Forcasting US Elections using Polls-of-Polls*

Using a Linear Regression Model to Adjust Pollster Biases and Predict Kamala Harris' Vote Percentage at 47.78%

Shamayla Durrin Denise Chang Krishna Kumar

October 22, 2024

This paper uses a polls of polls approach to predict voter support for Kamala Harris in the 2024 U.S. election, aggregating multiple polls to reduce individual bias and improve accuracy. Through a regression model incorporating factors such as pollster, sample size, state, and recency, we predict her overall support to be 47.78%, just below the 50% threshold. Our analysis highlights the importance of poll quality and recency in understanding trends in public opinion. These findings are significant as they provide a more reliable estimate of electoral outcomes, helping to inform political strategies and public discourse.

1 Introduction

While individual polls provide snapshots of public opinion, they are often subject to biases and methodological differences. To address these limitations, this paper uses a "polls of polls" approach to predict support for Kamala Harris, aggregating multiple polls to smooth out inconsistencies. By combining data from various polling organizations, our aim is to create a more accurate forecast of her voter support leading up to the 2024 U.S. election.

We employ a linear regression model to estimate Kamala Harris' support percentage, incorporating key variables such as pollster, sample size, state, and recency. To account for the varying reliability of different polls, we weight predictions using each pollster's numeric grade. Our model also explores the role of recency, with recent polls showing a decline in support for Harris. The final weighted prediction for her support stands at 47.78%, indicating that her overall backing falls just below the 50% threshold, suggesting potential challenges in gaining majority support.

This paper contributes to the field of election forecasting by emphasizing the importance of poll quality and recency in prediction models. The structure of the paper proceeds as follows: we

*Code and data are available at: https://github.com/krishnak30/US_elections.

first introduce the dataset and methodology, followed by model selection and results. We then discuss the implications of our findings, and conclude with limitations and potential directions for future research.

2 Data

2.0.1 Measurement

In the dataset of our analysis, the process of measurement begins by capturing real-world public opinion through surveys. Polling organizations transform this observed phenomenon into structured data by recording support percentages, pollster details, and methodological choices, thus turning public sentiment into quantifiable entries in the dataset.

Public opinion polls are essential for understanding voter preferences and electoral dynamics, providing data that informs policy decisions and campaign strategies. Our data set captures public support for candidates at specific times, along with information on methodologies, sample sizes, and pollster ratings.

Polling organizations use various methods, such as App panels, IVR (Interactive Voice Response), Live phone calls, Text-to-web systems etc. These methods can influence the validity and reliability of results, with larger sample sizes generally yielding more reliable estimates. Some methods, like live phone interviews, may be seen as more trustworthy than others.

The dataset also includes important measures of pollster performance, such as numeric grades, poll scores, and transparency scores. These metrics are calculated by considering bias, race difficulty, predictive error, and transparency, allowing us to assess the accuracy of each poll. The "pollscore," for instance, reflects a pollster's past performance and reliability, adjusted for potential biases.

