Forecasting US Presidential Election Results Using Bayesian Prediction

YiZhuo Li Zhibin Huang

October 22, 2024

This article aims to use Bayesian prediction methods to predict the results of the US presidential election. By summarizing public opinion survey data, we can understand the candidate's support rate.

Table of contents

1 Introduction

In today's era of information explosion, predicting the outcome of the US presidential election has become a complex and delicate task. With the constantly changing opinions of voters and the influence of numerous variables, traditional single opinion polls are no longer able to accurately capture the dynamics of elections. As Pasek (2015) pointed out, a single opinion poll may produce misleading results due to various sources of error such as sampling errors, question wording, interviewer characteristics, and lack of response. In order to improve the accuracy of predictions, this analysis adopted the views of Blumenthal (2014) and Pasek (2015), using Bayesian prediction methods and poll aggregation strategies to obtain a deeper and more comprehensive understanding of candidate support rates. The study first introduced data preparation and variable description, with a focus on the percentage of support rates and indicators of opinion poll agencies. Then, we defined and fitted a Bayesian model to estimate the candidate's support rate, which can reduce the uncertainty of quantitative parameters through posterior distribution, enhance the credibility of analysis results, and allow the introduction of prior information to improve accuracy. The flexibility of Bayesian models enables them to effectively integrate data from multiple polling companies, reduce noise, and provide more stable predictions. Through posterior estimation of model parameters, we found that the overall average support rate was about 14.35%, with a standard deviation of about 15.39%, indicating significant variability between different survey results. Our model prediction shows that Trump is most likely to win with an average approval rating of 35.52%, followed by Biden with 28.90%. Although Biden has withdrawn from the election, his approval rating remains high, which has raised concerns about the effectiveness of the model and data processing. So, we discussed the possible limitations of the model, including the quality of input data, prior assumptions of the model, and biases in the survey data itself. We believe that when analyzing public opinion polls, it is crucial to use the latest and accurate data as well as a reasonable model structure.

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023), with the goal of studying poll of poll data to obtain more information about presidential elections. The Poll of Polls is a comprehensive dataset that combines public surveys from multiple pollsters, including key data such as methodology, states, and answers used by each pollster. Through simple data cleaning and processing, we obtained percentage data, which is the support rate of each candidate in different pollsters. By modeling this variable, I believe we can obtain more meaningful information.

2.2 Varibale

In this model, we mainly introduced two variables, percentage and pollster. Percentage is a vector that contains the percentage of support for each observation, reflecting the level of voter support for a candidate in a specific survey, typically ranging from 0% to 100%. Pollster is an integer vector representing the index of the survey institution corresponding to each observation value, ranging from 1 to K, where K is the total number of survey institutions used to distinguish the influence of different survey institutions. To identify the specific effects of different survey agencies on support rates, we calculated the pollster effect to capture the systematic bias that different survey agencies may have in overestimating or underestimating candidate support rates. Through these variables, the model can more accurately estimate support rates and consider differences in survey institutions.

3 Model

3.1 Bayesian

We will define and fit a Bayesian model to estimate the candidate support. Using Bayesian models to analyze 'poll of polls' has multiple advantages. Firstly, Bayesian methods can clarify the uncertainty of quantitative parameters through posterior distributions, thereby improving the credibility of analysis results. Secondly, it allows for the introduction of prior information, which can improve the accuracy of the model by combining historical data or expert knowledge in situations where there is limited or uncertain data. In addition, Bayesian models have strong flexibility and can adapt to different model structures and assumptions, especially suitable for complex voting data. At the same time, it can effectively integrate data from multiple polling companies, reduce noise, and obtain more stable predictions. Finally, Bayesian methods perform well in parameter estimation and inference tasks, providing a comprehensive analytical perspective for predicting future outcomes and hypothesis testing. Therefore, Bayesian models provide a systematic and reliable analytical tool for processing election predictions.

3.2 Modelling

The Bayesian model can be expressed with the following equations:

1. **Prior Distributions:**

$$\mu \sim \mathcal{N}(0, 1)$$

$$\sigma \sim \text{Cauchy}(0, 1)$$

2. **Likelihood Function:**

$$\text{percentage}_i \mid \mu, \sigma \sim \mathcal{N}(\mu, \sigma)$$

Combining these, the complete representation of the Bayesian model is:

$$P(\mu, \sigma \mid \text{percentage}) \propto P(\text{percentage} \mid \mu, \sigma) P(\mu) P(\sigma)$$

Where:

$$P(\mu, \sigma \mid \text{percentage})$$

is the posterior distribution given the data.

$$P(\text{percentage} \mid \mu, \sigma)$$

is the likelihood function.

$$P(\mu) and P(\sigma)$$

are the prior distributions.

