US Presidential Election Data Analysis*

My subtitle if needed

Xiaolu Ji Lequan Li Yuehan Dai Dannie

October 21, 2024

This paper explores the analysis of Donald Trump's polling performance across national and state-specific polls, focusing on data from high-quality pollsters with a rating of 2.7 or higher. Following the approaches laid out by Blumenthal (2014) and Pasek (2015), we examine the use of polling data to predict political outcomes. Polls, forecasts, and aggregators, as discussed by Blumenthal, serve as a foundation for understanding how different pollsters contribute to the accuracy of predictions. Pasek's work on pooling polls further informs the analysis, highlighting the importance of combining results from multiple sources to mitigate individual biases and enhance accuracy. The key variables in this study include Trump's polling percentage, the credibility of pollsters, and the poll type (national or state-specific). A histogram is used to illustrate the distribution of Trump's polling percentages, and a box plot analyzes variability across different pollsters. This study highlights the importance of understanding the relationships between pollster reliability, geographic specificity, and candidate performance. By building on existing literature, this analysis provides insights into how polling data can be used to assess political trends, enhancing our understanding of polling accuracy and its implications for electoral forecasting. Further research could extend these findings by examining additional candidates and broader polling data sets.

1 Introduction

Introduction

In the highly interconnected and fast-changing political landscape characterizing the United States, presidential elections turn out to be a focal point of global attention-they reflect broader socio-economic trends and shifts in public sentiment. In this regard, predictive analytics has grown exponentially in its role as a means for stakeholders to gain accurate insights

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

into voter behavior and election outcomes, from the many political analysts to the campaign advisors. Add to that the task of integrating these divergent signals into a coherent forecast that accurately predicts not only the popular vote but also the ultimate electoral college outcome.

This paper proposes a sound statistical method of election forecasting, which is done by aggregating data from many different polls-a "poll-of-polls"-in order to create such a predictive model of the U.S. presidential election. The key challenge will be to transcend the individual biases and variances of the polls, which so often distort perception and decision processes. While polling data is plenty, there exists a need to apply appropriate linear and generalized linear models, with due attention to aggregated poll data peculiarities, particularly in forecasting electoral college outcomes.

Our work fills this void by building a model considering the popular vote but deeply integrating an analysis of the electoral college system, hence providing a dual perspective on the election outcome. We employed state-of-the-art statistical methods to aggregate and analyze polling data, combined with demographic trends, economic indicators, and historical voting patterns, as a way of refining our predictions. These results present key determinants of voter preference and changes in the possible electoral landscape that are crucial in the prognosis of election results.

The importance of this study consists in the contribution to the strategic planning of political campaigns and a deeper understanding of democratic engagement in an era of extreme polarization. Our work enables stakeholders to make informed choices that best resonate with varied voter constituencies by providing them with a truer reflection of the electoral process and possible outcomes.

The paper first describes the data sources and methodology used to develop the forecasting model, then describes the model per se. Further sections review the results, trying to derive insights from their implications for future elections, and conclude with an overall analysis of the strengths and weaknesses concerning the model itself. Appendices provide extensive detail in the technical documentation of methodologies used, along with a hypothetical survey designed for further validation of our model's predictions.

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)…. Our data (Toronto Shelter & Support Services 2024)…. Following Alexander (2023),

We use the statistical programming language R (R Core Team 2023) to analyze polling data for Donald Trump, focusing on his support across various national and state-specific polls. Our data consists of key variables such as pollster, display_name, numeric_grade, pct (Trump's polling percentage), and state. These variables give insight into the credibility of pollsters, the level of support Trump received, and whether the poll is national or state-specific. By filtering the data to include only high-quality pollsters, with a numeric grade of 2.7 or higher, and focusing solely on Donald Trump, we ensure that the data is reliable and focused on a specific outcome.

Summary statistics (see Plot 2) for Trump's polling percentage (pct) show both the central tendency and variability of his support across these high-quality pollsters. The mean polling percentage provides a snapshot of Trump's average support, while the median highlights the middle point, and the standard deviation reflects the variation in polling percentages. The numeric_grade variable further adds credibility to the analysis, allowing us to filter out lower-rated pollsters and focus on those that meet higher standards of accuracy. Higher ratings reflect more reliable polls, influencing how we interpret Trump's polling performance.

A histogram of Trump's polling percentages (Plot 1) visualizes the distribution of his support across different polls. This plot reveals the concentration of Trump's polling percentages and indicates whether his support is consistently falling within a particular range or varies widely across different polls. A more concentrated distribution would indicate uniform support across pollsters, while a wider spread suggests greater variability. This histogram sets the stage for a deeper exploration of the relationships between variables in the dataset.

