Forecasting the 2024 US Presidential Election: A Poll-of-Polls Approach to Predict the Outcome*

Utilizing Poll Aggregation and Statistical Modeling to Project a Clear Path for the Presidential Race

Ruiying Li Lorina Yang

November 1, 2024

This paper forecasts the 2024 U.S. Presidential election by analyzing polling data on support for Kamala Harris and Donald Trump. Using Bayesian linear and logistic regression models, we examine the impact of poll characteristics like sample size and quality on support levels and predict each candidate's winning probability. Our results show minimal influence from poll quality on support percentages but highlight a recent surge in Harris's support, narrowing Trump's previous lead. These findings provide insight into polling reliability and evolving voter sentiment as the election nears.

1 Introduction

The ongoing U.S. presidential election is a major focus in the media, finance, and politics, making it essential to provide reliable forecasts and understand the factors influencing varying levels of candidate support. This study utilizes Bayesian linear and logistic regression models to analyze and predict support for Kamala Harris and Donald Trump, integrating poll-level data on support percentages, poll quality, sample size, and time trends.

Aim

Our primary objectives are to examine how poll characteristics—sample size and quality—influence reported support percentages and to track shifts in support over time for Harris and Trump. The linear regression model evaluates these poll factors, assessing their impact on support stability. Additionally, time-trend analysis identifies evolving voter sentiment,

*Code and data are available at: https://github.com/Lorina-Y/US-Presidential-election.git.

capturing a recent change in support rate for candidates. And second, to forecast aggregate support rates for Harris and Trump, facilitating comparison. We use the Bayesian linear regression model to estimate overall support rate aggregate by cadidates, a logistic model complements this by estimating each candidate's probability of success directly, adding a probabilistic perspective.

Estimand
The primary estimands include the impact of poll characteristics on support percentage and each candidate's probability of winning. The linear model provides insights into how polling factors influence reported support. For the Bayesian forecasting component, The key estimand in this analysis is the probability of support for Harris versus Trump for each poll, aggregated to estimate an overall probability of each candidate's support. This probability is informed by predictors including poll-specific support percentage (pct), poll quality (pollscore) and sample size, providing a detailed estimate of each candidate's likelihood of winning.

Results
The analysis shows that sample size and poll quality have minimal effects on support levels, supporting the reliability of aggregated poll data. The linear model found near-identical support levels for Harris and Trump (44.75% and 44.76%, respectively), indicating a close race. The logistic model, however, gave Harris a slight edge with a support probability of 0.2925 versus Trump's 0.2451, though wide prediction intervals suggest caution. Additionally, a recent time-trend analysis reveals Harris's support has surged past Trump's, hinting at a potential momentum shift. This combination of analyses offers a nuanced view of the race, balancing poll stability with the evolving dynamics of voter sentiment.

1.1 Software and packages used

We use the statistical programming language R (R Core Team 2023) and the following packages: tidyverse (Wickham et al. 2016), dplyr (Wickham 2023a), readr (Wickham 2023c), ggplot2 (Wickham 2023b), janitor (Baumer et al. 2023), lubridate (Wickham et al. 2023), broom (Robinson et al. 2023), modelsummary (P. 2023), rstanarm (Goodrich et al. 2022), here (Müller and Bryan 2023), bayesplot(Cepeda et al. 2023), patchwork(Pedersen 2020).

1.2 Paper Structure

The remainder of this paper follows a structure Section 2,Section 3 about linear regression model and logistic model,Section 4,Section 5,Section 6, Section A about pollster methodology,Section B about our methodology on survey.

2 Data

2.1 Overview

Our analysis, conducted in statistical programming language R (R Core Team 2023), utilizes polling data from FiveThirtyEight's repository on the 2024 U.S. Presidential election polls (FiveThirtyEight 2024). This dataset includes variables such as pollster, methodology, sample size, poll end_dates, and support percentages 'pct' for candidates, allowing us to examine how support fluctuates over time and in response to poll characteristics for each candidate. To align with our study's focus, we cleaned the dataset to remove inconsistencies and retained only relevant variables, with only focus on data of 2 Candidate "Trump" and "Harris", preparing the support rate for forecasting the possible winner and examine the trend.

2.2 Measurement

This dataset reflects public sentiment on the 2024 Presidential race, gathered through various polling methods,which involves translating public opinion polling into measurable data for analysis.The polling methods including online surveys and phone interviews by organizations like Morning Consult and TIPP. Each poll captures respondent preferences for candidates, sampling strategies, and timing, along with poll related factors, such as sample_size and pollscore, forming structured data points. These variables represent the underlying polling practices and voter sentiment, after cleaning, the data enables us to analyze factors influencing support rates and model time trends in voter preferences, crucial for accurate election forecasting.

2.3 Response variables

In our analysis, we use two response variables to capture different aspects of candidate support. For the linear model, the response variable is `pct`, representing the percentage of respondents indicating support for each candidate in individual polls. This continuous variable allows us to explore how factors like poll quality (`pollscore`) and sample size influence reported support percentages. For the logistic model, we use `candidate_dummy`, a binary variable coded as `1` for Harris and `0` for Trump. This binary outcome enables us to estimate the probability of support for each candidate, providing a probabilistic forecast of each candidate's chance of winning based on aggregated poll characteristics.

2.4 Predictor variables

The model incorporates a set of key predictor variables aimed at refining forecast accuracy by capturing unique attributes of each poll. **Average Support (`pct`)** is a predictor in the second Bayesian logistic model, it provides the support percentage reported by each poll for

