Forecasting the 2024 U.S. Presidential Election Using a Model-Based Approach*

Harris expects to win on 49%, leads Trump by about 2% in 2024 US election

Diana Shen Jinyan Wei Jerry Yu

October 31, 2024

This paper predicts the outcome of the 2024 U.S. presidential election using a statistical model based on aggregated polling data for Kamala Harris and Donald Trump. We employ multi-level regression with post-stratification (MRP) using demographic predictors to estimate voter support. The analysis addresses polling biases and proposes an idealized survey methodology with a $100,000 budget. Our results offer insights into voter behavior and suggest improvements for future election forecasting models.

1 Introduction

Understanding voter sentiment is essential for both political campaigns and analysts, especially with the upcoming U.S. election on the horizon. Public opinion is highly dynamic and can change swiftly due to various influences, including media coverage, campaign tactics, and major events. This study aims to forecast the percentage of support for Kamala Harris, offering insights into the elements that shape voter preferences as the election nears. By examining data from multiple polling sources, we intend to pinpoint the key factors influencing support, such as the poll's end date, the polling organization, geographic location, and the poll's score. Our research seeks to fill a gap in existing literature that often fails to address the complexities of polling data, thereby enhancing our understanding of voter behavior within the electoral context.

The estimand of our analysis is the true percentage of voter support for Kamala Harris in the upcoming U.S. presidential election.

The main focus of our analysis is the percentage of support for Harris, which we will model using various predictor variables. We are particularly interested in how the end date, polling

*Code and data are available at: https://github.com/DianaShen1224/Forecast-2024-US-election.

organization, state, and poll score affect voter support. Using a linear regression framework, we can quantify the relationships between these predictors and the support outcome, providing clarity on how each factor influences overall support for Harris. By estimating the coefficients for each predictor, we aim to draw significant conclusions about their respective impacts on voter sentiment.

Our findings reveal a notable positive correlation between the end date and the percentage of support, indicating that as the election approaches, voter support tends to increase. We also observed considerable variability in support levels based on the polling organization and state, with certain pollsters consistently reporting higher support for Harris. The quality of the polls significantly affected results, with more reputable polls correlating with higher levels of support. These insights emphasize the necessity of considering both the timing of polls and the characteristics of different polling firms when analyzing public opinion.

This research is important because precise predictions of voter support are crucial for effective campaign strategies. By identifying the main factors influencing support for Harris, campaign teams can customize their outreach and messaging to resonate better with voters. Furthermore, recognizing the differences across various polling organizations and states can assist in resource allocation and strategic focus during the campaign. Given that elections can be decided by narrow margins, having trustworthy insights into voter preferences can substantially influence the final outcomes.

The structure of this paper is organized as follows: (**third-data?**) provides details on the data sources and variables used in our analysis. (**forth-model?**) explains the modeling approach, including the assumptions and specifications of our linear regression framework. In (**fifth-results?**), we present our findings, emphasizing the key predictors of Harris's support. Finally, (**sixth-discussion?**) explores the implications of our results and suggests potential directions for future research. (**appendix-a?**) provide external data detail, (**appendix-b?**) provide model detail, (**appendix-c?**) provide a exploration for pollster methodology, (**appendix-d?**) provide a idealized methodology.

2 Data

2.1 Overview

We conduct our polling data analysis using the R programming language (R Core Team 2023). Our dataset, obtained from FiveThirtyEight (FiveThirtyEight 2024),based on polling as of 27 October 2024, provides a detailed overview of public opinion in the lead-up to the election. Adhering to the guidelines presented in *Telling Stories with Data, Author = Rohan Alexander, Year = 2023, Publisher = Chapman; Hall/CRC, Url = https://tellingstorieswithdata.com/* (n.d.), we explore various factors that influence voter support percentages, including the timing of the polls, the traits of polling organizations, and regional differences.

In this study, we utilized several R packages to enhance our data manipulation, modeling, and visualization capabilities. The tidyverse package offered a comprehensive set of tools for data wrangling and analysis, improving workflow efficiency ("Welcome to the tidyverse, Author = Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy D'Agostino McGowan and Romain François and Garrett Grolemund and Alex Hayes and Lionel Henry and Jim Hester and Max Kuhn and Thomas Lin Pedersen and Evan Miller and Stephan Milton Bache and Kirill Müller and Jeroen Ooms and David Robinson and Dana Paige Seidel and Vitalie Spinu and Kohske Takahashi and Davis Vaughan and Claus Wilke and Kara Woo and Hiroaki Yutani, year = 2019, journal = Journal of Open Source Software, volume = 4, number = 43, pages = 1686, doi = 10.21105/joss.01686, ," n.d.). The here package aided in managing file paths, allowing for easy access to our data files (Müller 2020). We relied on janitor to perform data cleaning, as it provides functionalities to identify and rectify quality issues within the dataset (Firke 2023). For handling date-related operations, the lubridate package proved invaluable, simplifying the manipulation of time variables (Grolemund and Wickham 2011). Lastly, arrow supported efficient data input and output in a performance-oriented format, essential for managing larger datasets (Richardson et al. 2024). Our coding practices and file organization were informed by the structure outlined in *Telling Stories with Data, Author = Rohan Alexander, Year = 2023, Publisher = Chapman; Hall/CRC, Url = https://tellingstorieswithdata.com/* (n.d.).

2.2 Measurement

The process of translating real-world events into our dataset requires a systematic approach to measurement and data gathering. In this research, we aim to assess public opinion regarding Kamala Harris as the upcoming U.S. presidential election approaches. Polling agencies formulate surveys featuring specific questions designed to capture voters' attitudes, including their likelihood of supporting Harris and their views on prevailing political issues.

Once the survey items are established, a representative sample is selected through stratified random sampling methods, ensuring a diverse demographic representation. Respondents are reached using various techniques, such as telephone interviews and online questionnaires.

After gathering the responses, the data is subjected to thorough cleaning and validation procedures to rectify inconsistencies and handle any missing information. This step is crucial for ensuring that the dataset accurately mirrors the electorate's sentiments. Each entry in the finalized dataset reflects an individual's viewpoint at a given moment, enabling a detailed analysis of the factors that shape public opinion as the election nears. This structured methodology effectively converts subjective opinions into measurable data, providing valuable insights into voter behavior and preferences.

2.3 Outcome variable

2.3.1 The support percentage of Kamala Harris.

Figure **??** illustrates the distribution of percentage support for Kamala Harris based on polling data, where support reflects the proportion of respondents favoring Harris in each survey. Most polls report support clustered around 50%, with the majority of values falling between 40% and 55%. This central peak suggests moderate, consistent support levels among respondents, with fewer instances of higher support levels above 55%. The right-skew in the distribution indicates occasional polls with elevated support, though these are less common. Overall, this visualization highlights the general sentiment and variability in support for Harris as captured across multiple polls.

