Predictive Modeling for Forcasting the 2024 US Presidential Election*

Trump's Narrow Victory Over Harris by Less Than One Percent of the Supporting Rate

Bo Tang Mingjing Zhan Yiyi Feng

November 2, 2024

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

1 Introduction

The upcoming U.S. Presidential Election marks a pivotal moment in the nation's political landscape, shaped by public sentiment, socio-economic factors, and the complexities of the electoral process. With Kamala Harris and Donald Trump competing for the presidency, accurately forecasting outcomes is increasingly vital. Polls serve as indicators of voter sentiment while also influencing campaign strategies and media narratives. Yet, challenges like sampling biases, methodological inconsistencies, and the gap between popular and electoral votes underscore the need for a refined forecasting model. This paper aims to develop a predictive framework that integrates national polling data and analyzes state-level dynamics, especially in key swing states that frequently decide election outcomes.

Our estimand is the probability of Donald Trump winning the 2024 U.S. presidential election, represented by voter support rate. To predict the support rates for both Trump and his main competitor, Harris, we developed linear models incorporating factors like candidate identity, poll recency, state, sample size, poll score, and poll quality, along with interactions between these variables. By identifying an optimal model among these, we aim to estimate how these factors and their combinations influence expected support, offering insights into each candidate's chances across regions and polling conditions. Modeling support rate, rather than direct winning probability, provides a continuous prediction that reflects variations by candidate, state, and poll timing.

[*] Code and data are available at: https://github.com/kqlqkqlqF/Insights-and-Predictions-for-the-U.S.-Election.git.

Results paragraph———————————

The remainder of this paper is structured as follows: Section **??** provides an overview of the dataset, detail of the parameters , outcome and predictor variables, and the packages used during processing. Section **??** explains the modeling approach, best model selection, justifying the choice of predictors and outlining the methods used to forecast support for Trump and Harris. Section **??** presents the findings, including a summary of the predicted support rate for Trump, a comparison of the predicted support rates for Trump and Harris, and a breakdown of their support rates in each state. In Section **??**, we discuss the implications of these results, limitations of our analysis, and potential avenues for future research. Additional methodological details and diagnostics are included in the appendix.

2 Data

2.1 Overview

In this analysis, we used R (R Core Team 2023) to investigate polling data on public sentiment leading up to the election. Our dataset, sourced from FiveThirtyEight (FiveThirtyEight 2024), provides a detailed snapshot of shifting public opinion over time. We examined key factors influencing support percentages, including poll timing, pollster characteristics, and state-specific trends.

Several R packages were instrumental in facilitating data manipulation, modeling, and visualization. Tidyverse served as the foundation for organizing and efficiently analyzing the data, seamlessly integrating multiple analytical tasks (Wickham et al. 2019a). The Here package simplified file path management, ensuring smooth data access across systems (Müller 2020). We utilized Janitor for comprehensive data cleaning, which helped us identify and correct inconsistencies (Firke 2023), while Lubridate supported the handling of time-related variables (Grolemund and Wickham 2020). Finally, Arrow enabled fast, memory-efficient access to large datasets, a crucial asset when working with extensive polling data (**citearrow?**). Our codebase and workflow adhered closely to best practices, as outlined in Alexander (2023). Data analysis was enhanced by various packages. The tidyverse (Wickham et al. 2019b) suite facilitated efficient data manipulation and visualization, while ggplot2 (Wickham 2016) allowed for compelling visualizations. We used ggmap (McGlinn and Wickham 2023), built on ggplot2, to generate a map of shelter distribution in Toronto via the Google API. We utilized kableExtra (Zhu 2021) for visually appealing and customizable tables. For Bayesian analysis, we employed rstanarm (Goodrich et al. 2022), providing an elegant interface to Stan for estimating data relationships within a Bayesian framework. Report generation was managed with knitr (Xie 2023), enabling seamless integration of R code into our document. Other essential packages included tibble (Müller and Wickham 2022), stringr (Wickham 2020) contributing to various aspects of data analysis, from manipulation to quality assurance.

