25 pollsters in overall polling data

Figure 2 details the frequency of various polling methods used in the individual polling data, showing online panels are the most popular methodology used to deploy these polls, followed by live phone. The polling method plays a crucial role in reducing certain response biases. For instance, if the poll is conducted via an online panel with anonymous responses, participants may feel more comfortable expressing their true voting intentions, potentially reducing social desirability bias.

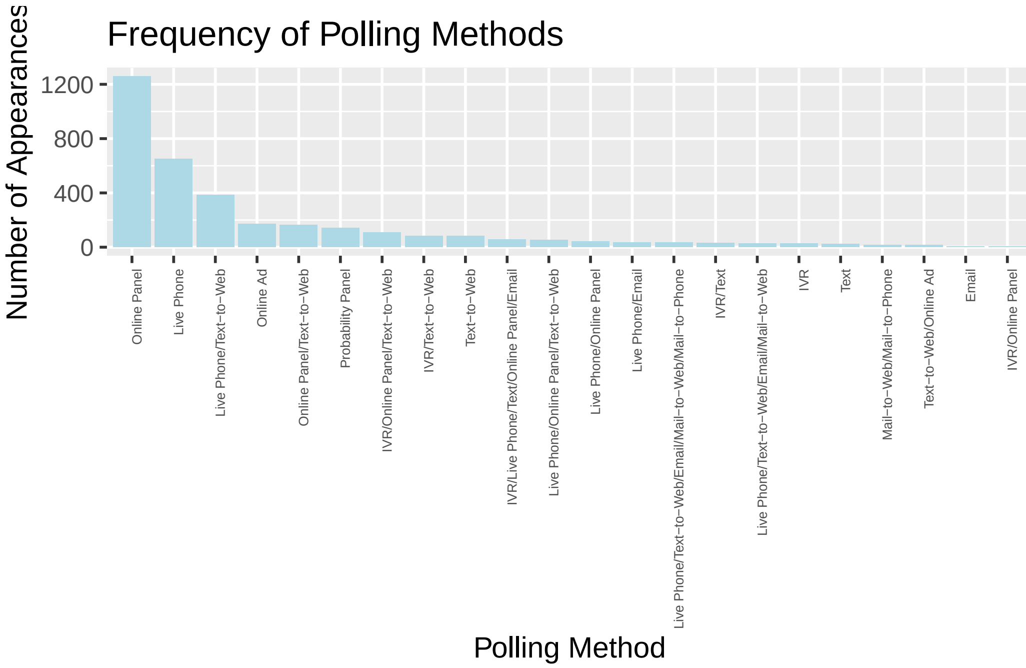

Figure 2: Frequency of Various Polling Methods used by FiveThirtyEight in their “pool of polls”

Finally, Figure 3 shows the number of votes per candidate. As expected, Harris and Trump have the highest number of votes, with Harris having slightly more votes based on the surveyed polls than Trump. Other candidates with a noticeable number of votes in the graph include Robert F. Kennedy and Jill Stein. However, their totals are significantly lower than those of Harris and Trump, making them likely insignificant in the election. It is also important to mention that Biden still appears in these polls, and therefore in Figure 3, due to the lag time in poll data processing by FiveThirtyEight. However, since he has dropped out of the race, he will not be included in our model.

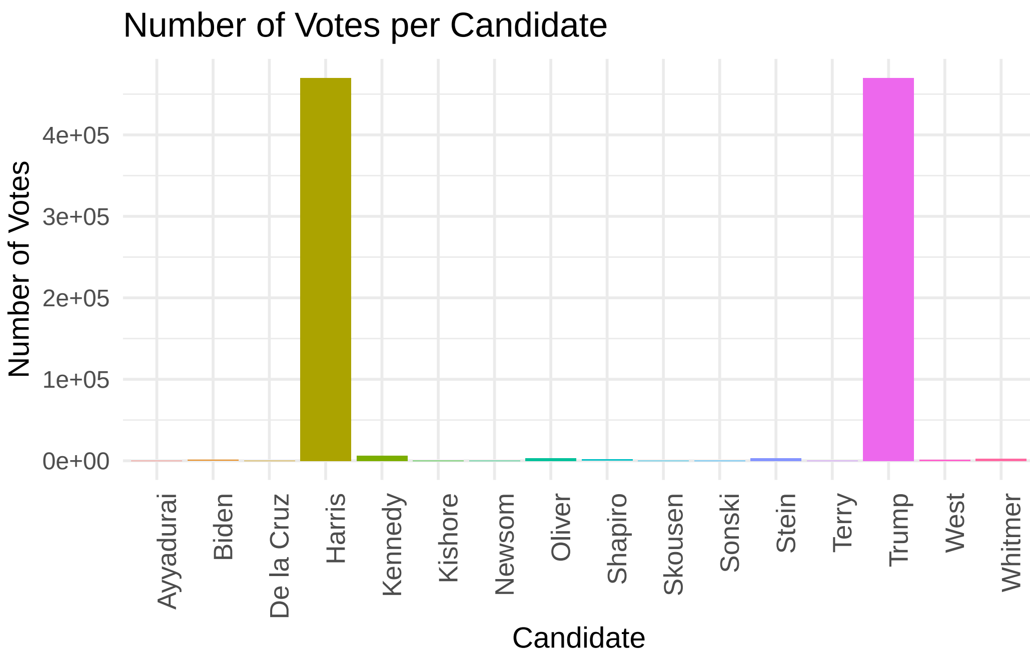

Figure 3: Number of Votes per Candidate. This reflects the popular vote for each candidate and does not account for the electoral college sytem.

3 Model

All model construction and analysis was conducted using R (R Core Team (2023)) with the following key packages:

- **tidyverse:** For data manipulation and visualization. (Wickham et al. (2019))
- **rstanarm:** For Bayesian model fitting using Stan. (Goodrich et al. (2024), Brilleman et al. (2018))
- **brms:** Alternative package for Bayesian regression modeling. (Bürkner (2017), Bürkner (2018), Bürkner (2021))
- **bayesplot:** For comprehensive model diagnostics and visualization.(Gabry and Mahr (2024), Gabry et al. (2019))

To predict state level support and calculate the expected Electoral Votes (EV) for each candidate in the upcoming US election, we employed a Bayesian linear regression model. The model is specified as follows:

$$\text{Support}_i = \beta_0 + \beta_1 \times \text{Sample Size}_i + \beta_2 \times \text{Days to election}_i + \beta_3 \times \text{Transparency Score}_i + \beta_4 \times \text{Pollscore}_i + \gamma \times \text{State}_i + \epsilon_i$$

- **Support**
  Definition: The support level for candidate (i) in a given poll.
  Justification: Measures the proportion of respondents supporting candidate (i), providing insight into their current standing.

- **Sample Size**
  Definition: Represents the number of respondents in poll (i).
  Justification: Larger sample sizes generally yield more accurate estimates of support, reducing sampling variability.

