Forecasting the 2024 US Presidential Election*

Tianning He Julia Lee Shuangyuan Yang

November 4, 2024

1 Introduction

2 Data

To simulate, test, download, clean, and model data throughout this analysis, the statistical programming language R was used (R Core Team 2023). Specific libraries that assisted the analysis include `tidyverse` (Wickham et al. 2019), `dplyr` (Wickham et al. 2023),`tinytex` (Xie 2019), `ggplot2` (Wickham 2016), `knitr` (Xie 2015), `arrow` (Richardson et al. 2024), `here` (Müller and Bryan 2020), and `testthat` (Wickham 2011).

2.1 Presidential General Election Polls Data

The present analysis uses Presidential Election Polls data provided by 538 (FiveThirtyEight 2024). The data contains records of various pollsters, several aspects of each pollster (e.g. a poll identification number, poll quality characterized by a numeric grade, sample size, etc.), candidates for the presidential election, and the percentage of support (pct) each candidate received based on the polls. Poll results are periodically updated to the 538 website and can be downloaded by users as csv files. The types of pollsters that can be included in the data by 538 are determined by various methodological and ethical standards set by 538 (Radcliffe and Morris 2023).

As the original data acquired by 538 is an extensive dataset, the following will provide details for the variables that are considered within the present analysis.

- **Pollster**: This variable contains the names of polling organizations that conducted each poll (e.g. YouGov, AtlasIntel, Siena, etc.).

*Code and data are available at: https://github.com/JuliaJLee/Forecasting_US_Election_2024.git

- **Numeric Grade**: This variable reflects the quality or reliability of a pollster. It ranges from 0.5 to 3.0, increasing by 0.1. A larger number indicates a higher-quality pollster.
- **Population (pop.)**: This variable represents the voting status of those who respond to a given pollster. Respondents for different pollsters can be likely voters (i.e. voters who are more likely to vote on election day) or registered voters (i.e. everyone who can vote).
- **State**: Polls can be state-specific (i.e. conducted within a specific state) or they can be national. In the original data, blank values indicate national polls.
- **Sample Size**: This variable reflects the total number of respondents for each poll within the data.
- **Hypothetical (hypo.)**: Polls can ask respondents about a hypothetical match-up between different candidates. With this, polls that do not consider hypothetical match-ups are denoted as "False" in the data.
- **Candidate Name**: This variable contains the names of all the candidates that were asked about in the poll. For this analysis, the presidential candidates, Kamala Harris and Donald Trump, are considered.
- **Poll Start and End Dates**: The start and end dates indicate when a poll was launched and when that same poll was closed.
- **Percentage of Support (pct)**: The percentage of support (pct) reflects the percentage of support a candidate received in a given poll.

Table 1: Organized Presidential General Election Poll Data

Pollster	Numeric Grade	State	Start Date	End Date	Sample Size	pop.	hypo.	Candidate Name	pct
AtlasIntel	2.7	North Carolina	10/30/24	10/31/24	1373	lv	FALSE	Kamala Harris	46.7
AtlasIntel	2.7	North Carolina	10/30/24	10/31/24	1373	lv	FALSE	Donald Trump	50.7
AtlasIntel	2.7	North Carolina	10/30/24	10/31/24	1373	lv	FALSE	Kamala Harris	47.0

Table 1 displays the data that is used throughout the analysis and illustrates each of the variables described above. Summary statistics for these variables can be found in the Appendix (SECTION).

Putting these variables together, the analysis is able to construct a model that forecasts the percentage of support (pct) for Harris and Trump by considering high-quality pollsters, a population of voters who a more likely to vote on election day, polls from various states, poll sample sizes, and the times at which polls were conducted. An in-depth explanation as to why the variables above are selected is provided in (SECTION).

2.2 An Account on Measurement

From the day of the week to different survey methods (e.g. email, phone, mail, etc.), there are many things that can influence how voters' opinions or preferences are shaped and how those opinions are reduced down to a value. The percentage of support (pct) for a candidate is the value that is meant to reflect voters' opinions on who should be president. As a single percentage is the end result of a survey that was conducted by a pollster, voters' thoughts and opinions can be translated into an entry in a dataset by considering the ways in which pollsters build their surveys along with how they analyze and interpret the responses. Opinions may be converted into data by calculating average or percentages of responses to closed-ended questions. Preference towards a particular candidate or party could be measured through the coding and counting of themes or key words within responses to open-ended questions. Additionally, how responses are interpreted (e.g. how themes are generated for open-ended responses) can impact how voters' thoughts are measured. Thus, complex thoughts and opinions can be measured through careful consideration about what is being asked, how it is being asked, and how the responses are interpreted.

