Race to the Ballot Box: Will Demographics Decide a Winner This Year?*

A Regressional Approach to Race and Political Preferences in the 2024 US Elections and Insights on the Next

Sameeck Bhatia Sean Chua Tanmay Shinde

November 1, 2024

This paper analyzes polling data for the 2024 U.S. Presidential election to predict the winner, while also capturing demographic sentiment toward the two main contenders - Donald Trump and Kamala Harris. We found that ….[NEED TO COMPLETE]. This study highlights that considering factors such as the reputation of the pollster, quality of the poll, bias, sample size, and demographic area are essential for analyzing polling data and thereby forecasting election outcomes. Highlighting these trends is crucial for understanding the evolving political landscape in the U.S., as they reveal how voter demographics and sentiments shape election outcomes.

Table of contents

*Code and data are available at: https://github.com/SameeckBhatia/2024-US-Elections.

1 Introduction

As the 2024 U.S. Presidential election draws near, the race between former President Donald Trump and Vice President Kamala Harris has garnered widespread attention worldwide. With both candidates representing contrasting ideologies and policy priorities, appealing to distinct voter bases, understanding the public sentiment surrounding them becomes crucial in predicting the election's outcome. Polls and demographic analyses provide a lens through which we can examine the factors influencing voter preferences and gauge the political landscape as it evolves. This paper aims to analyze the 2024 election polling data to not only forecast the winner but also to capture the demographic-wise public sentiment surrounding the two contenders.

The primary estimand for this analysis is the predicted support percentage for each of the two candidates, which represents the proportion of the electorate that is expected to favor one candidate over the other. This estimand will be calculated based on survey data or polling results, and it will allow us to estimate the level of support each candidate has within the population. This analysis will also provide insights into how various demographic factors or voter preferences may influence the final outcomes.

We use polling data from FiveThirtyEight, focusing on surveys from organizations such as YouGov, Siena/NYT, and Washington Post. We used a Bayesian approach to model the polling support percentages for both candidates, and understand how polling factors such as numeric grade, poll score, state, and sample size affect the predicted support percentage.

The results show that … [TO-DO]. These results are important because … [TO-DO].

The rest of the paper is structured as follows: Section 2 details the data and measurement process. Section 3 presents the model and justifies the choices made in the building of the chosen model. Section 4 presents the results, highlighting the relationship between different variables and the polling support percentage, and Section 5 discusses the implications of the findings for this as well as future election forecasting.

2 Data

2.1 Overview

The dataset used for this paper is the 2024 national-level presidential general election dataset from FiveThirtyEight (FiveThirtyEight 2024), which compiles polling data from various reputable pollsters, including YouGov, Siena/NYT, Ipsos, and more. This dataset offers a comprehensive snapshot of public opinion leading up to the 2024 U.S. Presidential election, capturing support for the two main contenders: Donald Trump and Kamala Harris. By focusing on national polling, this dataset provides insight into voter sentiment across a wide range of demographic groups and regions.

The FiveThirtyEight dataset is part of a broader landscape of polling data used in electoral predictions. It includes surveys conducted at the national level, focusing on general election matchups, and is continuously updated as new polls are released. The dataset covers polls taken from early in the election cycle to the final days before the election, capturing the shifts in public opinion over time. While there are other polling datasets available, such as those from RealClearPolitics (RealClearPolitics 2024) or individual pollsters, we selected the FiveThirtyEight dataset due to its high level of transparency, detailed methodology grades for each pollster, and advanced polling aggregation techniques, which adjust for biases and historical pollster performance. These features provide a more refined and reliable dataset compared to others that might not incorporate such rigorous adjustments.

We use the statistical programming language R (R Core Team 2023) to perform data cleaning and analysis on the dataset to ensure consistency and reliability. Polls were filtered to include only those related to Donald Trump and Kamala Harris, with a pollster grade of 2.8 or higher, and those with a negative pollscore, ensuring reliability and unbiasedness. The dataset was further narrowed by selecting polls conducted on or after July 21, 2024, to focus on the shift in sentiment over time after the Democratic Party announced Harris as their nominee for president in the 2024 elections.

2.2 Measurement

The process of converting real-world electoral phenomena into data involves careful measurement of public sentiment toward presidential candidates, in the context of the 2024 U.S. presidential election. The dataset is built from polling data collected by various established pollsters, such as YouGov and Siena/NYT, who use a variety of survey techniques to measure voter preferences. These polls aim to capture critical variables that reflect voter sentiment, such as the percentage of respondents favoring each candidate, the sample size of the poll, and the demographics of the voters surveyed.

The measurement begins with pollsters designing surveys aimed at accurately reflecting voter behavior and sentiment. A subset of the population is sampled using stratified random sampling to ensure that the sample accurately represents the diverse demographic makeup of the electorate. Responses are collected through various channels, including phone calls, online surveys, and face-to-face interviews. These polls capture voter sentiment around a candidate as the percentage of respondents favoring the candidate.

