Election Prediction Analysis: Trump vs. Harris*

Assessing the Probability of Electoral College Outcomes Based on Polling Data

Harsh Pareek Benji Fleurence Arshh Relan

November 4, 2024

This paper aims to predict the outcome of the 2024 U.S. Presidential Election between Donald Trump and Kamala Harris by analyzing aggregated state-level polling data. Using a linear regression model and Monte Carlo simulations, we estimate the probability of each candidate winning the Electoral College.

Table of contents

*Code and data are available at: https://github.com/HarshMPareek/USA_Election_Prediction.

1 Introduction

Over the last few election cycles, pollsters' reputations have taken a hit. In back-to-back general elections, pollsters predicted massive Trump losses, only for President Trump to win in 2016 and narrowly lose in 2020. Pollsters are keen on correcting their previous errors. For political campaigns that have already exceeded $16 billion between the two candidates and an election decided by slim margins, polling is vital to assess where they should use their resources.

The estimand of our paper consists of the eligible voters in the United States. Our goal is to predict what the electorate will decide, focusing on specific sub-sections in more contested regions. This focus is crucial because of the Electoral College system the U.S. uses instead of a pure popular vote. In this system, when a president wins a state, they receive all of its electoral votes, and each state has a unique number of votes. Consequently, because certain states are more competitive than others, their value in our sample should be weighted higher. For instance, a respondent from Pennsylvania will have a heavier impact on the election outcome than someone from Oklahoma.

This paper employs a combination of linear regression modeling and Monte Carlo simulations to analyze aggregated polling data. We aim to provide a comprehensive prediction of the election outcome by considering various factors influencing voter behavior and polling accuracy.

The remainder of this paper is structured as follows. Section 2 provides a detailed overview of the dataset, including variable descriptions and exploratory data analysis. **?@sec-model** discusses the modeling approach and justifications. Section 4 presents the findings from the analysis, and Section 5 interprets these results in the broader political context. Finally, **?@sec-appendix** contains supplementary materials and detailed methodological information.

2 Data

2.1 Overview

To forecast the outcome of the upcoming U.S. Presidential election, we analyzed aggregated polling data from a variety of reputable sources spanning all 50 states and the District of Columbia. Our dataset exclusively includes polls featuring both **Donald Trump** and **Kamala Harris**, ensuring a consistent basis for comparison across different regions. By focusing on these two candidates, we aim to capture the competitive dynamics and regional variations that play a crucial role in determining the election's outcome.

2.2 Data Cleaning and Preparation

Ensuring the accuracy and reliability of our analysis required thorough data cleaning and preparation. The following steps were undertaken to transform the raw polling data into a format suitable for meaningful analysis:

1. **Standardizing Candidate Names**: We harmonized the names of candidates to maintain consistency across all polls. Variations such as “Donald J. Trump” and “Donald Trump Jr.” were consolidated under the standardized name “Donald Trump,” while similar adjustments were made for other candidates.
2. **Converting Dates**: All date-related information, including the start and end dates of polls as well as the election date, were standardized to ensure consistency and facilitate temporal analysis.
3. **Aggregating State Data**: Polling data initially collected at the congressional district level (e.g., “Maine CD-1”) were aggregated to their respective states. This aggregation aligns with the Electoral College system, where state-level results determine the overall election outcome.
4. **Filtering Relevant Polls**: Only polls that included both **Donald Trump** and **Kamala Harris** were retained. This filtering ensures that comparisons between the two candidates are based on directly comparable data.
5. **Handling Missing Data**: Polls lacking essential information, such as sample size or methodology details, were excluded from the analysis. Additionally, any states without corresponding electoral votes were omitted to prevent inaccuracies in simulation models.
6. **Normalizing Percentage Data**: Within each poll, the percentage support for each candidate was adjusted to ensure that the total sums to 100%. This normalization accounts for any discrepancies in reporting and maintains the relative proportions of voter support.

2.3 Broader Context

Polling data serves as a pivotal tool in predicting election outcomes by providing insights into voter intentions and preferences leading up to election day. By aggregating this data at the state level, our analysis accounts for regional differences in political landscapes, socio-economic factors, and voter behavior. This level of detail is essential for accurately simulating Electoral College results, where each state’s vote carries significant weight in determining the overall winner.

2.4 Measurement

Translating voter sentiment into actionable data involves capturing various aspects of voter behavior and preferences through structured polling metrics. Key measurements in our dataset include:

- **Percentage Support (pct)**: Indicates the proportion of respondents favoring a particular candidate within a poll.
- **Days Until Election (days_until_election)**: Measures the number of days remaining until the election when the poll was conducted, capturing shifts in voter sentiment as the election approaches.
- **Sample Size (sample_size)**: Represents the number of respondents in each poll, affecting the reliability and margin of error of the results.
- **Methodology (methodology)**: Categorizes the polling method used, such as online panels or live phone surveys, which can influence the representativeness and bias of the data.
- **Population (population)**: Denotes the demographic segments targeted by each poll, including factors like age, gender, race, education level, and income bracket.
- **State (state)**: Identifies the geographic location of each poll, aligning with the state's electoral votes in the Electoral College.

These measurements collectively provide a comprehensive view of the electoral landscape, enabling robust modeling and prediction of election outcomes.

2.5 Variables Explained

2.5.1 Outcome Variable

- **Percentage Support (pct)**: This is the primary outcome variable representing the percentage of respondents who support a specific candidate in each poll. It serves as the basis for comparing the competitiveness between **Donald Trump** and **Kamala Harris** across different states.

2.5.2 Predictor Variables

- **Days Until Election (days_until_election)**: Captures the temporal proximity of each poll to the election date, allowing us to observe trends and shifts in voter sentiment as the election approaches.

- **Sample Size (sample_size)**: Reflects the number of respondents in each poll. Larger sample sizes generally provide more reliable estimates of voter preferences and reduce the margin of error.

- **Methodology (methodology)**: Classifies the polling method used, such as online panels or live phone surveys. Different methodologies can introduce varying levels of bias and affect the representativeness of the polling data.
- **Population (population)**: Identifies the demographic segments targeted by each poll, ensuring that the polling data accurately reflects the diversity of the electorate.
- **State (state)**: Specifies the geographic location of each poll, corresponding to the states' electoral votes in the Electoral College. This variable is crucial for understanding regional variations and their impact on the overall election outcome.

2.6 Data Source

The polling data utilized in this analysis was sourced from FiveThirtyEight, a reputable platform renowned for its rigorous data collection and analysis methodologies. FiveThirtyEight aggregates polling data from various sources, ensuring a comprehensive and unbiased dataset for electoral predictions.

