Predicting Kamala Harris's Victory in the 2024 US Election*

An Investigation into Poll Data Using Multiple Linear Regression

Xuanle Zhou Yongqi Liu Yuxuan Wei

October 25, 2024

The study predicting the 2024 U.S. presidential election is important as it provides the predicted outcomes for the US election. Analysis is based on state-by-state electoral votes, and these outcomes can influence campaign strategies and shape public discourse. The results indicate Kamala Harris will win by over 270 votes in this election cycle.

Table of contents

*Code and data are available at: https://github.com/wyx827/2024USpresidentialelection.git.

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

Section 2 expains the data we used to build up the prediction mode.

2 Data

2.1 Overview

This study uses R packages (R Core Team 2023) to clean and analyz the dataset , including libraries from tidyverse [], ggplot2[].

After cleaning the data, which included grouping and removing missing values, the analysis dataset consists of 1,683 observations, focusing on the following 11 variables: pollster name, methodology, numeric grade, start date, end date, sample size, candidate name, percentage, transparency score, and population group.

2.2 Data Measurement and Considerations

The dataset for this analysis is sourced from FiveThirtyEight, which rigorously collects and aggregates polling data from a variety of firms to reflect public opinion. To ensure the integrity of the data, only polls that adhere to specific criteria are included in the dataset. Each poll must provide essential information, such as the pollster's name, survey dates, sample sizes, and methodological details (e.g., polling medium, voter files, weighting criteria). Polls that are deemed nonscientific, that blend data from multiple sources, or that are conducted by hobbyists are excluded.

Once a poll meets these stringent standards, it is incorporated into the database, enabling it to inform polling averages, forecasts, and political coverage. This careful selection process ensures that the dataset accurately captures and reflects the nuances of public sentiment and behavior.

2.3 Outcome variables

The outcome variable of interest for this research is the percentage, representing the level of public support for Donald Trump. The distribution in Figure 1 indicates that most observations cluster around a support percentage of approximately 48%, suggesting moderate backing from the electorate. Additionally, a smaller proportion of polls show support exceeding 55%, indicating that while Trump has a core base, many voters remain either indifferent or opposed to him.

2.4 Predictor variables

2.4.1 Numeric Grade

The numeric grade reflects the quality of the pollster, with FiveThirtyEight defining a scale from 0 to 3. A grade of 0 indicates a low-quality poll, while a grade of 3 signifies a high-quality pollster. After filtering for pollsters with a numeric grade higher than 2.5, we identified a total of 30 distinct pollsters, with half of them scoring between 2.6 and 2.8, as shown in Figure 2 .

2.4.2 Sample Size

The sample size indicates the number of respondents in each poll. The distribution in Figure 3 exhibits a right-skewed shape, suggesting that there are more observations with smaller sample sizes compared to larger ones. The peak of the distribution is around 1,000, indicating that this is the most common sample size used in the polls. Overall, the sample size data highlights that each poll contains a sufficient number of respondents to provide reliable insights.

