A Prediction of the 2024 U.S. Election Outcome*

A Bayesian Spline Regression Analysis of Kamala Harris's Winning

Shanjie Jiao Aman Rana Kevin Shen

November 4, 2024

The 2024 U.S. Presidential Election, set against a backdrop of economic and political challenges, features Donald Trump and Kamala Harris as the main candidates. This study uses Bayesian spline regression on polling data from FiveThirtyEight to predict the election outcome, capturing dynamic voter trends over time. Our analysis suggests a likely win for Kamala Harris, however a possible surprise by Donald Trump. Given the U.S.'s global influence, this result could significantly impact international security, trade, geopolitics, and provide a clear insight of forthcoming policy directions.

Table of contents

1 Introduction

The United States, as the world's largest economy and most influential country, will hold its presidential election on November 5, 2024. Despite facing domestic challenges such as inflation, low employment rates after the pandemic, and increasing political polarization due to conflicts between the Democratic and Republican parties, which have accelerated internal tensions.As of October 2024, former President Donald Trump has secured the Republican Party nomination. Meanwhile, President Joe Biden withdrew from the election on July 21, leading to the nomination of Kamala Harris as the Democratic candidate. These events have made the 2024 U.S. election increasingly complex and unpredictable, further intensifying uncertainty about the future.

*Code and data are available at: [https://github.com/Jie-jiao05/2024_US_Election.git).

The estimand of this study is the predicted outcome of the 2024 U.S. Presidential Election for either Donald Trump or Kamala Harris, based on aggregated polling averages. Our analysis employs Bayesian spline regression to effectively capture the dynamic changes in voting trends over time, allowing us to forecast the likely election result. By comparing the predicted probabilities for both candidates, we aim to identify the probable winner based on current trends in voter support.

Using Bayesian spline regression analysis on the dataset provided by FiveThirtyEight (FiveThirtyEight 2024a), this study forecasts the outcome of the 2024 U.S. election, predicting a victory for Kamala Harris.

Due to the United States' hegemonic position and unparalleled global influence, the outcome of the U.S. election will serve as a guiding light for global developments over the next four years, significantly impacting international security, trade, cooperation, and geopolitics. This article aims to provide a clearer analytical framework for political scientists, the general public, journalists, corporate strategists, and anyone globally concerned with the U.S. election.

The remainder of this paper is structured as follows. In the Data Section we will discuss the data used in the modelling process and the result of the Bayesian Spline fit with its predictor variables and the response after transformation. In the Discussion we delve into the shortcomings of the study and areas for improvement will be described.Appendix is a A Deep Dive to a pollster will be conducted, and Appendix B we will given a idealized survey and methodology

2 Data

2.1 Overview

We use R (R Core Team 2023) and Python (Van Rossum and Drake Jr 1995) to analyze a panel of poll data from Project FiveThirtyEight (FiveThirtyEight 2024b). It consists of voter support data from major pollsters in the US, with national and state-level polls. The data includes labels for pollster robustness as constructed by FiveThirtyEight. Polling date data is included, which gives us snapshots of the progressing state of the general election.

Our analysis is focused on determining whether Donald Trump or Kamala Harris will win the election. We drop all other candidates and parties on the belief that they are not likely to win given how close the race is between the two leading candidates.

For every observation in our data we have percentage of voter support, state, start and end dates of the poll, pollster quality, and party.