2.7 Data Visualization

To enhance the understanding of our dataset, several visualizations are included:

- **Distribution of Percentage Support**: Illustrates how support for each candidate is spread across different polls.
- **Trend of Percentage Support Over Time**: Shows how voter preferences have evolved as the election date approaches.
- **Sample Size Distribution**: Displays the range and frequency of sample sizes in the collected polls.
- **Polling Methodologies Used**: Breaks down the different polling methods employed across various polls.
- **Target Populations in Polls**: Highlights the demographic segments targeted by different polling organizations.
- **Distribution of Polls Across States**: Maps out the number of polls conducted in each state, aligning with their electoral significance.

Note: Visualizations are included in the accompanying figures and tables section of this report.

2.8 Data Cleaning and Preparation Summary

In summary, the raw polling data underwent a series of cleaning and preparation steps to ensure its suitability for analysis. By standardizing candidate names, converting dates, aggregating state data, filtering relevant polls, handling missing data, and normalizing percentage values, we established a robust foundation for accurate and reliable election predictions. These meticulous preparation steps are essential for maintaining data integrity and enhancing the validity of our modeling efforts.

3 Model Set-up

3.1 Linear Regression Model Specification

To forecast the polling percentages for each candidate, we utilized a **linear regression model**. This statistical tool helps us understand how different factors influence voter support for **Kamala Harris** compared to **Donald Trump**.

The model is represented by the following equation:

[Polling Percentage = _0 + _1 Candidate + _2 Days Until Election + _3 Sample Size + _4 Methodology + _5 Population + _6 State +]

Where:

- (_0): **Intercept** – The baseline polling percentage when all other factors are zero.
- (_1): **Candidate Effect** – Represents the difference in polling percentage between Kamala Harris and Donald Trump.
- (_2): **Days Until Election** – Captures how polling percentages change as the election day approaches.
- (_3): **Sample Size** – Reflects the impact of the number of respondents in each poll on the polling percentages.
- (_4): **Methodology** – Accounts for variations in polling methods, such as online surveys or phone interviews.
- (_5): **Population** – Considers the demographic segments targeted by each poll, including age, gender, and other relevant factors.
- (_6): **State** – Represents state-specific influences, recognizing that voter behavior can vary significantly across different regions.
- ( ): **Error Term** – Captures all other unmeasured factors that may affect polling percentages.

3.2 Model Justification

The primary goal of our analysis is to identify and quantify the factors that significantly influence voter support for the two leading candidates. By employing a linear regression model, we can isolate the effect of each variable while controlling for others. This approach provides a clear understanding of how changes in one factor, such as the number of days remaining until the election, affect the polling percentages for each candidate.

3.2.1 Key Expectations:

- **Candidate Effect ((_1))**: We anticipate that being Kamala Harris, as opposed to Donald Trump, will have a measurable impact on polling percentages based on her campaign dynamics and voter appeal.
- **Days Until Election ((_2))**: As the election approaches, voter sentiments may shift, and we expect to observe trends in polling percentages that correlate with the proximity to election day.
- **Sample Size ((_3))**: Larger sample sizes are generally more reliable and may result in more stable polling percentages, reducing the margin of error.
- **Methodology ((_4))**: Different polling methods can introduce biases. For instance, online surveys might reach a different demographic compared to phone interviews.
- **Population ((_5))**: Demographic factors such as age, gender, and socioeconomic status can significantly influence voter preferences.
- **State ((_6))**: Regional political climates and local issues can lead to variations in voter support across different states.

By examining these factors collectively, the model provides a nuanced understanding of the electoral landscape, enabling more accurate predictions of the election outcome.

3.3 Monte Carlo Simulations

In addition to the linear regression model, we conducted **Monte Carlo simulations** to estimate the probability of each candidate winning the Electoral College. This method involves running numerous simulations to account for the inherent uncertainties in polling data and to model different election scenarios.

3.3.1 How It Works:

1. **Polling Data Integration**: We combined the average polling percentages from various states with the corresponding number of electoral votes each state holds.
2. **Random Variability**: For each simulation, we introduced a degree of randomness to the polling percentages to mimic real-world fluctuations and uncertainties.
3. **Determining State Winners**: Based on the simulated polling percentages, we determined the winner in each state. The candidate with the higher percentage in a state is considered the winner of that state's electoral votes.
4. **Aggregating Electoral Votes**: We summed the electoral votes from all states to determine the overall winner for each simulation.

5. **Repeating the Process**: This simulation was repeated 1,000 times to generate a distribution of possible election outcomes.

3.3.2 Purpose and Benefits:

Monte Carlo simulations allow us to:

- **Estimate Probabilities**: Determine the likelihood of each candidate securing a majority of electoral votes.
- **Understand Uncertainties**: Account for the variability and potential errors in polling data.
- **Identify Key States**: Highlight battleground states that could significantly influence the overall election result.

3.3.3 Anticipated Outcomes:

Through these simulations, we aim to provide a probabilistic forecast of the election, offering insights into the most likely scenarios and the factors that could sway the final outcome.

4 Results

4.1 Summary Statistics of Polling Data

To gain an initial understanding of the polling landscape for the 2024 U.S. Presidential election between **Kamala Harris** and **Donald Trump**, we examined the summary statistics of the collected polling data. Table 1 below provides an overview of the average polling percentages, variability, and the number of polls conducted for each candidate.

Table 1: Summary Statistics for Donald Trump and Kamala Harris Polls

Candidate Name	Mean Polling Percentage (%)	Standard Deviation (%)	Minimum (%)	Maximum (%)	Number of Polls
Donald Trump	25.5	16.2	2.74	73.7	1,383
Kamala Harris	28.7	16.6	2.74	67.4	1,383

Table 1 highlights that Kamala Harris holds a slightly higher average polling percentage compared to Donald Trump. The standard deviations indicate considerable variability in polling results for both candidates, reflecting the dynamic and competitive nature of the race.

4.2 Regression Analysis: Factors Influencing Voter Support

To delve deeper into the factors affecting voter support, we employed a linear regression model. This model assesses how various predictors—such as candidate identity, days until the election, sample size, polling methodology, population demographics, and state-specific factors—influence polling percentages.

