Forecasting the 2024 U.S. Presidential Election Results: Linear Regression, Bayesian and Spline-Fit Gaussian Model by Using Pollsters' Data From Five Thirty Eight*

Contrasting Harris's Lead: Strong in Linear Regression and Bayesian Model, Modest in Spline-Fit Gaussian Model

Ziqi Zhu Yuanchen Miao Claire Chang

November 4, 2024

We utilize FiveThirtyEight's 2024 U.S. presidential poll-of-polls data in 2024 to analyze support trends for the leading candidates, Kamala Harris and Donald Trump. By focusing on variables shows the reliability of a pollster, such as sample size, numeric scores, poll scores, and state-level factors, the data is fit into three models: a linear regression model, a Bayesian model, and a spline-fit Gaussian model. The results of the analysis indicated a strong lead for Harris in both the linear regression and Bayesian models, with a more modest lead in the spline-fit Gaussian model. This paper predicts the potential trajectory of the election, helping the public anticipate the policy directions that might follow a Harris administration.

Table of contents

*Code and data are available at: https://github.com/zzq20010617/2024_USelection_prediction

1 Introduction

Nowadays, U.S. presidential elections have been marked by intense public interest and polarization. The race between Donald Trump and Kamala Harris, both of them are having a significant leads over other candidates, is currently the most closely watched contest. Polling data is now playing a very important role in showing the public perceptions and support rate to each candidates. Analyzing polling data with statistical models, we are able to have a statistical approach into candidate support rate across different states and parties. Analysis

on polling data could help the public to predict the shifts in national, economic and social directions. Since the stances hold by Trump and Harris on issues like healthcare, immigration, and the international relations are in distinct positions, analyzing on polling data to predict the winner of the 2024 U.S. presidential election is very important to the public. This paper is contributing to this understanding using the latest polling data and statistical models to capture the trends in support rates of these two candidates.

In this paper, polling data of 2024 U.S. presidential election from FiveThirtyEight (FiveThirtyEight, n.d.), focusing specifically on the support rate of the leading candidates, Kamala Harris and Donald Trump, is analyzed. The polling data from FiveThirtyEight (FiveThirtyEight, n.d.) aggregates polls from various sources, thus, to ensure the reliability on the forecast result and analysis, we reduced the variables of the data to only retain variables that reflect polling reliability, such as sample size, numeric scores and poll scores. Additionally, the state-level data is kept as a factor influencing support rates, acknowledging the geographic variations in voter preferences.

The primary estimand in this study is the expected percentage of support (pct) for a U.S. presidential candidate, given polling data and relevant predictor variables like rating of the pollster and sample size of the poll.

The analysis employs three statistical models: linear regression model, Bayesian model and spline-fit Gaussian model. The findings indicates that Kamala Harris maintains a strong lead in both linear regression model and Bayesian model, while her leads becomes modest when comparing the trend of support rate of Donald Trump in spline-fit Gaussian model. Comparing to Donald Trump, Kamala Harris has a 2.6 percent lead in linear regression model, 23 more supporter out of 1200 in Bayesian model and only 1.06 percent lead in spline-fit Gaussian model.

These results are significant since they show the evolving dynamics of the election and how different statistical models and approaches could yield varying interpretations of candidate support rate. This paper is giving public an idea of who is the likely winner of the election that will have the potential implications on future political strategies and the overall direction of U.S. policy.

The remainder of this paper is structured as follows. Section 2 discusses the data used for this analysis, including key variables and sources, with particular attention to the quality metrics that affect polling accuracy. Section 3 outlines our modeling approach for each candidate, incorporating lessons learned from recent electoral cycles. Our predictions are underr section of each model. Section 4 discusses the implications of our findings and suggests directions for future research. Finally, Section A evaluates Trafalgar Group's polling methodology, and our idealized methodology and survey copy.

2 Data

2.1 Overview

Our data is download from [Five Thirty Eight] (https://projects.fivethirtyeight.com/polls) (FiveThirtyEight, n.d.), the website gathered survey data from different pollsters of 2024 US president election, We use the statistical programming **language R** (R Core Team 2023), and packages **lubridate** (Grolemund and Wickham 2011), **dplyr** (Wickham et al. 2023), **tidyverse** (Wickham et al. 2019), **model summary** (Arel-Bundock 2022), and **arrow** (Richardson et al. 2024) to process the data. Also, we use **tibble** (Müller and Wickham 2023) to create the table for simulate table and save it as csv files. Graphs are created by using ggplot2 in from package **tidyverse** (Wickham et al. 2019), packages **broom** (Robinson, Hayes, and Couch 2024) and **knitr** (Xie 2014) were used to generate clean summary and table in this paper. Package **patchwork** (Pedersen 2024) is been used to combine two graphs together. Following **Telling stories with data** (Data 2024), we create different models for the candidates from different parties in general state and forecasting the US president election in 2024.

2.2 Cleaned Data

The raw data obtained from FiveThirtyEight (FiveThirtyEight, n.d.) initially contains 17,440 observations, which we reduced to 12,185 by only retaining polls conducted after January 1, 2024. As our analysis focuses specifically on Kamala Harris and Donald Trump, data collected before January 1, 2024, holds limited relevance due to the inclusion of outdated candidates. Reducing the data further to only include polls after Harris officially announced her candidacy would result in too few observations to support a robust model. Additionally, the "state" column in the dataset was adjusted by replacing all "NA" values with "National," as these "NA" entries represent polls conducted nationwide in the raw data. A view of the first five rows of cleaned data is shown in Table 1

Table 1: First Five Rows of Cleaned Data

pollster	sample size	poll grade	pollscore	state	transparency score	poll end date	party	candidate	percentage
MassINC Polling Group	582	2.8	-0.8	Massachusetts	7	2024-11-01	DEM	Harris	61
MassINC Polling Group	582	2.8	-0.8	Massachusetts	7	2024-11-01	REP	Trump	31
MassINC Polling Group	582	2.8	-0.8	Massachusetts	7	2024-11-01	IND	Kennedy	2
MassINC Polling Group	582	2.8	-0.8	Massachusetts	7	2024-11-01	GRE	Stein	1
MassINC Polling Group	582	2.8	-0.8	Massachusetts	7	2024-11-01	IND	West	0

2.3 Measurement

Polling data is collected by different pollsters using various survey methods reflect public opinion on the 2024 U.S. presidential election. Information of individuals collected by pollsters is being separate to different groups based on age, gender, race, occupation and geographic

to reflect a more accurate supporting trend of each candidates involved in the presidential elections. From FiveThirtyEight, poll-of-polls dataset included the pollscore and numeric grade of the pollster which influence the reliability of the pollster on the collection of support rate of each candidates. Polling methodology is also various among pollsters, such as online panels, app panels, online Ad and phone interviews. Different methodology are having its own target audiences which the data collected can represent the opinion of specific age group.

