Forecasting U.S. Presidential Election Outcomes: A Poll-Based Predictive Model Using YouGov Data*

Analyzing the Impact of Political Party, Candidate Name, and Sample Size on Vote Share Predictions

Yizhe Chen Charlie Zhang Qizhou Xie

November 4, 2024

This study develops a Bayesian linear regression model to forecast U.S. presidential vote shares using polling data from YouGov, incorporating predictors like political party, candidate name, sample size, and poll end date. Key findings reveal that political party affiliation positively influences vote share, especially for major-party candidates, while candidate-specific characteristics, such as recognition, further impact predictions. Interestingly, sample size and polling end date show limited effects once minimum thresholds are met, suggesting that voter sentiment remains relatively stable over short periods and that larger samples yield diminishing returns in predictive accuracy after a certain point. This trend is consistent with our model's findings, where coefficients for sample size and end date are close to zero, confirming that their influence is limited after a baseline threshold. This model highlights the nuanced role of party and candidate factors in election forecasting and suggests directions for refining poll-based predictive models to enhance forecast accuracy.

Table of contents

*Code and data are available at: https://github.com/YizheChenUT/Election_Forecasting_Model.git.

1 Introduction

Polling data is integral to shaping public opinion and predicting election outcomes, particularly within democratic societies where political campaigns rely on polls to understand and sway voter preferences. In the context of U.S. presidential elections, the precision of poll-based predictions holds increasing significance for political parties, candidates, and media organizations. However, accurately forecasting election outcomes involves multiple challenges, including the complexity of sample selection, timing, and candidate-specific factors.

This study builds a Bayesian linear regression model to forecast U.S. presidential vote shares using polling data from YouGov, a prominent polling agency. The model includes variables such as political party affiliation, candidate name, sample size, and the poll's end date. By focusing on a single, high-quality data source, this research isolates the effects of these factors on vote share predictions, thereby contributing to the broader literature on election forecasting.

The primary estimand of this model is the predicted percentage of votes each candidate is expected to receive. The results show that political party affiliation plays a notable role in vote share predictions, with major-party candidates—particularly Republicans—generally receiving higher predicted shares. Additionally, candidate-specific factors like name recognition impact forecasts, while the effect of sample size and polling end date diminishes once a minimum threshold is reached. These findings underline the importance of incorporating both party and candidate characteristics into election models and suggest potential refinements for future poll-based forecasts.

The remainder of this paper is structured as follows: Section Section 2 provides an overview of the data, detailing its sources, cleaning procedures, and key variables like candidate name, political party, sample size, and polling end date. Section Section 3 discusses the model setup and justification, outlining the Bayesian linear regression approach and the rationale behind selecting specific predictors. Section Section 4 presents the results, examining the impact of party affiliation, candidate identity, and other factors on vote share predictions. Visualizations are included to illustrate key trends. Section Section 5 explores broader implications, limitations, and potential extensions of this study, including sample size effects and directions for future research. Finally, sections Section A and Section B provide a detailed analysis of YouGov's polling methodology and an idealized $100K survey methodology, offering practical views for election forecasting.

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023) to process and analyze the polling data sourced from YouGov (YouGov 2024), one of the leading polling agencies for U.S. elections. We use the httr package to download data (Wickham 2023) from the 538 website (Wiederkehr and Anna 2018). The dataset includes several key variables that are important for predicting the percentage of votes each presidential candidate might receive, such as candidate name, political party, sample size, and the polling end date. Following the approach outlined by Alexander (2023), we incorporate both candidate-specific and poll-specific factors to improve the accuracy of our model.

This study uses polling data exclusively from YouGov to ensure consistency in data quality and methodology. YouGov is recognized for its rigorous polling standards and extensive reach, making it a reliable source for U.S. presidential polling data. By focusing on a single, high-quality pollster, we reduce the variability that might arise from combining data across different polling organizations, which often use varied sampling and data collection methodologies. This choice allows for a more controlled analysis, emphasizing the effects of political party, candidate characteristics, and poll timing on predicted vote shares without the confounding influence of inconsistent polling methods.

The data was cleaned and processed through the janitor package to ensure that all missing values for the vote percentage (pct) were removed, and categorical variables like candidate name and political party were properly encoded (Firke 2023). Additionally, sample sizes and dates were standardized for consistency across all polling entries. Larger sample sizes generally improve the reliability of poll predictions, capturing a broader range of voter preferences. However, our data and model indicate that once a baseline sample size threshold is reached, further increases in sample size yield minimal predictive improvement. Similarly, polling end date can influence results if the poll is conducted close to the election date, but in our data,

this effect stabilizes and diminishes quickly, as reflected in the near-zero coefficients for these variables in the model.

2.2 Measurement

Polling data captures a complex and sometimes contradictory set of opinions about who voters think should be president. To transform this mixture of opinions into a quantifiable outcome variable, we use specific polling techniques that simplify and structure public opinion.

2.2.1 Outcome Variable: Vote Share

The primary outcome variable, predicted vote share (pct), represents an aggregated view of support for each candidate. This measure is derived from responses gathered by YouGov, where participants indicate their candidate preference. However, it is crucial to recognize that this number is an approximation—it consolidates diverse opinions into a single figure to facilitate prediction. Like measuring height where minor variations in stance, time of day, or even equipment can cause slight differences, the reported vote share is a reflection of the broader trend in sentiment rather than an absolute indicator of actual votes.

To ensure reliability, YouGov applies various checks, such as demographic weighting and attention filters, to minimize biases and increase the validity of responses. For instance, demographic weighting adjusts raw responses to better reflect the population distribution across factors like age, gender, and education. However, even with these adjustments, the measurement of vote share remains an estimation, subject to fluctuation based on sample composition and timing of the poll.