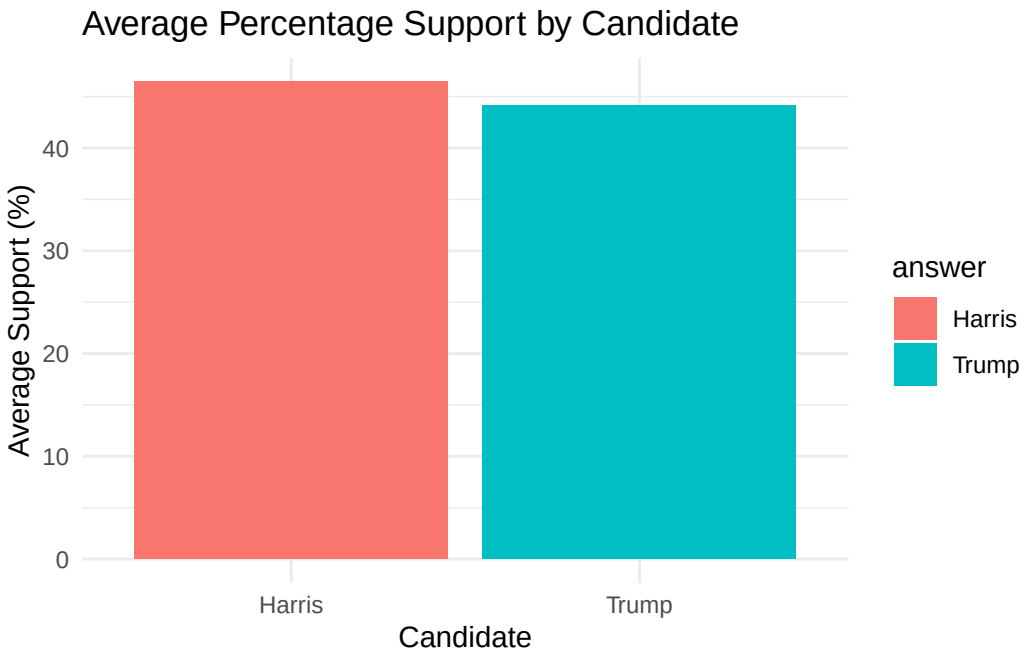

Figure 1: Average Percentage Support(pct) for Harris and Trump

the candidate in question, serving as a direct indicator of voter preference. **Poll Quality (`pollscore`)** reflects the reliability of each poll based on historical accuracy and transparency, with higher scores indicating greater reliability. **Sample Size (`sample_size`)** accounts for the number of respondents per poll, controlling for the variability that smaller samples may introduce. Together, these predictors enable a nuanced analysis of support that accounts for each poll's quality, methodology, and respondent base, leading to a robust and interpretable model of voter preference.

3 Model

3.1 Model Overview

This analysis employs a Bayesian linear regression model and a Bayesian logistic regression model to predict support for each candidate, Harris and Trump, in the 2024 U.S. presidential election. By using a "poll-of-polls" approach, we aggregate data from multiple polls, creating a comprehensive view of voter support. The linear model estimates how poll characteristics like sample size and poll quality influence reported support percentages (pct), capturing trends over time. Meanwhile, the logistic model provides a probabilistic forecast, estimating each

candidate's likelihood of winning based on aggregated support metrics. Both models incorporate Bayesian inference, allowing us to account for variability and uncertainty in polling data and offering a robust election forecast.

3.2 Model selection

We selected the Bayesian linear and logistic models based on their capacity to integrate prior information and address different aspects of our research objectives. The linear model was chosen for its ability to assess predictor effects on support percentages (pct), making it ideal for understanding poll characteristics' impact on support rates. The Bayesian logistic model was selected for its suitability in estimating support probabilities, capturing the competition between candidates. Model fit and Bayesian methods was evaluated use posterior predictive checks, also allow for rigorous model diagnostics, validating the models against observed polling data and enhancing forecast reliability.

3.3 Model set-up

3.3.1 Bayesian linear model

In the Bayesian linear regression model, the response variable 'pct' represents the percentage of respondents supporting each candidate in individual polls. This model can be express as below, In the Bayesian linear regression model, the outcome variable, `pct`, represents the percentage of respondents supporting each candidate in individual polls. This model is expressed as below, where Beta_0 is the intercept, Beta_1 and Beta_2 are coefficients for sample size and poll quality (pollscore), respectively, and epsilon_i represents the error term, capturing random variability in support percentages.

$$\text{pct}_i = \beta_0 + \beta_1 \cdot \text{sample\_size}_i + \beta_2 \cdot \text{pollscore}_i + \epsilon_i$$

3.3.2 Bayesian logistic model

In the Bayesian logistic regression model, the outcome variable is "candidate_dummy", a binary indicator of support for each candidate, where `1` represents Harris and `0` represents Trump. Predictor variables include **support percentage (`pct`)**, **poll quality (`pollscore`)**, **sample size (`sample_size`)**, each representing critical aspects of poll characteristics that impact voter support. The model is specified as follows:

Define y_i as the political preference of the respondent and equal to 1 if Biden and 0 if Trump. Then each Beta_i coefficient measures the effect of the corresponding predictor on the log-odds of

$$\log(\frac{P(y_i = 1)}{1 - P(y_i = 1)}) = \beta_0 + \beta_1 \cdot \text{pct}_i + \beta_2 \cdot \text{sample\_size}_i + \beta_3 \cdot \text{pollscore}_i + \epsilon_i$$

We run the model in R (R Core Team 2023).

3.3.3 Bayesian prior

In our Bayesian framework, priors are set for each model coefficient based on historical data and reasonable assumptions about each predictor's influence. In the Bayesian linear model, The prior for intercept of pct for Harries and Trump is set as 'Normal(45,8)", by historical data, the mean support rate should be around 45, and a variance of 8 yield to great flexibility. In the Bayesian logistic model, a prior for intercept of dummy variable"Candidate_dummy" is set as 'Normal(0, 5)', assumes that baseline support is close to evenly split but allows the data to drive the estimate without excessively constraining it to initial assumptions, and improve flexibility. For both 2 model, we chose "Normal(0,2.5)" as the prior distribution for each predictor. This prior centering around zero (implying no effect) but with a standard deviation of 2.5, allowing flexibility to capture both small and moderate effects without imposing strong assumptions. This choice reflects a balance between letting the data primarily drive coefficient estimates and acknowledging some prior uncertainty. Using Markov Chain Monte Carlo (MCMC) sampling, the model estimates posterior distributions for each coefficient, yielding both mean estimates and credible intervals that capture uncertainty in support probabilities.

$$\begin{aligned}
\text{pct}_i|\mu_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \beta_0 + \beta_1 \cdot \text{sample_size}_i + \beta_2 \cdot \text{pollscore}_i \\
\beta_0 &\sim \text{Normal}(45, 10) \\
\beta_1 &\sim \text{Normal}(0, 2.5) \\
\beta_2 &\sim \text{Normal}(0, 2.5) \\
\sigma &\sim \text{Exponential}(1)
\end{aligned}$$

$$\begin{aligned}
y_i|\mu_i &\sim \text{Normal}(\pi_i) \\
\text{logit}(\pi_i) &= \beta_0 + \beta_1 \cdot \text{pct}_i + \beta_2 \cdot \text{sample_size}_i + \beta_3 \cdot \text{pollscore}_i \\
\beta_0 &\sim \text{Normal}(45, 10) \\
\beta_1 &\sim \text{Normal}(0, 2.5) \\
\beta_2 &\sim \text{Normal}(0, 2.5) \\
\beta_3 &\sim \text{Normal}(0, 2.5)
\end{aligned}$$

3.4 Model justification

Fitness of our Bayesian linear and logistic regression models are examine by the pp_check() function, which use for posterior predictive checks, allowing us to visually assess model fit by comparing observed data to simulated data from the posterior distribution. The (Figure 2) and (Figure 3) show that the two model fit well. And we also checking the convergence of the MCMC algorithm of the Linear regression model(Figure 4). This step helps validate model assumptions and detect any potential misfit.

