Predicting US Presidential Elections*

Analysing Electoral Polls to Model Forecast the Winner of the Upcoming US Presidential Election

Aamishi Avarsekar Gauravpreet Thind Divya Gupta

November 4, 2024

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

Table of contents

[*] Code and data are available at: https://github.com/aamishi/marriages_weekly_reflection/tree/master

1 Introduction

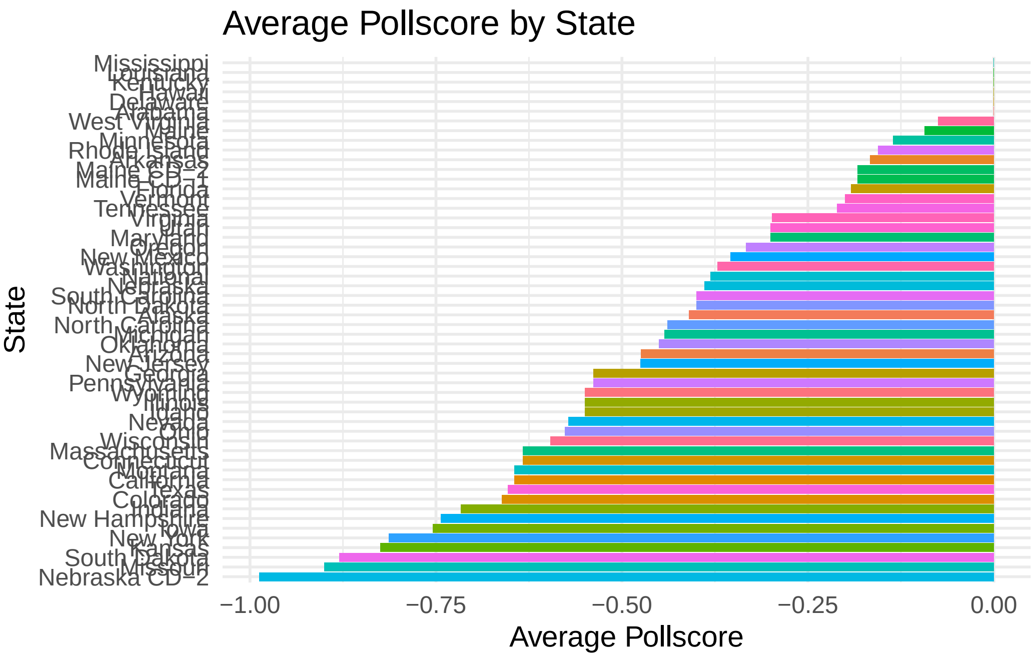
2 Data

2.1 Data Overview

The data that we are working with for this paper has been obtained from 385(R Core Team (2023)), using the Presidential General Election Polls (Current Cycle) data. The website also contains the data for the previous US Presidential Election from 2020 and Polling averages for predicting the next President. The website does not offer an API so the data must be downloaded manually and is available in .csv format. We have employed the dplyr package (Wickham et al. (2023)) to select the variables of our interest by omitting unused unnecessary variables. We used the janitor package (Firke (2023)) to clean all variable names for uniformity. For the creation of new variables or the mutation of existing variables, we used Grolemund and Wickham (2011) package's mdy() function. The csv data was converted to a dataframe in and subsequently converted back to a parquet file using Richardson et al. (2024) for ease of use.

The original data set provides 52 variables that reflect the data that each pollster has collected. These include each pollstor's sponsor, their respective identification numbers, the duration of the poll, the methodology used, the percentage of support received by each leading candidate, and numerics that evaluate the reliability and strength of the poll. 385 determines "major" candidates against a matrix of evaluation and aggregates the polls that focus on these candidates ("538's Polls Policy FAQs" 2024).

The website uses a poll-of-polls method to calculate the support for a major candidate. Using this method increase reliability by reducing the effect of individual bias for each pollster and noise across the data.