3.3 Hypothesis

In the model, the parameter mu represents the mean of the overall voting percentage, which is the average level of different pollster voting results. And the parameter sigma represents the standard deviation of the voting percentage, while pollster_fect represents the impact of each pollster on the voting percentage. The prior distribution of parameters defines the initial assumptions for these parameters. Specifically, the prior distribution of mu is a normal distribution with a mean of 0 and a standard deviation of 1 (mu~normal (0,1)), which means that in the absence of data, we consider the central value of the overall voting percentage to be approximately 0 and there is some fluctuation. The prior distribution of sigma is a Cauchy distribution with a mean of 0 and a scale parameter of 1 (sigma~Cauchy (0,1)), indicating that we hold a relatively loose view on the variability of voting percentages, allowing for greater uncertainty. Finally, the prior distribution of the pollster_ effect is a normal distribution based on sigma (pollster_ effect~normal (0, sigma)), implying that the influence of each pollster on the voting percentage is distributed around the overall effect mu and has a certain degree of variability.

4 Figure results

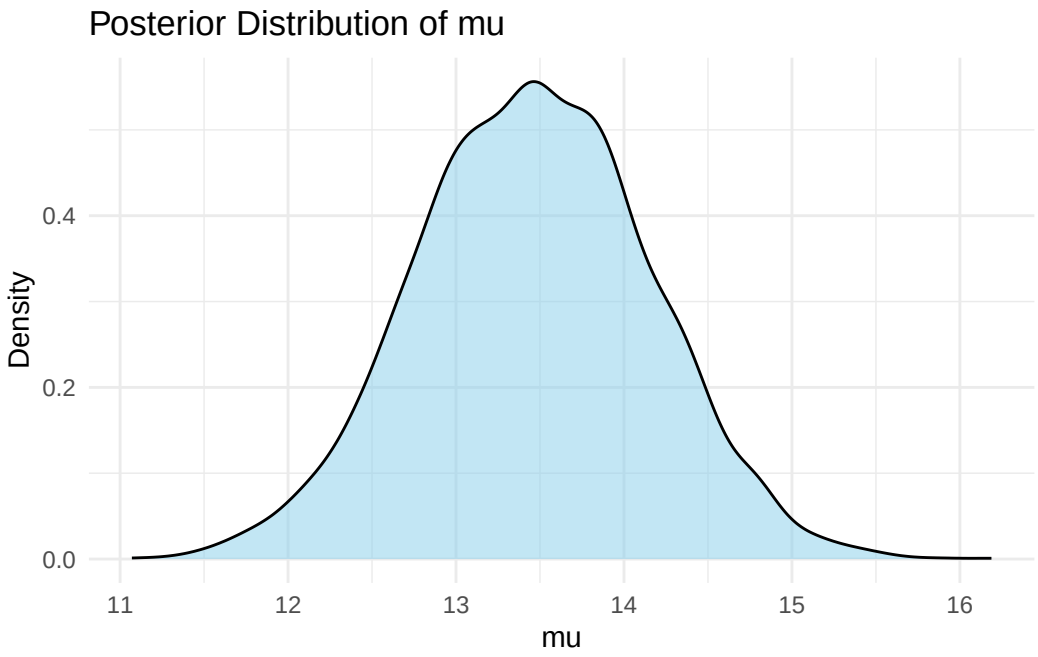

Figure 1

Figure 1 show the posterior estimate for the parameter shows that the posterior mean is -0.5724523. The 95% credible interval for ranges from -1.414763 to 0.5034533, indicating the range within which the true value of is likely to fall with 95% probability.