2.0.2 Data Cleaning

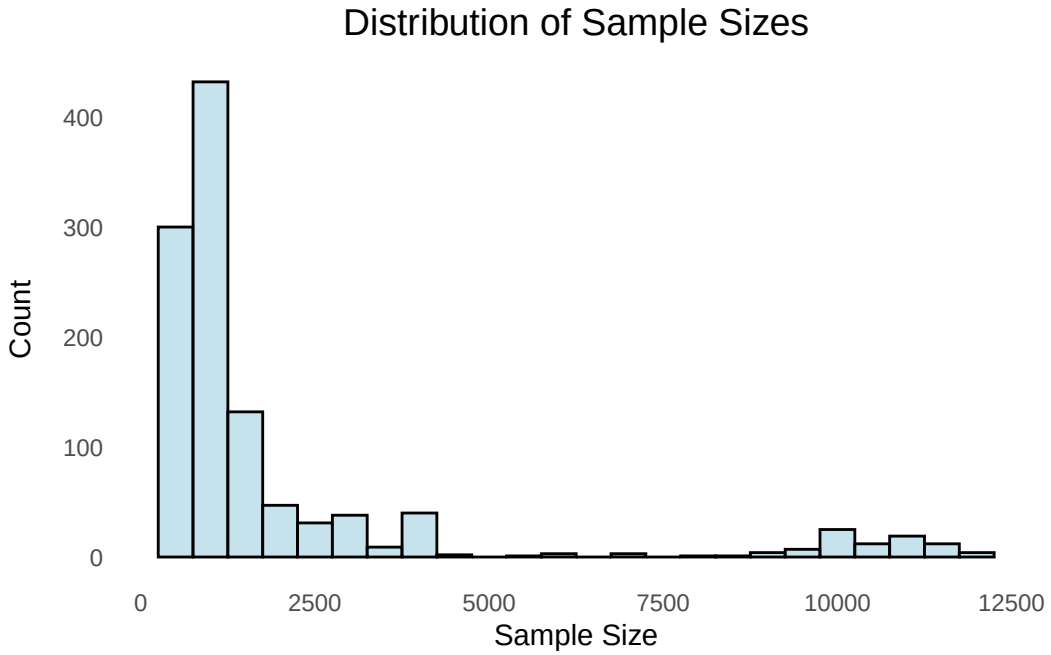
The raw data for this project, sourced from FiveThirtyEight, (FiveThirtyEight 2024) underwent a series of cleaning steps to prepare it for analysis. Initially, duplicate rows were removed to ensure that only unique observations remained, facilitated by the janitor package [@janitor]. A new binary variable, 'national', was created to indicate whether a poll was conducted at the national or state level. Missing values in the 'state' column were replaced with "Not Applicable," and numeric grades were evaluated to filter out low-quality pollsters, keeping only those with a numeric grade above 1. This cutoff was selected to retain mid to high-level pollsters for more reliable results. These steps were performed using functions from the dpylr package [@dplyr]. Furthermore, dates were standardized and converted into a proper format for analysis using the lubridate package [@lubridate]. Polls related to Kamala Harris were retained for further analysis, and percentage support values were transformed into actual numbers of

supporters based on sample size. Additionally, pollster counts below five were excluded to focus on more reliable data sources. Polls regarding Kamala Harris were filtered to include only those conducted after her official candidacy announcement on July 21, 2024, ensuring the data reflects post-announcement public sentiment.The cleaned dataset was saved in Parquet format for efficient storage and retrieval, using the arrow package [@arrow].

2.0.3 Summary Statistics of Variables of Interest

The summary statistics table presents key descriptive measures for the numeric variables of interest. The Numeric Grade variable has a mean of 2.3, with values ranging from a minimum of 1.1 to a maximum of 3.0. It has a standard deviation of 0.6, indicating relatively low variability. The Pollscore variable averages -0.5, with a standard deviation of 0.6 and a range between -1.5 and 1.7. The Sample Size has a wider range, with an average of 1962.5 and a standard deviation of 2608.0, indicating significant variation across polls, with values spanning from 382 to 12084. Lastly, Num Support, calculated as the number of respondents supporting Kamala Harris, shows a mean of 949.2, with values ranging from 172 to 6042 and a standard deviation of 1290.0, highlighting substantial differences across the polls.

2.0.4 Distribution of Sample Size of Polls



The histogram of sample sizes reveals a clear right-skewed distribution. Most of the sample sizes are clustered between 0 and 3000, with a sharp peak around 1000-1500. This indicates

that the majority of polls are conducted with smaller sample sizes. As sample size increases, the frequency significantly drops, with very few polls conducted with sample sizes larger than 5000, though there are a few outliers with sizes approaching 10,000 or more. This wide range in sample sizes might affect the precision of estimates in different polls.

2.0.5 Most Frequent Pollsters

Table 1: Top 10 Pollsters by Frequency

Pollster	Count	Average Pollscore	Average Numeric Grade
Morning Consult	235	-0.3	1.9
Siena/NYT	94	-1.5	3.0
Redfield & Wilton Strategies	88	0.4	1.8
Emerson	58	-1.1	2.9
YouGov	54	-1.1	3.0
Ipsos	45	-0.9	2.8
TIPP	44	-0.4	1.8
Beacon/Shaw	38	-1.1	2.8
AtlasIntel	34	-0.8	2.7
Quinnipiac	32	-0.5	2.8

This table lists the top 10 pollsters based on the frequency of polls they conducted. Morning Consult leads with 235 polls, followed by Siena/NYT and Redfield & Wilton Strategies with 94 and 88 polls, respectively. The table also shows each pollster's average poll score and average numeric grade. Most pollsters have negative poll scores, indicating a lower perceived reliability, except for Redfield & Wilton Strategies. However, the numeric grades vary, with Siena/NYT, YouGov, and Emerson receiving relatively high grades around 3, while others like TIPP and Redfield & Wilton Strategies are rated slightly lower.

