Polling Insights for the 2024 U.S. Election: Predicting Support for Donald Trump*

Investigating the impact of timing, geography, and methodology on voter sentiment

Jing Liang Jierui Zhan

October 22, 2024

This paper analyzes the factors influencing Donald Trump's percentage support in the lead-up to the 2024 U.S. presidential election, focusing on variables such as poll end date, state, pollster, and poll score. Our findings reveal significant variations in support levels based on geographic location and poll quality, indicating that higher-quality polls tend to report lower support for Trump. This research underscores the importance of critically evaluating polling methodologies to avoid misleading interpretations of public opinion. Ultimately, our work enhances understanding of the complex dynamics shaping voter sentiment, contributing to more informed political discourse in an election year marked by intense scrutiny of polling accuracy.

1 Introduction

In recent years, polling has become a critical tool for measuring public opinion and predicting political outcomes, particularly in U.S. presidential elections. As public reliance on polling data has grown, so has the need for scrutiny around its reliability and accuracy. Factors such as timing, geographic variation, pollster methodology, and poll quality can significantly affect reported levels of candidate support. For example, the timing of a poll relative to key campaign events or the methods used by different pollsters can cause variations in the support reported for candidates. These inconsistencies raise important questions about how polling differences, particularly across states and polling organizations, impact public opinion forecasts. In the context of the 2024 U.S. election, a clear understanding of these dynamics is crucial for accurate poll interpretation, particularly regarding Donald Trump's percentage support.

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

This paper aims to fill a gap in the literature by systematically analyzing how polling variables such as end date, state, pollster, and poll score influence Trump's support. The estimand in this analysis is the percentage support for Donald Trump as reported in polls. While previous studies have examined isolated factors like poll timing or specific pollsters, few have investigated how these variables interact. Using a linear regression model, we assess the contributions of each variable, exploring how factors such as time, geographic region, and poll quality influence polling results. Additionally, we account for potential non-linearities and omitted variables, such as demographic and economic factors, that may further explain variations in Trump's support.

Our findings identify several key influences on polling outcomes. State-level differences in support are significant, with some states consistently showing higher or lower levels of support for Trump. Poll score also plays a critical role, with higher-quality polls typically reporting lower support, indicating that less rigorous polls may overestimate Trump's backing. Additionally, the analysis reveals pollster-specific biases, where certain organizations systematically report higher or lower support, necessitating careful consideration of pollster effects. Diagnostic tests show potential violations of key model assumptions, such as heteroscedasticity and non-normality, suggesting that more sophisticated models may be needed to better capture these relationships. These results are crucial for improving the accuracy of election forecasts and ensuring more reliable interpretations of polling data.

The structure of this paper is organized as follows. Section 2 introduces the data sources and the key variables utilized in our analysis, offering a comprehensive overview of the dataset and how the variables were selected. Section 3 outlines the modeling strategy, including the linear regression framework, along with its underlying assumptions and the rationale for the inclusion of specific predictors like state, pollster, and poll score. Section 4 presents the model results, emphasizing the significant factors driving Trump's percentage support and analyzing any diagnostic issues such as heteroscedasticity and residual normality. Section 5 delves into the broader implications of our findings, discussing their relevance to polling accuracy, potential biases, and offering suggestions for future research to enhance predictive modeling in political contexts.

2 Data

2.1 Overview

For this analysis, we employed the R programming language (R Core Team 2023) to examine polling data on public sentiment before the election. Our dataset, sourced from FiveThirtyEight (FiveThirtyEight 2024), offers a detailed snapshot of evolving public opinion. We explored key influences on percentage support, including the timing of the polls, pollster characteristics, and geographic differences.

Several `R` packages were instrumental in performing data manipulation, modeling, and visualization. `Tidyverse` was the backbone for organizing and analyzing the data efficiently, allowing seamless integration of multiple tasks (Wickham et al. 2019). `Here` streamlined file path handling, ensuring smooth data access across systems (Müller 2020). We relied on `Janitor` for robust data cleaning, helping identify and fix potential inconsistencies (Firke 2023), and `Lubridate` facilitated the manipulation of time-related variables (Grolemund and Wickham 2011). Additionally, `Arrow` provided fast and memory-efficient access to large datasets, a critical factor for managing extensive polling data (Richardson et al. 2024). The structure of the codebase and workflow adhered to the best practices outlined in Alexander (2023).

2.2 Measurement

In the data section, we elaborate on the process of transforming raw polling data into a structured dataset for our analysis. The dataset originates from real-world polling events where respondents across the United States share their levels of support for Donald Trump. Various polling organizations collect these responses, each employing distinct methodologies that include sampling techniques, question phrasing, and the timing of polls. Such methodological differences can introduce significant variability in the data, which is critical to our investigation of public opinion trends.

Once the polls are completed, the results are aggregated into a comprehensive dataset, such as the one provided by FiveThirtyEight (FiveThirtyEight 2024). This dataset includes essential variables like the poll's end date (indicating when data collection concluded), the identity of the pollster (the organization conducting the poll), the state in which the poll was conducted, and a poll score that assesses the quality and reliability of each poll based on its methodology. These variables are crucial for understanding the reported percentage support for Trump, as they capture the design and context of each poll, allowing us to evaluate how these factors might influence public sentiment.

The entries in the dataset provide a quantified measurement of public opinion at specific points in time and space, enabling us to analyze patterns and trends in Trump's support. This transformation from raw polling data to structured entries facilitates the exploration of relationships between the predictors—such as state, poll score, and pollster—and the outcome of interest (Trump's percentage support). By grounding our analysis in data derived from actual polling processes, we ensure a reliable foundation for examining the various factors that shape public opinion, ultimately contributing to a more nuanced understanding of the electoral landscape in the lead-up to the 2024 U.S. election.

