Prediction about 2024 U.S. Presidential Election Outcome Using Bayesian Modeling*

Vote Percentage of Donald Trump over Time

Yunkai Gu Anqi Xu Yitong Wang

November 1, 2024

This study aims to forecast the outcome of the 2024 U.S. Presidential Election using aggregated polling data and a Bayesian modeling approach. The study also includes an analysis of a selected pollster's methodology and an idealized survey design for forecasting elections within a limited budget.

Table of contents

*Code and data are available at: https://github.com/Kylie309/2024-US-election-prediction.

1 Introduction

The U.S. presidential election is a critical event that attracts substantial public and media interest. Various tools and methodologies have then been used for making predictions about the final electoral outcome.
Curious about the election outcome as well, this paper employs the "poll-of-polls" approach to make predictions, which aggregates multiple polls to minimize biases and enhance prediction accuracy. Two Bayesian models are built to forecast the winner of the 2024 U.S. Presidential Election. In addition, in-depth analysis on certain pollster's methodology is also included, discussing their sampling approach, strengths, and weaknesses.

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

The remainder of this paper is structured as follows. Section 2 introduces the overview of the data (Section 2.1), measurement (Section 2.2), as well as explanations of outcome (Section 2.3) and predictor variables (Section 2.4). Section 3 explains the modeling process in detail, including procedure of model set-up (Section 3.1), model justification (Section 3.2) and model validation (Section 3.3). Then, Section 4 presents the prediction outcome, and Section 5 discusses the results and models. Appendix involves three parts: Appendix A presents methodology overview and evaluation of one certain pollster: Emerson; Appendix B provides detailed idealized methodology and survey design for the poll; Appendix C presents additional details during modeling process.

2 Data

2.1 Overview

The analysis uses the dataset of national presidential general polls from FiveThirtyEight (FiveThirtyEight 2024). Following Alexander (2023), we consider to make predictions about the election outcome based on the data.

The analyses presented in this paper were conducted using R programming language (R Core Team 2023). The `tidyverse` packages (Wickham et al. 2019) were used in the process of data simulation, testing beforehand. After the original raw data was downloaded by using `tidyverse` package (Wickham et al. 2019), data cleaning process was done by using `tidyverse` package (Wickham et al. 2019), `lubridate` package (Grolemund and Wickham 2011), and `arrow` package (Richardson et al. 2024). Then, models were constructed using `tidyverse` package (Wickham et al. 2019), `lubridate` package (Grolemund and Wickham 2011), `rstanarm` (Goodrich et al. 2022) package, and `splines` package (R Core Team 2024). The model results are then presented by `modelsummary` (Arel-Bundock 2022) package, and graphs were made with `ggplot2` package (Wickham 2016).

2.2 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

2.3 Outcome variables

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

Plot basic relationship

2.4 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

The models constructed in the paper are Bayesian linear regression models designed to predict the percentage of support for candidate Donald Trump (referred to as pct) based on the number of days since May 5, 2024, i.e. half year before the election takes place (referred to as end_date_num).

Two Bayesian models are constructed separately.

Firstly, a Bayesian model with spline using response variable as the percentage of support (pct) and predictor variable as the number of days (end_date_num) is constructed to make predictions for vote percentage for Trump nationwide.

Then, a similar Bayesian model with the same response and predictor variables, but with an additional predictor "state", which contains information of the state in which the polling occurred, is constructed to include "state" as a fixed effect in the model with spline.

The goal of our modeling strategy is twofold.

Firstly, we aim to understand the overall trend in the percentage of support for Trump over time from a national perspective. This is accomplished through the first Bayesian model, which uses the number of days (end_date_num) as the sole predictor. This model allows us to investigate how support changes as we approach the election, providing insights into temporal dynamics in voter preferences nationwide.

Secondly, we seek to explore the impact of state-level differences on voter support through the second Bayesian model, which incorporates state as an additional predictor alongside the number of days. By treating state as a categorical predictor variable, this model enables us to assess how the percentage of support varies across different states overtime. This approach helps to understand regional influences on voter behavior, potentially uncovering important variations that the first model may overlook.

Following sections define, explain and justify each model and the variables, as well as discuss underlying assumptions, potential limitations, software used to implement the model, and evidence of model validation and checking.

Background details and diagnostics are included in Appendix C.

3.1 Model set-up

The Bayesian models were implemented using the R programming language (R Core Team 2023), specifically utilizing the `rstanarm` package of Goodrich et al. (2022). This package provides an interface for fitting Bayesian regression models using Stan, and the models are fit using the package.

3.1.1 Bayesian model with spline for vote percentage nationwide

Define y_i as the percentage of the vote that candidate Donald Trump received in the poll for each observation i, which is the response variable. Then define n_i as the number of days since May 5, 2024, i.e. half year before the election takes place.

The Bayesian model could be defined by the following mathematical expressions:

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + f(n_i) \tag{2}$$
$$f(n_i) = \sum_{k=1}^{K} \gamma_k B_k(n_i) \tag{3}$$
$$\alpha \sim \text{Normal}(50, 10^2) \tag{4}$$
$$\gamma_k \sim \text{Normal}(0, 5^2) \quad \text{for } k = 1, 2, \ldots, K \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

Combining all the components, the complete model can be expressed as:

$$y_i \sim \text{Normal}\left(\beta_0 + \sum_{k=1}^{4} \gamma_k B_k(n_i), \sigma^2\right)$$

This Bayesian model involves the vote percentage of Trump (y_i, pct) as a function of the number of days since half year before the election (n_i, end_date_num), modeled using natural splines.