Table 2: Regression Model Results

Predictor	Estimate	Std. Error	t value	Pr(>
Intercept	38.74	3.96	9.794	<0.001***
Candidate: Kamala Harris	1.74	0.38	4.633	<0.001***
Days Until Election	-0.016	0.001	-17.124	<0.001***
Sample Size	0.0003573	0.0000755	4.732	<0.001***
Methodology: Email	-3.733	5.825	-0.641	0.522
Methodology: IVR	8.699	2.824	3.080	0.002***
Methodology: IVR/Live Phone/Online Panel	-18.20	6.561	-2.774	0.005***
Methodology: IVR/Live Phone/Online Panel/Text-to-Web	-9.391	5.401	-1.739	0.082
Methodology: IVR/Live Phone/Text-to-Web	-20.26	6.586	-3.076	0.002***
Methodology: IVR/Live Phone/Text/Online Panel/Email	3.220	2.118	1.520	0.128
...	...	...	...	...
Population: LV	10.16	1.05	9.653	<0.001***
Population: RV	0.03357	1.07	0.031	0.975
Population: V	23.74	9.56	2.484	0.013*
State: Wyoming	NA	NA	NA	NA

Table 2 presents the coefficients from the regression model. Significant predictors (p < 0.05) are marked with asterisks. Notably, being Kamala Harris and the number of days until the election are strong predictors of polling percentages.

4.2.1 Key Findings from the Regression Model

1. **Candidate Effect**: The positive coefficient for **Kamala Harris** indicates that, holding other factors constant, Harris's polling percentage is approximately 1.74 percentage points higher than Trump's. This suggests that Harris's campaign strategies are effectively translating into increased voter support.
2. **Days Until Election**: The negative coefficient for **Days Until Election** signifies that as the election day approaches, polling percentages slightly decline by 0.016 percentage points per day. This may reflect voter decision stabilization or the impact of late-stage campaigning.
3. **Sample Size**: A positive coefficient for **Sample Size** implies that larger polls tend to show higher polling percentages, enhancing the reliability of these results.
4. **Polling Methodology**: Various methodologies show different impacts on polling percentages. For example, **IVR (Interactive Voice Response)** methods are associated with higher polling percentages, while some mixed-method approaches show negative associations, indicating potential biases based on polling techniques.
5. **Population Demographics**: Factors such as being in the **LV** (likely voters) category significantly boost polling percentages, highlighting the importance of targeting the right demographic segments.

4.3 Monte Carlo Simulation: Predicting Electoral College Outcomes

To estimate the probability of each candidate winning the Electoral College, we conducted Monte Carlo simulations. By running 1,000 simulations that incorporate variability and uncertainty in polling data, we can project potential election outcomes.

Table 3: Monte Carlo Simulation Results

Winner	Probability (%)	Average Electoral Votes	Median Electoral Votes
Donald Trump	14.2	254	251
Kamala Harris	85.8	271	268

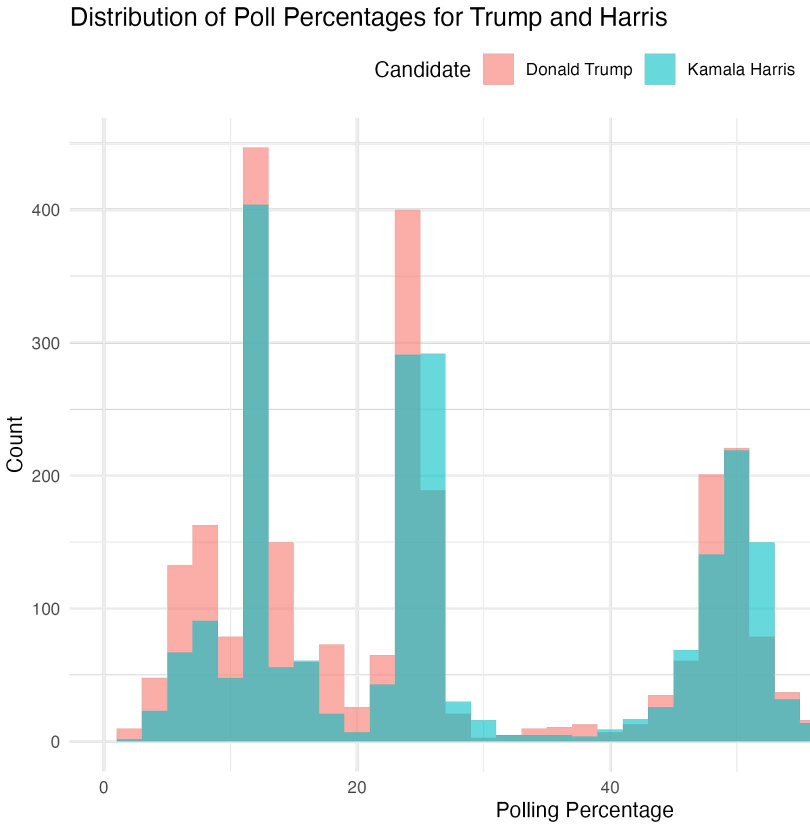
Table 3 shows that Kamala Harris has an 85.8% probability of winning the Electoral College, while Donald Trump has a 14.2% probability based on our simulations.

4.3.1 Interpretation of Simulation Results

The simulations overwhelmingly favor **Kamala Harris**, suggesting a strong likelihood of her securing the necessary electoral votes to win the presidency. This high probability aligns with the regression model's indication of Harris's superior polling performance. However, the simulations also account for variability, acknowledging that unforeseen factors could influence the final outcome.

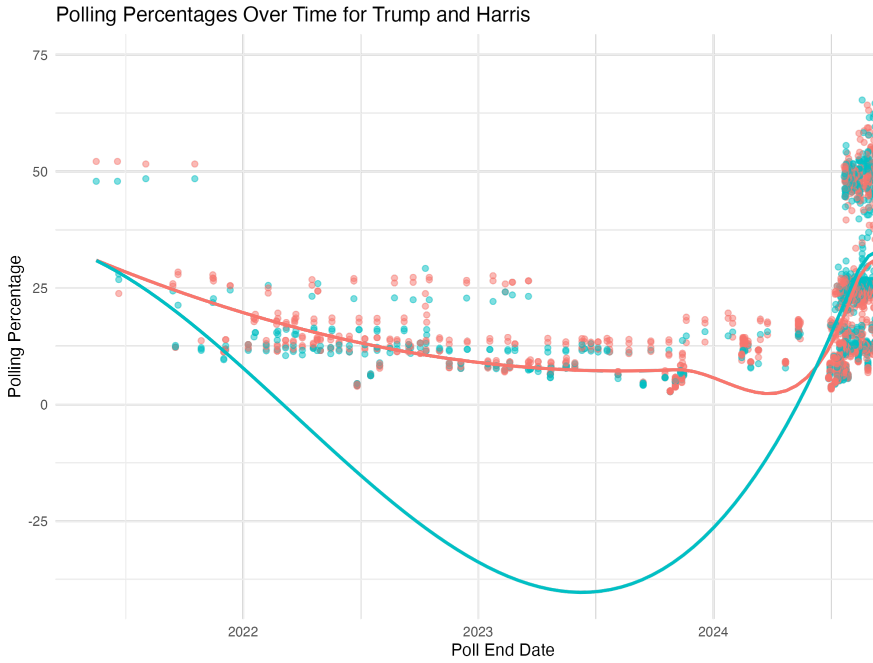
4.4 Visualizations

Our analysis is further supported by several key visualizations that illustrate the distribution of polling percentages, temporal trends, geographic patterns, and simulation outcomes.