- **Days to election**
  Definition: Denotes the number of days remaining until the election when poll (i) was conducted.
  Justification: Accounts for temporal proximity to the election, capturing potential fluctuations in support as the election approaches.

- **Transparency Score**
  Definition: A numerical score reflecting the pollster's transparency regarding their methodology, ranging up to 10, with higher scores indicating greater transparency.
  Justification: Higher transparency scores may correlate with the reliability and credibility of poll results, influencing voter trust.

- **Pollscore**
  Definition: A numeric value representing the reliability and bias of the pollster, where negative values denote better reliability.
  Justification: Accounts for systematic deviations in poll results based on pollster performance, ensuring more accurate support estimates.

- **State**
  Definition: Captures state-specific effects as fixed effects.
  Justification: Allows the model to account for regional variations in support that are not explained by other predictors, incorporating state-specific nuances.

An alternative approach considered was modeling State as a random effect to account for unobserved heterogeneity across states. However, given the focus on specific state-level predictions and the availability of sufficient data per state, fixed effects are chosen in this case. Ideally we would fit both models and use diagnostics to choose which has lower bais and variance, and thus better predictive quality. Additionally, simpler models excluding variables like Transparency Score and Poll Score were evaluated but found inadequate in capturing the state-by-state differences in the transparency and reliability of the polls, potentially leading to biased support estimates.

Process:

For each candidate and state, we aggregated polling data by calculating the mean values of each predictor. For example, here is the mathematical notation of Mean Days to Election, this value was calculated for each state for both candidates:

$\text{Mean Days to Election}_s = \frac{1}{N_s} \sum_{i=1}^{N_s} \text{Days to Election}_i$

where (N_s) is the number of polls for state (s).

Averaging predictors per state smooths out poll-to-poll variability and provides a representative set of predictors for each state, allowing use to make reliable state-level predictions. Instead of averaging, a training/testing split or regional aggregations based on 538's political regions could have been employed. However, a training/testing split is beyond the scope of this paper, and using political regions would have complicated the derivation of the model.

Using the averaged state level predictors, we generated posterior predictions by drawing 1,000 samples from the posterior distributions of the model parameters. These predictions simulate 1,000 possible election outcomes, and by taking the proportion of victories for Harris and Trump we estimate the probability each candidate has of winning the election. To generate an election map, we simply take the average outcome of each state across the 1,000 simulations. Alternative methods could include using point estimates from the posterior mean; however, simulating multiple outcomes offers a more comprehensive assessment of uncertainty and variability in election outcomes.

For states lacking recent polling data, we incorporated historical averages of Democratic and Republican support from elections since 2000. This data was collected from a github repository (Timm (2023)) who collected their data from Wikipedia. This decision is based on the assumption that certain states are very unlikely to change their voting patterns, reducing the necessity for current polling data. The ten states that do not have polling data have not changed parties in at least the last 20 plus years (Section A.4). Alternatively, imputation methods or political regional averages could have been used, but historical averages are used for their simplicity and relevance.

Though Maine and Nebraska allow for their Electoral College votes to be split. In Maine two of its four votes are awarded statewide and the other two for each of its congressional districts. In Nebraska two of its five votes are awarded statewide and the other three for each of its congressional districts. Our model simplifies this dynamic and all of their votes are assigned according the support overall in the state. The main reason for this is that some of the districts have little or no polling data, making it difficult to measure their support for either candidate.

The Bayesian models were implemented using the `rstanarm` package in R, which interfaces with Stan for efficient Markov Chain Monte Carlo (MCMC) sampling. Data manipulation and visualization were conducted using the `tidyverse` suite of packages, while model diagnostics used `bayesplot` and other related packages. This combination of tools facilitated a streamlined workflow for model fitting, prediction, and validation.

3.1 Diagnostics

Comprehensive diagnostics were performed to assess model convergence, fit, and predictive performance. Graphs and more details can be found in the relevant appendices. We evaluated model convergence using the R-hat statistic and trace plots. Trace plots, Section A.6 revealed well-mixed chains without discernible trends, further supporting the reliability of the MCMC sampling process.

Posterior predictive checks were conducted to assess the model's ability to replicate observed data, Section A.7. Density overlay plots for both Trump and Harris models indicated reasonable fits, although some discrepancies were noted near the peaks of the distributions. Specifically, the Trump model's replicated data tended to be slightly smaller around the peak, while the Harris model exhibited more extreme differences at the peak.

Residual plots (found in Section A.8, depicting residuals versus fitted values, showed notable clumping in the central region for both models, along with vertical lines. This pattern may indicate unmodeled heterogeneity or data limitations, suggesting areas where the model's fit could be improved. Despite these observations, the overall residual distribution did not reveal significant systematic errors, supporting the model's general adequacy.

Ideally several models would be created then using model selection techniques would be used to determine the best model for predicting the outcome of the US election. Due to time constraints the following models were considered but not implemented. A set of Bayesian models with state as a random effect and models excluding specific predictors like Transparency Score. More complex models with additional interaction terms or non-linear components were also considered

3.1.0.1 Assumptions and Limitations

The model operates under several key assumptions:

1. **Linearity:** The relationship between predictors and support is linear.
2. **Normality of Residuals:** The error terms follow a normal distribution.
3. **Independence:** Observations are independent given the predictors.
4. **Stable Partisan Leanings:** Historical averages adequately represent states without current polling data.

Limitations:

- **Model Misspecification:** If non-linear relationships exist between predictors and support, the linear model may not capture these dynamics effectively.
- **Reliance on Historical Data:** Using historical averages for certain states may not account for recent political shifts or emerging trends.
- **Residual Clumping:** Observed clumping in residual plots suggests potential areas for model refinement, such as incorporating additional predictors or interaction terms.

The final Bayesian linear regression model balances complexity and interpretability, incorporating relevant predictors to capture poll reliability and state-specific effects while maintaining computational efficiency. The inclusion of transparency_score and pollscore enhances the model's ability to account for poll quality, while fixed state effects ensure accurate regional support estimates. Validation through convergence diagnostics, posterior predictive checks, and residual analysis confirms the model's robustness and reliability within its assumptions.

4 Results

(**table-2?**) displays each candidates chance of winning and average number of Elector Votes they received over the 1000 simulated elections. Our model predicts that Harris has a 67.4% chance of winning the US presidential election. While Trump only has a 31.7% chance of winning. There is also a 0.9% chance of a tie On average across the 1,000 simulations Harris wins 279.33 Electoral Votes, while Trump wins on average 258.67 votes.

Table 1: Election Prediction Results

	Candidates	
Metric	**Harris**	**Trump**
Chance of Winning	67.4%	31.7%
Average Electoral Votes	279.33	258.67

Election Prediction Results

Based on the simulated elections, Figure 4 illustrates the most likely winner of each state. In this scenario Harris wins 276 Electoral Votes, and Trump wins 262. A table in Section A.5, shows how this map was constructed by taking the most probable winners of each state.