Formula (1) specifies that the response variable y_i is normally distributed, where μ_i is the predicted mean percentage of support for the i-th observation, and σ is the standard deviation of the normal distribution, representing the variability in the observed support percentages.

Formula (2) specifies that the mean μ_i is modeled as a linear combination of an intercept α and a function of n_i, represented by natural spline f. The intercept of α represents the expected support when all predictors are zero.

Formula (3) specifies that a natural spline function f is applied to variable n_i, which represents the non-linear effect of time, and allows the model to flexibly fit the relationship between time and support. In this function, K is the number of basis functions, $B_k(n_i)$ are the basis functions of natural spline, and γ_k are the coefficients for these basis function.

Then, formula (4), (5) and (6) presents the priors:

Formula (4) defines the prior normal distributions for intercept α, with mean equals 50 and variance equals 10^2. This prior for intercept suggests that the average vote percentage is expected to be around 50% with some uncertainty.

Formula (5) defines the prior normal distributions for parameter γ_k, with mean equals 0 and variance equals 5^2. This prior for splines indicates prior belief that the effects of number of days (n_i, end_data_num) are likely to be small but allowing for flexibility.

Formula (6) assigns a prior to the standard deviation σ, where 1 is the rate parameter of the exponential distribution. Using such prior for variance reflects the belief that smaller values are more likely but allows for larger variations if needed.

3.1.2 Bayesian model with spline for vote percentage and state as fixed effect

Similarly as before, define y_i as the percentage of the vote that candidate Donald Trump received in the poll for each observation i, which is the response variable. Then define n_i as the number of days since May 5, 2024, i.e. half year before the election takes place.

The Bayesian model could be defined by the following mathematical expressions:

$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (7)\\
\mu_i &= \alpha + f(n_i) + \beta_{\text{state}} \cdot \text{state}_i && (8)\\
f(n_i) &= \sum_{k=1}^{K} \gamma_k B_k(n_i) && (9)\\
\alpha &\sim \text{Normal}(50, 10^2) && (10)\\
\gamma_k &\sim \text{Normal}(0, 5^2) \quad \text{for } k = 1, 2, \ldots, K && (11)\\
\sigma &\sim \text{Exponential}(1) && (12)
\end{aligned}
$$

Combining all the components, the complete model can be expressed as:

$$
y_i \sim \text{Normal}\left(\beta_0 + \sum_{k=1}^{4} \gamma_k B_k(n_i) + + \beta_{\text{state}} \cdot \text{state}_i, \sigma^2\right)
$$

This Bayesian model involves the vote percentage of Trump (y_i, pct) as a function of the number of days since half year before the election (n_i, end_date_num), modeled using natural splines. The mean is modeled as a linear combination of an intercept, a natural spline function of n_i, and a fixed effect for the state.

Formula (1) specifies that the response variable y_i is normally distributed, where μ_i is the predicted mean percentage of support for the i-th observation, and σ is the standard deviation of the normal distribution, representing the variability in the observed support percentages.

Formula (2) specifies the mean structure. The linear combination is composed of three parts: a parameter α; a function $f(n_i)$ which represents the non-linear effect of the predictor n_i, modeled using natural splines; and the coefficient β_{state} for the categorical variable state, which captures the average difference in percentage across the different states. Treating state effects as levels (or random effects) may also include weakly informative priors to avoid overfitting and encourage regularization.

Formula (3) specifies that a natural spline function f is applied to variable n_i, which represents the non-linear effect of time, and allows the model to flexibly fit the relationship between time and support. In this function, K is the number of basis functions, $B_k(n_i)$ are the basis functions of natural spline, and γ_k are the coefficients for these basis function.

Then, formula (4), (5) and (6) presents the priors:

Formula (4) defines the prior normal distributions for intercept α, with mean equals 50 and variance equals 10^2. This prior for intercept suggests that the average vote percentage is expected to be around 50% with some uncertainty.

Formula (5) defines the prior normal distributions for parameter γ_k, with mean equals 0 and variance equals 5^2. This prior for splines indicates prior belief that the effects of number of days (n_i, end_data_num) are likely to be small but allowing for flexibility.

Formula (6) assigns a prior to the standard deviation σ, where 1 is the rate parameter of the exponential distribution. Using such prior for variance reflects the belief that smaller values are more likely but allows for larger variations if needed.

3.1.3 Assumptions of the Bayesian models

Underlying assumptions for two Bayesian models are discussed below:

1. Normality of residuals:

The models assume that the residuals of the model (the differences between observed and predicted values) follow a normal distribution. This assumption is important for the validity of inference, and violation of this assumption can lead to incorrect inferences and unreliable credible intervals.

2. Homoscedasticity:

The variance of residuals is assumed to be constant across all combinations of the predictors. If the residuals exhibit heteroscedasticity (i.e., non-constant variance), this may lead to inefficient estimates and biased conclusions about the relationships modeled.

3. Independence of observations:

It is assumed that the observations are independent of each other. If there are correlated observations (e.g., repeated measures from the same individual), this could lead to biased estimates.

4. Functional form:

The relationship between the response variable and the predictor variable is assumed to be modeled appropriately, using natural splines to allow for flexibility. If the functional form does not adequately capture the underlying relationship, this can lead to biased estimates. The choice of degrees of freedom in the spline affects how well the model fits the data.