2.2 Measurement

The primary measurement variables in this analysis are pollster, which represents the organization conducting the poll, and numeric_grade, which measures the credibility or reliability of these pollsters. The pollster variable is particularly relevant for examining whether specific organizations consistently report higher or lower levels of support for Trump. This could signal potential biases or differences in polling methodologies across organizations. Meanwhile, numeric_grade serves as a critical filter, ensuring that only credible pollsters are included in the analysis. This adds confidence in the accuracy of the results, especially when evaluating a sensitive outcome such as political support.

By focusing on these measurement variables, we can assess whether the pollster and pollster quality (as measured by numeric_grade) are related to the variation in Trump's polling percentages. Further analysis will reveal how these factors impact the outcome variable—Trump's polling percentage—and whether they help explain differences in support across various polls.

2.3 Outcome variables

The key outcome variable in this analysis is Donald Trump's polling percentage (pct), which is used to measure his level of support across different polls. To better understand this variable, we start by examining it through a box plot (Plot 2), which displays Trump's polling percentages across the top five pollsters based on the number of polls conducted. This box plot illustrates the variability in Trump's support across different pollsters, showing the central range of polling percentages for each pollster and highlighting any outliers. Pollsters with a narrower interquartile range indicate greater consistency in their polling results, while those with wider ranges show more variation. This analysis helps identify which pollsters are more reliable and whether there are significant differences in how Trump's support is measured across organizations.

Additionally, we examine whether Trump's support differs between national and state-specific polls. By comparing these two categories, we can determine whether Trump performs better in national polls versus state-specific ones. For instance, if national polls show consistently higher support for Trump, this could suggest that his popularity is stronger on a national level than in individual states, or vice versa. Understanding these distinctions helps interpret broader trends in Trump's polling data and provides insights into his regional appeal.

By organizing the analysis into the sections of overview, measurement, and outcome variables, we ensure a comprehensive understanding of Donald Trump's polling performance across high-quality polls. The visualizations and summary statistics contextualize the data, offering insights into how factors such as the pollster and poll type influence Trump's polling percentages. Further details and additional graphs could be included in appendices for a more in-depth examination of specific variables or relationships.

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

Talk more about it.

And also planes (Figure 2). (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.

