Forecasting the 2024 U.S. Presidential Election*

Jason Yang Another author

November 3, 2024

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2….

2 Data

2.1 Overview

The dataset used for this analysis was sourced from FiveThirtyEight's poll aggregation platform, which compiles high-quality polling data from various reputable pollsters across the U.S. (FiveThirtyEight (2024)). This analysis uses the statistical programming language R (R Core Team (2023)) and various libraries for data manipulation and visualization, including tidyverse for efficient data wrangling and ggplot2 for creating insightful visualizations (Wickham et al. (2019), Wickham (2016)). By employing robust statistical methods and tools, this analysis seeks to forecast potential outcomes of the 2024 U.S. Presidential Election.

Overview text

*Code and data are available at: [https://github.com/Jerryx2020/US_election_prediction/tree/main].

2.2 Measurement

2.3 Outcome variables

In this analysis, the outcome variable of interest is the percentage of support a candidate receives in each poll, represented by pct. This variable indicates the proportion of respondents within a poll who favor a particular candidate, measured as a percentage. Given that pct serves as a continuous variable bounded between 0 and 100, it is well-suited for modeling the levels of support across different polls and over time.

2.4 Predictor variables

To accurately predict the 2024 Presidential Election, we have identified several key variables that capture a range of factors influencing voter behavior. These predictor variables include:

- Polling Percentage (PCT): Represents the percentage of respondents in a poll who support a specific candidate
- Numeric Grade (numeric_grade): Represents the rating of the pollster to indicate their quality
- Sample Size (sample_size): size of the respondents participating in the poll
- State (state): the state of the poll was taken in
- End_date: the day the poll ended
- Start_date: the day the poll started

3 Model

This analysis employs predictive models to estimate the anticipated support for the two major candidates in the 2024 U.S. Presidential Election: a multiple linear regression model. These models use high-quality polling data from reputable sources, with relevant predictors to provide insights into public opinion trends.

3.1 Model set-up

The multiple linear regression model is designed to predict the candidate's percentage of support (pct) as a linear function of various predictors. Formally, we define the model as:

$$\text{pct}_i = \beta_0 + \beta_1 \cdot \text{numeric\_grade}_i + \beta_2 \cdot \text{sample\_size}_i + \sum_{k=1}^{K} \delta_k \cdot \text{state}_{ik} + \beta_4 \cdot \text{end\_date}_i + \epsilon_i,$$

- pct: The percentage of the vote or support that the candidate received in the poll (e.g., 51.0 for Kamala Harris, 48.0 for Donald Trump).
- numeric_grade: A numeric rating given to the pollster to indicate their quality or reliability (e.g., 3.0).
- sample_size: The total number of respondents participating in the poll (e.g., 2712).
- state: The U.S. state where the poll was conducted or focused, if applicable.
- end_date: The date the poll ended (e.g., 10/11/24).
- ϵ_i: The error term for poll accounting for variability not explained by the predictors.

3.1.1 Model justification

Predictors Selection: Variables such as sample_size, pollster, and state are included to capture systematic differences in polling methods and geographic variations in support. The date (end_date) of each poll controls for temporal effects on candidate popularity.

Linear regression was chosen for this analysis due to several key advantages. First, its simplicity and interpret ability make it highly suitable for examining the relationship between predictor variables and the response variable, specifically the candidate support percentage. Each coefficient in a linear regression model clearly represents the influence of a predictor on the outcome, which facilitates straightforward interpretation and communication of results. Moreover, linear regression is computationally efficient, making it ideal for rapid model estimation, tuning, and testing, even on relatively large datasets.

3.2 Regression Assumptions

- Linearity: The relationship between each predictor and the response (pct) is assumed to be linear.
- Normality: Errors are assumed to follow a normal distribution, which is tested during model diagnostics.

3.2.1 Checking Assumptions

In the Residuals vs. Fitted, there is no clear curved pattern, which suggests that the linear relationship might be appropriate. However, there are some clusters and a slight spread, which may indicate minor issues in capturing the relationship fully. Secondly, in the qqnorm plot, the points closely follow the diagonal line, indicating that residuals are approximately normal. This supports the normality assumption of the residuals, which is essential for reliable inference in linear regression.

