Election Predictions:A Data-Driven Approach to US Election Forecasting*

Ziheng Wang Manjun Zhu Dong Jun Yoon

November 4, 2024

This paper analyzes the distribution of voter preferences across demographics and regions in the United States using both simulated and survey data. Key findings reveal significant variations in political leanings based on race, gender, education level, and geographic location (urban vs. rural), with distinct approval patterns for candidates like Joe Biden and Kamala Harris. Additionally, comparisons between states highlight differences in safety and economic concerns among voters, which correlate with their past voting behavior. This study underscores the importance of post-stratification adjustments in survey data to achieve a more accurate representation of public opinion, shedding light on demographic-driven electoral dynamics that impact the political landscape.

1 Introduction

Overview Presidential elections are among the most significant events in American politics, shaping the political and economic landscape for the next four years. The outcome of the 2024 U.S. election will influence not only the American economy but also the global economy. In this context, accurately forecasting electoral outcomes is crucial for anticipating shifts in political power and predicting economic policies.

Polling is a commonly used tool to understand public opinion. However, it may lack accuracy due to various biases, such as social desirability bias, non-response bias, and sampling error. In this paper, we attempt to construct a model measures influencial factors to pct.

Estimand In this paper the estimand is the percentage of support of each candidate, Donald Trump and Hamala Harris. We construct a linear model to estimate various predictors, which are transparency score, poll score, numeric grade, state, sample size, number of days. The

*Code and data are available at: https://github.com/zero00499/2024_US_election_forecasting.git.

model is used to identify the most influential factor for preferences of pollsters and analyze relationships between variables.

Why it matter Our findings are helpful to political scholars who work on current political dynamics, economic policy analysts and sociologists, as we provide in-depth understanding of influential factors of pollster's preference and predict the winner of the election. Different candidates have different economic plans due to their different opinions on the economy.

2 Data

2.1 Overview

We obtained the "Presidential General Election Polls" dataset from FiveThirtyEight (**FiveThirtyEight2024?**) and conducted an in-depth analysis using the R programming language (R Core Team 2023). This analysis utilizes a dataset containing polling data for the 2024 U.S. Presidential Election. It includes key details such as the quality of each poll, sample size, geographic scope, and the timing of data collection. The dataset is designed to track voter support trends for the candidates Kamala Harris and Donald Trump. The data were obtained from FiveThirtyEight's national polling dataset for the 2024 U.S. presidential election. This dataset, published on FiveThirtyEight's platform, contains comprehensive polling data collected by various polling organizations. It includes information such as polling dates, candidate preferences, sample sizes, and margins of error for general election polling conducted across the United States. The data is collected from January 2024 to the present, capturing public opinion trends leading up to the 2024 election and allowing for analysis of voter sentiment over time and across different demographics and regions.

2.2 Measurement

For this analysis, the dataset comprises polling information gathered from multiple sources, each employing distinct methodologies and inherent biases. Although adjustments and weighting are applied to address these differences, limitations may persist, particularly regarding sample representation and measurement error. Polls with lower quality scores or limited transparency may contribute additional variability to the analysis.

2.3 Data Cleaning

To ensure high-quality and reliable insights from the polling data, the dataset was filtered based on the 'partisan' and 'pollscore'. Specifically, we retained only records where 'partisan' is missing (indicating no background support) and 'pollscore' is greater than 0, as this reflects

polls with better accuracy and reliability. This step refines the dataset to focus on more trustworthy polling data.

2.4 Outcome variables

The primary outcome variable is the percentage of support for each candidate, reflecting the proportion of respondents who express support for either Kamala Harris or Donald Trump at a specific point in time.

```
Rows: 2963 Columns: 11
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (4): pollster, state, party, candidate_name
dbl  (5): duration, numeric_grade, pollscore, transparency_score, pct
lgl  (1): hypothetical
date (1): end_date

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# A tibble: 10 x 11
   pollster state party hypothetical end_date   duration numeric_grade pollscore
   <chr>    <chr> <chr> <lgl>        <date>        <dbl>         <dbl>     <dbl>
 1 Patriot~ Nati~ DEM   FALSE        2024-09-03        2           1.1       0.6
 2 PRRI     Nati~ REP   TRUE         2024-09-04       19           1.8       0.4
 3 Big Vil~ Nati~ IND   TRUE         2023-11-05        6           1.2       1.7
 4 Big Dat~ Mich~ DEM   TRUE         2023-11-19        3           1.1       0.2
 5 Redfiel~ Nati~ REP   TRUE         2023-05-31        0           1.8       0.4
 6 Redfiel~ Nort~ LIB   TRUE         2024-07-10        2           1.8       0.4
 7 J.L. Pa~ Nati~ DEM   TRUE         2024-09-13        4           1.6       0.2
 8 Redfiel~ Mich~ DEM   TRUE         2023-11-29        2           1.8       0.4
 9 Targoz ~ Flor~ REP   TRUE         2024-07-24        5           1.8       0.1
10 Big Vil~ Nati~ REP   TRUE         2024-07-14        2           1.2       1.7
# i 3 more variables: transparency_score <dbl>, candidate_name <chr>, pct <dbl>
```

2.5 Predictor variables

2.5.1 Numeric Grade

"Numeric_grade" is a numerical rating assigned to each pollster, reflecting the reliability and overall quality of the pollster's data collection. A higher grade denotes stronger pollster cred-

ibility and consistent performance.

