Forecasting the 2024 U.S. Presidential Election*

Jason Yang   Another author

November 2, 2024

This paper presents a multiple linear regression model and a bayesian model for forecasting the outcome of the 2024 U.S. Presidential Election. Using the data collected from high-quality polls from various pollsters, we generate these models by selecting relevant variables such as sample size, percentage support for each candidate, and sate. By comparing the predictive power of both models, we aim to provide a reliable forecast of the election outcome and a deeper understanding of public sentiment and polling reliability.

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

---

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

((**R?**)) and various libraries for data manipulation and visualization, including tidyverse for efficient data wrangling and ggplot2 for creating insightful visualizations (Wickham et al. (2019), (**ggplot2?**)). By employing robust statistical methods and tools, this analysis seeks to forecast potential outcomes of the 2024 U.S. Presidential Election.

Overview text

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

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

3 Model

This analysis employs two 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:

$$\text{pct}_i = \beta_0 + \beta_1 \cdot \text{numeric\_grade}_i + \beta_2 \cdot \text{sample\_size}_i + \beta_3 \cdot \text{transparency\_score}_i$$

$$+ \sum_{k=1}^{K} \delta_k \cdot \text{state}_{ik} + \beta_4 \cdot \text{end\_date}_i + \epsilon_i,$$

- 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).
- transparency_score:A score reflecting the pollster's transparency about their methodology (e.g., 9.0). "A grade for how transparent a pollster is, calculated based on how much information it discloses about its polls and weighted by recency. The highest Transparency Score is 10."
- 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

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 **?@tbl-modelresults**.

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

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 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, in the survey, we 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.

B.2 Idealized Survey

The proposed survey is designed using Google Forms, https://docs.google.com/forms/d/e/1FAIpQLSeHHeKaryN ZgXwQYkNw9jBdxIbgeXWVj_t1ob2w/viewform?usp=sf_link

B.3 Survey Copy

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

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

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

- Very Certain
- Somewhat certain
- Unsure

3. **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:

4. **What is your age?**

- 18-24
- 24-34
- 35-44
- 45-54
- 55-64
- 65 or older

5. **What is your gender?**

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

6. **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:

7. **In which U.S. state do you currently reside**

- _____

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

- Joe Biden
- Donald Trump
- Other:

References

FiveThirtyEight. 2024. *FiveThirtyEight 2024 u.s. Presidential Election Poll Aggregation.* New York, USA: FiveThirtyEight. https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

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.