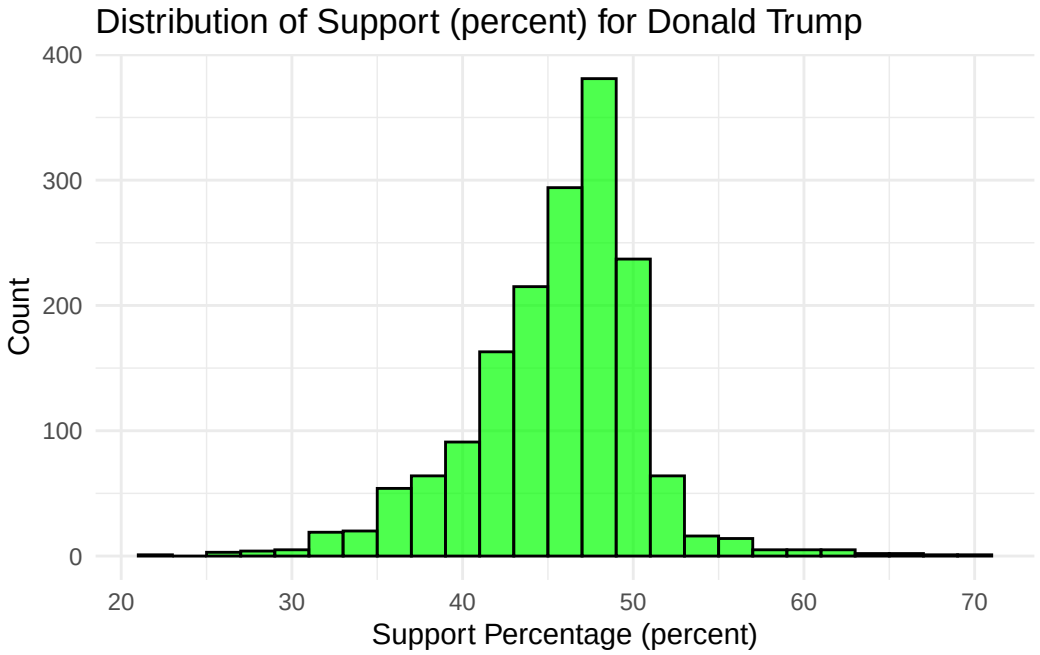

Figure 1: Distribution of Support (percent) for Donald Trump

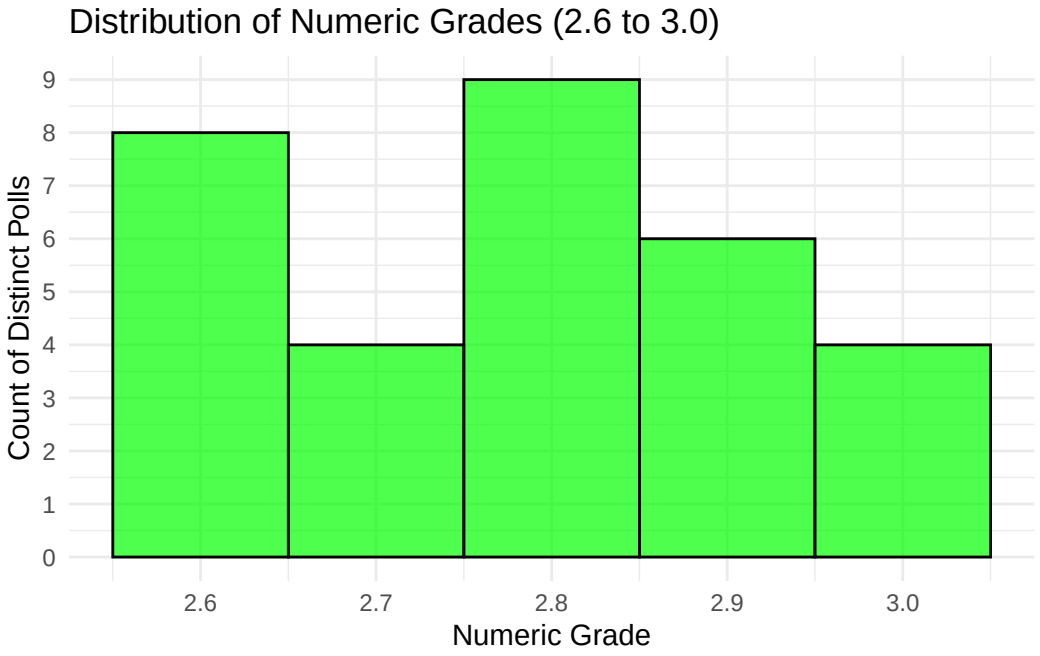

Figure 2: Distribution of Numeric Grades (2.6 to 3.0)

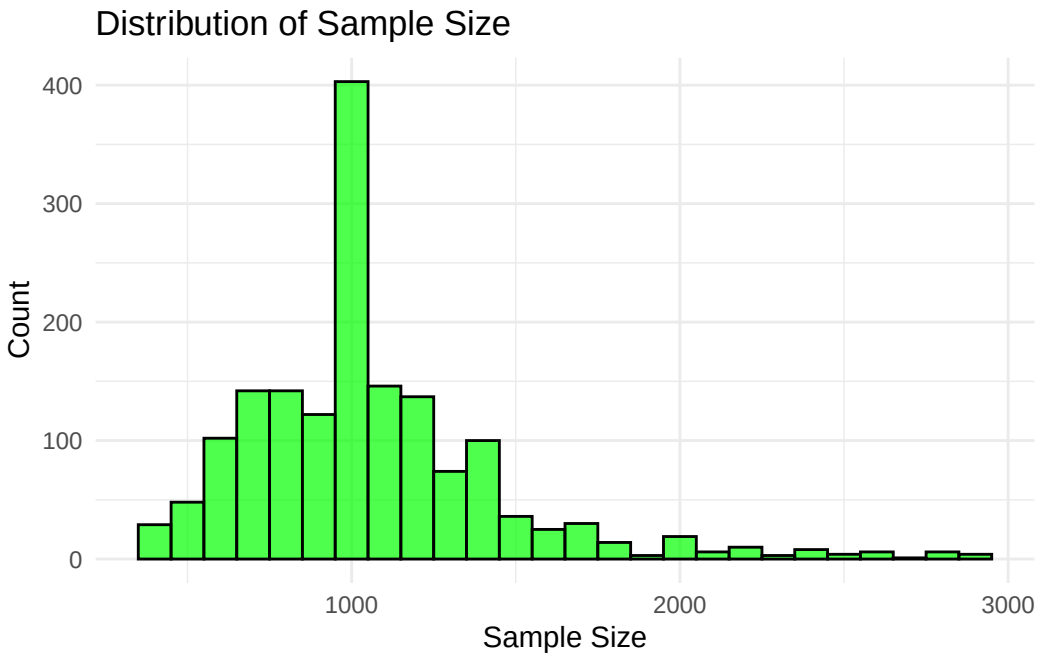

Figure 3: Distribution of Sample Size

2.4.3 Transparency Score

The Transparency Score measures how transparent a pollster is, calculated based on the amount of information disclosed about its polls, weighted by recency. The highest possible score is 10, while the lowest is 0. The distribution of Transparency Scores for the filtered pollsters shows a peak around 9 as presented in Figure 4, indicating that this is the most common score. This suggests that among the selected pollsters, there is a predominance of high transparency scores.

2.4.4 End Date

The end date indicates when each poll concluded. Figure 5 shows that as the final result of the U.S. election approached, more polls were conducted and completed. This trend reflects an increased interest in capturing public opinion and predicting electoral outcomes as election day neared, emphasizing the significance of real-time sentiment analysis during this critical period.