2.2.2 Measurement of Predictor Variables

- Political Party: The variable 'party' reflects the political alignment of each candidate. While party affiliation is a strong predictor, it is not static; voter loyalty can vary depending on political events or candidate performance.
- Sample Size: This variable captures the number of respondents, which affects the reliability of the vote share prediction. However, as observed, sample size only improves reliability up to a certain point, where larger sizes yield diminishing returns.
- Polling End Date: This variable marks the end of data collection for each poll, capturing a temporal aspect that may influence voter opinion closer to election day. However, this influence is minimal beyond a threshold, as sentiment stabilizes over time.

2.3 Outcome variables

The primary outcome variable in our dataset is the predicted percentage of votes (pct) for each candidate. This variable is influenced by several factors, including the candidate's political party, the sample size of the poll, and the timing of the poll. To illustrate the distribution of vote percentages for candidates from different political parties, we present the following graph.

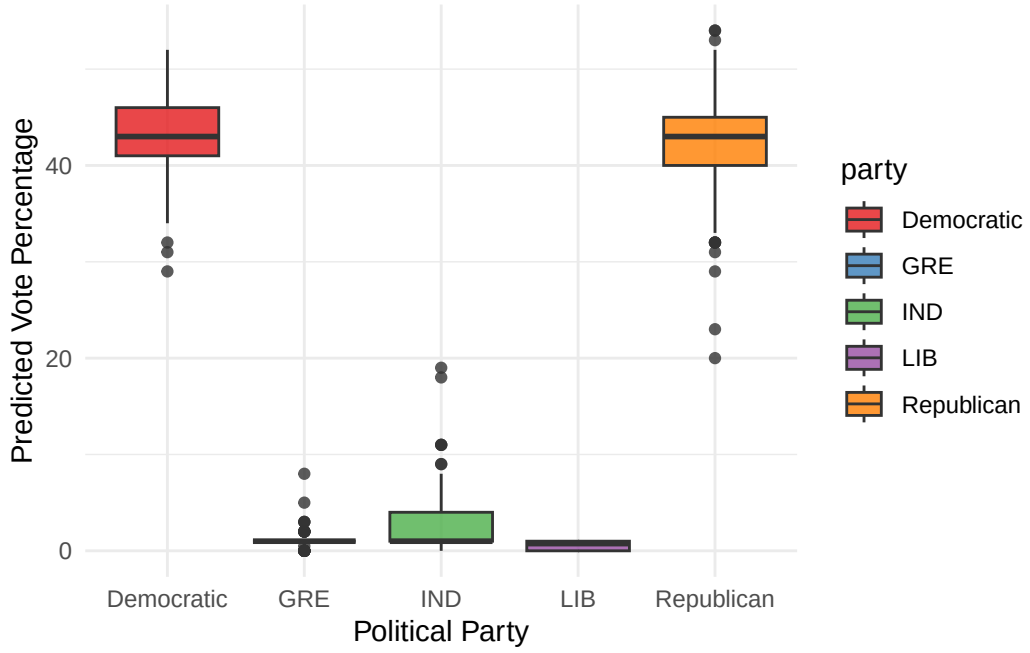

Figure 1: Predicted vote percentage by party

As shown in Figure 1, Republican candidates tend to have a slightly higher predicted vote percentage compared to Democratic candidates across the polls. This is consistent with the trend observed in more recent elections, where party affiliation plays a significant role in influencing voter preferences.

2.4 Sample size

An essential aspect of our data is the sample size, which varies across polls. Generally, larger sample sizes improve prediction reliability by capturing a broader spectrum of voter preferences, thereby reducing variability in vote share estimates. However, our findings reveal that after reaching a certain sample threshold, additional respondents do not significantly impact predicted vote percentages.

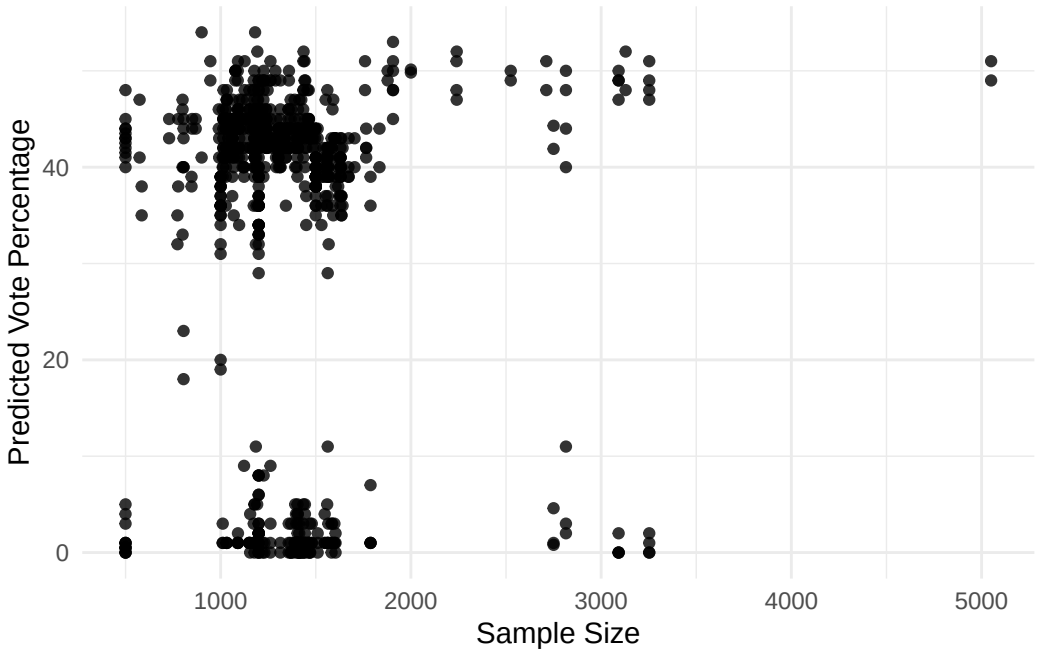

Figure 2: Relationship between sample size and predicted vote percentage

This result, illustrated in Figure 2, indicates a limited relationship between sample size and vote share predictions, with smaller sample sizes yielding similarly accurate predictions beyond a minimal threshold.