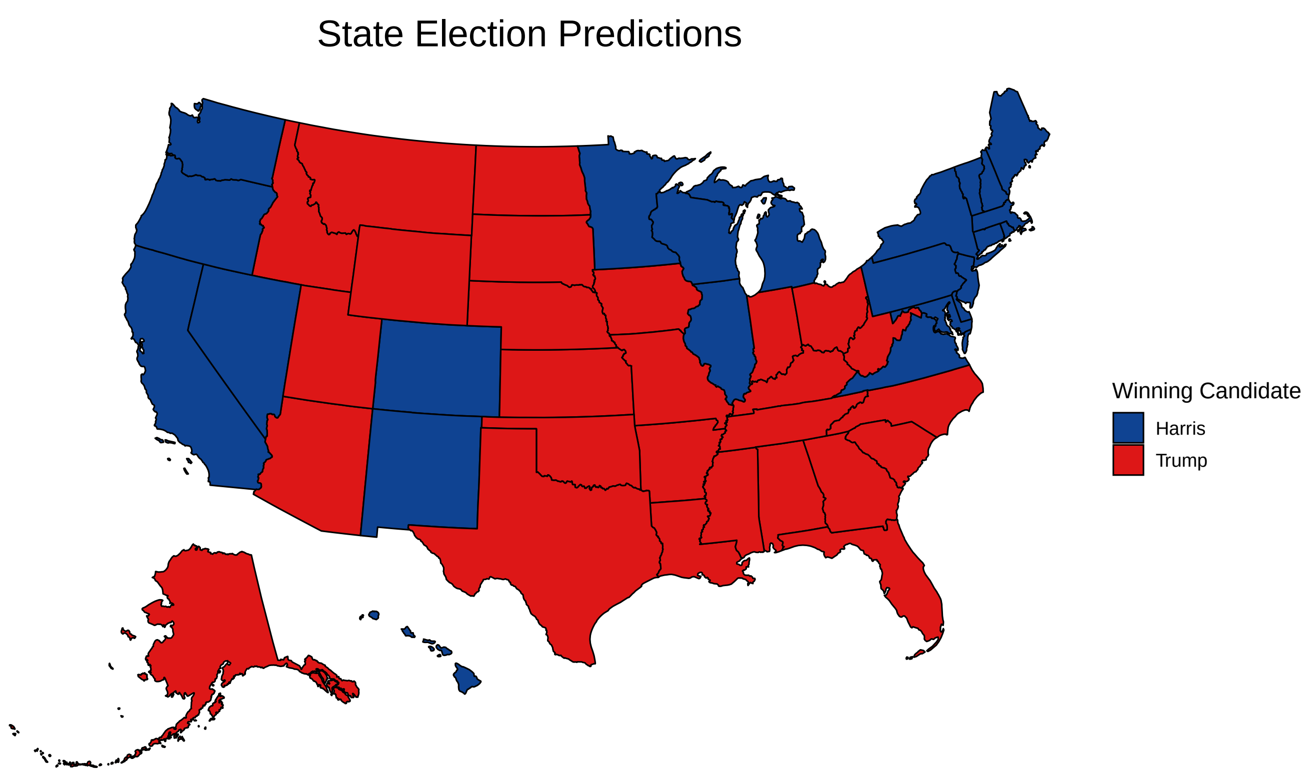

Figure 4: Election Map

The results of the US elections in large part hinge on the direction of the swing states. A swing state is a state that holds a significant number of electoral votes and could reasonable be won by either Trump or Harris. Figure 5 shows the probability that both candidates have of winning each of the 7 swing states.

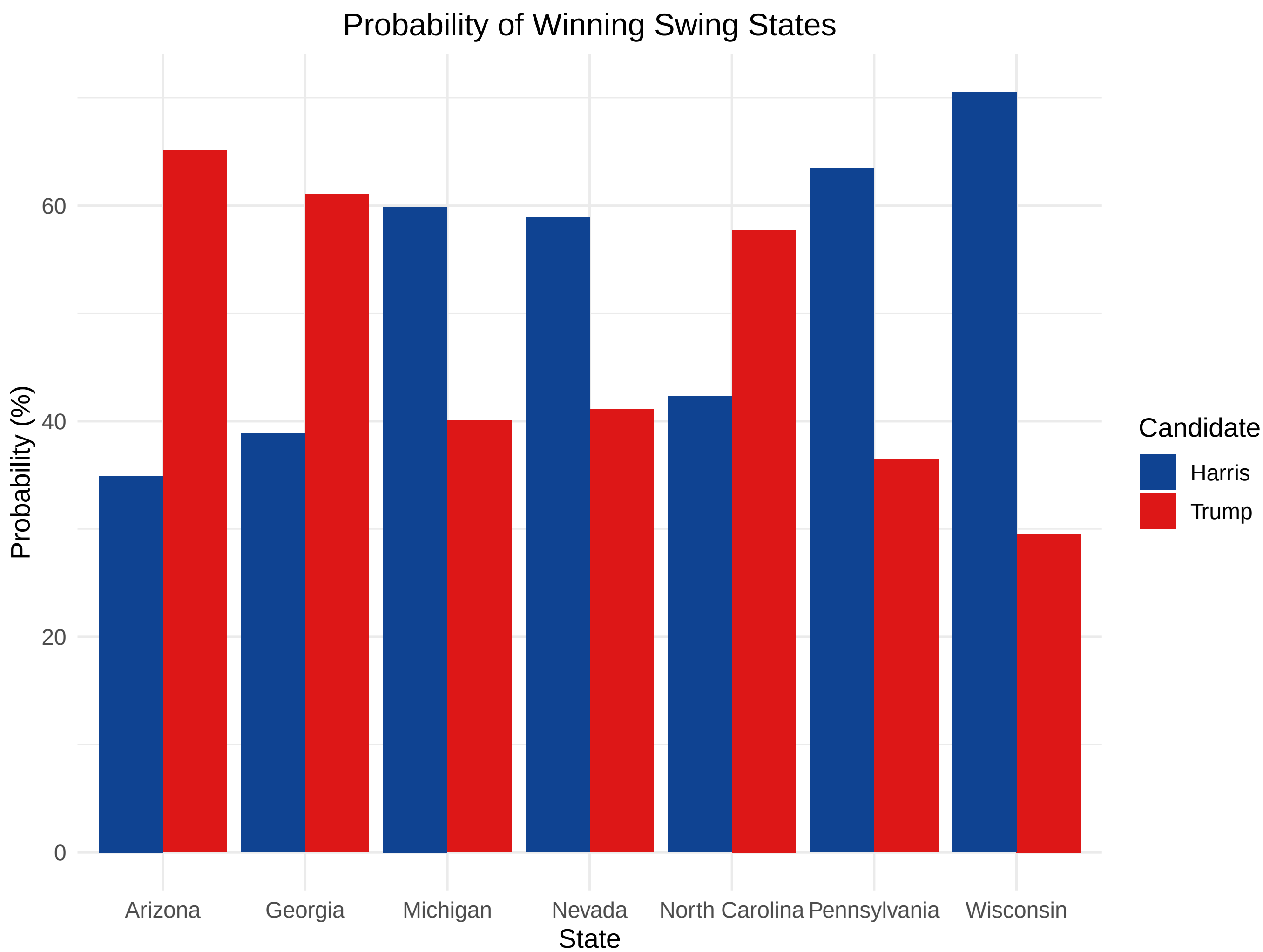
Probability of Winning Swing States
Probability (%)
0
20
40
60
Arizona
Georgia
Michigan
Nevada
North Carolina
Pennsylvania
Wisconsin
State
Candidate
Harris
Trump

