# A Prediction of the 2024 U.S. Election Outcome*

### A Bayesian Spline Regression Analysis of Kamala Harris's Winning

Shanjie Jiao Aman Rana Kevin Shen

November 4, 2024

The 2024 U.S. Presidential Election, set against a backdrop of economic and political challenges, features Donald Trump and Kamala Harris as the main candidates. This study uses Bayesian spline regression on polling data from FiveThirtyEight to predict the election outcome, capturing dynamic voter trends over time. Our analysis suggests a likely win for Kamala Harris. Given the U.S.'s global influence, this result could significantly impact international security, trade, geopolitics, and provide a clear insight of forthcoming policy directions.

## Table of contents



*Code and data are available at: [https://github.com/Jie-jiao05/2024_US_Election.git).

# 1 Introduction

The United States, as the world's largest economy and most influential country, will hold its presidential election on November 5, 2024. Despite facing domestic challenges such as inflation, low employment rates after the pandemic, and increasing political polarization due to conflicts between the Democratic and Republican parties, which have accelerated internal tensions.As of October 2024, former President Donald Trump has secured the Republican Party nomination. However, after assassination happened on July 14, 2024, many of his supporters have become even more committed to his legacy. Meanwhile, President Joe Biden withdrew from the election on July 21, leading to the nomination of Kamala Harris as the Democratic candidate. These events have made the 2024 U.S. election increasingly complex and unpredictable, further intensifying uncertainty about the future.

The estimand of this study is the predicted outcome of the 2024 U.S. Presidential Election for either Donald Trump or Kamala Harris, based on aggregated polling averages. Our analysis employs Bayesian spline regression to effectively capture the dynamic changes in voting trends over time, allowing us to forecast the likely election result. By comparing the predicted probabilities for both candidates, we aim to identify the probable winner based on current trends in voter support.

Using Bayesian spline regression analysis on the dataset provided by FiveThirtyEight (FiveThirtyEight 2024), this study forecasts the outcome of the 2024 U.S. election, predicting a victory for Kamala Harris.

Due to the United States' hegemonic position and unparalleled global influence, the outcome of the U.S. election will serve as a guiding light for global developments over the next four years, significantly impacting international security, trade, cooperation, and geopolitics. This article aims to provide a clearer analytical framework for political scientists, the general public, journalists, corporate strategists, and anyone globally concerned with the U.S. election.

The remainder of this paper is structured as follows. Section 2 we will discuss the MLR used in the prediction and the result of the MLR with its predictor variables and the response after transformation. **?@sec-dis** the shortcomings of the study and areas for improvement will be described. Section A.1 A Deep Dive to a pollster will be conducted, and Section A.2 We will given a idealized survey and methodology

# 2 Data

## 2.1 Overview

We use the statistical programming language R (R Core Team 2023) to analysis the data which are extracted from the 2024 U.S. General President Election collected and published by FiveThirtyEight (FiveThirtyEight 2024). This paper will conduct a Multiple Linear Regression to do further prediction. In this model, we take numeric_grade, pollscore, sample_size as predictors, and pct as response.

## 2.2 Data Cleaning

To Simplify the dataset we extract the data of the pollater_id, pollster, methodology, numeric_grade, pollscore, sample_size, pct, and transform Party and Answer into categorical (0=DEM, 1=REP and 0 = Harris, 1=Trump). In oreder to make the pollster more reliable we will only analysis that transparency are over 3 to increase the liability of this survey.