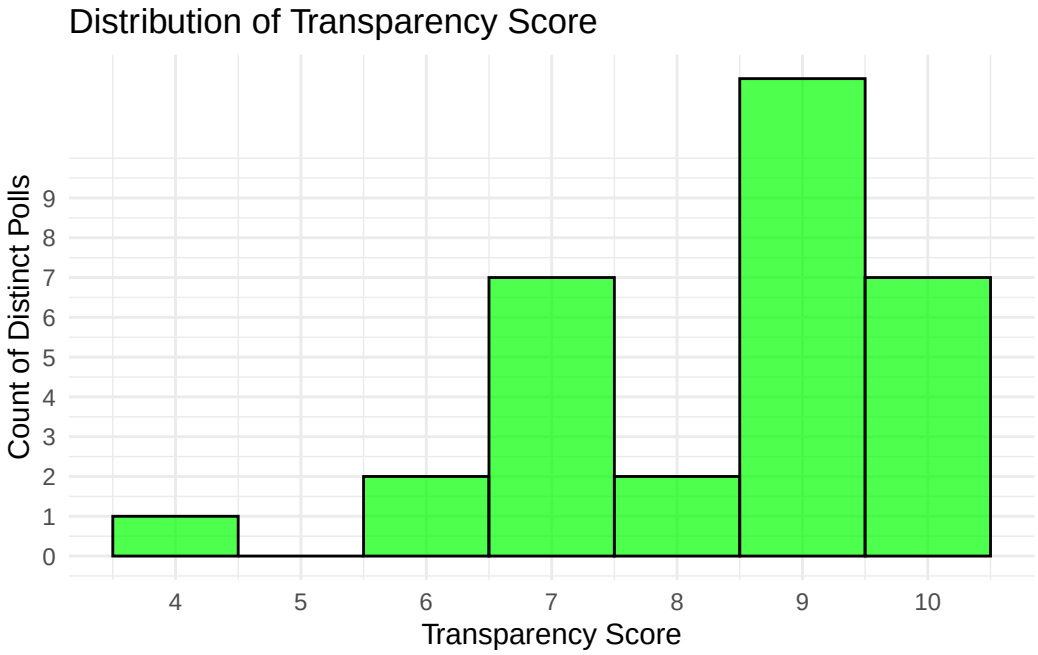

Figure 4: Distribution of Transparency Score

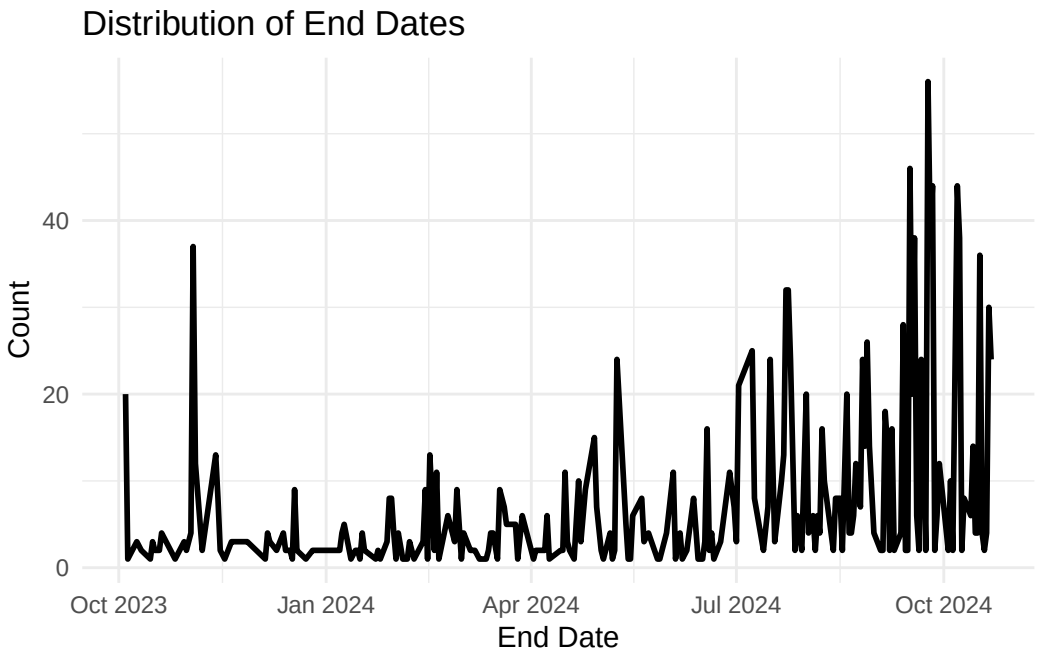

Figure 5: Distribution of End Dates

3 Model

To model Donald Trump's polling percentages over time, we employed a multiple linear regression framework. This approach estimates the relationship between polling percentages and various predictors by fitting a linear equation to the data. By analyzing the coefficients, we can quantify the impact of each predictor on Trump's polling percentages, while also assessing the overall fit of the model and making predictions based on the observed trends.

Here we briefly describe the multiple linear regression model used to investigate the winning probability of Trump. Background details and diagnostics are included in Appendix C.

3.1 Multiple Linear Regression Model Overview

The model now predicts Trump's polling percentage (percent) using the following predictors:

- Numeric Grade (numeric_grade): Reflects the quality rating of the pollster.
- Sample Size (sample_size): The number of respondents in the poll.
- State (state): A categorical variable for different U.S. states.
- Transparency Score (transparency_score): A measure of how transparent the polling data and methodology are.
- End Date (end_date): The date the poll was completed, which might capture trends over time.

The model takes the form:

$$\begin{aligned} \text{pct}_i &= \beta_0 + \beta_1 \cdot \text{numeric\_grade}_i + \beta_2 \cdot \text{transparency\_score}_i && (1) \\ &\quad + \beta_3 \cdot \text{sample\_size}_i + \beta_4 \cdot \text{state}_i + \beta_5 \cdot \text{end\_date}_i + \epsilon_i && (2) \\ \epsilon_i &\sim \text{Normal}(0, \sigma^2) && (3) \end{aligned}$$

Where:

$$\begin{aligned} \beta_0 &\text{ is the intercept term} && (4) \\ \beta_1, \beta_2, \beta_3, \beta_4, \beta_5 &\text{ are the coefficients for each predictor} && (5) \\ \sigma^2 &\text{ is the variance of the error term} && (6) \end{aligned}$$