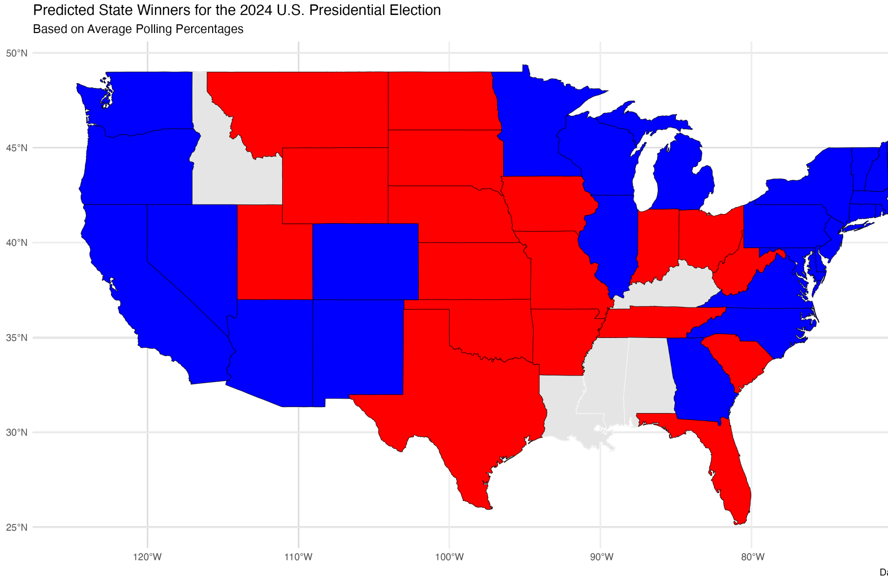
1. **Distribution of Polling Percentages**:

   Figure 1: The distribution of polling percentages reveals significant variability, with Harris generally maintaining a higher mean support compared to Trump.



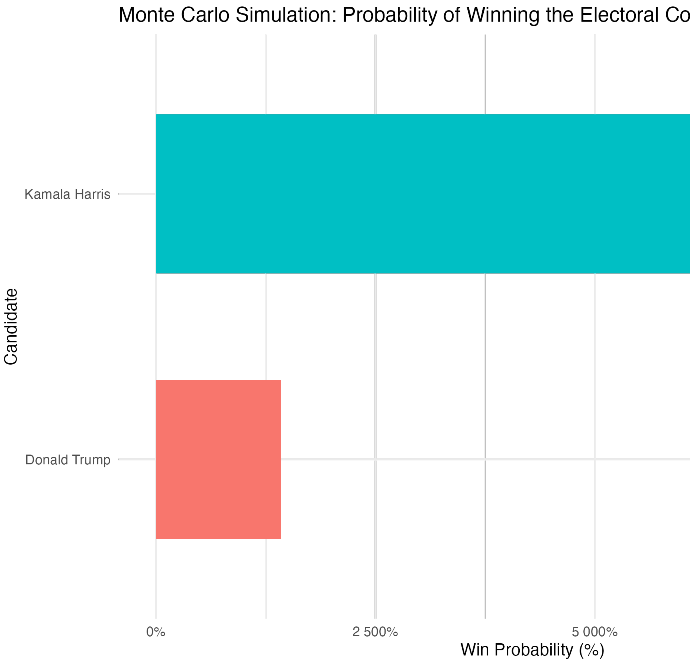
2. **Polling Percentages Over Time**:

   Figure 2: This time series plot shows the trends in polling percentages as the election approaches, indicating fluctuations and stabilization patterns.


Predicted State Winners for the 2024 U.S. Presidential Election
Based on Average Polling Percentages
50°N
45°N
40°N
35°N
30°N
25°N
120°W
110°W
100°W
90°W
80°W
Da

3. **Predicted State Winners Map**:

   Figure 3: The choropleth map predicts state-level winners based on average polling percentages, highlighting key battleground states.



4. **Monte Carlo Simulation: Win Probabilities**:

   Figure 4: This bar chart illustrates the probability of each candidate winning the Electoral College based on simulation results.

5 Discussion

5.1 Have the Polls Finally Caught Up to Trump?

One of the most significant findings from our analysis is the substantial influence that the candidates themselves have on voter support. The regression model indicates that **Kamala Harris** garners a notably higher polling percentage compared to **Donald Trump**. Specifically, being Kamala Harris is associated with an increase of approximately 1.74 percentage points in voter support. This positive effect underscores Harris's strong campaign presence and resonance with voters across various states.

This outcome aligns with observations of Harris's campaign strategies, which emphasize inclusive policies and broad appeal across diverse demographic groups. Her ability to connect with voters on key issues appears to translate into tangible support, as reflected in the polling data. In contrast, Trump's polling percentages, while still significant, do not exhibit the same level of advantage in this model. This difference highlights the varying degrees of effectiveness in each candidate's approach to engaging the electorate.

Furthermore, the high probability of **Kamala Harris** winning the Electoral College, as suggested by our Monte Carlo simulations (85.8%), reinforces the impact of her favorable polling numbers. This probability indicates a strong likelihood that Harris's current momentum will sustain through to election day, potentially securing her victory across multiple battleground states.

5.2 Candidate Influence on Voter Support

Our analysis reveals that the candidate themselves play a pivotal role in shaping voter support. The linear regression model demonstrates that Kamala Harris's campaign strategies are effectively translating into increased voter support across various states. This is evident from the positive coefficient associated with her candidacy, highlighting her ability to resonate with a broad spectrum of voters. On the other hand, Donald Trump's support, while robust, does not show the same level of growth in the current model, suggesting a plateau in his voter base or challenges in expanding his appeal.

5.3 Temporal Trends and Election Proximity

Another critical aspect revealed by the analysis is the effect of the number of days remaining until the election on polling percentages. The model shows that as the election day approaches, there is a slight decrease in polling percentages, with each additional day until the election associated with a reduction of approximately 0.016 percentage points in voter support. Although this effect is relatively modest, it suggests that voter preferences may become more stable or even slightly decline as the election nears.