5. Prior specification:

The prior distributions for the parameters (i.e. intercept and coefficients) are assumed specified based on prior beliefs about their likely values. The choice of priors can influence the posterior estimates, reflecting reasonable beliefs about the parameters.

3.2 Model justification

The Bayesian model with natural splines for predicting election outcomes with respect to the date is the most appropriate model because of several reasons. Firstly, the electoral dynamics for vote percentage could be highly complex and therefore non-linear. Various factors occurring before the election date could largely affect the outcomes. By employing natural splines, the model allows for the flexibility to capture these non-linear trends, presenting how support for a candidate evolves as the election date approaches, thus providing predictions about vote percentage.

Moreover, the Bayesian framework inherently accommodates uncertainty in parameter estimates, which is particularly valuable in political polling where outcomes can be unpredictable. The ability to incorporate prior information and update beliefs based on observed data enhances the robustness of the model, making it be the most appropriate model.

The above final model decision is made after considering several alternative models and variants. One alternative is a simple linear regression model (SLR), which assumes a constant linear relationship between the predictor (date) and the response (percentage support). Although the model is extremely straightforward and easy to interpret, it is far too simple for representing trends of and conducting predictions about election outcomes. SLR also applies plenty of strict assumptions to the population, the predictor and response variables, and violation of these assumptions, which is highly possible under this scenario, could lead to biased results.

Another option was to use a generalized linear model (GLM). While GLMs can handle various response distributions, they still rely on a linear assumption for the predictors, which may not adequately reflect the evolving dynamics of voter sentiment over time. Moreover, GLMs do not provide the same level of flexibility in modeling non-linear relationships as splines do.

After setting up the Bayesian model with splines, we expect a positive relationship between the percentage vote of Donald Trump and time passed since half year before the election. In other words, we anticipate that as the election date approaches, Donald Trump's support would increase. This expectation arises from the increasing support from minority voters, particularly from black voters (Post 2024b). Also, all seven battleground states (Pennsylvania, Wisconsin, Michigan, Arizona, Nevada, Georgia, North Carolina) are expected to shift to support Trump (Post 2024c). The vice-presidential debates that took place could be the major reason, which explain Trump's appeal of 'Unabashed machismo vibe', motivating especially young voters to support him (Post 2024a).

3.3 Model Validation

During the modeling process, posterior predictive checks (PPC) is used. PPC is the comparison between what the fitted model predicts and the actual observed data, which validates whether the fitted model is compatible with the observed data. The aim is to detect if the model is inadequate to describe the data (Andrés López-Sepulcre ORCID iD 2024).

In the PPC plots, the x-axis represents the values of the response variable (in this case, the vote percentage of Donald Trump), and the y-axis represents the density or probability of those values.

The dark black line presents the density of the observed data (y), which is a smoothed estimate of the distribution of actual concentrated data points. On the other hand, the light blue lines present the density of the replicated data (y_{rep}), which is generated from the posterior predictive distribution of the model.

Whether the black line matches well with the blue lines indicates how well the model fits the observed data. If the black line aligns well with the blue lines, it suggests that the model is performing well in capturing the actual data distribution, and vice versa.

Figure 1 shows the posterior predictive checks for the first model, which predicts Trump's percentage vote over time in the national perspective.

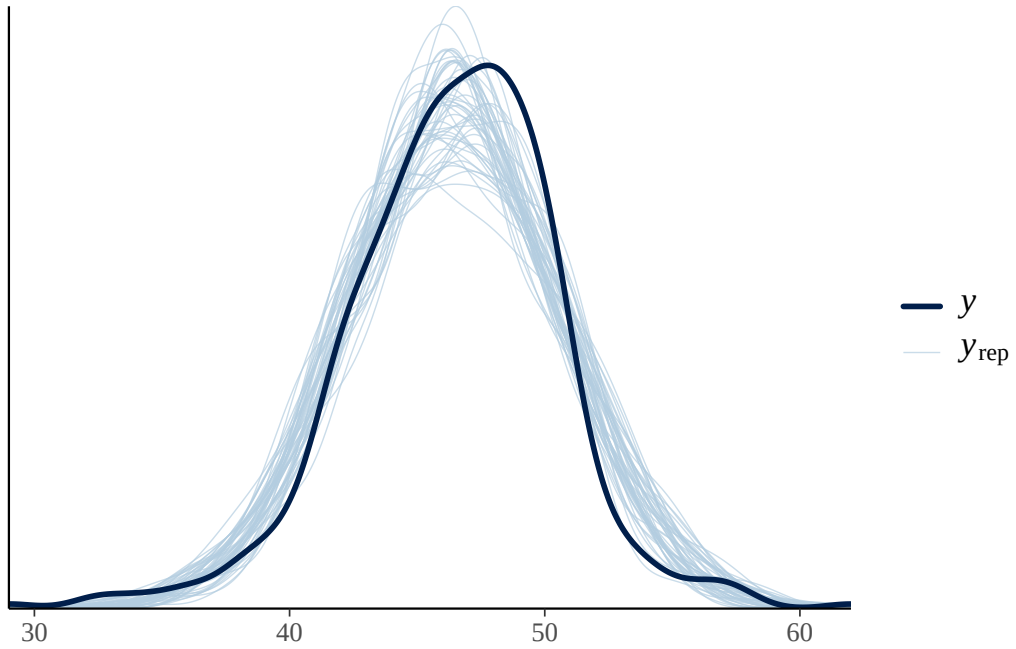

Figure 1: Posterior predictive checks for model 1 (national)

In Figure 1, the black line basically follows the trend of the light blue lines, indicating the observed data consistently falling within the range of simulated data. This suggests that the model provides a good fit, effectively predicting the actual trend of the data.

