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

October 23, 2024

This paper presents a linear regression model that predicts the percentage of votes for presidential candidates based on polling data from YouGov. The model incorporates key predictors such as political party, candidate name, sample size, and polling end date. Our results show that both political party affiliation and sample size significantly affect the predicted vote share, with Republican candidates often receiving higher predicted percentages. This study highlights the importance of integrating candidate-specific and poll-specific factors to improve election forecasts, providing useful insights for political analysts and pollsters.

Table of contents

1 Introduction

Polling data plays a critical role in shaping public opinion and forecasting election outcomes, particularly in democratic societies where political campaigns rely heavily on polls to gauge voter preferences. In the context of the U.S. presidential elections, the accuracy of poll-based predictions has become increasingly important for political parties, candidates, and media organizations. Despite the growing reliance on polling, there are several challenges in making accurate predictions, such as sample selection, timing of the poll, and candidate-specific factors.

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

This paper aims to build a predictive model using polling data from YouGov, one of the most recognized polling agencies, to forecast the percentage of votes for presidential candidates. The model incorporates variables such as political party, candidate name, sample size, and the poll's end date. By focusing on a single pollster, this study seeks to analyze the effect of these factors on vote share predictions and contribute to the broader literature on election forecasting.

The estimand of the model is the predicted percentage of votes that each candidate is expected to receive, which is influenced by the selected variables. Our findings suggest that both the political party of a candidate and the sample size of the poll significantly affect vote share predictions. These results are important as they provide valuable insights into how different factors influence the accuracy of polling predictions, potentially improving the quality of future election forecasts.

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2....

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, one of the leading polling agencies for U.S. elections. 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.

The data was cleaned and processed 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. Additionally, sample sizes and dates were standardized for consistency across all polling entries.

2.2 Measurement

Polling data, in essence, represents a snapshot of public opinion at a given time. In our case, the dataset captures public opinion on various presidential candidates based on responses collected via YouGov's online surveys. These responses are translated into numerical entries in the dataset, such as the predicted percentage of votes a candidate might receive (pct), the number of respondents (sample_size), and the poll's end date (end_date).

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.

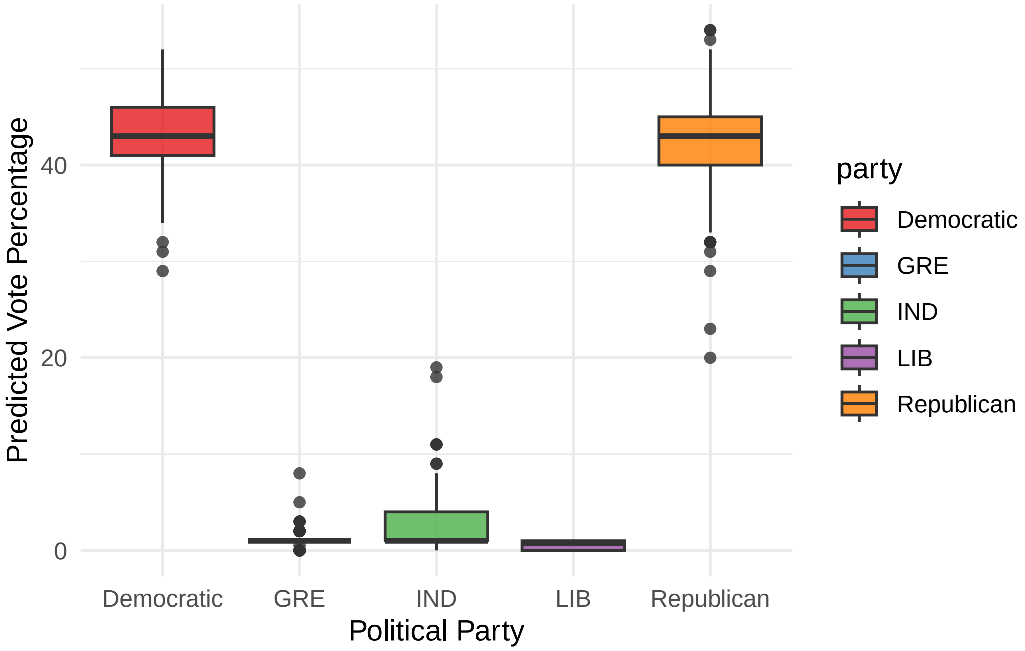


Figure 1: Predicted vote percentage by party

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

Another important aspect of our data is the sample size, which varies between polls. Larger sample sizes tend to produce more reliable predictions, as they better capture the diversity of voter preferences. Below, we show the relationship between sample size and predicted vote share.