3.3 Results

4 Discussion

4.1 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.

4.2 Second discussion point

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

4.3 Third discussion point

4.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Pollster Analysis

The Marquette Law School Poll employs a rigorous methodology to survey adults in the US on national issues, incorporating probability-based sampling that combines Address-Based Sampling (ABS) and Stratified Random Sampling (David Rodbart 2024). This approach, including online and offline respondents, ensures broad representativeness. The SSRS Polling Panel supports participant recruitment, and weighting adjustments enhance result accuracy.

A.1 What is the population, frame, and sample?

The population for the poll includes adults aged 18 and over across all 50 states and the District of Columbia. The sampling frame was constructed through ABS, leveraging the US Postal Service's Computerized Delivery Sequence (CDS), a comprehensive, regularly updated list of residential addresses. The primary source for participants is the SSRS Opinion Panel, which includes both ABS and telephone survey data to capture hard-to-reach demographics, like non-internet users. The survey ultimately included 1,005 adults, including registered and likely voters, with weighted samples to ensure proportional representation (David Rodbart 2024).

A.2 How is the sample recruited?

Participants were recruited via the SSRS Omnibus Panel, a probability panel formed from ABS and telephone survey data. This dual approach ensured that both internet and non-internet users were included, minimizing online-only bias. Stratified by demographics (age, gender, race, education, region, and partisan affiliation), this process produced a balanced representation across groups. Selected participants received an email invitation with a unique link to the survey. Non-responders received up to two reminder emails, and those who opted for text notifications also received reminders. Participants received an e-gift card incentive upon survey completion (David Rodbart 2024).

A.3 What sampling approach is taken, and what are some of the trade-offs of this?

The Marquette Law School Poll combines ABS and stratified random sampling. The ABS covers all residential addresses using USPS delivery sequences, ensuring broad geographic reach. To include hard-to-reach demographics, SSRS supplements ABS data with telephone survey responses. Stratified sampling by key demographics (age, gender, race, education, region, and political party) ensures representation across major groups (David Rodbart 2024). This

probability-based sampling minimizes bias but requires significant resources. Stratification and weighting to represent specific populations add to survey complexity and cost.

A.4 How is non-response handled?

To address non-response and improve representativeness, SSRS applies weighting adjustments based on demographic variables such as age, gender, race, education, region, and internet use (David Rodbart 2024). This process aligns sample characteristics with adult population parameters, reducing bias from demographic differences in response rates.

A.5 What is good and bad about the questionnaire?

The questionnaire is well-designed, with a collaborative approach that reduces respondent burden and improves comprehension. A thoughtfully structured skip pattern enhances flow and minimizes dropout rates. The probability-based sampling, combined with extensive weighting and stratification, enables a representative sample of adults. However, heavy reliance on web-based responses may introduce internet bias, potentially underrepresenting non-internet households. Additionally, the questionnaire's complexity may increase participant fatigue, affecting response completeness. The response rate of approximately 47.4% also indicates some level of attrition in the sample (David Rodbart 2024).

A.6 Idealized Methodology

In designing a forecast for the upcoming US presidential election with a $100,000 budget, our methodology focuses on creating a representative survey that captures the voting intentions of likely voters across the United States. The objective is to gather reliable, high-quality data that supports a well-informed forecast through effective sampling, data validation, and aggregation techniques.

Our target population is voters aged 18 and above, with a sample size of approximately 10,000 responses, 200 responses from each state, to ensure a statistically significance at the national level and within key demographic groups. To achieve balanced representation, we will employ stratified random sampling based on age, gender, race/ethnicity, education, region, and urban/rural status. This approach will involve weighting responses using recent census data and voter turnout statistics to address any demographic discrepancies.

For recruitment, we will combine online survey panels, social media outreach, and targeted telephone recruitment. These methods are chosen to increase accessibility for harder-to-reach demographics, such as older adults and rural residents. To encourage participation, we will offer modest incentives, such as a small monetary compensation or entry into a prize draw. All data collection will be conducted through a reliable survey platform—Google Forms.

The survey itself will be concise and straightforward, focusing on voting intentions, key issues, and demographic information. For data validation, the survey will include attention-check questions as respondents are very likely to complete the form for an entry into a prize draw. We will also conduct multiple waves of the survey, we will adopt a "poll-of-polls" approach, aggregating responses from different waves while applying weights based on relevance.