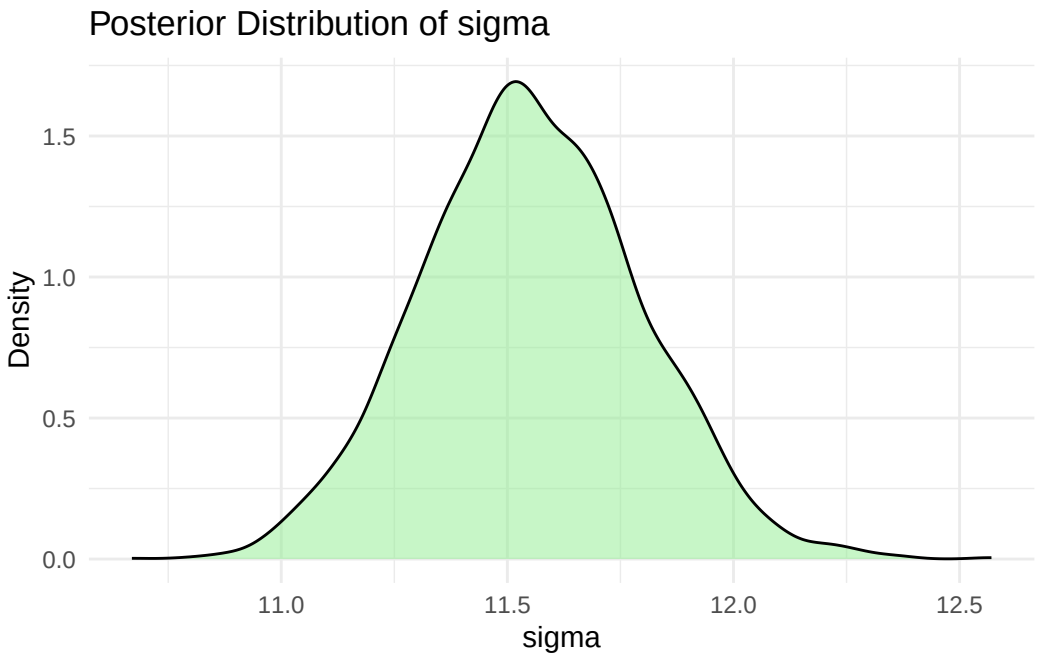

Figure 2

Figure 2 show the parameter , the posterior mean is estimated to be 3.533605. The 95% credible interval for extends from 0.8570568 to 10.27852, reflecting the uncertainty around the estimate and providing a range of probable values for .

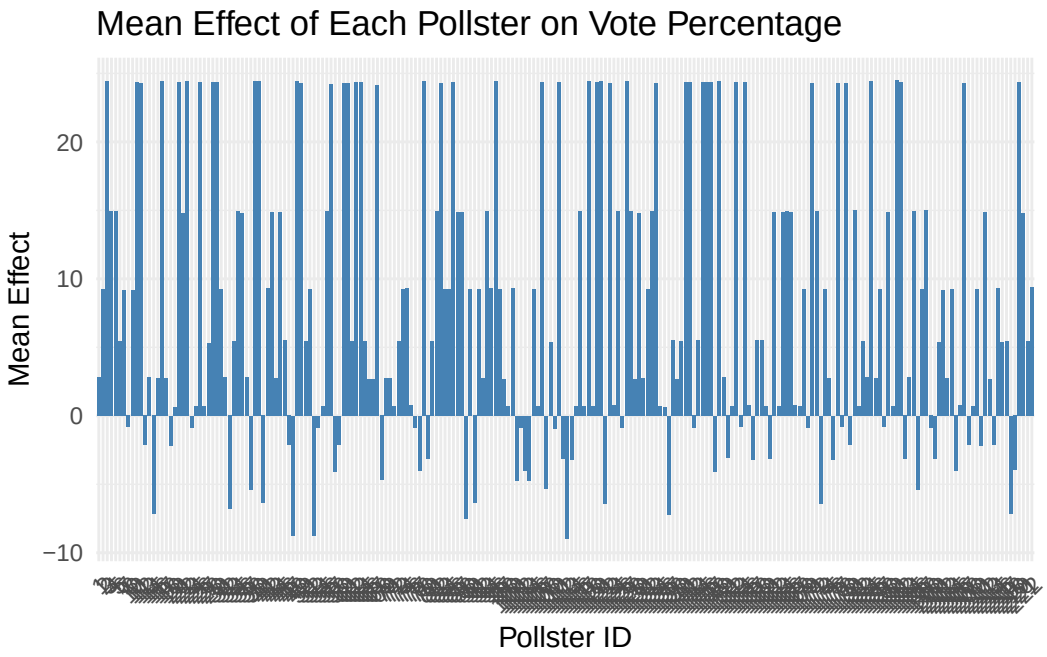

Figure 3

Figure 3 results indicate that the overall mean support rate for candidates is approximately 14.35%, and at a 95% confidence level, the true mean falls between 13.54% and 15.14%, showing some volatility. Meanwhile, the standard deviation of the support rate is about 15.39%, indicating a significant degree of variability between different survey results. The confidence interval (14.79% to 16.00%) further confirms the range of changes in support rates, reflecting possible differences in characteristics and methods among different survey institutions. Overall, these results indicate that the candidate's support rate is around 14% and there is significant variation.

Table 1: Top 4 Candidates Results Table

answer	mean_percentage
Trump	35.51560
Biden	28.90011
Harris	21.13103
Ballay	20.00000

Table 1 show trump is the most likely candidate to win with an average percentage of 35.5156. Vote Percentage Distribution Finally, we visualize the vote percentage distribution of the top candidates.

