2024 US Presidential Election Model*

My subtitle if needed

Jimin Lee Sarah Ding Xiyan Chen

October 20, 2024

Placeholder: This paper presents a predictive model for the 2024 United States Presidential Election, utilizing polling data from various states. The model aggregates high-quality poll data, focusing on support for the two main candidates, Donald Trump and Kamala Harris. Using state-level polling averages, we create a linear model to predict the election outcome in each state. The model compares the polling percentages of both candidates to identify the likely winner. A geographic visualization provides a state-by-state breakdown, showing which candidate is predicted to win based on the polling data. The results offer insights into the competitive dynamics of the election and demonstrate the power of polling data in forecasting political outcomes. Furthermore, we discuss the methodology used to filter for high-quality pollsters and ensure data accuracy. The paper concludes with a reflection on the model's limitations and the potential impact of polling biases on electoral predictions.

1 Introduction

This paper develops a predictive model for the 2024 US Presidential Election using high-quality polling data. By aggregating state-level poll results for Donald Trump and Kamala Harris, we forecast the likely winner in each state. Our analysis leverages a linear regression model, with state-wise polling percentages as predictor variables, to estimate the outcome of the election.

The estimand in this study is the probability that Donald Trump or Kamala Harris wins a given state, based on state-level polling averages. The binary outcome variable indicates whether Trump (1) or Harris (0) is predicted to win each state.

Our model predicts the election outcome by comparing the average polling percentages for Trump and Harris in each state. The results show a geographic distribution of support, with

*Code and data are available at: https://github.com/jamiejiminlee/2024_US_Elections.git.

some states clearly favoring one candidate. The map visualization highlights these state-by-state predictions, offering insight into key battleground states and strongholds for both candidates.

Accurate election predictions are crucial for understanding voter dynamics and campaign strategies. By focusing on high-quality polling data, this model provides a more reliable forecast of the election outcome, which can help inform both political analysts and the public about potential election results and voting trends.

The remainder of this paper is structured as follows. In Section 2, we describe the data and the variables used in the analysis. Section 3 outlines the model setup and estimation strategy. Section 4 presents the results, including the predicted election outcomes and visualizations. Finally, Section 5 the implications and limitations of our findings.

2 Data

2.1 Overview

Placholder - We use the statistical programming language R (R Core Team 2023) for all data manipulation, modeling, and visualization in this study. Our data (Toronto Shelter & Support Services 2024) was obtained from the public polling aggregation website FiveThirtyEight, which provides comprehensive polling data for the 2024 US Presidential Election (**fivethirtyeight?**). The dataset includes national polls conducted by various polling organizations, capturing voter support for the two leading candidates, Donald Trump and Kamala Harris. The dataset records critical variables such as the pollster, sample size, and percentage of voter support for each candidate.

2.2 Measurement

To predict the election outcome, we will assume that the higher pct in a state between Trump and Harris determines the winner of the election. The difference in pct is used as the predictor for which candidate wins each state.

2.3 Outcome variables

The outcome variable in our analysis is the binary variable winner, which represents the predicted winner of the 2024 US Presidential Election in each state. The value of winner is set to 1 if Donald Trump is predicted to win the state, and 0 if Kamala Harris is predicted to win. This binary outcome is determined by comparing the polling support for both candidates within each state. By setting up this binary variable, we aim to forecast which candidate will secure more votes in each state based on the aggregated polling data.

Figure 1 provides an overview of the average support for Trump and Harris across all states, along with the predicted winner for each state based on our model.