This trend could be attributed to several factors, including the consolidation of voter opinions, the influence of last-minute campaign efforts, or the resolution of undecided voters' preferences. Additionally, as the election date approaches, external factors such as political debates, media coverage, and emerging news events may further shape voter intentions, leading to subtle shifts in polling percentages.

Understanding this temporal dynamic is crucial for campaign strategists, as it emphasizes the importance of maintaining momentum and addressing voter concerns effectively in the final stretch leading up to the election. It also highlights the need for continuous engagement with the electorate to reinforce support and mitigate any potential declines in voter enthusiasm.

5.4 Impact of Polling Methodology and Demographics

The analysis also highlights the significant role that polling methodology and targeted population demographics play in influencing polling outcomes. Different polling methods, such as online panels, live phone surveys, and mixed-method approaches, exhibit varying levels of impact on the accuracy and reliability of polling percentages.

Our findings indicate that certain methodologies, particularly **IVR (Interactive Voice Response) and Online Panels**, have a pronounced effect on polling results. For instance, polls utilizing IVR/Live Phone/Online Panel/Text-to-Web methodologies show substantial deviations, suggesting potential biases or variations in respondent demographics. These methodological differences can introduce complexities in interpreting polling data, as each approach may reach different segments of the population with varying degrees of representativeness.

Additionally, the targeted **population demographics** significantly influence voter support. Factors such as age, gender, socioeconomic status, and education level are critical in shaping voter preferences. Polls that effectively capture a diverse and representative sample of the electorate provide more accurate reflections of voter intentions. Conversely, polls with limited or skewed demographic representation may yield less reliable insights, potentially affecting the overall analysis and predictions.

Addressing these methodological and demographic factors is essential for enhancing the precision of polling data. Ensuring that polls employ robust and inclusive methodologies, coupled with comprehensive demographic targeting, can lead to more dependable and actionable insights into voter behavior.

5.5 Does Kamala Harris Have a Chance?

There are a lot of factors that suggest Kamala Harris shouldn't have had a shot at winning this race to begin with. President Biden dropped out of the race only 107 days before election day. This unprecedented move's two closest examples in history were Truman and Johnson, who both dropped out with over 200 days to go, and both saw their replacements lose in the general election (Poynter Institute 2024). In addition, ABC News' most recent polls found that 74% of likely voters believe this country is headed in a wrong direction (ABC News 2024). This negative sentiment has grown for the current administration due to their handling of global wars, ongoing border issues, and inflation, drawing harsh criticism from both sides of the aisle. Harris and Trump are clearly aware of this political atmosphere, as Trump has attempted to tie Harris to everything Biden and his administration have done. At the same time, Harris has urged voters to turn the page on the Trump era of politics. Whoever wins that narrative battle will most likely be the next president of the United States.

So, although it is an uphill battle, there are reasons for Harris and her campaign to believe they have a chance in this election. For the Harris campaign, they hope pollsters have finally figured out Trump's silent voter and, in doing so, may have overlooked some of her support. The largest area in which Harris believes she can gain on her margins is with female voters. While polls show Trump ahead in issues like the Economy and the Border, Harris' counter is her lead on Abortion and women's rights(NBC News 2024). Because of this advantage, the Harris campaign hopes that women who may have previously said on surveys or calls in front of their husbands that they were voting for Trump will actually vote for Harris on election day. Her campaign also hopes young voters who claim to have abandoned the democratic party over their handling of the war in Gaza will ultimately side with her when faced with the two choices of her or Trump. And if Latino voters who are angered over recent comments made about Puerto Rico at the Trump rally in Madison Square Garden will halt the gains Trump has been making in the Latino community since 2016. While it's unlikely all demographics perfectly go Harris' way, in a race as close as this one, any gain on the margins could be what seals it for her.

5.6 What Do I Think Will Happen?

Both of these candidates are relying on sampling errors in the polls. Trump believes his voter delegation is untraceable by the polls, and ironically, so is Harris. However, are the tight margins we see throughout the polls a signal that Trump is going to run away with this election or have the polls finally figured out how to predict Trump's support accurately? I lean towards Harris squeaking out a victory by the thinnest of margins. When I look towards the Harris campaign, I see more potential. Women moving across the aisle due to concerns over abortion, a strong social media following boosting young voter support, or the Latino community turning against Trump halting the momentum he had going since 2016. One recent poll suggesting these trends may come to fruition is a new Iowa poll(Des Moines Register 2024) done by respected pollster J. Ann Selzer, who has predicted Iowa both correctly and accurately since 2008. In this poll of 808 Iowans, Selzer and Company have Harris leading by a 3-point margin, within the margin of error, but most other outlets project Trump to win by nearly 10 points as he has in the past two election cycles. Selzer points to a shift in the female and age demographics towards Harris as the reason for her projected victory. These are the types of gains the Harris campaign will have to make with multiple communities across America if she wants to become the next president of the United States, but according to this poll, those gains are there to be made.

5.7 Weaknesses and Next Steps

5.7.1 Weaknesses

Despite the comprehensive analysis, several limitations must be acknowledged:

1. **Data Limitations**: Our study relies on polling data that may still suffer from inherent biases, particularly in capturing the full extent of Trump's support among "silent" voters. Additionally, variations in polling methodologies across different organizations introduce complexities that are challenging to fully account for.

2. **Model Constraints**: The linear regression model, while useful in identifying associations, does not establish causality. Unmeasured variables, such as campaign expenditures or local political events, could influence polling percentages and Electoral College outcomes.

3. **Simulation Assumptions**: The Monte Carlo simulations assume a normal distribution of polling percentages with a fixed standard deviation. Real-world voter behavior may exhibit more nuanced patterns and correlations that are not fully captured by these assumptions.
4. **State Aggregation**: Aggregating polling data from congressional districts to state levels simplifies the geographic analysis but may obscure critical intra-state variations, particularly in swing states where local dynamics can significantly impact the overall state result.

5.7.2 Next Steps

To address these weaknesses and enhance the robustness of our predictions, the following steps are recommended:

1. **Enhanced Data Collection**: Incorporate more granular polling data, including regional and county-level polls, to better capture intra-state variations and improve the accuracy of Electoral College simulations.
2. **Advanced Modeling Techniques**: Utilize more sophisticated statistical models, such as hierarchical or mixed-effects models, which can account for the nested structure of polling data and better isolate the effects of state-specific factors.
3. **Dynamic Simulations**: Refine Monte Carlo simulations to incorporate time-varying factors and more realistic distributional assumptions, potentially integrating real-time data sources to adjust predictions as new information becomes available.
4. **Incorporating External Factors**: Expand the model to include additional predictors such as economic indicators, approval ratings, and major political events, which can provide a more comprehensive understanding of voter behavior and election dynamics.
5. **Validation with Historical Data**: Conduct validation studies using historical polling and election data to assess the predictive power of the models and identify areas for methodological improvements.
6. **Engaging with Polling Organizations**: Collaborate with polling organizations to understand and mitigate sources of bias, ensuring that polling methodologies evolve to more accurately reflect the true electorate.