Similarly, Figure 2 shows the posterior predictive checks for the second model, which predicts Trump's percentage vote over time by state.

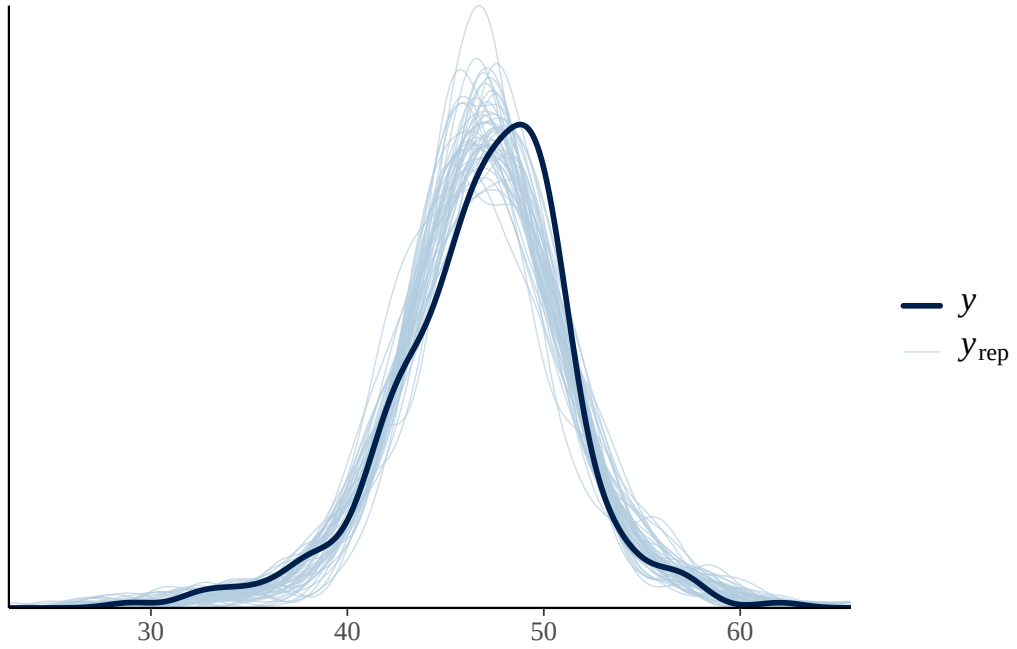

Figure 2: Posterior predictive checks for model 2 (by state)

In Figure 2, Although the black line has minor deviation away from the light blue lines, and a part of the black line falls outside the range of simulated lines, the majority part of the black line matches with the light blue lines.

This suggests that the model basically provides a good fit, effectively predicting the actual trend of the data. But the fit is not that well compared to model 1, since there is a larger deviation in this plot than the previous PPC plot for model 1.

4 Results

4.1 Overview

The results from two Bayesian models are summarized in Table 1.

Table 1: Model Results of Trump Percentage Vote based on Date and State

	Model 1	Model 2
(Intercept)	44.27	45.59
ns(end_date_num, df = 5)1	1.19	0.21
ns(end_date_num, df = 5)2	1.60	2.92
ns(end_date_num, df = 5)3	3.10	1.85
ns(end_date_num, df = 5)4	5.58	6.28
ns(end_date_num, df = 5)5	3.17	2.94
stateCalifornia		−9.85
stateFlorida		3.62
stateGeorgia		0.24
stateIndiana		7.34
stateMaryland		−12.69
stateMassachusetts		−15.02
stateMichigan		−1.79
stateMinnesota		−1.91
stateMissouri		6.07
stateMontana		7.55
stateNebraska CD-2		−6.10
stateNevada		−0.36
stateNew Hampshire		−5.26
stateNew Mexico		−3.40
stateNew York		−5.42
stateNorth Carolina		−0.25
stateOhio		1.75
statePennsylvania		−1.42
stateSouth Dakota		9.12
stateTexas		1.24
stateVirginia		−4.38
stateWisconsin		−1.45
Num.Obs.	472	307
R2	0.089	0.603
R2 Adj.	0.068	0.557
Log.Lik.	−1310.219	−740.269
ELPD	−1316.8	−768.5
ELPD s.e.	21.8	11.9
LOOIC	2633.7	1537.0
LOOIC s.e.	43.6	23.9
WAIC	2633.7	1533.1
RMSE	3.87	2.67

Table 1 shows positive coefficients for natural splines for both two models (nationwide and state-specific). This suggests that as the election date approaches (i.e. more days passed since half year before election), the percentage vote increases in a non-linear relationship.

Model 2 additionally discusses state effects, including coefficients for each state respectively. FOr instance, state Massachusetts has the smallest, negative coefficient (-15.02), indicating that the response variable of vote percentage is significantly lower in California compared to the reference state. On the other hand, state South Dakota has the largest, positive coefficient (9.12), indicating a relatively higher response for the state.

4.2 Predictions and spline fit for vote percentage nationally

Figure 3 demonstrates the results of the first Bayesian model, including predictions and spline fit for vote percentage of Donald Trump in the perspective of the whole nation.