3.2 Interpretation of Coefficients

- Intercept (β_0): This is the predicted Trump polling percentage when all predictors (numeric grade, sample size, state, transparency score, and end date) are at their baseline or zero value.
- Numeric Grade (β_1): This coefficient measures how much Trump's polling percentage changes as the pollster's numeric grade increases. A positive and significant coefficient would indicate that higher-rated pollsters report better polling numbers for Trump, while a negative coefficient would suggest the opposite.
- Sample Size (β_2): This measures the impact of the number of respondents on Trump's polling percentage. A positive coefficient would indicate that larger sample sizes are associated with higher polling percentages for Trump.
- State (β_3): The coefficients for the state variable represent differences in Trump's polling percentage in each state compared to the reference state (baseline category). For example, if the coefficient for Florida is negative, it means Trump polls lower in Florida compared to the reference state.
- Transparency Score (β_4): This coefficient shows how much Trump's polling percentage is affected by the transparency of the poll. A positive coefficient would indicate that polls with higher transparency tend to report higher polling percentages for Trump, whereas a negative coefficient would imply the opposite.
- End Date (β_5): The end date is a time-related variable, capturing trends over time. A positive and significant coefficient would suggest that Trump's polling percentage has increased as the election date approaches, while a negative coefficient would suggest a decrease in his polling numbers over time.

3.3 Interpretation

The posterior distributions of the parameters allow us to quantify the uncertainty around each effect:
- The coefficient for end_date informs us about how Trump's polling percentages have evolved over time. A positive coefficient would suggest an upward trend, while a negative coefficient would indicate a decline.
- The coefficient for numeric_grade captures the impact of pollster quality on the polling percentage. High-quality pollsters may produce different estimates compared to lower-quality ones.
- The state-level effects account for regional differences in Trump's support. Some states may show significantly higher or lower levels of support, even after adjusting for the time of the poll and pollster quality.

3.4 Model Justification

Table 1 shows the summary table for the model. And then evaluate it on the test set. It appears as though the model is having difficulty identifying Trump supporters. The reson why the multiple regression is applied is because…..

Table 1: Regression Model Estimate Summary

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-180.27	22.93	-7.86	0.00
numeric_grade	1.55	1.16	1.34	0.18
sample_size	0.00	0.00	-0.57	0.57
stateArizona	-5.48	2.61	-2.10	0.04
stateCalifornia	-18.15	2.80	-6.49	0.00
stateColorado	-13.81	3.61	-3.83	0.00
stateConnecticut	-17.94	4.40	-4.08	0.00
stateFlorida	-4.24	2.86	-1.48	0.14
stateGeorgia	-5.33	2.61	-2.04	0.04
stateIdaho	4.29	4.41	0.97	0.33

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

4.1 Predicted Electoral Outcomes

We applied a regression model to predict the percentage of votes Trump is expected to receive in each state. The model results, combined with each state's electoral vote allocation, allowed us to predict the winner in each state. Based on this, we calculated the total number of electoral votes for both Trump and Harris.

The table below (Table 2) summarizes the predicted results, showing Trump's predicted percentage, the number of electoral votes in each state, and the predicted winner (either Trump or Harris). For instance: - Alabama: Trump is predicted to win 53% of the vote, securing all 9 electoral votes. - California: Trump is predicted to receive 34.39% of the vote, resulting in a victory for Harris, who takes California's 55 electoral votes. - Florida: The model predicts a close race, with Trump at 48.89% of the vote, resulting in a Harris win in this critical battleground state.

Electoral Vote Count: - Trump Electoral Votes: 78 - Harris Electoral Votes: 416 These results indicate that based on the current model predictions, Harris is expected to win the 2024 U.S. Presidential Election, securing 416 electoral votes, compared to Trump's 78 electoral votes. The election outcome hinges on several battleground states, where the vote margins are predicted to be narrow.

```
[1] "Trump Electoral Votes: 45"
```

```
[1] "Harris Electoral Votes: 337"
```

Table 2: Prediction for Trump and Harris by Electoral College

State	Trump Predicted %	Electoral Votes	Winner
Arizona	47.64	11	Harris
California	34.50	55	Harris
Colorado	36.41	9	Harris
Florida	50.04	29	Trump
Georgia	47.38	16	Harris
Iowa	46.35	6	Harris
Maine	41.84	4	Harris
Maryland	34.50	10	Harris
Massachusetts	28.24	11	Harris
Michigan	45.78	15	Harris
Minnesota	43.92	10	Harris

Table 2: Prediction for Trump and Harris by Electoral College

State	Trump Predicted %	Electoral Votes	Winner
Missouri	54.11	10	Trump
Montana	53.93	3	Trump
Nebraska	45.44	10	Harris
Nevada	46.40	6	Harris
New Hampshire	42.57	4	Harris
New Jersey	39.40	14	Harris
New Mexico	42.57	5	Harris
New York	37.15	29	Harris
North Carolina	48.45	16	Harris
Ohio	48.59	18	Harris
Pennsylvania	46.18	20	Harris
Rhode Island	36.13	4	Harris
South Dakota	57.17	3	Trump
Texas	49.20	38	Harris
Vermont	27.00	3	Harris
Virginia	42.72	13	Harris
Wisconsin	45.62	10	Harris

4.2 Predicted Electoral Outcomes by State

The following map (Figure 6) shows the predicted winner for each state in the 2024 U.S. Presidential Election, based on the regression model's predicted vote percentages for Trump and Harris.

The predicted outcome in the regression model reflects the geographic voting patterns, with Trump winning in traditionally Republican-leaning states like Alabama, Missouri, and Wyoming, while Harris dominates in Democratic strongholds such as California, New York, and Illinois. However, key battleground states such as Florida and Arizona are predicted to favor Harris, potentially determining the overall election outcome.