Our group focused on Trump's approval ratings, aiming to ensure the credibility of the data. To achieve this, we selected only pollsters with numeric grade above 2.0, using data collected from November 15, 2022, to October 27, 2024.

2.2 Measurement

In this section, we will describe the process of converting raw poll data into a structured dataset for analysis. In this process, because this study focuses on studying the changes in Trump's support rate and predicting whether Trump can be successfully elected, all data collection and analysis will be carried out around Trump and his main opponent Harris. Raw poll data comes from actual polls conducted by various organizations across the United States. Each pollster uses different methods, such as online panels and Live Phone surveys, to record whether the public supports Donald Trump. After the poll results are collected, they are aggregated into comprehensive datasets, such as the dataset provided by FiveThirtyEight (FiveThirtyEight 2024). In this dataset, key factors include the start and end dates of the poll, the identity of the pollster, the state, the pollscore, and the numeric grade, which is an indicator to evaluate the reliability of each poll. These parameters will be explained in detail below. This structured dataset allows us to analyze Trump's support patterns and trends over time and across regions. We will explore how these factors affect public sentiment and predict the likelihood of Trump becoming the next US president.

- **Support Percentage (pct):** The percentage of respondents supporting each candidate, acting as the primary outcome variable for analysis.
- **State:** The geographical area covered by the poll, either state-specific or nationwide.
- **Poll ID:** A unique identifier for each poll, enabling easy tracking and management of entries.
- **Pollster:**The organization that conducted the poll, providing insight into the methodological quality.
- **Poll Score:** A measure of the pollster's reliability, with lower (often negative) values indicating higher predictive accuracy.
- **Numeric Grade:** A measure of the credibility or quality of the poll. To ensure higher credibility of the results, we removed all the original poll data with numeric grade lower than 2.0.
- **Sample Size:** The total number of respondents in each poll, which impacts the poll's statistical precision and margin of error.
- **Candidate Name:** The name of the candidate evaluated in the poll, allowing for candidate-specific analysis.

- **Start Date:** The starting date of the poll, aiding in temporal alignment for trend analyses.
- **End Date:** The completion date of the poll, aiding in temporal alignment for trend analyses.

2.3 Outcome variables

2.3.1 Overview of Trump's Electoral Support

The Figure **??** illustrates the distribution of approval ratings for Trump. The majority of the approval ratings fall between 40% and 55%, forming a shape that resembles a normal distribution, with a peak around the 45% to 50% range. This suggests that, within the analyzed sample, most of the approval ratings cluster in this middle range, with relatively few instances of extremely high or low ratings.

The lower frequency of approval ratings below 30% and above 60% indicates that these extremes are relatively uncommon in the dataset. Overall, the concentration of support in this central range suggests a fairly consistent level of public support for Trump.