3 Forcasting Election Outcome through Pooling Polls

3.1 Forecasting Approach

The polls of polls methodology is widely used in election prediction as it aggregates multiple polls to provide a more reliable estimate of voter support, rather than relying on any single poll. The goal is to reduce errors and biases present in individual polls by using a weighted average of many different polls.

In our approach, we will employ linear modeling of voter support percentage (pct) on pollster and other independent variables such as sample size, poll recency, and poll scope (state

vs. national). This will allow us to smooth out the inherent noise, biases, and variability across different pollsters. Once we obtain the predicted values from our model, we will weight these predictions based on the numeric grade (quality score) of each pollster to calculate an overall national estimate of the outcome. Additionally, we will separately compute estimates for key battleground states, where voter behavior can be more volatile and pivotal in deciding the final outcome of the election. This approach helps us capture both national trends and critical state-level dynamics.

3.2 Exploration of Independent Variables

3.2.1 Support for Kamala Harris by Sample Size and Polling Level

[@fig-sample] suggests that the relationship between support for Kamala Harris and sample size shows a slight positive trend, indicating a small increase in support as sample sizes grow. Both national (blue) and state (green) polls reflect this upward trend, although the effect is minimal. Since voting outcomes often depend on small margins, this subtle increase in support with larger sample sizes could still be relevant in predicting overall support.

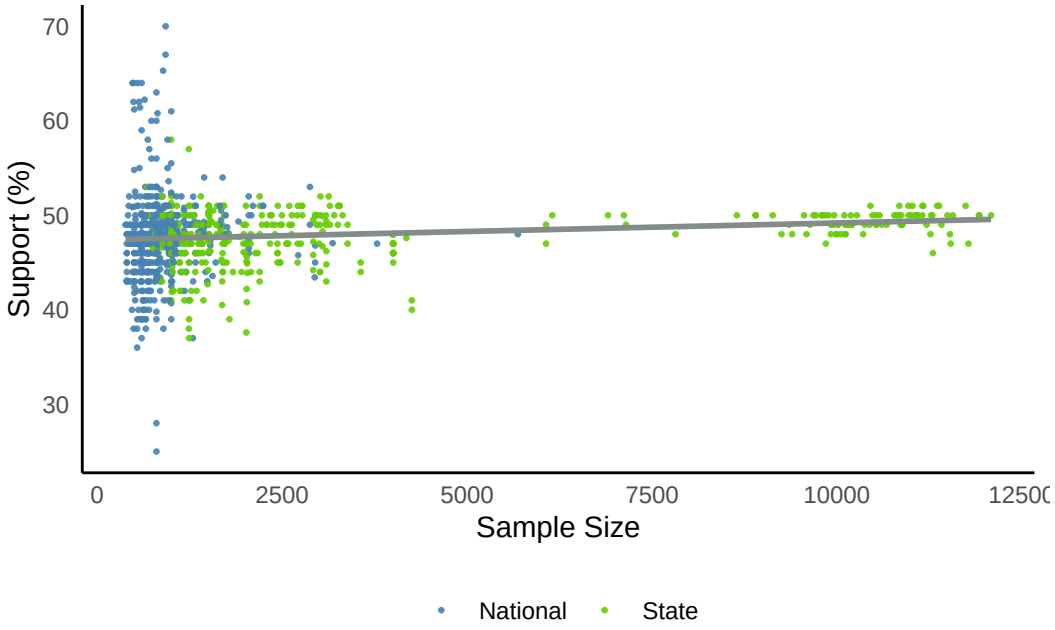


Figure 1: Sample Size vs. Support for Kamala Harris, Colored by Pollscope (National vs. State)

3.2.2 Variation in Support for Kamala Harris Among Most Frequent Pollsters

Figure 2 above shows the variation in support for Kamala Harris across the five most frequent pollsters. Notably, the median support levels differ between pollsters, indicating variability in central estimates of Harris's support. For instance, Siena/NYT reports a higher median compared to Redfield & Wilton Strategies. Additionally, the differing ranges of support estimates highlight potential discrepancies in the methodologies, sample sizes, and biases employed by each pollster. This suggests that different pollster have different interpretations of candidate support, underlining the importance of aggregating data from multiple sources to avoid overreliance on any single pollster.