2.5.2 Transparency Score

"Transparency_score" measures the level of openness a pollster has in disclosing their methodology, with scores up to a maximum of 10. Higher transparency scores are associated with greater reliability, as they indicate detailed data-sharing practices. This variable is used to examine how transparency influences polling outcomes.

2.5.3 State

The "state" variable represents the U.S. state where the poll is conducted or focused, capturing regional variations in voter support. Including state-level data allows the model to account for unique local trends that might affect voting behavior across different regions.

2.5.4 Duration

"Duration" represents the number of days from the start date of the polling period to the election date(Nov 4th). This predictor is intended to capture the temporal dynamics of voter sentiment and polling trends as the election approaches. By quantifying the time elapsed, duration allows for an analysis of how public opinion may shift over time, providing insights into the effectiveness of campaign strategies and the impact of external events on voter behavior. A shorter duration may indicate a more immediate influence of recent events on polling data, while a longer duration may reflect more stable trends in voter preferences.

2.5.5 Party

The "party" variable indicates the political affiliation of the candidate within each poll, such as "DEM" for Democrats or "REP" for Republicans. This variable helps in distinguishing the voting trends and support levels between the parties represented in the analysis.

2.5.6 Hypothetical

This variable indicates whether the poll reflects a real or hypothetical match-up scenario. Polls marked as $FALSE$ represent actual, live election match-ups. Including this variable in our linear regression model allows us to differentiate the reliability and predictive power of real versus hypothetical scenarios.