Figure 5

5 Discussion

5.1 Implications and Potential Weaknesses

Our model predicting the U.S. election carries important social and political implications. Accurate forecasts can shape voter perceptions, potentially affecting turnout by leading some individuals to feel their votes are less crucial if a particular outcome seems likely. Additionally, these predictions inform campaign strategies, prompting candidates to focus resources on undecided voters in critical swing states, which play a pivotal role in tight elections like the upcoming 2024 race. For instance, our model anticipates that key swing states such as Pennsylvania, Michigan, and Wisconsin will award their electoral votes to Kamala Harris, contributing to her victory. However, the forecast indicates that Harris is expected to win by narrow margins of approximately 1.98%, 1.47%, and 1.46% in these states, respectively.

Our model offers valuable insights into the landscape of the current U.S. election; however, it does face a few limitations. One of the primary challenges stems from gaps in our dataset, specifically the absence of polling data for several states. When the polling data was collected and compiled by FiveThirtyEight (2024a), it did not comprehensively include every state, leading to missing data points that affect our predictions. Our model relies on the electoral college system, where each state's popular vote determines its allocation of electoral votes, ultimately deciding the election's outcome. Missing data from certain states introduces an incomplete representation of the voting scene, affecting the accuracy of our predictions. Each state often exhibits unique voting behaviors and excluding certain states can introduce a bias in our model. This is especially problematic if the omitted states are swing states or tend to have historically unpredictable voting patterns, as these states can play a decisive role in election outcomes. Moreover, missing data for larger states further exacerbates the issue by potentially leading to a miscalculation of the Electoral College outcome. States with significant electoral votes carry substantial weight, and without accurate polling information from them, our model may struggle to forecast the election winner effectively.

Another limitation of our model is that we stopped pulling new poll data from FiveThirtyEight (2024a) as of October 19, 2024. Consequently, the model does not account for any last-minute events that could have impacted voter opinions from then until election day. This could potentially lead to misleading predictions if there has been a significant change in public sentiment. Furthermore, last-minute shifts in public opinion could result in different voter turnout than initially anticipated or introduce polling response biases, which predictive models may struggle to capture if they are based primarily on earlier data.

Our model does not take into account state correlations based on past election data and research based insights into voter preference. For example a model developed by Chernov, Elenev, and Song, posits if Harris were to win Pennsylvania, then her chances of winning Nevada would increase from 24% to 68% [https://anderson-review.ucla.edu/wp-content/uploads/2024/10/voter_preferences_updated.pdf]. These correlations between states could have a large impact on election predictions and our model is limited by not modelling these interactions.

5.2 Next Steps

Consequently, our forecast suggests that candidates should concentrate their efforts on swing states, as increased campaigning in these areas could significantly impact voter turnout and the popular vote. Additionally, enlisting high-profile figures or celebrities, like Elon Musk or LeBron James—who have recently endorsed Trump and Harris, respectively—could enhance their appeal in these swing states. These endorsements would help candidates reach diverse and niche demographics, broadening teheir voter base and potentially swaying undecided voters.

5.3 Betting Markets

Prediction betting markets provide a new data source that may be valuable for predicting the outcome of the US election. These online betting markets allow people to buy shares that represent an outcome of some event. If the event occurs then the share is worth $1, if not then the share's price falls to 0. This means the live share price reflects the probability of the event occurring. For example, for the outcome of the presidential election market, if Trump's share price is 65 cents this means the market is predicting there is a 65% chance that Trump will win the election. Recently the trading volume on these markets is significant, which indicates that these markets may be helpful in predicting election outcomes. Some papers have included data from these markets in their models to improve their prediction accuracy.[REF https://anderson-review.ucla.edu/wp-content/uploads/2024/10/voter_preferences_updated.pdf].

Some key advantages of these markets is that they are real time, so the market adjusts quickly to important and material news. Where as poll are sluggish to incorporate big changes as new polls need to be sent out and analysed. Another advantage is that these markets may attract sophisticated investors that could employ methods that differ from polling but still offer predictive capability. The larger the trade volume the more likely these actors are to be present and wash out any unsophisticated trading.

On the other hand the main disadvantage is that all of the investors in this market essentially completely anonymous. This means there is no way of checking whether a large investor, who has significant price setting power, is indeed sophisticated or not. Furthermore, because these markets have only recently become popular or even legal in some cases one might expect these markets to be full of amateurs. These markets may display a bais as well, where the largest market Polymarket [REF] is based on crypto-currency, and Trump is popular among crypto-enthusiasts, and large financial institutions are underrepresented in the crypto-market.

This is a new frontier in forecasting and including data from these markets may prove beneficial. Or these markets may be nothing more than a hobby, and hold no useful data.

A Appendix

A.1 Analysis into a Pollster's Methodology

One particular poll within our data sample is one conducted between October 11-14 by Beacon Research and Shaw & Company Research (Blanton 2024). Targeting the sample frame of registered American voters, this poll collected data on 1,110 participants' responses on 51 questions through either live phonecall interviews or an online survey. The sample was recruited by applying the probability proportionate to size method on the nationwide voter file of registered voters' phone numbers (Blanton 2024). This sampling method is a type of probability sampling in which each unit's chance of being included is dependent upon their size, a measure that is based off of some characteristic that is known about every single unit and often related to the main variable of interest. By giving a higher probability of inclusion to units of greater importance based on its size during sampling, a more accurate estimate can be obtained (Latpate et al. 2021). In the case of this poll, the number of voters per state region was used to determine the proportion of individuals contacted (Blanton 2024). This ensures that a greater number of residents in larger states, such as California, would be contacted compared to smaller states, such as Hawaii. This method of sampling allows for a better match between the voices represented by poll results to the ones involved with the real election.

The key finding from this poll is that 48% of participants favored Harris while 50% favored Trump, with the lead being consistent in the larger sample of registered voters and smaller subsample of likely voters. Despite this, responses also indicate that Harris is narrowly leading in seven swing states, meaning that Democrats could potentially win by electoral college while losing the popular vote (Blanton 2024). This report also notes that current results are Trump's highest approval ratings since Biden dropped out of the race, while support for Harris is at its lowest.