2.4 Outcome variables

Our primary outcome variable is **support percentage (pct)**, which represents the percentage of respondents who support each candidate. This variable is modeled as the response variable in the linear regression model, Bayesian model and spline-fit Gaussian model analysis.

2.5 Predictor variables

2.5.1 Sample Size

Sample size is an important predictor because it influences the reliability of a poll. Larger sample sizes tend to reduce sampling error, providing more accurate reflections of voter sentiment and more reliable predictions of the election. In our models, the sample size is log-transformed to account for diminishing returns—larger polls do not necessarily offer proportionally better accuracy.

2.5.2 Poll Score and Numeric Grade

The poll score is a measure of the error and bias we can attribute to a pollster, negative number is better. The numeric grade is an aggregate score of the poll based on a numeric scale. It serves as an indicator of the pollster's historical performance and the methodology used. 3 is the maximum and some pollsters have no rating. A lower poll score and higher numeric grade of a pollster indicate the data collected by the pollster has higher reliability and the methodology used by the pollster is more consistent.

The left side of Figure 1 presents the distribution of poll scores among all of the pollsters, most of the pollsters have negative poll scores which indicates these pollsters are reliable. Among these pollsters having negative poll scores, more than 2000 pollsters have poll scores lower than -1 showing higher reliability of the data collected by them than other pollsters. Some of the pollsters have positive poll scores, data collected by these pollsters is not reliable.

The right side of Figure 1 presents the distribution of numeric grades of all pollsters, most of the pollsters have the numeric grades higher than the middle number, 1.5. More than 3000 polling organizations have numeric grades higher than 2.8 reflects a higher reliability of data collected by them.

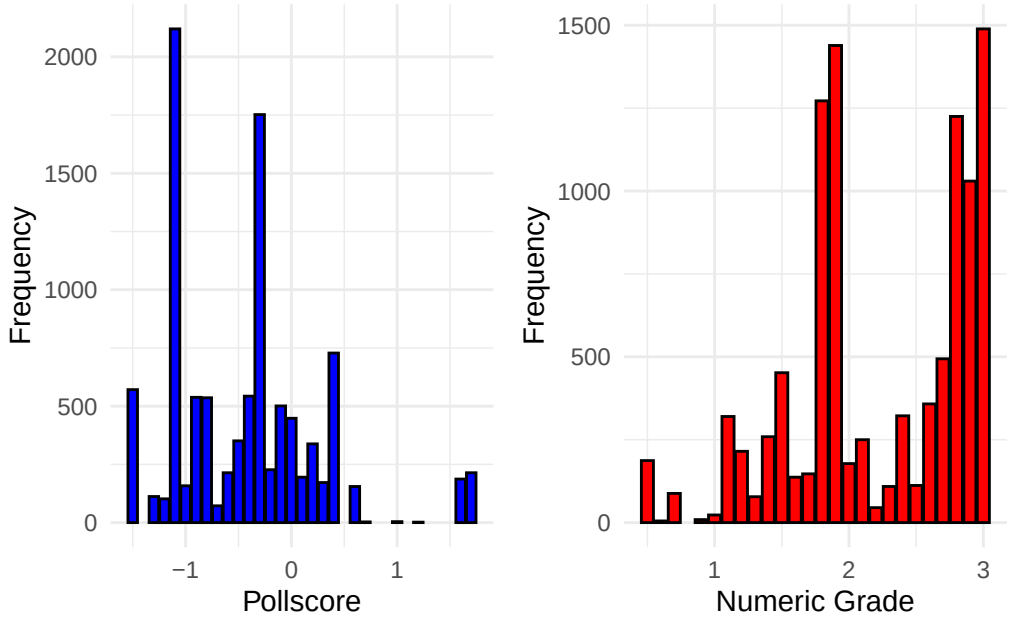


Figure 1: Distribution of Poll scores and Numeric Grade of Pollsters in the 2024 U.S. Presidential Election

2.5.3 Transparency Score

Transparency score measures how openly a pollster reports their methodology and results. A higher transparency score suggests that the pollster has disclosed key details about how the poll was conducted, improving trust in the poll results.

From the distribution graph of transparency score among all of the pollster shown in Figure 2, this reflects the differences in the methodologies used by different pollsters. Over 3,000 pollsters have a transparency score above 9, indicating that they have more open and trustworthy methodologies. In contrast, some pollsters have a transparency score below 5, which suggests that their methodologies are not disclose the details of their data collection processes, thereby diminishing the credibility of these polls."