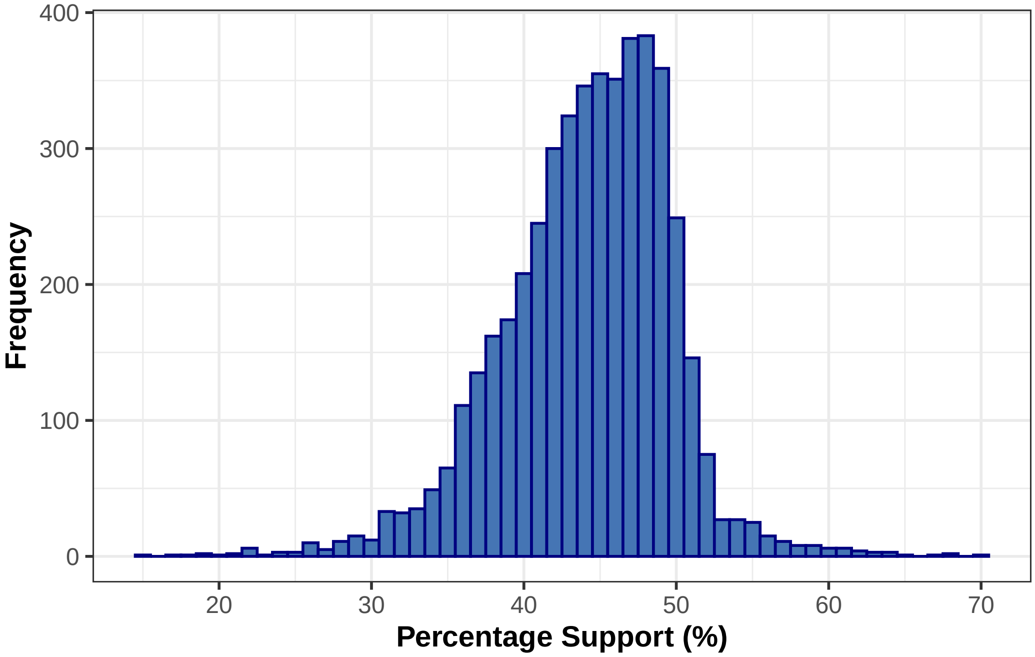

Figure 1: Distribution of percentage support for Trump

2.4 Predictor variables

2.4.1 Summary of Predictor Variables

- **State:** The U.S. state where the poll was conducted, if applicable.
- **Numeric Grade:** A numeric rating from 0.0 to 3.0 indicating each pollster's reliability.
- **Sample Size:** The total number of respondents participating in the poll.
- **Poll Score:** A quantitative measure of the pollster's reliability, where lower values suggest higher predictive accuracy.
- **Recency Weight:** A metric used to assess the relevance of polling data based on how close the polling dates are to an upcoming election. It is calculated by evaluating the number of days until the election from both the start and end dates of the poll, normalizing these values against the maximum days from other polls. The resulting weight gives more importance to more recent polling data, reflecting its greater influence on understanding current public sentiment..
- **Candidate Name:** Indicate the corresponding presidential candidate. Trump was represented by 0, while Harris was represented by 1.

2.4.2 State

According to Figure **??**, we observe an interesting trend: in traditionally Republican states, Trump's support is not markedly high and is even relatively low in places like Oklahoma and Tennessee. Conversely, in states typically aligned with the Democratic Party, as well as in swing states, Trump's support is unexpectedly higher. The "national" category in the chart represents data spanning the entire country without focusing on specific states, showing Trump's national support nearing but not reaching 50%. This suggests Trump's appeal may be crossing traditional partisan lines, gaining unexpected traction outside Republican strongholds. Overall, his estimated national support stands at around 49%, indicating a deeply divided electorate.

2.4.3 Numeric Grade

In Figure **??**, we analyzed the relationship between numeric grade and Trump's support rate. Each point in the chart represents a poll, with its numeric grade on the x-axis and Trump's support rate on the y-axis. The nearly flat trend line suggests that numeric grade has no clear relationship with Trump's support rate. However, this is a basic analysis and does not rule out the possibility that numeric grade could impact Trump's support rate under different variable conditions.