5 Discussion

5.1 What Is Done in This Paper?

This paper uses a regression model to predict the outcome of the 2024 U.S. Presidential Election between Donald Trump and Kamala Harris. By analyzing historical voting data, the model forecasts the percentage of votes Trump is expected to receive in each state. These predictions

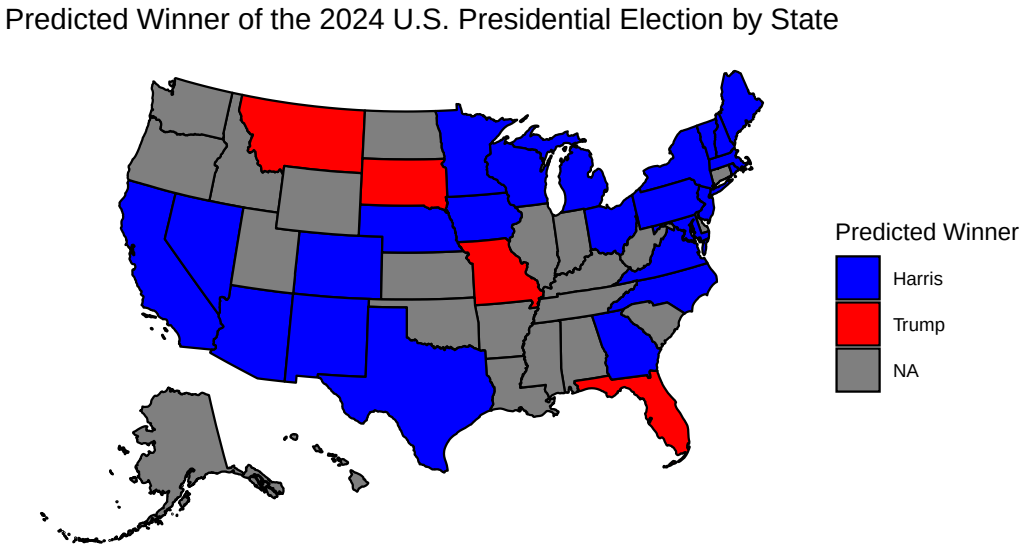

Figure 6

are combined with the state's electoral vote allocation to determine the overall winner. The findings indicate a significant electoral victory for Harris, projecting 416 electoral votes for her compared to 78 for Trump. A geographic map is also presented to visualize state-level outcomes and highlight regional voting trends.

5.2 What Do We Learn About the World?

The study highlights the persistence of geographic voting patterns in U.S. elections. Even with changes in candidates, many states follow their historical voting tendencies. Republican-leaning states like Alabama and Wyoming are projected to support Trump, while Democratic-leaning states such as California and New York favor Harris. This pattern shows the strong influence of regional political loyalties, which continue to shape election outcomes.

5.3 Another Thing We Learn About the World

The analysis underscores the importance of battleground states in determining the election's outcome. States like Florida and Arizona, predicted to lean toward Harris, play a decisive role in the projected electoral victory. This illustrates how close contests in a few states can

influence the overall election result, emphasizing the significant impact these states have in shaping the presidency.

5.4 What Is Left to Learn or How Should We Proceed in the Future?

Future research should incorporate more detailed data, including demographic shifts and economic factors that may influence voter preferences. It would be valuable to examine how emerging political movements and changes in communication strategies affect elections, particularly in closely contested states. Post-election analysis comparing predictions with actual results will also be essential to improving the accuracy of models and understanding how unforeseen events may alter outcomes.

5.5 Weaknesses and Next Steps

While the model offers a useful prediction, it has limitations. It relies heavily on past voting data, which may not fully capture future changes in voter behavior or external factors that could affect the election. The model also simplifies the election by focusing on state-level percentages, without accounting for variables such as turnout, third-party candidates, or unexpected political shifts. Additionally, the assumption that past voting patterns will continue might not be accurate in a rapidly changing political landscape. The logistic regression model is limited by the lack of interaction terms between states, which may obscure regional trends or external factors that influence neighboring states. Additionally, the model assumes that polling data accurately reflects voter intentions, but biases or missing data in key regions could affect the predictions.

Future iterations of this model should incorporate voter turnout predictions, demographic variables, and real-time polling updates to better capture the dynamics of voter behavior. Including social media sentiment analysis and analyzing cross-state interactions could refine the prediction model, offering more nuanced insights into election forecasts.

Appendix

A Pollster Methodology Overview and Evaluation

A.1 Overview of SurveyUSA

SurveyUSA is a privately held opinion research company that operates nationwide, across all 50 U.S. states. Since its founding, the company has conducted over 40,000 research projects, serving a client base of 400 organizations, including media outlets, corporations, non-profits, government agencies, and academic institutions. Known for its expertise in localized opinion research, SurveyUSA focuses on gathering data at the city, county, and regional levels. The company offers timely, cost-effective surveys tailored to meet specific client needs, distinguishing itself from larger global firms.

A.2 Population, Frame, and Sample

- Target Population: U.S. citizens eligible to vote in the 2024 presidential election.
- Sample Frame: U.S. households with either home telephones or access to devices such as phones or tablets.
- Sample Size: Sample sizes vary across different polls. For the 2024 U.S. presidential election cycle, SurveyUSA conducted 49 polls, with sample sizes ranging from 507 to 2,330 for registered voters or likely voters. The average sample size for these polls is approximately 1,045 households.

A.3 Recruitment

SurveyUSA employs a mixed-method approach to recruitment, including online panels, telephone calls, and a text-to-web method. Some respondents are recruited through Random Digit Dialing (RDD) using telephone samples purchased from Aristotle, while others, who do not use home telephones, are invited to complete the survey on an electronic device such as a phone or tablet. Respondents from non-probability online panels are selected randomly by Cint/Lucid Holdings LLC.

A.4 Sampling approach and Trade-offs

SurveyUSA uses a blend of probability and non-probability sampling methods. Some respondents are drawn from non-probability online panels, while others are recruited using probability-based telephone sampling. Responses are weighted based on the latest U.S.