However, our approach has limitations. We assume that polling errors are normally distributed and that predictors like sample size and poll quality have a linear impact on support rates. This may overlook nonlinear effects or unobserved confounders, such as regional bias or demographic shifts. Additionally, the models assume independence across polls, potentially ignoring correlations from repeated polling by the same pollster or temporal trends. These limitations highlight areas for refinement, such as including more predictors or accounting for pollster-specific effects, to enhance forecasting accuracy.

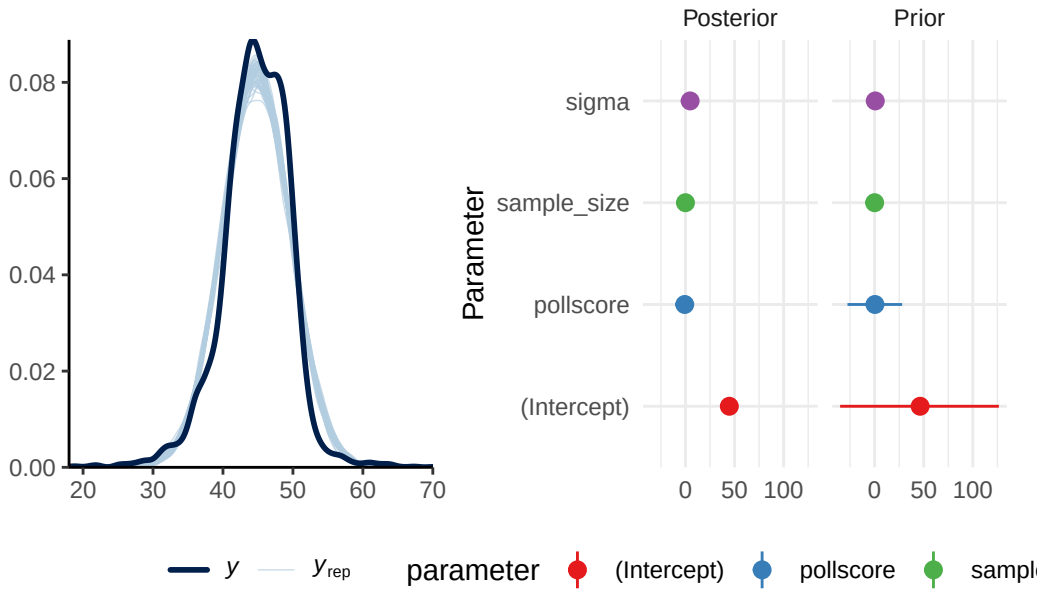

Figure 2: Examining how the Bayesian linear regression model fits, and is affected by, the data

The package rstanarm uses Markov chain Monte Carlo (MCMC) sampling algorithm to obtain samples from the posterior distributions of interest,therefore we can check whether the linear model converge to MCMC algorithm to access the quality of fit by trace plot and Rhat plot. Since the trace plot show horizontal lines that appear to bounce around, with a nice overlap between the chains. The trace plot in (Figure 4) does not suggest anything out of the ordinary.

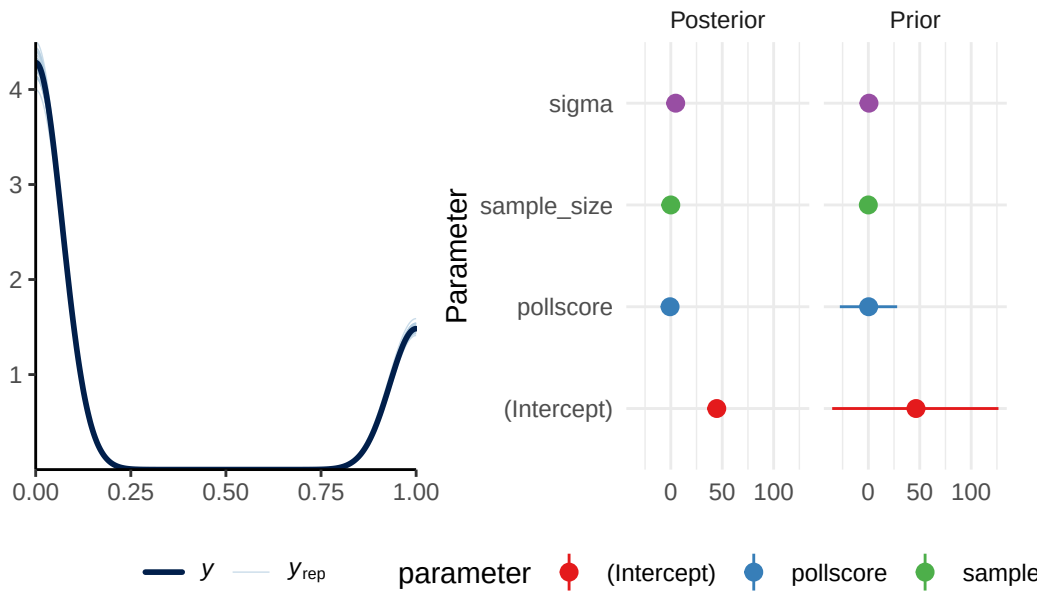

Figure 3: Examining how the Bayesian logistic regression model fits, and is affected by, the data

Similarly,Rhat plot i (Figure 4) show the R hat close to 1, and no more than 1.1. Therefore, the Bayesian linear model seem work well.