2.3 Outcome variable

2.3.1 The percentage support for Trump in the poll

The percentage support for Trump in polls represents the portion of respondents who favor Donald Trump in a given survey. This measure is crucial in gauging his popularity, especially in the context of upcoming elections. Expressed as a percentage, it ranges from 0 to 100, with higher values indicating greater support. The percentage of support can vary across different polls depending on factors such as sample size, region, and methodology. A strong polling result with higher percentages suggests robust backing, while lower percentages indicate weaker favorability among the surveyed population.

Figure 1 shows Trump's percentage support across polls, mostly clustered between 40% and 50%. There's a peak around 45%, with a slight right skew indicating a few polls showing higher support, up to 55%. Very few polls show support below 30% or above 50%, suggesting moderate and stable backing with minimal outliers. This distribution reflects consistent mid-range support for Trump across the sampled polls.

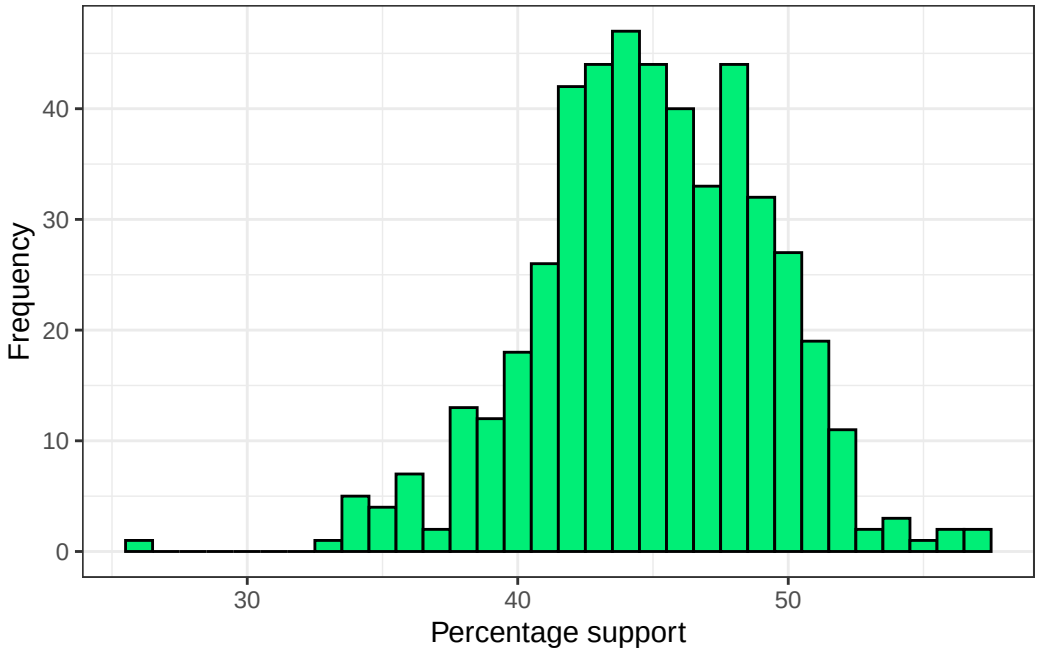


Figure 1: Distribution of percentage support for Trump

2.4 Predictor variables

2.4.1 End date

The end date marks the final day of data collection for a poll, signaling the close of the survey period. It provides critical context by showing when the poll reflects public opinion, as sentiments can evolve rapidly due to factors like news cycles, political events, or campaign strategies. The earliest end date is 2022-11-22 and the latest end date is 2024-10-16.

2.4.2 State

The state variable indicates where the poll was conducted, either in a specific U.S. state (e.g., Arizona or California) or across the nation ("National"). State polls focus on regional voter preferences, while national polls provide an overall view of public sentiment across the country. This distinction helps in analyzing both local and nationwide trends.

Figure 2 shows the number of polls conducted, with national polls leading by a wide margin, surpassing 250. Wisconsin, Pennsylvania, and Arizona have the highest number of state-specific polls, each below 60. Other key states like Michigan, Georgia, and Texas follow, while states like Missouri and Colorado have the fewest. The focus on national and battleground states highlights their electoral importance.