In Figure 2, we observe that larger sample sizes tend to produce a wider range of vote percentages. This suggests that larger polls may capture more nuanced variations in voter preferences,

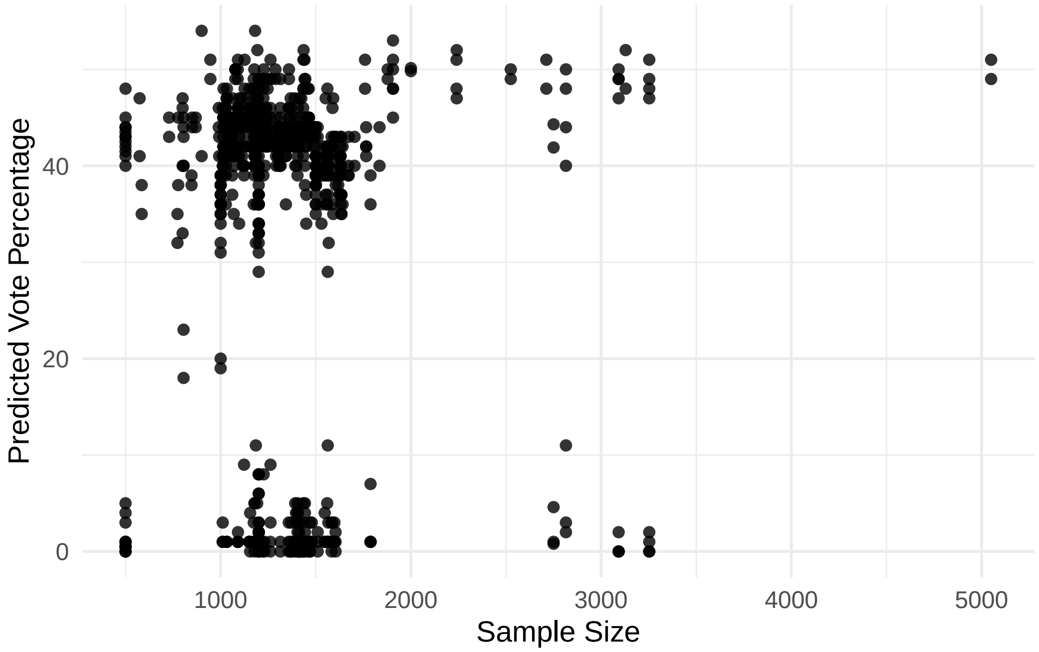


Figure 2: Relationship between sample size and predicted vote percentage

which could be critical in tight election races.

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.

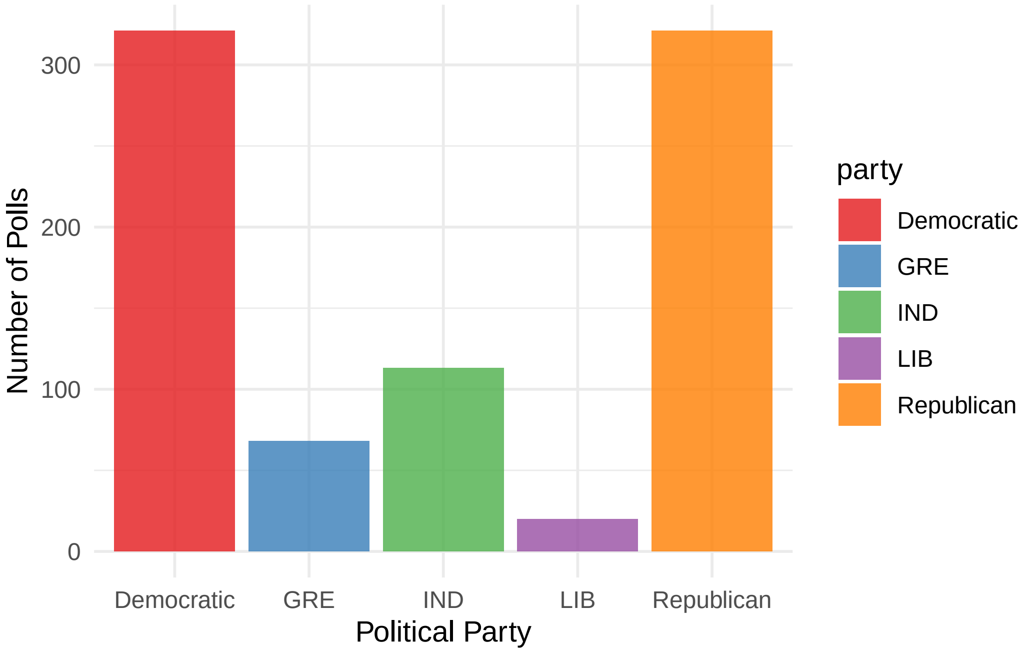


Figure 3: Distribution of political parties in the polling data

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 | _i, Normal(_i,)]