2.2 Data Quality Variable: Numeric_Grade

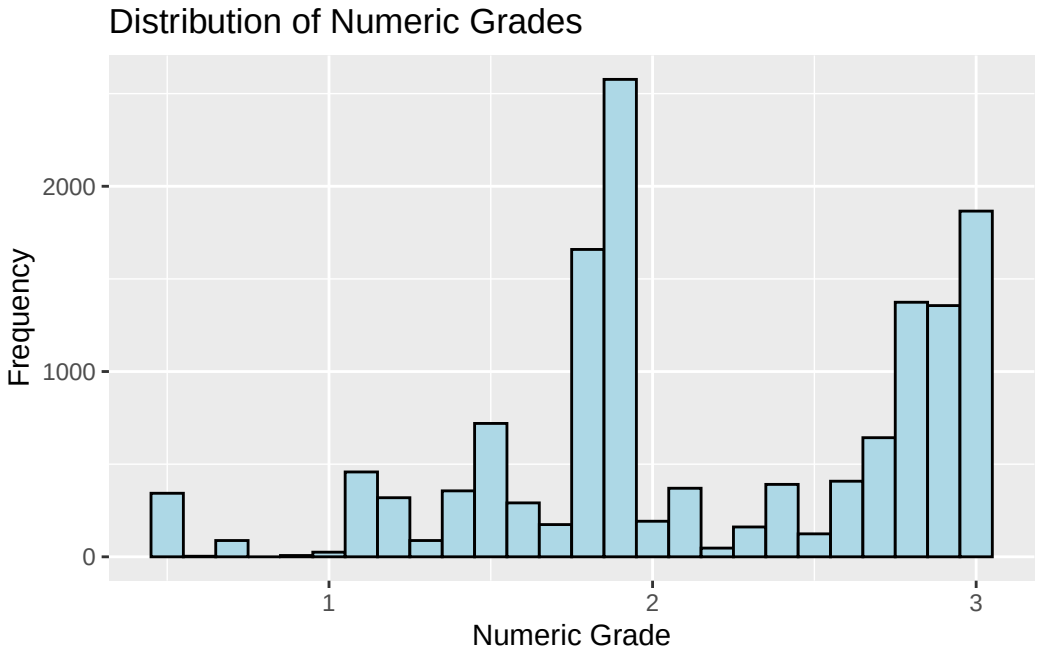


Figure 1: Distribution of Numeric Grades

Numeric_Grade is a numeric rating given by FiveThirtyEight to indicate a pollsters quality or reliability. The rating is based on methodology, transparency and historical accuracy which is important for our analysis as we evaluate eahc poll observation. The raw data set has the following distribution with numeric grades ((**numeric_grade_dist?**)). We choose to drop all polls with numeric grades less than 3 to keep only the highest quality polls.

2.3 Predictor Variable: State

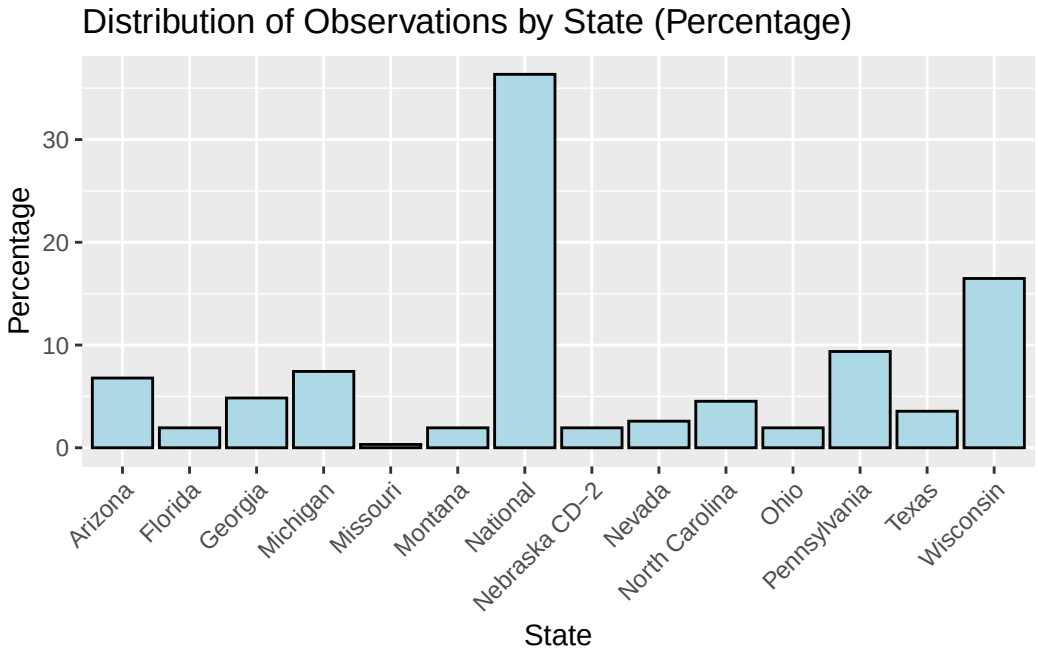


Figure 2: Distribution of Observations by State

The state where the polling took place. This is important for state-level fixed effects, some regions show more supoprt to one party over another. Regarding the data itself, ee notice that over half of all observations are NA, in our analysis we take these to be national level polls. Since this is not explicit in the documentation (FiveThirtyEight 2024b), this is an assumption which is a caveat in our outcome analysis. We notice in the figure ((**state_dist?**)) that not every state is represented, which limits us to the states we can make state-level predictions for.

2.4 Predictor Variable: Party

The party variable identifies whether the observation is for the Democratic or Republican party. What we are interested in is the most-likely presidential candidate so we remove any data before Kamala Harris was announced on July 21 2024, which allows us to use party as a proxy for candidate.

Table 1: Distribution of Observations by Party

party	n
CON	6
DEM	172
GRE	92
IND	101
LIB	74
PSL	2
REP	172

Distribution of Observations by Party

2.5 Predictor Variable: end_date_num

End_date is the end date of the poll and represents the informational cut off date that the results of a poll indicate. We include this as one of the variables in our regression in order to capture time effects such as the momentum and changing support of a candidate over time. End_date_num is the number of days since the start of the period, it increases as we approach the election period. We include it to measure the time effects, we might expect Trump's roportion to increase as end_date_num increases as we have seen historically due to shy voters Cohn (2024).

Plotting a histogram of end dates, we see how polls become much more frequent over time as the election approaches.

