A Polls-of-Polls Forecast for the 2024 United States Presidential Election Through Generalized Linear Model*

The Impact of Pollster Differences, Population Type, and Poll Regency on Predicting Support for Kamala Harris

Jiwon Choi Kevin Roe

November 4, 2024

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

Table of contents

[*] Code and data are available at: https://github.com/jwonc4602/2024_US_Election_Forecast.

1 Introduction

The United States Presidential Election is one of the most consequential events of 2024. As a key country in international relations and the largest economy in the world, the results of the U.S election determines the country's domestic and foreign policy for the next four years, which will have a significant impact on important initiatives such as tackling climate change or international conflicts. In anticipation of the 2024 U.S election, this paper aims to predict the possible outcomes of the election by analyzing the level of support that Kamala Harris will gain.

We forecast support for Kamala Harris based on polling results and apply a Bayesian generalized linear model. The main parameter of interest is the proportion of vote or support that Harris received in surveys, which is traced over time. By considering the results from different poll-making organizations and other demographic factors, our objective is to correct for variation across different voter bases with

Factoring in the different results of poll-making organizations, our model found that specific pollsters or demographic characteristics increase or decrease Harris' predicted voter share. Understanding the trajectory of the US might make in foreign policy, economic or global politics prediction allows global stakeholders to take preemptive measures for policy changes of the newly elected government. Thus, this study is a strong tool to navigate uncertainty around the 2024 election.

The paper is structured as follows: Section 2 and Section 3 explores the data and methodology used, highlighting the filtering and modeling techniques applied to the data; Section 4 presents the results from the Bayesian generalized linear model; Section 5 discusses the broader implications of our findings; and Section A highlights YouGov's methodology, outlines an idealized survey methodology, and shows model diagnostics.

2 Data

We have used a poll-of-polls of the upcoming US presidential election dataset obtained from FiveThirtyEight (FiveThirtyEight, n.d.). Data was collected and analyzed using R statistical programming software (R Core Team 2023), with additional packages like tidyverse (Wickham et al. 2019), rstanarm (Goodrich et al. 2022) knitr (Xie 2020), here (Müller 2020), and many others for support. The dataset includes a wide range of poll results from various polls, with keyv ariables such as pollster, sample size, percentage of support for Harris, and the date conducted for polls.

To ensure high data quality, we filtered the dataset to include polls with a numeric grade of 3.0 and focused only on polls conducted after July 21, 2024, the date of Harris' declaration of candidacy. The filter helped limit bias from older polls from when Joe Biden was the Democratic nominee, which may not reflect the current voter sentiment.

In performing the analysis, we used several R pacakages. We used tidyverse (Wickham et al. 2019) for data manipulation and visualization, and rstanarm (Goodrich et al. 2022) for Bayesian modeling. To visualize results, we used ggplot2 (Wickham 2016) and kableExtra (Zhu 2020) to format tables for presentation.

2.1 Measurement

FiveThirtyEight aggregates various poll results from national and state-wide polls, showing 16867 observations (FiveThirtyEight, n.d.). The poll takes a sample of the electorate and asks for the voters' candidate of choice. By factoring polls from a state and national level, FiveThirtyEight aims to represnet public perception on the two candidates during the election to predict the election's outcome.

Each poll aims to predict an actual event. The raw data is susceptible because each pollster has varying polling methods. There are other limitations such as sampling error, response error, or distortion from participants misunderstanding the question.

Beyond filtering for high-quality pollsters, we filtered for various criteria to prepare the analysis data.From the raw data, we filtered out any missing numbers, cleaned names, and removed all polls on non-Democrat or Republican candidates, effectively on comparing Harris and Trump. After determining state-level analysis for electoral votes and close races, we created different datasets for national and state level polls. Regardless, selection bias and sampling bias is still a concern because polls only represent a part of the population. Finally, polling methodology difference can introduce bias in the study. Overall, we use the data from individual responses to election polls, and filter the data for analysis to make a prediction on who will win the popular vote based on data from the pollsters.

2.2 Outcome variables

2.2.1 The Predicted Proportion of Support a Candidate Received in a Poll

The main variable we aim to forecast is the pct variable, which represents the proportion of the vote a candidate received in a poll. Table 1 and Figure 1 shows the summary statistics and distribution of pct variable in a filtered dataset that only comprises of votes from relatively high-quality polling organizations. We also get the popular vote predictions for each candidate using the predict() function and add this to the dataset as the predicted_pct variable. We have showed the summary statistics and distribution for predicted_pct in Table 2 and Figure 2, respectively.

Comparing our predictions to the data, our predictions have a smaller range than the cleaned data. While the difference in mean is 2% lower for our prediction than our cleaned data. The results suggest there is less variation in our predictions than the cleaned dataset.

Table 1: Summary statistics for the proportion of support candidates received in the poll