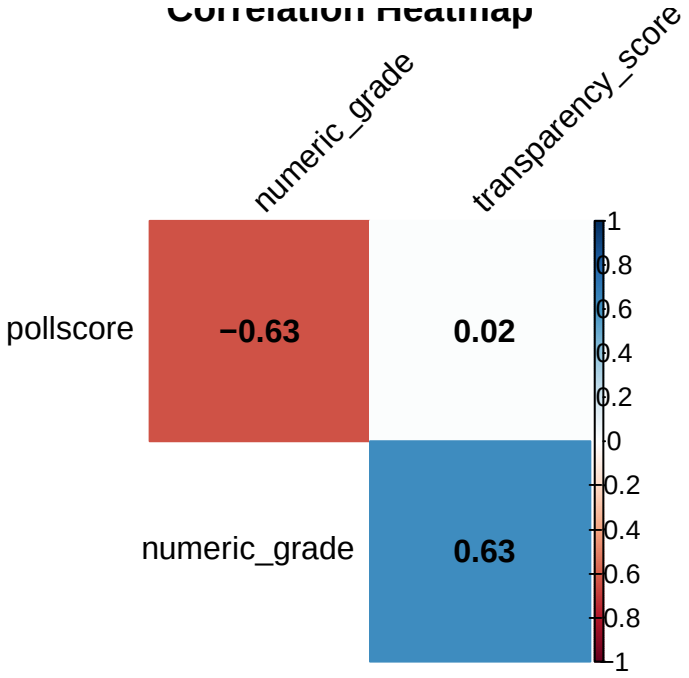

Figure 1: Correlation Matrix of Numeric Grade, Transparency, and Pollscore

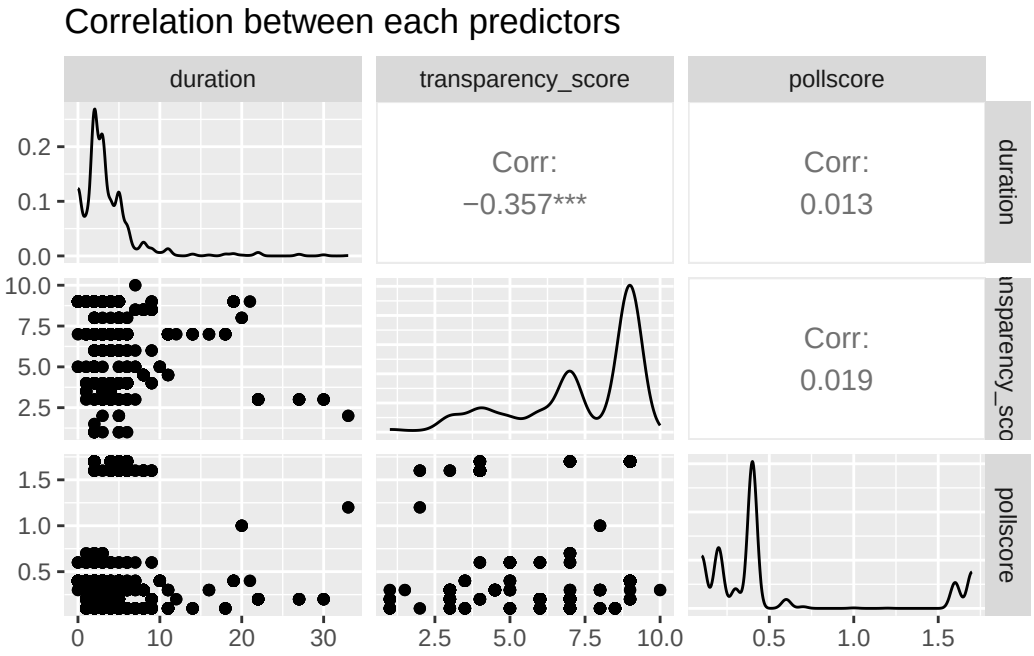

Figure 2: Correlation Plot

When selecting predictors for the model, addressing the assumption of uncorrelated errors was critical. Uncorrelated errors is violated when two or more predictor variables are highly correlated, which can inflate the variance of coefficient estimates and diminish model reliability. If this assumption is violated, it could lead to inefficient coefficient estimates and incorrect inferences about the model's predictors. The Correlation Matrix(Figure 1) indicates a significant correlation between *numericgrade* and *pollscore*, and *numericgrade* and *transparencyscore*, suggesting that these variables assess similar aspects of polling quality. To avoid the violation of uncorrelated errors, we decided to exclude *numericgrade* from our model, which also avoids redundancy and enhances the stability of coefficient estimates.

In the correlation plot (Figure 2),the correlations between *pollscore*, *duration*, and $transparency_score$ are relatively low. This suggests that these predictors do not exhibit significant linear relationships with one another, which is an important consideration in regression analysis, so these predictors will be considered in the further analysis.

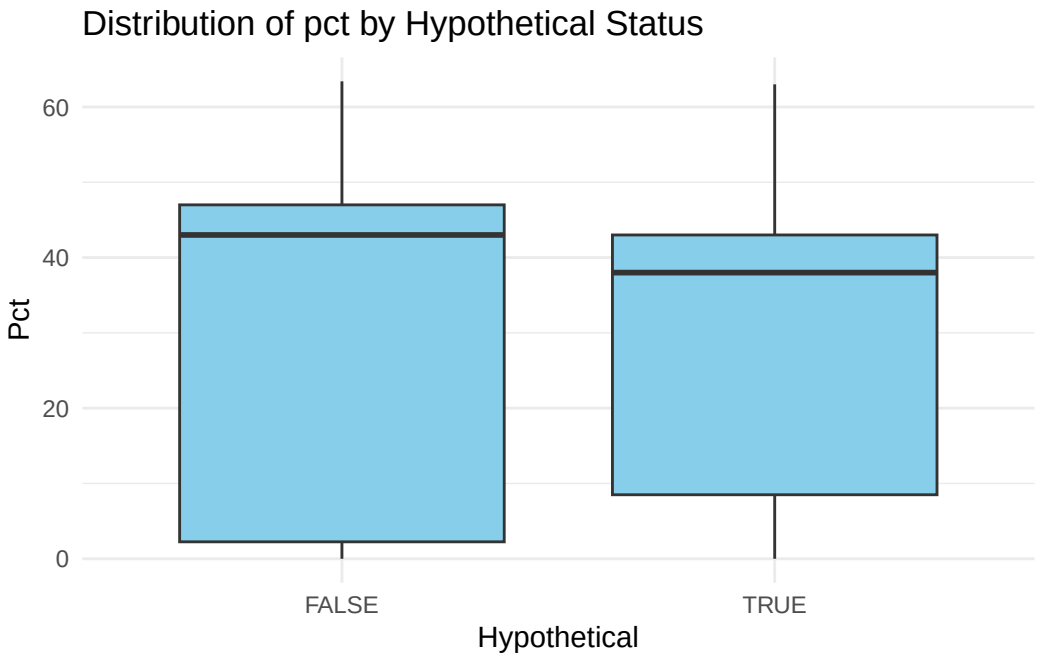

Figure 3

From the boxplot (Figure 3), voters tend to express stronger preferences when the hypothetical scenario is presented as True, leading to more polarized voting behavior. This could indicate that voters are more decisively aligned with their preferred candidate in hypothetical matchups compared to more generalized scenarios. The larger difference in the hypothetical scenario might also indicate that voter preferences are more susceptible to change based on specific campaign narratives or media portrayals, suggesting that candidates could capitalize on this by crafting targeted messages in their campaigns.