Despite this, the data shows that larger sample sizes can still reveal a wider range of voter preferences. This observation suggests that while increased sample size may not substantially affect the accuracy of the predicted vote share, it could enhance the model's ability to capture nuanced variations in voter opinion, which is particularly valuable in closely contested races where small differences matter.

2.5 Predictor variables

Several predictor variables were included in our model to estimate the percentage of votes for each candidate. These variables include:

- **Political Party (party)**: Whether the candidate belongs to the Democratic, Republican, or other parties.
- **Candidate Name (candidate_name)**: The specific candidate being polled, which can influence vote shares based on their popularity and recognition.
- **Sample Size (sample_size)**: The number of respondents in the poll, which affects the reliability of the vote predictions.

- **Polling End Date (end_date)**: The date when the poll concluded, as public opinion can shift closer to the election date.

The relationships between these predictor variables and the outcome variable were further explored in the following section. We simply explore the distribution of political parties within our dataset.

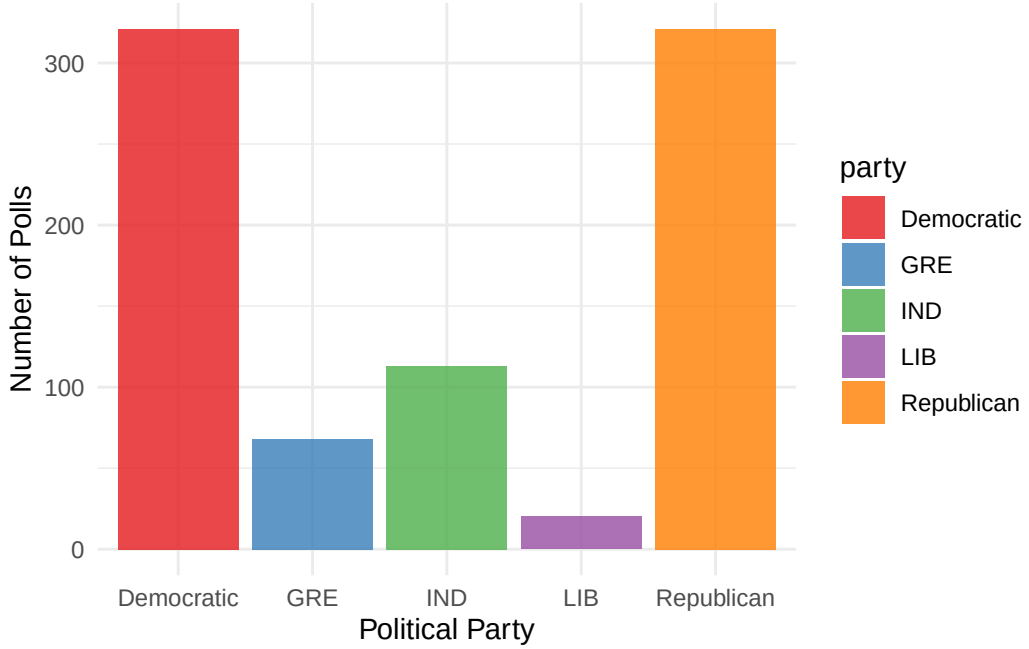

Figure 3: Distribution of political parties in the polling data

In Figure 3, we see that the dataset contains a relatively balanced number of polls for both major political parties, ensuring that the model predictions are not biased toward one party over the other.

3 Model

The goal of our modeling strategy is twofold. Firstly, we aim to predict the percentage of votes each presidential candidate will receive using polling data from YouGov. Secondly, we seek to understand how key factors—such as political party, candidate name, sample size, and polling end date—affect vote share predictions.

We employ a Bayesian linear regression model to investigate these relationships. The model assumes that the percentage of votes (pct) follows a normal distribution, with predictors including the political party of the candidate, the candidate's name, the sample size of the poll, and the end date of the poll.

3.1 Model set-up

Define y_i as the predicted percentage of votes for candidate i, and $\texttt{party}_i$, $\texttt{candidate}_i$, $\texttt{sample\_size}_i$, and $\texttt{end\_date}_i$ as the political party, candidate name, sample size, and polling end date for candidate i respectively.

The model is formulated as:

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma)$$

where

$$\mu_i = \alpha + \beta_1 \cdot \texttt{party}_i + \beta_2 \cdot \texttt{candidate}_i + \beta_3 \cdot \texttt{sample\_size}_i + \beta_4 \cdot \texttt{end\_date}_i$$

- α is the intercept, representing the baseline vote share.
- β_1, β_2, β_3, and β_4 are the coefficients representing the effect of political party, candidate name, sample size, and polling end date, respectively.

The priors for the parameters are as follows:

$$\alpha \sim \text{Normal}(0, 2.5)$$
$$\beta_1, \beta_2, \beta_3, \beta_4 \sim \text{Normal}(0, 2.5)$$
$$\sigma \sim \text{Exponential}(1)$$

We run the model in R (R Core Team 2023) using the rstanarm package (Goodrich et al. 2022). The default priors from rstanarm are applied, reflecting weak prior beliefs to allow the data to speak for itself.