Once the polling is done, the dataset undergoes extensive cleaning and processing to handle missing values and any discrepancies in the data. Each row in the dataset represents the outcome of a poll, including details such as the date it was conducted, the pollster's methodology, and the percentage of support for each candidate. Platforms like FiveThirtyEight then assign a numeric grade and poll score to each poll to evaluate the reliability, quality, and historical accuracy of their polling methods. The numeric grade is a standardized rating assigned to pollsters based on their historical performance, transparency, and polling methodology. The poll score measures the accuracy, consistency, and bias in each poll, pollsters that show bias toward a particular party receive a positive score.

Thus, each entry in the dataset corresponds to the results of a specific poll, providing information on when it was conducted, the pollster's methodology, and key metrics such as biasdness, sample size, and geographic coverage. By converting complex voter opinions into structured data, we are able to systematically analyze trends in public sentiment, making it possible to predict the election's outcome with greater precision.

2.3 Outcome variables

We aim to predict our response variable `candidate_trump`, which is a binary variable with 0 and 1 representing a loss and win in a state respectively. The plot of our outcome variables is shown below:

TO-DO: It is important to understand what the variables look like by including graphs, and possibly tables, of all observations, along with discussion of those graphs and the other features of these data. Summary statistics should also be included, and well as any relationships between the variables.

2.4 Predictor variables

TO-DO: All variables should be thoroughly examined and explained. [PLOTS OF PREDICTOR VARIABLES NEEDED] - **Sample Size (`sample_size`)**: The number of respondents in the poll. - **State (`state`)**: A categorical variable for different U.S. states. - **Days to Election (`days_to_election`)**: Counts the number of days until Election Day starting from today's date (it also uses `end_date` from the original dataset). - **Voter Percentage of Candidate (`pct`)**: A continuous variable for the percentage of voters who plan to vote for Trump based on polling data.

3 Model

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam auctor quam et justo efficitur, sit amet eleifend elit euismod. Mauris non libero vel ligula tincidunt consequat. Aenean gravida, risus id auctor aliquam, orci leo auctor tellus, non tincidunt est arcu at lectus. Vivamus feugiat et quam ut consequat. Aenean mollis ullamcorper facilisis. Vivamus scelerisque lectus et elementum vulputate. Nulla et arcu vehicula, iaculis felis sit amet, semper justo. Suspendisse quam augue, hendrerit a tortor non, tristique feugiat sem. Curabitur vitae tortor nec ligula scelerisque rutrum.

3.1 Model set-up

A Bayesian approach is used in constructing our model. As such, an informative prior is assumed, and logistic regression is implemented due to the nature of the outcome variable (binary).

The predictor variables used in the model are as follows:

- **Sample Size (`sample_size`)**: The number of respondents in the poll.
- **State (`state`)**: A categorical variable for different U.S. states.
- **Days to Election (`days_to_election`)**: Counts the number of days until Election Day starting from today's date (it also uses `end_date` from the original dataset).
- **Voter Percentage of Candidate (`pct`)**: A continuous variable for the percentage of voters who plan to vote for Trump based on polling data.

The model takes the form of the following equation [TO-DO]:

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

Where:

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

3.2 Interpretation

[WORDS TO-DO]

3.3 Model Justification and Evaluation

[tbl-summary] shows the summary table for the model. F1-Score blabla

4 Results

[WORDS]

State	Predicted Harris %	Predicted Trump %	Electoral Votes	Predicted Winner
Alaska	44.10	56.26	3	Trump
Arizona	45.13	55.98	11	Trump
Arkansas	26.97	NA	6	Trump
California	69.52	NA	54	Harris
Connecticut	NA	33.95	7	Harris
Florida	41.26	58.14	30	Trump
Georgia	48.93	49.50	16	Trump
Indiana	27.38	72.62	11	Trump
Iowa	42.35	42.95	6	Trump
Maine	48.41	34.04	4	Harris
Maryland	84.28	21.05	10	Harris
Massachusetts	71.17	23.90	11	Harris
Michigan	42.46	46.84	15	Trump
Minnesota	51.49	34.77	10	Harris
Missouri	25.94	59.88	10	Trump
Montana	26.64	69.62	4	Trump
Nebraska	42.24	63.40	5	Trump
Nevada	46.20	51.13	6	Trump
New Hampshire	48.27	26.74	4	Harris
New Mexico	55.75	39.01	5	Harris
New York	64.70	37.88	28	Harris
North Carolina	47.25	47.83	16	Trump
North Dakota	26.25	57.83	3	Trump
Ohio	36.21	60.08	17	Trump
Oklahoma	NA	79.83	7	Trump
Oregon	39.57	NA	8	Trump
Pennsylvania	50.93	50.40	19	Harris
Rhode Island	62.76	52.15	4	Harris
South Carolina	31.10	51.70	9	Trump
South Dakota	15.72	NA	3	Trump
Tennessee	14.80	NA	11	Trump
Texas	34.27	55.38	40	Trump
Utah	24.12	60.48	6	Trump
Vermont	86.92	NA	3	Harris
Virginia	54.70	43.48	13	Harris
Washington	72.04	23.24	12	Harris
West Virginia	23.05	NA	4	Trump
Wisconsin	47.88	46.81	10	Harris