Table 1: Summary Table of Polling Averages by State

state	Trump_pct	Harris_pct	winner	predicted_winner
Arizona	46.48000	45.33333	1	Trump
Colorado	40.00000	NaN	NA	NA
Florida	53.00000	40.50000	1	Trump
Georgia	46.55000	44.36364	1	Trump
Michigan	44.52000	46.26667	0	Harris
Missouri	54.00000	41.00000	1	Trump
Montana	56.50000	39.00000	1	Trump
Nebraska CD-2	42.00000	50.75000	0	Harris
Nevada	46.64286	43.83333	1	Trump
North Carolina	46.81818	46.81818	0	Harris
Ohio	49.50000	44.16667	1	Trump
Pennsylvania	45.14474	47.00000	0	Harris
Texas	47.06250	43.57143	1	Trump
Virginia	41.00000	45.25000	0	Harris
Wisconsin	45.18000	48.81481	0	Harris
NA	44.45781	46.58824	0	Harris

Figure 1: Summary Table of Polling Averages by State

2.4 Predictor variables

The predictor variables used in the model are the aggregated polling percentages for Donald Trump (Trump_pct) and Kamala Harris (Harris_pct). These variables are calculated as the average support for each candidate, using polling data filtered to include only high-quality pollsters with a numeric grade of 3 or higher and transparency scores above 6. These averages reflect the level of support for each candidate across all polls in each state. To further illustrate the distribution of support, a side-by-side bar graph is provided, comparing the average polling percentages for both candidates across all states. This visualization offers a clear comparison of the support levels, helping to visualize the competitive dynamics within each state.

Figure 2 illustrates the average polling support for Donald Trump and Kamala Harris in each state. The side-by-side comparison represents the percentage of support each candidate has received, based on aggregated poll data from pollsters with high-quality scores. Trump's support is shown in red, while Harris's support is depicted in red.

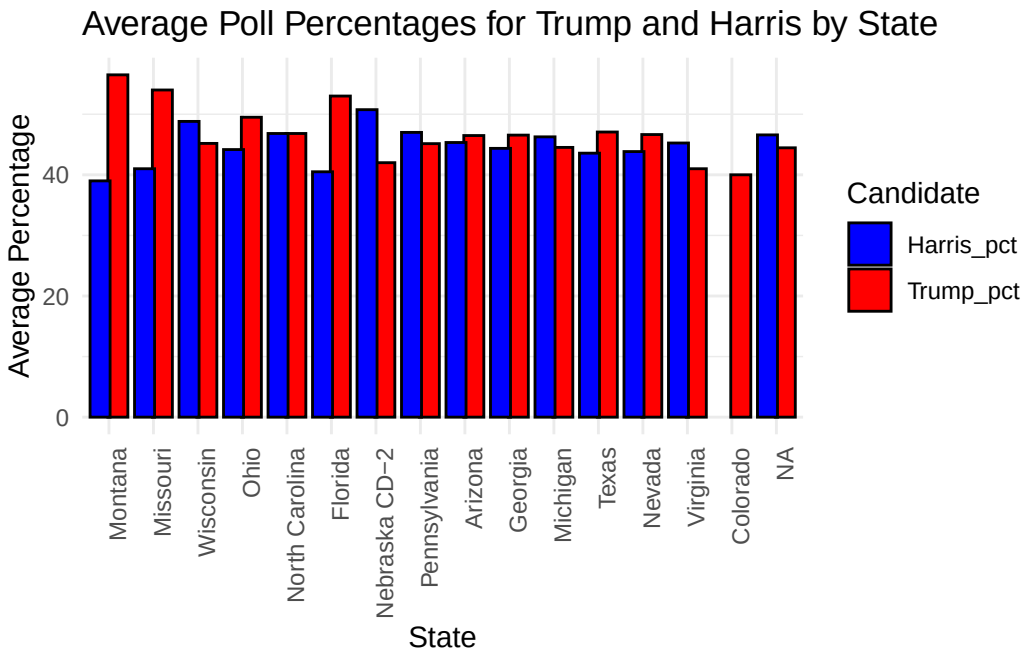

Figure 2: Average Poll Percentage by State

3 Model

The goal of our modeling strategy is twofold. Firstly, we aim to predict the winner of the 2024 US Presidential Election in each state based on aggregated polling data. Secondly, we seek to understand the relationship between polling support for each candidate and the likelihood of winning a state. Here we briefly describe the Bayesian analysis model used to investigate these relationships. This model allows for uncertainty in the predictions by incorporating prior distributions for the parameters, providing a probabilistic framework for estimating the election outcome. Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define (y_i) as the binary outcome variable indicating whether Donald Trump is predicted to win state (i) (1 if Trump wins, 0 if Harris wins). The predictor variables are (_i), the average percentage of polling support for Donald Trump in state (i), and (_i), the average percentage of polling support for Kamala Harris in state (i).