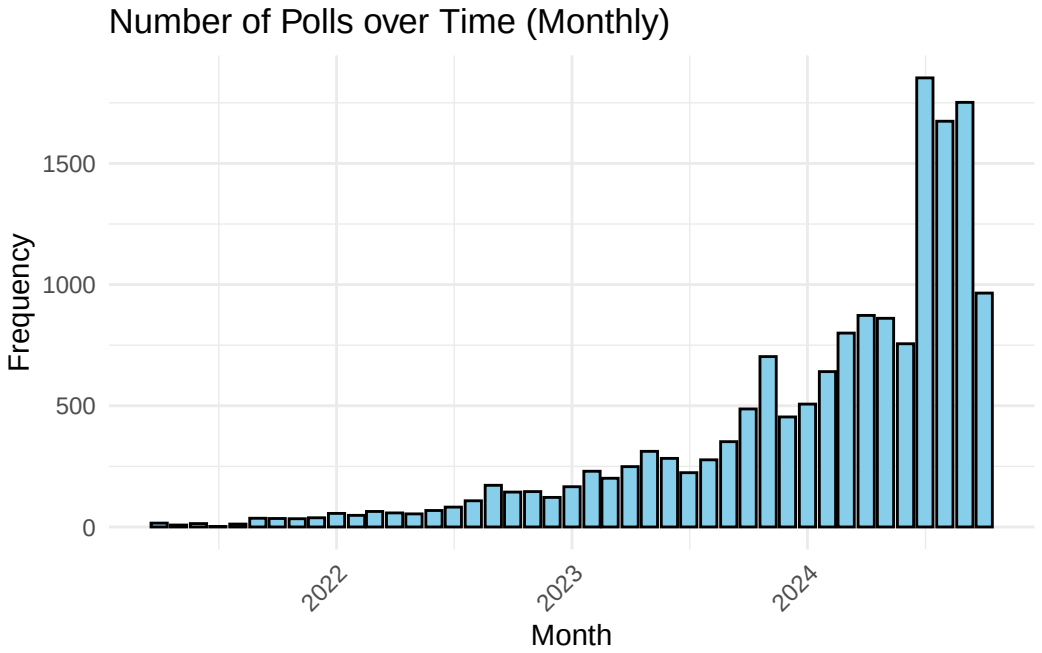


Figure 3: Number of Polls over Time

2.6 Predictor Variable: pollster

Table 2: High-Grade Poll counts by Pollster

pollster	n
Marquette Law School	88
Siena/NYT	314
YouGov	217

High-Grade Poll counts by Pollster

Pollsters are the agencies that do the polls, in our datasetwe find only a few pollsters remain as seen below. We include pollsters in our regression to account for any pollster effects, bias can come from the pollsters themselves, or the demographics they serve. While we do not expect this effect to be strong since we limit ourselves to the highest grade of polls we retain this vairbale for completeness. Additionaly we notice one of the remaining pollsters 'YouGov/Center for Working Class Politics' has very few observations and we drop this because it is not well represented in the regression. The polls are from a collaboration with YouGov and a different entity (2020) and so we see it as distinct form a YouGov poll and do not merge the two.

##Outcome Variable: pct

Pct represents percentage of support, the percentage of respondents that support each party. This is our primary outcome of interest, and will be the target of our regression. To note is that for a particular poll the sum of our pct does not add up to 100. Since we drop other parties, they are not represented in our data and their omission causes the sum to not be 100.

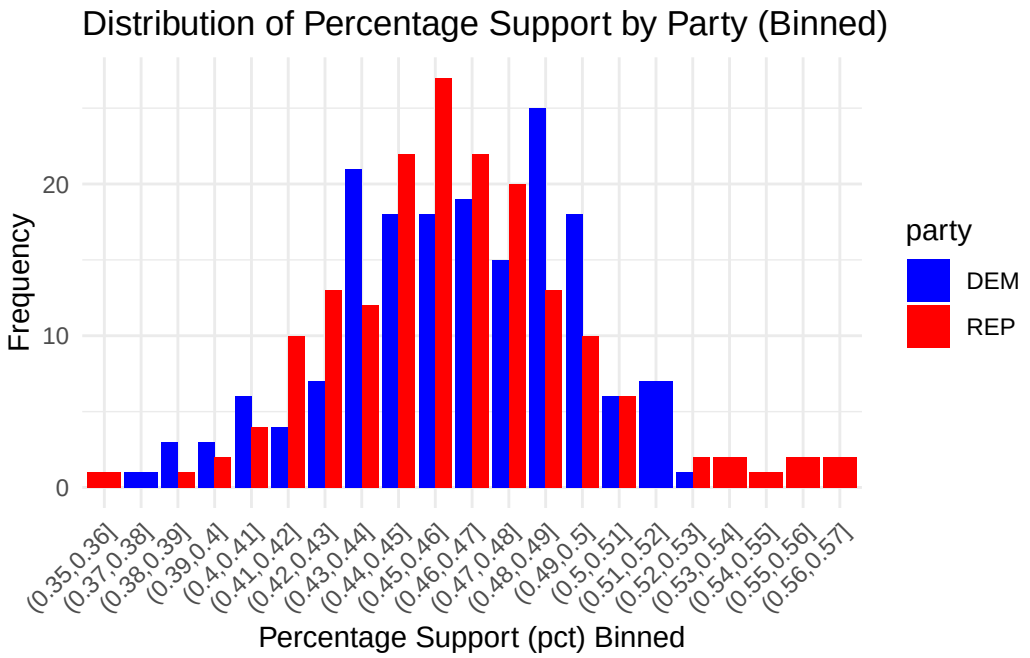


Figure 4: Distribution of Percentage support shows us how each party's support is distributed around 46%, showcasing the close nature of the election