5 Discussion

5.1 First discussion point

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam auctor quam et justo efficitur, sit amet eleifend elit euismod. Mauris non libero vel ligula tincidunt consequat. Aenean gravida, risus id auctor aliquam, orci leo auctor tellus, non tincidunt est arcu at lectus. Vivamus feugiat et quam ut consequat. Aenean mollis ullamcorper facilisis. Vivamus scelerisque lectus et elementum vulputate. Nulla et arcu vehicula, iaculis felis sit amet, semper justo. Suspendisse quam augue, hendrerit a tortor non, tristique feugiat sem. Curabitur vitae tortor nec ligula scelerisque rutrum. Nullam feugiat odio metus. Cras id convallis ante, ut ornare velit. Mauris turpis purus, porttitor eu leo quis, suscipit euismod ligula.

5.2 Second discussion point

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam auctor quam et justo efficitur, sit amet eleifend elit euismod. Mauris non libero vel ligula tincidunt consequat. Aenean gravida, risus id auctor aliquam, orci leo auctor tellus, non tincidunt est arcu at lectus. Vivamus feugiat et quam ut consequat. Aenean mollis ullamcorper facilisis. Vivamus scelerisque lectus et elementum vulputate. Nulla et arcu vehicula, iaculis felis sit amet, semper justo. Suspendisse quam augue, hendrerit a tortor non, tristique feugiat sem. Curabitur vitae tortor nec ligula scelerisque rutrum. Nullam feugiat odio metus. Cras id convallis ante, ut ornare velit. Mauris turpis purus, porttitor eu leo quis, suscipit euismod ligula.

5.3 Third discussion point

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam auctor quam et justo efficitur, sit amet eleifend elit euismod. Mauris non libero vel ligula tincidunt consequat. Aenean gravida, risus id auctor aliquam, orci leo auctor tellus, non tincidunt est arcu at lectus. Vivamus feugiat et quam ut consequat. Aenean mollis ullamcorper facilisis. Vivamus scelerisque lectus et elementum vulputate. Nulla et arcu vehicula, iaculis felis sit amet, semper justo. Suspendisse quam augue, hendrerit a tortor non, tristique feugiat sem. Curabitur vitae tortor nec ligula scelerisque rutrum. Nullam feugiat odio metus. Cras id convallis ante, ut ornare velit. Mauris turpis purus, porttitor eu leo quis, suscipit euismod ligula.

5.4 Weaknesses and next steps

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam auctor quam et justo efficitur, sit amet eleifend elit euismod. Mauris non libero vel ligula tincidunt consequat. Aenean

gravida, risus id auctor aliquam, orci leo auctor tellus, non tincidunt est arcu at lectus. Vivamus feugiat et quam ut consequat. Aenean mollis ullamcorper facilisis. Vivamus scelerisque lectus et elementum vulputate. Nulla et arcu vehicula, iaculis felis sit amet, semper justo. Suspendisse quam augue, hendrerit a tortor non, tristique feugiat sem. Curabitur vitae tortor nec ligula scelerisque rutrum. Nullam feugiat odio metus. Cras id convallis ante, ut ornare velit. Mauris turpis purus, porttitor eu leo quis, suscipit euismod ligula.

Appendix

A Additional Data Details

B Additional Model Details

C Polling Methodology Overview and Evaluation for Siena/NYT

The poll was conducted from Sept. 29 to Oct. 6 and surveyed 3,385 likely voters nationwide and found that Harris led Trump by 49 percent to 46 percent, a slight lead that is within the poll's margin of error (2.4 points). According to (The New York Times 2024c), the national poll includes separate polls of 622 voters in Florida, and 617 voters in Texas. The weight given to each of these groups in the national poll has been adjusted so that the overall results are reflective of the entire country. Polls were also conducted by telephone, using live interviewers, in both English and Spanish, with about 98% of respondents were contacted on a cellphone (The New York Times 2024b). For battleground polls, voters were called from Arizona, Georgia, Michigan, Nevada, Pennsylvania and Wisconsin (The New York Times 2024a) — swing states that have a significant share in electoral college votes). In light of this, responses were also weighted to consider over- and under-represented voters to ensure that each demographic is adequately represented.