mean	median	min	max	sd	n
45.25	46	18	70	4.82	7268

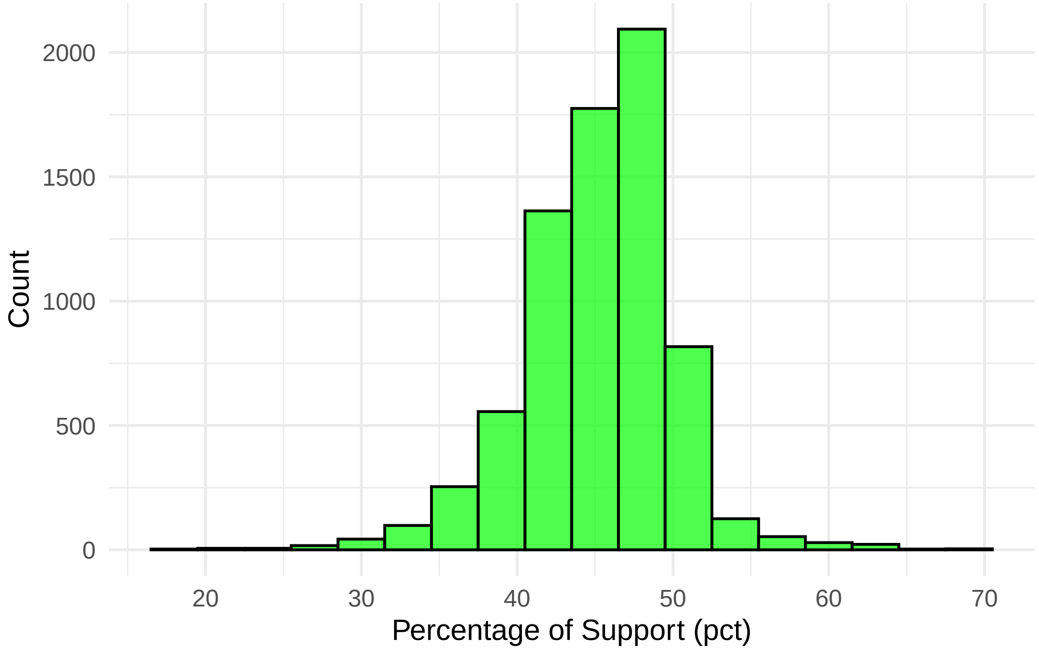

Figure 1: Distribution of the proportion of support candidates received in the poll

Table 2: Summary statistics for the predicted proportion of support candidates received in the poll

mean	median	min	max	sd	n
47.07	46.83	45.67	50.44	1.39	311

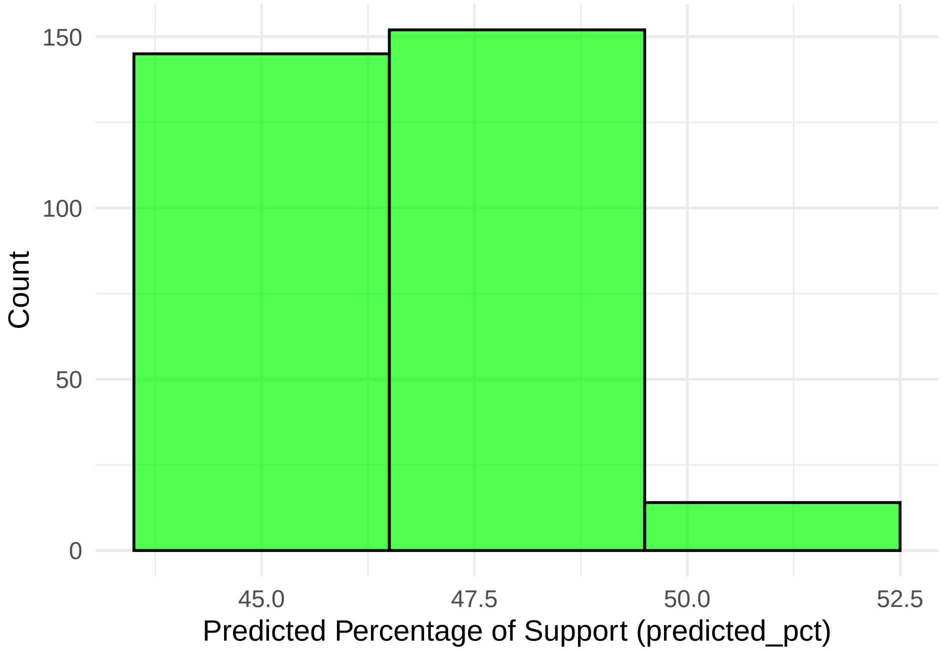

Figure 2: Distribution of the predicted proportion of support candidates received in the poll

2.3 Predictor variables

2.3.1 Type of Pollster

The pollster variable was selected to consider the effect of changes in pollster on the observed variable. The pollster variable represents the polling organization that conducted the poll. The distribution of polling counts for different pollsters in Figure 3 suggests that the data is dominated by two pollsters: YouGov and Siena/NYT. Further analysis is needed in their polling methodology to determine potential biases.

Table 3: Number of unique high-quality polling organizations

pollster	Count
CES / YouGov	28
Marquette Law School	24
McCourtney Institute/YouGov	1
Siena/NYT	136
The Washington Post	10
YouGov	110
YouGov Blue	1
YouGov/Center for Working Class Politics	1

2.3.2 Population Type

Population Type distinguishes among different groups surveyed into two groups surveyed: voters and adults. Adults is a more general poll that includes voters, likely voters and non-voters. Figure 4 shows that majority of the polls surveyed voters rather than the general adult population. The discrepancy between number of polls who surveyed voters and adults could introduce potential biases in responses, as those who are voters may have stronger opinions than those who are not.

Table 4: Number of unique high-quality polling organizations

population	Count
Adults	22
Voters	289

2.3.3 Poll Recency

Poll Recency, shown through the variable `recent_poll`, labels any polls conducted in the last 30 days as Recent and any polls collected before that are labeled Older. Based on our results shown in Table 5, 19 more polls have been collected in the last 30 days than before. This introduces biases in responses as recent responses are more represented in our model. In rapidly changing political environments, public opinion can shift dramatically due to major events, debates, or crises, making recent data crucial for accurate modeling. The distribution of this variable can also be found in Figure 5.

Table 5: Number of polls that were collected within 30 days and after

recent_poll	Count

Older	146
Recent	165

3 Model

For our analysis, we employ a Bayesian generalized linear model (GLM) to forecast the popular vote percentage for Kamala Harris. This approach allows us to capture polling characteristics and account for known variations between pollsters, population types, and the recency of polling data. By incorporating these factors, we aim to model a nuanced estimate of Harris's projected popular vote.