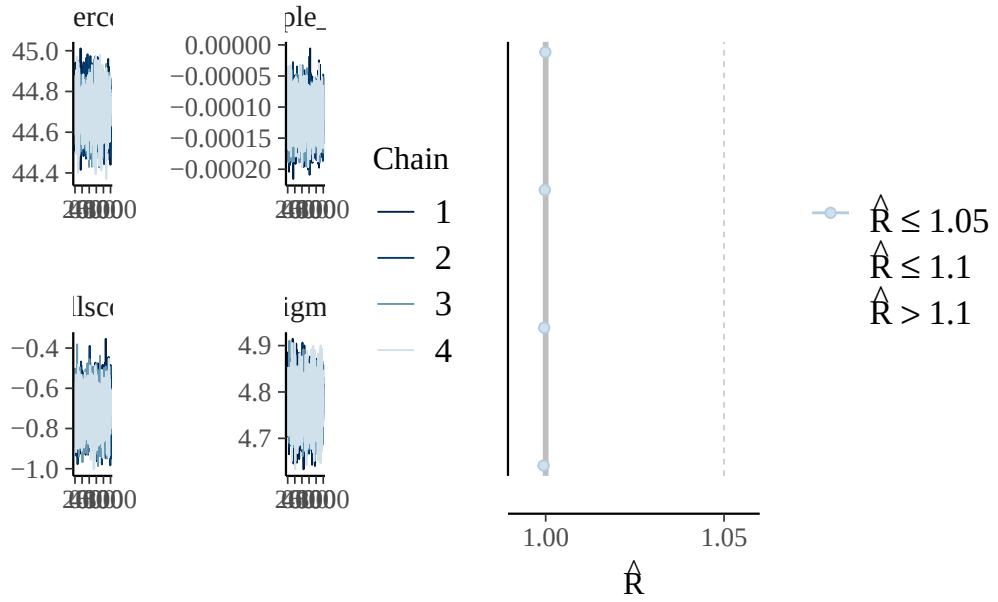

Figure 4: Checking the convergence of the MCMC algorithm by trace plot and rhat plot on linear model

4 Results

```
Model Info:
 function:     stan_glm
 family:       gaussian [identity]
 formula:      pct ~ sample_size + pollscore
 algorithm:    sampling
 sample:       4000 (posterior sample size)
 priors:       see help('prior_summary')
 observations: 5942
 predictors:   3

Estimates:
              mean   sd   10%   50%   90%
(Intercept) 44.7    0.1 44.6  44.7  44.8
sample_size  0.0    0.0  0.0   0.0   0.0
pollscore   -0.7    0.1 -0.8  -0.7  -0.6
sigma        4.8    0.0  4.7   4.8   4.8
```

```
Fit Diagnostics:
           mean   sd   10%   50%   90%
mean_PPD 44.8    0.1 44.6  44.8  44.9

The mean_ppd is the sample average posterior predictive distribution of the outcome variable

MCMC diagnostics
              mcse Rhat n_eff
(Intercept)   0.0  1.0  4990
sample_size   0.0  1.0  4521
pollscore     0.0  1.0  5064
sigma         0.0  1.0  4664
mean_PPD      0.0  1.0  3868
log-posterior 0.0  1.0  1920

For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effectiv
```

The Bayesian linear regression model provides insights into how poll characteristics, specifically sample_size and pollscore, influence reported support percentage (pct). The posterior mean for sample_size is close to zero with a 95% credible interval that centers around zero, indicating that sample size has no significant impact on support percentage in this model. The coefficient for pollscore has a mean of -0.7 with a 95% credible interval from -0.8 to -0.6, suggesting a small but consistent negative effect. This means that as poll quality increases (higher pollscore), reported support slightly decreases, though this effect is modest.

The (Figure 5) MCMC areas plot provides a visualization of the 95% credible intervals for the coefficients in our Bayesian linear regression model. For the `pollscore` coefficient, the credible interval is entirely below zero, aligning with the model summary's indication of a modest negative effect. This suggests that higher poll quality is associated with slightly lower reported support percentages. In contrast, the `sample_size` coefficient is centered around zero, with its credible interval covering zero, indicating no significant effect on `pct`. This visualization reinforces the findings from the summary by highlighting which predictors significantly influence the support percentage.

In (Figure 6), each point represents a poll's support percentage (pct) plotted against its poll quality score (pollscore). The trend line shows a slight negative slope, declining from a support percentage of around 46 to 44, suggesting a weak negative relationship between poll quality and reported support. This trend implies that as poll quality increases, the reported support percentage marginally decreases. However, this effect is minor, indicating that poll quality has limited influence on support levels, possibly reflecting the stability of support trends despite variations in poll quality across different polls.

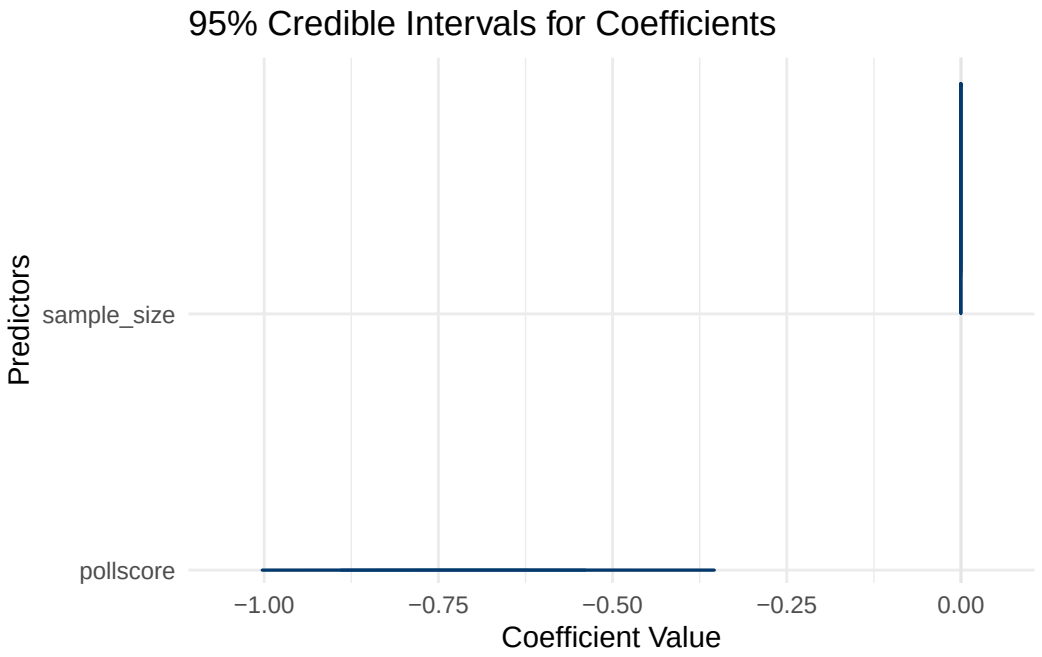

Figure 5: 95% CI for Coefficients of sample_size, pollescore in linear regression model-a MCMC areas plot

In (Figure 7), the relationship between `sample_size` and `pct` is displayed with a log scale on the x-axis to handle the wide range of sample sizes. The trend line is nearly flat, indicating minimal association between sample size and support percentage. This flat trend suggests that larger sample sizes do not significantly impact reported support levels. The result implies that variations in sample size among polls do not meaningfully shift support percentages, reinforcing the stability of support trends across different sample sizes.