2.5.4 Days Since Start

This variable represents the number of days since the beginning of the election polling period. It is calculated as Days Since Start = number of days from the beginning of the election to the end date of the specific poll. It helps capture the dynamic nature of voter preferences over time, accounting for shifts in public opinion as the election date nears. We only keep the polls with end dates after January 1 2024 to keep the variety of the data but avoid candidates and

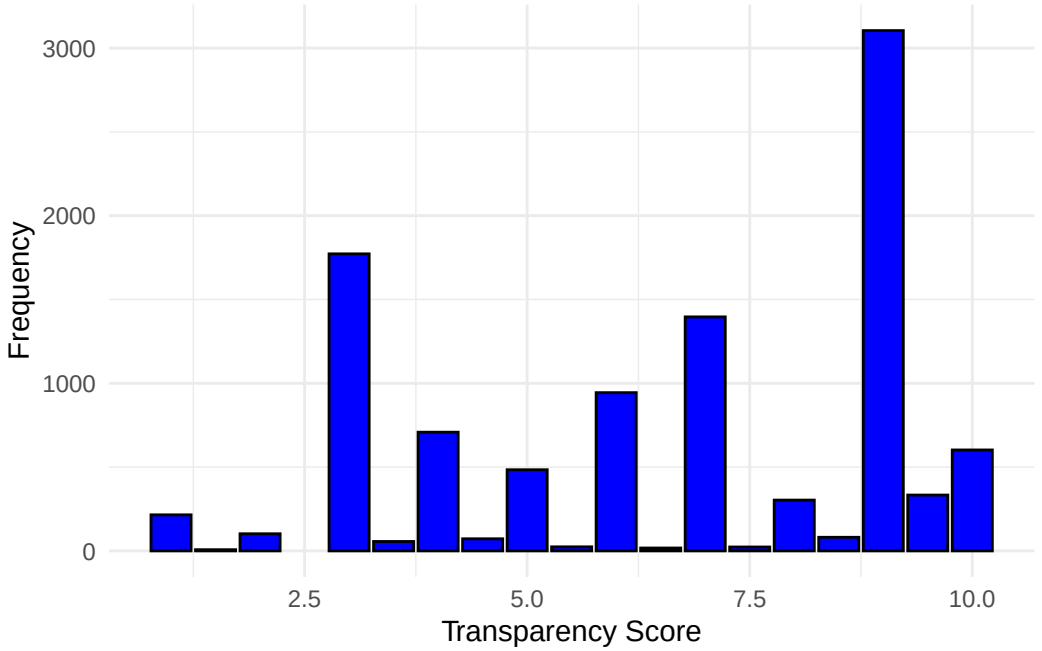


Figure 2: Distribution of Transparency Score of Pollsters in the 2024 U.S. Presidential Election

parties that are not involved in the final race of the election to mainly analyze the support rate percentage between Donald Trump and Kamala Harris.

3 Models

The goal of our modelling strategy is try to capture the trend of support percentage for different parties in 2024 Presidential election as much as possible. In the following section we briefly describe the models we used to investigate. Background details and diagnostics are included in Appendix B.

3.1 MLR Model set-up

We first filtered our data by different parties, as the single linear model but include other 3 parties(GRE, IND, and LIB, Conservative only has two entries in cleaned nationwide data and has 0 pct so its been removed), and fit linear model for each of them. The formula for Multiple linear regression is as follow

$$\text{pct}_i = \beta_0 + \beta_1 \log(\text{sample\_size}_i) + \beta_2 \,\text{pollscore}_i + \beta_3 \,\text{numeric\_grade}_i \tag{1}$$

$$+ \beta_4 \,\text{transparency\_score}_i + \beta_5 \,\text{days\_since\_start}_i + \epsilon_i \tag{2}$$

Table 2: Explanatory models of pct based on pollscore, sample size, and time

	DEM	REP	IND	GRE	LIB
(Intercept)	37.15	50.18	−4.45	3.01	4.36
	(1.46)	(1.42)	(3.47)	(0.77)	(1.41)
log(sample_size)	0.41	−0.25	2.26	−0.14	−0.21
	(0.16)	(0.16)	(0.43)	(0.10)	(0.17)
pollscore	−0.71	−3.33	−1.09	−0.04	−0.86
	(0.35)	(0.34)	(0.55)	(0.12)	(0.26)
numeric_grade	0.12	−3.55	−1.73	−0.07	−1.04
	(0.45)	(0.44)	(0.77)	(0.16)	(0.32)
transparency_score	−0.32	0.06	0.12	0.00	0.07
	(0.06)	(0.06)	(0.14)	(0.03)	(0.05)
days_since_start	0.03	0.01	−0.03	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Num.Obs.	1348	1348	636	307	126
R2	0.395	0.150	0.204	0.136	0.109
R2 Adj.	0.392	0.147	0.198	0.122	0.072
AIC	7320.7	7249.2	3792.0	619.1	313.9
BIC	7357.1	7285.7	3823.2	645.2	333.8
Log.Lik.	−3653.347	−3617.616	−1888.994	−302.568	−149.968
RMSE	3.64	3.54	4.72	0.65	0.80

Response Variable is pct. Predictors are log(sample_size), pollscore, numeric_grade, transparency_score, and days_since_start, detail can be find in Section 2.4, and Section 2.5

3.1.1 Result

Our results for the MLR model are summarized in Table 2.

First, we checked the residual vs. fitted plot as shown in Appendix B and saw no obvious violation of assumptions. According to the summary, DEM and REP models are based on the largest number of observations (both 1235). Smaller parties have fewer observations, which may lead to less robust models. The r^2 shows that the model explains the highest amount of variance on the Democratic Party than others with a value of 0.357, and other parties less than 0.2. Based on the summary of models for DEM and REP Section B.2, we are not sure

if those predictors that measure the quality of polls are significant, so we keep them for now in the Bayesian model. The prediction of MLR models is shown in Figure 3, in which we can see the blue line represents the Democratic party has a steeper slope and starts to take the lead between day 100 and day 200, and the average support percentage on the last day of data shows that Democratic (Harris) has a slight advantage of 2.6 percent.