One strength of this questionnaire is that information regarding the respondent's confidence level and commitment level was also collected. With questions such as "How often do you make a point to read or listen to the news?" and "Are you certain to support that candidate, or do you think you may change your mind and support someone else?", this poll is able to gain a more indepth view of how easily these participants may be swayed in the time between the polling and the real election (Blanton 2024). As well, probabilistic statistical models based on past voting history, interest in current election, age, education, race, ethnicity, church attendance, and marital status were used to predict the respondents' likeliness to vote; interestingly, results found that the results for the full sample vs subsample of 870 likely voters varied by a $\pm 3\%$, which can be quite significant given how tight the race is currently (Blanton 2024).

On the other hand, one potential weakness of this poll is that it was sponsored by Fox News, a news site that has historically been Republican-leaning. Though from the reported methodology alone, no evident biasing of the pollster can be observed, this potential source of biased reporting must be noted as it can affect the type of analyses that occurred and which key findings are the focus of news reports. As well, another potential issue is the lack of indication regarding how non-responses of "Don't know" were handled during analysis (Blanton 2024). This is a critical area to pay attention to as it is an option offered on every question, selected by up to 4% of participants on key questions such as "If the presidential election were today, how would you vote?" (Blanton 2024).

A.2 Idealized Methodology and Survey

In an ideal world, where a $100k in funds and all necessary resources could be obtained from a neutral source for the purpose of polling the population, the idealized methodology for a survey that aims to forecast the results of the US election should be one that uses quota sampling and the probability proportionate to size method to target the states of interest specifically. Historically, certain states have very strong and consistent party preferences; for example, we can be almost entirely certain even without conducting any current research that California would be voting Democrat, simply because this has been how they voted in the past eight elections (USAFacts Team 2024). Out of the 51 states, there are 20 states which voted for the same party each time in the past nine elections and therefore there is a high probability that the pattern will be upheld with this year's election (USAFacts Team 2024). On the other end of the spectrum, there are seven states which have voted for each party three of more times within the last nine elections——these would be the states of greater interest to us (USAFacts Team 2024).

Using the non-probability sampling method of quota sampling, these swing states will first be selected. This targetted sampling process allows us to focus our polling efforts on gathering more data from the areas where there is higher uncertainty, but the tradeoff is bias is easily introduced with the assumptions we are making, potentially impacting the accuracy of our forecast (Chen, Felt, and Henry 2018). This means that more work is required to adjust for these biases while also minimizing the inflation of variance (**quotasamping?**).

Next, within these states, probability proportionate to size will be used to select participants from the nationwide file of registered voters to ensure that the number of people polled is proportional to the number of voters per region. This is the sampling method used in Blanton (2024), as discussed above in Section A.1, and it allows for greater accuracy of prediction due to each unit's strategically adjusted probability of inclusion (Latpate et al. 2021).

After the sample has been selected with the above methods, the poll itself will be conducted through a survey sent to the chosen participants in order to reduce interviewer bias, or the effect of the interviewer's own views on the measured responses, which can affect results when face-to-face or phone interviews are utilized (Alexander 2023). As well, to reduce social desirability effect, the survey must emphasize in the beginning that the respondents' identity will be kept anonymous to the researchers and the final reported data (Stantcheva 2023).

- questions (wording, order, options)
- treatment of non-response
- offer of incentive

After analysis of poll responses, the votes of the consistent states should be added back into consideration before making the final forecast for election results.

A sample implementation of such poll, created using Google form, can be found here. In this version, the order of questions presented is consistent for all respondents due to the limitations of Google form, but for an idealized methodology with more professional surveying platform, such as Survey Monkey, the question order would be randomized when being presented to different respondents with the Question Randomization feature (SurveyMonkey, n.d.). This will help reduce response order bias, which is when the answer of a question is affected by the order in which it appeared in the survey (Stantcheva 2023).

A.3 Table Detailing First Few Observations of Dataset

Table 2: First few observations of cleaned data set

pollster	numeric_grade	pollscore	end_date	transparency_score	question_id	method
AtlasIntel	2.7	-0.8	2024-10-31	6	215182	Online Ad
AtlasIntel	2.7	-0.8	2024-10-31	6	215182	Online Ad
AtlasIntel	2.7	-0.8	2024-10-31	6	215182	Online Ad
AtlasIntel	2.7	-0.8	2024-10-31	6	215182	Online Ad

Table 2: First few observations of cleaned data set

pollster	numeric_grade	pollscore	end_date	transparency_score	question_id	method
AtlasIntel	2.7	-0.8	2024-10-31	6	215183	Online Ad

Table 3: First few observations of cleaned data set

state	sample_size	candidate	support	election_date	days_to_election	initial_weight	weight	pol_region
North Carolina	1373	Harris	46.7	2024-11-05	5	0.9486833	1.094691	Southeast
North Carolina	1373	Trump	50.7	2024-11-05	5	0.9486833	1.094691	Southeast
North Carolina	1373	Stein	0.7	2024-11-05	5	0.9486833	1.094691	Southeast
North Carolina	1373	Oliver	0.3	2024-11-05	5	0.9486833	1.094691	Southeast
North Carolina	1373	Harris	47.0	2024-11-05	5	0.9486833	1.094691	Southeast

A.4 Deep Red/Blue State Voting History

State	Last Time Voted for a Different Party	State	Last Time Voted for a Different Pa
Alabama	1980	Kentucky	2000
Delaware	1992	Louisiana	2000
Hawaii	1998	Mississippi	1980
Idaho	1976	Tennessee	2000
Illinois	1992	Wyoming	1976

Note:

Note: Data collected only went back to 1976, some states may have been voting for the same party farther back tha

Number of Wins in 1000 Simulations by State

A.5 State By State Win

Table 4: Election Results by State

state	trump_wins	harris_wins	winner
Alabama	1000	0	Trump
Alaska	955	45	Trump
Arizona	651	349	Trump
Arkansas	1000	0	Trump
California	0	1000	Harris
Colorado	2	998	Harris
Connecticut	2	998	Harris
Delaware	0	1000	Harris
Florida	946	54	Trump
Georgia	611	389	Trump
Hawaii	0	1000	Harris
Idaho	1000	0	Trump
Illinois	0	1000	Harris
Indiana	999	1	Trump
Iowa	809	191	Trump
Kansas	967	33	Trump
Kentucky	1000	0	Trump
Louisiana	1000	0	Trump
Maine	3	997	Harris
Maryland	0	1000	Harris
Massachusetts	0	1000	Harris
Michigan	401	599	Harris
Minnesota	35	965	Harris
Mississippi	1000	0	Trump
Missouri	994	6	Trump
Montana	1000	0	Trump
Nebraska	926	74	Trump
Nevada	411	589	Harris