A detailed account about the methodology of a pollster found within the data obtained from 538 along with an idealized methodology and survey can be found in the Appendix (SECTION).

3 Model

The objective of the present analysis is to forecast the percentage of support both presidential candidates, Kamala Harris and Donald Trump, will receive in the final week (October 27, 2024 to November 2, 2024) leading up to the election. The Presidential Poll data provided by 538 (FiveThirtyEight 2024) reflects voters' opinions and preferences about who should be the next President of the United States across time. As those opinions or preferences are subject to change as time goes on, this model seeks to account for this variability by building "seasonal indexes" and using them with a linear model to forecast the percentage of support for both presidential candidates.

In this model, a "season" is referred to as a 7-day week that is found between Sunday August 4, 2024 and Saturday October 26, 2024. The start date is August 4, 2024 because this date allows there to be enough data to observe the percentage of support for both candidates over several weeks. The end date is October 26, 2024 as this leaves roughly a week (October 27, 2024 to November 2, 2024) before the election on Tuesday November 5, 2024 to ensure that a forecast for this final week can be made.

Further, this model considers the following variables:

- **Pollsters with a numeric grade of 2.7 and above**: This model uses a cut-off of 2.7 for the numeric grade to strike a balance between the amount of data and the quality

of the data. The design of this model requires that there is at least one poll that was conducted in each "season" or week, and this could not have been satisfied if only pollsters with a numeric grade of 3 are considered. A numeric grade of 2.7 and above allows this model to have sufficient data as well as data of high quality.

- **The likely voter (lv) population**: Likely voters are defined as voters who intend to vote on election day. With this, the model focuses on this particular voter population to generate a forecast that more closely resembles election day.
- **States**: This model looks at polls that were state-specific rather than national polls. Though each state is not considered individually in this analysis, this model considers these state-specific polls as they also allow for sufficient data to be used within the model.
- **Sample Size**: Poll sample size is used within the model to pool the poll data for each week within the time period defined by the model above.
- **Polls that did not ask about hypothetical match-ups**: As this model aims to forecast and compare support for the presidential candidates in the current 2024 election, hypothetical match-ups are not considered.
- **Presidential Candidates**: This model looks at the percentage of support for both Kamala Harris and Donald Trump throughout the analysis.
- **Poll Start and End Dates**: Start and end dates are important variables in the model as they are used to categorize poll data into the different "seasons" or weeks between August 4, 2024 and October 26, 2024.
- **Percentage of Support (pct)**: This is the target variable to be estimated by the model.

3.1 Model Process

Code that runs through the following steps can be found in the repository linked on page 1.

3.1.1 Step 1: Organize Poll Data for each Candidate by Week

For this model, a total of 12 weeks of poll data is analyzed. The exact start and end dates of each week (defined by the model) can be found below in Table 2. Every four weeks corresponds to a month. The first four weeks are in August 2024, the next four weeks are in September 2024, and the last four weeks are in October 2024.

Table 2: Weeks Defined by the Model

Week	Dates
1	Aug. 4-10
2	Aug. 11-17
3	Aug. 18-24
4	Aug. 25-31

Table 2: Weeks Defined by the Model

Week	Dates
5	Sept. 1-7
6	Sept. 8-14
7	Sept. 15-21
8	Sept. 22-28
9	Sept. 29 - Oct. 5
10	Oct. 6-12
11	Oct. 13-19
12	Oct. 20-26

Using the start and end dates of the polls, the model first filters on the polls that were conducted between August 4, 2024 and October 26, 2024 for each candidate. Then, it assigns each poll to a week as outlined in Table 2. An example outcome is shown below for Kamala Harris (Table 3).

Table 3: Poll Data Organized by Week For Harris

Sample Size	Candidate	Percentage of Support (pct)	Week
1,000	Kamala Harris	42.5	1
619	Kamala Harris	50.0	1
661	Kamala Harris	50.0	1
693	Kamala Harris	50.0	1
1,738	Kamala Harris	50.0	2
1,000	Kamala Harris	49.3	2

Each row in the outcome above (Table 3) represents a poll, and the "Week" column indicates the week in which that poll occurred. For example, the first four rows show polls that were conducted in the week of August 4 to 10, 2024. The last two rows show polls that occurred between August 11 to 17, 2024 (i.e. Week 2). Each poll's sample size and pct estimate are also included.

3.1.2 Step 2: Pool Poll Data by Week for both Candidates

With polls organized by week, the model now pools all the polls that occurred within a single week to generate a weighted average estimate of pct for that week.