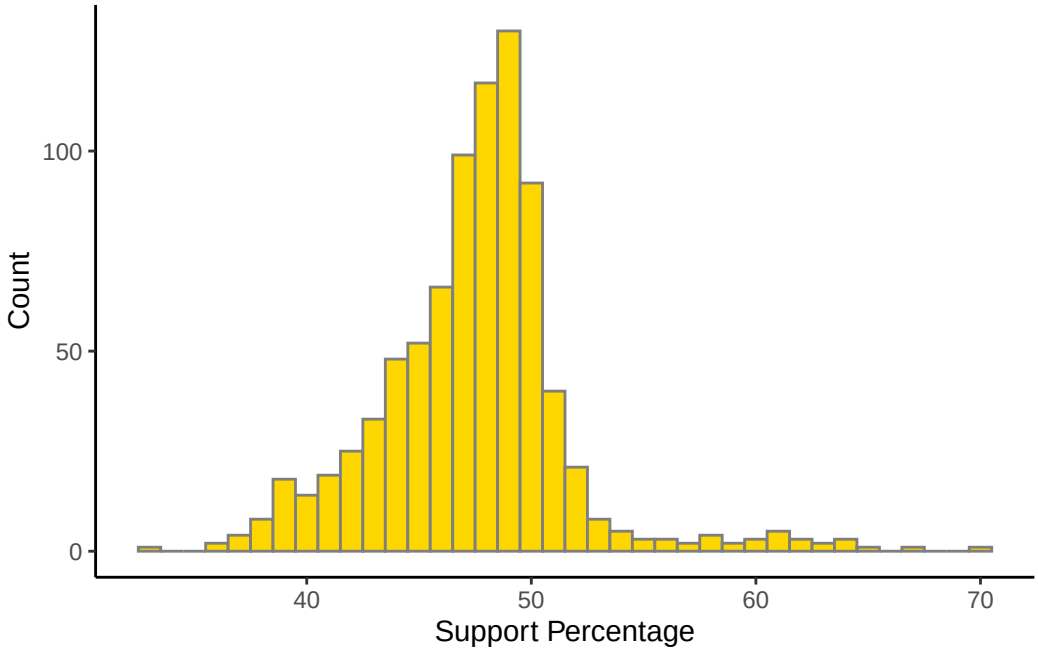


Figure 1: Distribution of support percentage of Kamala Harris

2.4 Predictor variables

2.4.1 End Date

The end date is the final day of data collection for a poll, indicating when the survey period concluded. This date provides crucial context for the poll results, as public opinion may change over time in response to events, campaign actions, or other influencing factors.

2.4.2 State

Figure **??** displays the distribution of polls by state, showing how frequently polling organizations conducted surveys across various U.S. states and at the national level. The "National" category has the highest number of polls, indicating a strong emphasis on capturing overall U.S. sentiment. Certain states, such as Pennsylvania, Wisconsin, North Carolina, Arizona, Georgia, and Michigan, also show higher polling frequencies, likely because these are battleground states with the potential to influence the election outcome significantly. Toward the right side of the chart, states with minimal polling activity, including South Carolina, Iowa, and Washington, appear less frequently, possibly due to their historically predictable or less competitive nature. This distribution reflects the strategic focus of polling efforts, with organizations prioritizing both national sentiment and swing states where public opinion is more volatile. Overall, the chart provides insight into where polling resources are allocated as election day nears, emphasizing areas that could sway the final result.

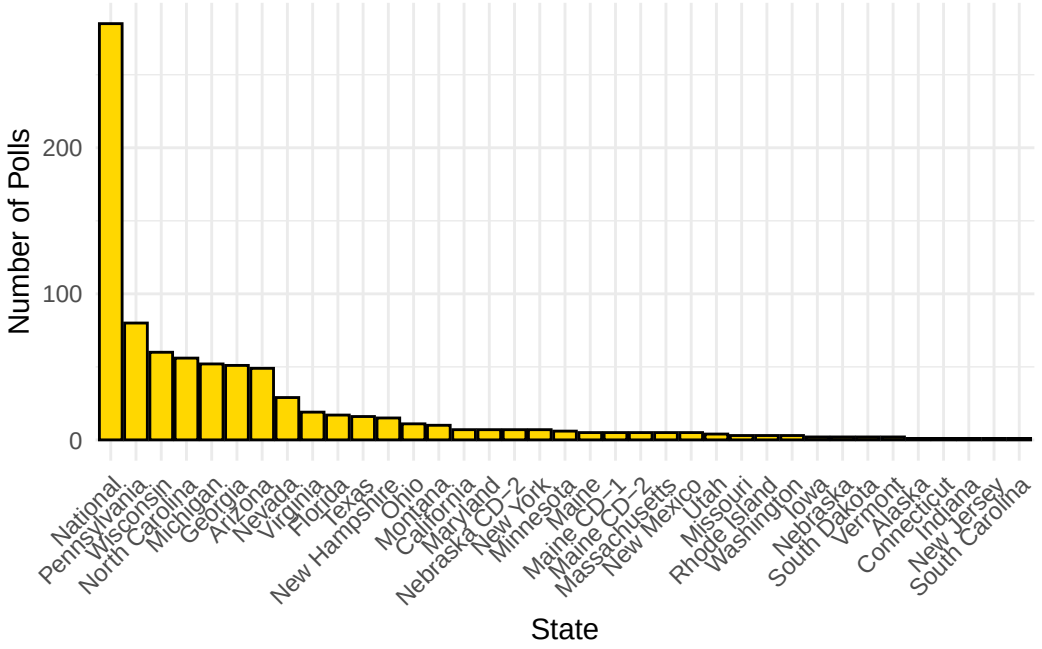


Figure 2: Count of polls by state

2.4.3 Pollster

Figure **??** displays the frequency of polls conducted by different pollsters. Each bar represents a pollster, with the height indicating the number of polls they conducted. Siena/NYT has the highest count, followed by YouGov and Emerson. Pollsters to the right have conducted significantly fewer polls, with some showing only one or two entries.

The pollster is the organization or firm that conducts the surveys, gathering and analyzing public opinion data on voter preferences. In this context, each pollster's count reflects its level of polling activity related to the election.

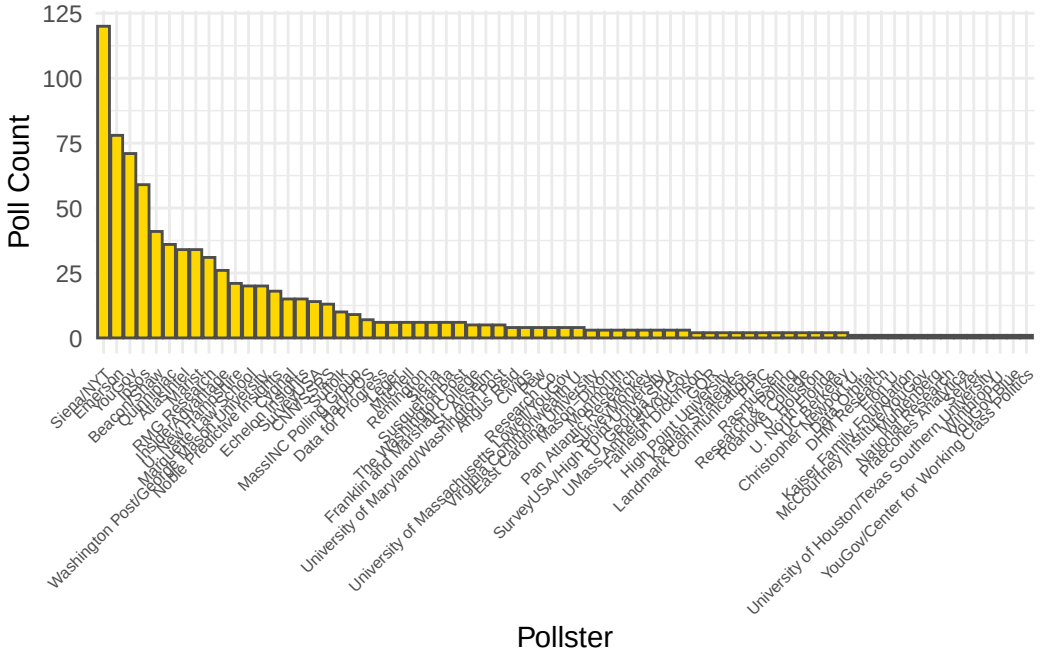


Figure 3: Frequency of Polls by Pollster"

2.4.4 Numeric Grade

Figure **??** displays the distribution of numeric grades assigned to various pollsters, with the x-axis representing the numeric grade values and the y-axis indicating the frequency of each grade. The numeric grade is a metric that evaluates the quality or reliability of a pollster, taking into account factors such as methodology, historical accuracy, sample size, and pollster reputation. In this chart, we observe that the most common numeric grades are concentrated around 2.75 and 3.00, with a large spike at these values, suggesting that a significant number of polls are conducted by highly-rated pollsters. Fewer polls have grades below 2.5, indicating a relatively lower occurrence of polls from less-reliable pollsters in this dataset. This distribution underscores the emphasis on high-quality pollsters within the dataset, ensuring a more reliable and consistent source of polling data for subsequent analysis.