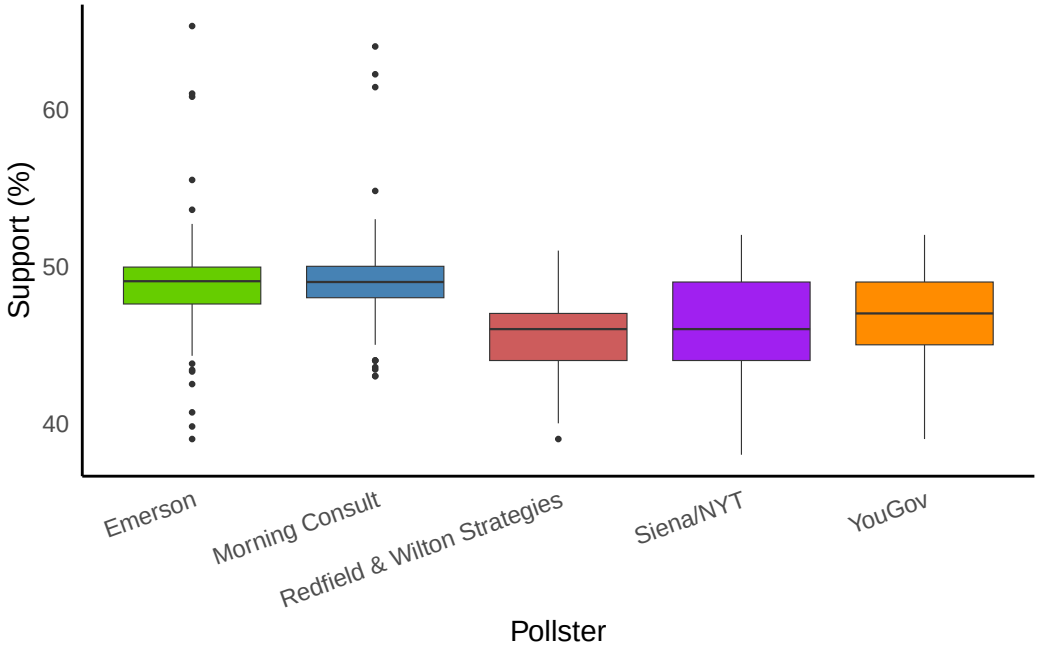


Figure 2: Support for Kamala Harris across the top 5 pollsters by poll count.

3.2.3 Support for Kamala Harris Based on Recency of Polls and Pollster Quality

In [@fig-enddate], we observe a slight increasing trend in support for Kamala Harris as the end date of polls progresses towards October. This may indicate a gradual improvement in her polling performance over time. The color gradient, reflecting the pollster's error and bias (Pollscore), reveals variations in bias: greener points represent more negative biases, while redder points indicate more positive biases.

While the recency of the polls shows a stable trend in Harris's support, the wide distribution of bias (Pollscore) values suggests that pollster quality can vary substantially. These variations

could influence the model's predictions if left unaddressed. Hence, including both recency and pollster quality as predictor variables in the model might be valuable to account for such variations. Further model diagnostics and validation will help decide if these variables are significant.