To pool the polls for each week, this model first creates a weight for each poll by:

- (1) Finding the sum of the sample sizes of all polls in a given week

- (2) Dividing each sample size of each poll within that week by the sum found in the previous step

An example of these two steps would be to find the sum of the first four sample sizes for week 1 in Table 3, and then divide 1000, 619, 661, and 693 from the first four rows of Table 3 by the sum of the first four sample sizes.

Next, each weight for each poll within a given week is multiplied to the corresponding pct estimate. For example, by using Table 3, the model would multiply the quotient of (1000/sum of sample sizes) by 42.5, which is the pct estimate that corresponds to the poll with a sample size of 1000 in row 1 of Table 3.

Lastly, by taking the sum of the products of each weight and the corresponding pct estimate, the model produces a weighted average pct estimate for each week. An example outcome is shown below for Kamala Harris (Table 4).

Table 4: Weighted Average Percentage of Support for Harris By Week

Week	Weighted Average Percentage of Support (pct)
1	47.5
2	47.8
3	47.0
4	48.7

Table 4 shows the average percentage of support that Harris received in the first four weeks. For instance, in Week 1 (i.e. during the week of August 4 to 10, 2024), Harris received an average support of 47.5% across the polls that occurred in within this time period.

3.1.3 Step 3: Fit a Regression Model for Each Candidate using the Pooled Poll Data

Using the weighted average percentage of support (pct) for each week as the response variable, the model performs two regression analyses to predict the support for each candidate during each of the 12 weeks. The linear models are structured as follows:

$$\hat{y_i} = b_0 + b_1 \cdot w_i + \epsilon_i$$

where

- $\hat{y_i}$ represents the percentage of support (for Harris or Trump),
- b_0 represents the intercept of the linear models,
- b_1 represents the effect of each week,

- w_i represents the time period of a week (i = 1, 2,..., 12)
- ϵ_i captures the error within the linear models

Summary outputs for each model (one for Harris and another for Trump) along with model diagnostics to validate these models can be found in the Appendix (SECTION).

3.1.4 Step 4: Find Seasonal (i.e. weekly) Indexes to Forecast Support for Harris and Trump

With the linear models fitted above, the model now creates a "seasonal" or weekly index for each week so that the week leading up to the election can be predicted while accounting for differences in voter opinions across different time periods.

A seasonal index for each week is calculated by first computing the ratio,

$$\frac{y}{\widehat{y}_i}$$

.

For each week, the model takes the weighted average percentage of support (pct) that was found in Step 2 (Section 3.1.2) and divides it by the predicted value that is found using the linear model. This produces an outcome like the following in Table 5.

Table 5: Ratios for Each Week

Week	Weighted Average Percentage of Support (pct)	Predicted Average Percent of Support (pct)	Ratio
1	47.48	48.08	0.988
2	47.76	48.11	0.993
3	47.00	48.13	0.977
4	48.67	48.15	1.011

Now, the since the data is manipulated such that every four weeks corresponds to a month (August, September, and October), it follows that the final week (October 27, 2024 to November 2, 2024) that this model aims to forecast is Week 13 and the first week of the next month, November. So, this model computes the average of the ratios for the first, second, third, and fourth weeks across each month to obtain a "seasonal" or weekly index that can be used to forecast Week 13. An example outcome is shown below Table 6.

Table 6: Seasonal (Weekly) Index for Each Week Across 3 Months

Week 1	Week 2	Week 3	Week 4
0.993	0.995	1.011	1.001

Using the seasonal index for Week 1 (0.993) presented in Table 6, this model can forecast the percentage of support that Harris will receive in Week 13 by:

- (1) Plugging in $w = 13$ to the linear model for Harris to predict her percentage of support
- (2) Multiplying the predicted value from the linear model by the seasonal index

The example outcomes provided throughout this section are for Kamala Harris only. It is important to note that the same process is also repeated for Donald Trump within the model.

3.2 Evaluating the Model

By pooling the polls for each of the defined weeks, the model assumes that the polls are unbiased – which is often not the case. While pooling polls that have occurred in a similar time period provides more precision than a single poll, a limitation of this model is that it overlooks the potential biases that can exist within the polls. Biases within polls can arise from their methodology, their audience, and the location in which the poll was conducted. As these variables are not explicitly considered by the model, it would not be appropriate to apply this model to forecast percentage of support as a function of different methodologies, voter populations, or states.

