# My title*

**My subtitle if needed**

Abdullah Motasim Elizabeth Luong

November 3, 2024

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

## Table of contents



*Code and data are available at: https://github.com/abdullah-motasim/Forecasting-2024-US-Presidential-Election.

# 1 Introduction

Overview paragraph: On November 5 2024 the U.S. Presidential Elections will be held to determine the 47th President of the United States. Since the U.S. is a powerhouse within the world, often affecting crucial international affairs, the result of this election will have a global impact on all countries and as such there is much interest in which of the top two candidates Donald Trump or Kamala Harris will end up winning the election. As such, leading up to the election time there are many polls conducted by different pollsters to determine the winner of the election, but these polls often vary significantly between each other due to biases and sampling limitations. Poll aggregation is a method which aims to mitigate these differences by combining or aggregating the results of multiple polls to provide a more robust estimate of the election outcome.

This study seeks to forecast the 2024 U.S. Presidential election outcome by developing a generalized linear model based on poll aggregation data. Specifically, we evaluate predictors including demographics, polling trends, and historical voting patterns. While the popular vote does not directly determine the presidential outcome due to the electoral college system, our model's findings offer a clearer picture of voter preferences and potential trends in key states. This paper outlines our modeling approach, data processing techniques, and results, with a final section discussing model limitations and implications for future election forecasting.

Estimand paragraph: The primary estimand in this study is the probability of a Democratic or Republican victory in the popular vote, based on current polling trends and demographic predictors.

Results paragraph: State results of who won in which state (the 5 in exploratory) and also talk about the final model and who it predicted to win in each state on the day of the election than also who won overall.

Why it matters paragraph:

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

# 2 Data

## 2.1 Overview

The aggregated dataset containing the polling data for the 2024 U.S. election was sourced from FiveThrityEight (). We use the statistical programming language R (R Core Team 2023) and packages the … to download, analyze, and model the data.

The initial raw data contained about 17000 observations of 52 variables from multiple pollsters and after cleaning the data we were left with about 9000 observations of 7 variables. The data cleaning involved removing missing values and only selecting variables which were of importance to us, these variables were:

- **pollster_name** - The name of the polling organization that conducted the poll (e.g., YouGov, RMG Research).
- **candidate** - The name of the candidate in the poll (e.g., Kamala Harris, Donald Trump).
- **percentage** - 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).
- **party** - The political party of the candidate in the poll (e.g., DEM for Democrats, REP for Republicans).
- **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.
- **numeric_grade** - A numeric rating given to the pollster to indicate their quality or reliability (e.g., 3.0).

These variables were selected due to their direct relevance to assessing candidate support, polling reliability, and geographic voting trends. Specifically, pollster_name and numeric_grade are crucial for tracking potential biases and variations across polling organizations, as different methodologies can impact results. Candidate/Party and percentage provide direct indicators of support levels for each contender, which is essential for making accurate projections of election outcomes. Also, it is worth noting that there were more Sample_size is key to assessing the reliability of each poll, as larger samples tend to yield more accurate representations of public opinion. Lastly, state enables us to account for regional variations, crucial in a system like the U.S. Electoral College, where state-level results play a determining role in the election. Collectively, these variables offer a focused and comprehensive view of the polling landscape, supporting a more reliable aggregation and interpretation of the data.

## 2.2 Measurement

Since different pollsters utilize various methods to collect their data and convert real-world phenomena into entries in their datasets, we have chosen to focus on the method utilized by Emerson, as it is the most frequent pollster within our cleaned data, accounting for 823 of the observations. Analyzing Emerson's methodology provides valuable insights into how polling results are generated, thereby enhancing our understanding of the dataset.

Emerson College Polling employs a combination of Interactive Voice Response (IVR) and online panel methodologies for data collection. The recent survey of Michigan sampled 1,000 likely voters. Emerson developed a series of survey questions to ask these potential candidates such as do you prove or disapprove of the job Joe Biden is doing as president? What is your party registration? etc. After this the results of the survey are analyzed to get the percentages of votes for each option and these are than presented within the aggregated data set.