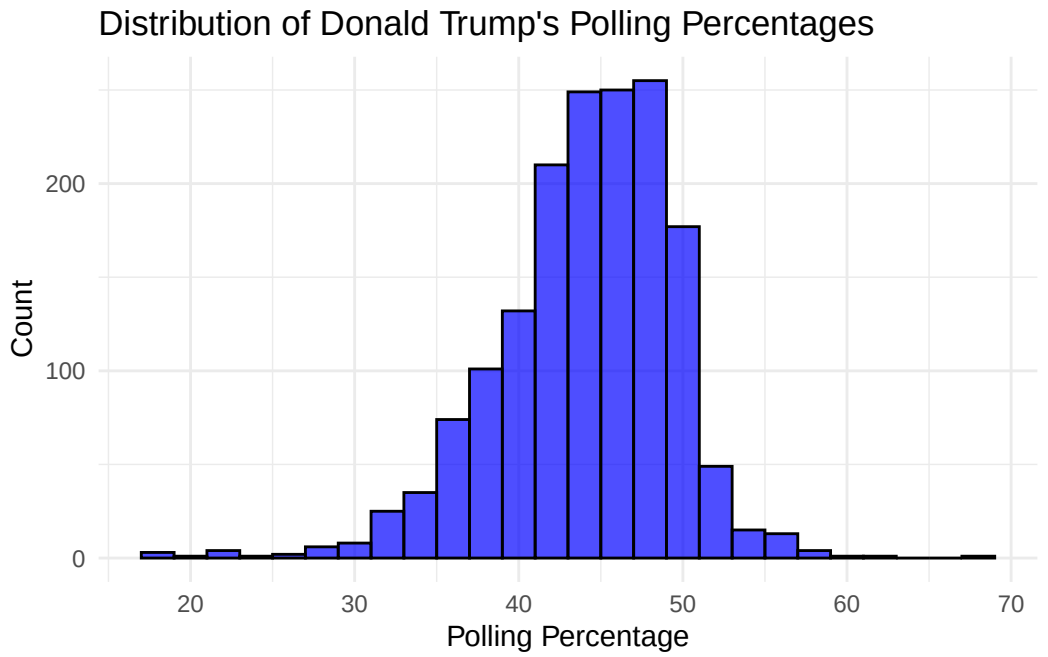

Figure 1: Bills of penguins

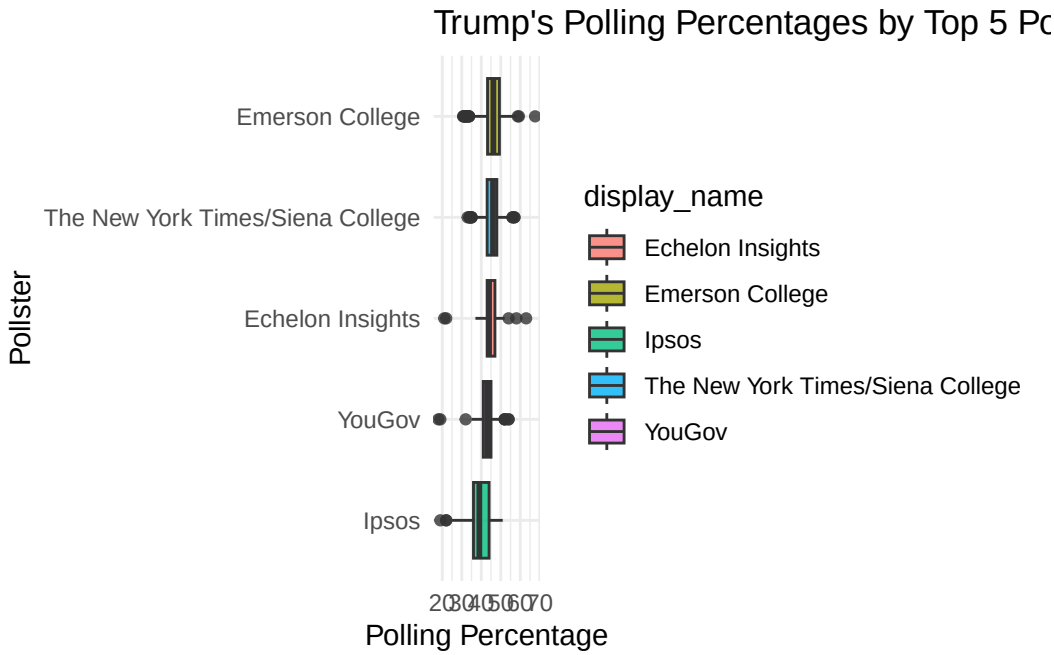

Figure 2: Relationship between wing length and width

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 **??**.

5 Discussion

5.1 Research framework and research design In this paper, we analyzed polling data about Donald Trump's support across various high-quality pollsters by using R language for statistical computations. Filtering high-quality pollsters and analyzing several key variables, such as approval ratings and different states, allowed us to understand the percentage of support Trump received, and how that support in states and across the whole country, while keeping the data reliable. Some charts like histograms and box plots are used to show the distribution and variability in Trump's polling percentages.

5.2 An insight gained from the research From this analysis, we learn that Donald Trump's polling percentages demonstrate variability depending on the different kinds of pollsters. This demonstration reflects potential differences in investigation methods and internal biases existing among all of the polling organizations. This insight shows how the reliability of polls and the context in they are conducted (nation versus specific states) can influence public attitudes and the final results of political support.

5.3 Another insight gained from the research Another important thing we learned is the importance of focusing on pollster credibility. The decision that we only used the data of high-grade pollsters in this analysis helped filter unreliable data. This could be generated as a broader conclusion that appropriate and strict selection of credible data sources is crucial in the process of reducing noise and improving the accuracy of analysis of prediction in political polling.

5.4 Weaknesses and next steps Even though the paper uses credible pollsters and provides a detailed examination of Trump's polling percentages, an obvious limitation is that the only research candidate is Donald Trump, which potentially overlooks wider trends in the overall political situation. Additionally, filtering data sources with low scores for the variable numeric_grade may ignore smaller but potentially insightful pollsters that have yet to establish a strong track record. Moreover, the paper could benefit from a more active interaction of different factors. For example, how regional political situations or population changes interact with polling data. By aiming at the weaknesses we mentioned in the last part, future research should expand from a single candidate to explore how our research models could be generated to multiple candidates, including their political movements. Additionally, by combining tracking polling data over time, our research could lead to deeper insights into how specific events that happened during a particular period influence the trends of public polling. In conclusion, analyzing the impact of demographic, geographic factors, and the credibility of pollsters would help improve the degree of comprehensive understanding of the main affecting factors behind polling variability.

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

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 algo-
rithm

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/.

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