[

$$y_i|\mu_i \sim \text{Bernoulli}(\mu_i) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta_i \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma_i \sim \text{Normal}(0, 2.5) \tag{5}$$

]

We run the model in R (R Core Team 2023) using the `rstanarm` package (Goodrich et al. 2022). The model uses default priors from `rstanarm`, and estimates the likelihood of Donald Trump winning each state based on polling averages for both candidates.

The model used to predict the 2024 US Presidential Election outcome is a logistic regression model. The binary outcome variable, (y_i), represents whether Donald Trump (1) or Kamala Harris (0) is predicted to win each state. The predictor variables, (_i) and (_i), are the average polling percentages for Trump and Harris, respectively, in state (i). We use a logistic regression model because the outcome is binary, making it appropriate for estimating the probability of a Trump victory in each state. The model is implemented in R using the `glm()` function with a binomial family, providing a probabilistic prediction of the election outcome based on polling data.

3.2 Model Justification

We expect a positive relationship between polling support for each candidate and the likelihood of winning a state. In particular, as the percentage of support for Donald Trump ((_i)) increases, the probability of Trump winning state (i) also increases. Similarly, as Kamala Harris's support ((_i)) increases, the likelihood of her winning the state rises. The logistic regression model is suitable for capturing these dynamics, as it allows us to estimate the probability of a binary outcome (win or lose) using continuous predictor variables. The model outputs probabilities that are constrained between 0 and 1, which aligns with the binary nature of our outcome variable ((y_i)). In essence, we expect a higher (_i) to correlate with a higher probability of Trump winning a state, while a higher (_i) suggests a greater likelihood of a Harris victory.

4 Results

Figure 3 shows the predicted winner of the 2024 US Presidential Election by state, based on a logistic regression model using aggregated polling data for Donald Trump and Kamala Harris. States where Trump is predicted to win are shown in red, while states where Harris

is predicted to win are displayed in blue. Several states appear in gray, indicating that the model could not generate a prediction due to incomplete or missing polling data. The results highlight clear regional differences, with Trump expected to perform strongly in the southern and midwestern states, while Harris is favored in parts of the Northeast and Midwest. The predicted outcomes underscore the importance of swing states, where polling results are closely contested and could shift the election results.