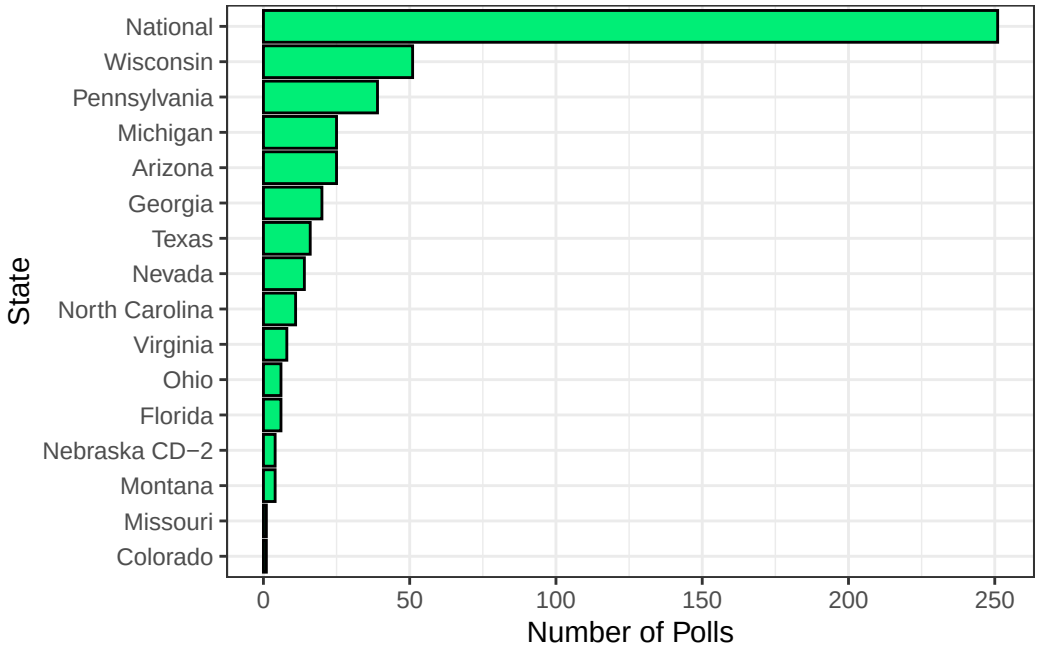


Figure 2: Number of polls by state

2.4.3 Poll score

The poll score is a rating that measures the reliability and quality of a poll based on various criteria, including the pollster's history, the clarity of their methodology, and the sample's representativeness. Higher scores indicate greater credibility, suggesting the poll follows rigorous methods and provides an accurate reflection of public opinion. In contrast, lower scores may reveal flaws such as bias, insufficient sample sizes, or poor transparency, making the results less dependable for drawing conclusions.

Figure 3 displays the distribution of poll scores, with two distinct peaks. Most polls cluster around scores of -1.5 and -1.1, indicating that the majority of polls have similar ratings. There is a small number of polls with scores around -1.3, suggesting less frequent variation in the middle of the range.

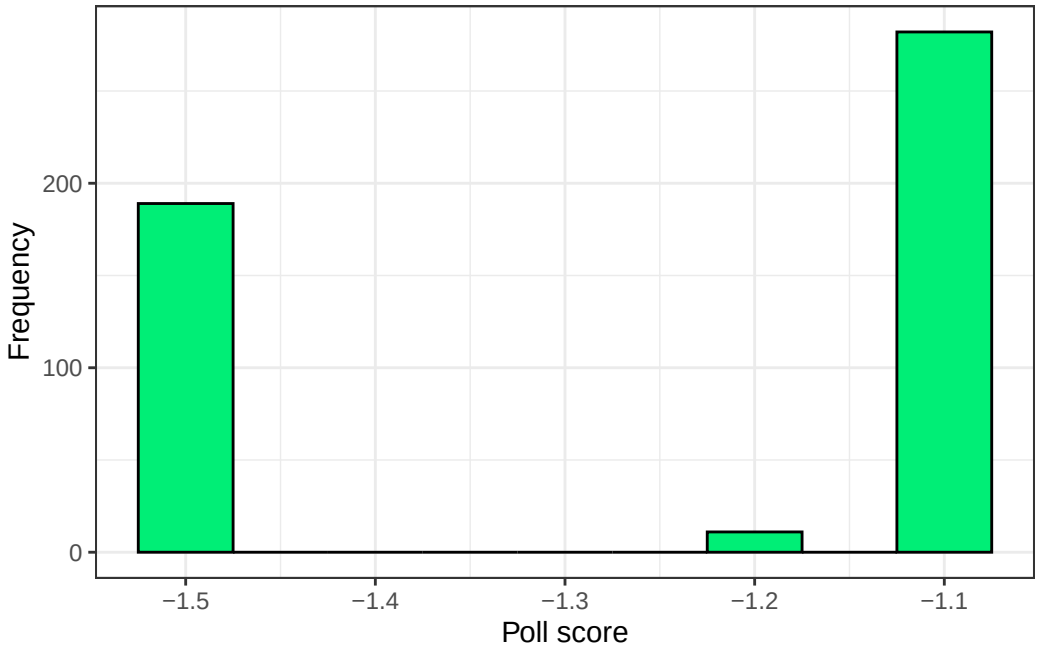


Figure 3: Distribution of poll scores

2.4.4 Pollster

The pollster are known as the organization responsible for conducting the poll, such as Emerson, YouGov, or Quinnipiac. These organizations collect data to gauge public opinion on various topics, including political preferences. Each pollster employs its own methodologies, sampling, and geographic coverage, which can lead to variations in results and influence the overall reliability of the poll.

Figure 4 illustrates the number of polls conducted by various polling organizations. YouGov leads with over 200 polls, followed closely by Siena/NYT, which has just under 200. Other pollsters, such as Marquette Law School, have significantly fewer polls, and the remaining organizations conducted only a handful of surveys.