The first step in our process involved selecting a reliable dataset for model development. Here, we utilized high-quality national polling data gathered after Harris's campaign announcement. We filtered the dataset to include polls with a numeric grade of 3.0, ensuring data quality, and focused only on polls conducted after July 21, 2024, the date of Harris's declaration of candidacy. This filter helps limit bias from older polls that may not reflect the current voter sentiment.

The GLM is specified as follows:

$$Y_i = \beta_0 + \beta_1 x_{pollster_i} + \beta_2 x_{population_i} + \beta_3 x_{recentpoll_i} + \epsilon_i \tag{1}$$

In equation 1, each β represents a coefficient determined through regression analysis. The variables chosen for this project are pollster, population type, and recency of the poll. Each predictor variable was carefully selected based on its significance in polling analysis and its correlation with voting trends. The identity of the polling organization is important to our model, as each pollster may exhibit unique biases. Including Pollster as a fixed effect allows us to account for these variations without introducing unnecessary complexity. Population Type distinguishes among different groups surveyed (e.g., voters, likely voters) and helps in capturing generalizability. Categorizing polls as either recent or not ensures that more recent polls, which are better predictors of voting behavior closer to the election, receive appropriate emphasis. Y_i denotes the predicted popular vote percentage for Kamala Harris in the i-th poll. ϵ_i is the Gaussian-distributed error term, accounting for residual variation in the model.

The selection of predictors in this model is based on observable patterns in the data, such as the consistent tendencies of certain pollsters and the impact of population type on polling outcomes. To enhance the model's robustness, Bayesian priors were applied, introducing regularization and incorporating plausible ranges grounded in previous election data and polling analysis. For the coefficient priors β, a normal distribution with a mean of 0 and a scale of 2.5 (autoscaled) was chosen to provide flexibility while mitigating overfitting. Similarly, the intercept uses a normal prior with a mean of 0 and scale of 2.5 to stabilize model estimates. For the error term (sigma), an exponential prior with a rate of 1 was selected to constrain the residuals,