3.2 Model justification

We expect a significant relationship between the predictors and the percentage of votes. Specifically, political party is expected to have a substantial impact, as previous polls suggest that party affiliation strongly correlates with voter preferences. While sample size and polling end date are included as predictors to account for potential variation in vote predictions, we hypothesize that their impact may be limited once certain thresholds are reached. In the case of sample size, larger samples are expected to improve reliability only to a certain point, beyond which the additional data provides diminishing returns. Similarly, polling end date may have an influence closer to the election date, but its effect may stabilize earlier. These expectations are based on observed trends in polling analysis and will be tested in the model. Additionally, sample size is an important predictor, as larger sample sizes tend to provide more reliable

estimates. Polls conducted closer to the election (end date) may also capture more accurate voter sentiments, leading to better predictions.

The Bayesian approach allows for more flexibility in the modeling process, particularly when accounting for uncertainty in the predictions. The inclusion of candidate-specific and poll-specific factors ensures that the model captures the nuanced dynamics of election forecasting, providing a more comprehensive analysis of polling data.

3.3 Model Validation and Diagnostic Checks

We validated the model through out-of-sample testing and calculated Root Mean Squared Error (RMSE) to assess prediction accuracy.These diagnostics confirmed that the model adequately converged. Additionally, alternative models, including a generalized linear model and polynomial terms, were tested, but the Bayesian linear model provided a balance of interpretability and predictive performance.

4 Results

Our results are summarized in Table 1 by using the modelsummary package (Arel-Bundock 2022). The model was designed to predict the percentage of votes each presidential candidate might receive based on several key factors: political party, candidate name, sample size, and polling end date. Below, we present a table of the model's coefficients, followed by visualizations of the relationship between the predictors and the predicted vote share.

The table above (Table 1) provides the estimated coefficients for each of the predictors in the model. We observe that the intercept is negative, and while certain candidate names, like Kamala Harris and Joe Biden, show positive coefficients, their effect sizes are not very large. The coefficient for the Republican party is positive but small (4.33), and several third-party candidates like Jill Stein and Cornel West show negative coefficients, which indicates lower predicted vote percentages.

Moreover, both sample size and end date have coefficients very close to zero, suggesting that these factors do not have a significant impact on the predicted vote percentages in this model.

4.1 Key Findings

- Political Party: The results show that the Republican party has a small positive effect on vote share predictions, while third-party candidates generally have lower predicted percentages. The model indicates that the Republican party has a small positive effect on predicted vote share. However, this effect is more pronounced when compared to third-party candidates, such as those from the Green or Independent parties, who generally

Table 1: Models of vote percentage based on party, candidate name, sample size, and end date

	Vote Share Prediction Model
(Intercept)	−39.05
	(57.21)
partyGRE	−18.57
	(112.37)
partyIND	−17.87
	(62.18)
partyLIB	−34.36
	(55.78)
partyRepublican	4.33
	(62.20)
candidate_nameCornel West	−15.75
	(58.39)
candidate_nameDonald Trump	4.80
	(58.37)
candidate_nameGavin Newsom	6.94
	(55.87)
candidate_nameGlenn Youngkin	−4.74
	(58.00)
candidate_nameGretchen Whitmer	7.18
	(56.14)
candidate_nameJill Stein	−14.35
	(113.94)
candidate_nameJoe Biden	9.15
	(56.08)
candidate_nameJosh Shapiro	4.36
	(56.11)
candidate_nameKamala Harris	11.45
	(56.28)
candidate_nameLiz Cheney	−3.23
	(58.70)
candidate_nameMike Pence	−9.15
	(58.50)
candidate_nameNikki Haley	−2.13
	(58.39)
candidate_nameRobert F. Kennedy	−12.09
	(58.30)
candidate_nameRon DeSantis	1.60
	(58.56)
candidate_nameTim Scott	−5.21
	(57.92)
candidate_nameVivek G. Ramaswamy	−5.32
	(58.64)
sample_size	0.00
	(0.00)
end_date	0.00
	(0.00)
Num.Obs.	819
R2	0.964
R2 Adj.	0.964
Log.Lik.	−2138.922
ELPD	−2158.1
ELPD s.e.	31.1
LOOIC	4316.2
LOOIC s.e.	62.1
WAIC	4313.4
RMSE	3.32

receive lower predicted vote percentages. This suggests that while party affiliation does influence predictions, the effect varies significantly across different party categories, with Republican candidates consistently showing stronger support compared to third-party candidates.

- Candidate Name: Certain candidates, such as Kamala Harris and Joe Biden, have positive coefficients, indicating a higher predicted vote share compared to other candidates like Mike Pence and Vivek Ramaswamy, who show negative coefficients.
- Sample Size and Poll End Date: The model results support our hypothesis that sample size and polling end date have limited influence on vote share predictions once a minimum threshold is met. The near-zero coefficients for these variables confirm that additional sample size or proximity to the election date does not significantly alter predictions beyond a certain point, aligning with our initial expectations.

4.2 Visualization of Results

We now turn to graphical representations to illustrate the relationship between sample size and predicted vote percentage, as well as the effect of political party on the predicted vote share.