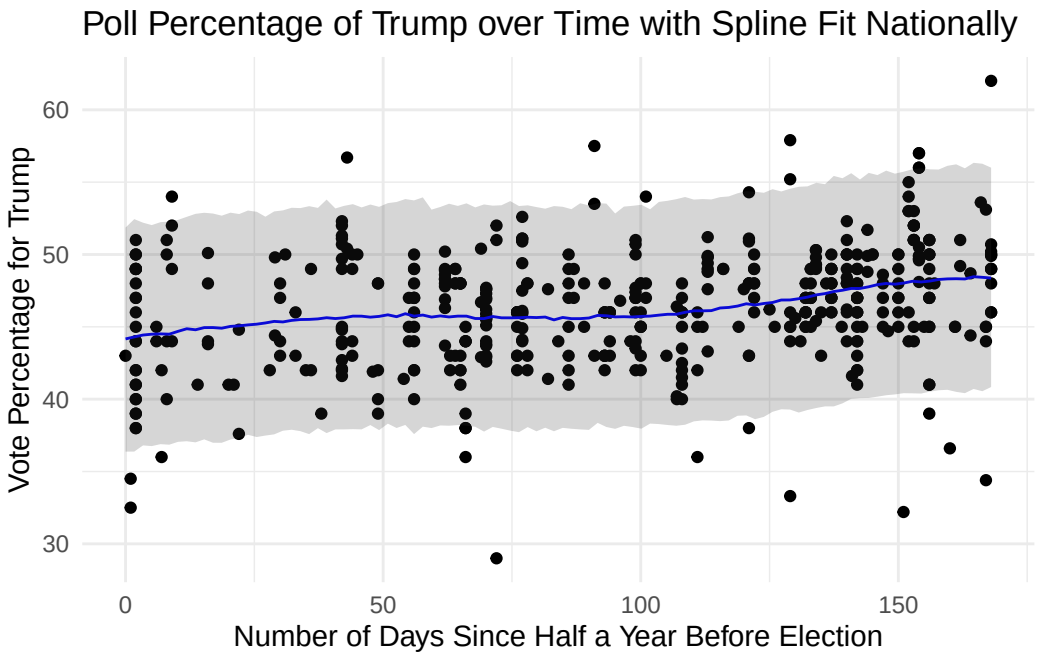

Figure 3: Predict Posterior Draws and Spline Fit for Vote Percentage Nationally

The blue line drawn through the scatter points in Figure 3 indicates an upward trend in the data. It represents that as the time approaches the election (i.e. more day passed since half year before election), Trump's vote percentage is expected to be higher. Though the intercept is below 45%, indicating that half year ago Trump's support was lower than Harris, the percentage vote has climbed to nearly 50% with a steady trend over time. The black spots on the graph are individual observations, and the shaded gray area represents the confidence interval for the predicted values.

4.3 Prediction and spline fit for vote percentage by state

Figure 4 demonstrates the results of the second Bayesian model, including predictions and spline fit for vote percentage of Donald Trump in each state respectively.

The blue line drawn through the scatter points in each plot in Figure 4 presents the trend of Trump's vote percentage over time. For all 23 states listed, the blue lines indicate upward relationship between the response and predictor variables. For some states such as Indiana, Misourri, Montana and Ohio, the vote percentage for Trump is higher than 50% for most of the time and is stably increasing. For other states, though the support is lower half year ago, the upward-sloping trend indicates final prediction about vote percentage to be potentially higher when the election date comes.

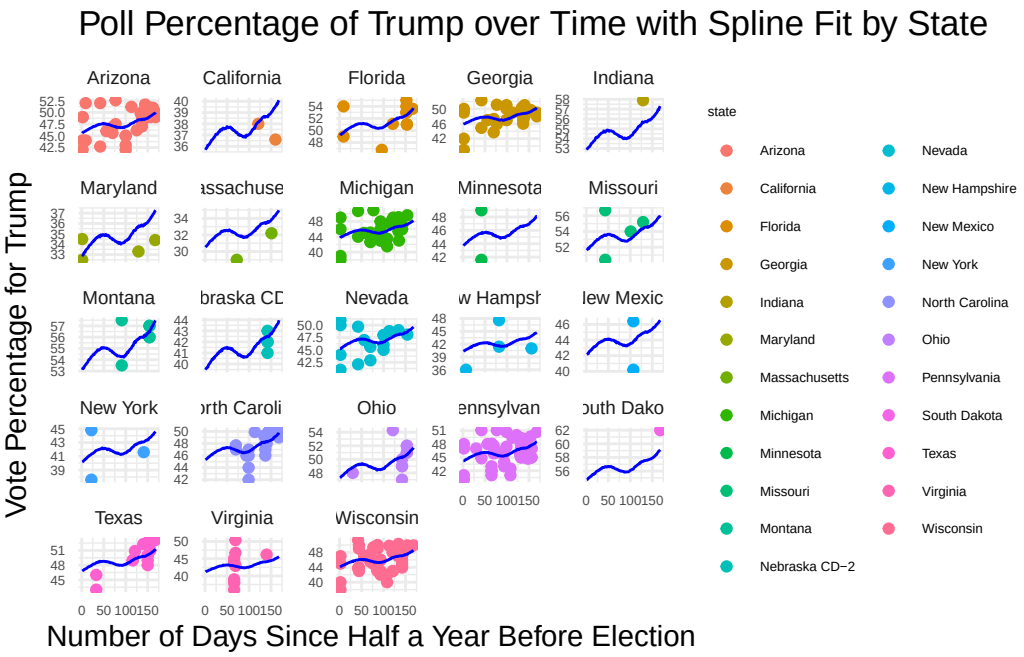

Figure 4: Predict Posterior Draws and Spline Fit for Vote Percentage in Each State

5 Discussion

5.1 First discussion point

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 Second discussion point

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

5.3 Third discussion point

5.4 Weaknesses

Potential limitations of the Bayesian models are described below:

1. Overfitting with complex models:

Using splines with high degrees of freedom may lead to overfitting, where the model captures noise rather than the underlying trend. Overfitting reduces the model's generalizability to new data.