Note this is a general overview of the measurement method utilized by Emerson, a more thorough breakdown of measurement can be found in [Appendix A].

## 2.3 Outcome variables

Add graphs, tables and text. Use sub-sub-headings for each outcome variable or update the subheading to be singular.

Some of our data is of penguins (**?@fig-bills**), from Horst, Hill, and Gorman (2020).

Talk more about it.

And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

Talk way more about it.

## 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

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix D.

## 3.1 Model set-up

Define $y_i$ as the number of seconds that the plane remained aloft. Then $\beta_i$ is the wing width and $\gamma_i$ is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

### 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 $\theta$.

# 4 Results

Our results are summarized in Table **??**.

# 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 Pollster Methodology Overview and Evaluation

The Emerson College Polling method utilizes a hybrid approach, combining Interactive Voice Response (IVR) with an online panel to gather data from likely voters in Michigan. This methodology is crucial in translating individual opinions into quantifiable data that can inform electoral forecasts and insights.

**Population, Frame, and Sample**
The target population for this survey consists of likely voters in Michigan. The sampling frame is constructed to include diverse demographic groups within this population, ensuring a representative sample of voter sentiments.

**Sample Recruitment and Sampling Approach**
Emerson employs both IVR and online panel methodologies to recruit participants. The IVR method facilitates automated phone interviews, reaching demographics that may be less likely to participate in online surveys, such as older individuals or those with limited internet access. In contrast, the online panel is sourced from CINT, engaging participants who are more willing to complete surveys digitally. This dual approach captures a broader spectrum of voter sentiments and demographics, balancing the advantages of both methods. However, while IVR can access hard-to-reach populations, it may also yield lower response rates compared to the more engaged online panel.

**Non-response Handling**
While the IVR method can reach various demographics, it is essential to consider potential non-response bias. Respondents who opt not to participate may share similar characteristics, leading to skewed results. Emerson may implement follow-up strategies or incentives to encourage participation, but these measures are not explicitly detailed in the methodology.

**Questionnaire Evaluation**
Emerson crafted a comprehensive questionnaire that included a range of questions, such as "Do you approve or disapprove of the job Joe Biden is doing as President?" and "What is your party registration?" Respondents provided their answers using multiple-choice formats, offering options like "Approve," "Disapprove," "Neutral or no opinion," as well as party affiliations such as "Democrat," "Republican," and "Independent/other." The full questionnaire and its results can be accessed at this link. After collecting the responses, the results were tallied to calculate the percentage for each option, enabling pollsters to convert individual opinions into broader trends. While the questions are generally clear and allow for straightforward responses, careful consideration should be given to the wording to avoid any potential bias. Questions that lead respondents toward a particular answer could impact the validity of the data collected.

**Overview of Methodology**

- **Data Collection Techniques:** Emerson employs both IVR and online panel methodologies. IVR allows for automated phone interviews, which can reach a diverse demographic, including those less likely to participate in online surveys. The online panel, sourced from CINT, complements this by engaging participants who are more accessible and willing to complete surveys digitally. This dual approach is intended to capture a broad spectrum of voter sentiments and demographics.
- **Sample Size and Credibility:** The survey includes a sample of 1,000 likely voters, providing a reasonable basis for statistical inference. Emerson states a credibility interval of $\pm 3$ percentage points, which indicates a moderate level of precision in their estimates. This credibility interval should be considered when interpreting results, as it reflects the potential variability in public opinion.
- **Weighting Adjustments:** To ensure that the sample accurately reflects the population of likely voters, the data is weighted by gender, education, race, age, party registration, and region. This weighting is critical, as it aligns the sample with demographic characteristics of the actual voter population, enhancing the reliability of the findings.