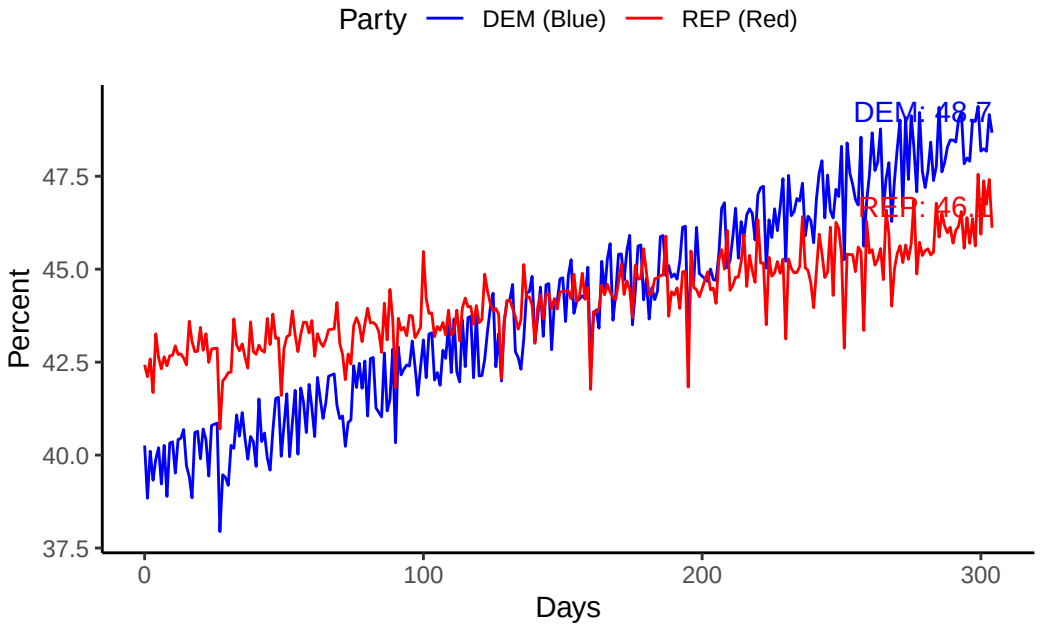


Figure 3: Prediction with MLR models for DEM and REP party

3.2 Bayesian Model set-up

The Bayesian model is built from the MLR models from the above section. With inspiration from the example R code, we introduce the random effects for Pollsters, and we use a logistic link to predict the probability of support, modeling count data which is in a binomial distribution. We choose to use a default prior and enable autoscale, the formula is shown below.

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \log(\text{sample\_size}_i) + \beta_2 \,\text{pollscore}_i + \beta_3 \,\text{numeric\_grade}_i \tag{3}$$

$$+ \beta_4 \,\text{transparency\_score}_i + \beta_5 \,\text{days\_since\_start}_i + \alpha_j \tag{4}$$

$$\alpha_j \sim \text{Normal}(0, \sigma_{\text{pollster}}) \tag{5}$$

$$\beta_0 \sim \text{Normal}(0, 2.5) \tag{6}$$

$$\beta_1 \sim \text{Normal}(0, 2.5) \tag{7}$$

$$\beta_2 \sim \text{Normal}(0, 2.5) \tag{8}$$

$$\beta_3 \sim \text{Normal}(0, 2.5) \tag{9}$$

$$\beta_4 \sim \text{Normal}(0, 2.5) \tag{10}$$

$$\beta_5 \sim \text{Normal}(0, 2.5) \tag{11}$$

$$\sigma_{\text{pollster}} \sim \text{Exponential}(1) \tag{12}$$

y_i represents the number of individuals in the sample that support the Democratic party (this corresponds to num_party). The response is modeled as binomial: $y_i \sim \text{Binomial}(\text{sample\_size}_i, p_i)$, where p_i is the probability that an individual in poll i supports the Democratic party, and sample_size_i is the total number of individuals surveyed in poll i. We also create a model for the Republican party to compare the trend in prediction results in the following section Section 3.2.1.

3.2.1 Result

To get a prediction by the Bayesian model, we create a data frame and generate posterior predictions by the new data frame, this table Table 4 shows the first several rows, and we set the quality scale of predicted data to the best to perform a poll that has high quality. A summary of the Bayesian model is shown in Table 3, and the posterior predictive check is in appendix Section B.3. The predicted result is shown in Figure 4.

The prediction results of the Bayesian model closely align with those of the linear model, with both showing the mean predicted number of supporters for each party. Notably, the Democratic Party takes the lead around day 210 (late July). By early November, the predictions indicate that the Democratic Party has an average of 601 supporters, compared to the Republican Party's 578. This suggests a slight advantage for the Democrats as the final prediction period concludes.

3.3 Spline Fit model set-up

We first run the model in R (R Core Team 2023) using the `rstanarm` package of (Goodrich et al. 2022), and use priors Normal$(0, 5)$ to allow for more flexibility of predictors effects

Table 3

	(1)
(Intercept)	−0.67
log(sample_size)	0.04
pollscore	−0.10
numeric_grade	−0.07
transparency_score	0.00
days_since_start	0.00
Sigma[pollster × (Intercept),(Intercept)]	0.01
Num.Obs.	1348
ICC	0.9
Log.Lik.	−11 224.203
ELPD	−11 500.4
ELPD s.e.	326.4
LOOIC	23 000.9
LOOIC s.e.	652.8
WAIC	23 018.0
RMSE	0.03

Table 4

```
  num_party pollscore sample_size numeric_grade transparency_score
1         0      -1.5        1200             3                  10
2         0      -1.5        1200             3                  10
3         0      -1.5        1200             3                  10
  days_since_start pollster
1         0.000000   YouGov
2         3.070707   YouGov
3         6.141414   YouGov
```