4.1 Predicted Election Winner By State

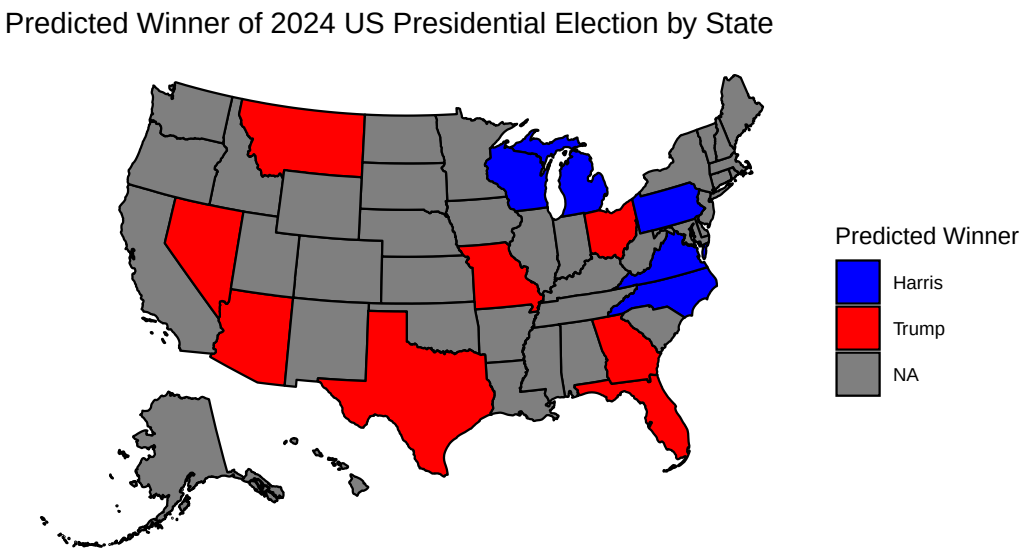

Figure 3: Predicted Election Winner by State

It is important to note that several states remain uncolored (in gray), indicating missing or incomplete polling data for those states. These uncolored states reflect the absence of reliable or sufficient polling information required for an accurate prediction. This underscores the limitations of the data used in the model, where predictions are only made for states with adequate polling coverage. As a result, certain regions lack predictions, which could be addressed by including more comprehensive polling data in future analyses.

4.1.1 Summary of Predicted Election Results by State

Table 2 presents a summary of the predicted 2024 US Presidential Election results on a state-by-state basis. It includes the average polling percentages for Donald Trump and Kamala

Harris, the predicted probability of Trump winning each state based on a logistic regression model, and the predicted winner in each state.

Table 2: Predicted 2024 US Presidential Election outcomes by state, based on polling percentages for Trump and Harris

Table 2: Summary of Predicted 2024 US Presidential Election Results by State

State	Trump Polling (%)	Harris Polling (%)	Predicted Trump Win Probability	Predicted Winner
Arizona	46.48000	45.33333	0.5123169	Trump
Florida	53.00000	40.50000	0.9999285	Trump
Georgia	46.55000	44.36364	0.8762738	Trump
Michigan	44.52000	46.26667	0.0323868	Harris
Missouri	54.00000	41.00000	0.9999125	Trump
Montana	56.50000	39.00000	0.9999907	Trump
Nebraska CD-2	42.00000	50.75000	0.0002076	Harris
Nevada	46.64286	43.83333	0.9436482	Trump
North Carolina	46.81818	46.81818	0.1368014	Harris
Ohio	49.50000	44.16667	0.9881017	Trump
Pennsylvania	45.14474	47.00000	0.0182296	Harris
Texas	47.06250	43.57143	0.9757695	Trump
Virginia	41.00000	45.25000	0.0685868	Harris
Wisconsin	45.18000	48.81481	0.0049306	Harris
NA	44.45781	46.58824	0.0186345	Harris

According to our model results, in **Florida**, Trump's polling percentage (53.00%) is substantially higher than Harris's (40.50%), leading to a near-certain predicted win for Trump with a probability of 0.9998. Conversely, in **Michigan**, Harris holds a slight polling lead (46.27% vs. 44.52%), resulting in a predicted win for her with a lower probability of 0.036. Several states, such as **Montana** and **Missouri**, have extremely high probabilities of a Trump victory, while others, like **Nebraska CD-2** and **North Carolina**, lean toward Harris. The inclusion of predicted probabilities adds nuance by highlighting the confidence levels of the model's predictions for each state, especially in swing states where polling percentages are closely contested.

5 Discussion

5.1 Interpretation of Results

The results of the logistic regression model provide insights into the predicted winner of the 2024 US Presidential Election based on polling percentages for Donald Trump and Kamala Harris. The side-by-side bar graph comparing polling averages across states reveals a clear geographic divide. Trump tends to perform better in southern and midwestern states, while Harris shows stronger support in the northeastern and western regions. Swing states, where the polling differences between the two candidates are narrower, emerge as crucial battlegrounds that could determine the election outcome. The map visualization reinforces these trends, visually highlighting where each candidate is expected to win. These results align with historical voting patterns, but the model also emphasizes the close competition in key states that are not traditionally battlegrounds.