Figure (Figure 8) shows the trend in support percentages (pct) for Harris and Trump from 2021 to 2024. Throughout most of this period, Trump maintained a slight lead, with an average support rate around 3% higher than Harris's between August 2022 and June 2024. However, in recent months, Harris's support surged significantly, overtaking Trump with a notable upward spike. Aside from this recent increase, both candidates' support rates generally remain close, showing no large differences except for the recent jump in Harris's support. Interpretation The recent rise in Harris's support suggests a shift in voter preference leading up to the 2024 election, potentially influenced by recent political events or changes in campaign strategy. The overall similarity in support rates indicates that the race remains competitive, with support fluctuating within a narrow range for both candidates. Harris's recent surge, however, may reflect growing momentum, making it a key trend to monitor as the election approaches. This pattern also highlights the importance of temporal trends, as public opinion can shift markedly in the months before an election.

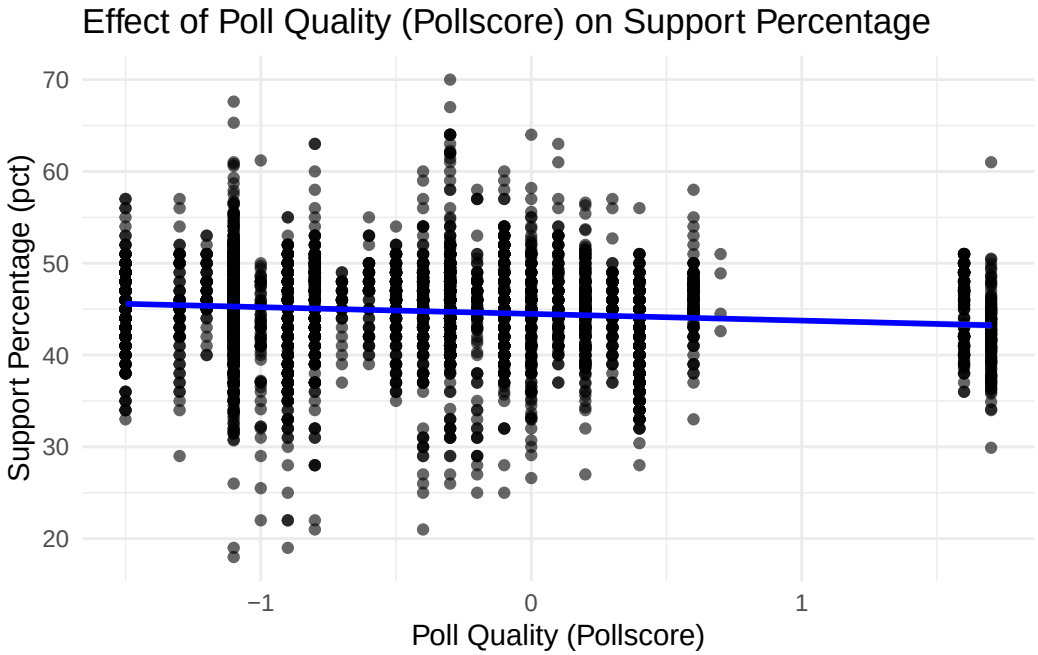

Figure 6: Effect of poll quality (pollscore) on support percentage (pct), showing a minor negative trend suggesting slightly lower support percentages in higher-quality polls.

4.0.1 Forecasting and Prediction Intervals

After fitting the model, we generate forecasts by calculating predicted probabilities for each candidate across polls. To assess forecast reliability, we use 95% prediction intervals for each candidate's aggregated support, providing a range within which actual support levels are likely to fall. Additionally, posterior predictive checks allow us to compare model predictions against observed poll results, enhancing forecast robustness. This setup not only estimates the probability of support but also offers interpretable intervals that reflect uncertainty, making the model's forecasts more reliable for understanding potential election outcomes.

According to (Figure 9), for the linear regression model, the average predicted support percentage for Harris is 44.75% with a 95% prediction interval from 35.41% to 54.10%, while Trump's average predicted support percentage is 44.76%, with a similar 95% interval from 35.42% to 54.11%. These values reflect the average level of support that each candidate garners across polls, without translating this support directly into a probability of winning. Both candidates show nearly identical support percentages (with Trump slightly higher by only 0.01%), indicating that support levels are virtually tied. The prediction intervals, while narrower than the logistic model, still show considerable overlap, underscoring that both candidates have comparable levels of support within the polling data. This model provides insight into the aggregate support each candidate has across polls, rather than directly predicting a winner. It