To mitigate biases such as non-response bias, social desirability bias, and order bias, our team will implement several strategies. We will send reminders to non-respondents to encourage participation, emphasizing the prize draw incentives to increase response rates. Survey questions will be carefully crafted to maintain neutral wording and avoid sensitive topics that might discourage honest responses. Additionally, within each wave of polls, we will randomize the order of questions to minimize the impact of question order on responses. These measures aim to enhance the reliability and accuracy of our survey results.

Ethically, our approach prioritizes respondent privacy and data security. We will ensure that participants' responses remain anonymous and that all data is handled securely. Transparency is also essential; respondents will be informed about the survey's sponsorship, methodology, and data use.

The survey will be implemented on Google Forms, with a link provided in the appendix. We also included the questions in the appendix for better understandings.

A.6.1 Budget Allocation

-

A.7 Idealized Survey

The proposed survey is designed using Google Forms, https://forms.gle/Vhmm1b1XnqTSoBWq8

A.8 Survey Copy

Welcome to our survey! We are conducting research to understand public perceptions and opinions related to the upcoming 2024 Presidential Election. Your responses are valuable in helping us gain insight into voter perspectives and preferences. By participating, you will gain the chance to enter a prize draw. Prizes include a chance to win gift cards, exclusive merchandise, and special deals with our sponsors.

The survey should take about 10-15 minutes to complete.

Please note:

- Your Privacy are our top concern. All responses are confidential and will be used solely for research purposes.
- Participation is entirely voluntary.

Contact Information

If you have any inquiries or need assistance with this survye, please contact us at:

- jzc.yang@mail.utoronto.ca
- 123-456-7891

1. **Are you eligible to vote in the upcoming 2024 U.S. Presidential Election?**

- Yes
- No
- Not sure

2. **Who do you intend to vote for in the 2024 U.S. Presidential Election?**

- Kamala Harris (Democratic Party)
- Donald Trump (Republican Party)
- Undecided
- Other:

3. **How certain are you about your choice?**

- Very Certain
- Somewhat certain
- Unsure

4. **Which of the following issues are most important in deciding your vote? (Select up to 3)**

- Economy and jobs
- Healthcare
- Climate Change
- Education
- Immigration
- National Security
- Social justice and Equality
- Abortion rights
- Gun control
- Other:

5. **What is your age?**

- 18-24

- 24-34
- 35-44
- 45-54
- 55-64
- 65 or older

6. **What is your gender?**

- Male
- Female
- Non-binary
- Prefer not to say
- Other:

7. **What is your race/ethnicity? (Select all that apply)**

- White
- Black or African American
- Hispanic or Latino
- Asian
- Native American or Alaska Native
- Native Hawaiian or Other Pacific Islander
- Other:

8. **In which U.S. state do you currently reside?**

- [Drop Down With All the States]

9. **If you voted in 2020, who did you vote for?**

- Joe Biden
- Donald Trump
- Other:

10. **Source of Information: What are your primary sources for political news?**

- Television
- Newspapers/Magazines
- Online News Websites
- Social Media
- Radio
- Friends and Family
- Other (Please specify): _____

11. **What is your highest level of education?**

- High school graduate, equivalent, or less
- College Degree
- Associate degree
- Bachelor's degree
- Graduate degree

12. **Would you be willing to participate in follow-up surveys regarding the election?**

- Yes
- No

13. **Do you have any additional comments or concerns about the 2024 election?**

- [Open Text Response]

Thank you for taking the time to complete our survey. Your input is greatly appreciated and will contribute to our understanding of public opinion for the upcoming 2024 Presidential Election.

References

David Rodbart, Julia Dalagan. 2024. *Marquette Law School Poll Methodology Statement*. 1215 W Michigan St, Milwaukee, WI 53233, United States: Marquette University Law School. https://law.marquette.edu/poll/wp-content/uploads/2024/10/MLSPSC22Methodology.pdf.

FiveThirtyEight. 2024. *FiveThirtyEight 2024 u.s. Presidential Election Poll Aggregation*. New York, USA: FiveThirtyEight. 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/.

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.