Despite these limitations, this model's strength lies in its ability to account for variations across time. This approach of using seasonal indexes and regression to forecast the percentage of support (pct) for both presidential candidates is able to capture seasonal (i.e. weekly) variation within the percentage of support candidates received and assess long-term trends. As such, this model can provide both a numerical outcome (i.e. a forecasted percentage of support) for each candidate along with a means to observe how the percentage of support for the presidential candidates has changed over time. As these strengths align with the objective of the analysis to forecast the percentage of support the presidential candidates will receive in the final week (October 27, 2024 to November 2, 2024) leading up to the election, this model is employed to obtain the findings presented in the next section (Section 4).

4 Results

4.1 Change in the Percentage of Support (pct) for Both Presidential Candidates Over Time

This initial analysis aims to understand how the percentage of support for both candidates has changed over the course of the 12 weeks defined by the model.

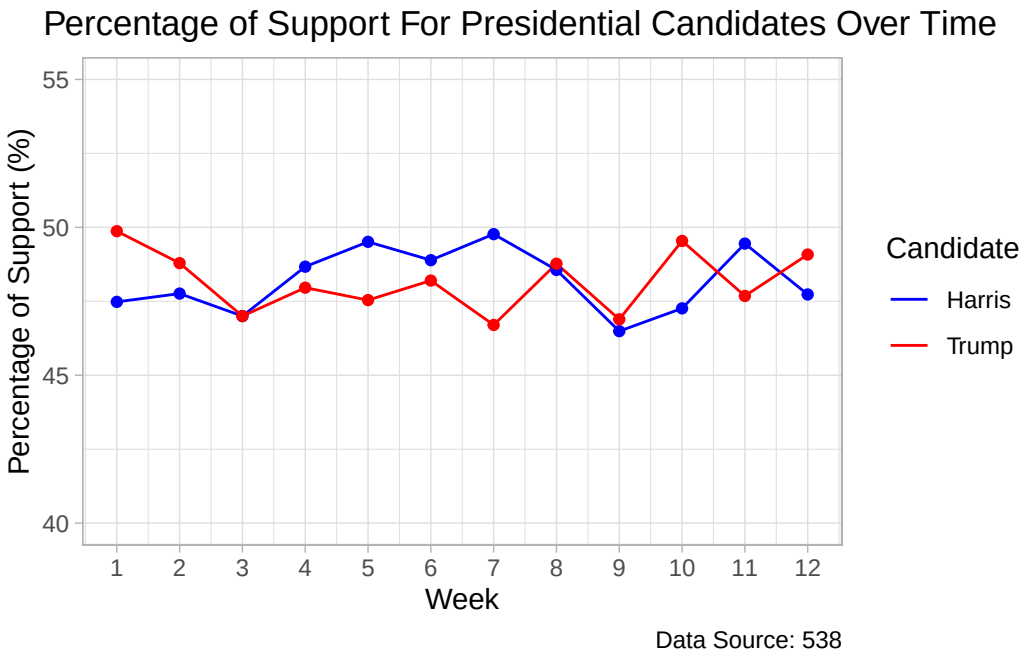

Figure 1: Percentage of Support for Both Candidates Over 12 Weeks (August 4 to October 26, 2024)

Figure 1 presents the weighted average percentage of support for each presidential candidate over time. While both candidates appear to have had fluctuations in the percentage of support they received, the difference between their percentage of support is smaller in Week 12 than it was at Week 1. In week 1, Trump received an additional 2.39% of support compared to Harris, and in Week 12, his percentage of support was 1.35% higher than Harris'. This pattern further indicates that as support for Trump has dropped, support for Harris has increased over the course of 12 weeks.

4.2 Predicting the Percentage of Support (pct) for Both Presidential Candidates Over Time

Figure 2 and Figure 3 provides a comparison of the predicted percentage of support from the linear model and the weighted average percentage of support that was computed with the poll data for Harris and Trump. Though both models do not seem to accurately capture the fluctuations in support each candidate received over time, they do appear to capture the overall increase in support for Harris along with the overall decrease in support for Trump over the 12 weeks. This further gives rise to the idea that while the percentage of support for Trump decreased, the percentage of support for Harris increased as the weeks went by.