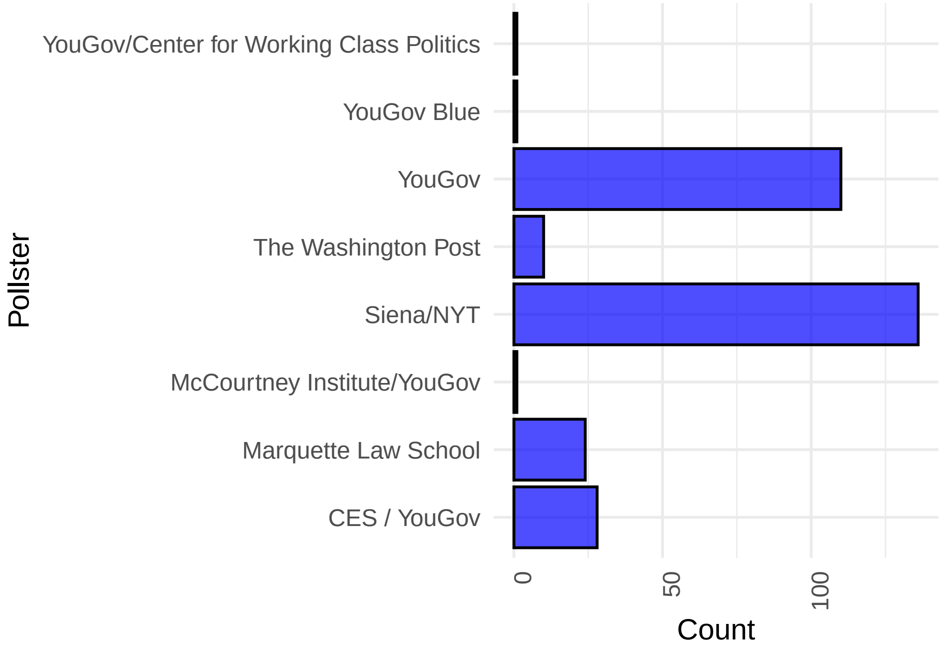

Figure 3: Distribution of high-quality polling organizations

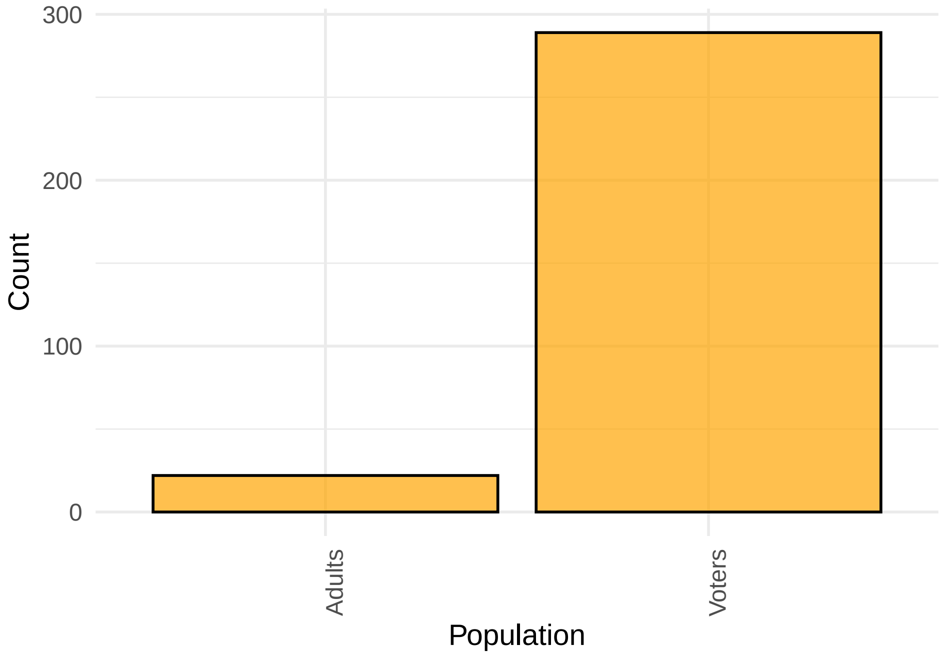

Figure 4: Distribution of population type

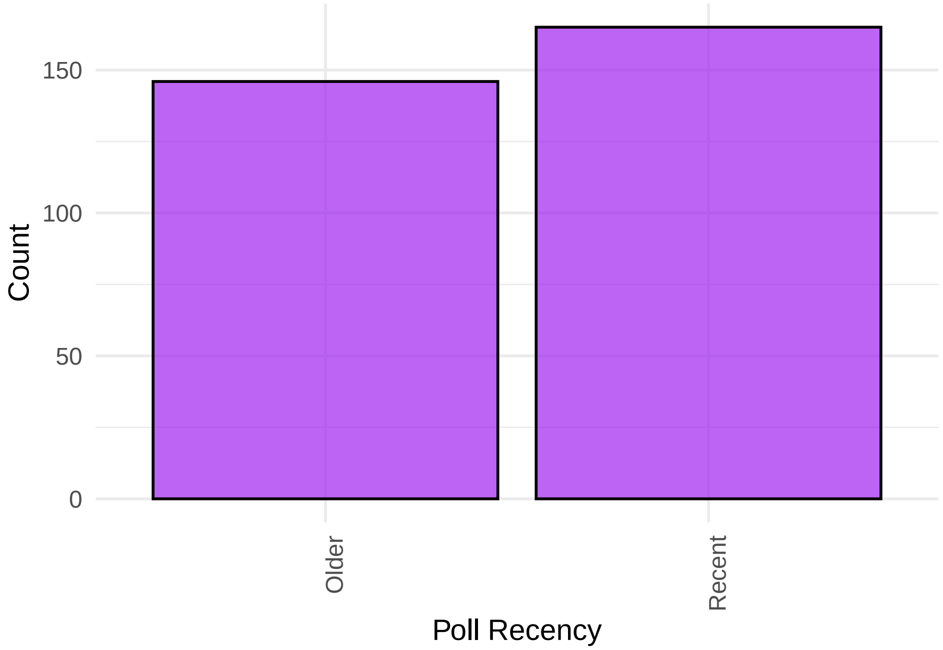

Figure 5: Distribution of on the amount of polls collected in 30 days

aligning with Gaussian assumptions. These priors offer a balance between model flexibility and constraint, informed by established trends in polling data and comparable electoral forecasts.

The model was implemented in R (R Core Team 2023) using the `rstanarm` package, which offers an accessible interface for Bayesian generalized linear models (GLMs), allowing specification of priors and customization of model parameters. Once the logistic regression model for predicting the popular vote is developed, we apply the `predict()` function in R (R Core Team 2023) to generate popular vote percentage predictions for each candidate, using filtered national data. These predictions are then added to the dataset as a new variable (`predicted_pct`) for further analysis. The results are saved in a CSV file (`popular_vote_predictions.csv`), enabling an in-depth examination of potential outcomes in the popular vote. This approach facilitates understanding of vote distributions at a national level and provides a foundation for forecasting electoral outcomes based on demographic and polling data.

4 Results

To assess model reliability, we examined several key diagnostics. Convergence metrics, such as Rhat values, were close to 1 for all parameters, indicating strong convergence. Additionally, the effective sample size (n_eff) was high across parameters, suggesting low autocorrelation and contributing to model stability. For validation, we conducted out-of-sample testing and

Table 6: Coefficients from the GLM Model

term	estimate	std.error	conf.low	conf.high
(Intercept)	48.64	0.88	47.18	50.12
pollsterMarquette Law School	-1.66	1.12	-3.56	0.23
pollsterMcCourtney Institute/YouGov	-2.48	3.66	-8.44	3.56
pollsterSiena/NYT	-4.76	0.92	-6.28	-3.25
pollsterThe Washington Post	-3.67	1.42	-5.99	-1.39
pollsterYouGov	-2.56	0.90	-4.04	-1.12
pollsterYouGov Blue	-4.48	3.60	-10.54	1.26
pollsterYouGov/Center for Working Class Politics	-3.76	3.63	-9.69	2.23
populationVoters	1.87	0.95	0.33	3.41
recent_pollRecent	-0.07	0.43	-0.77	0.63

calculated the Root Mean Square Error (RMSE) to assess predictive accuracy. This test data, which includes polls not used in training, provides an unbiased estimate of model performance. See more details of model diagnostic here: Section A.3.

The model operates under the assumption that residuals follow a Gaussian distribution, though this may not entirely capture extreme polling variances. Additionally, pollster effects are treated as fixed, which, while limiting the model's ability to reflect dynamic shifts in polling methodologies, simplifies the model and reduces the risk of overfitting. Alternative specifications, including a hierarchical model with random intercepts for pollsters, were considered. However, the added complexity of this approach did not much improve predictive accuracy. The chosen GLM specification thus offered a balanced approach, optimizing both interpretability and performance, making it the preferred model for our study.

Table 6 presents the estimated coefficients for the predictors in our GLM model. These coefficients fit into the GLM equation, allowing us to interpret the impact of each predictor on Harris's predicted vote percentage. Positive values indicate a higher predicted percentage, while negative values indicate a decrease. Key predictors, such as specific pollsters and population types, show distinct effects on the forecasted outcomes.