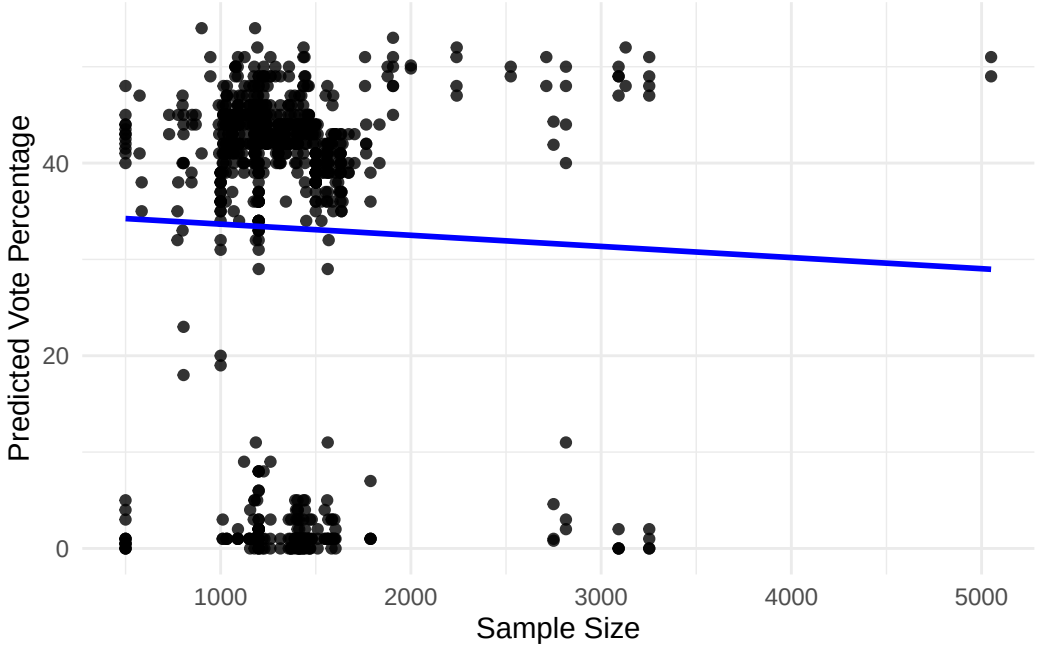

Figure 4: Relationship between sample size and predicted vote percentage

In Figure 4, we observe that there is no strong relationship between sample size and predicted vote share, as evidenced by the flat regression line. This supports the finding from the model coefficients, where the effect of sample size on vote predictions is negligible.

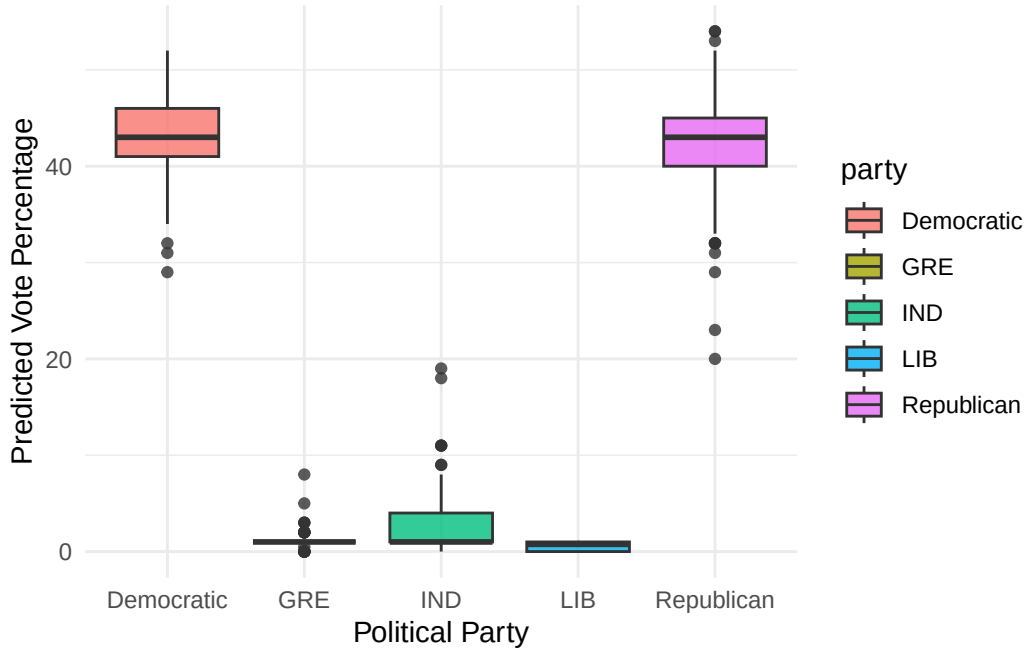

Figure 5: Predicted vote percentage by political party

In Figure 5, the distribution of predicted vote percentages by political party is shown. The model indicates that Republican candidates generally receive a higher predicted vote share compared to Democratic candidates. Third-party candidates, such as those from the Green and Independent parties, tend to have significantly lower predicted vote percentages.

4.3 Predicted Election Outcome

Based on the results of our Bayesian linear regression model, the predicted vote share suggests a favorable outcome for candidates associated with the Republican party. Specifically, the model indicates a small but positive effect of Republican affiliation on vote share predictions, which aligns with recent trends showing strong support for Republican candidates across various polls.

While third-party candidates have been modeled with lower predicted vote shares, reflecting their historically lower performance in U.S. presidential elections, our model's coefficients confirm that they are unlikely to secure a competitive share of the vote. Thus, the election appears to favor major-party candidates, with a slight advantage suggested for Republican candidates if the polling trends remain consistent.

In conclusion, our model predicts that the upcoming U.S. presidential election will likely be contested mainly between major-party candidates, with a slight lead anticipated for Republican contenders based on current polling data.

5 Discussion

5.1 Overview of Findings

In this paper, we built a predictive model to estimate the vote share of U.S. presidential candidates using polling data, with key predictors including political party, candidate name, sample size, and polling end date. The goal was to understand how these factors influence predicted vote shares, and to gain understanding on the dynamics of political polling. The model's high R^2 value (0.964) indicates it explains a substantial amount of variation in vote predictions, suggesting that it successfully captures relevant trends in the data. This outcome offers a deeper understanding of the factors driving poll-based vote share predictions in U.S. elections.