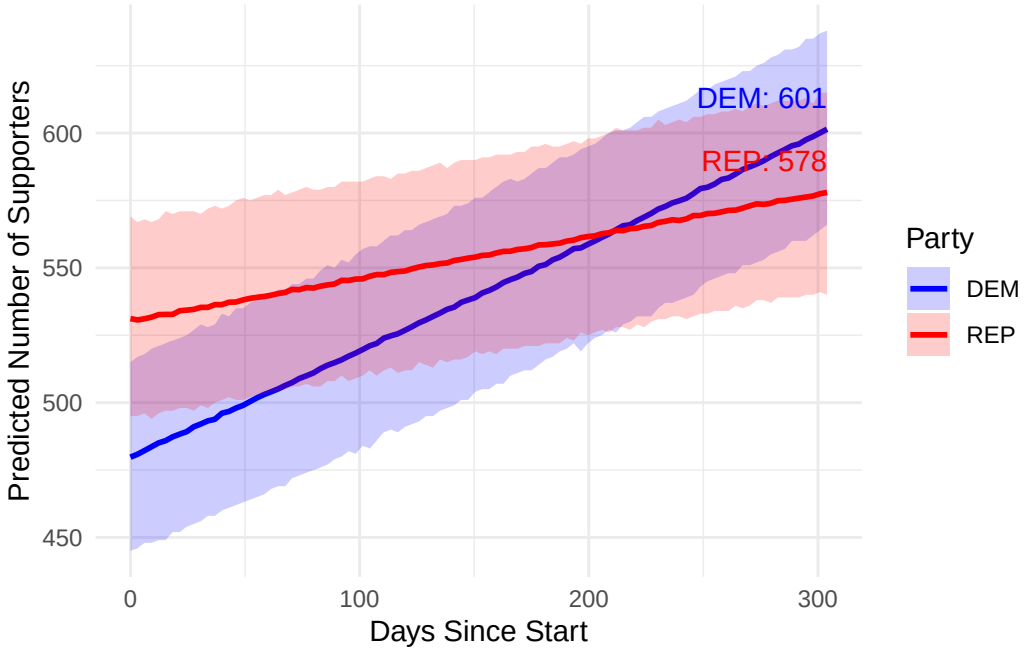


Figure 4: Predicted number of supporters for DEM and REP party

with "autoscale = TRUE". Then we use the same data frame that was used to predict for the Bayesian model Table 4, the result is shown as Section 3.3.1

3.3.1 Result

The prediction graph Figure 5 contains a prediction of the spline fit model for two parties, the blue line represents the DEM party's predicted percentage of support, and the red line represents the REP party's predicted percentage of support over time (measured in days since the start of 2024). We can tell that the predicted support of the DEM party is taking the lead by around 2 percent until the last day of the dataset.

4 Discussion

4.1 Crossing point in the prediction

From the prediction of spline fit model Figure 5, There is a clear increase in the support for the DEM party starting around day 200, while the REP party's support remains more stable with only a slight increase over time. The increase in support for the DEM party in the predictions could be attributed to Harris's rise as the Democratic candidate after Biden's exit

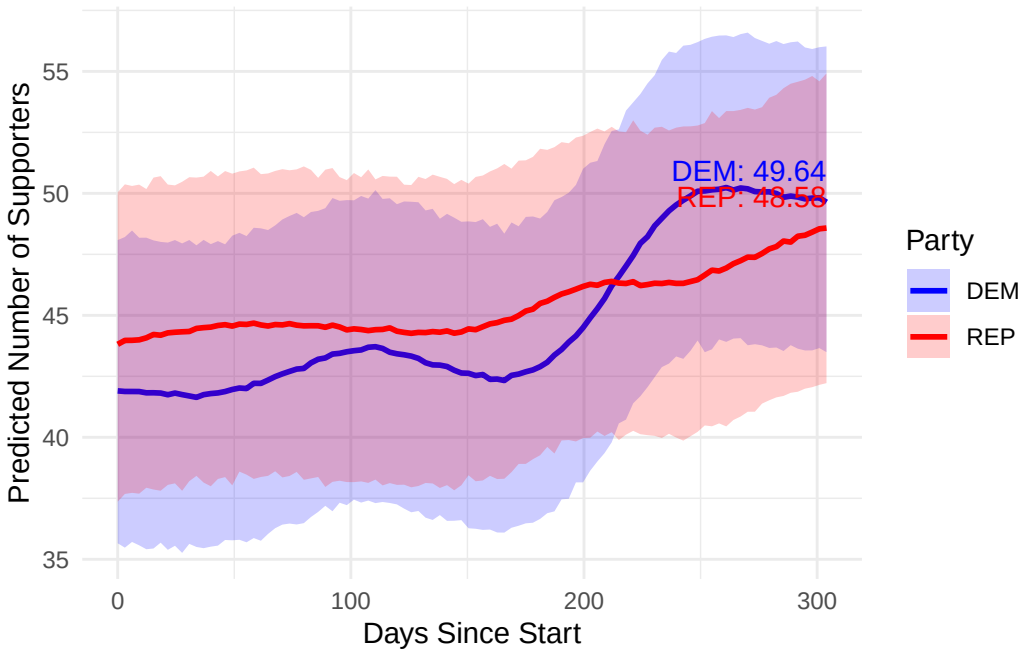


Figure 5: Predicted support percentage over Time with Spline Fit for DEM and REP party

from the race On July 21, 2024, which is about 230 days from start of 2024. By analyzing the support percentages of Biden and Harris from the Democratic poll data Figure 6, it becomes evident that their combined impact likely contributes to the double peak seen in the posterior predictive check Figure 9. This dual pattern challenges the model's ability to capture trends accurately

4.2 Cross-Model Observations

In both the MLR and Bayesian models, Harris demonstrates a significant lead over Trump. The MLR model projects an approximate 2.6% lead for Harris, as indicated by the slope for the Democratic (DEM) variable, which is steeper and surpasses the Republican (REP) trend line between day 100 and day 200 (see Figure 3). This early crossover point indicates a shift in support dynamics, suggesting that Harris's policies or campaign events during this period may have positively influenced public sentiment. The Bayesian model, which accounts for random effects among pollsters, reinforces this lead by predicting 601 supporters for Harris versus 578 for Trump towards the end of the prediction period (see Figure 4). This stability across models suggests that Harris's support is likely durable, despite the model-specific variations in lead magnitude. The Spline-Fit Gaussian model, however, presents a more conservative lead for Harris, with a smaller difference of approximately 1.06% in favor of the Democratic candidate by the end of the dataset (see Figure 5). This model, designed to capture smoother trends,