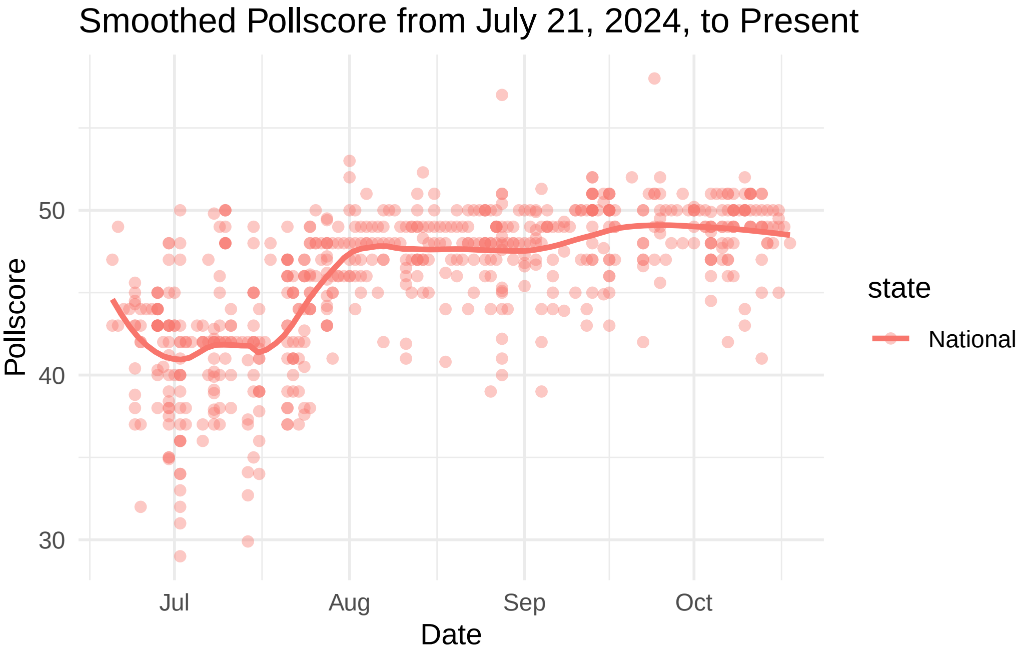
2.2 Variables

2.2.1 Outcome Variable

The outcome variable that we decided to focus on is `pct`, which is the proportion of support that a major candidate is projected to receive through that poll. This proportion accounts for the candidate's support against all other candidates. We use this variable to estimate which candidate is likely to win based on popular vote.

```
# Filter data from July 21, 2024, until today
filtered_data <- kh_data %>%
  filter(end_date >= as.Date("2024-07-21") & end_date <= Sys.Date())

cutoff_days <- max(kh_data$end_days) - 120

# Filter the data to include only the past 4 months
filtered_data <- kh_data %>%
  filter(end_days >= cutoff_days, state=="National")

# Plot with the filtered data

ggplot(filtered_data, aes(x = end_date, y = pct, color = state)) +
  geom_point(alpha = 0.4) +  # Add points for reference
  geom_smooth(method = "loess", span = 0.3, se = FALSE) +  # Adjust span for more or less smo
  labs(title = "Smoothed Pollscore from July 21, 2024, to Present",
       x = "Date",
       y = "Pollscore") +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```



2.2.2 Predictor Variables

Our predictor variables are as follows:

- **pollster**: The pollster conducting the poll. Each pollster also had an associated sponsor candidate, but this occurrence was rare, so we chose to omit it.
- **pollscore**: The score assigned to the reliability of the respective pollster. It includes error and bias that can arise within each pollster's methodology; hence, negative numbers are better. The original website describes it as "Predictive Optimization of Latent skill Level in Surveys, Considering Overall Record, Empirically." `pollscore` is used to assign weight, as we will see later in the model section.

- **transparency_score**: A numerical value, graded on a scale of 1 to 10, that indicates how transparent a pollster is based on the amount of information they disclose and how recent their polls are.
- **end_days**: The number of days that each poll was conducted until November 3rd. This number was generated using `start_date` and `end_date`, and it is used as the spline term for our model with 3 degrees of freedom. This variable is used to assign greater weight, as we will see in the model section.
- **state**: The United States state where the poll was conducted. This variable has been modified to include a "national" category if the poll was not specific to any particular state.

3 Model

3.1 Model Overview

This model employs Bayesian analysis to fit the predictive model based on the current polling data. Spline functions have been used to fit the `end_days` variable, capturing the weight of more recent polls to reflect their potentially greater relevance in predicting support for Vice President Kamala Harris. The dependent variable, `pct`, represents the likelihood of Harris being elected as the next President of the USA, serving as the primary measure of voter support.

3.2 Model Equation

$$pct_i = \beta_0 + \beta_1 \cdot \text{pollscore}_i + \beta_2 \cdot \text{transparency\_score}_i + f(\text{end\_days}_i) + \beta_3 \cdot \text{state}_i + \epsilon_i$$

3.2.1 Where:

- pct_i = support for Kamala Harris in the (i^{th}) poll
- β_0 = intercept (baseline support when all predictors are zero)
- β_1 = coefficient for the pollster reliability score (`pollscore`)
- β_2 = coefficient for the pollster transparency score (`transparency_score`)
- $\text{end\_days}_i)$ = spline function capturing the non-linear effect of the recency of the poll (measured in days)
- β_3 = coefficients for the categorical variable representing state (each state would have its own coefficient)
- ϵ_i = error term (captures unexplained variance)

3.3 Predictor Variables

The predictor variables selected for this model include `pollscore`, `transparency_score`, and `state`, each chosen for their significance in understanding polling dynamics.

3.3.1 Pollscore