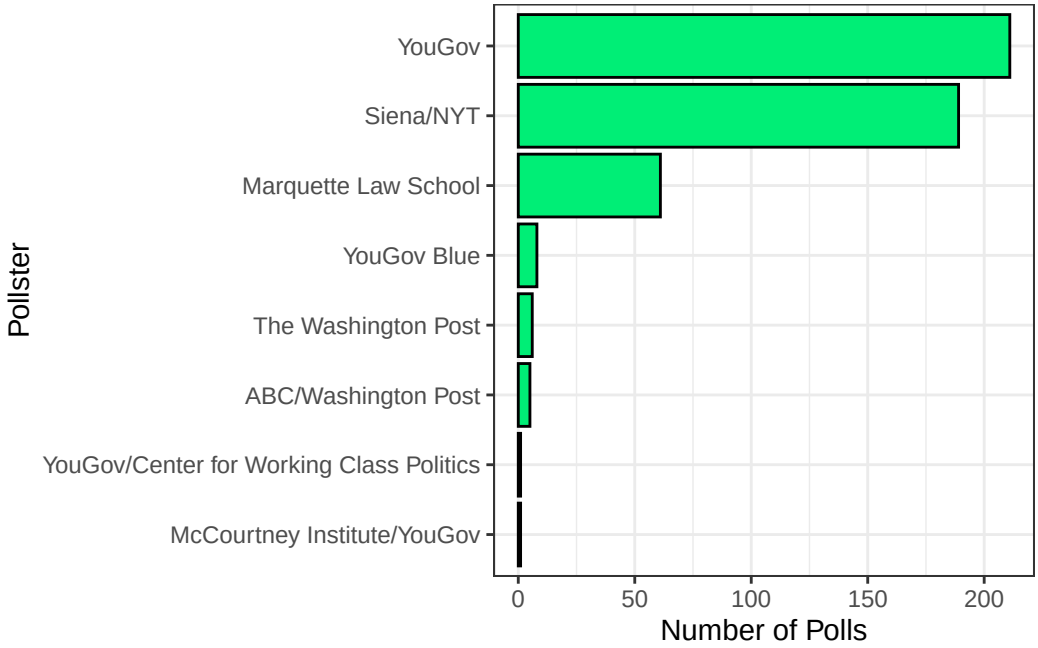


Figure 4: Distribution of pollster

2.5 Variable associations

Figure 5 displays the distribution of percentage support for various pollsters, with the poll score on the vertical axis. Each box represents the interquartile range of the percentage support, while the whiskers extend to the minimum and maximum values, with outliers indicated by dots. Pollsters like YouGov and Siena/NYT show more variability in support, while Marquette Law School and ABC/Washington Post have tighter ranges. Some pollsters, such as YouGov Blue, have more consistent results with smaller interquartile ranges, indicating more stable percentage support.

Figure 6 displays the distribution of percentage support across different states. Colorado, Virginia, and Nebraska CD-2 show the highest median support, while states like Missouri and Montana have the lowest median support levels. The variability in the percentage support is noticeable, with several states, such as Michigan and Wisconsin, displaying wider interquartile ranges, indicating greater variation in support. Outliers in states like Arizona and National suggest some polls reported either higher or lower support compared to the majority.