This poll employs a random sampling approach of registered voters using telephone interviews, with a focus on both landlines and cellphones. As such, this generally allowed for a representative sample. In addition, NYT/Siena also used voter registration files help ensure proper balance between political parties, and telephone polls have historically proven to be effective in gauging public opinion from recent elections. However, one particular drawback is that telephone response rates are extremely low; in fact, for this poll, "fewer than 2% of contacted individuals participate in the survey."

In handling non-responses within their questionnaire, for some questions another "subquestion" related to the original is asked only once in order to get a definitive answer; this usually occurs in questions related to leaning (not outright) candidate support. By doing this, it allows NYT to more accurately gauge the public's opinions. However, at times, they leave non-responses as "Refused" or "Don't know" often on questions targeted to the individual person (e.g. How has _____ affected you personally?). At this point, there is little no use prodding further in case of a non-response. The overall questionnaire is of high quality (NYT/Siena is one of the most reputable polls after all), and it has its strong and weak points. First, most questions, often those on respondent identification or demographic, are unbiased and fair; in particular, questions on candidate preferences did not contain words suggestive of one candidate over another. However, some questions contain gender-coded language that may skew respondents' answers in favor of one candidate (e.g. In the question

"cares more about people like you" the word "cares" is more often associated with females and so respondents are more likely to choose Harris as the answer, which was seen in the poll results with Harris receiving 49% of responses versus Trump's 41%).

D Appendix B: Idealized Methodology and Survey

D.1 Sampling Approach and Strategy

In predicting U.S. presidential election outcomes through polling, an ideal methodology would utilize stratified sampling to ensure a representative sample of the electorate. This approach would survey registered voters from all states, then the sample will be adjusted proportionally, reflecting key demographic factors such as education, race, gender, and the urban-to-rural population ratio. To achieve this, voter registration records and census data would serve as the basis for selecting respondents. Telephone polling would be employed, with calls made at different times of the day to mitigate potential biases from non-response patterns, as certain groups of individuals may be more reachable at specific times. Given that modern telephone polling often encounters low response rates, this methodology assumes a 10% response rate, consistent with findings, highlighting the challenges in obtaining representative samples from telephone-based surveys (Center 2017). Therefore, around 15,000 individuals will be contacted to obtain a desired sample size of 1500. Despite the low response rate, carefully deploying methods should help mitigate biases and improve the accuracy of election predictions.

D.1.1 Stratification Variables [TO-DO]

- **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

D.2 Budget Allocation

- **Marketing and Outreach (Ads and Incentives)**: $50,000
- **Survey Design and Data Validation**:
  - Questionnaire Development: $5,000
  - Pilot Testing (survey run on small group of people): $5,000
  - Google Forms (with built-in validation rules): $5,000

- **Data Analysis and Software**:
  - Statistical Software: $5,000
  - Staff Fees (Analysts and Statisticians): $15,000
- **Financial Incentives**: $2,000
- **Miscellaneous Expenses and Contingency Fund**: $8,000

D.2.1 Survey Structure:

The survey will begin by introducing each potential participant to the topic and providing necessary context about the purpose of the research. This introduction aims to make the participants feel informed and comfortable with the process. The first question will ask whether the individual is willing to participate in the survey, emphasizing that participation is voluntary. It will also be clearly communicated that respondents may drop out of the survey at any point if they choose. Once participants agree, they will proceed to answer a series of demographic questions that will help ensure a diverse and stratified sample. These questions include:

- "What is your gender?"
- "What is your age?"
- "Which city do you live in?"
- "What race do you identify as?"
- "What is your highest degree of education?"
- "Will you be voting in the upcoming election?"
- "Which candidate will you be voting for?"

D.3 Data Validation

References

Center, Pew Research. 2017. "What Low Response Rates Mean for Telephone Surveys." 2017. https://www.pewresearch.org/methods/2017/05/15/what-low-response-rates-mean-for-telephone-surveys/.

FiveThirtyEight. 2024. *FiveThirtyEight: 2024 US Presidential Election Polls*. https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

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

RealClearPolitics. 2024. "RealClearPolitics Polling Data." https://www.realclearpolitics.com/epolls/latest_polls/.

The New York Times. 2024a. "Times/Siena Poll Methodology." https://www.nytimes.com/article/times-siena-poll-methodology.html.

———. 2024b. "Times/Siena Poll: Florida Toplines." https://www.nytimes.com/interactive/2024/10/13/us/elections/times-siena-poll-florida-toplines.html.

———. 2024c. "Times/Siena Poll: Likely Electorate Crosstabs." https://www.nytimes.com/interactive/2024/10/13/us/elections/times-siena-poll-likely-electorate-crosstabs.html.