Census estimates for age, gender, ethnicity, and region, ensuring alignment with the target population. Questions and answer choices are rotated to reduce order bias, recency effects, and latency effects.

- **Advantages:**
  The diverse sampling approach not only ensures a broad range of opinions is captured but also complements probability-based sampling, which accurately reflects the overall population. Furthermore, reweighting the data according to U.S. Census demographics strengthens the credibility of the results by ensuring demographic accuracy. Additionally, rotating questions and answer choices helps mitigate bias, further improving the reliability of the data. Finally, the use of online surveys offers a cost-effective solution for efficient data collection.

- **Disadvantages:**
  Phone-based data collection tends to be time-consuming and can be affected by interviewer effects during telephone interviews. Additionally, challenges like non-response issues, such as busy signals or refusals to participate, can hinder the effectiveness of the data collection process.

A.5 Non-response Handling

In cases of non-response, SurveyUSA attempts follow-up calls if interviews are interrupted by answering machines or busy signals. Weighting is applied to adjust for non-response bias, although this doesn't completely eliminate challenges posed by unreachable or unwilling participants.

A.6 Questionnaire Evaluation

- **Positive Aspects:** A logical flow between questions facilitates easy navigation for respondents throughout the survey, while simple wording promotes inclusivity by enabling individuals from diverse backgrounds to comprehend the questions. Furthermore, all questions are directly relevant to analyzing the 2024 U.S. presidential election, and providing predefined response options simplifies the choices for participants.

- **Negative Aspects:** Static options for party affiliation and ideology may fail to capture the nuances of respondents' political beliefs. These rigid categories could oversimplify complex political identities.

A.7 Summary Evaluation

SurveyUSA's methodology reflects a balanced approach, leveraging various sampling approach and method to reach a representative sample. While its blend of probability and non-probability methods has strengths, such as cost-effectiveness and broad reach, it faces challenges related to telephone interview logistics, potential interviewer bias, and the limitations of fixed questionnaire options. Nevertheless, the inclusion of data weighting and question rotation adds credibility to its results, making SurveyUSA a reliable pollster for localized opinion research.

B Appendix B: Idealized Methodology and Survey

B.1 Objective and Overview

The goal of this survey methodology is to accurately forecast the outcome of the U.S. presidential election by collecting high-quality, representative data from a diverse set of respondents across the country. With a budget of $100,000, this methodology incorporates sophisticated sampling techniques, robust respondent recruitment strategies, and rigorous data validation protocols. The approach is designed to maximize accuracy, reduce bias, and account for various demographic, geographic, and political factors that influence voting behavior.

B.2 Core Objectives

- Obtain a representative sample of the U.S. electorate.
- Ensure data quality through rigorous validation.
- Leverage statistical modeling and poll aggregation for an accurate prediction.

B.3 Sampling Strategy

The sampling strategy is designed to ensure that the survey reaches a broad, representative section of the voting population. To achieve this, we will use **stratified random sampling** combined with **quota sampling** for key demographics. This ensures that each important subgroup within the population is adequately represented.

B.3.1 Stratification Variables

- **Age Groups**: 18-29, 30-44, 45-64, 65+
- **Gender**: Male, Female, Non-binary/Other
- **Race/Ethnicity**: White, Black, Hispanic/Latino, Asian, Indigenous, Other

- **Education Level**: No high school, High school graduate, College graduate, Post-graduate
- **Income Bracket**: <$30,000, $30,000-$60,000, $60,000-$100,000, >$100,000
- **Geographic Region**: Northeast, Midwest, South, West

B.3.2 Sample Size

A total of **10,000 respondents** will be surveyed, providing a margin of error of approximately $\pm 1\%$ at a 95% confidence level. This sample size will allow for detailed subgroup analysis (e.g., by state, demographic group), yielding statistically robust predictions.

B.3.3 Weighting

Post-stratification weights will be applied to adjust for any oversampling or undersampling of specific demographic groups. For example, younger voters or underrepresented minorities will be weighted to reflect their true proportions in the voting population.

B.4 Recruitment Strategy

To maximize respondent diversity and ensure accurate sampling, the survey will employ **multi-channel recruitment**:

- **Digital Advertisements**: Targeted ads on platforms like Facebook, Instagram, and Google will recruit respondents based on their demographic profiles (age, gender, location, political interest).
- **Email Outreach**: If permissible, voter registration databases will be accessed to send email invitations to registered voters.
- **Partnerships with Civic Organizations**: Partnering with non-profits and civic organizations that engage diverse communities will further boost respondent diversity.
- **Incentives**: Each participant will be entered into a lottery with a chance to win a $100 gift card to encourage participation.

B.5 Data Validation and Quality Assurance

Maintaining data integrity and ensuring high-quality responses are critical to the accuracy of the election forecast. Several measures will be put in place to validate responses and reduce noise in the dataset.

B.5.1 Data Validation Protocols

- **Real-time Captcha Verification**: This will prevent automated bots from submitting responses.
- **Email/Phone Verification**: Respondents will verify their email or phone number to ensure authenticity and prevent duplicate submissions.
- **Time on Task Monitoring**: The survey platform will monitor the time respondents spend on each question. Responses completed suspiciously quickly will be flagged for review.
- **Voter Registration Cross-Check**: If feasible, respondents will be cross-referenced with voter registration records to ensure eligibility.
- **Response Audits**: Randomly selected respondents will be contacted to verify the accuracy of their responses, ensuring dataset integrity.

B.6 Poll Aggregation and Data Analysis

B.6.1 Poll Aggregation

This survey will be combined with results from reputable polling firms (e.g., YouGov, Ipsos, Gallup) to strengthen the forecast through a **poll-of-polls** approach.