where

[_i = + _1 party_i + _2 candidate_i + _3 sample_size_i + _4 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:

[Normal(0, 2.5)] [_1, _2, _3, _4 Normal(0, 2.5)] [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. 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.

4 Results

Our results are summarized in Table 1. 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.

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

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

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

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

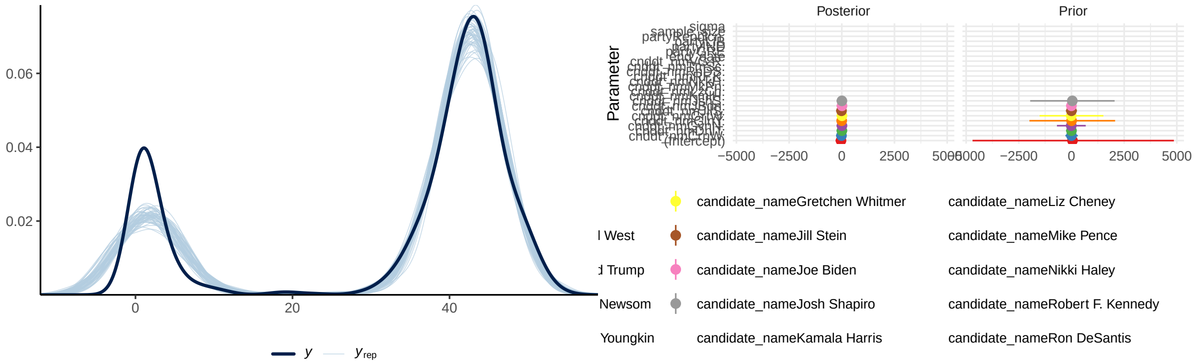
A Additional data details

B Model details

B.1 Posterior predictive check

In Figure 4a we implement a posterior predictive check. This shows...

In Figure 4b we compare the posterior with the prior. This shows...



(a) Posterior prediction check

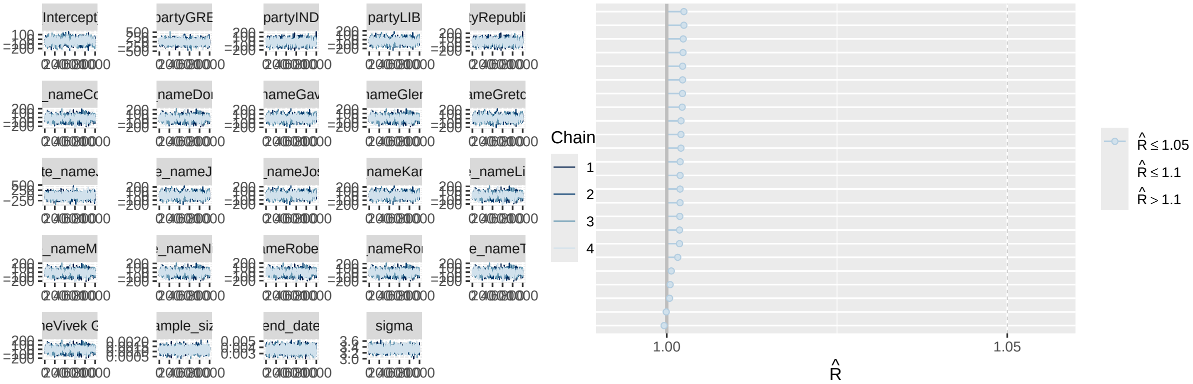
(b) Comparing the posterior with the prior

Figure 4: Examining how the model fits, and is affected by, the data

B.2 Diagnostics

Figure 5a is a trace plot. It shows... This suggests...

Figure 5b is a Rhat plot. It shows... This suggests...



(a) Trace plot

(b) Rhat plot

Figure 5: Checking the convergence of the MCMC algorithm

References

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

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/.