2. Missing data:

Some data included to make predictions nationwide using the first Bayesian model have missing values in the predictor "state", and is cleaned out during the process of making predictions about how the vote percentage varies across different states with the second Bayesian model. These missing values in key variables could lead to biases in the model.

3. Assumptions violation:

If the assumptions and priors of the models are violated, situations where the models may not be appropriate could occur. For instance, if the population of the response variable (pct) exhibits a skewed distribution instead of the normal distribution assumed, the models would be inappropriate and provide biased results.

Appendix

A Pollster methodology overview and evaluation: Emerson

A.1 Background of Emerson College Polling

Emerson College Polling is a famous U.S. polling organization as known for its precise and timely predictions, particularly in political forecasting such as the U.S, election. Emerson uses multiple methods approach, combining both traditional and modern techniques to capture a representative outcome of the election. Emerson College uses a mix of landline calls, text-to-web links, and online panel surveys to reach a broad range of people. This approach helps them collect decisions from a diverse group of likely voters.

A.2 Population, Frame, and Sample

1. Population: Emerson's target population consists of likely U.S. voters, determined based on their past voting records, registration status, and stated intention to vote. For their 2024 surveys, they specifically target people who most likely to participate in upcoming elections, making adjustments to closely model the anticipated voter's tendency
2. Frame: Emerson's sampling frame includes U.S. census parameters, voter registration and exit polls(pre validated online panels). These lists are accurate to reflect key demographics, such as age, gender, race, education, and geographic region. By using stratified sampling and comprehensive voter files, Emerson ensures broad coverage across significant voting populations, including those who might hard to reach, such as rural or younger voters https://emersoncollegepolling.com/october-2024-national-poll-harris-50-trump-48/
3. Sample: Emerson College uses stratified sampling in their methodology to set quotas that match the actual demographic breakdown of the U.S. electorate. They use a combination of probability and non-probability sampling techniques, primarily from an online panel, MMS-to-web(text-to-web) surveys, and landline calls via Interactive Voice Response (IVR). For example, their sample size for national surveys typically consists of 1,000 likely voters, with each sample calibrated to reflect key demographic features.

A.3 Sample Recruitment Methods

1. Cell Phone MSS-to-Web: Emerson applies a text-to-web approach, sending potential respondents a survey link through MMS, this method is especially effective in reaching younger or mobile-first voters, they are more likely to response by one click rather than receiving an half-hour long call.

2. Landline IVR: IVR calls are used to reach older demographics or rural population who may be less engaged online or not familiar with new generation technology. The traditional method such as the landline calls are easier to accepted by them, and increases response rates among older adults and those in rural areas, which ensuring balanced representation across age groups.
3. Online Panels: Emerson also utilizes verified online panels, where respondents' eligibility is cross-checked with voter file data, in order to exclude not potential voter's response, this step reinforces sample accuracy.

These recruitment methods help Emerson reach a wide audience, though self-selection bias may arise, particularly with online panel participants, as individuals joining these panels are often more politically engaged.

A.4 Sampling Approach's Trade-offs

Emerson's sampling approach uses stratification to set quotas based on demographic factors, such as age, race, gender, education, and location. Post-stratification weighting further adjusts these quotas to align the sample with Census and voter registration data.

A.4.1 Advantages

1. Inclusivity: By applying a combination of recruitment methods, Emerson reaches a range of voter groups across demographics, which avoid the incomplete context representation bias.
2. Budget Saving: Digital methods such as text-to-web and online survey are generally less expensive than the traditional methods (landline call).
3. Effectiveness: Using text-to-web and online panels allows a quick receiving of the answers, for rapid adjustments and data collection.

A.4.2 Disadvantages

1. Non-Probability Limitations: Non-probability sampling means not all individuals have an equal chance of being selected, in some of Emerson's methods such as landline call, targeted people are not randomly choose, which can increase potential bias.
2. Access Bias: Voters without internet or mobile access may be underrepresented, particularly among lower-income or rural demographics, which can increase the incomplete representation bias.

A.5 Handling Non-Response

To address non-response bias, Emerson applies demographic weighting to underrepresented groups, such as younger voters or certain racial demographics, they put more weight for these groups. Emerson occasionally provides incentives to respondents, which encourages survey completion. This is especially effective in boosting participation rates for online panel members, who may be less likely to complete surveys without additional motivation. However, people with lower political engagement levels may still be less likely to respond, which could impact representativeness.

A.6 Questionnaire Design's Advantages and Disadvantages

A.6.1 Advantages

1. Clarity and Style of Question: Emerson focuses on straightforward, unbiased questions to ensure respondents understand each item consistently, closed-ended questions make voters are easy to give the answer without too much thinking, which decrease the drops-out.
2. Relevance and Adaptability: Questions are updated regularly to reflect current political events, enhancing the survey's relevance and accuracy.

A.6.2 Disadvantages

1. Simplicity: Online surveys typically rely on multiple-choice or straightforward questions, which may limit deeper insights into complex opinions or voter's tendency.
2. Question Order Bias: The order of questions presented can unintentionally influence voter responses. For example, asking participants to rate their favorability towards Kamala Harris before asking about other candidates could mislead respondents, subtly framing how they perceive subsequent individuals. This can lead to "order effects," in polls, order effects can create a bias, especially if the initial question involves a polarizing candidate, it may cause strong emotions that influence answers to following questions.
3. Potential Response Fatigue: Frequent participation in polls may lead to rushed responses, impacting quality.