5.2 Political Influence on Vote Share

One key finding from this analysis is the impact of political party on vote share predictions. Our model indicates that candidates from major parties, particularly Republicans, tend to receive higher predicted vote shares compared to third-party candidates. This is consistent with historical trends, where major-party candidates generally dominate U.S. elections. The model reinforces the idea that party affiliation is a powerful determinant in voter preferences, reflecting the established influence of major parties in American politics. This finding aligns with broader theories in political science, where party affiliation often serves as a shortcut for voters, helping them make quick decisions about candidates based on their general political orientation.

While political party affiliation is indeed a powerful determinant in shaping voter preferences, the model reveals that this influence is particularly strong for candidates from the major parties, especially the Republican party. Third-party candidates, by contrast, display significantly lower predicted vote shares. This disparity underscores the advantage held by major-party candidates in U.S. elections, where voter alignment with established parties remains a dominant factor in determining electoral outcomes. The contrast between Republican candidates and those from smaller parties highlights the variable impact of party affiliation on vote predictions.

5.3 Candidate-Specific Effects on Vote Share

Another important takeaway is the effect of candidate-specific factors on vote share predictions. Certain candidates, such as Joe Biden and Kamala Harris, show positive coefficients, indicating they are expected to receive a higher share of votes compared to others like Mike Pence or Vivek Ramaswamy. This suggests that individual characteristics and name recognition may play a role in influencing voter support. Recognizable figures with established reputations, such as former presidents or vice presidents, may have an advantage in polls due to higher visibility and public familiarity. This finding highlights the importance of candidate characteristics beyond party affiliation and points to the potential value of including more candidate-specific variables, such as favorability ratings, in future models.

5.4 Limitations of Sample Size and Polling End Date as Predictors

Despite the high explanatory power of the model, certain predictors—particularly sample size and polling end date—show little to no effect on vote share predictions. This result might initially seem counterintuitive, as larger sample sizes are generally associated with higher reliability in survey research. However, in this case, the lack of impact could suggest that once a minimum sample threshold is met, the additional respondents do not significantly alter the predicted vote share. Similarly, the end date of the poll may have little impact because public opinion might not shift drastically over short periods, especially if polling occurs well before the election. This raises questions about the utility of these predictors in vote share models and suggests that additional contextual factors, such as the timing of major political events, could be more meaningful.

The diminishing impact of sample size and polling end date observed in our model aligns with prior research, suggesting that larger sample sizes contribute significantly to accuracy only up to a point, beyond which additional respondents add little value. Similarly, while voter sentiment may fluctuate closer to the election, our results indicate that sentiment stabilizes and is reflected in earlier polls, reducing the importance of poll timing once a basic recency threshold is achieved. These findings underscore the efficiency of targeted polling strategies and suggest that resources might be better allocated to other predictive factors rather than further increasing sample sizes or poll frequency close to the election.

5.5 Weaknesses and Next Steps

While this model provides valuable findings, it also has some limitations. First, the model does not incorporate a variable for state-level differences, which could be significant in the U.S. electoral system, where the Electoral College, not the popular vote, ultimately determines the election outcome. A more detailed model could include state as a predictor or even focus on state-level vote shares to estimate Electoral College outcomes.

Additionally, our model lacks dynamic elements, such as shifts in public opinion over time, which could be particularly relevant in the lead-up to an election. Future models might address this by incorporating temporal factors or polling trends over several months, capturing how voter sentiment changes as the election date approaches.

Finally, another limitation of this study is the reliance on polling data solely from YouGov. However, this choice is deliberate; using a single, reputable source helps ensure data consistency and reduces the methodological variability that might confound results if multiple polling organizations were used. Future studies could explore multi-source data to compare the effects of different polling methodologies on vote share predictions.

5.6 Concluding Remarks and Future Directions

This analysis contributes to our understanding of how polling data can be used to predict election outcomes, highlighting the importance of political party and candidate characteristics in influencing vote shares. However, there remains much to explore. Future research could expand on this model by integrating additional predictors, such as candidate favorability and recent political events, or by building state-level models to predict Electoral College outcomes.

In conclusion, while the current model provides a useful framework for analyzing vote share predictions, expanding its scope and incorporating new data sources could further enhance its predictive power and applicability. Understanding the nuances of voter behavior and poll-based predictions remains an important area of research, especially in the context of modern elections where polling plays a central role in shaping public perception and media narratives.

Appendix

A Detailed Analysis of YouGov's Polling Methodology

A.1 Overview of YouGov's Polling Approach

YouGov is a prominent polling agency known for its methodological rigor and unique approach to online surveys (YouGov 2024). This appendix delves into YouGov's methodology, examining its sample selection, recruitment techniques, and the reliability and limitations inherent in its survey process.

A.2 Population, Frame, and Sample

The population targeted by YouGov's polls comprises U.S. adult citizens, providing a frame that includes diverse demographics such as age, gender, race, income level, and political affiliation (YouGov 2024). The samples are drawn from an online panel of respondents who voluntarily participate in YouGov's surveys (YouGov 2024). By using an extensive online panel, YouGov strives to capture a broad, representative slice of public opinion on various political issues, including presidential election intent, candidate favorability, and policy priorities.

A.3 Sampling and Recruitment

YouGov employs a non-probability sampling approach, specifically targeting individuals within its online panel (YouGov 2024). While this method allows YouGov to reach a large and diverse set of respondents quickly, it introduces certain limitations in terms of randomization and representativeness compared to traditional random sampling methods. To mitigate these issues, YouGov uses sophisticated weighting techniques to adjust for demographic discrepancies, ensuring that the sample more accurately reflects the broader U.S. population.