2.5 Relationships between key varaibles

Figure **??** shows the trend of percentage support for Kamala Harris over time, with data points and smoothed trend lines for each pollster. The x-axis represents the dates from August

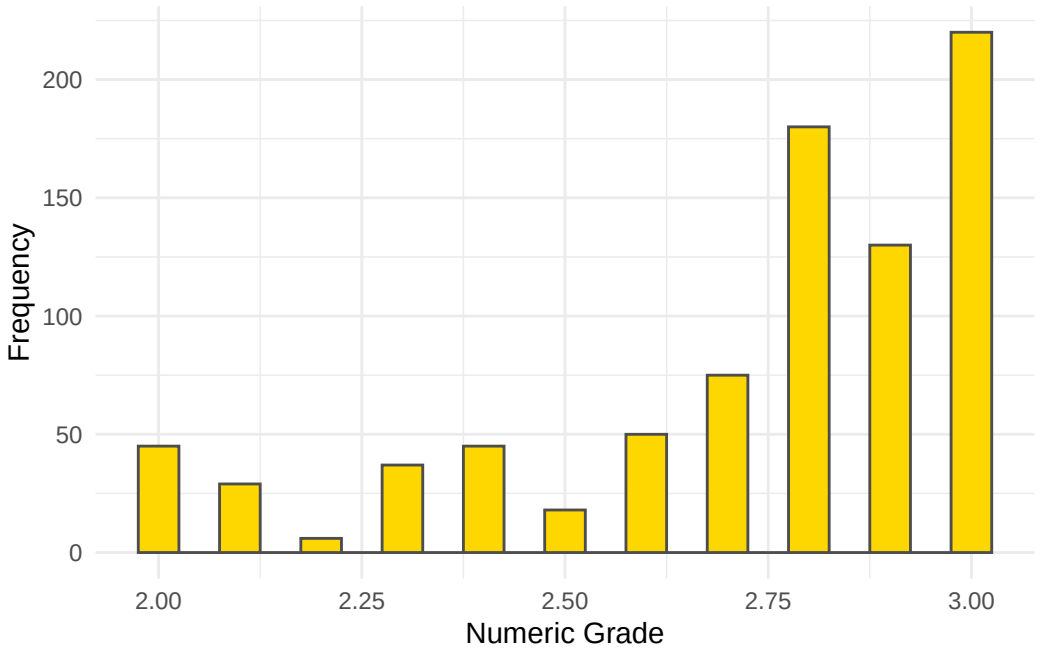


Figure 4: Distribution of numeric grade

through October, while the y-axis indicates the percentage of respondents supporting Harris. Each color corresponds to a specific polling organization, with prominent pollsters such as Beacon/Shaw, Ipsos, Siena/NYT, Emerson, Quinnipiac, and YouGov. The smoothed lines reveal subtle trends over time, with some pollsters like Emerson and Beacon/Shaw showing a slight upward trend, while others like Siena/NYT display a downward trend. Figure **??** highlights the variability in poll results across different organizations, reflecting each pollster's methodology and sample.

Figure **??** illustrates Harris's support percentages over time, with each data point representing poll results colored by the numeric grade of the pollster. The smoothing lines for each numeric grade remain subtle, indicating minor variations in support trends over time across different pollster quality levels. Most points are clustered around the 50% support level, suggesting a generally stable voter sentiment for Harris, with only slight fluctuations across numeric grades. By limiting the y-axis, outliers and extreme variations are minimized, resulting in a clearer and more interpretable visualization. This approach emphasizes the central trend, allowing for clearer comparisons of support levels across pollsters of varying reliability without distraction from extreme values.

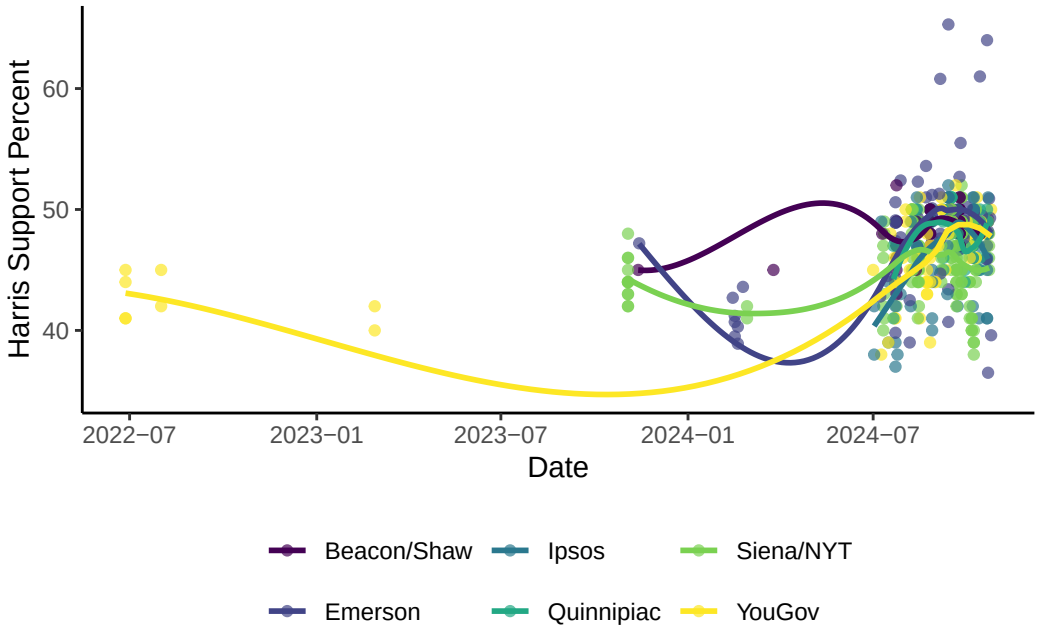


Figure 5: Harris Support Over Time by Pollster (Top 6 Pollsters)

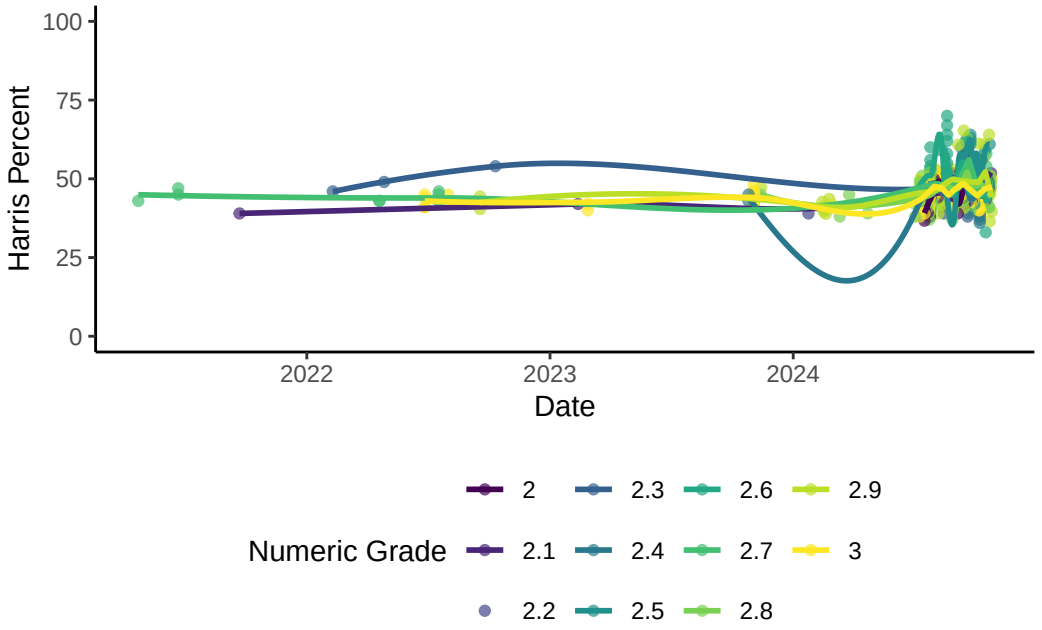


Figure 6: Harris Support Over Time by Numeric Grade

2.6 Model

We conducted a regression analysis to predict support for Kamala Harris and Donald Trump in the 2024 U.S. presidential election, including models for Trump for comparison. Our analysis includes four models: **unweighted linear models** and **weighted linear models** for both candidates. These models predict each candidate's support as a continuous outcome across different states, using **pollster**, **national poll**, and **population** as key predictors. The weighted models additionally incorporate adjustments for recency, sample size, and poll quality to account for potential biases. This setup aligns with the New York Times methodology for election polling, which assigns weights based on poll quality, sample size, and recency to better reflect voter sentiment ("How We Conduct Our Election Polls" 2024).