Overview text

## 2.3 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

## 2.4 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 (Figure 1), from Horst, Hill, and Gorman (2020).

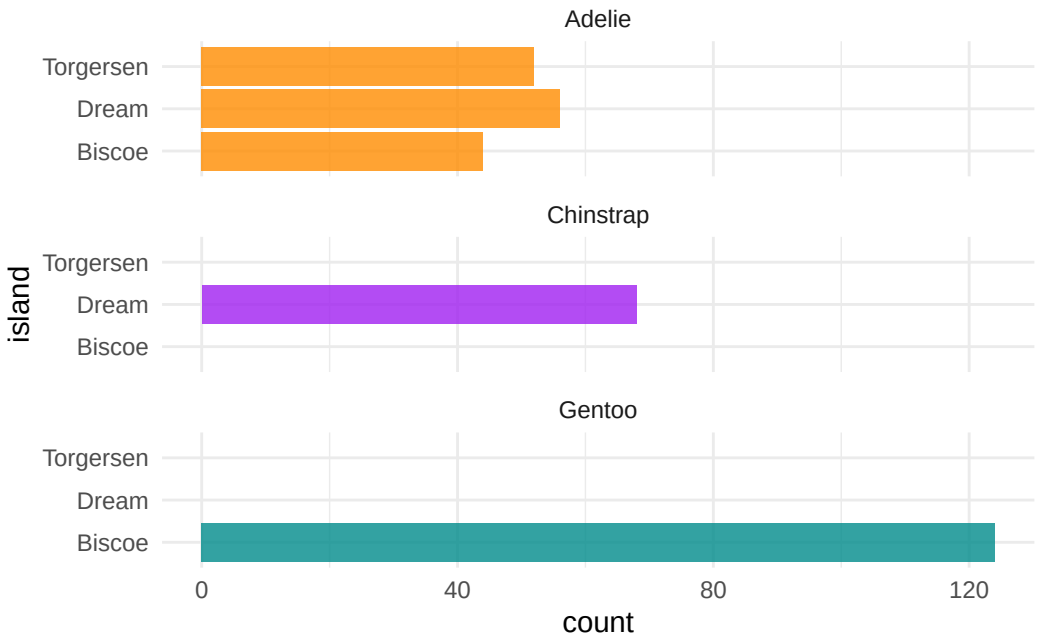



Figure 1: Bills of penguins

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.

### 2.5 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 B.

### 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.

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

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 $\theta$.

Table 1: Explanatory models of flight time based on wing width and wing length

| | First model |
|---|---|
| (Intercept) | 1.12 |
| | (1.70) |
| length | 0.01 |
| | (0.01) |
| width | −0.01 |
| | (0.02) |
| Num.Obs. | 19 |
| R2 | 0.320 |
| R2 Adj. | 0.019 |
| Log.Lik. | −18.128 |
| ELPD | −21.6 |
| ELPD s.e. | 2.1 |
| LOOIC | 43.2 |
| LOOIC s.e. | 4.3 |
| WAIC | 42.7 |
| RMSE | 0.60 |

# 4 Results

Our results are summarized in Table 1.

# 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.

# A Appendix

## A.1 A: Emerson College Polling Methodology Analysis

### A.1.1 Overview

Emerson College Polling (ECP) is a nationally recognized, non-partisan polling organization (Emerson College Polling 2024a). ECP conducted a national survey from October 30 to November 2, 2024, targeting 1,000 likely voters in the 2024 U.S. presidential election. The poll measured voter preferences between candidates Kamala Harris and Donald Trump, indicating a tie at 49% each, with third party candidates and undecided voters at one percent each (Emerson College Polling 2024b). Appendix A examines the methodology used by ECP for this poll.

### A.1.2 Survey Population and Sampling

When conducting a poll, it is crucial to define the target population, sampling frame, and sample in order to provide a clear direction and objective (Whaley 2024).

The target population is the entire group about which we aim to understand and draw conclusions from (Alexander 2023). For this poll, the target population consists of likely voters in the 2024 U.S. presidential election. ECP determines likely voters through a combination of voter history, registration status, and self-reported demographic data (Emerson College Polling 2024b).

Since it is oftentimes infeasible to gather data from the entire target population, sampling frames and samples are utilized. A sampling frame is the list of all units from which samples can be taken (Alexander 2023). In this poll, the sampling frame consists of voters available on Aristotle's database and the online panel provided by CINT (Emerson College Polling 2024b). Aristotle maintains comprehensive voter and consumer data (Aristotle 2024), while CINT operates as a global research marketplace connecting researchers to survey respondents (CINT 2024). The sample consists of 1,000 likely voters selected from this frame.