Figure 6 represents the model coefficients, with error bars indicating the confidence interval for each estimate. Positive coefficients suggest that specific pollsters or demographic characteristics increase Harris's predicted vote share. For example, the 'populationVoters' shows a positive impact, while most of the pollsters has a slight negative effect. These error bars help contextualize the reliability of each predictor.

5 Discussion

5.1 How has the Forecast Changed Over Time?

5.2 Which States Prefer Harris?

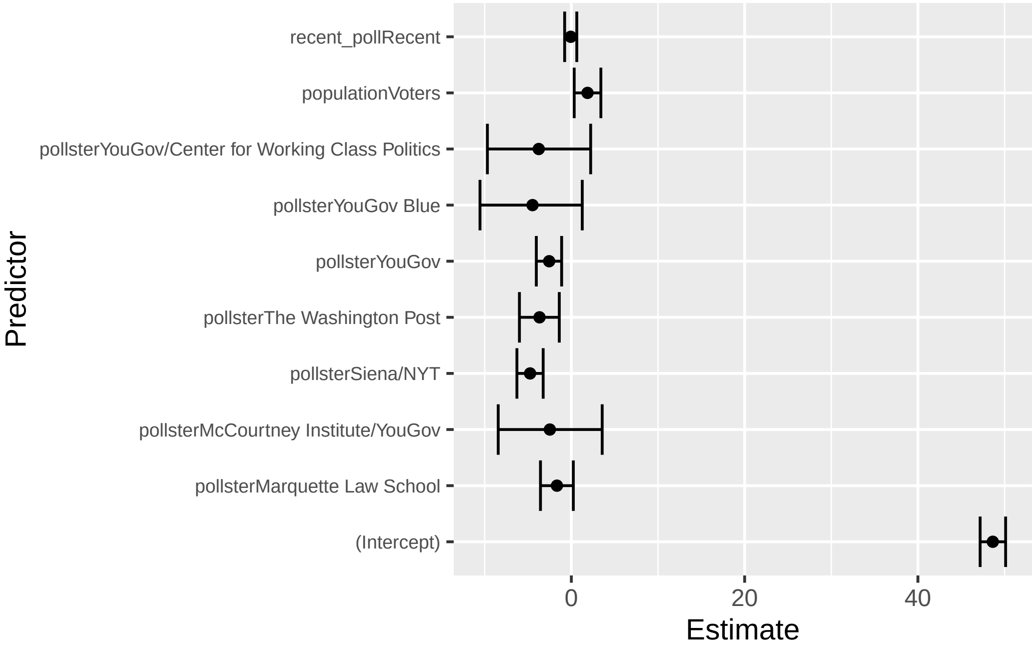

Figure 6: Coefficient Estimates for Predictors

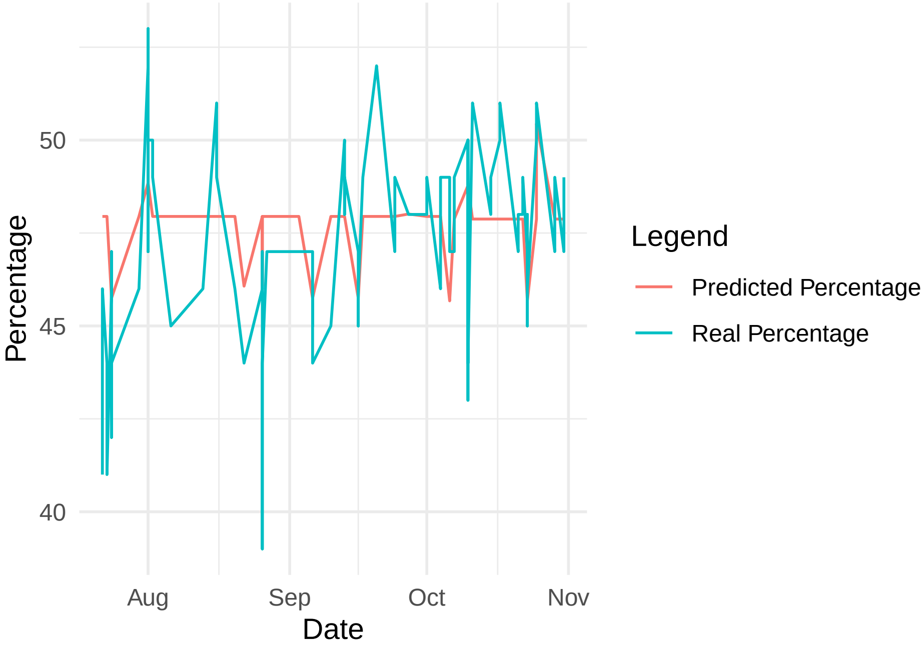

Figure 7: Comparison of Predicted vs Actual National Poll Percentages for Harris Over Time

Table 7

state	missing_dates	total_dates
Arizona	0	23
California	0	2
Florida	0	8
Georgia	0	19
Michigan	0	23
Minnesota	0	2
Missouri	0	1
Montana	0	4
Nebraska	0	14
Nevada	0	14
New Hampshire	0	2
New York	0	2
North Carolina	0	21
Ohio	0	8
Pennsylvania	0	33
Texas	0	15
Virginia	0	2
Wisconsin	0	37
NA	0	81

Summary of missing and total dates across key states

```
`geom_line()`: Each group consists of only one observation.
i Do you need to adjust the group aesthetic?
```

5.3 Implications

asdfasdfasdf

5.4 Weaknesses and Next Steps

- if filter thorugh state, data not collected easily

Weaknesses and next steps should also be included.