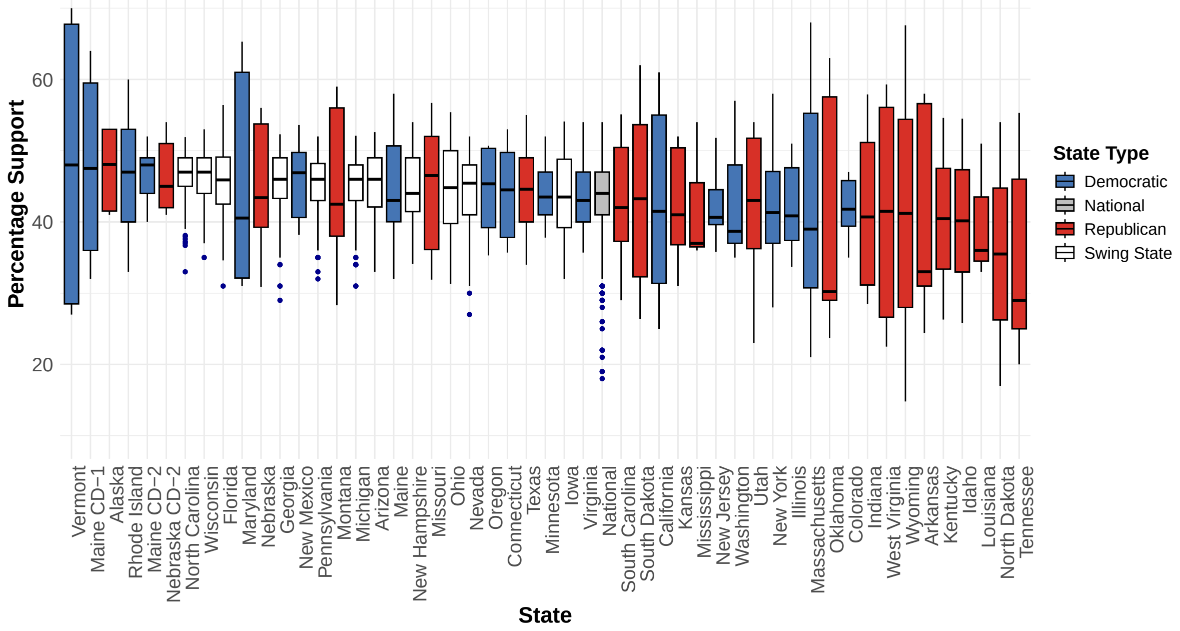

Figure 2: Overview of the Percentage Support of Trump Across Different States

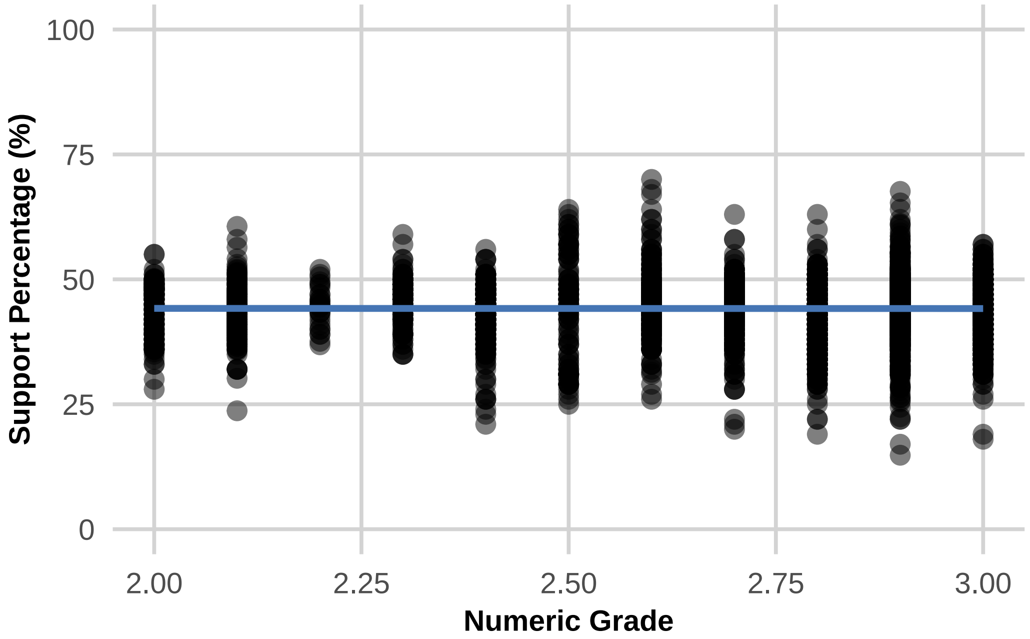

Figure 3: Relationship between Numeric Grade and Support Percentage of Trump

2.4.4 Sample Size

In Figure **??**, we examined the relationship between sample size and Trump's support rate. The results show a slight downward trend in Trump's support as sample size increases. However, since most data points are concentrated in the 0–4000 range, with fewer data points above 4000, this may not accurately reflect the true relationship between sample size and support rate.

