2024 US Election Prediction*

An Analysis of Donald Trump's Polling Trends and Predictive Outcomes

Betty Liu Jingchuan Xu Dingshuo Li

November 4, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

The 2024 U.S. presidential election is one of the most closely watched events in recent political history. Polling provides a glimpse into the dynamics of the race, offering insights that can shape campaign strategies and public opinion. Among the candidates, former President Donald Trump consistently draws significant attention and debate from the public across different political perspectives. Tracking Trump's support through the poll reveals his current popularity and potential electoral outcomes. The paper dives into Trump's polling data, analyzing data from various pollsters to see how his support has changed over time. The study explores whether there is an upward or downward trend using linear modelling in Trump's polling percentage, while also accounting for differences in his support reported by various pollsters. By investigating these patterns, the analysis aims to estimate Trump's likely level of support as election day approaches.

Estimand: The estimand of this study is the expected level of voter support for Donald Trump in the 2024 U.S. presidential election, based on observed patterns in polling data. This estimated support level provides a snapshot of his projected standing among voters with the adjective by pollster-specific effects.

In this study, we apply statistical models, including a linear model of Trump's percentage over time, to quantify changes in his support level. Additionally, we adjust for pollster variations to provide a more accurate reflection of his support trajectory across different sources. Our findings reveal both temporal trends and variability among pollsters, highlighting the complexities of interpreting polling data in a polarized political environment. These insights contribute to a better understanding of Trump's standing and the potential outcomes in 2024.

*Code and data are available at: https://github.com/dawsonlll/2024_US_Election_Analysis

The remainder of the paper is structured as follows: Section 2 outlines the dataset selection and cleaning process to ensure transparency in data preparation. Section 3 introduces the model used, explaining its components with clear notation and connecting modeling decisions to the data section with focusing on linear models and pollster-specific adjustments. Section 4 presents the results of the analysis, including both temporal and cross-pollster trends in Trump's support. Finally, Section 5 interprets the results, and Section 6 addresses study limitations and future research directions.

2 Data

2.1 Overview

The dataset, sourced from FiveThirtyEight (2024), compiles polling data from multiple pollsters who applied various methodologies—such as online panels, live phone surveys, and text-to-web approaches—to gauge voter support for presidential candidates in the 2024 U.S. election. Each entry captures a distinct polling event, with details on the pollster, methodology, sample size, and public support percentages for each candidate, specifically focusing on the public's stance toward Donald Trump following his declaration to run for president. This dataset provides a snapshot of public opinion across different pollsters, methodologies, and timeframes, offering insights into the shifting landscape of voter sentiment.

All data analysis was conducted using R, including statistical computing and graphics. And the following R packages were used: tidyverse Wickham et al. (2019), Palmerpenguins Horst, Hill, and Gorman (2020), ggplot2 Wickham (2016), Dplyr Wickham et al. (2023), Knitr Xie (2023), modelsummary Arel-Bundock (2022), arrow Richardson et al. (2024), and here Müller (2020). Following Alexander (2023a) we conducted data simulate, test simulated data, data cleanning, test analysis data, EDA, data modelling. The R code in scripts were adapted from Alexander (2023b)

For accuracy and relevance, we filtered the data to include only polls conducted by high-quality pollsters with a numeric grade greater than 2.7. This threshold ensures that the dataset primarily represents credible polls, aligning with industry standards for reliability and transparency. Additionally, we factored in only polls conducted after Trump and Harris' 2024 presidential campaign announcement, allowing a focused analysis of his support trajectory post-announcement. Through these data selection and cleaning criteria, the dataset provides a high-quality, methodologically consistent foundation for examining voter support trends for Trump and Harris in the 2024 election.

2.2 Measurement

The measurement approach in this dataset translates real-world polling events into structured data entries, capturing shifts in public opinion on presidential candidates. Each entry represents a specific poll by a polling organization, providing a snapshot of support levels for the 2024 U.S. presidential election. Polls are conducted through varied methodologies—such as online panels and live phone surveys—which can impact reliability. Polls are sponsored by institutions like news organizations, with the pollster and sponsor information recorded for transparency and credibility.

Key columns include poll scores and candidate-specific percentages, translating public sentiment into measurable values that reflect candidates' standings over time. Sample size and population type give insight into the scope of each poll, helping to contextualize its representativeness. Temporal data, such as start and end dates, allow us to analyze trends over time, correlating shifts in support with major events.

