Predicting the 2024 US Presidential Election with a Model-Based Forecast*

Using Generalized Linear Models to Predict Election Outcomes

Yuanyi (Leo) Liu Dezhen Chen Ziyuan Shen

October 21, 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) to process and analyze polling data for the 2024 US Presidential election. The data, obtained from a repository of publicly available polls, includes variables from numerous pollsters across the country (Ryan Best 2024). The aim is to create a reproducible forecast model by focusing on high-quality polls and narrowing our scope to Kamala Harris as a candidate of interest. Following methodologies discussed by "Telling Stories with Data" (Alexander 2023), we examine how polling data reflects electoral behavior and voter preferences.

*Code and data are available at: Forecasting the 2024 US Presidential Election.

The dataset includes 15,891 rows and 52 columns, covering various pollster attributes such as pollster name, state, methodology, and polling results. We filter the dataset to include only high-quality polls, i.e., those with a numeric grade of 3.0 or higher, ensuring that only reputable and well-documented pollster results are used in our model. Additionally, we focus solely on polls that include Kamala Harris as a candidate, which helps narrow our analysis to a specific electoral scenario.

2.2 Measurement

Polling data captures voter preferences and electoral forecasts by collecting responses from a representative sample of the population. In this case, pollsters attempt to gauge public opinion by surveying individuals on their preferred candidate, adjusting for demographic factors and political trends. The raw data entries reflect the outcomes of these surveys.

Pollsters often rely on random sampling methods to recruit participants, as discussed in [@ todo], ensuring a diverse and representative group. Some pollsters use online platforms, while others use phone interviews or in-person surveys. The dataset captures these variations through variables such as `methodology` and `sample_size`. Sampling errors, response biases, and adjustments (e.g., weighting respondents based on age, race, or geographic region) are part of how this world phenomenon is translated into data entries.

The process begins with identifying the population of interest—likely voters or registered voters—and applying statistical adjustments to the raw responses. Pollsters handle non-response through weighting or imputation methods, ensuring that missing data or low response rates do not skew the results significantly. The column `numeric_grade` reflects how well these pollsters adhere to rigorous methodology, allowing us to filter for only the most reliable sources in our analysis.

2.3 Outcome variables

2.3.1 Percent Support (pct)

The primary outcome variable in this dataset is the percentage of respondents who support a given candidate. This is represented by the pct variable, which is reported by pollsters for each candidate they survey. For Kamala Harris, this percentage reflects her standing in polls, and variations in this metric can indicate shifts in voter preference over time.

Below is a summary table showing the average, minimum, and maximum percent support for Kamala Harris across high-quality polls:

Some of our data is of penguins (Figure 1), from Horst, Hill, and Gorman (2020).

Talk more about it.

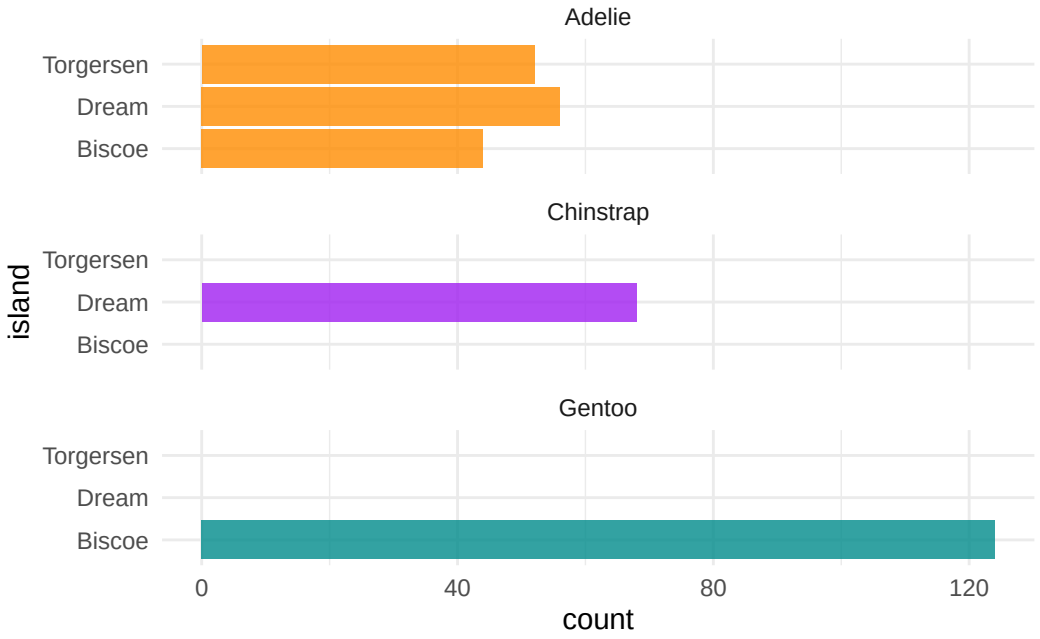

Figure 1: Bills of penguins

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

2.4.1 Pollster Name (display_name)

This variable represents the pollster responsible for each entry in the dataset. Different pollsters may use varying methods and possess differing levels of reliability, which the numeric_grade helps quantify. High-quality pollsters (numeric_grade 3.0) are emphasized in this analysis to ensure the robustness of the forecast.

2.4.2 Internal Polling (internal)

The variable internal is a binary indicator that identifies whether a poll was commissioned by a political campaign or interest group. Internal polls can sometimes be biased, either in how respondents are selected or how results are reported. However, the filtering for high-quality pollsters ensures that even internal polls meet certain standards of rigor.

2.4.3 Partisan Polling (partisan)

Like internal polling, partisan polling can introduce bias, as it may be sponsored by organizations with a vested interest in the election outcome. The partisan variable, similar to the internal variable, flags such polls. In the dataset, these are rare but still important to identify, especially when making impartial predictions.

2.4.4 Party Affiliation (party)

The party variable indicates the political party that the poll respondent supports. This is key to understanding the partisan leanings of the sample. For Kamala Harris, her affiliation with the Democratic Party is represented in this column.

Below is a table summarizing the distribution of party affiliations in polls that include Kamala Harris:

2.4.5 Race ID (race_id)

This variable identifies the electoral race or contest being polled. In this analysis, the focus is on the 2024 presidential election. Tracking results across different races (i.e., primary versus general election) helps in understanding shifts in voter opinion throughout the campaign.

In summary, the dataset offers a rich variety of variables that capture voter sentiment, pollster reliability, and electoral dynamics. By carefully filtering and analyzing this data, we can build a robust model to forecast the outcome of the 2024 US Presidential election.

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 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 Additional data details

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

Figure 2: **?(caption)**

B.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

Figure 3: **?(caption)**

References

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

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/.

Ryan Best, Aaron Bycoffe. 2024. "National: President: General Election: 2024 Polls." *FiveThirtyEight*. https://projects.fivethirtyeight.com/polls/president-general/2024/national/.