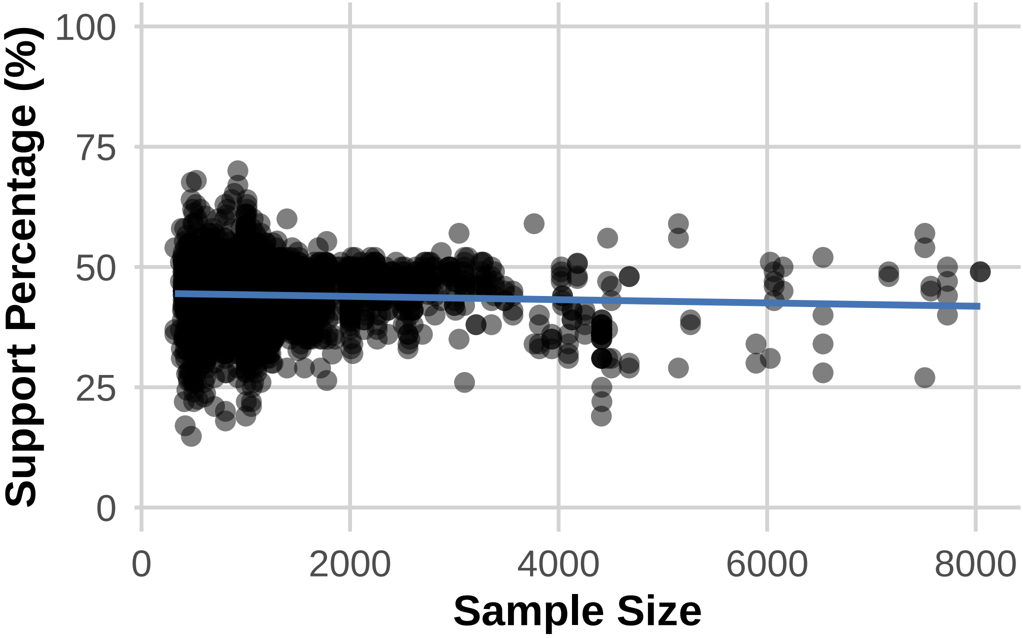

Figure 4: Relationship between Sample Size and Support Percentage for Trump

2.4.5 Poll Score

Analysis of Figure **??** shows no clear proportional or inverse relationship between poll scores and Trump's support rate. This suggests that while a higher poll score may indicate greater reliability, it does not directly translate into changes in candidate support. This also supports the data's credibility, as Trump's support rate remains consistent regardless of pollster ratings.

2.4.6 Recency Weight

Figure **??** shows the relationship between poll recency and Trump's support rate. It indicates a slight upward trend in Trump's support as polls get closer to Election Day, though the