- **Weighting by Methodology and Recency**: Poll results will be weighted based on the rigor of their methodology and the recency of the poll.
- **Handling Bias and Variability**: Aggregated results will adjust for pollster biases and variability between polls to ensure that no single poll dominates the prediction.

B.6.2 Modeling Approach

Bayesian hierarchical models will account for variability across different states, demographics, and regions. This will allow for modeling the popular vote and potentially translating it into **Electoral College predictions**.

B.7 Budget Allocation

- **Respondent Recruitment (Targeted ads, outreach)**: $70,000
- **Incentives (e.g., lottery prizes)**: $10,000
- **Survey Platform (Google Forms, Qualtrics subscription)**: $5,000
- **Data Validation Tools**: $5,000
- **Poll Aggregation & Analysis Software**: $10,000

B.8 Survey Implementation

The survey will be implemented via **Google Forms**, which offers a cost-effective platform for data collection. You can access the survey at the following link: Google Form Survey

B.8.1 Survey Structure:

Introduction:

Thank you for taking part in this survey aimed at predicting the outcome of the 2024 US Presidential election. Your insights are valuable to our research.

Please note:

- **All responses will be kept strictly confidential.**
- **Your participation is entirely voluntary.**
- **We kindly request that you answer all questions honestly and to the best of your knowledge.**
- **The survey is estimated to take approximately 10 minutes to complete.**

If you have any inquiries or concerns regarding this survey, please don't hesitate to contact the research team at shaw.wei@mail.utoronto.ca.(Yuxuan Wei, Xuanle Zhou, Yongqi Liu)

Your contribution to this study is greatly appreciated! Each participant will be entered into a lottery with a chance to win a $100 gift card!

Section 1: Eligibility Screening:

Are you a U.S. citizen? - Yes - No [If No, end survey]

Will you be 18 or older by Election Day (November 5, 2024)? - Yes - No [If No, end survey]

Are you registered to vote in the United States? - Yes - No - Not sure - Plan to register before the election

Section 2: Demographic Information:

What is your age group? - 18-29 - 30-44 - 45-64 - 65 or older - Prefer not to say

What is your gender? - Male - Female - Non-binary/Other - Prefer not to say

What is your race/ethnicity? (Select all that apply) - White - Black or African American - Hispanic or Latino - Asian - American Indian or Alaska Native - Native Hawaiian or Pacific Islander - Prefer not to say - Other: [Short text answer]

What is your highest level of education completed? - No high school - High school graduate or equivalent - Some college, no degree - Bachelor's degree - Graduate or professional degree - Prefer not to say

What was your total household income in 2023? - Less than $30,000 - $30,000 - $59,999 - $60,000 - $99,999 - $100,000 - $149,999 - $150,000 or more - Prefer not to say

In which region of the United States do you currently reside? - Northeast (ME, NH, VT, MA, RI, CT, NY, NJ, PA) - Midwest (OH, IN, IL, MI, WI, MN, IA, MO, ND, SD, NE, KS) - South (DE, MD, DC,

VA, WV, NC, SC, GA, FL, KY, TN, AL, MS, AR, LA, OK, TX) - West (MT, ID, WY, CO, NM, AZ, UT, NV, WA, OR, CA, AK, HI)

Section 3: Political Views and Voting Intentions:

How likely are you to vote in the 2024 Presidential election? - Definitely will vote - Probably will vote - Might or might not vote - Probably will not vote - Definitely will not vote

Generally speaking, do you usually think of yourself as a: - Democrat - Republican - Independent - Prefer not to say - Other: [Short text answer]

If the 2024 Presidential election were held today, who would you vote for? - Kamala Harris (Democrat) - Donald Trump (Republican) - Undecided - Prefer not to say - Other: [Short text answer]

How certain are you about your choice? - Very certain - Somewhat certain - Not very certain - Not at all certain - Prefer not to say

Which THREE issues are most important to you in deciding your vote? (Select exactly three) - Economy and jobs - Healthcare - Immigration - Climate change - National security - Education - Gun policy - Social justice/racial equality - Taxes - Crime and public safety - Foreign policy - Other: [Short text answer]

Section 4: Information Sources and Engagement:

What is your primary source of political news? (Select all that apply) - Network TV news (ABC, CBS, NBC) - Cable TV news (CNN, Fox News, MSNBC) - News websites - Social media - Radio - Print newspapers - Friends and family - Other: [Short text answer]

How closely have you been following news about the 2024 Presidential election? - Very closely - Somewhat closely - Not too closely - Not at all - Not sure

Section 5: Validation and Consent:

Please verify that you are a human by selecting "Blue" from the following options: - Red - Green - Blue - Yellow

Consent Statement: "I understand that my participation in this survey is voluntary and that my responses will be kept confidential. I agree that my responses may be used for research purposes." - Yes, I agree - No, I do not agree

Email Address: [Email field]

End Message:

"Thank you for completing this survey. Your response has been recorded. If you have any questions about this survey or would like to be informed about the results, please contact at shaw.wei@mail.utoronto.ca."

B.9 Survey Design Considerations

- **Question Wording**: All questions are designed to avoid bias or leading responses.
- **Neutrality**: Political questions are framed neutrally to avoid influencing respondents' answers.
- **Pilot Testing**: The survey will undergo a pilot test to identify and resolve any issues before full deployment.