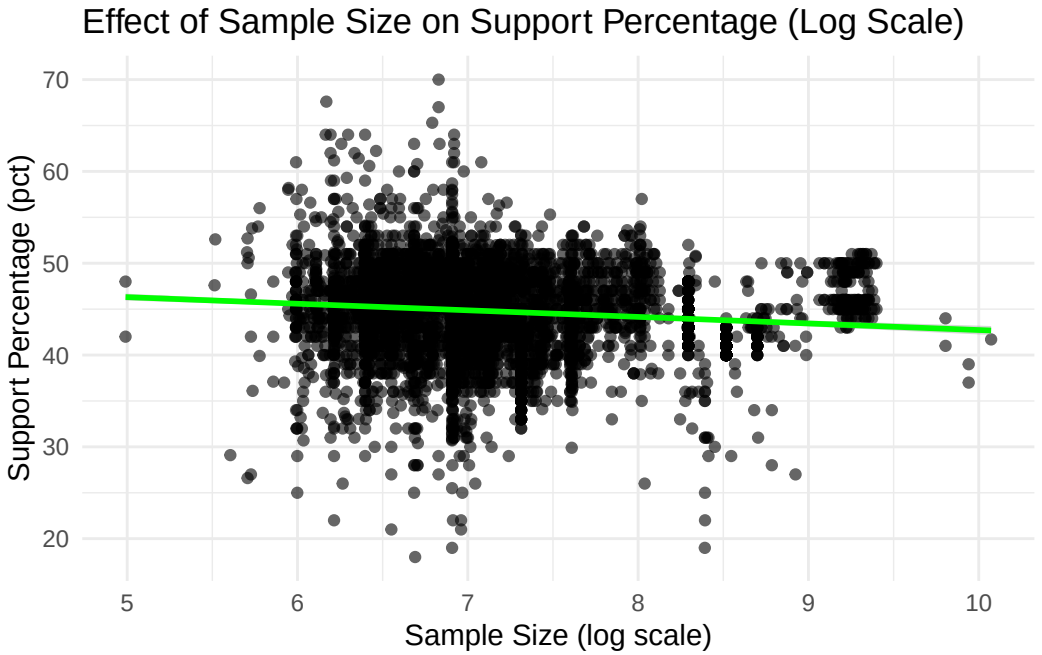

Figure 7: Relationship between log(sample siz)e and support percentage (pct), indicating minimal effect of sample size on rep

suggests stability in support levels and implies that neither candidate has a clear advantage in overall support percentage, highlighting a competitive election landscape.

According to (Figure 10).The Logistic Regression Model with Response "candidate_dummy" yields average predicted probabilities of support for each candidate. Harris has an average predicted probability of support at 0.2925, with a 95% prediction interval spanning from 0 to 1, while Trump has an average probability of 0.2451, with a 95% interval from 0 to 0.998. The probabilities from this model are indicative of the likelihood of each candidate receiving support in a head-to-head comparison. The slightly higher probability for Harris (0.2925 vs. 0.2451) suggests that she has a marginally stronger likelihood of support. However, the wide prediction intervals (spanning almost the full probability range for both candidates) reflect high uncertainty in the forecasts, indicating that both candidates are nearly equally likely to receive support within the observed data variability. Given the high uncertainty captured in the intervals, the model does not strongly favor one candidate as the definitive winner. Instead, it highlights that support is closely matched, with Harris having a slight edge. This approach meets the primary goal of forecasting potential election outcomes based on individual-level preferences.

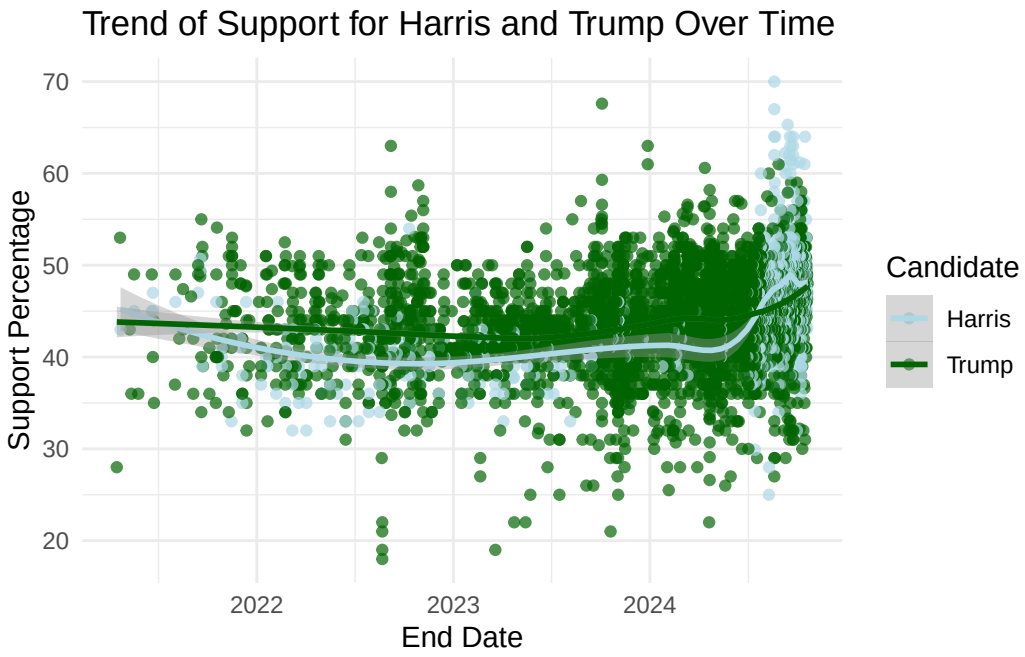

Figure 8: Time trend of support percentage for Harris and Trump in the 2024 U.S. presidential election, showing a recent increase in Harris's support, surpassing Trump.

5 Discussion

5.1 Summary of Key Findings

In this study, we applied Bayesian linear and logistic regression models to forecast the likely outcome of the 2024 U.S. presidential election between Kamala Harris and Donald Trump. Using poll-level data that includes support percentages, sample sizes, and poll quality scores, we aimed to capture a detailed picture of voter preferences. The Bayesian linear regression model examined how poll characteristics impact reported support percentages, providing insight into polling data reliability and potential biases. The Bayesian logistic regression model then estimated each candidate's probability of winning in head-to-head scenarios, allowing us to explore relative candidate advantage based on voter sentiment. Together, these models provide complementary perspectives on support trends and probabilities, enhancing the robustness of our election forecast.

5.2 Bayesian Linear Regression Model Insights (Average Support Levels)

The linear model, which calculates average support percentages, found near-equal levels of support for Harris and Trump, with Harris at 44.75% and Trump at 44.76%. This model