The validation and diagnostic details of the model are provided in Appendix B

2.7 Model Set-up

We implemented our linear regression models using the `lm()` function in R.

2.7.1 Mathematical Expressions

Both the unweighted and weighted models share the same mathematical expression format:

2.7.1.1 Unweighted Model

The unweighted model for Harris provides a baseline by treating all polls equally without adjustments for recency, sample size, or poll quality:

$$\text{Support\_Harris}_i = \beta_0 + \beta_1 \cdot \text{National\_Poll}_i + \beta_2 \cdot \text{Pollster}_i + \beta_3 \cdot \text{Population}_i + \epsilon_i$$

2.7.1.2 Weighted Model

The weighted model for Harris incorporates adjustments for recency, sample size, and pollster quality:

$$\text{Support\_Harris}_i = \beta_0 + \beta_1 \cdot \text{National\_Poll}_i + \beta_2 \cdot \text{Pollster}_i + \beta_3 \cdot \text{Population}_i + \epsilon_i$$

2.7.2 Coefficient Explanations

- β_0: Intercept of the model, representing the expected support when all predictors are zero.
- β_1: Coefficient for the national poll, indicating how much support changes with a one-unit increase in the national poll percentage.
- β_2: Coefficient for the pollster, reflecting the impact of different polling organizations on support.
- β_3: Coefficient for the population, accounting for the influence of the demographic context on support levels.

2.7.3 Model Justification

1. **Weighted Least Squares Estimation**

In our analysis, we utilize weighted least squares estimation to account for the varying quality of polling data. The weights are crucial in the estimation process, leading to the following expression for the estimates of the coefficients ($\widehat{\beta}$):

$$\widehat{\beta} = (X^T W X)^{-1} X^T W y$$

Where: - X is the design matrix of predictors. - W is the diagonal matrix of weights, which incorporates factors such as recency, sample size, and pollster quality to enhance the reliability of our estimates.

The weighted model is essential for accurately reflecting voter sentiment. By incorporating weights, we adjust for the reliability of the polling data, ensuring that more credible polls have a greater influence on the estimates.

2. **Calculation of Variables**:

   - **Combined Weight**: The combined weight is calculated as follows:

     $$\text{combined\_weight} = \text{recency\_weight} \times \text{sample\_size\_weight} \times \text{poll\_frequency\_weight} \times \text{pollster\_qualit}$$

   - **Recency Weight**: This weight uses an exponential decay function:

     $$\text{recency\_weight} = \exp(-\text{Recency}_i \cdot 0.1)$$

     This reflects the diminishing influence of older polls, with $\lambda = 0.1$ for the exponential decay.

- **Sample Size Weight**: This weight adjusts the significance of each poll based on the number of respondents, capping the weights at a maximum of 2,300 responses to reflect the reliability of larger sample sizes ("How We Conduct Our Election Polls" 2024).
- **Poll Frequency Weight**: This weight considers how often a pollster conducts polls, with higher weights assigned to pollsters with a greater number of recent surveys.
- **Pollster Quality Weight**: Based on the historical performance of the pollster, this weight emphasizes the reliability of their polling methods.

2.7.4 Alternative Models

While this analysis employs linear regression with weighted least squares, alternative methods such as Bayesian modeling could provide additional insights. However, we opted for the weighted linear a regression approach to maintain interpretability and ease of implementation given the scope and resources of this analysis. The Bayesian approach was not chosen primarily due to the complexity involved in defining priors and the additional computational requirements.

2.7.5 Importance of Combined Weights

By incorporating these combined weights, the weighted model offers a more nuanced estimation of voter support for both Harris and Trump. While the mathematical expression for the models remains consistent, the weighted models effectively adjust for variations in data quality and recency, leading to more accurate predictions of voter sentiment.

3 Result

3.1 Recent Three Months Support Trends Prediction

The recent three-month support trends for Kamala Harris and Donald Trump across key battleground states—Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin—show a highly competitive race. Each state's panel displays support percentages over time from August to November 2024, with Harris shown in blue and Trump in red. The plot uses point size to represent poll weights, accounting for factors such as recency, pollster quality, and sample size, where larger points indicate higher-weighted polls, emphasizing more reliable data.

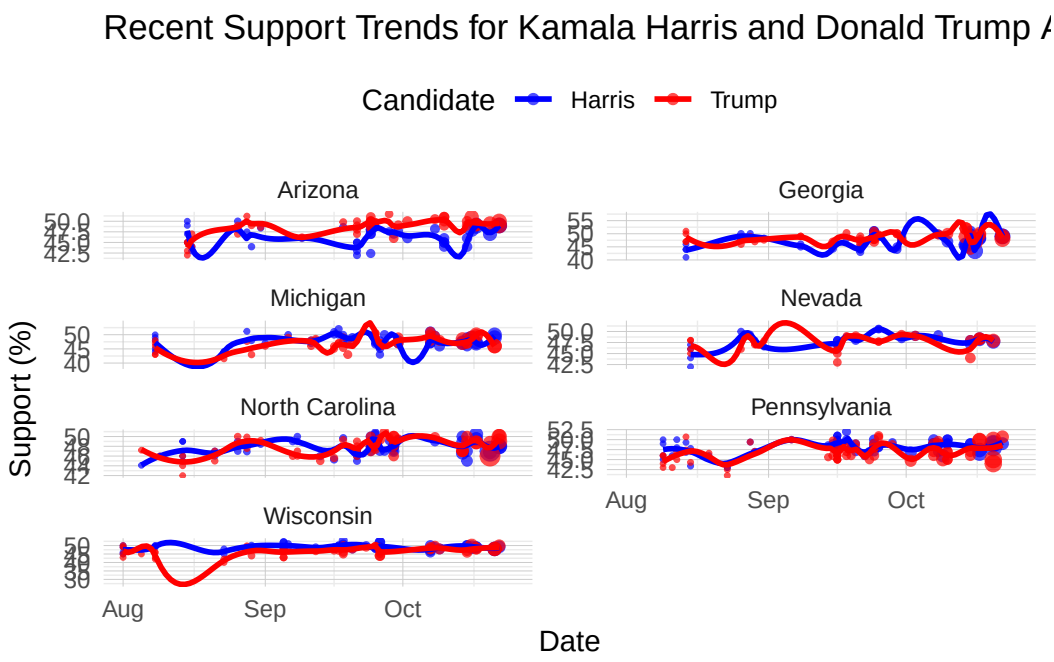


Figure 7: Support Trends for Kamala Harris and Donald Trump Across Selected States

Figure **??** illustrates the recent support trends for Kamala Harris and Donald Trump across selected battleground states from August to November 2024. Each subplot represents a different state, with support percentages for each candidate displayed over time. Key features include Harris shown in blue and Trump in red, with smoothed trend lines to capture changes over time. The plot's y-axis spans from approximately 30% to 50% support, while the x-axis highlights monthly intervals from August to November. For example, in Arizona, Trump maintains a slight lead with support fluctuating around 50% in October, compared to Harris around 48%. In Georgia, Trump's support reached just above 49% in late September, whereas Harris hovers closer to 47%. In Wisconsin, both candidates show tightly aligned trends near 48%, highlighting a close race. Each point's size reflects the poll weight—calculated using recency, pollster quality, and sample size—indicating the relative importance of each poll. This visualization emphasizes recent trends in candidate support across key battleground states.