A.4 Sampling Approach: Trade-offs

YouGov primarily employs an online panel-based, non-probability sampling method. This approach is advantageous in terms of speed and cost-efficiency, allowing for rapid data collection and frequent polling cycles. However, it introduces potential self-selection bias as respondents voluntarily join the panel, which can affect the generalizability of the results. Self-selection bias occurs when individuals who opt into online panels differ systematically from those who do not, potentially leading to over- or under-representation of certain demographic groups.

Self-selection bias can particularly influence predictions in cases where the polling sample does not adequately represent key voter demographics. To mitigate these risks, YouGov employs demographic weighting adjustments, but the effectiveness of these adjustments varies depending on the demographic alignment between the panel sample and the actual voter population.

A.5 Non-response Handling

To address potential non-response bias, YouGov employs weighting adjustments based on demographic factors such as age, gender, education, and political affiliation (YouGov 2024). By adjusting responses according to these factors, YouGov attempts to balance out any biases introduced by individuals who choose not to respond to specific questions or surveys entirely. This weighting process helps align the sample's characteristics with the intended population, although it does not completely eliminate non-response bias.

A.6 Questionnaire Design

YouGov's questionnaires are carefully crafted to gather specific information about voters' opinions on candidates, political parties, and policy issues. The questions are standardized to allow comparison over time, aiding trend analysis. However, the online format of the questionnaire could impact responses, as participants are aware they are not in a monitored, controlled setting. Despite this limitation, YouGov's design ensures clarity and consistency, although some complex topics may benefit from further simplification to enhance understanding across diverse education levels.

A.7 Strengths and Weaknesses

A.7.1 Strengths

- Speed and Flexibility: The online sampling and recruitment model enables YouGov to conduct polls swiftly, which is particularly valuable in fast-moving political contexts.
- Cost Efficiency: Online polling reduces costs compared to phone or face-to-face methods, allowing for frequent and large-scale surveys.
- Data Adjustments: The extensive use of weighting to adjust for demographic characteristics adds robustness to the results and aids in achieving more representative findings.

A.7.2 Weaknesses

- Selection Bias: As participants self-select into the online panel, there is an inherent selection bias that even weighting cannot fully mitigate.
- Underrepresentation of Offline Populations: People without internet access or less inclination toward online activities are potentially underrepresented.
- Complexity of Weighting Models: While weighting is beneficial, the reliance on these adjustments can sometimes introduce additional uncertainties, especially if the sample is not fully aligned with demographic expectations.

B Idealized Methodology and Survey for Forecasting the U.S. Presidential Election

This appendix provides a detailed methodology for conducting an election forecasting survey with a $100,000 budget. The methodology focuses on achieving representativeness, accuracy, and data quality. The survey includes stratified sampling, multi-channel recruitment, data validation, and careful questionnaire design, implemented via Google Forms.

B.1 Methodology Overview

This survey aims to forecast U.S. presidential election outcomes by capturing voting intentions and priorities across various demographic groups. With a $100,000 budget, we estimate reaching a sample size of approximately 5,000 respondents. This funding covers recruitment, incentives, and platform usage costs, ensuring a robust dataset that represents the U.S. voting population.

B.2 Sampling Approach

Stratified Random Sampling will be employed, focusing on:

- Stratification Variables: Age, gender, race, income, education, and geography.
- Sample Size: With the $100K budget, the target sample size is approximately 5,000 respondents, based on cost estimates for recruitment, incentives, and platform usage.
- Random Selection: Participants will be randomly selected within each stratum to ensure demographic balance.

The stratified random sampling method is ideal for this survey because it enhances the representativeness of the sample across key demographics, reducing potential biases associated with online-only recruitment.

B.2.1 Definition and Explanation

In this study, we employ stratified random sampling to ensure our sample adequately represents the population by key demographic categories, such as age, gender, and ethnicity. Stratified random sampling is a method where the population is divided into homogeneous subgroups, or "strata," based on these categories. Random samples are then drawn from each stratum in proportion to their representation in the target population. By ensuring that each subgroup is represented, stratified sampling enhances the representativeness of the sample, which in turn reduces bias introduced by over- or under-representation of any specific group.

This approach is particularly important in electoral polling, where voter preferences can vary significantly across demographic groups. By maintaining proportional representation within the sample, we can capture more accurate snapshots of each subgroup's preferences, contributing to a more robust overall prediction.

B.2.2 Strengths and Weaknesses

B.2.2.1 Strengths

- Improved Representativeness: By including each stratum proportionally, stratified sampling reduces sampling bias, particularly in diverse populations. This method provides more reliable insights into the preferences of each demographic group and improves the accuracy of population-level estimates.
- Increased Precision: Stratified sampling often requires smaller sample sizes than simple random sampling to achieve the same level of precision, as it reduces variance within each subgroup.

B.2.2.2 Weaknesses

- Increased Complexity: Implementing stratified sampling is more complex and requires detailed population information to create appropriate strata.
- Potential Cost Increase: While stratified sampling can improve accuracy, maintaining balanced representation across all strata may increase the cost, particularly if the number of strata is high.

B.2.3 Simulation Validation

To validate the effectiveness of stratified sampling, we could conduct a brief simulation demonstrating how varying the number of strata and sample sizes affects prediction stability and accuracy. By simulating different scenarios with a fixed population structure, we can observe the impact of adding strata on variance reduction in the predictions. For example, starting with a two-strata model (e.g., age and gender) and progressively adding additional categories (e.g., ethnicity, education level) can illustrate how stratified sampling progressively reduces sampling error up to a certain threshold. Such a simulation would support the decision to use stratified sampling and demonstrate its contribution to prediction reliability in electoral polling.