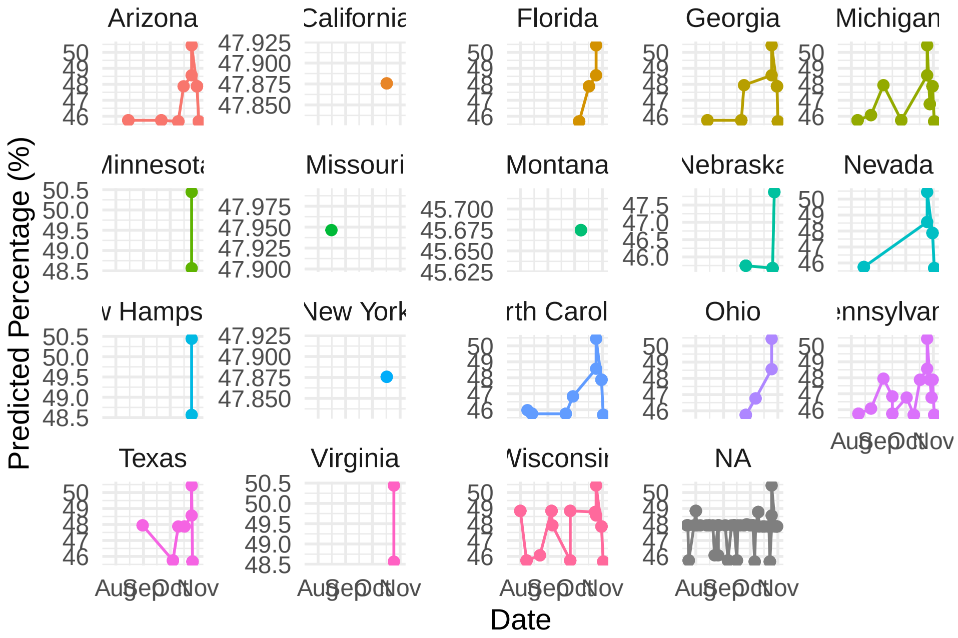

Figure 8State-by-State comparison of predicted percentages over time

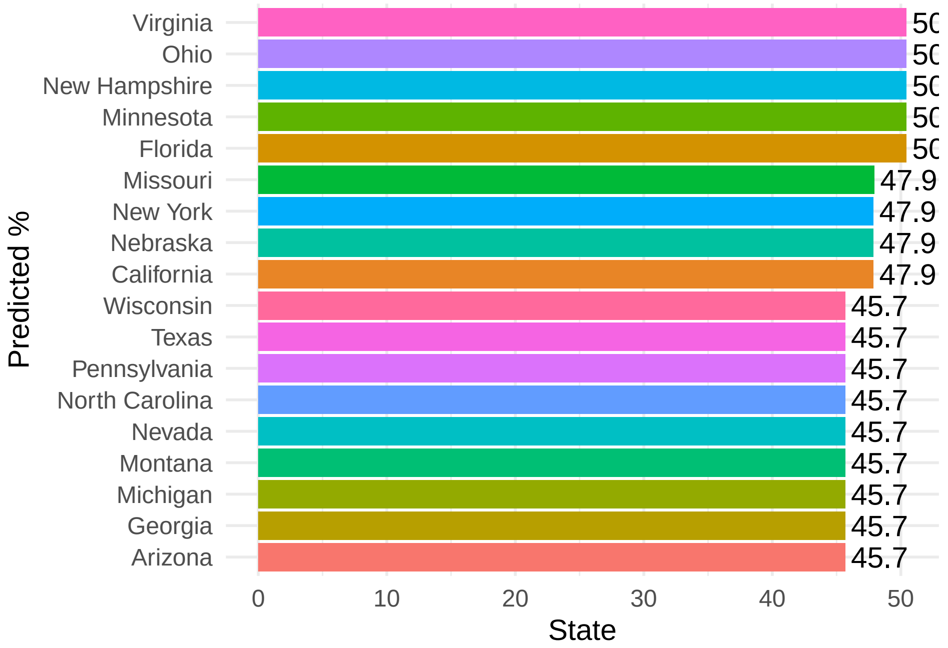

Figure 9Summary of Kamala Harris' predicted popular vote by key states

A Appendix

A.1 YouGov Methodology Analysis

YouGov employs a panel-based methodology for conducting online surveys, recruiting participants who voluntarily join their platform. Their surveys span topics like politics, policy opinions, and voter behavior. In the context of election polling, YouGov applies Multilevel Regression and Post-Stratification (MRP) models to predict outcomes by blending respondent data with external data like voter files (YouGov 2024).

The target population for YouGov's surveys typically consists of U.S. adults, with specific samples drawn for each survey depending on the topic. YouGov's sample frame consists of individuals who have opted into their online panel. To ensure representativeness, YouGov sets demographic quotas, stratifying by variables such as age, gender, and political affiliation. Additionally, YouGov monitors and adjusts their sample post-survey using MRP to correct for demographic imbalances and potential sampling bias.

YouGov's panel is recruited through various online methods, including digital advertisements and direct outreach. This recruitment strategy leads to a convenience sample, meaning participants self-select to join the panel, which can introduce selection bias. Although YouGov attempts to offset this bias using quotas and post-stratification weighting, the sample still primarily consists of individuals comfortable with online surveys. This method may underrepresent certain demographics, such as older adults or individuals with limited internet access.

YouGov's use of non-probability sampling—drawing respondents from a convenience sample—provides cost efficiency and quick data collection. However, the trade-off lies in the introduction of potential biases due to self-selection. While YouGov's MRP model corrects some biases by leveraging population-level data for small samples, the non-random nature of the sample may still skew results. For instance, respondents who engage in online surveys may systematically differ from those who avoid them, such as individuals with lower levels of social trust.