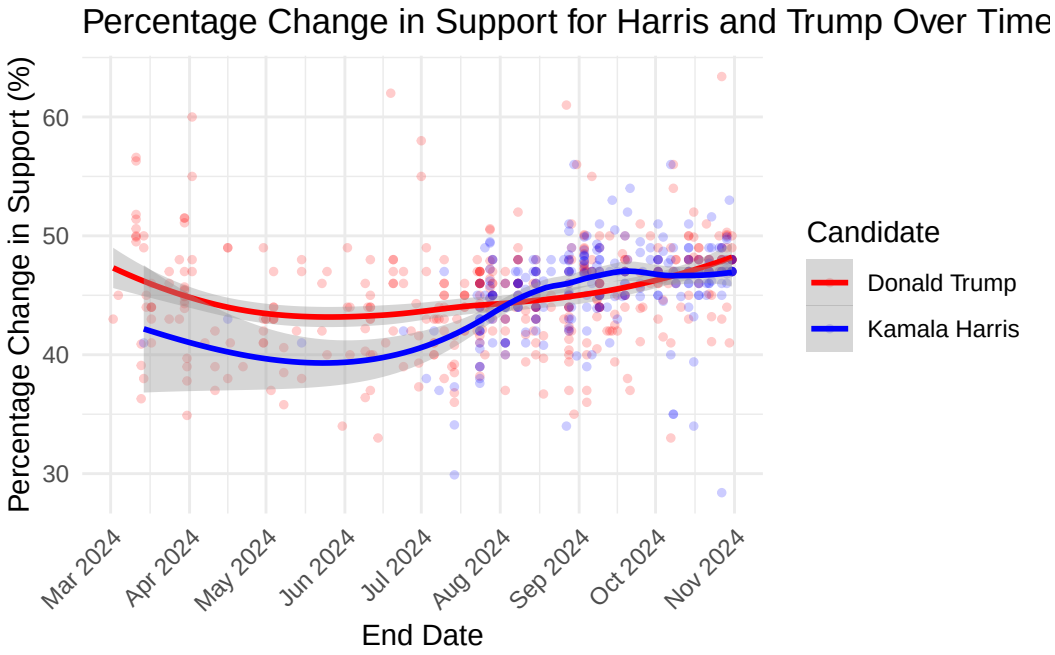

Figure 4: Support Rate over Time

To gain insights into the overall trends in the data, we performed exploratory data analysis(**?@fig-overall**) through summary statistics and visual representations. Figure 1 (below) depicts the polling percentages for each candidate over time, providing a clear view of the shifts in voter support as the election approaches."

This figure (Figure 5) presents the polling percentages across key swing states, highlighting the varying levels of support for candidates in different regions that are crucial for the election result. This visualization allows us to assess the competitive states and identify trends that may influence electoral outcomes.

```
    Welch Two Sample t-test

data:  pct by candidate_name
t = 0.21586, df = 714.48, p-value = 0.8292
alternative hypothesis: true difference in means between group Donald Trump and group Kamala
95 percent confidence interval:
 -0.5162469  0.6437917
sample estimates:
 mean in group Donald Trump mean in group Kamala Harris
                   43.90021                     43.83644
```

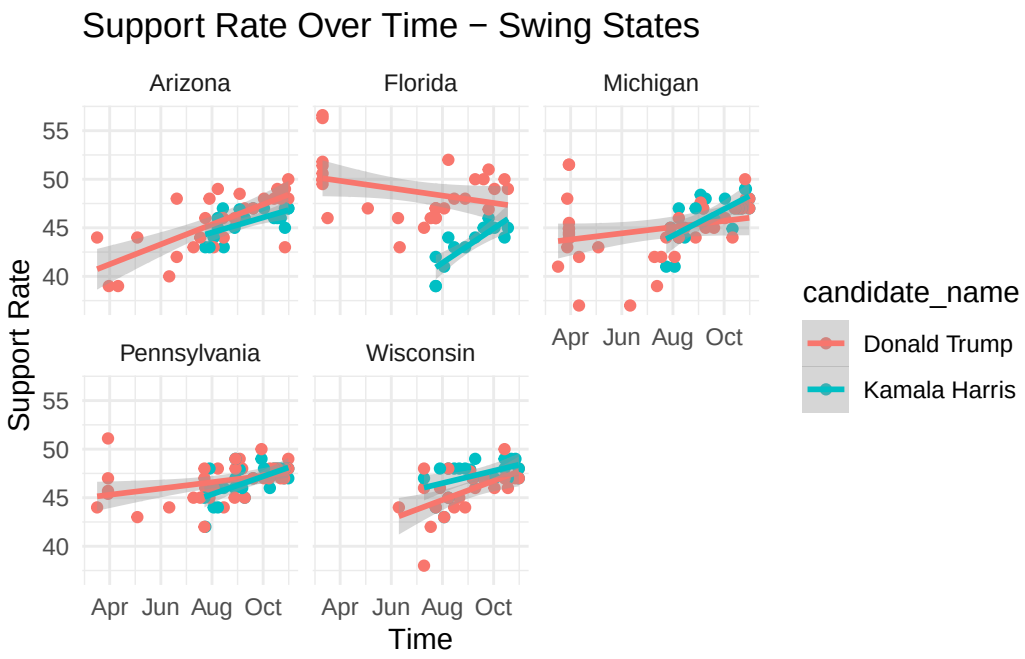

Figure 5: Interaction of a Specific Predictor and Response by Party

p-value (0.1607): This is the probability of observing a test statistic at least as extreme as the one obtained, under the null hypothesis (no difference in means between groups). Since the p-value (0.1607) is higher than the standard significance level (e.g., 0.05), it suggests that we do not have strong evidence to reject the null hypothesis. In other words, there isn't a statistically significant difference in polling percentages between Donald Trump and Kamala Harris in this dataset.

95% Confidence Interval (-0.181, 1.087): This interval estimates the range within which the true difference in means likely lies, with 95% confidence. Since 0 is within this range, it's consistent with the conclusion that there's no significant difference between the two candidates' polling percentages.