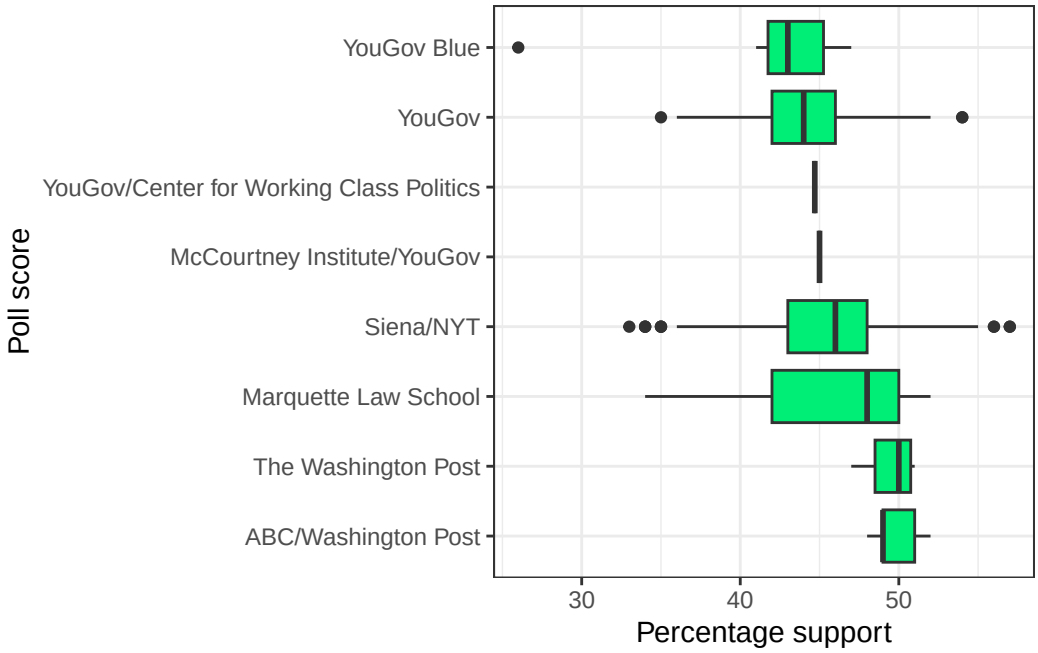


Figure 5: Association between percentage support and pollster

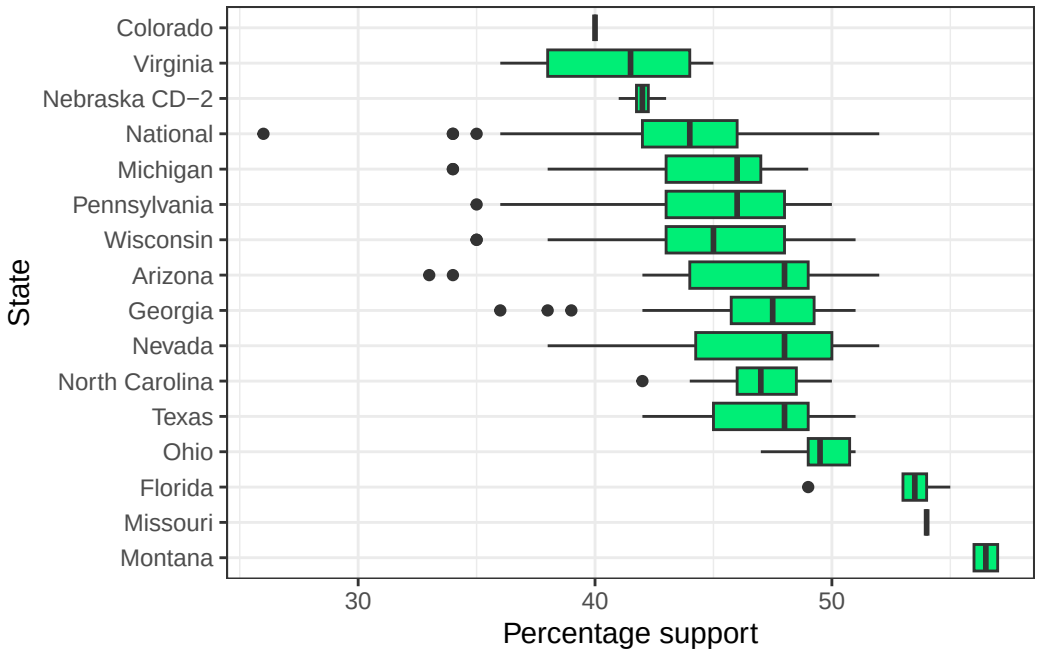


Figure 6: Percentage support for Trump by state

3 Model

Our aim is to quantify the relationship between key variables, such as the end date of the poll, the state, the poll score, and the pollster, and the percentage support for Trump. We employ a linear regression model to investigate how each of these factors influences support levels. Our linear regression model includes the end date, pollster, state, and poll score as predictors, allowing us to quantify their individual impact. By estimating the coefficients, we can assess the direction and strength of their effects, providing insights into which variables play the most significant role in shaping public opinion. More comprehensive details on the model's specification, underlying assumptions, and diagnostic checks are provided in Appendix .1 and Appendix .2. For a full description of the validation procedures used, please refer to Appendix .3.

The modeling choices align with the structure of the data. We treat the end date as a continuous variable to capture any linear trends over time. Pollster and state are treated as categorical variables to reflect inherent group differences without assuming any specific order. Poll score remains continuous to preserve its detail and capture the nuanced impact of poll quality on support levels. These decisions ensure that essential characteristics of the data are preserved, allowing for a more accurate and informative analysis.

3.1 Model set-up

3.1.1 Model 1: Percentage support as a function of end date

The first model investigates how the end date of a poll impacts the percentage of support for Trump. Mathematically, it is represented as:

$$y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i$$
$$\epsilon_i \sim \text{Normal}\left(0, \sigma^2\right)$$

Where:

- y_i denotes the percentage support for Trump in poll i
- x_{1i} is the end date of poll i
- β_0 represents the intercept, which is the baseline level of support
- β_1 captures the effect of the end date on percentage support
- ϵ_i is the error term, assumed to follow a normal distribution with a mean of 0 and variance σ^2

3.1.2 Model 2: Percentage support as a function of end date, state, poll score, and pollster

In the second model, we extend the analysis by including additional variables: the state in which the poll was conducted, the poll score, and the pollster. The equation is as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \epsilon_i$$
$$\epsilon_i \sim \text{Normal}\left(0, \sigma^2\right)$$

Where:

- y_i represents the percentage support for Trump in poll i,
- x_{1i} is the end date of the poll
- x_{2i} corresponds to the state where the poll was conducted
- x_{3i} is the poll score, indicating the quality of the poll
- x_{4i} represents the pollster who conducted the poll
- β_0 is the intercept
- $\beta_1, \beta_2, \beta_3, \beta_4$ are the coefficients corresponding to each predictor variable, measuring their individual effects
- ϵ_i is the error term, which follows a norma [1] - istribution with mean 0 and variance σ^2

We run the models in R (R Core Team 2023).

3.1.3 Modeling justification

The first model examines how the end date of a poll influences Trump's percentage support, based on the idea that public opinion shifts over time due to external events or campaign developments. By treating the end date as a continuous variable, the model captures trends in support, where polls conducted closer to the election are expected to reflect more accurate voter preferences. The linear regression approach provides a simple way to quantify this relationship.

The second model expands by incorporating state, poll score, and pollster as additional predictors. These variables are key factors influencing polling outcomes, as state-level differences, poll quality, and methodologies of various pollsters can all affect the results. Including these predictors allows the model to control for geographic and methodological variations, offering a clearer picture of the factors driving Trump's support. The use of linear regression ensures that the model remains interpretable while accounting for these important influences.

Table 1: Model results

	Model 1	Model 2
(Intercept)	−131.653	−223.206
	(20.651)	(52.682)
end_date	0.009	0.008
	(0.001)	(0.001)
stateColorado		−4.920
		(3.613)
stateFlorida		5.903
		(1.604)
stateGeorgia		0.316
		(1.055)
stateMichigan		−1.836
		(0.997)
stateMissouri		7.056
		(3.612)
stateMontana		8.925
		(1.899)
stateNational		−1.803
		(0.816)
stateNebraska CD-2		−5.475
		(1.898)
stateNevada		0.779
		(1.178)
stateNorth Carolina		−0.732
		(1.314)
stateOhio		1.194
		(1.688)
statePennsylvania		−1.623
		(0.910)
stateTexas		0.732
		(1.165)
stateVirginia		−5.839
		(1.430)
stateWisconsin		−2.280
		(0.918)
pollscore		−92.960
		(39.349)
pollsterMarquette Law School		3.173
		(3.616)
pollsterMcCourtney Institute/YouGov		0.521

4 Results

Our results are summarized in Table 1.

In Model 1, the intercept is estimated at -131.653 , meaning that when the end date is at zero (far back in time), the predicted percentage support for Trump would be negative, though this has no real-world interpretation due to the range of actual dates. The coefficient for end date is positive (0.009), indicating that as time progresses (later end dates), Trump's percentage support slightly increases by 0.009 percentage points per day. This is a statistically significant result given the low standard error (0.001). The model explains 13.2% of the variation in support ($R^2 = 0.132$), which is relatively modest, and has a root mean square error (RMSE) of 3.96, meaning there is some error in the predictions.