3 Model

Our model is designed to estimate temporal trends in party support as reflected in polling percentages for the Democratic and Republican parties. We employ a Bayesian framework to capture uncertainty and provide flexibility in trend estimation over time. Based on our understanding of state-level variances FiveThirtyEight (2024b), and possible pollster bias, we include them as variables.

3.1 Model Set-up

The response variable (y_i) represents the percentage voter support for a given poll (i). The explanatory variables: - (end_date_num_i): the number of days since the earliest poll date

for the given party, - (pollster_i): the pollster conducting the poll, and - (state_i): the U.S. state where the poll was conducted (or "National" if the poll was nationwide).

For each party, we fit a separate Bayesian model with spline smoothing on (end_date_num) to capture time trends, alongside fixed effects for both pollster and state. We implement the model in R using the `rstanarm` package (R Core Team 2023).

The model's hierarchical structure and notation are as follows:

[

$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (1)\\
\mu_i &= \alpha + f(\text{end\_date\_num}_i) + \text{pollster}_i + \text{state}_i && (2)\\
\alpha &\sim \text{Normal}(50, 10) && (3)\\
f(\text{end\_date\_num}) &\sim \text{B-spline basis expansion} && (4)\\
\text{pollster}_i &\sim \text{Normal}(0, 5) && (5)\\
\text{state}_i &\sim \text{Normal}(0, 3) && (6)\\
\sigma &\sim \text{Exponential}(1) && (7)
\end{aligned}
$$

]

3.1.1 Explanation of Components

- **Response variable (y_i)**: Observed polling voter support percentage for each poll (i). This represents the percentage of voter support.
- **Mean structure (_i)**: Consists of four components:
  - **Intercept ( )**: Represents the baseline average percentage across all polls.
  - **Spline component (f(end_date_num_i))**: Captures temporal trends in polling percentages over time, modeled using a B-spline basis expansion. This allows the trend to adjust flexibly over time, fitting the non-linear trends typically observed in polling data.
  - **Pollster effect (pollster_i)**: A fixed effect to account for systematic differences across pollsters. Due to varying methodologies and sampling practices, some pollsters may produce consistently higher or lower estimates.
  - **State effect (state_i)**: A fixed effect accounting for differences across states. Given the diversity in political landscapes across U.S. states, this term allows for state-specific adjustments, capturing the influence of state-level variability on polling trends. For example, large or politically heterogeneous states may exhibit unique polling behaviors that could impact results.

3.1.2 Priors

- **Intercept (Normal(50, 10))**: Centered around 50%, allowing flexibility given historical polling percentages range from 0% to 100%.
- **Pollster effect prior (Normal(0, 5))**: This moderate prior accounts for typical systematic differences by pollster, usually less than 5% on average.
- **State effect prior (Normal(0, 3))**: Reflects an expected range of moderate variation across states, accounting for demographic or political differences that may influence polling results.
- **Spline basis priors**: Each spline coefficient has a `Normal(0, 5)` prior, allowing moderate variation in temporal trends without overfitting.
- **Error term (Exponential(1))**: This prior allows positive values with most mass around lower standard deviations, yet still accommodates higher values in the presence of greater noise.

3.1.3 Model Implementation

The model is implemented with `rstanarm` and `stan_glm`, Goodrich et al. (2024) with spline terms on (end_date_num) and fixed effects for both pollster and state. This structure enables us to capture time trends while also adjusting for pollster and state-level variability. The Bayesian framework in `rstanarm` allows us to derive credible intervals around predictions, aligning with our focus on uncertainty quantification.

3.1.4 Diagnostics and Validation

Model diagnostics are crucial to ensure reliability and robustness: - **Posterior Predictive Checks**: We use `pp_check` functions to visually compare observed and model-predicted polling percentages. These checks help assess whether the model captures the data patterns effectively. And gives us an interpretable belief system. - **Convergence Diagnostics**: R-hat values close to 1 and trace plots are manually inspected to confirm the convergence of the Markov Chain Monte Carlo (MCMC) sampling. - **Sensitivity Analysis**: We tested several values for the spline degrees of freedom (`df = 4` was chosen) to avoid overfitting. - **Out-of-Sample Validation**: Future work will include a test-train split for validating predictions on unseen data, using metrics like Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) to assess predictive performance.