Figure **??** illustrates the polling support trends for Kamala Harris and Donald Trump across the country from August to October. The scatter points represent individual poll results, with blue for Harris and red for Trump, while each dot's size reflects the weight of the poll, accounting for factors like sample size, quality, and recency. Harris's support trend line generally remains above Trump's, with an average support level around 49%, compared to Trump's approximate average of 47% over this period. Toward the end of October, both candidates show a slight increase, with Harris's support reaching just over 50% and Trump's approaching 49%. This trend suggests a potential narrowing in support levels as the election approaches, providing a quantitative snapshot of national polling dynamics.

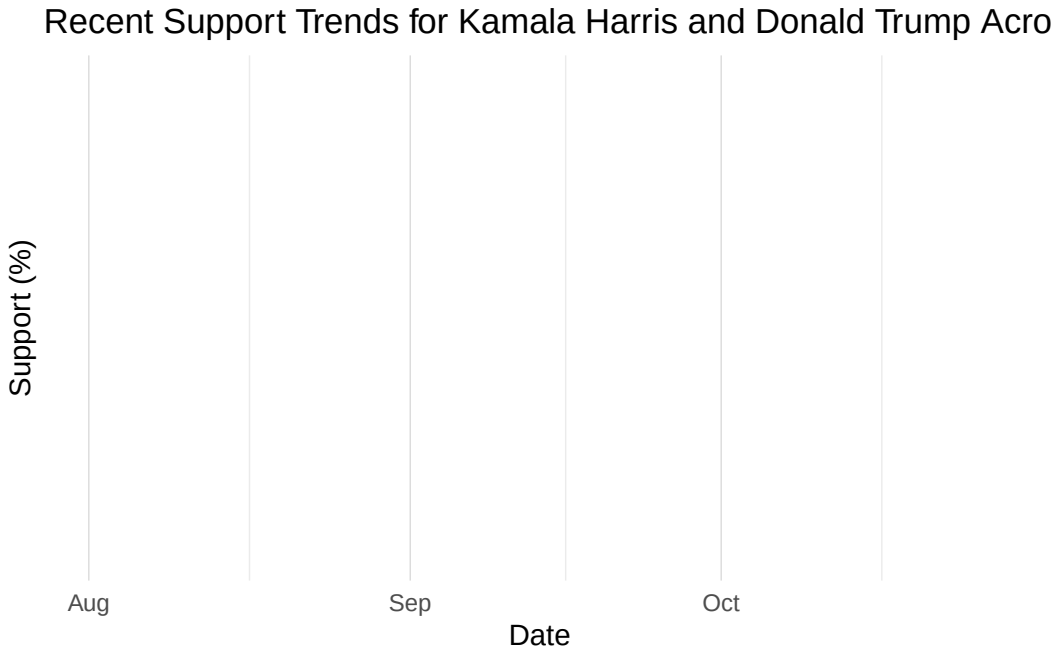


Figure 8: Support Trends for Kamala Harris and Donald Trump Across Selected States

3.2 Predicted Electoral Outcomes for Recent Three Months by Model

The predicted support for Kamala Harris and Donald Trump in the upcoming 2024 U.S. presidential election is illustrated in Figure 1. This analysis spans the period from August 1, 2024, to October 27, 2024, and uses multiple linear regression models to project voter support.

The unweighted model predictions indicate that Kamala Harris maintains a consistent support level, with projections hovering around 49%. The blue dashed line represents this trend, showing only minor fluctuations throughout the polling period. Conversely, the weighted model for Harris, indicated by the solid blue line, suggests a slight upward trend in predicted support, reflecting a potential positive shift in voter sentiment as the election date approaches.

In contrast, Donald Trump's predicted support is represented by the red lines. The unweighted model forecasts his support to remain around 48%, while the weighted model predicts a gradual decline to approximately 46%. The red dashed line indicates the unweighted prediction, whereas the solid red line shows the weighted prediction. This downward trajectory may indicate challenges for Trump in sustaining voter enthusiasm amid changing public opinions.

At the conclusion of the polling period, the models suggest that Kamala Harris leads Donald Trump by approximately 2.5% in the un-weighted (49% vs. 46.5%) and 1% in weighted models (49% vs. 48%). Average leading by about 2%.

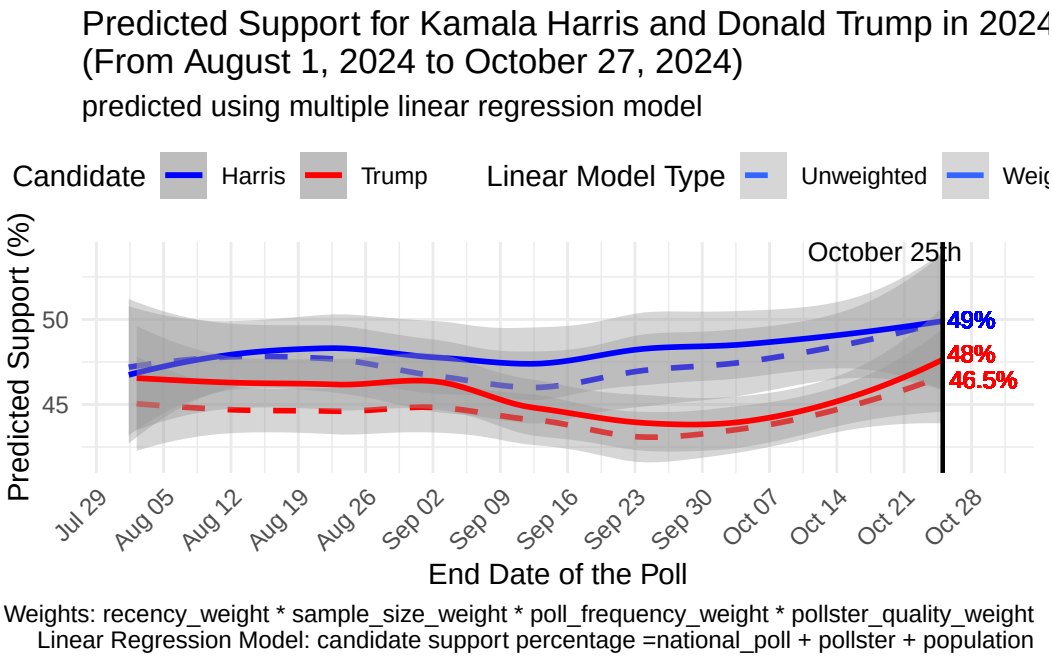


Figure 9: Predicted Support Trends for Kamala Harris and Donald Trump using Fitted Linear Model

4 Discussion

4.1 Key Findings and Real-World Implications

Our model reveals a portrait of a nation deeply divided, with Harris holding a narrow 49% edge over Trump's 47% in national support. However, as recent elections have shown, the national popular vote does not always translate into an Electoral College victory. The U.S. electoral system, which awards each state's electoral votes to the candidate with the majority of votes in that state, means that winning the popular vote nationally might still leave Harris short of the presidency. This structural quirk played a defining role in the 2016 election, where Hillary Clinton's popular vote win failed to deliver the electoral majority, paving Trump's path to the White House. Our findings underscore this enduring tension: while Harris may have a slight national advantage, the true battle will be fought state by state, in a handful of battlegrounds that could flip the election either way.

In states like Wisconsin and Arizona, our model highlights just how close the race remains. Wisconsin, for instance, shows both candidates locked in a near tie around 48% support—a statistical dead heat that brings the state's significance into sharp relief. Arizona, meanwhile, leans modestly toward Trump, a reflection of the unique demographic and political nuances shaping each battleground. These state-level insights illuminate a critical truth: while national polls offer a snapshot of overall sentiment, they risk obscuring the specific, local dynamics that

will ultimately decide the outcome. The stakes are high; these are the states that could swing, the margins that will be watched closely on election night, as even slight changes in turnout or last-minute shifts in opinion could tip the balance.