Model 2 expands on the predictors, including state, poll score, and pollster. The intercept decreases to -223.206, with the coefficient for end date still positive but slightly smaller (0.008), showing a similar upward trend in support over time. Many state-level coefficients show how different states' support for Trump compares to the baseline category (Colorado). For instance, Florida has a positive and statistically significant coefficient (5.903), suggesting higher support for Trump compared to Colorado, while Nebraska CD-2 has a negative coefficient (-5.475), indicating lower support.

The poll score has a negative coefficient (-92.960), meaning higher poll quality (or score) tends to show lower support for Trump. Among pollsters, Siena/NYT has a notably negative coefficient (-35.982), indicating that this pollster tends to report significantly lower Trump support compared to the baseline.

Model 2 improves overall fit, explaining 35% of the variation in support ($R^2 = 0.350$) and reducing the RMSE to 3.43, indicating better predictive accuracy. The AIC and BIC values also decrease (AIC = 2605.4, BIC = 2709.8), suggesting that Model 2 is a better fit compared to Model 1.

Overall, the expanded Model 2 captures more variation in support, with state, poll score, and pollster contributing significantly to the predictive power.

5 Discussion

5.1 Influence of State-Level Variability on Support

Model 2 reveals significant variability in Trump's percentage support across different states. For instance, states like Florida and Missouri show a positive and statistically significant increase in support compared to the baseline (Colorado), while Nebraska CD-2 and Virginia exhibit lower support. This demonstrates the importance of geographic factors in understanding voter preferences, as state-specific dynamics heavily influence the overall level of support.

Campaign strategies might need to be tailored for individual states, especially swing states or regions showing significant divergence in support.

5.2 Impact of Poll Score on Reported Support

The poll score coefficient in Model 2 is negative, suggesting that polls with higher methodological rigor or better quality report lower support for Trump. This is an important finding, as it indicates that lower-quality polls might overstate Trump's support, potentially leading to biased or skewed interpretations of his popularity. This result emphasizes the need to critically evaluate poll quality when analyzing and aggregating polling data to avoid misleading conclusions.

5.3 Improvement in Model Performance with Additional Predictors

The improvement in model fit (R^2 increasing from 0.132 in Model 1 to 0.350 in Model 2) and the reduction in RMSE (from 3.96 to 3.43) highlights the importance of including multiple factors in the analysis. Adding state, poll score, and pollster as predictors significantly enhances the model's ability to explain the variability in Trump's percentage support. This suggests that relying solely on temporal trends (as in Model 1) may miss key contextual factors, such as the influence of geographic and methodological differences, in explaining public opinion.

5.4 Overall Influence of Factors on Percentage Support

The inclusion of multiple factors in Model 2—such as state, poll score, and pollster—demonstrates that Trump's percentage support is influenced by a combination of geographic, methodological, and temporal factors. The model highlights that no single variable fully explains the variation in support, but rather, a combination of these factors work together to shape public opinion. The coefficients for state-level variables suggest that regional differences play a significant role, while the poll score and pollster coefficients indicate that the methodology and quality of the polls themselves also have a measurable impact. This underscores the importance of considering multiple dimensions when analyzing polling data, as relying solely on one factor (such as time) may overlook key influences such as regional trends or poll quality.

5.5 Weaknesses and next steps

Model 2, while an improvement over Model 1, may still suffer from omitted variable bias as key demographic factors like age, education, and economic conditions are not included. These variables likely influence Trump's support across different states and their absence could limit

the model's ability to fully explain variation. Additionally, the residual plot and the Normal Q-Q plot suggest potential heteroscedasticity and deviations from normality (Figure 7a; Figure 7b), indicating the model struggles with capturing non-linear dynamics or outliers, particularly during significant shifts in public opinion or extreme polling outcomes. The limited pollster-specific effects also fail to account for interactions between pollsters and factors like state or poll quality, which could lead to oversimplified interpretations of polling data.

To address these issues, future models should consider including demographic and economic variables to improve accuracy and capture a more comprehensive picture of Trump's support. Additionally, exploring non-linear relationships using polynomial terms or splines could help better capture the complexity of public opinion changes over time. It would also be valuable to investigate interaction effects between pollster, state, and poll score to understand how these factors jointly influence results. Finally, addressing the heteroscedasticity and potential non-normality identified in the residual plots, through transformations or the use of robust standard errors, will improve the model's reliability and interpretation.

Appendix

.1 Model specification

The model specification assumes a linear relationship between predictors (such as end date, state, poll score, and pollster) and Trump's percentage support. It also assumes homoscedasticity, meaning the variance of errors is constant across all fitted values, and that errors are independently and normally distributed. The model also treats pollster and state as categorical variables, assuming no interaction effects between them and other factors. These assumptions provide a straightforward, interpretable model.

.2 Model diagnostics

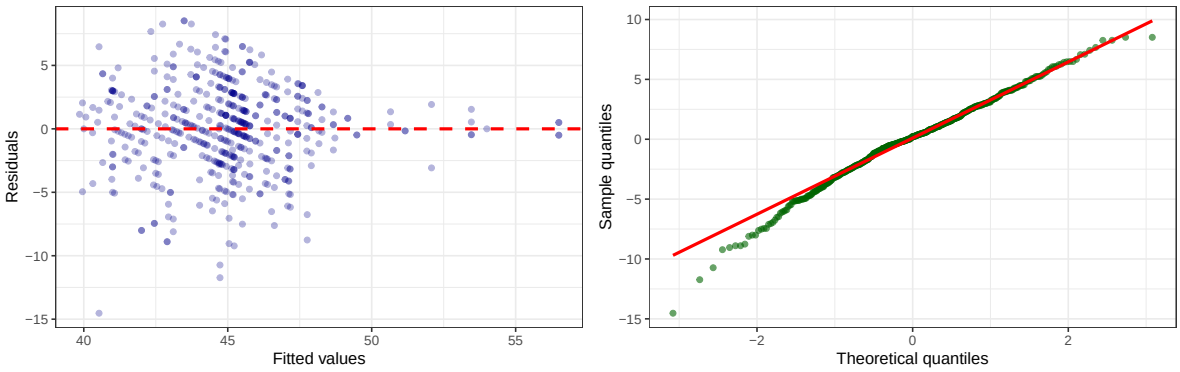
Figure 7a shows the relationship between the fitted values (predicted Trump support) and the residuals (errors). Ideally, the residuals should be randomly distributed around zero without any clear pattern or clusters, which would indicate a linear good model fit and that the independent errors assumption is satisfied. However, this plot shows some visible structure and a potential funnel shape, suggesting possible heteroscedasticity, meaning that the variance of the residuals changes with the fitted values. This may indicate that the model's assumptions of constant variance are violated, warranting further investigation into model improvements or the need for transformation of variables.