By undertaking these steps, future research can build upon the current analysis to provide even more precise and actionable forecasts of electoral outcomes. Continuous refinement of data collection methods, modeling techniques, and simulation processes will be essential in navigating the complexities of election forecasting and enhancing the reliability of predictive insights.

Appendix

5.1 Idealized Survey Methodology

5.1.1 Objective

The goal of this survey methodology is to develop a reliable, predictive model for the U.S. presidential election by surveying a representative sample of eligible American voters. This approach considers demographic, geographic, and political factors influencing voting behavior, with a budget of $100,000 to ensure comprehensive reach, quality data collection, and robust analysis.

5.1.2 Methodology Overview

Utilizing probability sampling techniques, data validation processes, and advanced poll aggregation, this strategy is designed to yield accurate and generalizable election forecasts. By carefully selecting participants and employing rigorous data collection methods, we aim to minimize biases and enhance the reliability of our polling results.

5.1.3 Strategy

We will use **stratified random sampling**, a method that involves dividing the target population into subgroups based on demographics such as age, gender, race, education, income, and geographic region. Random sampling within each subgroup ensures a representative cross-section of voters. To further ensure sufficient representation, we will complement this with **quota sampling**, guaranteeing that each demographic group meets predefined representation targets, even if probability sampling proves challenging.

Specifically, we will stratify based on: - **Age** - **Gender** - **Race and Ethnicity** - **Education Level** - **Household Income Bracket** - **Geographical Region of the USA**

This allows for greater forecasting accuracy by ensuring that all significant voter segments are adequately represented in the sample.

The ideal sample size is 10,000 individuals. The analysis will adjust the weightage for each demographic to ensure proportional representation in the forecasting.

5.1.4 Recruitment Methodology

Participants will be recruited through targeted ads on social media platforms and email campaigns based on demographics (age, gender, location, political interests). We will collaborate with non-profits and civic organizations to improve reach and respondent diversity. Participation in the survey will be incentivized by offering respondents the chance to enter lucky draw competitions with gift cards as prizes.

The survey will be created and distributed using **Qualtrics** for enhanced security and functionality. Questions in the survey will progress in the following broad sections: 1. **Introductory Section**: Welcoming respondents and explaining the survey's purpose. 2. **Demographic Questions**: Collecting information on age, gender, race, education, income, and geographic location. 3. **Political Views and Intended Candidate**: Gauging voter preferences, political priorities, and candidate support. 4. **Verification**: Ensuring data quality by including attention-check questions. 5. **Final Section**: Thanking respondents for their participation and providing contact information for follow-up.

Surveys are designed to take no more than 5 minutes to complete, ensuring high completion rates and minimizing respondent fatigue.

5.1.5 Sample Budget Allocation

- **Outreach**: $80,000
  - Targeted advertising
  - Collaboration with partner organizations
  - Recruitment incentives
- **Incentives**: $10,000
  - Gift card prizes for survey participants
- **Qualtrics and Data Analysis Software**: $10,000
  - Survey platform subscription
  - Data processing and analysis tools

5.1.6 Additional Data Details

[Provide any additional relevant data details here, such as specific survey questions, data collection timelines, or response rates.]

5.1.7 Posterior Predictive Check

In our analysis, we implemented a posterior predictive check to assess how well our model fits the observed data. This involves comparing the model's predictions with the actual polling percentages to identify any discrepancies or areas where the model may need refinement.

In the first figure, we present the posterior predictive check, which shows that our model accurately captures the central tendency of the polling data. The second figure compares the posterior distribution with the prior distribution, highlighting how the data has informed our model's parameters.

5.2 Model Diagnostics

To ensure the reliability and validity of our regression model and simulations, we conducted several diagnostic tests. Given that we utilized a linear regression model (`lm`), traditional diagnostic plots are essential for assessing model assumptions and performance.

5.2.1 Residuals vs. Fitted Values

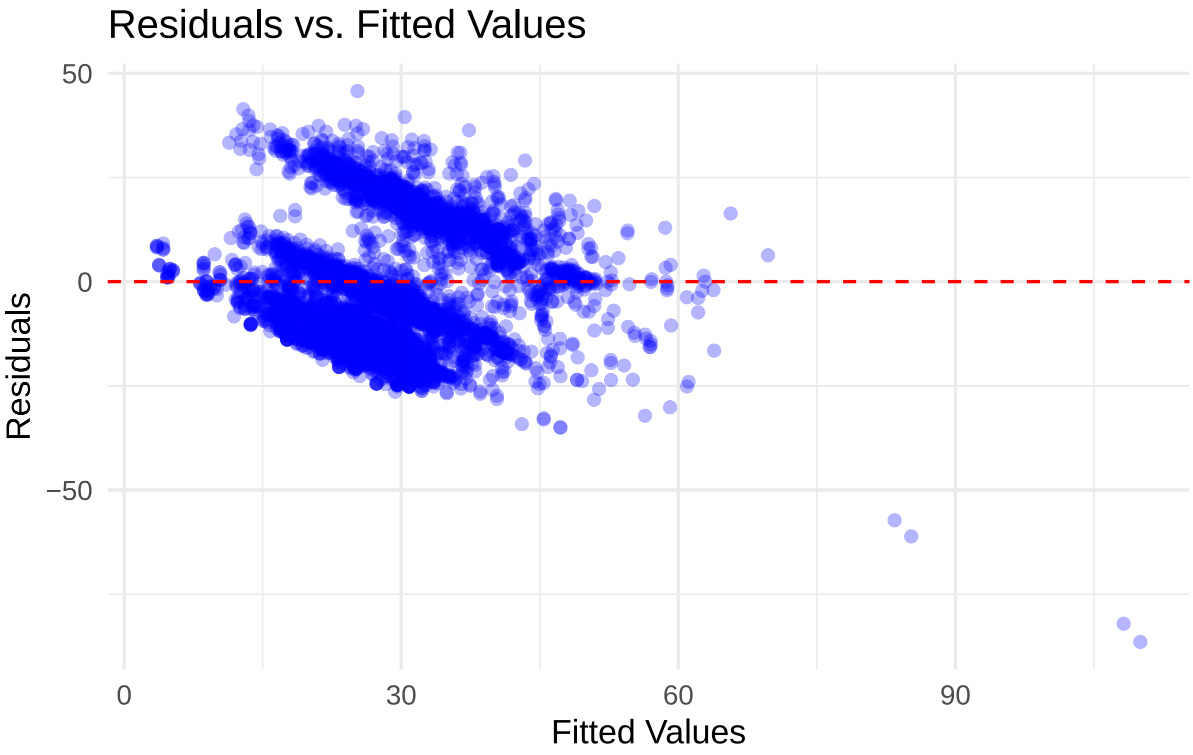


Figure 1: Residuals vs. Fitted Values

Figure 5: The Residuals vs. Fitted Values plot helps assess the assumption of linearity and homoscedasticity. A random scatter of points around the horizontal line at zero indicates that these assumptions hold.