A.7 Conclusion

Emerson College's approach combines traditional and digital methodologies to reach a comprehensive and representative sample of U.S. likely voters. Even the inclusion of diverse communication methods enhances accessibility and speed, challenges such as self-selection bias and non-probability sampling limitations still appear. Despite these, Emerson's effective stratification and weighting strategies have made its poll significant in prediction of U.S. elections.

B Idealized Methodology and Survey

B.1 Overview

The idealized survey methodology is designed to forecast the outcome of the upcoming U.S. presidential election accurately. This approach will focus on a robust sampling strategy, targeted recruitment, comprehensive data validation, and effective poll aggregation. A budget of $100,000 will be allocated to ensure that each aspect of the methodology is fully supported and that the resulting data is accurate and representative.

B.2 Sampling Approach

B.2.1 Target Population

The target population consists of U.S. registered voters eligible to vote in the upcoming presidential election. This population includes diverse demographic groups, varying by age, race, gender, geography, and political affiliation, all of which are critical to accurate representation.

B.2.2 Sampling Frame

To build a representative sampling frame, recent national voter registration data from reliable sources, such as the U.S. Election Assistance Commission, will serve as the foundation. This data will be enhanced by voter profile databases from third-party aggregators like Catalist, which consolidate demographic, geographic, and political affiliation data. To ensure the sample reflects the diversity of the U.S. electorate, quotas will be set based on demographic data from the U.S. Census and Bureau of Labor Statistics, including age, race, gender, income, and urban-rural distribution.

B.2.3 Sample Size

We will take a sample of at least 3,000 respondents to balance statistical accuracy with budget constraints.

B.2.4 Stratified Random Sampling

To achieve the goal of accurate and representative sample, we will apply stratified random sampling by using key demographic categories to divide the sample into strata that reflect the U.S. voting population's diversity. Stratify as followed:

- Age: Use age brackets (18-29, 30-44, 45-64, 65+) based on voter turnout trends, as younger age groups are often underrepresented.
- Gender: Stratify by gender to capture any potential disparities in voting behavior.
- Race and Ethnicity: Reflect population proportions of racial and ethnic groups, as provided by census data.
- Geography: Stratify by state and, within each state, by rural, suburban, and urban classifications to capture regional voting patterns.
- Political Affiliation: Where available, include political party affiliation, such as Democrat, Republican and Independent, to account for differing levels of partisan engagement.

Within the division, we will randomly select respondents within each stratum to match the proportions of these demographic categories in the general population.

B.2.5 Trade-offs

This approach minimizes bias but can be costly due to increased recruitment needs in underrepresented groups. Random sampling may also limit the ability to reach specific subsets of voters.

B.3 Data Validation

As data quality is imperative to reach a reliable prediction, we will apply the following examinations to assure the validation:

- Verification of Responses: Employ CAPTCHA for online responses to limit bot entries and cross-check voter registration data to confirm respondent eligibility.

- Consistency Checks: Highlight and review responses with inconsistencies, such as duplicate IP addresses or extreme response times, and validate through follow-up when feasible.
- Data Cleaning: Remove incomplete or invalid responses and use imputation techniques for minor missing data if the omission rate is low.

B.4 Poll Aggregation Methodology

We will pick the poll-of-polls approach aggregates results from multiple polling sources to produce a more stable and representative forecast of the U.S. presidential election. This method helps to mitigate individual poll biases, smooth out fluctuations, and capture a broader picture of voter sentiment. Polls will be weighted by the following factors:

- Sample Size: Larger sample size polls will receive greater weight to reflect their higher statistical reliability.
- Recency Emphasis: The most recent polls are given greater influence in the aggregation to reflect real-time changes in voter preferences. Time-weighting is applied to each poll within the last month, with those closest to the current date contributing most to the overall estimate.
- Outlier Identification: Identify outlier polls by analyzing their deviation from the rolling average of other polls. Polls with extreme deviation from the average will be assessed individually to determine if they represent unique insights or sampling anomalies.

B.5 Survey

This survey will be conducted using Google Forms, which is an effective platform for data collection. The survey can be accessed by the link, Google Form Survey.

B.5.1 Survey Structure

Title:

2024 U.S. Presidential Election Forecast Survey

Introduction:

We appreciate your participation in this survey, which aims to forecast the outcome of the 2024 U.S. Presidential election. Your responses are essential for our research.

Please note:

- Your answers will be treated with complete confidentiality.

- Participation in this survey is voluntary.
- We encourage you to provide honest and thoughtful responses.
- The survey is estimated to take about 5 minutes to complete.
- If you have any questions or concerns, feel free to reach out to our research team at anjojoo.xu@mail.utoronto.ca (Angel Xu, Yunkai Gu, Yitong Wang).

Thank you for your valuable contribution! As a token of our appreciation, each participant will receive $5 upon completion of the survey.

Section 1: Eligibility

Are you a U.S. citizen?

- Yes
- No (End Survey)

Do you meet your state's residency requirements?

- Yes
- No (End Survey)

Will you be 18 years old or elder by the Election Day?

- Yes
- No (End Survey)

Are you registered to vote by your state's voter registration deadline. (North Dakota does not require voter registration.)