Figure 7b compares the residuals to a theoretical normal distribution. If the residuals are normally distributed, the points should closely follow the red line. In this case, most of the points lie along the line, suggesting that the residuals are approximately normal. However, there is some deviation at the tails, particularly at the lower end, where the residuals fall below the line. This indicates potential non-normality in the distribution of residuals, particularly in the extremes, which may suggest that the model struggles to fit extreme values as accurately.

.3 Model validation

Model validation via Root Mean Square Error (RMSE) is an important aspect of assessing the accuracy and predictive power of regression models. RMSE provides a measure of how well the model's predicted values align with the actual observed values, specifically focusing on the magnitude of prediction errors.

The RMSE tells us how far off the model's predictions are from the actual values, in the same units as the dependent variable (percentage support). In this context, an RMSE of 3.43 means that, on average, Model 2's predictions deviate from the true percentage support by approximately 3.43 percentage points. This is an improvement over Model 1, where the average deviation is about 3.96 percentage points. While both RMSE values are relatively low, the reduction in RMSE from Model 1 to Model 2 suggests that including additional predictors



(a) Residuals vs. fitted values

(b) Normal Q-Q plot of residuals

Figure 7: Checking assumptions of the linear regression model (model 2)

(such as state and poll score) meaningfully enhances the model's precision. This improvement is particularly important for making more informed predictions, as Model 2 captures more variability in the data.

Appendix a.

.1 Deep Dive into YouGov's Methodology

YouGov has established itself as a leading polling organization, renowned for its systematic and transparent approach to gathering public opinion data. The methodology employed by YouGov transforms individual opinions into aggregate results, reflecting broader societal sentiments. Central to this process is the organization's extensive online panel, which aims to represent the U.S. adult population. This panel is formed from a diverse recruitment strategy, including digital advertising and partnerships with various platforms, ensuring a wide-ranging demographic representation that encompasses varying age groups, races, and political affiliations. The primary population of interest includes all adults aged 18 and older residing in the United States, with the sampling frame limited to those who have opted into YouGov's panel.

One significant aspect of YouGov's methodology is its nonprobability sampling approach. While this method allows for rapid and cost-effective data collection, it raises concerns about representativeness. Nonprobability sampling can lead to self-selection bias, where individuals who choose to join the panel may inherently differ from the general population. To mitigate this issue, YouGov employs rigorous demographic matching processes. Upon joining the panel, participants provide detailed demographic information that helps ensure future survey samples align closely with national demographic benchmarks. The organization continually monitors the characteristics of its panelists to maintain a balanced representation, drawing from sources such as the U.S. Census and other reputable datasets.

When conducting surveys, YouGov prioritizes quality by employing a well-structured questionnaire design. Surveys typically include a variety of question types, such as multiple-choice, Likert scales, and open-ended formats. This diversity allows for capturing nuanced opinions, which can lead to richer data analysis. However, the effectiveness of the questionnaire is highly dependent on clarity and neutrality; poorly phrased or leading questions can skew results. YouGov addresses this concern by emphasizing neutrality in its wording and frequently randomizing the order of questions to reduce response bias. Furthermore, the organization utilizes techniques such as pre-tests and focus groups to refine questions before full deployment.

To handle non-response, YouGov implements strategic management of its panelists. The organization tracks individual participation history, inviting respondents based on their previous engagement levels and survey categories. This approach helps maintain a high response rate while ensuring that the final sample remains representative of the target population. Nonetheless, the reliance on an opt-in panel raises the question of whether certain demographic groups may be underrepresented, particularly those less likely to engage with online surveys.

Once data collection is complete, YouGov employs a thorough weighting process to adjust the results so that they accurately reflect the broader population. Weighting involves assigning

different levels of importance to respondents based on demographic characteristics, such as age, gender, race, and political affiliation. By comparing the sample's demographics to established benchmarks, YouGov can adjust the influence of individual responses to align more closely with the actual population distribution. This method is critical in ensuring that the final reported results provide an accurate portrayal of public opinion.

In terms of ensuring data quality, YouGov employs a multifaceted strategy involving monitoring, testing, and refinement. The organization conducts rigorous checks on panelists' identities, verifying email addresses and running IP address checks. Furthermore, they utilize response quality assessments to evaluate the reliability of participant responses, which helps identify those providing unreliable or fraudulent answers. Respondents who consistently fail quality checks may be removed from the panel, thus preserving the integrity of the data collected.

YouGov is also committed to protecting the privacy of its panelists. Participants retain control over their personal data, with options to opt out of data sharing and to request modifications or deletions of their information. Such measures help foster trust among panelists, encouraging participation while safeguarding their rights.

Finally, YouGov employs multilevel regression with post-stratification (MRP) as a key method for estimating voting behavior, particularly in major elections. This model leverages data collected from their panelists to predict voting patterns across different demographics, allowing for sophisticated analysis of potential election outcomes. By applying this method to the national voter file, YouGov can project support levels for candidates with a higher degree of accuracy.