**Strengths of the Methodology**

- **Diverse Reach:** The combination of IVR and online panel methods allows Emerson to capture data from a wide range of demographics, minimizing the risk of bias that may arise from using a single method. This is particularly important in a politically diverse state like Michigan.
- **Rigorous Weighting:** The meticulous approach to weighting the data helps to correct any imbalances in the sample, ensuring a more accurate reflection of voter sentiments across different demographic groups.
- **Clarity in Reporting:** Emerson provides detailed demographic breakdowns and cross-tabulations, allowing for a nuanced understanding of how different groups perceive key issues and candidates.

**Limitations of the Methodology**

- **Potential Non-response Bias:** While the IVR method can reach various demographics, those who opt not to participate may share similar characteristics, potentially leading to non-response bias. Additionally, the online panel may attract a skewed sample if certain demographics are overrepresented or underrepresented in the panel.
- **Higher Credibility Intervals for Subgroups:** Subsets based on demographics (e.g., age or party registration) often come with larger credibility intervals due to smaller sample sizes. This can diminish the reliability of conclusions drawn from those specific groups.

- **Temporal Limitations:** The survey results reflect a snapshot of opinions at a specific time (October 25-27, 2024), which may not account for rapidly changing political landscapes, particularly as Election Day approaches.

In summary, Emerson College Polling employs a robust methodology that effectively combines IVR and online panel methodologies to gauge voter sentiment. While there are inherent limitations, such as potential biases and the credibility of subgroup analysis, the strengths of this approach lie in its comprehensive reach and careful weighting adjustments. Overall, this methodology offers a valuable tool for understanding the current political climate in Michigan.

# B Idealized Survey

In this appendix, I outline an idealized methodology and survey for forecasting the U.S. presidential election, utilizing a budget of $100,000. This approach combines robust sampling techniques, diverse recruitment strategies, and thorough data validation to ensure accurate and reliable insights into voter sentiment.

**Sampling Approach**

1. **Target Population**: The target population consists of registered voters in the United States. The survey will focus on likely voters, including individuals aged 18 and older across various demographic groups, including age, gender, race, and political affiliation.
2. **Sample Size**: A sample size of approximately 5,000 likely voters will be utilized. This sample size is designed to provide a high level of statistical confidence, with a credibility interval of $\pm 2$ percentage points.
3. **Sampling Technique**:
    - A **stratified random sampling** method will be employed to ensure representation across key demographic groups. The population will be divided into strata based on demographic variables, such as age, gender, race, and region.
    - A systematic selection will then be conducted within each stratum to achieve a balanced representation, reducing sampling bias.

**Recruitment of Respondents**

1. **Data Sources**:
    - Voter registration databases will be utilized to identify potential respondents. These databases will provide contact information for registered voters across the United States.

- Partnerships with organizations that specialize in voter engagement will also be established to enhance the recruitment process.

2. **Recruitment Methods**:
   - **Online Panel**: An online panel sourced from reputable survey platforms (e.g., SurveyMonkey, Qualtrics) will be used to recruit respondents who opt-in to participate in surveys.
   - **IVR and SMS Outreach**: Interactive Voice Response (IVR) and SMS outreach will be implemented to reach demographics that may be less likely to engage in online surveys. This will ensure inclusivity and broaden the reach of the survey.
3. **Incentives**: Participants will be incentivized to complete the survey with small monetary rewards (e.g., $5) or entry into a raffle for a larger prize, thereby increasing response rates and engagement.

**Data Validation and Quality Assurance**

1. **Pre-Survey Testing**: A pilot test of the survey will be conducted with a smaller group of respondents (approximately 200) to identify any ambiguities or biases in the questions. Feedback will be incorporated to refine the survey instrument.
2. **Quality Checks**:
   - During data collection, real-time monitoring will be implemented to identify and address any inconsistencies in responses.
   - Automated checks will flag any duplicate responses, incomplete surveys, or inconsistent answer patterns.
3. **Post-Survey Validation**: After data collection, a validation process will be employed to cross-reference responses with demographic data to ensure the sample reflects the broader population accurately.