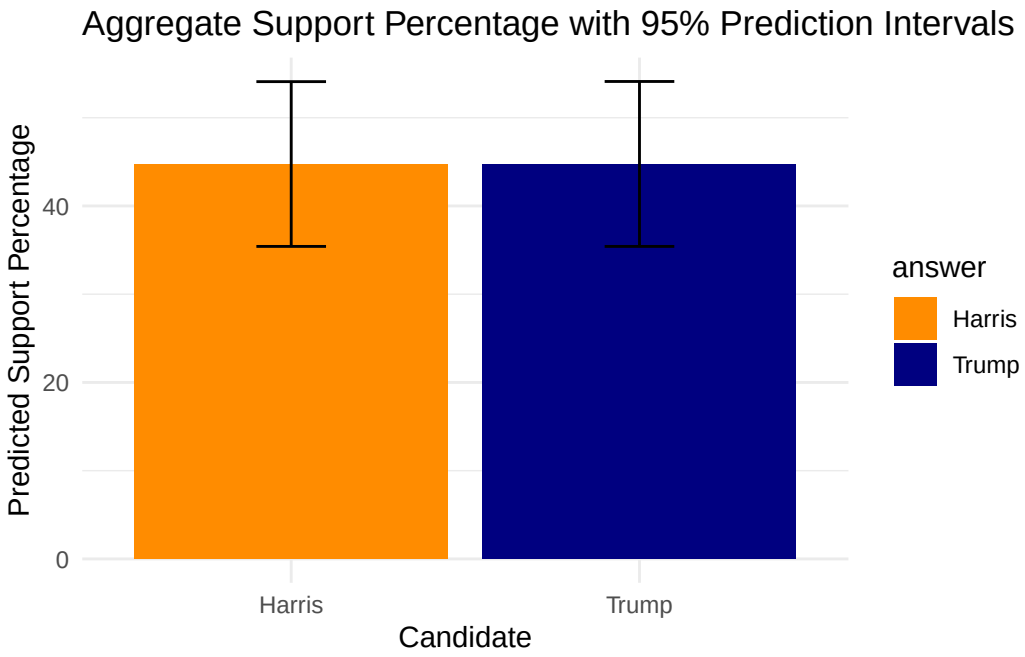

Figure 9: Aggregate predicted support percentages for Harris and Trump with 95% prediction intervals, illustrating the estimated support levels and associated uncertainty for each candidate

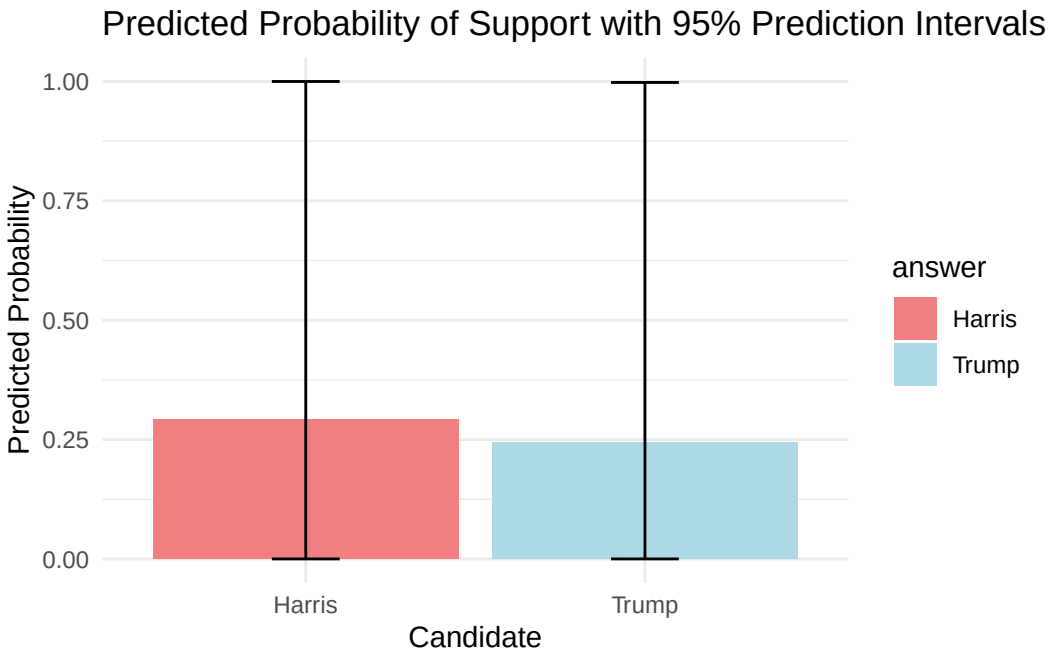

Figure 10: Aggregate predicted probabilities of support for Harris and Trump with 95% prediction intervals, reflecting the estimated likelihood of support for each candidate and the associated uncertainty in the logistic model

suggests that, across polls, the aggregate level of support for each candidate is almost identical, indicating a highly competitive race without a clear frontrunner. These results underscore the stability of overall support levels across various polls and suggest that differences in poll sample sizes and quality scores do not significantly impact the aggregate support. The linear model thus emphasizes the close nature of this election, showing no substantial difference between the candidates in terms of general support. However, while useful for understanding stability in public sentiment, this model does not account for head-to-head dynamics and may miss nuances in individual voter preference shifts.

5.3 Bayesian Logistic Model Insights (Probability of Support)

The logistic model provides a more direct assessment of each candidate's probability of winning by estimating the likelihood of support for Harris over Trump in binary terms. In this model, Harris holds a slight edge with an average probability of 0.2925, compared to Trump's 0.2451. This difference, though small, hints at a slight advantage for Harris in terms of head-to-head support likelihood. However, the wide prediction intervals (0 to 1 for Harris and 0 to 0.998 for Trump) indicate considerable uncertainty, highlighting the limitations of polling data and forecasting in capturing true electoral dynamics. This uncertainty suggests that while Harris may have a marginal lead, the probabilities are close enough that either candidate could still prevail. The logistic model, therefore, supports the notion of a competitive race but suggests that Harris may have a slight edge, albeit with caution due to the high level of forecast variability.

5.4 Time Trend Analysis and Recent Shift

The time-trend analysis provides further context for interpreting these findings. Historically, Trump maintained a slight lead over Harris from mid-2022 to mid-2024, indicating a consistent edge in public support during that period. However, recent polling data reveals a significant rise in Harris's support, allowing her to overtake Trump in the months leading up to the election. This late increase in support for Harris suggests a shift in momentum that could influence the final outcome. When viewed alongside the logistic model's slight probability advantage for Harris, this recent uptick implies that she may be gaining favor as the election nears. The time trend's emphasis on late-campaign dynamics supports the view that Harris has potential momentum, though this trend remains tentative and subject to change as additional data emerge.

5.5 Limitations and Model Assumptions

Despite the strengths of this approach, several limitations should be acknowledged. First, while Bayesian models offer flexibility and the ability to incorporate prior information, the priors chosen may influence the results. We used relatively uninformative priors to minimize bias, but

these selections inherently shape the posterior distributions, potentially affecting probability estimates. Another limitation lies in the data source: the reliance on publicly available polling data from various organizations introduces heterogeneity in sampling methods and question wording, which may contribute to variability that is difficult to account for fully in the models. Additionally, while we used posterior predictive checks to validate model fit, the wide intervals in the logistic model suggest high uncertainty, reflecting the challenge of forecasting elections with inherently noisy data.