In summary, YouGov's methodology combines innovative sampling techniques, rigorous data quality measures, and a strong commitment to transparency and participant privacy. While the organization's approach offers significant advantages, including speed and adaptability, it also faces inherent challenges associated with nonprobability sampling and self-selection bias. Understanding these strengths and limitations is essential for interpreting YouGov's polling results accurately and making informed conclusions about public opinion in the lead-up to elections.

Appendix b.

.1 Idealized Methodology for Forecasting the U.S. Presidential Election

Budget Overview

With a budget of $100,000, this project will focus on a comprehensive methodology designed to gather, analyze, and interpret public opinion data regarding the U.S. presidential election. The aim is to construct a robust survey framework that allows for precise forecasting of voter behavior and sentiment. The investment will primarily be allocated toward panel recruitment, survey development, data collection, and statistical modeling. This financial backing will ensure high-quality data acquisition and analysis, contributing significantly to the accuracy of the forecasts.

Sampling Approach

Target Population:

The survey will target all eligible voters in the United States, specifically adults aged 18 and older. This broad demographic will enable a comprehensive understanding of voter preferences and sentiments across different segments of the population. Within this overarching target, the sampling will be stratified to ensure representation from key demographics, including age, gender, race, income level, education, and geographic location. Stratified random sampling is crucial, as it helps to capture the diversity of the U.S. electorate, allowing for a more nuanced analysis of how different factors influence support for presidential candidates.

Sampling Technique:

A probability sampling method will be employed, focusing on stratified random sampling. This approach allows for a representative selection of respondents based on predetermined demographic criteria. By ensuring that various segments of the population are adequately represented, we can minimize selection bias and enhance the reliability of the results (Groves et al., 2009). The sample size will aim for approximately 2,000 respondents, which strikes a balance between statistical reliability and resource efficiency, targeting a margin of error of $\pm 2\%$ at a 95% confidence level. This sample size is sufficient to capture variations in voter support across different demographic and regional groups.

Respondent Recruitment

Panel Development:

To assemble the sample, respondents will be recruited through a multi-channel strategy. This includes targeted online advertising, partnerships with civic organizations, and outreach via social media platforms. The objective is to attract a diverse array of participants, encompassing various socioeconomic and demographic backgrounds. This method not only increases the pool of potential respondents but also enhances the inclusivity of the sample, allowing for a more accurate reflection of the electorate.

Incentives for Participation:

Respondents will receive monetary incentives or gift cards for completing the survey. These incentives are crucial for encouraging participation, especially among demographics that might be less inclined to engage in online surveys. Additionally, the recruitment process will emphasize the importance of each individual's contribution to understanding public sentiment, which can motivate participation based on a desire to influence research outcomes.

Online Survey Platform:

Surveys will be conducted using Google Forms, allowing respondents to complete them on their smartphones, tablets, or computers at their convenience. The design will incorporate a mix of question types, including multiple-choice, Likert scales, and open-ended questions. This variety will facilitate a comprehensive understanding of voter attitudes and preferences, making it easier to analyze the data effectively.

Data Validation

Quality Control Measures:

Ensuring data integrity will be paramount. YouGov will implement a robust quality control system that includes various verification techniques. Each respondent's identity will be confirmed through email verification and IP address checks to prevent duplicate or fraudulent responses. Additionally, response time will be monitored to ensure participants are thoughtfully engaging with the survey rather than rushing through it.

Demographic Weighting:

After data collection, statistical weighting will be applied to adjust for any demographic imbalances in the sample compared to the national population, using U.S. Census data as a benchmark. This weighting process is critical for enhancing the accuracy of the findings, ensuring that the results reflect the broader electorate's opinions.

Poll Aggregation and Modeling

Aggregation Techniques:

The survey results will be aggregated with existing polling data from reputable sources, such as FiveThirtyEight and Gallup. This will involve a weighted polling model that combines results from various polls to produce a more reliable overall estimate. By pooling data from multiple sources, we can mitigate the impact of any single poll's biases, resulting in a more robust analysis of voter sentiment.

Statistical Modeling:

A multilevel regression with post-stratification (MRP) approach will be used to estimate candidate support across different demographics and geographic segments. This technique has been effective in previous elections, allowing us to account for complex interactions between

variables (Gelman & Hill, 2007). By training the model with data from our survey and integrating it with aggregated polling data, we can refine our forecasts and provide a more nuanced view of voter behavior leading up to the election.

Reporting and Transparency

Results Dissemination:

The findings from this research will be compiled into a comprehensive report detailing the methodology, demographic breakdown, key insights, and margin of error. This report will be made publicly available to ensure transparency and build trust in the results. Engaging with civic organizations and stakeholders will be essential to communicate findings and foster discussions about voter engagement and turnout strategies.

In summary, this idealized methodology for forecasting the U.S. presidential election combines rigorous sampling techniques, effective respondent recruitment, and thorough data validation. By utilizing a well-structured online survey platform and employing advanced statistical modeling, the project aims to deliver accurate insights into voter sentiment, ultimately contributing to a better understanding of the electoral landscape in the lead-up to the election. This comprehensive approach ensures that the data collected will provide valuable information for both political analysts and decision-makers.

Survey Implementation

The survey has been developed using Google Forms, allowing for easy access and completion across various devices. This platform facilitates user-friendly interaction, enabling respondents to provide their insights conveniently. The survey can be accessed through the following link: https://forms.gle/4VjHGLDEbmwiy9ex6

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

FiveThirtyEight. 2024. "Dataset: US Presidential General Election Polls." https://projects.fivethirtyeight.com/polls/data/president_polls.csv.

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/.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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://CRAN.R-project.org/package=arrow.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.