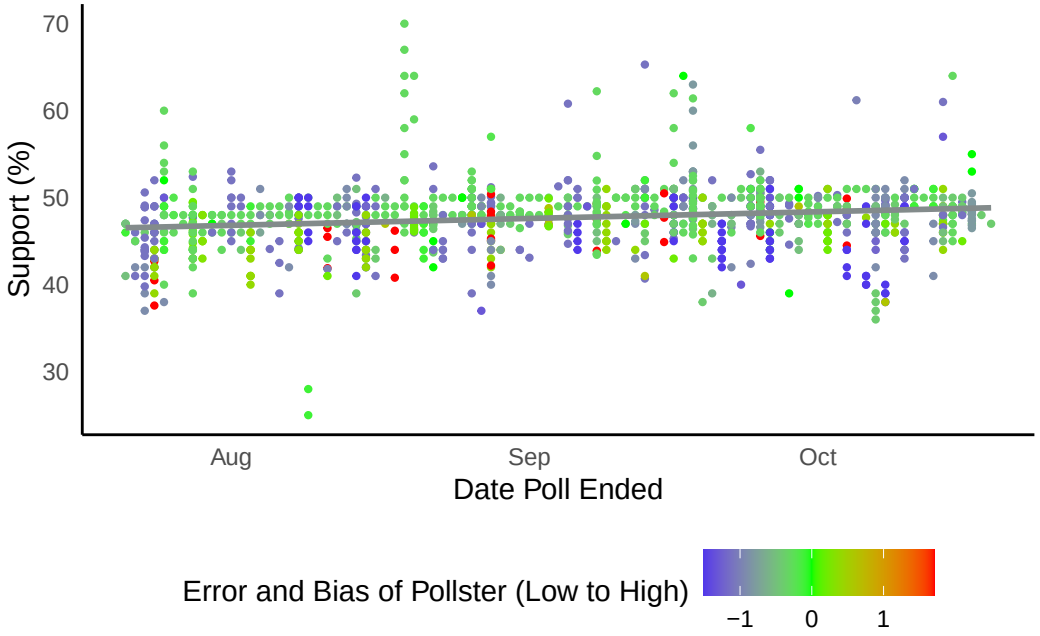


Figure 3: Average Support for Kamala Harris in Different States estimated by Polls

4 Model

In this section, we aim to address the inherent biases and differences present in various polling data to arrive at a robust prediction model. The core challenge lies in selecting a model with an optimal balance between complexity and fit, ensuring it accurately captures the dynamics of polling data while avoiding overfitting. To this end, we carefully evaluated different model specifications to determine the most appropriate one for our forecasting purpose.

Given that variables like numeric grade and pollscore are perfectly collinear with pollster, they were excluded from the regression analysis to avoid multicollinearity issues. These variables, however, remain integral to our weighting strategy, where they will be used to adjust for differences in polling accuracy and reliability. Instead, we focus on key features such as pollster, sample size, state, and recency, gradually adding complexity to the model.

By systematically comparing model specifications that incorporate these variables, we aim to select the model with the right balance between predictive accuracy and generalizability,

ultimately providing the best possible forecast.

4.1 Model Set Up

We aim to model the percentage of support for Kamala Harris in each poll as a function of the pollster the sample size, the state, and the recency of the poll.

$$y_i = \alpha + \beta_1 \cdot \text{pollster}_i + \beta_2 \cdot \text{sample\_size}_i + \beta_3 \cdot \text{recency}_i + \beta_4 \cdot \text{state}_i + \epsilon_i$$

Where

- y_i is the percentage of support for Kamala Harris in poll iii,
- α is the intercept,
- β_1 captures the effect of the polling organization,
- β_2 captures the effect of the sample size,
- β_3 captures the effect of recency (how recent the poll is),
- β_4 capture the effects of the different states
- ϵ_i represents the error term, assumed to follow a normal distribution with mean 0.

4.2 Model Justification

We compared three models with different sets of predictors, gradually increasing the model complexity. Model 1 only included **pollster** and **sample size** as predictors, while Model 2 introduced **state** as an additional factor. Finally, Model 3 incorporated **recency** as another key predictor. Among these models, Model 3 demonstrated the highest R^2 value (0.731), indicating the best fit and ability to explain the variation in the data. Additionally, Model 3 had the lowest AIC (4786.815) and BIC (5203.787) values, suggesting that it provided the best balance between model complexity and predictive performance. Therefore, Model 3 was selected as the most appropriate model for smoothing out polling biases and differences, offering the most accurate prediction of support for Kamala Harris.

Model	Variables Included	R^2	Adjusted R^2	F-statistic	Sample Size Significance	Recency Significance	States Significant
Model 1	Pollster + Sample Size + State	0.255	0.224	8.21	Not significant (p = 0.2059)	Not included	Not included

Model	Variables Included	R^2	Adjusted R^2	F-statistic	Sample Size Significance	Recency Significance	States Significant
Model 2	Pollster + Sample Size + State	0.666	0.641	26.00	Significant (p = 0.0032)	Not included	Several states
Model 3	Pollster + Sample Size + State + Recency	0.731	0.710	34.91	Significant (p = 0.0088)	Highly significant (p < 2e-16)	More states

Model	AIC	BIC
Model 1	5857.616	6093.732
Model 2	5026.643	5438.592
Model 3	4786.815	5203.787

4.3 Model Results

The table below shows the model output, where only a few states and pollsters are displayed to keep the output concise, as the full model includes many variables. We included "State California," "State Texas," "Pollster Redfield & Wilton Strategies," and "Pollster Echelon Insights" to provide examples of how different states and pollsters influence support for Kamala Harris.