New Hampshire	33	967	Harris
New Jersey	0	1000	Harris
New Mexico	18	982	Harris
New York	0	1000	Harris
North Carolina	577	423	Trump
North Dakota	1000	0	Trump
Ohio	981	19	Trump
Oklahoma	1000	0	Trump
Oregon	31	969	Harris
Pennsylvania	365	635	Harris
Rhode Island	0	1000	Harris
South Carolina	999	1	Trump
South Dakota	1000	0	Trump
Tennessee	1000	0	Trump
Texas	961	39	Trump
Utah	1000	0	Trump
Vermont	0	1000	Harris
Virginia	30	970	Harris
Washington	0	1000	Harris
West Virginia	1000	0	Trump
Wisconsin	295	705	Harris
Wyoming	1000	0	Trump
District Of Columbia	0	1000	Harris

Number of Wins in 1000 Simulations by State

A.6 Trace Plots

Figure 6: Trump Trace Plots

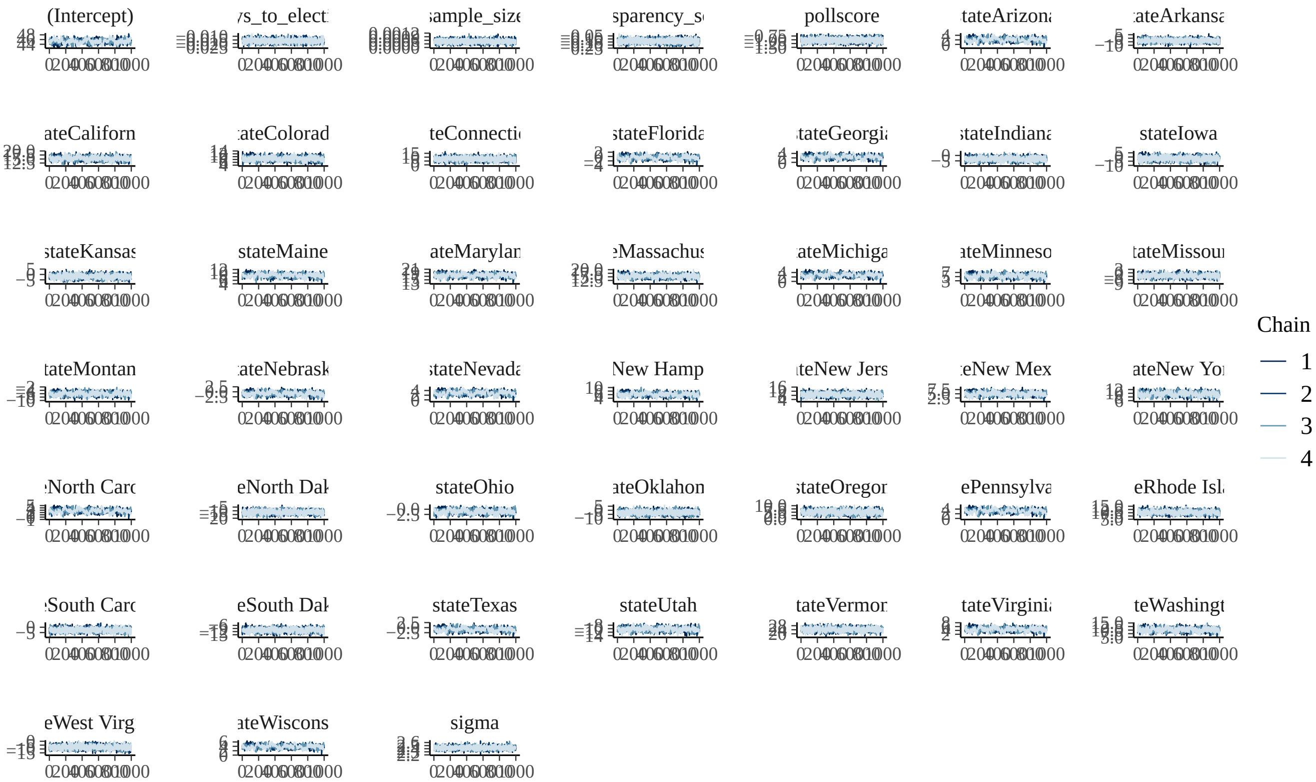

Figure 7: Harris Trace Plots

A.7 Posterior Predictive Checks

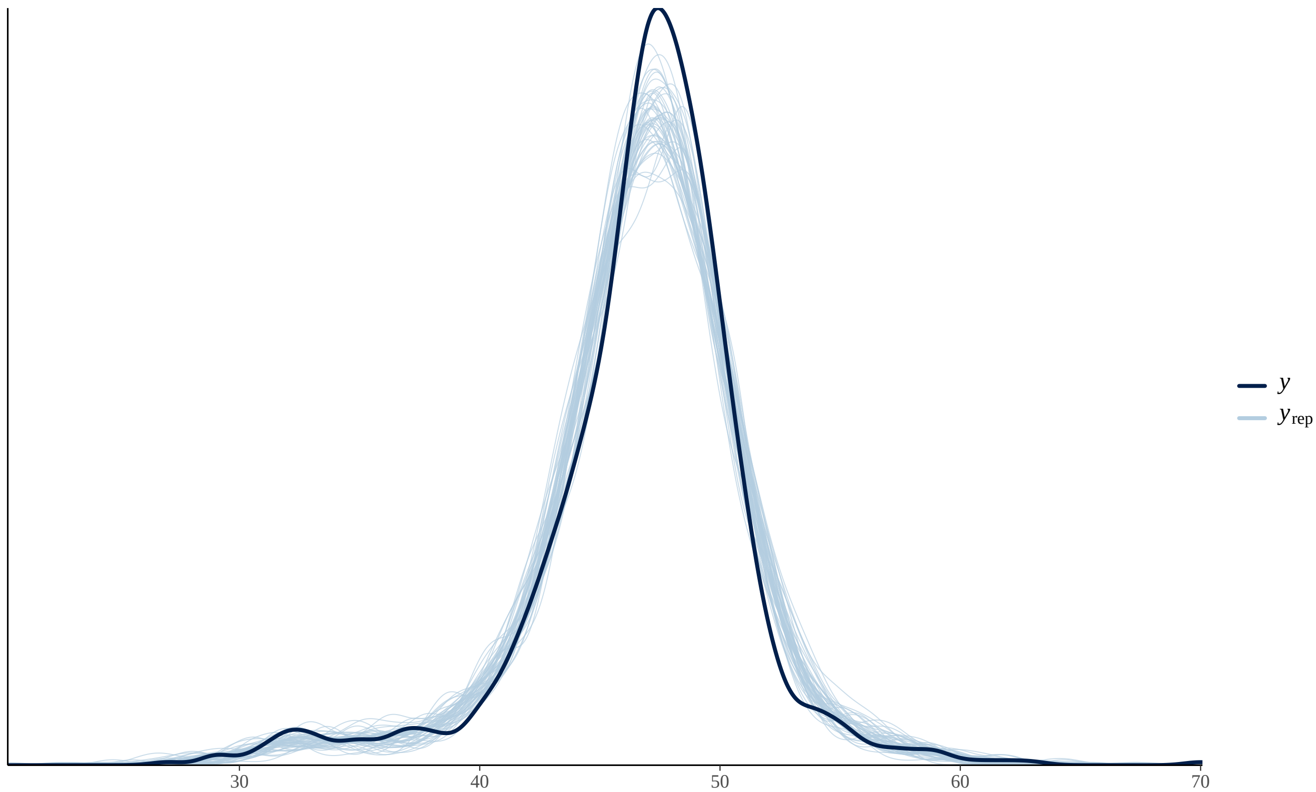

Figure 8: Trump Posterior Predictive Checks

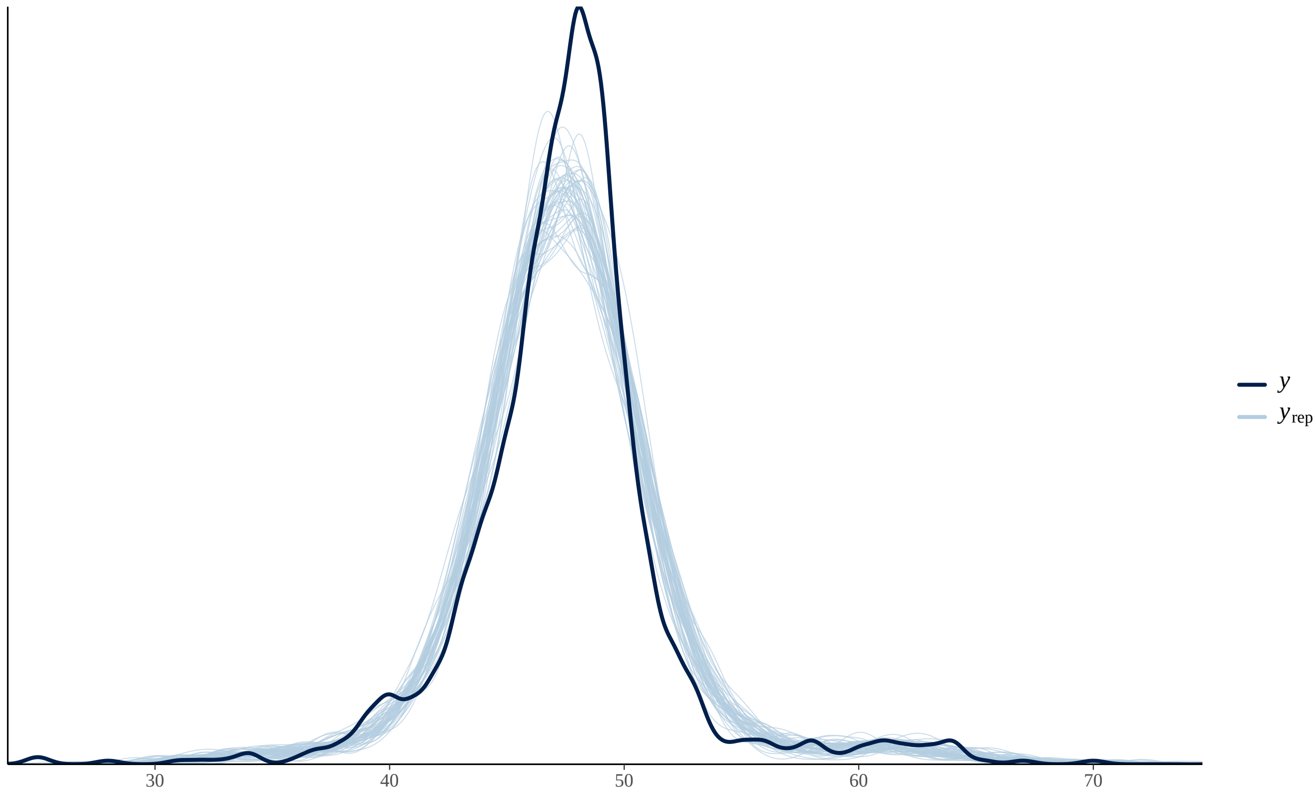

Figure 9: Harris Posterior Predictive Checks

A.8 Residual Plots

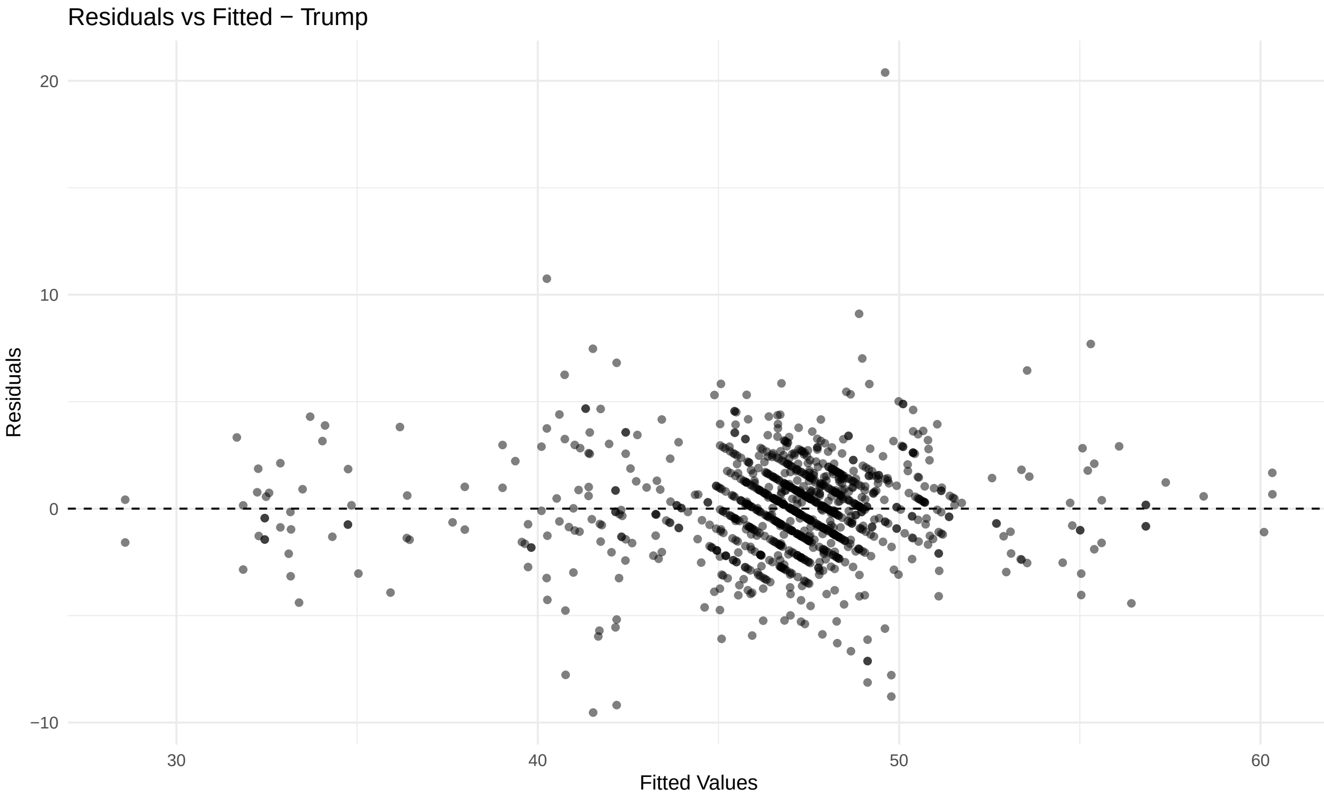

Figure 10: Trump Residuals

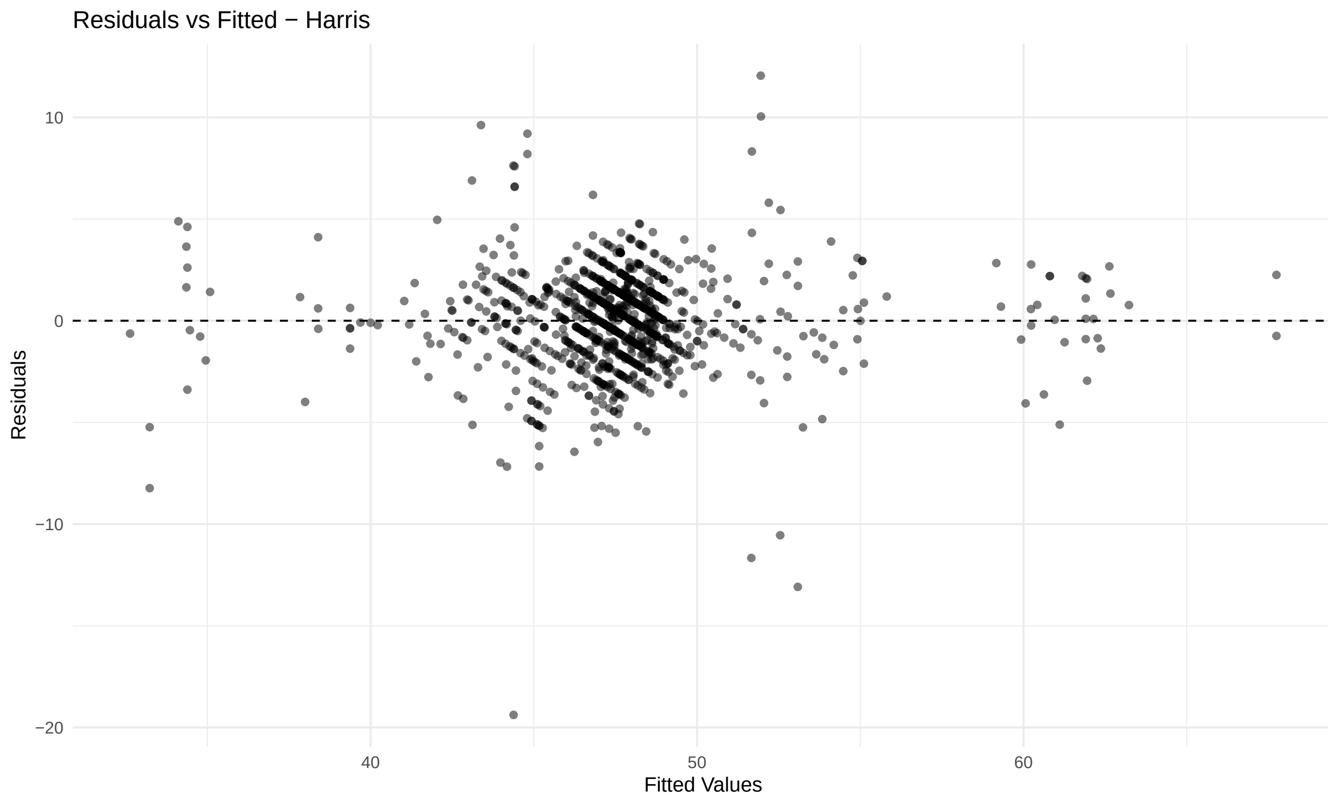

Figure 11: Harris Residuals

References

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

Blanton, Dana. 2024. "Fox News Poll: Trump Ahead of Harris by 2 Points Nationally." https://www.foxnews.com/official-polls/fox-news-poll-trump-ahead-harris-2-points-nationally.

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/.