Based on the data(**?@fig-ab-test**), there's no statistically significant difference in the average polling percentages between Donald Trump and Kamala Harris. The observed difference is small and could reasonably have occurred by chance.

Since the t-test showed no significant difference, to explore and gain deeper insights, we will include an interaction between pollscore and transparency_score could indicate if support for each candidate varies by state. By including interaction terms, some significant differences that simple group comparisons miss will be noticed.

3 Model

The goal of our modelling strategy is twofold. Firstly, we introduce a primary model provide a detailed Current Votes Overview and a Historical Trends Analysis for two candidates in the upcoming election. The Current Votes Overview aggregates real-time polling data to offer an up-to-date snapshot of voter sentiment. This component emphasizes the latest trends in candidate support, enabling us to identify fluctuations and emerging patterns that could influence the election outcome. The Historical Trends Analysis examines past voting behaviors for the two candidates, analyzing historical polling data to uncover patterns and trends over time. By investigating how voter preferences have shifted in previous elections and during significant events, this analysis aims to provide insights into the factors that have historically influenced electoral outcomes.

Secondly, we have second model specifically focused on understanding voting behavior in swing states, which are critical in determining the outcome of elections. The model incorporates several key predictors, including pollscore, transparency score, duration, and party, allowing for a nuanced analysis of voter preferences. By integrating these variables, we aim to capture the complex dynamics that drive electoral decisions in swing states. Additionally, the model explores interaction terms, enable us to investigate how the relationships between these variables may vary across different contexts, revealing insights into how voter sentiment is shaped by local political landscapes and candidate messaging.

3.1 Model set-up

3.1.1 Model 1: Percentage of 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. The linear regression model is specified as follows:

Where:

- 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

```
Call:
lm(formula = pct ~ end_date + candidate_name, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-16.8382  -2.9833   0.1386   2.4028  18.8928
```

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -5.061e+01  8.044e+00  -6.292 4.33e-10 ***
end_date                     4.803e-03  4.087e-04  11.752  < 2e-16 ***
candidate_nameKamala Harris -5.859e-01  2.877e-01  -2.037   0.0419 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.596 on 1254 degrees of freedom
Multiple R-squared:  0.09923,   Adjusted R-squared:  0.0978
F-statistic: 69.07 on 2 and 1254 DF,  p-value: < 2.2e-16
```

3.1.2 Model 2: Percentage of support as a function of multiple predictors and interaction

We analyze the percentage of support by including more variables: the states that conduct the polls, the interaction between poll score and transparency score, and the polls' duration. The model is as follows:

Where:

- $beta_0$ represents the percentage support for Trump in poll i,
- $beta_1$, $beta_2$, $beta_3$, $beta_4$, $beta_5$, $beta_6$, $beta_7$, $beta_8$ are the coefficients corresponding to each predictor variable, measuring their individual effects
- ϵ_i is the error term, assumed to follow a normal distribution with a mean of 0 and variance σ^2

```
Call:
lm(formula = pct ~ pollscore * transparency_score + party + state +
    duration + candidate_name, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-18.597  -2.295   0.057   2.446  18.096

Coefficients: (1 not defined because of singularities)
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 47.65618    0.97877  48.690  < 2e-16 ***
pollscore                    1.03841    0.92197   1.126 0.260364
transparency_score          -0.06860    0.10343  -0.663 0.507357
partyREP                    -1.19168    0.30583  -3.897 0.000105 ***
```