**Poll Aggregation and Analysis**

1. **Data Aggregation**: The data will be aggregated and analyzed using statistical software (e.g., R or Python) to compute margins of error and establish confidence intervals for key questions, such as candidate approval ratings and voting intentions.
2. **Cross-Tabulation**: The survey results will be cross-tabulated by demographic variables to provide insights into how different groups perceive candidates and issues. This will enable nuanced analysis of voter sentiment across demographics.
3. **Reporting**: Findings will be compiled into a comprehensive report summarizing key insights, trends, and forecasts for the upcoming presidential election. The report will be shared with stakeholders, including political analysts and campaign teams.

### Survey Implementation

The survey will be implemented using Google Forms. Below is a link to the actual survey, which includes an introductory section, well-constructed questions, and a closing section thanking respondents for their participation:

**U.S. Presidential Election Forecasting Survey**

### Survey Copy

### U.S. Presidential Election Forecasting Survey

### Introduction

Thank you for participating in our survey! Your insights are invaluable for understanding voter sentiment leading up to the upcoming US presidential election. This survey will take approximately 5-10 minutes to complete. Your responses are anonymous and will be used solely for research purposes.

If you have any questions or concerns, please contact Elizabeth Luong at elizabethh.luong@mail.utoronto.ca.

### Section 1: Demographic Information

1. **What is your age?**
    - Under 18
    - 18-24
    - 25-34
    - 35-44
    - 45-54
    - 55-64
    - 65 or older
2. **What is your gender?**
    - Male
    - Female
    - Non-binary/Third gender
    - Prefer not to say
3. **What is your race or ethnicity:**
    - White

- Black or African American
- Asian
- Hispanic or Latino
- Other (please specify): [Open text field]

4. **In which state do you reside?**
    - [Open text field]
5. **What is your highest level of education completed?**
    - Some high school
    - High school diploma or equivalent
    - Some college
    - Bachelor's degree
    - Graduate degree
6. **What is your party affiliation?**
    - Democrat
    - Republican
    - Independent
    - Other (please specify): [Open text field]

**Section 2: Political Opinions**

6. **How likely are you to vote in the upcoming presidential election?**
    - Very likely
    - Somewhat likely
    - Unsure
    - Somewhat unlikely
    - Very unlikely
7. **Which candidate do you plan to vote for in the upcoming presidential election?**
    - Donald Trump (Republican)

- Kamala Harris (Democrat)
- Robert F. Kennedy Jr. (Independent)
- Undecided
- Other (please specify): [Open text field]

8. **What are the most important issues influencing your vote?** (Select up to three)
   - Economy
   - Healthcare
   - Education
   - Climate Change
   - Social Justice
   - National Security
   - Immigration
   - Other (please specify): [Open text field]
9. **How do you feel about the job performance of the current President?**
   - Strongly approve
   - Approve
   - Neutral
   - Disapprove
   - Strongly disapprove

**Section 3: Media Consumption**

10. **Which sources do you primarily use to get news about the election?** (Select all that apply)
    - Television
    - Online news websites
    - Social media
    - Radio
    - Newspapers
    - Podcasts

- Other (please specify): [Open text field]

11. **How often do you discuss politics with friends and family?**

    - Daily
    - Weekly
    - Monthly
    - Rarely
    - Never

**Conclusion**

Thank you for taking the time to complete this survey! Your feedback is crucial for understanding voter sentiment as we approach the election. If you have any additional comments or thoughts you'd like to share, please reach out to Elizabeth Luong at elizabethh.luong@mail.utoronto.ca.

**Thank you again for your participation!**

# C Additional data details

# D Model details

## D.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…

Examining how the model fits, and is affected
by, the data

## D.2 Diagnostics

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

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

Checking the convergence of the MCMC algorithm

## References


Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

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

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