The variable "Sample Size" has a positive coefficient, suggesting that polls with larger sample sizes are associated with slightly higher support percentages. "Recency" has a negative and highly significant coefficient, indicating that more recent polls tend to show lower support levels. Additionally, certain states like California and Texas show significant effects on the percentage of support, with California showing a notable increase. Similarly, different pollsters also display varying biases, as seen with the coefficients for "Pollster Redfield & Wilton Strategies" and "Pollster Echelon Insights." The model shows strong goodness-of-fit, with an R^2 of 0.731, indicating that 73.1% of the variability in support can be explained by the included variables.

```
=================================================
                          Dependent variable:
                      ---------------------------
                                  pct
-------------------------------------------------
Sample Size                  1.373 (0.892)
Recency                    -1.681*** (0.412)
State California          0.0001** (0.00004)
```

```
State Texas                      14.376*** (2.493)
Pollster Redfield                Wilton Strategies
Pollster Echelon Insights        -0.040*** (0.003)
-------------------------------------------------------
Observations                            1,123
R2                                      0.731
Adjusted R2                             0.710
=======================================================
Note:                        *p<0.05; **p<0.01; ***p<0.001
```

5 Prediction

5.0.1 Prediction

To predict Kamala Harris' overall support, we used a weighted average approach based on the quality of each poll. The weights are calculated using each poll's `numeric_grade`, which reflects the reliability and transparency of the polling methodology.

We define the weight for each pollster w_i as follows:

$$w_i = \frac{\text{numeric\_grade}_i}{\sum_{i=1}^{n} \text{numeric\_grade}_i}$$

where:

- w_i represents the weight assigned to poll i,
- $numericgrade_i$ is the numeric grade of poll i, and
- n is the total number of polls used in the analysis.

Using these weights, the overall weighted prediction of Kamala Harris' support is calculated by summing the weighted predicted values from our regression model:

$$\text{Overall Weighted Support} = \sum_{i=1}^{n} w_i \cdot \hat{y}_i$$

- $\hat{y}_i$ is the predicted percentage of support for Kamala Harris from poll i.

Applying this approach, our model predicts Kamala Harris' overall support to be **47.78%**. This prediction indicates that, after adjusting for biases in the polls, Harris' support is slightly below the 50% threshold. The weighted approach accounts for the differences in poll reliability and thus provides a more accurate estimate than using raw averages of the polls.

This result suggests that Kamala Harris is facing some challenges in gaining majority support based on the current polling data. However, the model reflects the weighted average from multiple sources, smoothing out individual poll biases to offer a comprehensive view of her support landscape.

6 Discussion

In this paper, we set out to predict Kamala Harris' support using a polls-of-polls approach. By aggregating multiple polls, we aimed to smooth out the inherent biases present in individual surveys and create a more accurate and reliable forecast. Our analysis was driven by a linear regression model that used key predictors, such as pollster, sample size, state, and recency of the polls. After obtaining predicted values from the model, we applied a weighting scheme based on the numeric grade of each pollster, reflecting the quality and reliability of the poll. By assigning higher weights to more reliable pollsters, we generated a final prediction for Kamala Harris' overall support, accounting for variations across different states and polling organizations.

One key insight from our analysis is the significance of poll recency. The negative and highly significant coefficient for recency indicates that **older** polls show lower support for Kamala Harris, while more recent polls tend to show higher support. This suggests that her popularity may be increasing over time, or that certain factors have shifted voter sentiment in her favor in more recent months. Inncorporating recency into the model also greatly improved its overall performance. Adding recency as a predictor significantly increased the model's explanatory power, raising the R^2 value and lowering both AIC and BIC scores compared to models that excluded it. This indicates that accounting for the timing of polls is essential to accurately predicting voter sentiment, as public opinion is highly dynamic and evolves rapidly in response to events. Including recency strengthens the model's ability to reflect these temporal shifts, leading to more reliable forecasts.