Yet, as with all models, limitations persist. While our weighting system, with its emphasis on recent data, aims to capture a timely snapshot of voter sentiment, it may miss sudden changes sparked by campaign events or unexpected shifts in public discourse. The variability in our state projections—ranging from a low estimate of 363 electoral votes for Harris to a high of 471—reflects the inherent uncertainty in polling-based models, particularly given the constraints of our data sources. This wide range suggests that while data-driven insights can illuminate trends, they cannot fully account for the unpredictable nature of electoral outcomes. Moving forward, incorporating real-time sentiment from social media or demographic-specific insights could offer a more responsive understanding of voter behavior. Ultimately, these findings highlight the intricate dance between popular sentiment and the electoral mechanics of American democracy—a system where every vote counts, but some states matter just a bit more than others.

4.2 Second discussion point

4.3 Third discussion point

4.4 Weaknesses and next steps

Appendix

A Additional data details

A.1 Dataset and Graph Sketches

Sketches depicting both the desired dataset and the graphs generated in this analysis are available in the GitHub Repository `other/sketches`.

A.2 Data Cleaning

In this data-cleaning process, we focus on refining raw polling data for Kamala Harris and Donald Trump to enhance its quality and relevance for analysis. The process begins by loading the dataset and using the janitor package to standardize column names, ensuring consistent naming conventions throughout. We then filter the data to retain only essential columns and remove rows with missing values in key fields, including numeric_grade, pct, sample_size, and end_date.

For each candidate, we isolate polls specifically for Kamala Harris and Donald Trump, retaining only high-quality polls with a numeric_grade of 2 or higher—given that the average numeric_grade is approximately 2.175, with a median of 1.9. We also handle placeholder values in state information by setting entries marked as "–" to NA, and create a national_poll indicator, assigning a value of 1 for national polls and 0 for state-specific ones. Dates are standardized using the lubridate package to facilitate accurate time-based analysis.

Recency weights are calculated based on the days elapsed since the poll's end date, applying an exponential decay function to prioritize more recent polls. Weights based on sample size are capped at a maximum of 2,300 responses to maintain balanced representation. Additionally, categorical variables, including pollster, state, candidate_name, population, and methodology, are converted to factors to prepare for analysis.

The cleaned datasets for both candidates are then saved as Parquet files for efficient storage and access in further modeling and analysis. This structured approach ensures that the data is accurate, complete, and optimized for insightful statistical analysis.

A.3 Attribution Statement

B Model details

B.1 Model validation: K-Fold Cross-Validation

```
For the Harris model, the RMSE is 3.46, R-squared is 0.34, and MAE is 2.50.
```

```
For the Trump model, the RMSE is 4.44, R-squared is 0.33, and MAE is 3.19.
```

We use a 10-fold cross-validation on two linear regression models—one for Harris and one for Trump. The models use three predictors: national_poll, pollster, and population. The output provides key metrics, which breaks down here:

RMSE (Root Mean Square Error): Measures the average magnitude of prediction errors (lower is better).

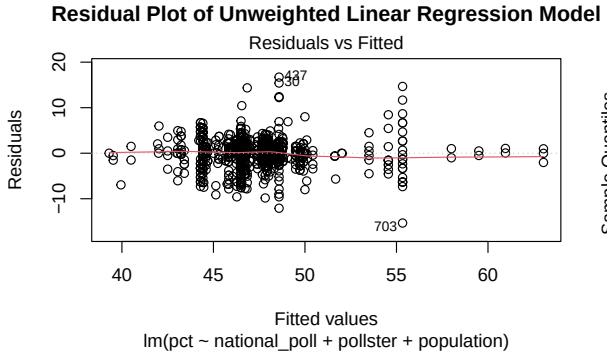
Harris model: RMSE of 3.44, indicating an average prediction error of around 3.44 percentage points.
Trump model: RMSE of 4.40, showing a slightly higher prediction error on average.

R-squared: Represents the proportion of the variance in the response variable explained by the model (higher is better).
Harris model: R-squared of 0.34, meaning the model explains about 34.6% of the variance in Harris's polling data.
Trump model: R-squared of 0.34 as well, indicating similar explanatory power for Trump's polling data.

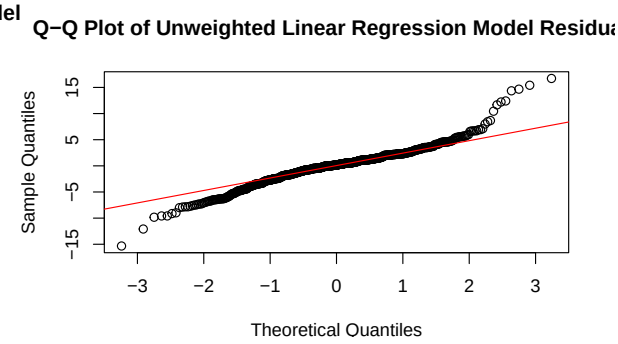
MAE (Mean Absolute Error): Shows the average absolute difference between observed and predicted values (lower is better).
Harris model: MAE of 2.48, meaning that, on average, the predictions are off by 2.48 percentage points.
Trump model: MAE of 3.15, indicating slightly less precise predictions compared to the Harris model.

Interpretation Summary Predictive Accuracy: The Harris model has slightly better predictive accuracy than the Trump model, as reflected by its lower RMSE and MAE values.
Model Fit: Both models explain roughly 34% of the variance in their respective datasets. This suggests that other factors not included in the model may play a significant role in explaining the remaining variance.
This summary indicates the models are moderately predictive, with room for improvement in accuracy and fit, potentially by adding more predictors or adjusting model specifications.

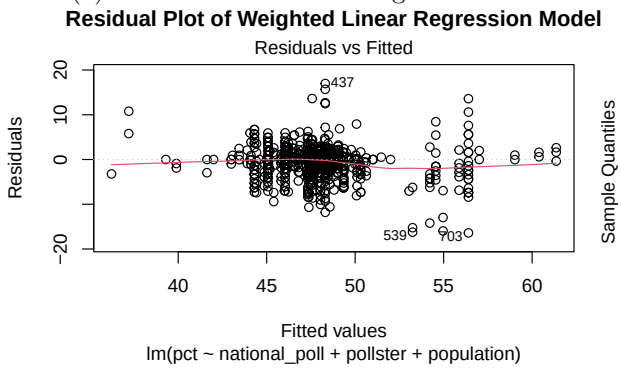
B.2 Diagnostics



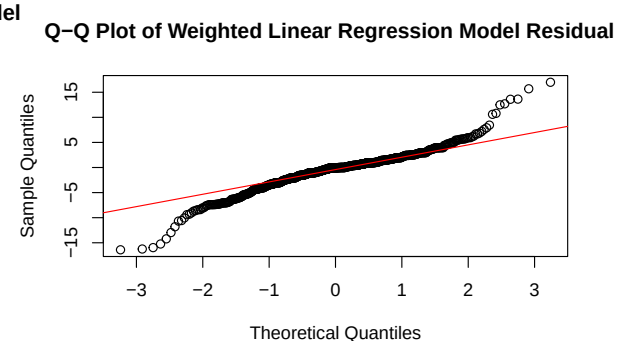
(a) Residual Plot for Unweighted Model



(b) Q-Q Plot for Unweighted Model



(c) Residual Plot for Weighted Model



(d) Q-Q Plot for Weighted Model

Figure 10: Diagnostics of model using residual vs fitted plot and norm Q-Q plot -Support for Harris

Generally, we use Residual vs Fitted plot and Q-Q Plot diagnostic our model. Residual vs Fitted plot aare Residuals (differences between observed and predicted values) plotted against fitted values. Ideally, these residuals should be randomly scattered around the zero line to indicate that the model does not have systematic errors. The Q-Q plot for the unweighted model shows how the residuals align with a theoretical normal distribution. Ideally, residuals should follow a straight line in this plot if they are normally distributed, which is an assumption of linear regression.