Non-response bias is one of the main concerns in YouGov's methodology. Individuals who opt out of surveys may systematically differ from those who participate. YouGov employs several strategies to mitigate this issue, including post-survey weighting, which adjusts the survey data to reflect population characteristics, and imputation techniques, which estimate missing responses based on patterns in the available data. Despite these efforts, non-response bias can still affect the accuracy of predictions, particularly when certain demographic groups (e.g., less-educated or low-trust individuals) are less likely to participate.

YouGov focuses on clear and concise survey design, reducing participant fatigue and improving data quality. Surveys typically avoid overly complex or leading questions, which can skew the results. However, question-wording remains an important area, as even subtle changes in expression can affect responses. For example, YouGov's research shows how question-wording can affect voter perceptions and reported support for policies. Furthermore, prolonged surveys can lead to increased dropout rates, potentially resulting in non-response bias. To counteract this, YouGov often dynamically adjusts questions based on previous responses to keep participants engaged by incorporating survey logic.

YouGov's panel-based methodology, combined with statistical techniques like MRP, offers a robust approach to polling. Their methods effectively address many challenges associated with non-random sampling and non-response bias. However, limitations persist, particularly regarding the biases inherent in self-selection and question framing. Overall, while YouGov's methodology is well-suited to large-scale surveys, particularly in politics, careful consideration of their sampling and non-response strategies is necessary to fully interpret their findings.

A.2 Idealized Methodology

Because the U.S. operates on the Electoral College system, where state-affiliation matters, I would use Stratified Random Sampling to ensure each state is represented proportionally in the survey, improving accuracy and representativeness. My target population would be voter-eligible citizens in the United States who are above 18 years old, with a sampling frame utilizing voter registration databases stratified by key demographics, including age, race, gender, geographic region, income, and education level. I aim for a sample size of around 10,000 respondents to achieve a reasonable margin of error, ensuring robust findings, though I expect the actual sample may be less due to various factors.

To recruit respondents, I would primarily utilize online methods, such as surveys on social media platforms like Facebook and Instagram to engage younger audiences, and interactive voice response systems for older or less tech-savvy individuals. To ensure the data collected is valid, I will cross-check responses with voter registration databases and implement logic checks within the survey to detect inconsistencies. For instance, if a respondent claims to have already voted but indicates they're unlikely to vote, this response will be flagged as unreliable.

Addressing non-response bias is crucial, and I plan to over-sample underrepresented groups while employing multiple attempts to contact individuals online and conducting weekly polls. Frequent surveys will allow me to aggregate results using a moving average to smooth out short-term fluctuations in responses. I will also implement checks to prevent duplicate entries in the online survey and track phone responses to ensure unique participants. Given that the U.S. does not conduct elections based on a popular vote, I will weight the survey results to reflect the voting population as indicated by U.S. Census data and voter turnout estimates, considering demographic factors such as race, gender, and age.

To allocate the $100,000 budget, I will spend $20,000 on a survey platform subscription, $50,000 on targeted digital ads and phone surveys for recruitment, $20,000 on data analysts and poll aggregation services, and $10,000 for re-contact surveys and miscellaneous costs. However, I recognize that this budget may be on the lower end to ensure a high-quality poll and that the actual costs might exceed this amount. Most expenses will arise from data collection, and I have accounted for this in my budget planning.

Please find the proposed question list for the online survey attached here.

A.3 Diagnostics for model

Figure 10 compares observed data (dark line) with replicated posterior predictions (lighter lines). The close alignment suggests that the model accurately captures the data's central tendency and variability. Figure 11 and Figure 12 show that the sampling algorithm used, the Markov chain Monte Carlo (MCMC) algorithm, did not run into issues as the posterior distribution for the model was created. Using the checks presented by Alexander (2023), both graphs do not show anything abnormal since he trace plots in Figure 11 display substantial horizontal fluctuation across chains, indicating good mixing, while the Rhat values in Figure 12 are close to 1 and well below 1.1, further supporting convergence.