3.1.5 Alternative Models Considered

We considered the following alternatives: 1. **Non-Spline Linear Trend Model**: A simpler model assuming a linear trend in (end_date_num). However, this approach did not capture the observed non-linear trends over time. 2. **Pollster Only Model (without State Effects)**:

A model without state effects was tested, but we felt that fundamentally state effects are important and should be included

The selected model—using a spline with both pollster and state fixed effects—was chosen for its ability to balance flexibility with robustness, capturing essential polling trends without excessive complexity.

3.1.6 Assumptions and Limitations

This model assumes that polling trends over time follow a relatively smooth pattern and that pollster and state effects are additive and consistent over time. We note that interactions between states, their co movement with each other FiveThirtyEight (2024b) are not included, and that this is a simplistic model.

4 Results

Here we present the results of our analysis of the election cycle. First we look at a cross-sectional summary of the landscape, state-wise polling averages of each party. Then we look at the prediction from our regression of the possible outcomes and evaluate whether the Democrats or Republicans have a higher voter support percentage as per our Bayesian model. We finally look at the posterior beliefs generated by our model and use them to make a conclusion on election outcomes.

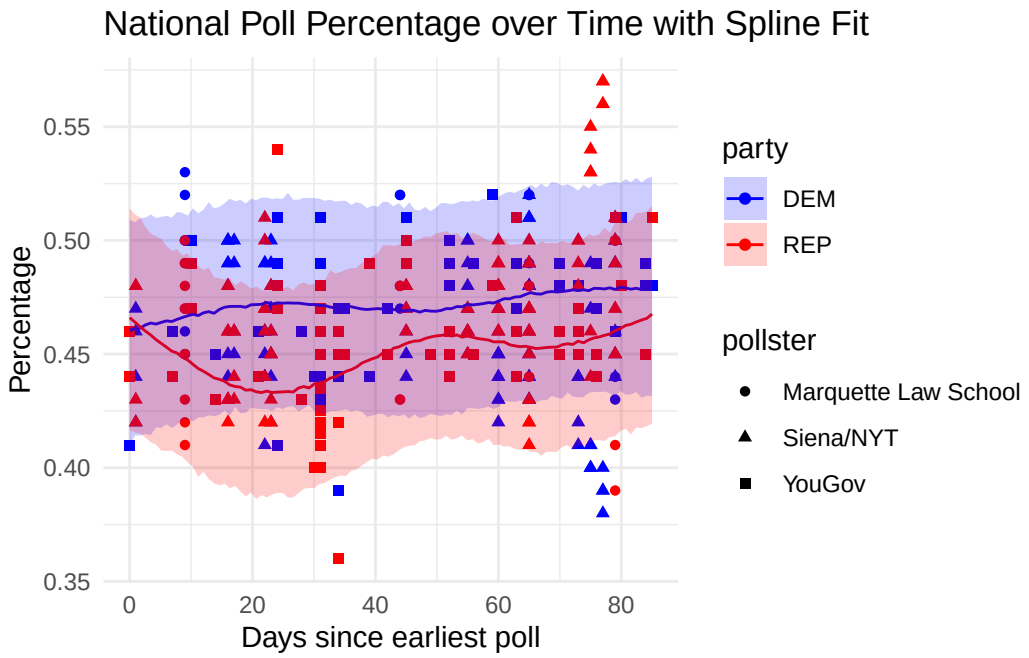
(**cross_summary_stats?**) shows us the cross sectional stats for the polling data. We determine that on average across all time, the Democrats under Harris have a marginal lead. However, given that these polls have confidence intervals of 3% on average, this is well within a margin of error.

Table 3: Summary Statistics of Polling Data

Avg % DEM	Avg % REP	Min % DEM	Max % DEM	Min % REP	Max % REP	Poll Count	Earliest Date	Latest Date
0.466	0.464	0.38	0.53	0.36	0.57	619	Inf	-Inf

4.1 National Polling Trends for Democratic and Republican Parties

First we observe the results from the 'National' state, which are the polls with no state label, that we assumed to be national polls. Our Bayesian spline model allows us to examine trends and quantify the uncertainty.



4.1.1 Bayesian Spline Model

- **Smoothed Prediction Line**: The Bayesian spline model generates the smoothed trend line (blue for Democrats and red for Republicans) on the plot. This line represents the posterior mean, the expected polling percentage over time after accounting for individual poll variability.
- **Uncertainty Interval**: The shaded area around each trend line represents the 95% credible interval derived from the Bayesian model's posterior predictions. This interval provides a measure of uncertainty, highlighting where actual polling values are likely to fall based on the model's posterior distribution.