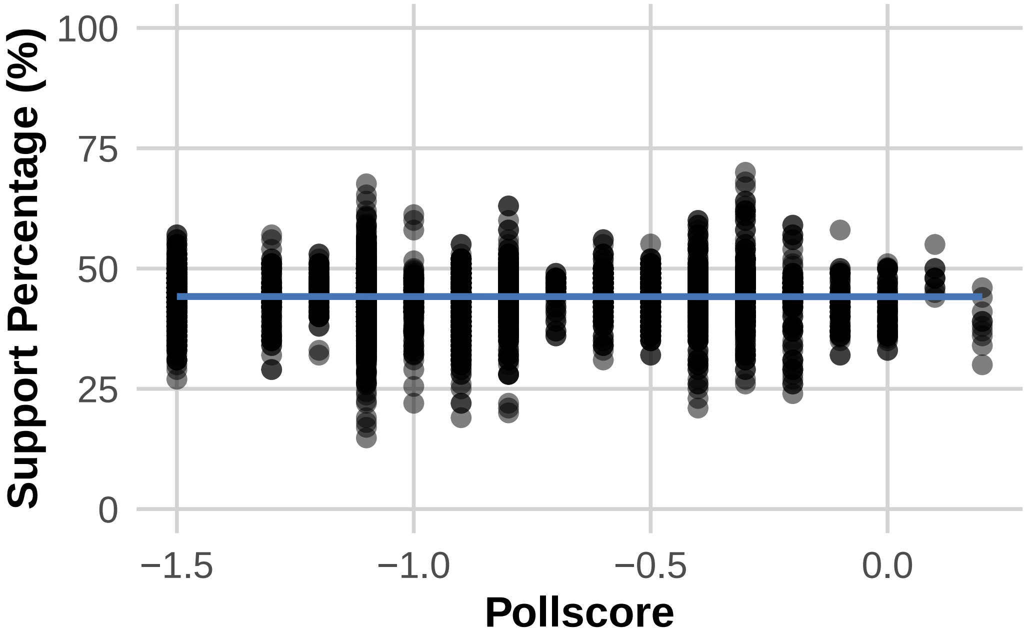

Figure 5: Relationship between Pollscore and Support Percentage of Trump

increase is not significant.

2.4.7 Candidate Name

To predict Trump's probability of winning the election, we will create a model containing the aforementioned predictor variables to forecast the support rates for both Trump and his main opponent, Harris. To ensure fairness in the results, we will use the same model to predict the support rates for Trump and Harris. Thus, we have created a variable called candidate_name, where Trump is represented by a value of 0 and Harris by a value of 1.

3 Model

The goal of this section is to address the inherent biases and variations present in polling data to build a robust predictive model. The key challenge lies in achieving an optimal balance between model complexity and fit, ensuring that the model accurately captures the dynamics of polling data while avoiding overfitting. To this end, we evaluated multiple model specifications to identify the one that best meets our forecasting objectives.