### A.1.3 Data Collection and Sampling Approach

ECP employs a mixed-mode sampling methodology for its polling (Emerson College Polling 2024a). When collecting data for its October 30 to November 2, 2024 U.S. Presidential Election poll, three primary methods were used (Emerson College Polling 2024b):

1. MMS-to-web text survey: Respondents receive text messages with custom graphics inviting them to take online Qualtrics survey. Selected randomly from Aristotle voter files.
2. Online panel surveys: CINT panel respondents screened for voter registration and demographics, then directed to survey with quality checks.

3. Interactive Voice Response (IVR): Automated calls to landlines where permitted. Respondents use touch-tone phones. Selected randomly from Aristotle voter files.

This mixed-mode approach offers several advantages, primarily reducing coverage bias by reaching different demographic groups within the target population. Coverage bias occurs when there is a discrepancy between the sampling frame and the target population (Stantcheva 2023). By employing multiple methods, ECP minimizes this bias: MMS-to-web surveys typically capture younger voters, IVR targets older and rural voters who prefer landlines, and online panels engage tech-savvy individuals. A mixed-mode approach can also reduce fieldwork costs by maximizing the use of lower cost methods, and using higher cost methods when necessary (Wilkinson and McTiernan 2020). However, this approach introduces measurement error as each method brings its own biases (Emerson College Polling 2024a), potentially increasing the overall margin of error.

### A.1.4 Non-response Handling

In polling, non-response occurs when subjects either refuse to participate in a survey or skip specific questions, limiting the survey's representativeness beyond its respondents. To address this issue and ensure representativeness, ECP implements a weighting system based on multiple demographic variables:

- Gender
- Education
- Race
- Age
- Party registration
- Region

These weights are calculated based on 2024 likely voter modeling, adjusting for under- and over-represented groups in the sample (Emerson College Polling 2024b).

### A.1.5 Questionnaire Design

The survey's questionnaire demonstrates both strengths and limitations. Its strengths include:

- Clear, unambiguous question formulation
- Efficient format
- Topical relevance focusing on vote preferences and key demographic splits

However, the questionnaire's limitations include:

- Limited exploration of underlying voter motivations

- Potential variation in response quality across different survey modes
- Basic preference capture without deeper attitudinal investigation
- All data collection was conducted in English, potentially limiting representation of non-English speaking voters.

The survey maintains a credibility interval of +/- 3 percentage points, with higher intervals for demographic subsets due to reduced sample sizes.

### A.1.6 Conclusion

Emerson College Polling's mixed-mode methodology effectively balances cost with demographic reach. While their weighting system helps ensure representativeness, the multi-mode approach presents trade-offs: reduced coverage bias through diverse voter outreach, but increased measurement error from combining different survey methods (Mora 2011). These methodological considerations are crucial when interpreting the poll's results.

## A.2 B: Idealized Survey Methodology – $100K Budget

### A.2.1 Introduction

In this appendix, we outline an idealized survey methodology for forecasting U.S. presidential elections within a budget of up to $100,000. The proposed design aims to maximize accuracy and cost-efficiency by strategically selecting sample populations, targeting key demographic groups, and effectively aggregating results.

### A.2.2 Sampling Strategy

The ideal survey methodology employs a stratified random sampling approach, given the diversity of the American population as the target group. This method is designed to better analyze the varying effects of different races, age groups, genders, and education levels on voting trends by distributing the sample into multiple subgroups. By categorizing these key variables, this approach ensures more concise and clearer statistical results, which in turn facilitates in-depth analysis. The representative sample is listed below:

- **Sampling Frame**: For the information security, valid entitlement holders registered on the voter list and registered with the Census Bureau will receive the survey.
- **Sampling Method**: Stratified random sampling will be implemented, ensuring better analyze the differential impact of different key factors, including races, ages, genders, and levels of education on voting trends.