4.1.2 Key Observations

- **Trend for Democratic Party**: The smoothed trend line for the Democratic Party shows that support for the Democrats has been stable and above that of the Republicans.
- **Trend for Republican Party**: For the Republican Party, the trend line indicates more variability in their support, but a more increasing trend in the most recent past.
- **Comparative Insights**: By comparing the two we find that the Democrats are in the lead but not by a significant margin, and that the Republicans are showing some recent momentum. However, the credibility intervals are so large, they overlap. As such, there is no significant conclusion. Within the credibility intervals are scenarios where the

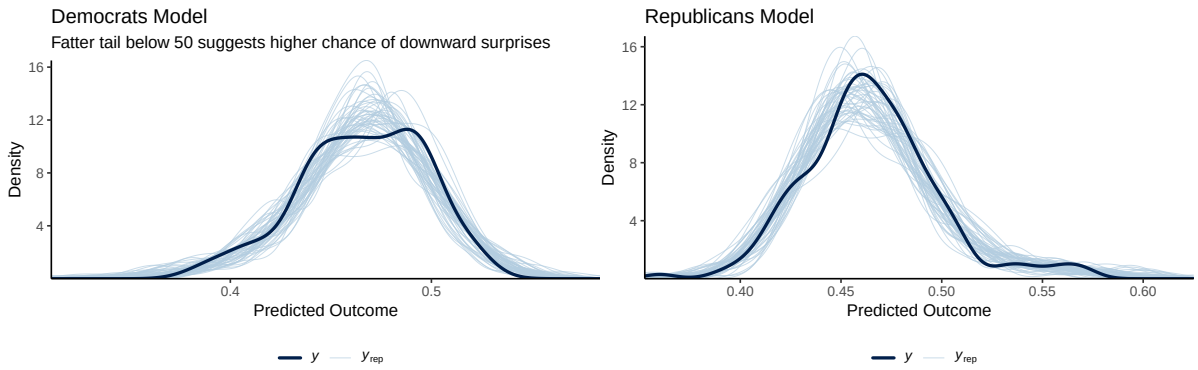
Republicans lose with 42% and the Democrats win with 52%, as well as scenarios where the Republicans win with 51% and the Democrats lose with 42%.

4.2 Posterior predictive check

In (**combined_pp_check?**) we implement a posterior predictive check. Our prior belief is a normal distribution around 50, and our posterior shows how much that has been updated by the data.

We find that the democrat posterior forms a table with a fatter tail below 50, indicating that the uncertainty is more on the lower side, our confidence in the voter support estimate is not strongly concentrated around a single mean value, but decays to the left.

The inverse is true for the Republican posterior where the spike is lower, indicating a higher confidence in their possible voter support, but a fatter right tail, indicating a higher surprise probability.



(a) Comparison of Posterior Predictive Checks: Democrats vs Republicans Models. The Democrat model shows a fat tail below 50, suggesting higher probability of downward surprises. The Republican model has a fatter right tail, indicating a higher probability of upward surprises.

(a) Comparison of Posterior Predictive Checks: Democrats vs Republicans Models. The Democrat model shows a fat tail below 50, suggesting higher probability of downward surprises. The Republican model has a fatter right tail, indicating a higher probability of upward surprises.

A model summary for the Republican and Democrat fits can be found in the Appendix Section **??**.

5 Discussion

5.1 Discussion of Empirical Results

Our results section allows us to make some inferences. First, we see that the Democrats are ahead, and we predict a win for Kamala Harris. However, the time series of the spline shows us that their support has been flat over time, whereas the Republican party has been showing some recent momentum which could contribute to a surprise. This is in-line with some of the literatue, with Cohn (2024) showing how Trump brings his strength later in the election cycle as voters are generally shy to show their support for him.

Our posterior checks show the variability in our uncertainty. Since we do not have normal posteriors, we can make nuanced inferences on the confidences of our prediction. The Democrats do have a higher projected voter support, but the distribution from which this came has a fatter tail on the left, and drops sharply above 49%. Which means the mean is high, but there is a substantial portion of the distribution that is in a lower voter support percentage. Additionally, if we look at the Republican posterior, it has a sharp spike near its mean, which means it is likely that they will get the losing voter support of 45%, but, it has a fat tail that shows that there is an unlikely but possible outcome that they win with a surprise.

We can conclude, that the democrats are strongly favoured to win, but the Republicans do also have a stronger probability of an upward surprise, whereas the democrats have a stronger probabiltiy of a downward surprise.

5.2 Limitations and Weaknesses

Our model, while providing valuable insights into voter support dynamics, has several important limitations:

- **Smoothness Assumption:** The spline regression framework assumes smooth changes in voter preferences over time, potentially missing abrupt shifts after major campaign events or breaking news.
- **Static Pollster Quality:** Pollster quality is treated as a static measure in our model, ignoring that polling methodology and accuracy may vary over the campaign period - pollsters could improve or become less reliable over time.
- **Limited Geographic-Temporal Interaction:** While our model includes state-level effects, it doesn't account for time-state interactions - patterns of changing support may vary across states in ways our current specification cannot capture.
- **National Poll Classification:** We assume that polls with a NaN state value are National level polls, but this is a spurious assumption. We could be introducing state biases to the outcomes.

