Forecasting the 2024 U.S. Presidential Election*

Jason Yang Jerry Xia Peter Fan

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
- Pollster: Represents the name of the polling organization that conducted the poll.
- Candidate: The name of the candidate in the poll
- 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:

$$\text{pct}_i = \beta_0 + \beta_1 \cdot \text{pollster}_i + \beta_2 \cdot \text{sample\_size}_i + \sum_{k=1}^{K} \delta_k \cdot \text{state}_{ik} + \beta_4 \cdot \text{end\_date}_i + \beta_5 \cdot \text{candidate}_i + \epsilon_i,$$

Where

- β_0 is the expected value of PCT when all predictor variables are zero, intercept
- β_1 captures the effect pollsters have on the response variable
- β_2 captures the effect sample_size have on the response variable
- δ_k captures the difference in PCT between states
- β_4 captures the expected change in PCT for each additional unit increase in the end date variable
- β_5 captures the difference in PCT associated with difference candidates

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

Appendix

A Pollster Analysis

The Marquette Law School Poll utilizes a rigorous probability-based methodology to conduct a national question survey of adults across the United States. This method combines address-based sampling (ABS), which uses the U.S. Postal Service's Computerized Delivery Sequence (CDS), and stratified random sampling, which covers virtually all residential addresses and is able to reach the offline population, thereby reducing the bias associated with surveys conducted solely via the Internet (David Rodbart 2024). The stratified random sampling method divides the population into subgroups (e.g., age, gender, education), to make sure that each subgroup is proportionately represented in the sample. This structure improves representativeness and accuracy, thereby supporting valid inferences about the total U.S. adult population.

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

The population surveyed by the Marquette Law School Poll includes all U.S. adults aged 18 and older, living in the 50 states and D.C. The sampling frame was developed using Address-Based Sampling (ABS) by using the U.S. Postal Service's Computerized Delivery Sequence (CDS) file—a regularly updated, near-complete listing of all residential addresses in the U.S., excluding business addresses. This frame can cover almost every U.S. household, including non-internet and traditionally hard-to-reach areas (David Rodbart 2024).

The sample itself was drawn from the SSRS Opinion Panel, a probability-based panel that combines participants from ABS and a bilingual telephone survey conducted through the SSRS Omnibus. This dual recruitment strategy, which includes participants from both internet and non-internet households, improves demographic inclusivity. To achieve accurate representation, the survey sampled 1,005 adults, including registered and likely voters, and applied post-stratification weighting to adjust for demographic imbalances. This weighting aligns the sample with key population parameters, ensuring that the poll's findings are as representative as possible of the broader adult population.

A.2 How is the sample recruited?

Marquette Law School Poll using the SSRS Opinion Panel for the recruitment, a probability-based panel integrating Address-Based Sampling (ABS) and telephone survey data from the SSRS Omnibus, a nationally representative bilingual survey platform. This dual recruitment approach ensures that both internet and non-internet users are included, addressing the online-only bias often seen in web-based surveys and thereby enhancing sample inclusivity (David Rodbart 2024).

The Marquette Law School Poll uses stratified sampling to obtain a balanced and representative sample by differentiating participants based on key demographic variables such as age, gender, race, education, region, and political affiliation. The poll then ensures proportional representation of key demographic groups, which reduces the risk of sampling bias and increases the generalizability of the results.

Selected participants received a unique survey link through email invitations. Non-responders were sent up to two reminder emails, and, for those who opted in, a text notification reminder. To encourage completion and maintain high response rates, participants were offered an e-gift card as an incentive. This use of reminders and incentives is critical in minimizing non-response bias, thereby strengthening the reliability of the poll's findings.

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

The Marquette Law School Poll uses both sampling probability sampling methods, address-based sampling (ABS) and stratified random sampling, this can increase representativeness and inclusiveness. ABS uses the U.S. Postal Service's Computerized Delivery Sequence (CDS), which covers nearly all residential addresses in the U.S., including homes with no Internet access. This approach is critical to minimizing the coverage bias associated with web-only surveys (David Rodbart 2024).

The poll also applies stratified random sampling to ensure that key demographic groups—such as age, gender, race, education, region, and political affiliation—are proportionally represented within the sample. Samples are taken from each subgroup once the population is divided into subgroups using stratified sampling. Consequently, sampling error is reduced and the sample findings accurately represent the views of each subgroup.