Figure **??** is a residual plot of un-weighted model for Harris support. It showsthe residuals are generally spread around zero, with no clear pattern. This suggests the model is relatively well-specified. However, there is a slight curvature, indicating potential non-linearity that the model may not fully capture. A few notable outliers with larger residuals might be influencing the model, indicating that some data points have more significant prediction errors.

Figure **??** is a Q-Q plot of un-weighted model plot for Harris support. It shows most residuals fall along the line, especially in the middle range. This suggests that our model satisfy the

normality assumption. However, some points at the tails deviate, indicating potential outliers or non-normality in the extreme residual values. This slight deviation at the ends suggests the model might have some issues with extreme predictions but performs reasonably well overall.

Figure **??** is a residual plot of weighted model for Harris support.It shows residuals are again plotted against fitted values. Similar to the unweighted model, the residuals are mostly centered around zero, indicating that the weighted model captures the general trend without significant systematic bias. The curvature is slightly reduced compared to the un-weighted model, suggesting that weighting has helped in addressing some of the non-linearity observed in the un-weighted model. However, some residuals are still notably large, which may indicate outliers that influence the model despite the weighting scheme. This suggests that while the weighted model performs better in terms of capturing non-linearity, further refinement might still be beneficial.

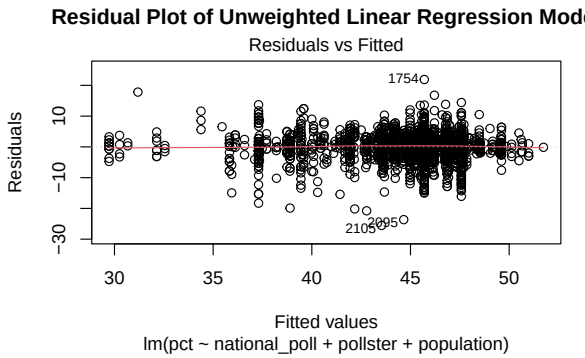
Figure **??** is a Q-Q Plot of weighted model plot for Harris support.The residuals generally align with the theoretical normal distribution line, particularly in the central range, indicating that the residuals of the weighted model are close to normal. Similar to the unweighted model, there are deviations at the tails, though they appear less pronounced. This suggests that the weighting scheme has slightly improved the distribution of residuals, making the model's predictions more robust. However, some extreme values remain, which could still affect model

In summary, both models show a reasonably good fit, with the weighted model offering slight improvements in handling non-linearity and extreme values. However, both models exhibit minor deviations from normality and a few notable outliers, which may warrant further model adjustments for improved prediction accuracy.

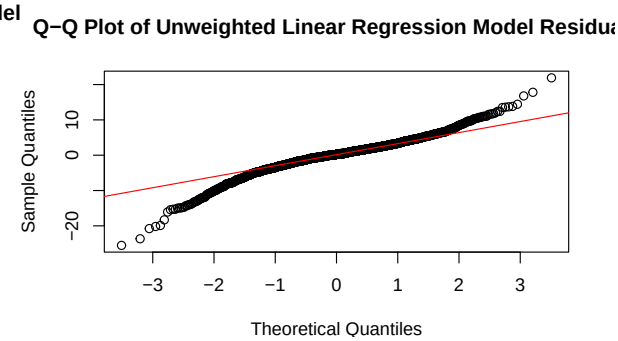
Figure **??** shows the residuals plotted against the fitted values for the unweighted model. It shows that the residuals are generally spread around zero with no clear pattern, suggesting that the model is relatively well-specified. However, there is a slight curvature, indicating possible non-linearity that the model may not fully capture. Some notable outliers with larger residuals suggest that certain data points have significant prediction errors, potentially influencing the model.

Figure **??** shows the Q-Q plot for the unweighted model, showing that most residuals fall along the line, especially in the middle range, suggesting that the model satisfies the normality assumption. However, some points at the tails deviate, indicating potential outliers or non-normality in extreme residual values. This deviation at the ends suggests the model may face issues with extreme predictions, though it performs reasonably well overall.

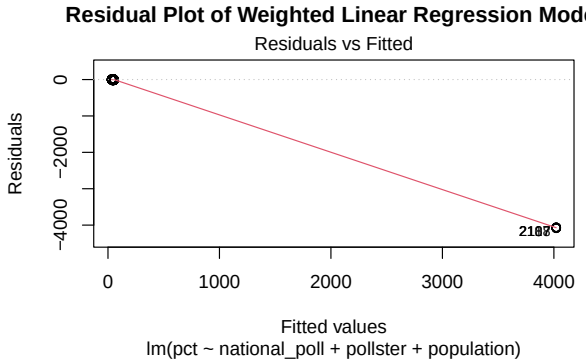
Figure **??** shows residuals plotted against fitted values. Similar to the unweighted model, the residuals are centered around zero, indicating that the weighted model captures the overall trend without significant systematic bias. The slight curvature seen in the unweighted model is reduced here, suggesting that weighting has addressed some of the non-linearity. However, some large residuals remain, which could indicate outliers that affect the model even with



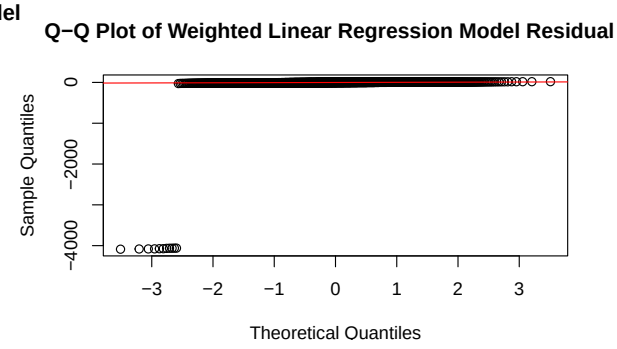
(a) Residual Plot for Unweighted Model



(b) Q-Q Plot for Unweighted Model



(c) Residual Plot for Weighted Model



(d) Q-Q Plot for Weighted Model

Figure 11: Diagnostics of model using residual vs fitted plot and norm Q-Q plot -Support for Trump

the weighting scheme. This suggests that while the weighted model has improved in handling non-linearity, further refinements could enhance accuracy.

Figure **??** shows the Q-Q plot for the weighted model compares residuals with a theoretical normal distribution. Here, the residuals generally align with the theoretical line, especially in the central range, indicating that the residuals of the weighted model are close to normal. Similar to the unweighted model, some deviations occur at the tails, though they appear less pronounced, suggesting that the weighting scheme has slightly improved the normality of residuals. Nonetheless, some extreme values persist, which may impact model robustness in cases of outliers.

Summary Both models exhibit a reasonably good fit, with the weighted model offering slight improvements in managing non-linearity and extreme values. Despite this, both models show minor deviations from normality and some notable outliers, suggesting that further model adjustments may be beneficial for improved prediction accuracy.

C The New York Times/Siena College Polling Methodology

This appendix provides a comprehensive overview of the methodology employed by the Siena College Polling Institute in conducting its surveys.Siena College is renowned for its methodologically rigorous approach to political polling, focusing on accurately capturing voter sentiment during elections. Siena College has conducted polls in three key states: Michigan, Wisconsin, and Ohio.

In this section, we will delve into the key components of Siena's polling methodology, including the target population, sampling frame, recruitment processes, and the sampling strategies used. We will also address how non-response is managed and evaluate the strengths and weaknesses of the questionnaire design. By exploring these elements, this appendix aims to clarify how Siena College ensures the reliability and validity of its polling results, contributing valuable insights to the understanding of voter behavior and election outcomes.

C.1 Pollster Overview

Siena College Polling Institute is a prominent pollster known for its comprehensive and methodologically rigorous surveys. It specializes in political polling and is particularly recognized for its work in understanding voter sentiment during elections. Established in 1980 at Siena College in New York's Capital District, the institute carries out both expert and public opinion polls. (Siena College Research Institute 2024).

