The Battle for America: Forecasting about the 2024 Presidential Election*

My subtitle if needed

Yi Tang Jin Zhang Siyuan Lu

November 1, 2024

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

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

We use the statistical programming language R (R Core Team 2023).There are totally 17749 observations in the original dataset. Variables in the dataset help analyze and expect the final outcome of US Presidential Election in 2024.

Below is the key variable from dataset that we used for our analysis:

- poll_id: Uniquely identifies each poll.
- pollster: The organization that conducted the poll.

*Code and data are available at: https://github.com/RohanAlexander/starter_folder.

- state: The state in which the poll was conducted.
- numeric_grade: A grade assigned to the pollster based on their historical accuracy and methodology, where grades equal to or greater than 3 have been included for analysis to ensure data reliability.
- methodology: The method used for polling, such as phone, online and so on.
- start_date and end_date: The duration over which the poll was conducted.
- sample_size: The number of respondents in the poll.
- population: The abbreviation of the respondent group indicating their voting status, such as "likely voters" or "adults"
- party: The political party of the candidate.
- candidate_name: The name of the candidate to whom the poll response pertains, originally recorded under 'answer' and renamed for clarity.
- pct: The percentage of respondents favoring the candidate.

After studying the data and cleaning the data, we decided to select pollsters with a numeric grade greater than or equal to three because the results of such polling organizations are more meaningful. And we set candidate names to Harris and Trump, which are the two most widely watched candidates this year.

2.2 Measurement

In order to show how political developments affect public opinion, variables like State, Start Date, End Date, and Candidate Name link the data to particular electoral events and timelines. Methodological specifics guarantee that the technique used to gather the data is suitable for obtaining an accurate representation of voter sentiment.

The degree to which these attitudes reflect the electorate's actual views can vary depending on the data collection method (methodology), whether it be online, in-person, or over the phone. This is because different methodologies are susceptible to biases such as selection bias and nonresponse bias.

2.3 Outcome variables

The dataset primarily revolves around understanding and predicting voter preferences and trends through various polling metrics. The key outcome variables in this dataset are:

- candidate_name: This variable captures the preference expressed by the polled individuals, indicating which candidate they intend to vote for or support. It serves as the primary outcome variable for analyses aimed at gauging candidate popularity and electoral viability.

- pct: This variable represents the percentage of respondents who support a given candidate in each poll. It is crucial for conducting detailed trend analysis over time and across different demographic segments within each state.

To comprehend the dynamics determining voter preference, these result factors are compared to a variety of predictor variables (independent variables), including state, sample size, and pollster numeric_grade. Through the start_date and end_date columns, which specify the length of each poll, the dataset also enables time-series analysis, which aids in monitoring shifts in public sentiment before the election. Analyzing these result variables can reveal information about regional preferences, the success of campaign tactics, and possible changes in the voting base brought about by current political developments and socioeconomic determinants.

Some of our data is of penguins (**?@fig-bills**), from (**palmerpenguins?**).

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 C.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_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 θ.

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.

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

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Methodology

1200 sample of registered voters

#Online survey

B Additional data details

C Model details

C.1 Posterior predictive check

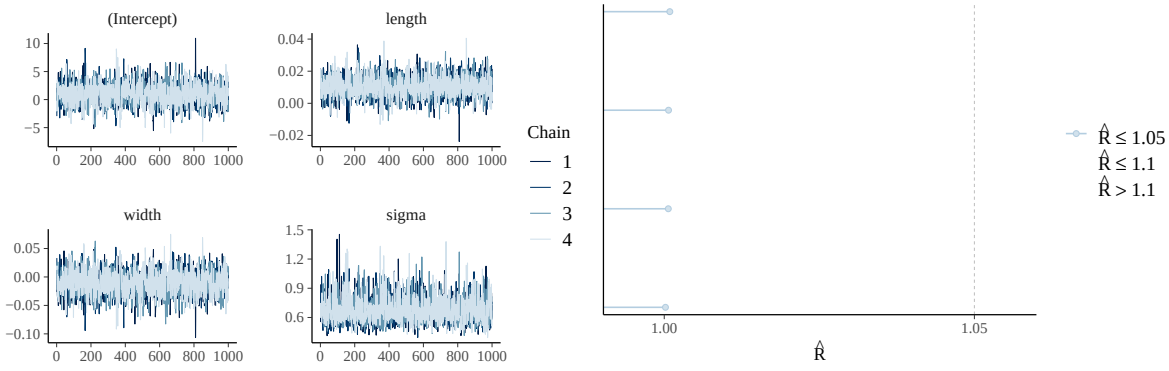
In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

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

C.2 Diagnostics

Figure 1a is a trace plot. It shows... This suggests...

Figure 1b is a Rhat plot. It shows... This suggests...



(a) Trace plot

(b) Rhat plot

Figure 1: Checking the convergence of the MCMC algorithm

D Appendix 1

D.1 Population, Frame, and Sample

YouGov conducted surveys with adults in the U.S. who have internet access. They used their online group of people as the sample frame. This frame covers a wide population. For each survey, the sample includes people from this group who fit specific requirements set for that survey. This helps collect data that targets certain information.

D.2 Sample Recruitment

YouGov worked actively to bring people into their panel. They used different ways of advertising and worked with various partners across many types of media. This method aimed to gather a wide range of people. This helps make sure the group of survey takers represents many different kinds of people in the U.S. However, it mostly includes people who can use the internet.

D.3 Sampling Methodology

YouGov used a non-probability method to find people for their online panel through various media sources. This helps improve how well the panel represents different groups of people. This method is fast and lets them interact with the survey takers through surveys that can include videos and pictures. But, relying on people who choose to join and using digital surveys might limit who takes part and affect the survey's accuracy. They made the survey results match the general population by adjusting weights according to certain standards.

D.4 Handling Non-Response

They managed the issue of people not responding to surveys by adjusting the survey sample's make-up. They did this using weights based on well-known data like the U.S. Census.

D.5 Questionnaire Design

YouGov designed their questionnaires to be easy to access and interesting, often using videos and pictures to give more information. However, using digital formats might stop people with limited internet skills or access from taking part.

References

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

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