While this approach increases the reliability and representativeness of the findings, it also demands significant resources. ABS requires comprehensive address data and ongoing maintenance, while stratification and post-stratification weighting add complexity to both survey design and data processing. These resource-intensive processes make ABS and stratified random sampling more costly than simpler sampling methods but are justified by the improved accuracy and generalizability of the poll's results.

A.4 How is non-response handled?

The Marquette Law School Poll applies weighting adjustments across demographic factors, including age, gender, race, education, region, and internet access, aligning the sample's characteristics with those of the broader adult population (David Rodbart 2024). This can represent the under-represented groups more accurately and increase the weight of these respondents. This process enhances the survey's representativeness, ensuring findings reflect the national population reliably despite disparities in response rates.

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

The Marquette Law School Poll is designed to boost respondent engagement and minimize survey fatigue. Using skip patterns that adapt questions based on previous answers, the survey reduces cognitive load, streamlines flow and lowers dropout rates by customizing the survey experience (David Rodbart 2024). This probability-based approach, supplemented by comprehensive weighting and stratification, further makes sure that the results accurately represent the U.S. adult population, with clear and straightforward questions that enhance accessibility across demographic groups.

However, despite ABS and telephone recruitment efforts, the reliance on web-based responses can cause some internet bias possibly underrepresenting households with no internet access. Additionally, the survey's length and complexity could add to respondent fatigue, potentially impacting response quality as participants progress. The response rate of approximately 47.4% reflects moderate attrition, typical in public opinion research, highlighting the necessity of post-stratification adjustments to address any demographic imbalances from non-response.

B 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 statistical 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.

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.

To ensure data validity, IP addresses would be recorded to prevent multiple submissions from the same source. Additionally, cookies would be implemented to prevent users from retaking the survey on the same device. Finally, we will verify respondents' email addresses to ensure uniqueness.

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.1 Budget Allocation

- Sampling and Data Collection (Advertisements and Recruitment Platforms): $40,000
- Participant Incentives (Prize Pool): $20,000
- Survey Design and Testing: $20,000
- Data Analysis and Validation: $20,000

C Idealized Survey

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

C.1 Survey Copy

2024 U.S. Presidential Election Voter Intention Survey

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

* Indicates required question

1. Email *

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

Mark only one oval.

- ◯ Yes
- ◯ No
- ◯ Other: _____

3. What is your age?

Mark only one oval.

- ◯ 18-24
- ◯ 25-34
- ◯ 35-44
- ◯ 45-54
- ◯ 55-64
- ◯ 65 or older

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

Check all that apply.

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

5. What is your gender?

Mark only one oval.

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

6. In which U.S. state do you currently reside?

Mark only one oval.

- Alabama
- Alaska
- Arizona
- Arkansas
- California
- Colorado
- Connecticut
- Delaware
- Florida
- Georgia
- Hawaii
- Idaho
- Illinois
- Indiana
- Iowa
- Kansas
- Kentucky
- Louisiana
- Maine
- Maryland
- Massachusetts
- Michigan
- Minnesota
- Mississippi
- Missouri
- Montana
- Nebraska
- Nevada
- New Hampshire
- New Jersey

- New Mexico
- New York
- North Carolina
- North Dakota
- Ohio
- Oklahoma
- Oregon
- Pennsylvania
- Rhode Island
- South Carolina
- South Dakota
- Tennessee
- Texas
- Utah
- Vermont
- Virginia
- Washington
- West Virginia
- Wisconsin
- Wyoming

7. What is your highest level of education?

Mark only one oval.

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

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

Mark only one oval.

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

9. How certain are you about your choice?

Mark only one oval.

- ◯ Very certain
- ◯ Somewhat certain
- ◯ Unsure

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

Check all that apply.

- ☐ Economy and jobs
- ☐ Healthcare
- ☐ Climate change
- ☐ Education
- ☐ Immigration
- ☐ National security
- ☐ Social justice and equality
- ☐ Abortion rights
- ☐ Gun control
- ☐ Other: _____

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

Mark only one oval.

- ◯ Joe Biden
- ◯ Donald Trump
- ◯ Other: _____

12. What are your primary sources for political news?

Mark only one oval.

- ◯ Television
- ◯ Newspapers/Magazines
- ◯ Online News Websites
- ◯ Social Media
- ◯ Radio
- ◯ Friends and Family
- ◯ Other: _____

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

Mark only one oval.

- ◯ Yes
- ◯ No

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

This content is neither created nor endorsed by Google.

Google Forms

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.