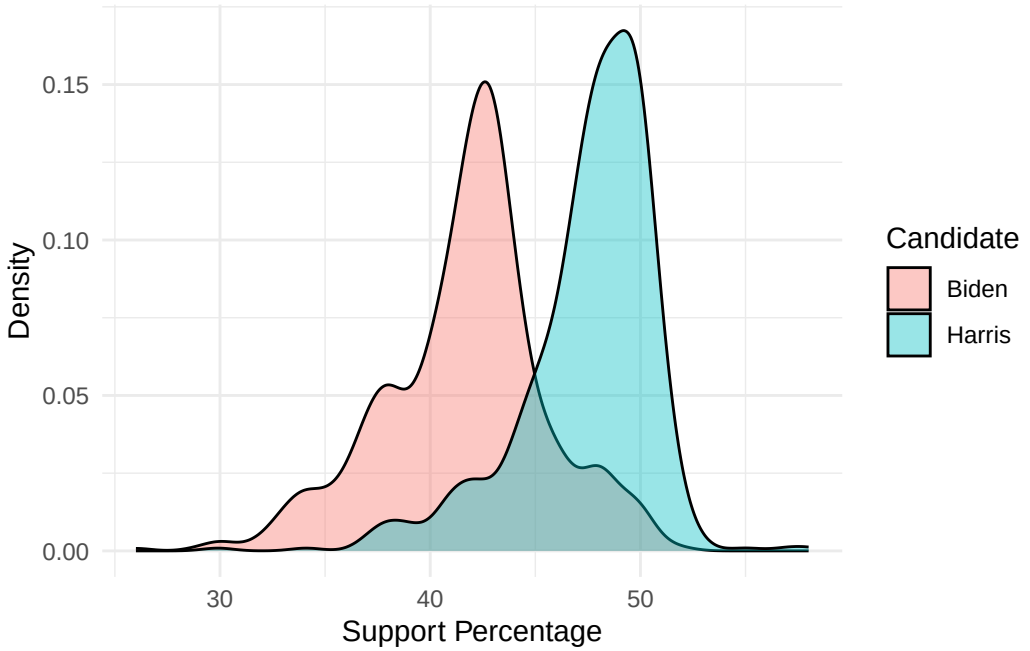


Figure 6: Density Plot of Support Percentage for Biden and Harris in 2024

suggests that Harris's lead, while present, may not be as robust as projected by the MLR and Bayesian models. This discrepancy emphasizes that different modeling approaches yield varied levels of confidence in support predictions, with the Spline-Fit model capturing subtle shifts that might indicate voter hesitancy or fluctuating preferences over time.

4.3 Differences in Modeling Nationwide Polls and Actual Election Results

In this paper, we used nationwide poll data for modeling, which contrasts significantly with using state-by-state election results. Nationwide polls offer general insights into aggregate support levels but do not account for the Electoral College system in U.S. presidential elections. This approach can overlook state-specific variations and the unique influence of each state's electoral votes, leading to potential bias where larger states disproportionately affect overall results. Conversely, state-level modeling captures the distribution of support across states, highlighting the impact of battleground states. To enhance prediction accuracy and align with the actual election structure, future models should integrate state-level data especially for the battleground states for more robust, applicable results.

4.4 Weaknesses and next steps

Firstly we focused only on nationwide polls, excluding state-specific data, due to limited and outdated information from certain states. For example, Mississippi had only four polls, with the most recent conducted in April 2024, making state-level modeling and vote counting challenging. This limitation restricts the accuracy of state-by-state analysis. With more reliable and updated data, using a state-based approach would enhance the model's precision and provide a more comprehensive prediction of the election outcome, especially for discovering the public opinion in battleground states. Furthermore modeling Biden and Harris together under one Democratic model introduces potential inaccuracies, especially influencing the slope of predictions in both linear and Bayesian models, making the observed upward trend less reliable.

A Appendix

A.1 Idealized Methodology

To forecast the 2024 U.S. presidential election with a budget of $100,000, this methodology combines stratified sampling, multimodal recruitment, and aggregation techniques that leverage the strengths of multiple data sources. The approach employs stratified random sampling with targeted oversampling of key subgroups, like younger and minority voters, to address their frequent underrepresentation in polling. By segmenting eligible U.S. voters by demographics—age, gender, race, party affiliation, region, and urban/rural residency—the method ensures broad representational coverage, aiming to capture diverse voting patterns accurately (PewResearchCenter 2019). Partnering with a voter database provider gives us access to a detailed list of registered voters, organized by demographics, to improve the accuracy of our sample. This approach also builds on public trust, as Americans generally trust polls from news organizations (43%) more than those from websites that combine multiple polls (30%) (Pasek 2015).

Respondents are recruited through multiple channels to reduce potential biases associated with any single mode. The multimodal approach includes Interactive Voice Response (IVR) to reach older and rural populations (30%), targeted online surveys to capture difficult-to-reach demographics (40%), text surveys directed at younger voters (20%), and live calls in key battleground states (10%). This mix helps address "coverage error" by compensating for the limitations associated with each mode (Blumenthal 2014).

Survey data is collected via a short, 5–7 minute questionnaire with 10 core questions covering demographics, voting intention, likelihood of voting, and key issues. A shorter survey length and mix of direct and indirect questions help reduce Social Desirability Bias, while randomizing question order minimizes positional bias. Given that any individual survey is likely to suffer from random and systematic errors—sampling errors, coverage errors, and response biases—aggregation across multiple polls can mitigate these sources of error, improving the accuracy of forecasts by averaging out errors specific to individual polls (Blumenthal 2014).

For data validation and quality control, screening questions confirm attentiveness, demographic cross-checks validate responses against voter registration data, and post-survey weighting adjusts for demographic imbalances. Weighting responses to match population demographics ensures that key groups are proportionally represented, while data cleaning removes low-quality responses to maintain data integrity (Blumenthal 2014). To further address biases, surveys are conducted anonymously in formats like online and SMS. Weighting adjustments are also made to address non-response bias, enhancing accuracy by reflecting the true population distribution. We will keep the process and result transparent by posting on our website for the public to view.