```
stateArkansas                  0.94998    2.99572   0.317 0.751238
stateColorado                  0.72473    2.17982   0.332 0.739615
stateDelaware                 -1.09385    3.00698  -0.364 0.716121
stateFlorida                   1.25264    0.80806   1.550 0.121474
stateGeorgia                   0.84587    0.84641   0.999 0.317907
stateIowa                     -3.25947    4.17770  -0.780 0.435489
stateKansas                    2.77428    2.39300   1.159 0.246650
stateMaine                    -0.60878    2.04319  -0.298 0.765812
stateMaine CD-1               -4.17603    2.89607  -1.442 0.149685
stateMaine CD-2                0.82397    2.89607   0.285 0.776086
stateMaryland                 -7.38601    2.16411  -3.413 0.000673 ***
stateMichigan                 -1.04816    0.74014  -1.416 0.157097
stateMinnesota                -1.74192    0.97304  -1.790 0.073784 .
stateMissouri                 -0.66381    2.98697  -0.222 0.824184
stateMontana                  -1.06554    3.08219  -0.346 0.729648
stateNational                 -1.03870    0.62968  -1.650 0.099404 .
stateNevada                   -0.12680    0.91033  -0.139 0.889256
stateNew Mexico               -1.09156    1.42890  -0.764 0.445132
stateNew York                 -9.13517    4.19920  -2.175 0.029874 *
stateNorth Carolina           -0.05585    0.80609  -0.069 0.944783
stateNorth Dakota             16.24908    4.18994   3.878 0.000113 ***
stateOhio                      1.44161    1.37031   1.052 0.293088
stateOklahoma                  4.62690    2.46688   1.876 0.061055 .
stateOregon                   -1.89650    2.15558  -0.880 0.379213
statePennsylvania              0.36082    0.73595   0.490 0.624061
stateRhode Island             -4.21323    2.45886  -1.713 0.086992 .
stateTennessee                 8.03627    1.24288   6.466 1.71e-10 ***
stateTexas                     1.21445    1.38890   0.874 0.382152
stateVirginia                 -0.05814    1.69774  -0.034 0.972691
stateWest Virginia             1.81686    3.05083   0.596 0.551649
stateWisconsin                -0.36453    0.79087  -0.461 0.644970
stateWyoming                   2.97885    3.49408   0.853 0.394156
duration                      -0.13716    0.06263  -2.190 0.028787 *
candidate_nameKamala Harris         NA         NA      NA       NA
pollscore:transparency_score  -0.30076    0.12086  -2.489 0.013021 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.139 on 840 degrees of freedom
  (379 observations deleted due to missingness)
Multiple R-squared:  0.2032,    Adjusted R-squared:  0.1681
F-statistic: 5.788 on 37 and 840 DF,  p-value: < 2.2e-16
```

3.1.3 Model justification

The use of multiple linear regression with interaction terms is justified for this analysis of candidate support for several compelling reasons. First and foremost, incorporating interaction terms allows for a nuanced examination of how the relationship between the predictors and the response variable—percentage of support for each candidate—varies across different groups.This is particularly important in political analysis, where the effect of factors like state or transparency scores on candidate support may not be uniform. By including these interaction terms, the model captures the complexities of voter behavior, reflecting how the impact of one predictor may change depending on the level of another.

Secondly, linear regression is well-suited for this analysis because it assumes a linear relationship between the predictors and the response variable. Given that the factors influencing voter support, such as polling quality and sample size, are anticipated to have a linear effect, this model aligns well with the data. The interactions add complexity but remain grounded in the linear framework, allowing for straightforward interpretation of the results. Moreover, the inclusion of both continuous and categorical predictors—such as "numeric_grade," "transparency_score," "state," and "party"—enhances the model's flexibility.

Furthermore, the use of linear regression facilitates the identification of potential issues such as multicollinearity among predictors. By carefully selecting which variables to include and excluding highly correlated predictors, the model enhances the reliability and interpretability of the coefficient estimates. This attention to model specification ensures that the insights drawn from the analysis are robust and meaningful.

In conclusion, the integration of interaction terms within the linear regression framework provides a powerful method for understanding and predicting candidate support in the context of the 2024 U.S. Presidential Election. This approach not only captures the additive effects of individual predictors but also reveals the intricate ways in which these factors interact to influence voter behavior, making it an ideal choice for this analysis.

4 Results

4.1 Result of Model 1

for model 1 we focus on two predictors, which are the end date and candidates. We build a linear regression model to identify the relationship between pct and two variables end date and candidate. The median is 0.1306, which is very small. In this context, if we only consider end date, their pct is very close.

Table 1: Overall-Trend Modell Result

4.2 Result of Model 2

for model 2 we added more predictors, including transparency_score, party, state, duration, candidate_name. We build a multivariable linear regression to identify predictors for pct in swing states. from the previous model(**?@fig-model2**), we can see transparency score, duration and few states have negative coefficients.

Generally, Model 2 has more predictors than model 1 and it is more predictive than model 1

5 Discussion

5.1 Interpretation of Model 1 Findings

Model 1 examined the relationship between the percentage (pct) and two primary predictors: the end date and the candidate. The linear regression analysis yielded a median effect size of 0.1306, indicating that both the end date and the candidate have minimal individual impacts on the pct. This suggests that when only these two variables are considered, the variation in pct across different candidates and end dates is relatively insignificant. The closeness of pct values implies that neither the timing of the campaign's conclusion nor the identity of the candidate alone are strong determinants of electoral performance within the scope of this model.

This limited influence observed in Model 1 underscores the necessity of incorporating additional variables to capture the complexities of electoral dynamics. Relying solely on temporal factors and candidate identity may overlook critical elements that drive voter behavior and election outcomes.

Enhanced Insights from Model 2 To build upon the initial findings, Model 2 introduced additional predictors: transparency_score, party, state, duration, and candidate_name. This multivariable linear regression aimed to provide a more nuanced understanding of the factors influencing pct in swing states.

