Red, Blue, or Purple? 2024's Battleground States Tell a Story*

Pennsylvania, Arizona, and Nevada Become Game Changers as Trump Gains Ground

Yingke He Ziheng Zhong

October 31, 2024

This paper analyzes polling data for the 2024 U.S. Presidential election, examining the key candidates notabily Donald Trump and Kamala Harris to gain a comprehensive perspective on the race. A Hierarchical Bayesian Model is applied in this study and the findings reveal significant variations across states, with [To be updated…] emerging as particularly influential battlegrounds that ultimately tilt in favor of [To be updated..].By highlighting these trends, this research underscores the dynamic and evolving nature of U.S. political landscapes and the crucial role that swing states play in determining election outcomes.

Table of contents

*Code and data are available at: https://github.com/iJustinn/Election_Prediction.

1 Introduction

In the lead-up to the 2024 U.S. Presidential election, understanding voter behavior across states and demographic groups is essential for accurately predicting electoral outcomes. Recent polling data from various sources suggests a close race among major candidates, including Donald Trump, Kamala Harris, and other key contenders. This study applies a Hierarchical Bayesian Model to analyze polling data across all U.S. states, with a focus on identifying critical swing states where even slight shifts in voter sentiment could prove decisive. By exploring trends in voter support across different states and demographic groups, this research aims to highlight the regions most likely to influence the final election outcome.

The primary objective of this study is to estimate state-level vote shares for each candidate while accounting for factors such as regional variations, pollster-specific biases, and the timing of polls relative to Election Day. Predictors including polling percentages, sample sizes, and days remaining until the election are incorporated into the model, which adjusts for state and pollster effects. The model's structure enables the estimation of vote shares that reflect the underlying distribution of polling data across diverse regions in the U.S.

Initial findings reveal variations in candidate support across states, with certain swing states emerging as pivotal in determining the Electoral College result. The model's projections suggest a likely lead for Trump in the Electoral College, underscoring the influence of large-sample, recent polls on forecast accuracy. This study's results emphasize the importance of swing states in the electoral process, demonstrating how regional dynamics and polling methodologies impact predictions.[To Be updated…]

This research contributes to the field of election forecasting by combining aggregated polling data with a robust modeling approach, offering valuable insights for political analysts, campaign strategists, and policymakers. By identifying critical swing states and accounting for potential polling biases, this study equips stakeholders with the tools to anticipate voter shifts and strategize effectively in an evolving political landscape.

The structure of the paper is organized as follows: following Section 1, Section 2 outlines the data collection and cleaning process, along with a description of the outcome and predictor variables used in the analysis.@sec-model, introduces the forecasting models and discuss the rationale behind choosing these models for election outcomes prediction. Section 4 then presents the main findings, including a breakdown of state-level and pollster-level random effects. Finally, Section 5 interprets the results, highlighting significant trends and predictions, and concludes with a discussion on the reliability of the forecasts and potential limitations of the models.

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023)…. Our data (Toronto Shelter & Support Services 2024)…. Following Alexander (2023), we consider…

Overview text [To be updated…]

2.2 Data Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.[To be updated…]

2.3 Outcome variables

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

Table 1

Variable
State
Pollster
Candidate Name
Percentage (pct)
Sample Size
Start Date
End Date

2.4 Predictor variables

This study includes several key predictor variables that provide insights into regional voting patterns, polling methodologies, and the timing of voter sentiment shifts. Each variable plays a unique role in forecasting vote shares for the 2024 U.S. Presidential election. Below are detailed descriptions of each predictor.

2.4.1 State

State (`state`) is a categorical variable that represents the U.S. state in which each poll was conducted. States have distinct political dynamics, influenced by regional issues, demographics, and historical voting patterns. Including the state variable allows the model to account for these differences, enabling predictions that are sensitive to each state's unique electoral landscape.

2.4.2 Pollster

Pollster (`pollster`) identifies the organization conducting each poll. Different polling organizations may use varied methodologies, such as sample selection, weighting, and question phrasing, which can introduce systematic biases. By including the pollster as a predictor, the model can adjust for potential biases or methodological differences, providing more accurate and standardized vote share estimates.

2.4.3 Candidate Name

Candidate Name (`candidate_name`) is a categorical variable that identifies the candidate for whom each poll measures support. Including candidate names allows the model to assess each candidate's baseline popularity and identify variations in support across different demographics

and states. This variable enables comparisons between major contenders, such as Donald Trump and Kamala Harris, as well as other candidates, to forecast overall vote shares.