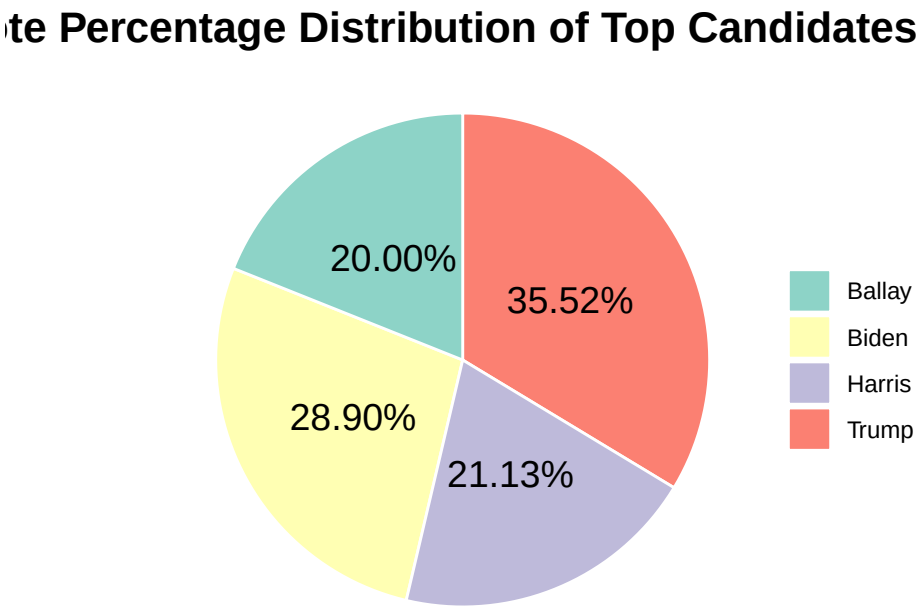

Figure 4

Figure 4 show in the popular vote of the 2024 US presidential election, the Republican Party has one representative, former President Donald Trump, who ranks first with 35.52% of the vote. The Democratic Party has two representatives, namely current Vice President Kamala Harris and President Joe Biden, with vote percentages of 21.13% and 28.9%, respectively. In addition, the Liberal Party has one representative, Charles Ballay, who received 20% of the vote, the lowest among all candidates. This distribution shows the diversity of the political landscape in the United States, where the Republican and Democratic parties remain the main political forces, while the Liberal Party, as a third-party force, still holds a place in elections despite its lower vote share. Trump's high vote share reflects his continued influence and support among Republican voters in the United States. The vote share of Harris and Biden shows the competitiveness within the Democratic Party and the different views of voters towards these two candidates. As a candidate for the Liberal Party, Charles Bale's participation represents the demand for third-party candidates and a challenge to the two party system in American politics, despite having the lowest vote share.

5 Discussion

According to the prediction results of your model, Trump's win rate is 35.52%, while Biden's win rate is 28.90%, despite Biden's withdrawal from the election, which has raised concerns about the effectiveness of the model and data processing. Firstly, Bayesian models rely on the quality of input data. If the data contains outdated or incorrect polling information, the model may produce inaccurate results. In this situation, Biden's approval rating may stem from early polls not reflecting his withdrawal in a timely manner. Secondly, the prior assumptions of the model may also affect the results. If the prior information is not appropriately adjusted to reflect candidate withdrawal during model construction, it may lead to inaccurate estimation of Biden's support rate. In addition, the polling data itself may have biases, especially after candidates withdraw, and many voters' opinions may not have been updated yet. Therefore, when integrating polling data from multiple sources, special attention needs to be paid to updating and reflecting the latest election situation. Finally, the design and parameter settings of the model may not be sufficient to handle dynamically changing election environments. Although Bayesian models are flexible, they also require a reasonable modeling strategy to capture changes in candidate support rates. Therefore, when analyzing the 'poll of polls', it is crucial to ensure the use of the latest and accurate data, as well as a reasonable model structure.

6 Appendix

6.1 Appendix A

Online panel is a method of using Internet technology to conduct research. It collects data by recruiting a group of volunteers who are willing to participate in multiple online surveys. These volunteers, i.e. panel members, typically receive survey questionnaires via email or other online channels and are expected to complete these questionnaires. The advantage of this method is that it can quickly and cost effectively obtain information from a large number of participants. However, due to the voluntary participation of panel members, this may lead to sample selection bias, meaning that these members may not fully represent the diversity of the entire population. For example, some groups may be more inclined to join the panel because they are more familiar with or interested in technology, while other groups may be unable to participate due to lack of Internet access or other obstacles. In the study of political attitudes and behaviors, this bias is particularly noteworthy as it may affect the accuracy of election predictions and policy analysis. Despite these potential biases, online panels still have the potential to collect representative data. If it can be ensured that the samples are obtained through probability sampling, that is, randomly selecting samples to ensure that each member has the same chance of being selected, then online panels can become a valuable tool in social science research. The transparency and replicability of this method make it particularly important when comparing the accuracy and representativeness of different survey

methods. By combining advanced technology and social science research, online panels can help researchers better understand the complexity of election dynamics, attitude formation, and social and political behavior.

6.2 Appendix B

References

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