- **Sample Size**: The poll is expected to include responses from 10,000 individuals. Assuming nonresponse bias is 10%, we estimate the final adjusted sample size of approximately 9,000.
- **Geographical Distribution**: The survey will be conducted across all 50 states and the District of Columbia. Considering that traditional "red" and "blue" states are less likely to change political alignments, an increased focus will be placed on polling within swing states.

### A.2.3 Recruitment Plan

This survey will conduct recruiting in following methods:

- **Online recruitment**: Advertisements and official information about the survey will be disseminated through various search engines (including Google and Firefox) and social media platforms such as X and Instagram. All advertisements will contain direct links to the Google Forms survey.

- **Phone recruitment**: Valid phone numbers will be contacted by staff to conduct the survey using the same set of questions as the online survey. For privacy and security, no conversations will be recorded, and the calling system will automatically dial numbers without disclosing them to staff members.
- **In person recruitment**: A tent will be set up in key high-traffic regions in swing states to conduct survey in-person.
- **Gift incentive**: To increase participation, all respondents will receive a specially designed pin, estimated to cost around $2 each. This incentive is intended to boost engagement and encourage broader participation in the survey.

### A.2.4 Budget Allocation

- **Online Recruitment Cost**: $20,000 allocated for advertising across various platforms (search engines and social media).
- **Phone Recruitment Cost**: $20,000 allocated for hiring staff to conduct phone surveys and cover the associated data usage charges.
- **In-Person Recruitment Cost**: $22,000 allocated for hiring field staff and setting up in-person recruitment locations.
- **Gift Cost**: $20,000 allocated for the total value of participant gifts
- **Data Analysis and Quality Control Cost**: $6,000 allocated for employing technical staff, purchasing software licenses, and utilizing required tools.
- **Administrative Cost**: $2,000 allocated for overall management, including miscellaneous expenses and logistical support.

Total Budget: $100,000

### A.2.5 Survey Design

This survey will only collect data that are accurate and unbigoted. Meanwhile, personal information will not be contained.

- **Close-ended Question**: Usage of multiple choice can make sure the answer is straightforward and easier for analysis.
- **Response Option**: Most questions will only contain option from A, B and "prefer not to say", all the option will be a neutral response.
- **Insensitive Information**: This survey will only contain personal information only and merely age, gender, race, education background, income, living state.

### A.2.6 Data Process

To ensure data we collected are valid and significant, we will conduct following methods to test our data.

- **One-times Submission**: For online surveys, only one submission will be allowed per unique IP address to prevent multiple entries that could skew results. The same principle will apply to phone recruitment, ensuring each number is only surveyed once.
- **Missing Data**: During the data cleaning process, responses with missing values (N/A entries) will be excluded to maintain the integrity and reliability of the analysis.
- **Swing States Data Process**: Data collected from swing states will be assigned a slightly higher weight in the estimation process, given their critical influence on election outcomes.
- **Stratified Analysis**: To gain insights into voter preferences within different demographic subgroups, stratified analyses will be conducted, focusing both on overall trends and trends within specific subgroups.
- **Poll Aggregation Approach**: The aggregated data will be weighted based on sample size (with larger samples receiving greater weight), and past reputable polls will be incorporated, giving priority to highly rated pollsters to strengthen the overall analysis.

### A.2.7 Sample Survey

Google Forms will be used to collect data efficiently and securely, ensuring no information is leaked. The survey will cover aspects such as gender, race, demographics, education, first-choice candidate, and views on the current United States.

A complete sample survey could be found in the following link https://forms.gle/1jrQP7bPqntJ9N2s9 or by clicking Sample survey.