Our analysis also highlights the variability in support for Kamala Harris across different states and pollsters. For example, certain states like California and Texas exhibit distinct effects, with California significantly boosting Harris' predicted support. Additionally, different pollsters show varying levels of bias, as seen with Redfield & Wilton Strategies and Echelon Insights, which have divergent effects on support levels. This underscores the geographic and methodological complexities in predicting election outcomes, demonstrating that voter preferences are not homogenous across regions or polling firms.

Despite the strengths of the polls-of-polls methodology, there are several limitations to our approach. First, our model assumes that the relationship between the predictors (pollster, sample size, state, recency) and support for Kamala Harris is linear, which may not always be the case. Non-linear trends or interactions between variables might exist that were not captured by the model. Additionally, while we weighted polls based on their numeric grade, this metric might not fully reflect the true accuracy or bias of each pollster. Moreover, we only included polls conducted after Harris' official candidacy announcement, potentially omitting relevant data from earlier in the election cycle that could provide a fuller picture.

Future research could delve into more sophisticated methods to capture non-linear relationships and interactions between variables. For instance, incorporating machine learning algorithms like random forests or neural networks could allow for more flexible modeling of polling data. Another area of improvement could be refining the pollster weighting mechanism by integrating other factors, such as the historical performance of the pollster or specific characteristics of the electorate surveyed. Additionally, conducting more in-depth analyses of key battleground states could reveal more granular insights into voter behavior, especially in regions where small shifts in support could have outsized impacts on election outcomes.

In conclusion, while our model provides valuable predictions of Kamala Harris' support, there is always room to enhance the predictive accuracy by refining methodologies and incorporating new data sources. The evolving nature of elections demands continuous adjustment of models to reflect changes in public opinion, candidate strategies, and electoral dynamics.

Appendix

A Emerson College Polling Methodology (October 14-16, 2024)

A.1 Overview

Emerson College Polling, known for its extensive online polling methods, conducted a survey from October 14 to 16, 2024, targeting 1,000 likely voters. The poll measured voter preferences between Kamala Harris and Donald Trump, revealing a near tie with Harris at 50% and Trump at 49%.

A.2 Population, Frame, and Sample

The target population is likely voters in the U.S. elections. These voters are determined by the voters' likeliness to vote in the upcoming elections and the voter's voting history. Both of these factors are self-reported.

The sampling frame is the voters that are on the Aristotle list and on the online panel provided by CINT. Aristotle is an online supplier of voter files and CINT is a software company that connects researchers, such as Emerson, to online panels. This approach helps ensure a representative sample by accurately reflecting the electorate.

The sample size is the 1,000 likely voters chosen arbitrarily from the sampling frame. Likely voters are determined by a combination of voter history, registration status, and demographic data. These factors are self-reported.

A.3 Sample Recruitment and Approach

Emerson College used a mixed-mode sampling approach for its polling. Specifically, three primary modes were used to collect data for its October 14-16, 2024 poll:

- MMS-to-web text surveys: Respondents were contacted via text messages sent to cell phones. These messages directed participants to complete the survey online. This method was conducted using voter lists provided by Aristotle.
- Interactive Voice Response (IVR): Landlines were targeted using automated phone calls where respondents could answer questions using their keypads. The contact information for this was also sourced from Aristotle's voter lists.
- Online Panel from CINT: Emerson utilized a pre-screened, opt-in online panel of voters provided by CINT.

A.4 Advantages and Trade-offs in Mixed-Mode Sampling

The multi-mode sampling approach used in this poll has several advantages:

- Wider demographic reach: Having a combination of data collection methods allows Emerson to gather responses from a broader demographic group. In this polling, MMS-to-web surveys mostly gather responses from younger voters and those who primarily use mobile devices. IVR target older and rural voters, who are more likely to respond via landlines. Online panels is more efficient in reaching tech-savvy individuals.
- Cost-efficient: Automated technologies reduce the cost compared to live interviews. In a polling context, this allows for a larger sample size.

However, the trade-offs include:

- Coverage bias: IVR only reaches individuals with landlines, which may skew results towards older demographics.
- Self-selection bias: Respondents in online panels tend to be more engaged, which can lead to over-representation of politically active individuals.

A.5 Non-response Handling

Emerson addresses non-response using a weighting system. The data is adjusted based on key demographic variables like age, gender, race, education, and party affiliation, making the sample more representative of the electorate. Weights are applied to compensate for under- or over-represented groups, ensuring balanced results.