Together, these components ensure each dataset entry reflects a distinct polling event, with variables like pollster reputation and methodology contributing to an accurate picture of public sentiment. This structured framework enables reliable trend analysis, supporting meaningful insights into polling data for presidential candidates.

2.3 Outcome variables

The primary outcome variable in this dataset is the approval rating (pct), which represents the percentage of poll respondents who support Donald Trump as a presidential candidate. This variable provides insight into Trump's popularity over time, helping to reveal changes in public sentiment across states and voting periods. The summary of table for both Trump and Harris can be observed Table 1.

Table 1: Summary Statistics for Trump and Harris

Candidate	Mean %	Median %	Min %	Max %	SD %
Trump	45.15432	46	19	70.0	5.089919
Harris	47.80643	48	25	65.3	3.695899

To understand the factors that influence this outcome, we include several predictor variables. The pollster variable identifies each polling organization, acknowledging that different pollsters may produce slightly different results due to unique methodologies and respondent demographics. Numerical poll scores provide a quality rating for each poll, with higher scores indicating greater credibility and predictive reliability. These variables ensure that only reliable sources influence our analysis of Trump's approval rating. The methodology and transparency scores reflect the transparency of each poll's survey methodology and reporting, respectively. Different methodologies, such as online panels or telephone surveys, may affect response patterns, while transparency scores indicate reliability, providing further context for accurately interpreting support levels.

Time and region data are also critical in our analysis. The start and end dates provide a timeline of the polling period, allowing us to observe trends in public opinion as it change over time. The state variable captures the geographic focus of each poll, as table Table 2 shows which help us identify regional differences in Trump's support. Finally, the candidate name variable specifies Trump as the focal candidate, ensuring that we accurately measure support for his candidacy alone.

Table 2: Summary Statistics for Trump and Harris

State	Harris %	Trump %
Arizona	46.78	47.66
California	61.95	34.01
Connecticut	53.00	37.00
Florida	44.09	50.57
Georgia	47.28	47.86
Indiana	40.70	53.40
Iowa	43.00	47.88
Maine	51.00	43.00
Maine CD-1	58.00	37.00
Maine CD-2	44.00	49.00
Maryland	64.65	33.02
Massachusetts	61.08	30.21
Michigan	48.01	45.97
Minnesota	50.38	42.52
Missouri	42.20	52.65
Montana	39.59	54.48
National	47.22	43.59
Nebraska	39.14	53.48
Nebraska CD-1	43.00	51.00
Nebraska CD-2	51.33	42.11
Nebraska CD-3	25.00	70.00
Nevada	48.20	47.01
New Hampshire	50.71	41.61
New Mexico	51.15	43.15
New York	54.72	36.40
North Carolina	48.04	47.98
Ohio	44.70	49.45
Pennsylvania	48.32	46.48
Rhode Island	54.50	41.50
South Dakota	36.50	55.87
Texas	44.28	49.08
Virginia	50.20	42.92

Table 2: Summary Statistics for Trump and Harris

State	Harris %	Trump %
Washington	57.00	36.00
Wisconsin	48.98	46.14

```
echo=TRUE
library(rstanarm)

#model_date <-
  #readRDS(file = here::here("models/model_date.rds"))
#model_date_pollster <-
   #readRDS(file = here::here("models/model_date_pollster.rds"))
```

3 Model

```
#modelsummary(models = list("Model 1" = model_date, "Model 2" = model_date_pollster))
```

3.0.1 Model justification

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

Our results are summarized in

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

6 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

Polling Methodology for CBS News/YouGov Survey (October 11-16, 2024)

1. Population, Frame, and Sample This CBS News/YouGov survey took place from October 11-16, 2024, with 1,439 registered voters in Arizona. The survey focused on registered voters in Arizona, and the sample was weighted to match key demographics like gender, age, race, and education. The weights were based on data from the U.S. Census American Community Survey, the U.S. Census Current Population Survey, and voter turnout data from the 2020 Presidential election.

2. Sample Recruitment The recruitment process focused on including respondents representative of Arizona's registered voter population by adjusting for demographic factors such as age, race, gender, and education. The sample was recruited primarily from various online panels, which included a mixture of respondents across demographic lines, to ensure a representative sample: 1,152 respondents were selected from YouGov's online panel. 212 respondents from Pure Spectrum's panel. 49 respondents from Dynata. 17 respondents from Cint's panel. 9 respondents from ROI Rocket's panel.

Surveys were conducted in both English and Spanish to account for language preferences among respondents. The weights applied to the data ranged from 0.1 to 5.0, with a mean of 1 and a standard deviation of 0.8, ensuring that the sample was representative of Arizona's voting population.