Proposed Survey Questions:

1. What sex were you assigned at birth, or on your original birth certificate?

- Male
- Female

2. How do you currently describe your gender?

- Male
- Female
- Transgender
- Non-binary
- other

3. What is your racial self-identification?

- White
- Black
- Asian
- Indigenous
- Prefer not to say
- Other

4. What is your current age?

- Under 18
- 30-45
- 46-51
- 52-66
- over 66

5. What is your highest level of education?

- High school or Under
- Undergraduate
- Graduate _ Doctor

6. Which candidate would you vote for in U.S. 2024 Presidential Election?

- Donald Trump (Republic)
- Kamala Harris (Democrat)
- Other

7. Would you vote for the 2024 U.S. Presidential Election?

- Yes
- No

- I'm not sure

8. Are you satisfied with the current situation in the United States?

- Pretty much
- Somewhat satisfied
- Neutral
- Somewhat dissatisfied
- Very Dissatisfied
- No opinion

9. What is your most concerned toward to current U.S?

- Economy
- Employment
- Education
- Environment
- Diplomacy
- Immigration
- Domestic safety

#### A.2.8 Conclusion

The processes outlined in this survey aim to provide valid and meaningful insights into the upcoming U.S. 2024 Presidential Election. Working within a limited budget, we have carefully designed the survey and data processing methods to achieve reliable and comprehensive analyses, both at the subgroup level and in the overall population.

## B Model details

### B.1 Posterior predictive check

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

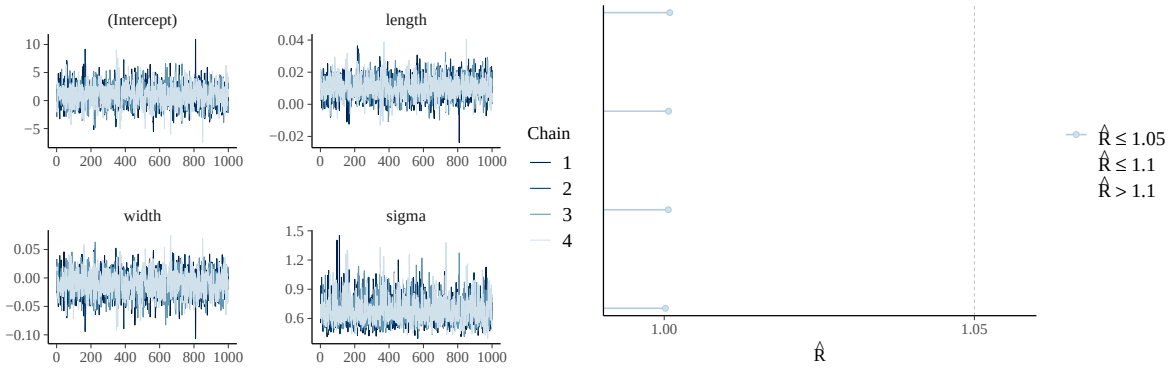
In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

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

## B.2 Diagnostics

Figure 2a is a trace plot. It shows… This suggests…

Figure 2b is a Rhat plot. It shows… This suggests…




(a) Trace plot (b) Rhat plot

Figure 2: Checking the convergence of the MCMC algorithm

## References


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

Aristotle. 2024. "Data." https://www.aristotle.com/data/.

CINT. 2024. "Cint." https://www.cint.com/.

Emerson College Polling. 2024a. "About Us." https://emersoncollegepolling.com/about/.

———. 2024b. "November 2024 Tracking National Poll: Trump and Harris Remain Locked in Tight Race." https://emersoncollegepolling.com/november-2024-national-poll-trump-and-harris-remain-locked-in-tight-race/.

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

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.

Mora, Michaela. 2011. "Understanding the Pros and Cons of Mixed-Mode Research." https://www.relevantinsights.com/articles/pros-and-cons-of-mixed-mode-research/.

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

Stantcheva, Stefanie. 2023. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." *Annual Review of Economics* 15 (1): 205–34. https://doi.org/10.1146/annurev-economics-091622-010157.

Whaley, Jim. 2024. "Understanding Target Population in Research." OvationMR. https://www.ovationmr.com/target-population-in-research/.

Wilkinson, Sara, and Leah McTiernan. 2020. "Mixed Mode Research: Reaching the Right People in the Right Way to Get the Data You Need." Ipsos. https://www.ipsos.com/sites/default/files/ct/publication/documents/2020-06/mixed-mode-research-ipsos.pdf.