This variable reflects the reliability of the pollster, where lower scores indicate a more accurate historical performance. Including `pollscore` allows the model to account for biases and errors associated with different polling organizations, ensuring that more reputable polls are given greater weight in the analysis.

3.3.2 Transparency Score

The transparency score indicates how openly pollsters disclose their methodologies. This variable is critical because it helps assess the trustworthiness of the polls. A higher transparency score suggests a more reliable polling process, which can enhance the credibility of the results.

3.3.3 Spline for End Days

By incorporating a spline function for `end_days`, the model captures non-linear effects associated with the recency of the polls. This approach allows for a more nuanced understanding of how support for Harris changes over time, recognizing that more recent polls may reflect current voter sentiment more accurately.

3.3.4 State

Including state as a categorical variable helps account for regional differences in voter support. Each state may have unique demographic and political contexts that influence support for Harris, and capturing these variations is essential for improving the model's accuracy.

3.4 Model Priors

All variables in this model utilize default priors, with an additional weight applied to recent polls with higher `pollscore` values, emphasizing the importance of both reliability and recency in predicting electoral support for Kamala Harris.

3.5 Limitations

This model predicts the likelihood of Vice President Kamala Harris winning the upcoming USA Presidential Elections in 2024. This model is a simple representation of Kamala Harris's win or no-win. The model does not take into consideration that independent candidates could account for the lack of support for Harris but that does not accumulate into direct support for Trump. This model also only considers specific support for Harris. There is no exploration of the relationship between support for the Democrats and support for Harris. A state may have an overall Blue majority but not necessarily for Harris directly.

This model assumes a linear relation for the scores provided for understanding the reliability of each poll. pollscore and transparency_score are a general guideline to assess the reliability of a poll but they are based on historical evaluations of each pollstor. Thus, biases may still exists regardless. This model also has a simplified weight distribution for recent polls and polls with higher pollscores and transparency_scores. For examples, any poll within the 6 month period sense writing this paper is weighted at 1 and any older polls are weighted to be 0. This does not capture the right sentiment associated with general people's opinions as the election day nears.

This model also does not take into consideration swing states and only uses them as categorical variables. This was done by design at this risk of overfitting.

3.6 Next steps

Add a more educated method to assign weights to more reliable and more recent polls. For example, as Election Day nears, Trump has been focusing more on swing states like North Carolina ("US Election Live: Polls Show Close Race as Trump, Harris Hit North Carolina" 2024) for his last set of campaign speeches. This could potentially weigh more than polls taken slightly earlier but still within the 6 month period.

References

"538's Polls Policy FAQs." 2024. https://abcnews.go.com/538/538s-polls-policy-faqs/story?id=104489193.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with Lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

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

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to Apache Arrow*. https://github.com/apache/arrow/.

"US Election Live: Polls Show Close Race as Trump, Harris Hit North Carolina." 2024. https://www.aljazeera.com/news/liveblog/2024/11/2/us-election-live-polls-show-close-race-as-trump-harris-hit-north-carolina.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.