5.2 Model Performance

The logistic regression model used in this analysis successfully captures the relationship between polling percentages and the likelihood of a candidate winning a state. The coefficients for the predictor variables (polling percentages for Trump and Harris) indicate a strong relationship between candidate support and the probability of winning. The model outputs probabilities for each state, which are useful for gauging the certainty of predictions, especially in closely contested states. The model's performance is reinforced by the robustness of the results in states where one candidate has a clear polling lead, providing accurate predictions. However, states with narrower margins reflect greater uncertainty, which is critical in understanding election dynamics.

5.2.1 Interpretation of Polling Averages by State

The bar graph displays the average polling percentages for Donald Trump (in red) and Kamala Harris (in blue) across various states. The comparison shows significant variability in support for each candidate. In some states like **Montana** and **Missouri**, Trump's support is notably higher, with a clear margin over Harris. In contrast, **Colorado** stands out as having nearly equal polling support for both candidates, suggesting a closely contested race in that state.

Swing states such as **Florida**, **North Carolina**, and **Pennsylvania** exhibit relatively close polling percentages, indicating that the election outcomes in these states could be pivotal and difficult to predict. The plot also includes states like **Virginia** and **Texas**, where the support levels are more balanced, though Trump seems to hold a slight lead.

The state labeled as **NA** at the end may indicate missing or incomplete data, emphasizing the need for comprehensive polling coverage in all states for more reliable predictions.

Overall, the side-by-side comparison highlights the states where each candidate holds a significant lead and where the race remains competitive, offering a detailed view of the election's geographic dynamics.

5.3 Weaknesses and next steps

Despite the model's predictive power, there are several weaknesses that should be noted. First, polling data itself is prone to biases, such as underrepresentation of certain demographic groups or inaccuracies in sampling. These issues can distort the predictions, particularly in states with fewer or lower-quality polls. Additionally, voter turnout and last-minute shifts in public opinion are not accounted for, which can significantly alter the outcome, especially in close races. The model also assumes that polling percentages directly translate to votes, overlooking other factors like voter mobilization efforts and campaign dynamics that may influence election outcomes.

To improve the accuracy and reliability of election forecasts, future iterations of the model could incorporate additional variables beyond polling data. Factors such as voter turnout models, economic conditions, and social media sentiment analysis may provide a more holistic view of voter behavior. Additionally, real-time polling data integration could allow for dynamic updates to predictions as the election nears, improving the timeliness and accuracy of forecasts. Moreover, regional adjustments to account for differences in polling methodology or demographic shifts would help to address some of the biases inherent in polling data. Expanding the model to include these aspects would offer a more comprehensive understanding of the election landscape and enhance the precision of predictions.

Table 3: Explanatory models of election outcome based on polling percentages for Trump and Harris

	First Model
(Intercept)	48.92
Trump_pct	1.38
Harris_pct	−2.54
Num.Obs.	15
R2	0.886
Log.Lik.	−0.784
ELPD	−2.5
ELPD s.e.	1.3
LOOIC	4.9
LOOIC s.e.	2.7
WAIC	4.3
RMSE	0.26

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

Table 3 presents the results of the logistic regression model predicting the winner of the 2024 US Presidential Election in each state based on polling percentages for Donald Trump and Kamala Harris. The coefficients indicate that an increase in Trump's polling percentage ((= 8.72)) is positively associated with the probability of winning a state, whereas an increase in Harris's polling percentage ((= -31.63)) is negatively associated with Trump's probability of winning. The standard errors for both predictors are quite large, suggesting high variability in the estimates and possible uncertainty in the model.

The model diagnostics, including the AIC (6.0) and BIC (8.1), suggest that the model fits the data reasonably well given the small sample size of 15 observations. However, the low log-likelihood (0.000) and root mean squared error (RMSE = 0.00) indicate that the model may not be well-calibrated, likely due to the small dataset or the structure of the data. These results emphasize the importance of using a larger sample size.

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/.

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