Bürkner, Paul-Christian. 2017. "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1): 1–28. https://doi.org/10.18637/jss.v080.i01.

———. 2018. "Advanced Bayesian Multilevel Modeling with the R Package brms." *The R Journal* 10 (1): 395–411. https://doi.org/10.32614/RJ-2018-017.

———. 2021. "Bayesian Item Response Modeling in R with brms and Stan." *Journal of Statistical Software* 100 (5): 1–54. https://doi.org/10.18637/jss.v100.i05.

Chad de Guzman and Koh Ewe. 2024. "A Guide to Kamala Harris' Views on Abortion, the Economy, and More." https://time.com/7001208/kamala-harris-views-abortion-economy-immigration-israel-gaza/.

Chen, Heng, Marie-Hélène Felt, and Christopher Henry. 2018. "2017 Methods-of-Payment Survey: Sample Calibration and Variance Estimation." https://doi.org/10.34989/tr-114.

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

———. 2024b. *FiveThirtyEight Presidential General Polls Data*. https://projects.fivethirtyeight.com/polls/data/president_polls.csv.

Gabry, Jonah, and Tristan Mahr. 2024. "Bayesplot: Plotting for Bayesian Models." https://mc-stan.org/bayesplot/.

Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. 2019. "Visualization in Bayesian Workflow." *J. R. Stat. Soc. A* 182: 389–402. https://doi.org/10.1111/rssa.12378.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2024. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Latpate, Raosaheb, Jayant Kshirsagar, Vinod Kumar Gupta, and Girish Chandra. 2021. "Probability Proportional to Size Sampling." In *Advanced Sampling Methods*, 85–98. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-16-0622-9_7.

McKenzie Beard and Abbie Cheeseman and Justine McDaniel. 2024. "Harris Vs. Trump on Abortion: Where They Stand on the Issue." https://www.washingtonpost.com/politics/interactive/2024/trump-harris-abortion/.

NBC. 2024. "Harris and Trump: Compare Where They Stand on Key Issues." https://www.nbcnews.com/politics/2024-election/harris-trump-stance-issues-policies-president-race-rcna150570.

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.

SurveyMonkey. n.d. "Question & Page Randomization." https://help.surveymonkey.com/en/surveymonkey/create/question-page-randomization/.

Timm, Jay. 2023. "Presidential Election Results." https://github.com/jaytimm/PresElectionResults.

USAFacts Team. 2024. "How red or blue is your state?" https://usafacts.org/articles/how-red-or-blue-is-your-state/.

Weber. 2024. "Electoral College Information." https://www.sos.ca.gov/elections/electoral-college#:~:text=Under%20the%20%22Electoral%20College%22%20system,more%20%22votes%22%20it%20gets.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

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.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2014. "knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.