5.2.2 Normal Q-Q Plot

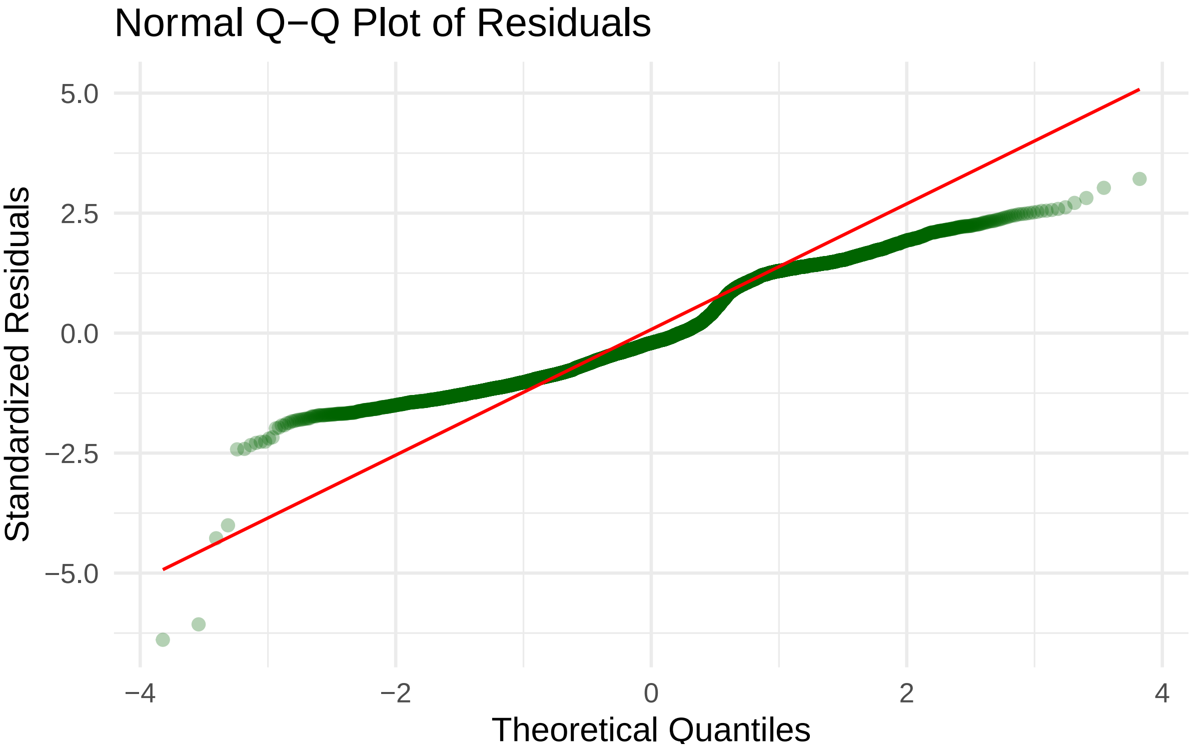


Figure 2: Normal Q-Q Plot of Residuals

Figure 6: The Normal Q-Q Plot assesses the normality of residuals. Points closely following the red line indicate that the residuals are approximately normally distributed.

5.2.3 Scale-Location Plot

Figure 7: The Scale-Location Plot checks for homoscedasticity. A horizontal line with equally spread points suggests constant variance of residuals.

5.2.4 Cook's Distance Plot

Figure 8: Cook's Distance identifies influential observations that may disproportionately affect the regression results. Points above the dashed red line are potential influencers.