Weekly aggregation serves as a "poll of polls" that balances fluctuations across samples and smooths out errors. By incorporating recent data more heavily, this rolling average adapts

to shifts in voter sentiment leading up to the election. Bayesian modeling, applied to smooth sample variations and incorporate prior election trends, stabilizes the forecast further. Poll aggregation, which is commonly employed in election analyses, effectively combines estimates from various samples, reducing random error and discounting non-universal biases (Blumenthal 2014). Given the variability in survey methodologies—such as IVR surveys missing cell-only populations or web surveys excluding offline individuals—aggregation allows errors in one type of survey to offset those in another. Weighting each survey according to sample size and precision further refines the accuracy of the aggregated forecast.

A.1.1 Budget Allocation:

- **Sample Acquisition**: Partnership with a voter database provider to access a list of registered voters across all states $10,000
- **Interactive Voice Response (IVR)**: Automated voice surveys for older and rural demographics $20,000
- **Online Surveys**: Targeted ads and opt-in digital surveys to capture harder-to-reach groups (40% of sample) $25,000
- **Text Messaging**: SMS-based surveys to engage younger demographics (20% of sample) $10,000
- **Live Phone Calls**: Calls in battleground states to reach underrepresented groups (10% of sample) $20,000
- **Data Validation & Cleaning**: Screening questions, demographic verification, and removal of low-quality responses to maintain data integrity $5,000
- **Data Analysis & Forecast Modeling** Bayesian modeling, poll aggregation, and analysis of survey trends to produce election forecasts $10,000

A.2 Idealized Survey

The proposed survey questionnarie design is in the following link: https://forms.gle/Zm5Kfj3kL58gwCz38

A.2.1 Survey Copy

1. **What is your age??**
    - 18-29
    - 30-44
    - 45-64
    - 65+

2. **What is your gender identity?**

    - Female
    - Male
    - Another Gender Identity

3. **Which of the following best describes your race/ethnicity?**

    - White
    - Black/African American
    - Hispanic/Latino
    - Asian
    - Native American
    - Prefer not to say
    - Other

4. **What is your highest level of education?**

    - High school or less
    - College
    - Bachelor's degree
    - Graduate degree
    - Prefer not to say

5. **In which U.S. region do you currently reside?**

    - Midwest
    - Northeast
    - South
    - West

6. **Are you registered to vote in the 2024 presidential election?**

    - Yes
    - No

7. **2024 Which candidate do you plan to vote for?**

    - Kamala Harris
    - Donald Trump
    - Undecided
    - Other

8. **2020 Which candidate did you vote for?**

    - Joe Biden
    - Donald Trump
    - Did not vote
    - Prefer not to say

- Other

9. **What is the most important issue influencing your vote?**

    - Economy
    - Healthcare
    - Immigration
    - Climate Change
    - Social Justice
    - National Security
    - Other

10. **On a scale of 1-5, how likely are you to vote in the 2024 U.S. presidential election?**

- 1 being Not Likely to Vote
- 5 being Likely to Vote

A.3 Trafalgar group's methodology overview and evaluation

The Trafalgar Group conducts polls ranging from major political campaigns to marketing surveys. The organization's 0.7 pollster rating indicates moderate accuracy and reliability compared to other pollsters. The population consists of all eligible voters in the U.S., while the sampling frame includes registered voters across states, segmented by demographics like age, gender, party affiliation, and ethnicity. Trafalgar Group typically samples likely voters,focusing on those most likely to participate in upcoming elections. Trafalgar Group recruits its sample using a mix of interactive voice response, live phone calls, text messages, emails, digital dial back interface and online targeted opt-in digital survey platforms (The Trafalgar Group 2024). Trafalgar places particular emphasis on reducing Social Desirability Bias, designing short questionnaires and using nontraditional question formats to help participants feel more comfortable expressing their true preferences, especially on sensitive topics (The Trafalgar Group 2024). They also adjust for non-response by weighting the sample, ensuring it reflects the broader population's demographic composition. While this helps to reduce bias from underrepresented groups, reliance on post-survey adjustments can introduce new biases, especially if response rates vary significantly among demographic groups.