3. Sampling Approach and Trade-offs Sampling Approach: This survey employed stratified random sampling and applied post-survey weighting to ensure accurate representation of key demographic groups. Stratified random sampling is a technique that divides the overall population into several subgroups (or strata) based on specific attributes such as age, gender, race, and education level. Random samples are then drawn from each subgroup. The advantage of this method is that it ensures adequate representation across each stratum, preventing certain groups from being underrepresented or overlooked in the sample. After the survey was completed, weighting was applied to further adjust the sample to match the demographic distribution of registered voters in Arizona. This process involved adjusting the weights of individuals in the sample to better reflect the true composition of the overall voter population. This allows for a more accurate capture of how different demographic factors (such as gender, age, race, and education) influence voting behavior, thereby improving the external validity of the results. By using stratified random sampling and weighting, the researchers were able to minimize sampling bias and increase the accuracy of predicting voter tendencies. Trade-offs: One of the main limitations of this sampling method is its reliance on online panels, which may exclude individuals without internet access, thus introducing selection bias. While online surveys are cost-effective and convenient for collecting large samples, this reliance may result in certain groups (such as older individuals, low-income households, or voters living in remote areas) being left out due to lack of internet access. This means that some groups

might be underrepresented in the sample, which can affect the overall representativeness of the results. Although the weighting process can help adjust the sample to better reflect demographic differences, some errors are still difficult to completely eliminate. For example, when filling out surveys, respondents might overstate or understate their voting intentions due to social pressure or personal emotions—this is known as self-reporting bias. Even after weighting adjustments, such biases may persist and impact the accuracy of the final survey results. Therefore, while weighting can improve the representativeness of the results to a certain extent, systematic biases like non-response bias or selection bias may still leave traces in the final outcomes. Researchers need to interpret these potential errors with caution.

4. Regression Model A regression model was used to estimate each respondent's likelihood of voting. This model combined self-reported voting intentions with demographic and historical voting data, such as: Age, gender, race, and education. Voting history from past elections. The regression model allowed the survey to distinguish "likely voters" from the broader pool of registered voters, improving the accuracy of the predictions. By analyzing both individual and aggregate data, the model offered a more reliable estimate of actual voter turnout, thus increasing the precision of the survey's results.

5. Handling Non-response In surveys, non-response can cause bias because these people might have different voting behaviors or opinions. To fix this potential bias, researchers use weighting to adjust the sample data. Specifically, they assign different weights to respondents based on key demographic factors like gender, age, race, and education. The main goal of this weighting process is to make sure that even if some voters didn't respond, the final sample still accurately represents the overall population of registered voters in Arizona. This adjustment helps make the sample more representative and reduces the bias caused by non-response, improving the reliability and accuracy of the survey results. Besides that, weighting helps balance the proportion of different groups in the sample, ensuring that certain groups (like those with less access to the internet) are properly reflected in the results. In the end, this process helps make sure the survey results are more valid and can be applied more effectively to predict real voter behavior.

6. Strengths and Weaknesses of the Questionnaire Strengths: This survey effectively covered the key issues that Arizona voters care about the most, such as the economy, immigration, abortion, and the state of democracy. By combining demographic factors and historical voting data, the reliability of the results was improved, making it more accurate in reflecting the voting preferences of different groups. Weaknesses: Since the survey relies on self-reported voting intentions, this might introduce some bias, as respondents could overestimate or underestimate their likelihood of voting. Also, because it depends on online panels, voters without internet access may have been excluded, which could affect the external validity of the results. The margin of error is ±3.3 points, showing that there's still some uncertainty in the findings.

7. Margin of Error The margin of error for this survey is ±3.3 points, within a 95% confidence interval. The formula to calculate the margin of error is: $\hat{p} \pm 100 \times \surd((1+CV\hat{}2)/n)$

Where CV is the coefficient of variation of the sample weights, and (n) is the sample size. This formula calculates the sampling error, meaning that 95% of the sample results should fall within this range. It's important to note that this margin doesn't account for non-sampling errors, such as biases from panel selection or respondent behavior.

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows... This suggests...

?@fig-stanareyouokay-2 is a Rhat plot. It shows... This suggests...

References

Alexander, Rohan. 2023a. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

———. 2023b. *Telling Stories with Data: With Applications in r*. Chapman; Hall/CRC.

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

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

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. https://doi.org/10.5281/zenodo.3960218.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.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.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.