A.6 Questionnaire Design

The Emerson survey features a straightforward, efficient design focused on key issues and voter preferences. Strengths of the questionnaire include:

- Simplicity: Clear, concise questions makes the survey easy for respondents to understand and to provide accurate answers.
- Topical relevance: The poll zeroes in on vote preference and demographic splits, delivering timely insights during election cycles.

However, weaknesses include:

- Lack of nuance: While the survey captures basic preferences, it does not delve deeply into voters' motivations or the factors driving their choices.

- Mode limitations: Different survey modes can affect how respondents engage with the questions. This can potentially influence the depth and accuracy of voter responses.

A.7 Conclusion

Emerson College Polling's mixed-mode methodology is effective in balancing cost with broad demographic reach. Its weighting techniques contribute to reliable and accurate results despite non-responses. While the use of multiple survey modes increases population representation, possible biases—such as lack of nuance and mode effects need to be taken into consideration when interpreting results.

In sum, this method is a strong predictor of election outcomes and continues to adapt to the evolving challenges of modern polling.

B Idealized Methodology

B.1 Overview

This methodology outlines an election forecasting plan with a budget of $100,000. Heavily inspired by TIPP Insights and Emerson College's hybrid approaches, the survey methodology balances cost-efficiency and data accuracy by using different modes of data collection and wide demographic reach.

This forecasting plan will predict the popular vote and Electoral College outcomes, emphasizing key swing states and underrepresented voter groups.

B.2 Sampling Approach

This idealized methodology uses a multi-mode hybrid sampling approach. This allows the use of both probability-based and non-probability-based methods:

- Probability-based methods:
  - Live phone interviews (landline and cell): As used by TIPP, live phone interviews is a direct and personal method of collecting responses. Voter lists would be sourced from Aristotle to ensure we contact likely voters. This mode reaches older voters and voters with less internet connection.
  - Interactive Voice Response (IVR): Similar to Emerson's approach, IVR for landline users balances costs and ensures demographic representation, particularly among older or less tech-savvy populations. This method captures quick responses and is less expensive than live interviews.

- Non-probability methods:
  - MMS-to-web (mobile): Following Emerson's MMS-to-web strategy, we would send multimedia messages to likely voters' cell phones, encouraging them to complete an online survey. This method works well for younger, mobile-first voters, helping fill in gaps where traditional methods might miss.
  - Online panel recruitment via CINT: To mirror TIPP's online panel approach, a portion of the survey would be conducted via an opt-in panel of pre-screened voters from CINT, capturing tech-savvy respondents. CINT's platform ensures that different voter groups, including hard-to-reach groups, are represented in the polling.

B.3 Recruitment Strategy

By using Aristotle's voter lists and CINT's platform as a sampling frame, the sample consists of likely voters based on registration status and past voting behavior. Quota sampling would be used to match the demographic profile of the U.S. electorate. This ensures that respondents in the sample reflect key demographic variables such as age, gender, race, and education level.

We would oversample key swing states like Pennsylvania, Wisconsin, and Florida. Key swing states are states where there is no overwhelming support for a particular candidate or political party. The outcome in these states often determine the overall result of the election, making them critical in securing a win in the Electoral College. Oversampling these states allows a more accurate forecasting in these battlegrounds, which are critical for predicting the Electoral College outcome.

B.4 Data Validation

Given that the data is collected in a hybrid format, we would use duplicate detection methods via IP addresses or phone numbers to eliminate multiple responses from the same individual. Also, we would implement attention checks to identify respondents who are not fully engaged. Both duplicate respondents and unengaged respondents bias results, and should be removed from the data set.

Following TIPP's iterative weighting process, the data will be weighted to ensure that the sample aligns with key demographics (race, gender, education) based on census data.

B.5 Conclusion

By combining TIPP's live phone/IVR hybrid and Emerson College's mixed-mode survey methods, this idealized methodology captures diverse voter demographics. It uses both probability-based and non-probability-based sampling methods to forecast the 2024 U.S. Presidential Election.

A link to the survey can be found at: https://forms.gle/haGLaQBPQhrXE6Xa7

References

FiveThirtyEight. 2024. "FiveThirtyEight u.s. Election Polls." https://projects.fivethirtyeight.com/polls/.