The inclusion of transparency_score and duration revealed significant negative coefficients. Specifically, higher transparency scores are associated with a decrease in pct, while longer campaign durations also correlate with lower pct values. These findings suggest that increased transparency, while generally perceived positively, may lead to voter skepticism or fatigue, thereby reducing electoral support. Similarly, extended campaign durations might contribute to voter disengagement or diminishing enthusiasm over time.

Additionally, the identification of certain states with negative coefficients highlights the unique political landscapes and voter behaviors inherent to these regions. These states may possess specific demographic, socioeconomic, or political characteristics that influence the effectiveness of campaign strategies and, consequently, the pct outcomes.

5.2 Implications for Campaign Strategy and Policy

The results from Model 2 offer valuable insights for political campaigns targeting swing states. The negative association between transparency and pct indicates a delicate balance that campaigns must maintain. While transparency is crucial for building trust, excessive disclosure or perceived overexposure might inadvertently erode voter support. Campaigns should therefore strategize on maintaining transparency without overwhelming the electorate.

5.3 The inverse relationship

between campaign duration and pct emphasizes the importance of timing and sustained engagement. Prolonged campaigns may lead to voter fatigue, suggesting that concentrated and strategic campaigning within optimal timeframes could be more effective in securing higher pct values.

Furthermore, the state-specific negative coefficients imply that tailored strategies are essential. Understanding the unique attributes of each swing state can inform targeted approaches that resonate with local voter concerns and preferences, enhancing overall campaign efficacy.

5.4 Weaknesses and next steps

While both models offer valuable insights, there are several limitations that should be acknowledged. The reliance on polling data are variability and potential biases, including factors like non-response bias, sampling errors, and discrepancies in data collection techniques among different polling organizations. Moreover, both models presuppose a linear trajectory of polling shifts leading up to election day, which may overlook abrupt changes in voter sentiment caused by unexpected events, shifts in campaign strategies, or emerging social issues.

Additionally, the Baseline Model's use of national aggregates can obscure the distinct political contexts within individual states, potentially glossing over the complexities of regional voting patterns. Conversely, the Primary Model's focus on state-specific data demands a wealth of local information that may not be uniformly accessible or trustworthy across all swing states, limiting the comprehensiveness of the analysis.

Future research can enhance this study by incorporating real-time data streams, such as social media sentiment and Google Trends, to dynamically capture shifts in voter opinion as events unfold. Machine learning techniques, like random forests or gradient boosting, could

reveal complex, non-linear interactions among variables, while Bayesian hierarchical modeling would improve post-stratification by factoring in prior knowledge of demographics and regional preferences. Including economic indicators, such as inflation and unemployment, could help explain shifts in voter sentiment, while geospatial analysis at district or county levels could provide more detailed insights into local influences and differences between urban and rural areas.These approaches would deepen the understanding of voter preferences and enhance the predictive power of election models.

Appendix

Appendix A: Emerson College Pollster Methodology Overview

Emerson College Polling is a non-partisan organization which established 25 years ago as classroom excercise and was transformed into an innovative, nationally-ranked polling center in 2012 ((**Aboutus?**)). In general, the emerson college define their population to be "registered voters, likely voters, or residents"((**Aboutus?**)), but for the 2024 US President Election Poll, the population is defined to be only "likely voters" based on "2024 national poll".

Emerson College Polling recruits respondents using a diverse recruitment strategy, including "MMS-to-Online", "Online Opt-in Panel", "IVR (Interactive Voice Response)" ((**Aboutus?**) & Polls). Generally there is an additional approach named "Emails"((**Aboutus?**)), but we did not find the sign that it was used in 2024 Election surveys(National polling), so we do not include it in this paper. "MMS-to-Online" is an approach which the target population receive text messages with a custom graphic that invite them to take the online-survey hosted on Qualtrics. The respondents are select randomly from "state voter files provided by Aristotle". In Online Opt-in Panel approach, respondents are invited to take a "screening questionnaire"((**Aboutus?**)) through an online opt-in panel "provided by CINT"((**Aboutus?**)), respondents who pass the screening questionnaire are directed to the survey. Data quality are measured using additional screening questions, respondents who do not meet data quality measures are removed from the survey. Based on "Polling", respondents are selected from "L2 voter file data provided by Rep Data". In IVR (Interactive Voice Response) approach, respondents reveive automated calls, they answer the survey using their telepones.IVR is not used in some states where it is prohibited ((**Aboutus?**)) Respondents are selected randomly from "state voter files, provided by Aristotle"((**Aboutus?**)).

Based on the method of recuiting people and the target population, we can deduce the sampling frame are likely voters who are able to use landline telepone and cellphone. The sample size of likely voters is 1000 according to "November 2024 National Poll: Trump and Harris Remain Locked in Tight Race". The data sets were "weighted by gender, education, race, age, party registration, and region based on 2024 likely voter modeling" according to "November 2024 National Poll: Trump and Harris Remain Locked in Tight Race".

Emerson College Polling employs a random sampling approach, specifically through MMS-to-online surveys and IVR (Interactive Voice Response) calls to landlines. In "MMS-to-Online" and "IVR (Interactive Voice Response)", a non-probability sampling approach is used, specifically, a random sampling approach. In "Online Opt-in Panel", a probability sampling approach is used, but we do not find the specific sampling approach. In this case, we conclude the general sampling approach is random sampling approach. A random sampling approach has notable advantages. It helps reduce selection bias by providing multiple ways for individuals to participate based on their preferences, increasing the chances of capturing a diverse range of respondents. This can enhance the representativeness of the sample, as each individual has an equal chance of being selected, potentially making the results more generalizable to the larger