B Model details

B.1 Assumption check for MLR models

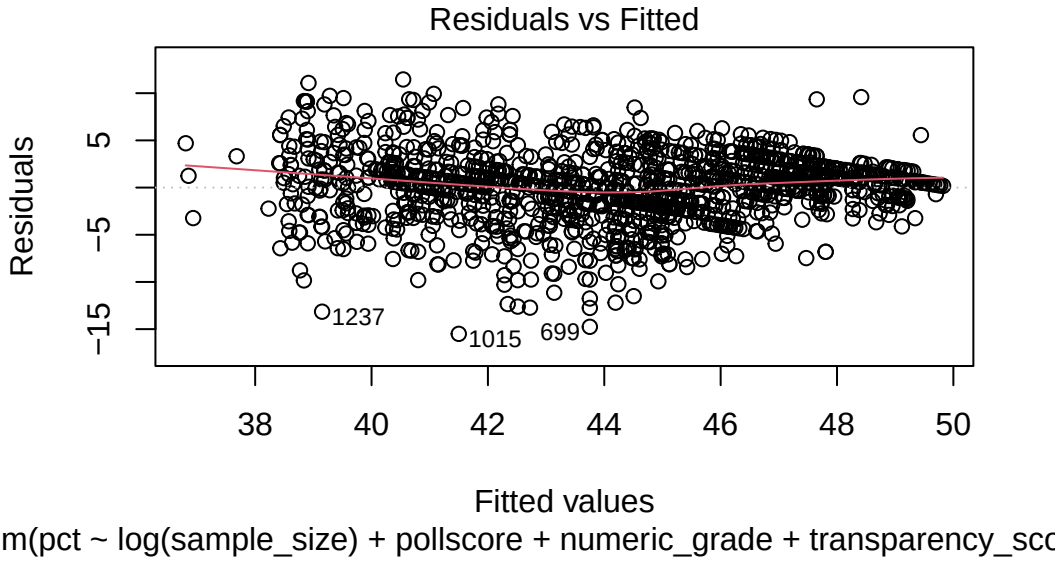


Figure 7: Residuals vs Fitted for DEM Party

Table 6: Summary of REP data fit by MLR model

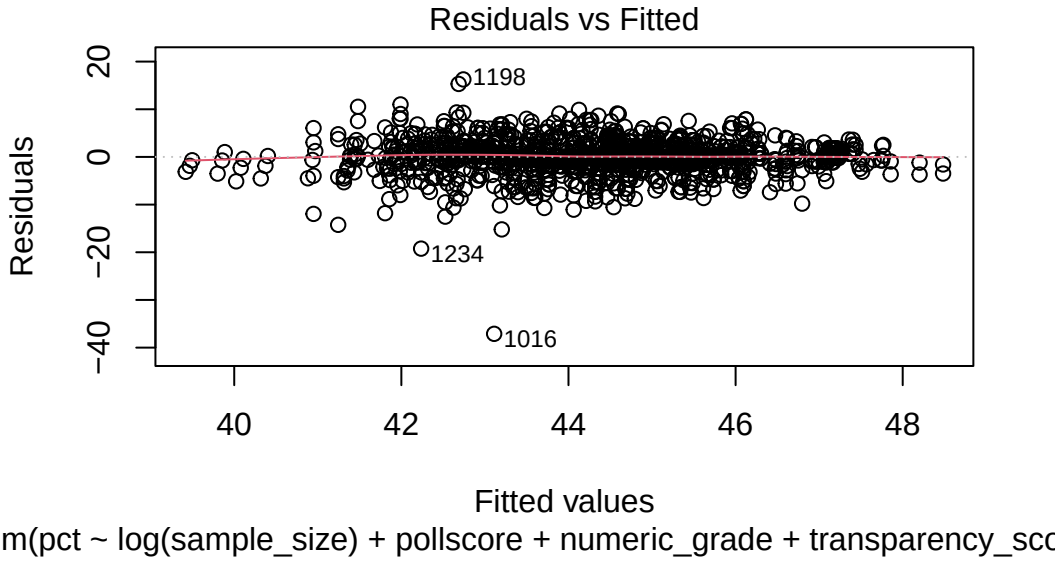


Figure 8: Residuals vs Fitted for REP Party

B.2 Significant check for MLR models

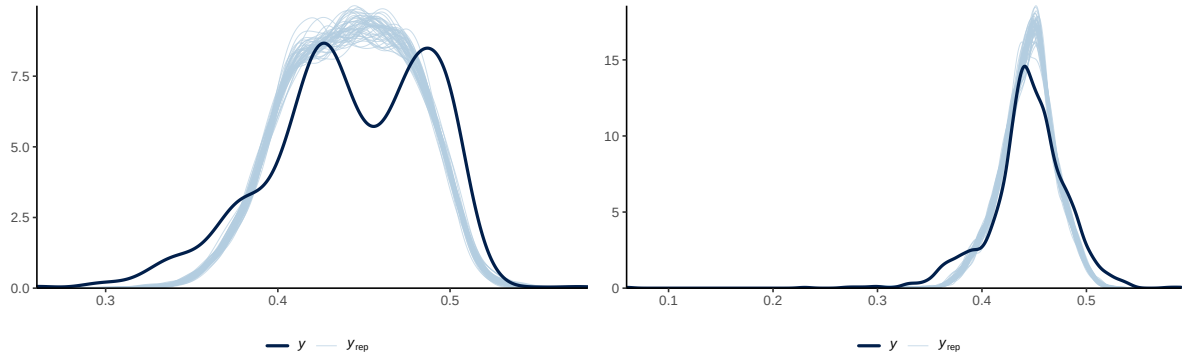
see Table 5, and Table 6

Table 5: Summary of DEM data fit by MLR model

term	estimate	std.error	statistic	p.value
(Intercept)	37.1463637	1.4606726	25.4310000	0.0000000
log(sample_size)	0.4095513	0.1599441	2.5605897	0.0105581
pollscore	-0.7077593	0.3491791	-2.0269236	0.0428676
numeric_grade	0.1221605	0.4485269	0.2723593	0.7853877
transparency_score	-0.3158432	0.0599232	-5.2707954	0.0000002
days_since_start	0.0315317	0.0011488	27.4478685	0.0000000

B.3 Posterior predictive check

In Figure 9 we implement a posterior predictive check. The check for DEM model shows the observed data has two peaks, indicating a possible bimodal distribution, but the predictive simulations also show significant variability around the peaks. While the model for REP performs better, capturing the main distribution of the data



(a) Democratic Model

(b) Republican Model

Figure 9: Posterior prediction check for Bayesian models

References

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.

Blumenthal, Mark. 2014. "Polls, Forecasts, and Aggregators" 47 (2): 297–300. https://doi.org/10.1017/S1049096514000055.

Data, Telling Stories with. 2024. "Telling Stories with Data." https://tellingstorieswithdata.com/.

FiveThirtyEight. n.d. "FiveThirtyEight Presidential General Election Polls - National (2024)." 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/.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

Pasek, Josh. 2015. "Predicting Elections: Considering Tools to Pool the Polls." *Public Opinion Quarterly*. https://academic.oup.com/poq/article/79/2/594/2277466?login=true#84746216.

Pedersen, Thomas Lin. 2024. *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

PewResearchCenter. 2019. "A Field Guide to Polling: Election 2020 Edition." *Pew Research Center*. https://www.pewresearch.org/methods/2019/11/19/a-field-guide-to-polling-election-2020-edition/#how-can-you-tell-a-good-poll-from-a-bad-one.

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

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Robinson, David, Alex Hayes, and Simon Couch. 2024. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

The Trafalgar Group. 2024. "Polling Methodology." https://www.thetrafalgargroup.org/polling-methodology/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.