The logistic model, in particular, assumes that support for each candidate can be predicted based on a binary outcome, reducing voter preferences to a simple choice between Harris and Trump. This assumption may overlook the complexity of voter behavior and preferences, especially in a multi-dimensional electoral landscape. The linear model, while useful for understanding aggregate trends, may not fully capture the probability of individual voters' support. These model assumptions, though reasonable for a first-pass analysis, indicate that our forecasts should be interpreted with caution.

5.6 Future Directions

Looking ahead, there are several avenues for enhancing this analysis. One potential improvement is to refine the prior distributions for each model based on historical election data, which could provide more informative predictions. Additionally, integrating demographic data, such as voter age, education level, or geographic region, could enrich the models by capturing the diversity within voter bases, making forecasts more reflective of real-world voting patterns. Future work could also incorporate temporal weighting, giving more recent polls higher influence, which may help capture late shifts in sentiment more accurately.

To address the limitations of binary support estimates, future studies could adopt a multinomial or ordinal approach, capturing a range of preferences rather than reducing support to a binary outcome. This would better reflect the complex decision-making process of voters and allow for the inclusion of third-party candidates if applicable. Finally, expanding the time-trend analysis could provide a more dynamic view of support trajectories, informing campaigns about how public opinion evolves in real-time. By refining these models and integrating more granular data, future research can build on the foundations laid here to create even more accurate and actionable election forecasts.

In summary, this paper provides a comprehensive approach to election forecasting through the combined use of Bayesian linear and logistic regression models, highlighting both the stability of support and the subtle competitive edge for Harris in the upcoming U.S. presidential election. While these findings offer valuable insights, continued refinement of the models and exploration of additional predictors would deepen our understanding of the evolving electoral landscape.

6 Conclusion

This study provides a comprehensive analysis of the 2024 U.S. presidential election polling data, examining how poll characteristics—specifically sample size and poll quality—impact reported support percentages and employing Bayesian regression models to forecast the likely winner between Kamala Harris and Donald Trump. By exploring each of these dimensions, we gain a multifaceted view of voter sentiment, poll reliability, and support trends over time. Together, these analyses contribute to a more robust understanding of how polling data can be interpreted in forecasting election outcomes.

Firstly, our analysis of poll characteristics revealed that sample size and poll quality had minimal impact on reported support levels. The Bayesian linear regression model showed only a slight negative association between poll quality and support, with support percentages decreasing marginally as quality increased. The relationship between sample size and support was nearly flat, indicating that variations in sample size did not meaningfully shift support levels. These findings highlight the robustness of aggregated poll data: methodological differences across polls introduce only minor variability, suggesting that aggregated polling data can reliably capture underlying voter sentiment. This insight reinforces the validity of using a "poll-of-polls" approach, where combining multiple polls mitigates the effects of individual poll characteristics and enhances the stability of support estimates.

Secondly, our Bayesian linear and logistic regression models offered complementary perspectives on candidate support. The linear model, which analyzed average support levels, showed that both Harris and Trump have near-identical levels of aggregated support, with Harris at 44.75% and Trump at 44.76%. This indicates an exceptionally close race, with no candidate having a significant edge based on overall support percentages. The logistic model, however, provided a probability-based forecast of each candidate's likelihood of winning in head-to-head scenarios. Here, Harris held a slight but uncertain lead over Trump, with probabilities of 0.2925 and 0.2451, respectively. While this probability difference is small, it suggests that, in competitive settings, Harris may have a slight advantage, though the wide prediction intervals caution against overconfidence in this result. The combination of these models underscores the nuanced nature of the race, with aggregate support levels indicating a tie but probability outcomes hinting at a possible edge for Harris.

Finally, the time-trend analysis added depth to these findings by revealing a significant shift in candidate support over the past two years. From mid-2022 until mid-2024, Trump consistently led Harris by a small margin. However, recent months have seen a substantial increase in Harris's support, surpassing Trump for the first time. This upward trend for Harris may reflect shifting voter dynamics or emerging factors influencing public opinion. When combined with the logistic model's slight probability advantage for Harris, this recent increase in support suggests that Harris may be gaining momentum as the election nears. This trend emphasizes the dynamic nature of voter sentiment and the importance of temporal analysis in capturing shifts that aggregate models alone might overlook.

In conclusion, this study highlights the strengths and limitations of using polling data to forecast election outcomes. By demonstrating that poll characteristics have minimal impact on aggregated support levels, validating the slight yet uncertain edge Harris may have, and revealing the evolving nature of voter sentiment, we provide a nuanced understanding of the 2024 presidential race. The findings underscore the importance of combining different analytical approaches—such as linear and logistic models, alongside time-trend analysis—to capture both stable and dynamic aspects of electoral support. This multidimensional approach enhances the credibility of polling-based forecasts and suggests that while close, Harris's recent momentum could shape the final outcome. Future studies may further explore how temporal shifts impact electoral probabilities, especially in competitive races, to improve the reliability of predictive models in capturing late-stage changes in public sentiment.

A Appendix1

B Appendix2

C Additional data details

D Model details

D.1 Posterior predictive check

References

Baumer, Ben et al. 2023. "Janitor: Simple Tools for Examining and Cleaning Dirty Data." https://cran.r-project.org/web/packages/janitor/index.html.

Cepeda, Gabriel A., Jonah Gabry, Matthew Kay, and Andrew Gelman. 2023. *Bayesplot: Plotting for Bayesian Models*. https://mc-stan.org/bayesplot/.

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

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Müller, Kirill, and Jennifer Bryan. 2023. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

P., Mikhael A. P. L. 2023. "Modelsummary: Create Summary Tables for Model Outputs." https://cran.r-project.org/web/packages/modelsummary/index.html.

Pedersen, Thomas Lin. 2020. *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

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

Robinson, David et al. 2023. "Broom: Convert Statistical Analysis Objects into Tidy Tibbles." https://cran.r-project.org/web/packages/broom/index.html.

Wickham, Hadley et al. 2016. "The Tidyverse." https://www.tidyverse.org.

Wickham, Hadley. 2023a. "Dplyr: A Grammar of Data Manipulation." https://cran.r-project.org/web/packages/dplyr/index.html.

———. 2023b. "Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics." https://cran.r-project.org/web/packages/ggplot2/index.html.

Wickham, Hadley et al. 2023. "Lubridate: Make Dealing with Dates a Little Easier." https://cran.r-project.org/web/packages/lubridate/index.html.

Wickham, Hadley. 2023c. "Readr: Read Rectangular Text Data." https://cran.r-project.org/web/packages/readr/index.html.