population. However, this approach has limitations. Each method, such as IVR calls versus online surveys, can introduce slight variations in response patterns, which can lead to inconsistencies in the data and reduce reliability. Furthermore, implementing a random sampling approach is often time-consuming and costly, as it requires extensive planning and resources to reach a wide, representative audience effectively

There is also no clear indication of how Emerson College Polling handled non-responses. Consequently, we cannot rule out the possibility of non-response bias, which arises when non-respondents differ meaningfully from respondents. Additionally, we cannot assess the impact of non-response or achieve more representative polling results.

Emerson's approach to questionnaire design(2024) emphasizes precision, objectivity, and participant engagement. By employing clear and straightforward language, Emerson minimizes the risk of misinterpretation, ensuring that respondents can easily understand and accurately respond to questions. The use of neutral wording is crucial in preventing any inadvertent bias, allowing participants to express their views without feeling led toward specific answers. In addition to clarity, Emerson utilizes a randomized order for both questions and response options. Such randomization enhances the survey's reliability and validity, contributing to a more accurate representation of public sentiment. Emerson's commitment to a balanced and methodologically sound approach ensures that findings are credible and actionable.

Quality control is integral to Emerson's survey process(2024), bolstering the integrity of the data collected. The organization implements rigorous measures to verify respondent identities and track response patterns. Instances of inconsistent or suspicious responses are flagged for further review, with problematic data excluded from analysis, which helps maintain a high standard of data quality.

While Emerson(2024) focuses on asking straightforward questions, it may limit the depth of insight into complex issues. Respondents might not have the opportunity to articulate their nuanced opinions, particularly on multifaceted topics such as political preferences or economic concerns. Although this approach yields valuable top-line data, there is a trade-off in terms of capturing the subtleties that drive voter behavior and opinion formation.

.1 Appendix B: Methodology and Survey Design for 2024 U.S. Presidential Election Forecast

.2 B.1 Sampling approach

Total Sample Size: 6,000 respondents (1,000 participants per targeted demographic). Target Population: Eligible voters across the following key demographics: urban areas, suburban regions, and rural communities in swing states: Florida, Ohio, Iowa, and Texas.

.3 B.2 Recruit respondents

Sampling Breakdown by Demographic Group: - Urban Voters: 500 participants per state (total 2,000) – reached through local community organizations and online forums. - Suburban Voters: 400 participants per state (total 1,600) – reached via neighborhood associations and targeted social media campaigns. - Rural Voters: 100 participants per state (total 400) – reached through local agricultural fairs and county events. - Additional Participants: 1,600 participants – evenly distributed among all demographics to reach the total sample size. Stratified Random Sampling: Stratify by demographic factors such as age, gender, ethnicity, and socioeconomic status. Ensure representation reflects the diversity of the targeted communities, based on recent census data. Weighting Strategy: Implement post-stratification weighting to correct for any overrepresentation or underrepresentation within the sample, ensuring it aligns with the overall voter demographic in the selected states.

.4 B.3 Data validation

Techniques for Data Quality: Eligibility Verification: Implement screening questions to confirm respondent eligibility (e.g., age, voter registration status). Data Cleaning Procedures: Utilize automated checks for inconsistencies, such as duplicate responses or incomplete surveys, followed by manual review of flagged entries. Follow-Up Verification: Conduct a follow-up survey with a random sample of participants to validate the accuracy of the original responses, ensuring reliability in the findings.

.5 B.4 Other relevant aspects of interest

Other relevant aspects of interest could be: Voter Registration Status: Confirm whether people are registered for the upcoming election. Candidates' Choices: Provide a brief overview of the major candidates and their platforms, helping voters consider their positions on essential issues like the economy, healthcare, and climate change. Personal Values and Future Vision: Define each person's vision for the country's future and the progress they wish to see in areas such as healthcare, education, or environmental policy. This can help identify which candidates' policies and values align with their vision.

.6 B.5 Google Form Link

This survey aims to leverage the strengths of online platforms to maximize participation and ensure a diverse respondent pool. Online surveys provide unique advantages, allowing respondents to participate at a time and place that suits them, which is particularly helpful for busy individuals with varying schedules. To enhance accessibility, I designed the survey as an online format, allowing people to complete it from any location with internet access. The

survey is intentionally concise, designed to be quick and easy to complete, which helps keep participants engaged and minimizes the risk of incomplete responses. This approach reflects a commitment to inclusivity and data quality, aiming to gather valuable insights from a wide array of perspectives while adhering to established best practices in survey design.

Survey Platform: Google Forms Link: https://docs.google.com/forms/d/e/1FAIpQLScllb1CYoM3EWHs8txFgn

Survey Structure: Demographics Section: Age, gender, ethnicity, income, education, and demographics residence. Age, gender, ethnicity, income, education, and place of residence. Voting Preferences Section: Voter Registration Status and Candidates choice. Personal Values and Vision: Public Sentiment and Values and Priorities (e.g., economy, climate change, Immigration) Thank You Message

A References

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

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

Wickham, Hadley. 2016. Ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. Readr: Read Rectangular Text Data. 27. https://readr.tidyverse.org.

Emerson. 2024. ECP_Swing_States_10.10.24. https://docs.google.com/spreadsheets/d/1KdGpVQ0P7AVCD

Emerson. 2024. October 2024 State Polls: Mixed Movement Across Swing States Shows Dead Heat. https://emersoncollegepolling.com/october-2024-state-polls-mixed-movement-across-swing-states-shows-dead-heat/.

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