Despite using only polls with numeric grades above a threshold, variations in pollster methodology and inherent biases in which demographics are sampled can introduce significant uncertainties. While a high numeric grade provides some quality assurance, it cannot guarantee that a pollster's methodology is free from systematic biases in terms of which population groups they reach and survey. This limitation becomes particularly apparent when considering differential voter turnout, as polls may not adequately capture certain demographic segments of the voting population, leading to coverage bias - a significant error that occurs when there is a mismatch between the sampling frame (the population being surveyed) and the actual target population of likely voters (Stantcheva 2023).

A Appendix

A.1 A: Emerson College Polling Methodology Analysis

A.1.1 Overview

Emerson College Polling (ECP) is a nationally recognized, non-partisan polling organization (Emerson College Polling 2024). ECP conducted a national survey from October 30 to November 2, 2024, targeting 1,000 likely voters in the 2024 U.S. presidential election. The poll measured voter preferences between candidates Kamala Harris and Donald Trump, indicating a tie at 49% each, with third party candidates and undecided voters at one percent each (**emerson2024november?**). Appendix A examines the methodology used by ECP for this poll.

A.1.2 Survey Population and Sampling

When conducting a poll, it is crucial to define the target population, sampling frame, and sample in order to provide a clear direction and objective (Whaley 2024).

The target population is the entire group about which we aim to understand and draw conclusions from (Alexander 2023). For this poll, the target population consists of likely voters in the 2024 U.S. presidential election. ECP determines likely voters through a combination of voter history, registration status, and self-reported demographic data (**emerson2024november?**).

Since it is oftentimes infeasible to gather data from the entire target population, sampling frames and samples are utilized. A sampling frame is the list of all units from which samples can be taken (Alexander 2023). In this poll, the sampling frame consists of voters available on Aristotle's database and the online panel provided by CINT (**emerson2024november?**). Aristotle maintains comprehensive voter and consumer data (Aristotle 2024), while CINT operates as a global research marketplace connecting researchers to survey respondents (CINT 2024). The sample consists of 1,000 likely voters selected from this frame.

A.1.3 Data Collection and Sampling Approach

ECP employs a mixed-mode sampling methodology for its polling (Emerson College Polling 2024). When collecting data for its October 30 to November 2, 2024 U.S. Presidential Election poll, three primary methods were used (**emerson2024november?**):

1. MMS-to-web text survey: Respondents receive text messages with custom graphics inviting them to take online Qualtrics survey. Selected randomly from Aristotle voter files.

2. Online panel surveys: CINT panel respondents screened for voter registration and demographics, then directed to survey with quality checks.
3. Interactive Voice Response (IVR): Automated calls to landlines where permitted. Respondents use touch-tone phones. Selected randomly from Aristotle voter files.

This mixed-mode approach offers several advantages, primarily reducing coverage bias by reaching different demographic groups within the target population. Coverage bias occurs when there is a discrepancy between the sampling frame and the target population (Stantcheva 2023). By employing multiple methods, ECP minimizes this bias: MMS-to-web surveys typically capture younger voters, IVR targets older and rural voters who prefer landlines, and online panels engage tech-savvy individuals. A mixed-mode approach can also reduce fieldwork costs by maximizing the use of lower cost methods, and using higher cost methods when necessary (Wilkinson and McTiernan 2020). However, this approach introduces measurement error as each method brings its own biases (Emerson College Polling 2024), potentially increasing the overall margin of error.

A.1.4 Non-response Handling

In polling, non-response occurs when subjects either refuse to participate in a survey or skip specific questions, limiting the survey's representativeness beyond its respondents. To address this issue and ensure representativeness, ECP implements a weighting system based on multiple demographic variables:

- Gender
- Education
- Race
- Age
- Party registration
- Region

These weights are calculated based on 2024 likely voter modeling, adjusting for under- and over-represented groups in the sample (**emerson2024november?**).

A.1.5 Questionnaire Design

The survey's questionnaire demonstrates both strengths and limitations. Its strengths include:

- Clear, unambiguous question formulation
- Efficient format
- Topical relevance focusing on vote preferences and key demographic splits

However, the questionnaire's limitations include:

- Limited exploration of underlying voter motivations
- Potential variation in response quality across different survey modes
- Basic preference capture without deeper attitudinal investigation
- All data collection was conducted in English, potentially limiting representation of non-English speaking voters.

The survey maintains a credibility interval of +/- 3 percentage points, with higher intervals for demographic subsets due to reduced sample sizes.

A.1.6 Conclusion

Emerson College Polling's mixed-mode methodology effectively balances cost with demographic reach. While their weighting system helps ensure representativeness, the multi-mode approach presents trade-offs: reduced coverage bias through diverse voter outreach, but increased measurement error from combining different survey methods (Mora 2011). These methodological considerations are crucial when interpreting the poll's results.

A.2 B: Idealized Survey Methodology – $100K Budget

A.2.1 Introduction

In this appendix, we outline an idealized survey methodology for forecasting U.S. presidential elections within a budget of up to $100,000. The proposed design aims to maximize accuracy and cost-efficiency by strategically selecting sample populations, targeting key demographic groups, and effectively aggregating results.