C Model details

C.1 Diagnostics

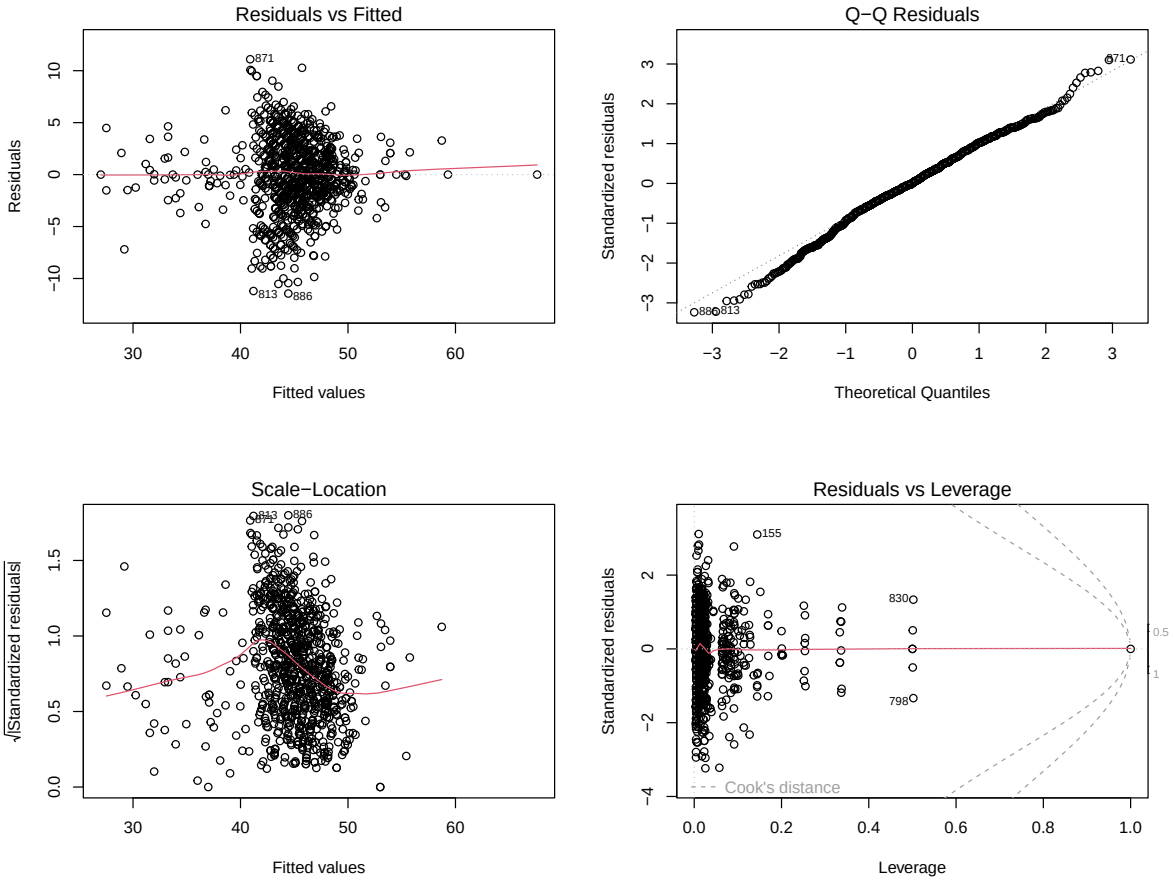

Figure 7

C.2 Calculate Mean Squared Error (MSE) on test data

```
[1] 11.30393
```

```r
library(caret)
```

```
Loading required package: lattice


Attaching package: 'caret'

The following objects are masked from 'package:rstanarm':

    compare_models, R2

The following object is masked from 'package:purrr':

    lift
```

```r
set.seed(856)
train_control <- trainControl(method = "cv", number = 10)
cv_model <- train(percent ~ numeric_grade + sample_size + state + transparency_score + end_da
                  data = analysis_data_train,
                  method = "lm",
                  trControl = train_control)
```

```
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
```

```r
print(cv_model$results)
```

```
  intercept     RMSE  Rsquared      MAE    RMSESD RsquaredSD     MAESD
1      TRUE 4.112088 0.3795555 3.035667 0.9986783  0.1514037 0.3949995
```

```r
baseline_prediction <- mean(analysis_data_train$percent)
baseline_mse <- mean((analysis_data_test$percent - baseline_prediction)^2)
print(baseline_mse)
```

```
[1] 25.72931
```

```r
# Calculate R-squared on test data
rss <- sum((predicted_values - actual_values)^2)  # Residual Sum of Squares
tss <- sum((actual_values - mean(actual_values))^2)  # Total Sum of Squares

r_squared <- 1 - (rss/tss)
print(r_squared)
```

```
[1] 0.5598326
```

```r
# Create a correlation matrix to check for collinearity
cor_matrix <- cor(analysis_data_train[, c("numeric_grade", "sample_size", "transparency_score
print(cor_matrix)
```

```
                 numeric_grade sample_size transparency_score
numeric_grade       1.00000000 -0.06464797         0.11854624
sample_size        -0.06464797  1.00000000         0.07612837
transparency_score  0.11854624  0.07612837         1.00000000
```

```r
# Check for sparse levels in 'state'
state_counts <- table(analysis_data_train$state)
print(state_counts)
```

```
      Alaska      Arizona   California     Colorado  Connecticut 
           2           52           10            2            1 
     Florida      Georgia        Idaho     Illinois      Indiana 
           8           49            1            1            3 
```

Iowa	Kansas	Maine	Maryland	Massachusetts
2	1	11	5	9
Michigan	Minnesota	Missouri	Montana	National
59	11	4	6	388
Nebraska	Nevada	New Hampshire	New Jersey	New Mexico
7	38	13	1	1
New York	North Carolina	Ohio	Pennsylvania	Rhode Island
14	47	13	64	4
South Carolina	South Dakota	Tennessee	Texas	Vermont
1	3	1	16	1
Virginia	Washington	West Virginia	Wisconsin	Wyoming
16	3	1	77	1

D Acknowledgements

R Core Team (2023) Thanks to Open AI (OpenAI 2024) and ChatGPT 4.0 is used to write the analysis of the paper.

References

OpenAI. 2024. "ChatGPT (Version GPT-4)." https://chat.openai.com/.

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