C.2 Population, Frame and sample

Refer from *Telling Stories with Data, Author = Rohan Alexander, Year = 2023, Publisher = Chapman; Hall/CRC, Url = https://tellingstorieswithdata.com/* (n.d.), we defined three key terms as:
Target population : The collection of all items about which we would like to speak/ the entire group about which we want to draw. conclusions
Sampling frame : A list of all the items from the target population that we could get data about.
Sample : The items from the sampling frame that we get data about.

The target population for Siena's polls includes registered voters eligible to vote in Michigan, Wisconsin, and Ohio.

The sampling frame is a comprehensive list of registered voters, which includes demographic information for each voter. This enables the pollsters to ensure an appropriate representation of voters across various parties, races, and regions (The New York Times 2024b).

The sample of registered voters sourced from the voter file maintained by L2, a nonpartisan vendor, and supplemented with additional cellular phone numbers matched from Marketing Systems Group. The sample for the poll totals 2,055 likely voters, with 688 from Michigan, 687 from Ohio, and 680 from Wisconsin, surveyed from September 21 to 26, 2024.

C.3 Sample Recruitment

Siena use phone poll to recruit sample. Telephone polling is a way to gather public opinion by contacting individuals via landlines and mobile phones, using live interviewers to improve data quality and capture nuanced responses. Through random digit dialing or voter registration databases, researchers achieve a representative sample across demographics.

According to The New York Times (2024a), the polls are conducted in both English and Spanish by live interviewers at call centers located in Florida, New York, South Carolina, Texas, and Virginia. The respondents are randomly selected from a national database of registered voters and are contacted via both landlines and cellphones.

C.4 Sampling Approach

Siena employs a response-rate-adjusted stratified sampling of registered voters sourced from the voter file maintained by L2, a nonpartisan vendor, and supplemented with additional cellular phone numbers matched from Marketing Systems Group. The New York Times selected the sample in multiple stages to address differences in telephone coverage, nonresponses, and notable variations in telephone number productivity by state.

Stratified sampling is typically utilized to ensure all strata of the population are represented. When considering our population, it typically consists of various groupings. These can range from a country being divided into states, provinces, counties, or statistical districts to a university comprising faculties and departments or even demographic characteristics groups among individuals. A stratified structure allows us to categorize the population into mutually exclusive and collectively exhaustive sub-populations known as "strata"(*Telling Stories with Data, Author = Rohan Alexander, Year = 2023, Publisher = Chapman; Hall/CRC, Url = https://tellingstorieswithdata.com/*, n.d.).

In this scenario, we want to collect the polls from all strata of our target population to balance our poll result. The sample was stratified by political party, race, and region, and screened by M.S.G. to ensure that the cellular phone numbers were active.

C.4.1 Strength and Weakness

Stratified sampling enhances sample representativeness by ensuring that smaller subgroups are adequately included, allowing researchers to allocate resources more efficiently and gain deeper insights into specific groups. However, this method can lead to **higher costs** due to the extensive data collection and analysis needed, especially when sampling large regions. Stratified sampling also introduces **complexity in data analysis**, requiring advanced techniques to accurately interpret subgroup data and appropriate weighting for each stratum. Additionally, poorly defined strata or imbalanced sampling can lead to sampling bias. While stratified sampling provides strong representation and analytical depth, it also brings challenges related to cost, complexity, and potential bias if not executed with care.

C.5 Non-response Bias

An interview was deemed complete for inclusion in the voting preference questions if the respondent stayed engaged in the survey after answering the two self-reported variables used for weighting—age and education—and provided responses to at least one question concerning age, education, or the presidential election candidate reference. If these conditions were not met, the interview was recorded as a non-response.

To handle the non-response bias, Siena choose to use weighting adjustments. Weighting is like balancing a scale to make sure each group in the survey counts the right amount. It changes the importance of each answer depending on how likely people are to skip the survey (Kinga Edwards 2024).

Siena use several steps to address nonresponse bias and ensure the reliability of the results. The weighting process was conducted by The New York Times using the R survey package and involved multiple adjustments. Siena's weighting process involved adjusting samples for unequal selection probabilities and turnout likelihood, based on 2020 data. Further adjustments aligned the sample with likely electorate targets from the L2 voter file. The final weight

combined modeled turnout (80%) and self-reported intentions (20%), mitigating nonresponse bias and ensuring the sample accurately reflected the characteristics and behaviors of likely voters, thereby enhancing result validity.

C.6 Questionnaire Design

C.6.1 Response bias defination

In the design of the questionnaire, there will be some common bias that may occur when running the questionnaire.

Stantcheva, Stefanie (2023) define these bias as:

- Moderacy response bias is the tendency to respond to each question by choosing a category in the middle of the scale.
- Extreme response bias is the tendency to respond with extreme values on the rating scale.
- Response order bias occurs when the order of response options in a list or a rating scale influences the response chosen. The primacy effect occurs when respondents are more likely to select one of the first alternatives provided, and it is more common in written surveys. This tendency can be due to satisficing, whereby a respondent uses the first acceptable response alternative without paying particular attention to the other options. The recency effect occurs when respondents choose one of the last items presented to them (more common in face-to-face or orally presented surveys).
- Social desirability bias typically stems from the desire of respondents to avoid embarrassment and project a favorable image to others, resulting in respondents not revealing their actual attitudes. The prevalence of this bias will depend on the topic, questions, respondent, mode of the survey, and the social context. For instance, in some circles, anti-immigrant views are not tolerated, and those who hold them may try to hide them. In other settings, people express such views more freely.
- Acquiescence is the tendency to answer items in a positive way regardless of their content, for instance, systematically selecting categories such as “agree,” “true,” or “yes”.

C.6.2 Strengths and Weakness

Strengths:

The questionnaire is concise and straightforward, reducing respondent fatigue and enhancing clarity, which is crucial for maintaining engagement. By incorporating both closed- and open-ended questions, it allows for both quantifiable data and rich qualitative insights. Clear

response categories help reduce moderacy bias, encouraging participants to choose decisively rather than defaulting to neutral answers. Additionally, varied question types help mitigate acquiescence bias by encouraging honest responses and avoiding leading language.

Weaknesses:

However, the questionnaire has some limitations. Its reliance on agree-disagree and yes-no formats may increase acquiescence bias, as respondents may lean toward favorable answers. Furthermore, some demographic nuances may be inadequately addressed, potentially leading to nonresponse bias from underrepresented groups. The risk of response order bias is also present, especially if randomization of options is not implemented, increasing the chance of recency effects in verbally-administered surveys.

Additionally, the absence of assured anonymity could lead to social desirability bias, where respondents alter answers to project a favorable image. Lastly, with over 50 questions, the length of the survey may increase dropout rates, especially in time-intensive formats like telephone surveys, thereby raising nonresponse bias.

In summary, while the questionnaire is clear and well-structured, it faces challenges from potential biases including acquiescence, nonresponse, social desirability, and order effects. Future improvements should focus on diversifying question types, ensuring demographic inclusivity, and refining question phrasing to reduce bias and enhance validity.

D Idealized Methodology for US Presidential Election Forecast

This appendix details the methodology and design for conducting a U.S. presidential election forecast survey with a budget of $100,000. The objective is to generate an accurate and reliable prediction of the election outcome while ensuring data quality through meticulous sampling, recruitment, validation, and aggregation of results.

D.1 Sampling Approach

To ensure a representative sample of likely voters, I will employ a Composite Measure sampling method based on past ballots cast data from the 2020 U.S. elections. After determining the sample size for each state, I will use stratified sampling based on demographics, dividing the population into subgroups and taking random samples from each subgroup. This Composite Measure sampling approach, as referenced in Clark Letterman (2021), enhances our chances of selecting respondents from states or regions that have historically exhibited higher voter engagement compared to the general population distribution. While some states may have larger populations, we aim to adjust the sampling to reflect higher turnout rates.