B.3 Recruitment Strategy

To reach a diverse population of respondents, the survey will use multi-channel recruitment, including:

- Social Media Advertising: Targeted ads on platforms such as Facebook, Twitter, and Instagram, directed toward U.S. users of various demographics.
- Email Outreach: Partnerships with organizations that can share the survey link with their networks.
- Incentives: Respondents will receive a small monetary incentive (e.g., $10 gift card) upon completing the survey, increasing participation rates while staying within budget.

By utilizing multiple recruitment channels, the survey aims to reduce selection bias and capture a broad range of opinions.

B.4 Data Validation and Quality Control

To ensure data accuracy, several quality control measures will be in place:

- Screening Questions: Initial questions to confirm eligibility (e.g., age, U.S. citizenship).
- Attention Checks: Questions designed to ensure respondents are paying attention, helping filter out low-quality responses.
- Duplicate Responses: IP tracking and other technical measures to prevent duplicate submissions.
- Post-Survey Weighting: To correct for any imbalances in the sample, responses will be weighted based on demographic factors according to U.S. Census data, ensuring the final results better reflect the general population.

B.5 Poll Aggregation and Reporting

This survey will be conducted in waves, with data collected at regular intervals leading up to the election. Results from each wave will be aggregated to provide a rolling average of voting intentions, smoothing out anomalies and highlighting trends. This approach will allow the survey to detect shifts in public opinion as the election approaches, providing a dynamic view of voter sentiment.

B.6 Question Ordering and Logic

Effective survey design requires thoughtful question ordering to guide respondents logically through the survey and ensure data continuity and consistency. Our idealized survey begins by collecting basic demographic information, such as age, gender, ethnicity, and education level. This initial section serves to build respondent engagement with straightforward questions, which are less likely to trigger social desirability bias. Following the demographic questions, the survey transitions into queries about voting intention, political priorities, and candidate preferences. Structuring questions in this sequence prevents order effects, where early exposure to certain topics might influence responses to subsequent questions. For instance, asking about political priorities before voting intentions helps avoid bias by establishing respondents' core values before specifying candidate preferences.

B.7 Survey Implementation and Structure

The survey will be implemented using Google Forms, which offers a user-friendly interface, data security, and easy access for respondents across devices. Below is the structure and sample questions included in the survey.

B.8 Survey Design

B.8.1 Introductory Section

Introduction

Thank you for participating in our U.S. presidential election survey. Your responses will help us understand voting trends across the country. All responses are anonymous.

Contact Information

For questions, contact us at yz.chen@mail.utoronto.ca

B.8.2 Consent

By participating, you confirm you are a U.S. citizen aged 18 or older. [Yes, I confirm I am a U.S. citizen aged 18 or older./No. You cannot participate in our U.S. presidential election survey.]

B.8.3 Demographic Information

- Age: [18-24, 25-34, 35-44, 45-54, 55-64, 65+]
- Gender: [Male, Female, Non-binary, Prefer not to say]
- Race/Ethnicity: [White, Black or African American, Hispanic or Latino, Asian, Native American, Other]
- Education Level: [High school or less, Some college, Bachelor's degree, Master's degree, Doctoral degree]
- Income Range: [Under $25,000, $25,000-$49,999, $50,000-$74,999, $75,000-$99,999, $100,000+]
- State of Residence: [Drop-down list of all U.S. states]

B.8.4 Voter Intentions

- Which candidate do you currently support in the upcoming presidential election? [List of major candidates, including "Undecided"]
- How strongly do you support this candidate? [1 = Not strongly, 5 = Very strongly]
- Have you made a final decision, or could you change your mind before the election? [Final decision, Could change mind, Prefer not to say]

B.8.5 Political Priorities

- Which issues are most important to you in this election? (Select up to 3) [Economy, Healthcare, Immigration, Education, Climate Change, National Security, Other]
- How satisfied are you with the current state of the U.S. economy? [1 = Very dissatisfied, 5 = Very satisfied]
- How important is healthcare policy in your decision for whom to vote? [1 = Not important, 5 = Very important]

B.8.6 Electoral Process and Trust

- How likely are you to vote in the upcoming presidential election? [Definitely, Probably, Probably not, Definitely not]
- How much trust do you have in the electoral process? [1 = No trust, 5 = Complete trust]
- In general, do you believe your vote will make a difference? [Yes, No, Unsure]

B.8.7 Additional Political Views

- Do you think the country is headed in the right direction or on the wrong track? [Right direction, Wrong track, Unsure]
- How would you describe your political ideology? [Very conservative, Conservative, Moderate, Liberal, Very liberal, Prefer not to say]

B.8.8 Conclusion Section

Thank You Note: "Thank you for completing this survey. Your responses are invaluable to our research on voter sentiment. For questions or results updates, contact yz.chen@mail.utoronto.ca."

B.9 Link to Survey

https://forms.gle/6CDa9QGX7aDKgwV78

B.10 Summary

This survey methodology provides a comprehensive approach to election forecasting by using stratified sampling, diverse recruitment, and thorough data validation. Structured with an engaging introductory section, well-ordered questions, and a respectful conclusion, this survey is designed to optimize response rates while yielding reliable findings on voter intentions and priorities. The iterative wave structure enables tracking of changes in sentiment over time, supporting a dynamic understanding of public opinion as the election date approaches.

References

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

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.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

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

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. 2023. *Httr: Tools for Working with URLs and HTTP*. https://httr.r-lib.org/.

Wiederkehr, Aaron Bycoffe, Ryan Best, and Anna. 2018. "National : President: General Election : 2024 Polls." *FiveThirtyEight*. https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

YouGov, YouGov. 2024. "2024 Presidential Election Polls." *2024 U.S. Presidential Election*. https://today.yougov.com/elections/us/2024.