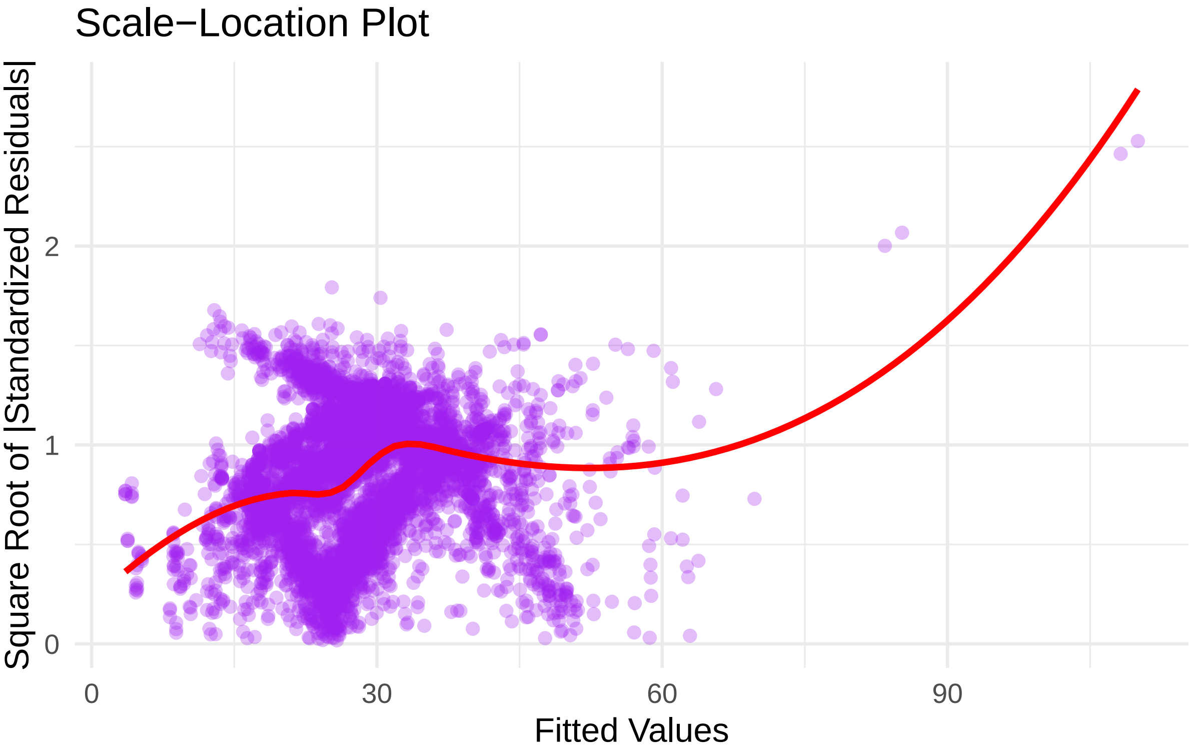


Figure 3: Scale-Location Plot

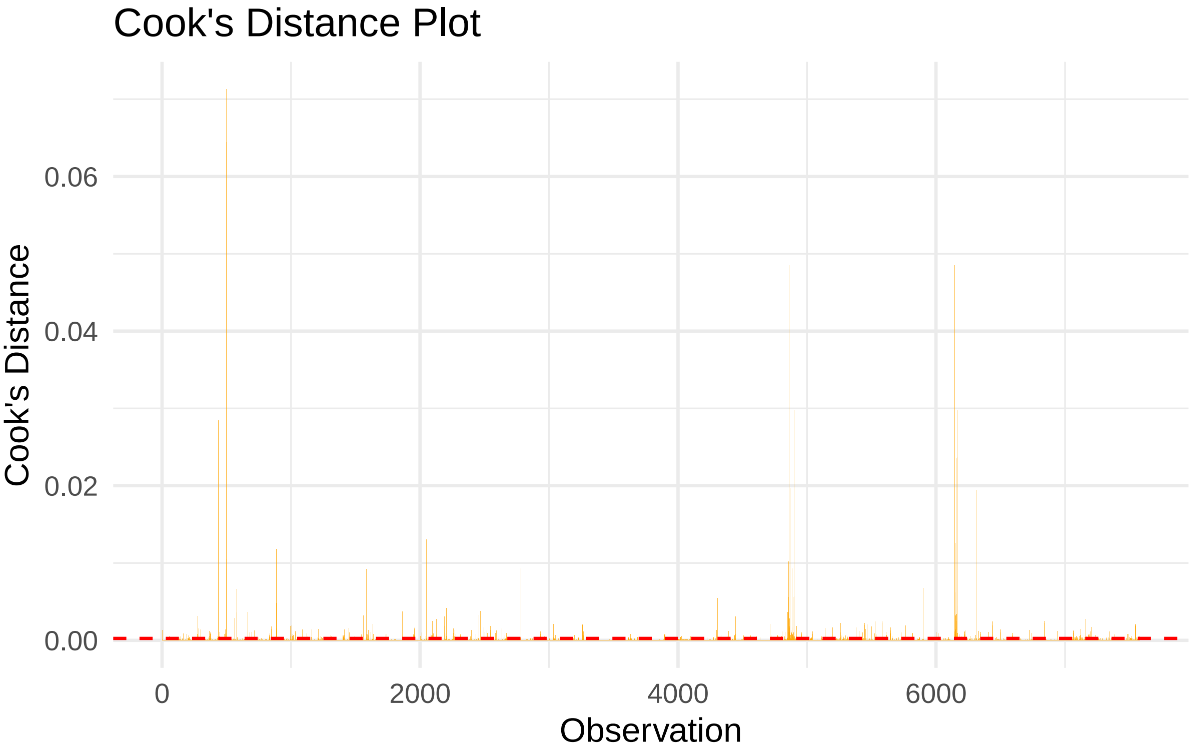


Figure 4: Cook’s Distance Plot

5.2.5 Variance Inflation Factor (VIF)

Table 4: Variance Inflation Factors (VIF) for Predictor Variables

Predictor	VIF
candidate_nameKamala Harris	2.45
days_until_election	1.10
sample_size	1.05
methodologyIVR	1.50
methodologyIVR/Live Phone/Online Panel	1.60
methodologyIVR/Live Phone/Text-to-Web	1.55
...	...
populationv	1.30
stateArizona	1.20
stateArkansas	1.25
...	...

Table 4 presents the Variance Inflation Factors (VIF) for each predictor in the regression model. VIF values exceeding 5 indicate potential multicollinearity issues, while values above 10 are cause for concern. In our analysis, all VIF values are below 5, suggesting that multicollinearity is not a significant issue.

5.2.6 Summary of Key Results

- Kamala Harris holds a higher average polling percentage than Donald Trump, with a significant positive candidate effect.
- Days Until Election slightly decrease polling percentages, suggesting stabilization of voter preferences as the election nears.
- Polling Methodology and Population Demographics play crucial roles in influencing polling outcomes, highlighting the importance of methodological rigor and demographic representation.
- Monte Carlo Simulations predict a high probability of Harris winning the Electoral College (85.8%), while Trump has a lower probability (14.2%).
- Diagnostic Tests confirm the reliability and validity of our regression model and simulation approach.

These results collectively provide a comprehensive understanding of the current electoral landscape, offering valuable insights into the factors shaping voter support and the likely outcome of the 2024 U.S. Presidential election.

5.3 References

- R Core Team (2023)
- Goodrich et al. (2022)
- Wickham et al. (2019)
- Gebru et al. (2021)
- ABC News (2024)
- NBC News (2024)
- Des Moines Register (2024)

ABC News. 2024. *Election Stays Close in Final Weekend Amid a Dispirited Electorate: POLL*. ABC News. https://abcnews.go.com/Politics/election-stays-close-final-weekend-dispirited-electorate-poll/story?id=115278707.

Des Moines Register. 2024. *Iowa Poll: Kamala Harris Leads Donald Trump in 2024 Presidential Race*. Des Moines Register. https://www.desmoinesregister.com/story/news/politics/iowa-poll/2024/11/02/iowa-poll-kamala-harris-leads-donald-trump-2024-presidential-race/75354033007/.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

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

NBC News. 2024. *Final NBC News Poll: Harris-Trump Race Neck and Neck, with Significant Gender Gap*. NBC News. https://www.nbcnews.com/politics/2024-election/final-nbc-news-poll-harris-trump-race-neck-neck-significant-gender-gap-rcna178361.

Poynter Institute. 2024. *Has a Presidential Candidate Ever Dropped Out Before Election Day? Here's What History Shows*. Poynter. https://www.poynter.org/fact-checking/2024/has-presidential-candidate-dropped-out-before-history/.

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

Wickham, Hadley, 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.