- Yes
- No
- Plan to register later
- Maybe

Section 2: Demographics

The following questions will help us understand the background characteristics of our respondents.

Which age group do you belong to?

- 18-29 years

- 30-44 years
- 45-64 years
- 65 years or older

What is your gender? - Male

- Female
- Other
- Prefer not to say

What is your race or ethnicity? Please select all that apply. - White

- Black or African American
- Hispanic or Latino
- Asian
- Native American or Alaska Native
- Native Hawaiian or Other Pacific Islander
- Prefer not to say
- Other

Which U.S. state do you currently reside in?

[answer box]

What region do you live in within your state?

- Rural
- Suburban
- Urban
- Prefer not to say

What is your political affiliation?

- Democratic
- Republican
- Independent
- Libertarian

- Green Party
- Prefer not to say
- Other

Section 3: Voting Behavior and Intentions

These questions focus on voting registration and plans for the upcoming election.

How likely is it that you will vote in the 2024 U.S. Presidential Election?

- Very likely
- Somewhat likely
- Somewhat unlikely
- Very unlikely

If the election were held today, which candidate would you most likely vote for?

- Donald Trump
- Kamala Harris
- Not sure
- Prefer not to say
- Other

How confident are you that your choice would remain the same by Election Day?

- Not at all confident
- Slightly confident
- Moderately confident
- Very confident
- Completely confident

Section 4: Engagement with the Election

This section aims to understand how actively our respondents follow and discuss the election.

How closely do you follow news and updates related to the 2024 U.S. Presidential Election?

- Very closely
- Somewhat closely

- Not very closely
- Not at all

How often do you discuss the 2024 U.S. Presidential Election with friends, family, or colleagues?

- Daily
- Weekly
- Occasionally
- Rarely
- Never

Section 5: Additional Insights

We appreciate any additional thoughts our respondents may have on the upcoming election.

Do you have any further comments or insights about the factors that might affect the 2024 U.S. Presidential Election? (optional)

[Answer box]

End Message:

Thank You for Completing the Survey!

We appreciate your time and thoughtful responses. Your participation is invaluable in helping us gather insights for our research on the 2024 U.S. Presidential Election forecast. Your answers will contribute to a more comprehensive understanding of voter trends and factors influencing this election.

If you have any further questions or would like to know more about this study, please feel free to reach out to our research team at anjojoo.xu@mail.utoronto.ca.

As a thank you, each participant will receive a $5 reward shortly after survey completion.

B.6 Survey Budget Allocation

- Sampling & Recruitment: $50,000. To ensure a large and diverse sample, the majority of the budget are distributed for online recruitment and sampling.
- Incentives: $15,000. Fund small incentives for respondent participation.
- Data Validation and Cleaning: $10,000. Employ validation mechanisms to ensure data quality.

- Poll Aggregation and Analysis: $15,000. Use advanced analytical methods for accurate aggregation.
- Miscellaneous Costs: $10,000. Cover unforeseen costs in recruitment, data cleaning, or analysis.

B.7 Survey Design Considerations

The survey was designed with a focus on clarity, respondent comfort, and data reliability. Questions flow logically, from demographic information to voting intentions, perceptions, and engagement, reducing cognitive load. Wording is conversational, neutral, and direct, minimizing confusion and potential bias. Multiple-choice questions provide efficiency in response and analysis, while open-ended questions are selectively used for unique insights. A pilot test will be conducted to refine clarity and flow, ensuring that all questions are relevant, easy to answer, and respectful of the respondent's experience.

B.8 Methodology Strength and Weakness

B.8.1 Strengths

This methodology leverages stratified random sampling and poll-of-polls aggregation, providing a balanced, representative forecast by combining demographic weighting and time-based adjustments. Outlier detection and confidence intervals further enhance reliability, making the forecast adaptable to real-time shifts in voter sentiment.

B.8.2 Weaknesses

Limitations include potential sampling and non-response bias, as underrepresented groups and undecided voters may be difficult to capture. Additionally, reliance on self-reported data and aggregated poll results introduces variability, while temporal limitations may impact accuracy close to Election Day. The forecast provides a likely outcome range but cannot fully account for unexpected events or behavioral shifts.

C Additional Model details

C.1 Posterior predictive check

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…

Examining how the model fits, and is affected by, the data

C.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

References

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

Andrés López-Sepulcre ORCID iD, Matthieu Bruneaux. 2024. "Posterior Predictive Checks." https://cran.r-project.org/web/packages/isotracer/vignettes/tutorial-100-posterior-predictive-checks.html.

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.

FiveThirtyEight. 2024. "Presidential General Election Polls (Current Cycle)." https://projects.fivethirtyeight.com/polls/data/president_polls.csv.

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/.

Post, New York. 2024a. "Male Voters Under 30 Voting MAGA in 2024 Explain Trump's Appeal: 'Unabashed Machismo Vibe'." https://nypost.com/2024/10/28/us-news/male-voters-under-30-voting-maga-in-2024-explain-trumps-appeal/.

———. 2024b. "Trump Leading Harris in All but One Swing State Thanks to Strong Black Support: Polls." https://nypost.com/2024/10/31/us-news/trump-takes-all-but-one-swing-state-thanks-to-strong-black-support-polls/.

———. 2024c. "Why Donald Trump May Still Have the Upper Hand as Polls Show Him Deadlocked with Kamala Harris." https://nypost.com/2024/10/27/us-news/why-donald-trump-may-still-have-the-upper-hand-as-polls-show-him-deadlocked-with-kamala-harris/.

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

———. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

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.