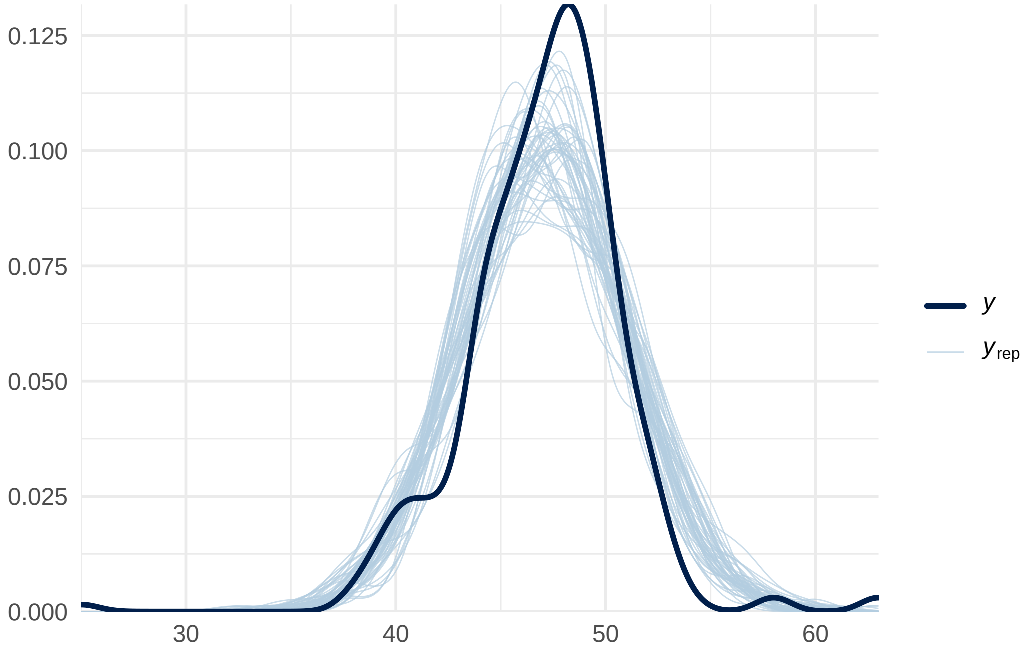

Figure 10Posterior Predictive Check: Comparison of Observed and Replicated Data

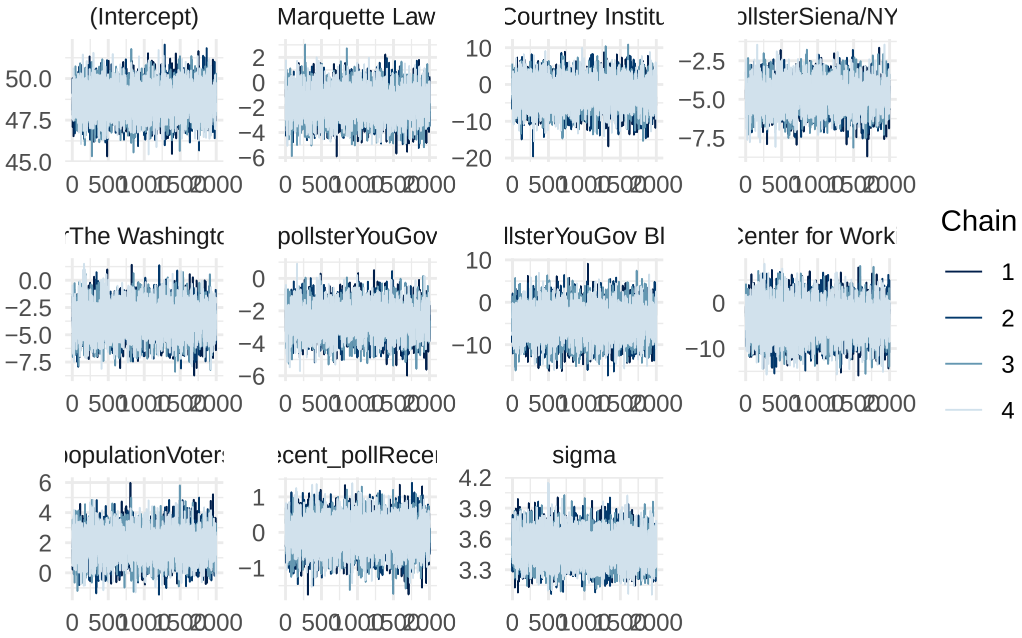

Figure 11Checking the convergence of the MCMC algorithm - Trace

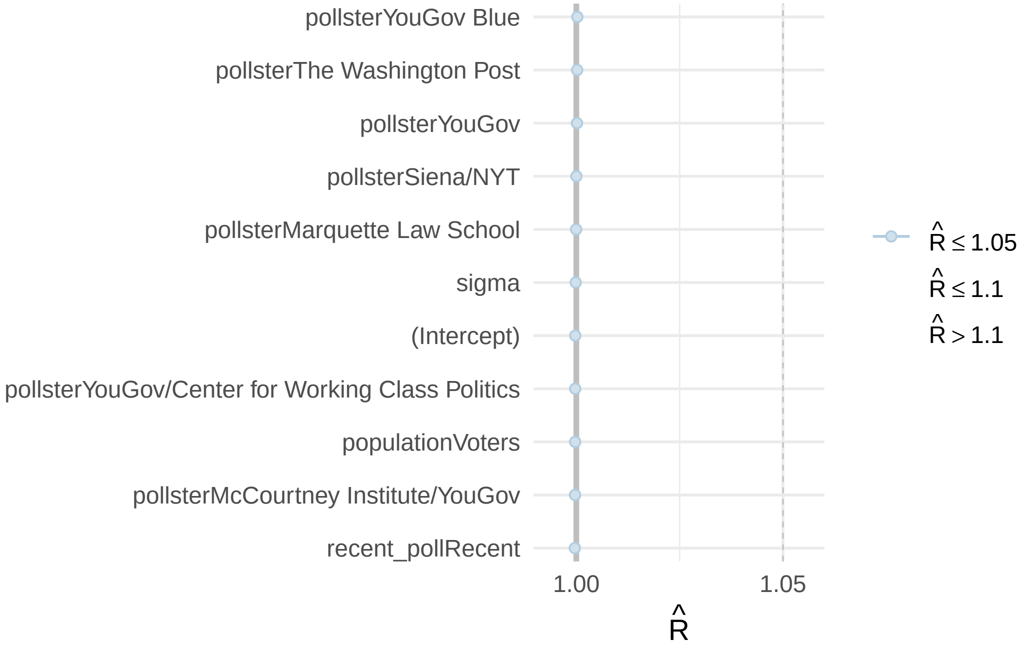

Figure 12Checking the convergence of the MCMC algorithm - Rhat

References

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

FiveThirtyEight. n.d. "Latest Polls - FiveThirtyEight." https://projects.fivethirtyeight.com/polls/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Müller, Kirill. 2020. *here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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

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.

Xie, Yihui. 2020. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

YouGov. 2024. "How YouGov's MRP Model Works for the 2024 Presidential and Congressional Elections." https://today.yougov.com/politics/articles/50587-how-yougov-mrp-model-works-2024-presidential-congressional-elections-polling-methodology.

Zhu, Hao. 2020. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.