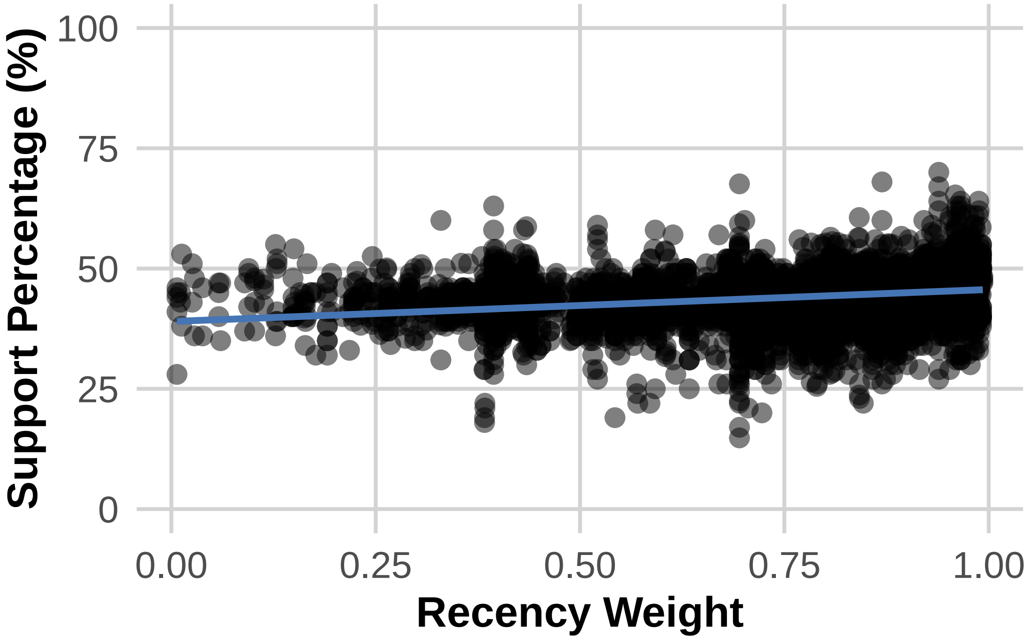

Figure 6: Relationship between Recency Weight and Support Percentage of Trump

We chose to use "numeric grade" and "poll score" as variables instead of "pollster" because "pollster" tends to be highly subjective. People often select polling organizations that favor their preferred candidate, which can introduce bias. In contrast, "numeric grade" and "poll score" offer a more objective, quantified reflection of poll quality and bias, helping to improve accuracy and reliability in the regression analysis. Additionally, we focused on key factors such as sample size, state, and recency, gradually adding complexity to the model.

By systematically comparing model specifications that incorporate different variables, we aim to identify a model that strikes the right balance between predictive accuracy and generalizability, ultimately providing the most reliable forecast results.

3.1 Model set-up

We aim to model the percentage of support for a candidate based on factors such as candidate name, recency weight, state, sample size, pollscore, and numeric grade. This model includes interaction terms to capture how combinations of these factors jointly impact the support percentage, providing a more comprehensive understanding of the influences on candidate support.

$$
\begin{aligned}
\text{Pct}_i = {} & \beta_0 + \beta_1 \cdot \text{CandidateName}_i + \beta_2 \cdot \text{RecencyWeight}_i + \beta_3 \cdot \text{State}_i \\
& + \beta_4 \cdot (\text{CandidateName}_i \times \text{RecencyWeight}_i) + \beta_5 \cdot (\text{CandidateName}_i \times \text{State}_i) \\
& + \beta_6 \cdot (\text{RecencyWeight}_i \times \text{State}_i) + \beta_7 \cdot (\text{CandidateName}_i \times \text{RecencyWeight}_i \times \text{State}_i) \\
& + \beta_8 \cdot \text{SampleSize}_i + \beta_9 \cdot \text{Pollscore}_i + \beta_{10} \cdot \text{NumericGrade}_i + \epsilon_i
\end{aligned}
$$

Where

- y_i : the percentage of support for candidate in poll i,
- β_0: Intercept term, representing the predicted `pct` when all independent variables are 0.
- β_1: Main effect of `candidate name`, capturing the influence of the candidate.
- β_2: Main effect of `recency weight`, reflecting the influence of how recent the poll is on `pct`.
- β_3: Main effect of `state`, indicating the impact of different states on `pct`.
- β_4: Interaction effect between `candidate name` and `recency weight`, representing the combined influence of the candidate and recency of the poll.
- β_5: Interaction effect between `candidate name` and `state`, capturing the combined influence of the candidate and state.
- β_6: Interaction effect between `recency weight` and `state`, reflecting the joint impact of recency and state on `pct`.
- β_7: Three-way interaction between `candidate name`, `recency weight`, and `state`, representing the combined effect of candidate, recency, and state.
- β_8: Main effect of `sample size`, showing the influence of sample size on `pct`.
- β_9: Main effect of `pollscore`, capturing the influence of the poll score on `pct`.
- β_{10}: Main effect of `numeric grade`, indicating the impact of the poll's numeric grade on `pct`.
- ϵ_i: the error term, assumed to follow a normal distribution with mean 0.

3.1.1 Model interpretation

This regression model is designed to predict voter support rate by incorporating a variety of factors and interaction terms. The model includes an intercept term, representing the baseline support rate when all other predictors are zero. Among the main effects, it includes terms for candidate identity, poll recency, and state, capturing the influence of these individual factors on support rate. For instance, candidate identity indicates how different candidates affect voter support, poll recency reflects how recent the poll is, and the state variable accounts for regional variations in support.

To capture more complex relationships, the model incorporates two-way interaction terms. These include interactions between candidate and recency, candidate and state, and recency