A.2.2 Sampling Strategy

The ideal survey methodology employs a stratified random sampling approach, given the diversity of the American population as the target group. This method is designed to better analyze the varying effects of different races, age groups, genders, and education levels on voting trends by distributing the sample into multiple subgroups. By categorizing these key variables, this approach ensures more concise and clearer statistical results, which in turn facilitates in-depth analysis. The representative sample is listed below:

- **Sampling Frame**: For the information security, valid entitlement holders registered on the voter list and registered with the Census Bureau will receive the survey.
- **Sampling Method**: Stratified random sampling will be implemented, ensuring better analyze the differential impact of different key factors, including races, ages, genders, and levels of education on voting trends.

- **Sample Size**: The poll is expected to include responses from 10,000 individuals. Assuming nonresponse bias is 10%, we estimate the final adjusted sample size of approximately 9,000.
- **Geographical Distribution**: The survey will be conducted across all 50 states and the District of Columbia. Considering that traditional "red" and "blue" states are less likely to change political alignments, an increased focus will be placed on polling within swing states.

A.2.3 Recruitment Plan

This survey will conduct recruiting in following methods:

- **Online recruitment**: Advertisements and official information about the survey will be disseminated through various search engines (including Google and Firefox) and social media platforms such as X and Instagram. All advertisements will contain direct links to the Google Forms survey.

- **Phone recruitment**: Valid phone numbers will be contacted by staff to conduct the survey using the same set of questions as the online survey. For privacy and security, no conversations will be recorded, and the calling system will automatically dial numbers without disclosing them to staff members.
- **In person recruitment**: A tent will be set up in key high-traffic regions in swing states to conduct survey in-person.
- **Gift incentive**: To increase participation, all respondents will receive a specially designed pin, estimated to cost around $2 each. This incentive is intended to boost engagement and encourage broader participation in the survey.

A.2.4 Budget Allocation

- **Online Recruitment Cost**: $20,000 allocated for advertising across various platforms (search engines and social media).
- **Phone Recruitment Cost**: $20,000 allocated for hiring staff to conduct phone surveys and cover the associated data usage charges.
- **In-Person Recruitment Cost**: $22,000 allocated for hiring field staff and setting up in-person recruitment locations.
- **Gift Cost**: $20,000 allocated for the total value of participant gifts
- **Data Analysis and Quality Control Cost**: $6,000 allocated for employing technical staff, purchasing software licenses, and utilizing required tools.
- **Administrative Cost**: $2,000 allocated for overall management, including miscellaneous expenses and logistical support.

Total Budget: $100,000

A.2.5 Survey Design

This survey will only collect data that are accurate and unbigoted. Meanwhile, personal information will not be contained.

- **Close-ended Question**: Usage of multiple choice can make sure the answer is straightforward and easier for analysis.
- **Response Option**: Most questions will only contain option from A, B and "prefer not to say", all the option will be a neutral response.
- **Insensitive Information**: This survey will only contain personal information only and merely age, gender, race, education background, income, living state.

A.2.6 Data Process

To ensure data we collected are valid and significant, we will conduct following methods to test our data.

- **One-times Submission**: For online surveys, only one submission will be allowed per unique IP address to prevent multiple entries that could skew results. The same principle will apply to phone recruitment, ensuring each number is only surveyed once.
- **Missing Data**: During the data cleaning process, responses with missing values (N/A entries) will be excluded to maintain the integrity and reliability of the analysis.
- **Swing States Data Process**: Data collected from swing states will be assigned a slightly higher weight in the estimation process, given their critical influence on election outcomes.
- **Stratified Analysis**: To gain insights into voter preferences within different demographic subgroups, stratified analyses will be conducted, focusing both on overall trends and trends within specific subgroups.
- **Poll Aggregation Approach**: The aggregated data will be weighted based on sample size (with larger samples receiving greater weight), and past reputable polls will be incorporated, giving priority to highly rated pollsters to strengthen the overall analysis.

A.2.7 Sample Survey

Google Forms will be used to collect data efficiently and securely, ensuring no information is leaked. The survey will cover aspects such as gender, race, demographics, education, first-choice candidate, and views on the current United States.

A complete sample survey could be found in the following link https://forms.gle/1jrQP7bPqntJ9N2s9 or by clicking Sample survey.

Proposed Survey Questions:

1. What sex were you assigned at birth, or on your original birth certificate?

- Male
- Female

2. How do you currently describe your gender?

- Male
- Female
- Transgender
- Non-binary
- other

3. What is your racial self-identification?

- White
- Black
- Asian
- Indigenous
- Prefer not to say
- Other

4. What is your current age?

- Under 18
- 30-45
- 46-51
- 52-66
- over 66

5. What is your highest level of education?

- High school or Under
- Undergraduate
- Graduate _ Doctor

6. Which candidate would you vote for in U.S. 2024 Presidential Election?

- Donald Trump (Republic)
- Kamala Harris (Democrat)
- Other

7. Would you vote for the 2024 U.S. Presidential Election?

- Yes
- No