2.4.4 Percentage of Votes

Percentage of Votes (`pct`) represents the percentage of respondents in each poll who indicated support for a specific candidate. As a primary predictor, this variable directly informs the model about the current level of support for each candidate at the time the poll was conducted. By converting these percentages into vote share estimates, the model can project likely election outcomes based on current polling data.

2.4.5 Sample Size

Sample Size (`sample_size`) denotes the number of respondents included in each poll. Larger sample sizes generally increase the reliability of a poll, as they reduce the margin of error and are more likely to represent the broader population. By incorporating sample size as a predictor, the model can weigh polls based on their reliability, giving more importance to larger, more robust samples.

2.4.6 Days to Election

The variable days to election represents the days remaining from the poll's end date to Election Day. It is calculated using the `end_date` variable, by subtracting the end_date from the election date, by assuming the election date is November 5 2024.

3 Model

To predict the actual election vote share for each candidate in each state while accounting for variations by pollster and other poll-specific factors. Background details and diagnostics are included in Appendix B. [To be updated…]

3.1 Model set-up

A Hierarchical Bayesian Model is utilized to predict the actual election vote for each candidate in each state while accounting for variables by pollster.

The model prediction utilizes the following predictor variables:

- **State** (`state`): Include as a categorical term to capture regional variations.

- **Pollster** (`pollster`): Include as a categorical variable for different polling effects.
- **Candidate Name** (`candidate_name`):Include as a categorical feature or model separately for each candidate.
- **Percentage of Votes by Poll in State** (`pct`): Use as a primary predictor of actual vote share, with a smooth term to allow for non-linear effects.
- **Sample Size** (`sample_size`): Include as a predictor or weight to reflect poll reliability.
- **Days to Election** (`end_date`):capture trends in support leading up to the election.

Let:

y_{ijk} : The target variable, representing the actual election vote share for candidate k in state

pct_{ijk} : The observed polling percentage for candidate k in state i by pollster j.

sample_size_{ijk} : The sample size of the poll, which helps in weighing the poll reliability.

$\text{days\_to\_election}_{ijk}$: Derived as the days remaining until the election from the poll's end date, capturing the

The model takes the form of the following equation:

$$y_{ijk} = \alpha_i + \beta_j + \gamma_k + \delta \cdot \text{pct}_{ijk} + \eta \cdot \text{sample\_size}_{ijk} + \theta \cdot \text{days\_to\_election}_{ijk} + \epsilon_{ijk} \qquad (1)$$

where:

$\alpha_i \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$: State-level random effect for each state i, capturing regional variations in voting patterns

$\beta_j \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2)$: Pollster-level random effect for each pollster j, accounting for systematic biases or differer

$\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$: Error term, accounting for random noise.

γ_k : Candidate fixed effect for each candidate k, representing baseline support across states and pollsters.

δ : Coefficient for Percentage of Votes by Poll (pct_{ijk}), reflecting how poll support translates to actual vot

η : Coefficient for Sample Size (sample_size_{ijk}), weighing polls based on their reliability.

θ : Coefficient for Days to Election ($\text{days\_to\_election}_{ijk}$), capturing the trend in support as the election c

3.1.1 Interpretation of Parameters

α_i : Captures state-specific effects, allowing the model to adjust the baseline vote share prediction based on

β_j : Accounts for systematic biases or differences in methodologies across pollsters.

γ_k : Provides an overall baseline effect for each candidate, independent of state or pollster.

δ : Measures how closely polling support translates to actual vote share.

η : Adjusts the model's sensitivity to polls based on their sample size, giving more weight to larger polls.

θ : Captures how support trends change as the election date approaches.

3.1.2 Prior Distributions

$$\begin{aligned}
\alpha_i \sim \mathcal{N}(0, 2.5): &\quad \text{State-level random effect prior for each state } i.\\
\beta_j \sim \mathcal{N}(0, 2.5): &\quad \text{Pollster-level random effect prior for each pollster } j.\\
\gamma_k \sim \mathcal{N}(0, 2.5): &\quad \text{Candidate fixed effect prior for each candidate } k.\\
\delta, \eta, \theta \sim \mathcal{N}(0, 1): &\quad \text{Coefficients for polling percentage, sample size, and days to election.}\\
\sigma \sim \text{Exponential}(1): &\quad \text{Prior for the standard deviation of the error term.}
\end{aligned}$$

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`.

3.1.3 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 **?@tbl-modelresults**.

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 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…

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…

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/.

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

Toronto Shelter & Support Services. 2024. *Deaths of Shelter Residents*. https://open.toronto.ca/dataset/deaths-of-shelter-residents/.