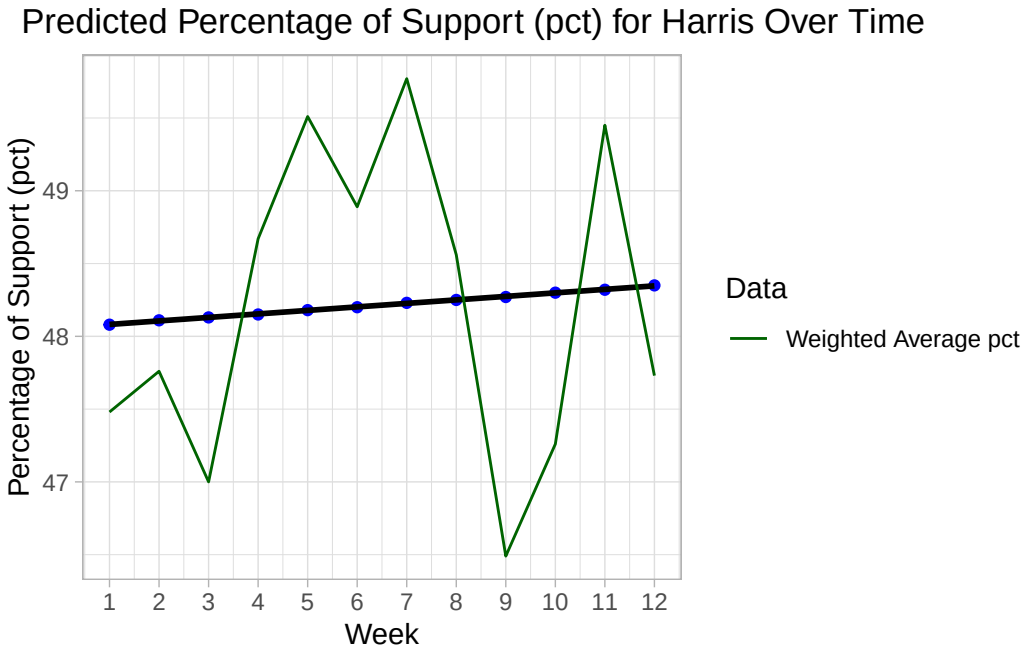

Figure 2: Predicted Percentage of Support for Harris Over 12 Weeks (August 4 to October 26, 2024)

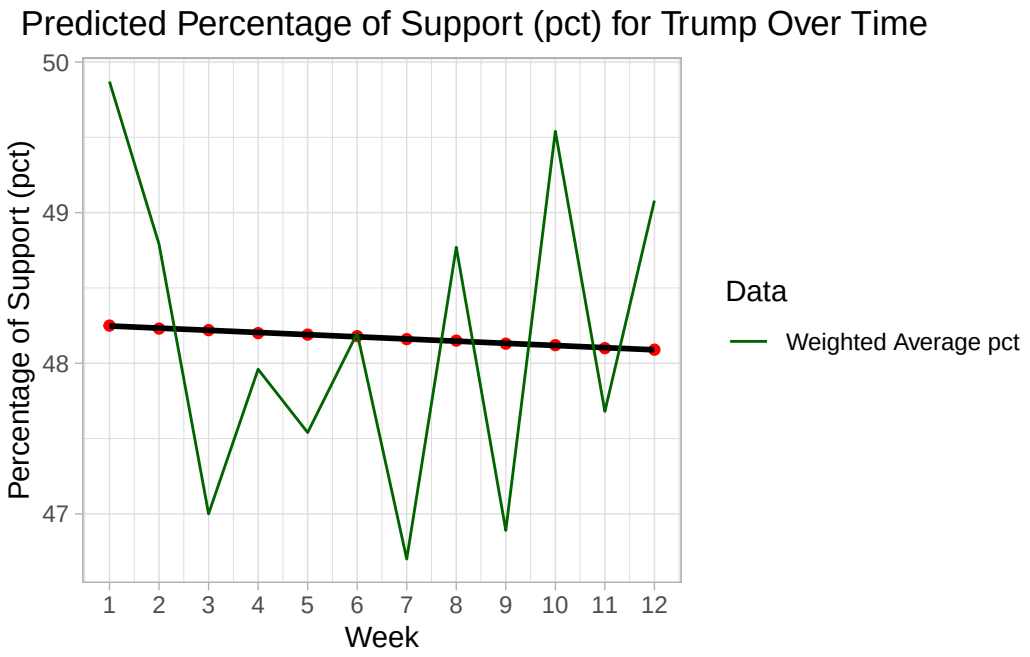

Figure 3: Predicted Percentage of Support for Trump Over 12 Weeks (August 4 to October 26, 2024)

4.3 Forecasting the Percentage of Support (pct) for Both Presidential Candidates For Week 13

The main objective of this model is to forecast the percentage of support Harris and Trump will receive in the final week before the election. This week - October 27, 2024 to November 2, 2024 - is Week 13, and the model uses a seasonal index along with a predicted value from the linear model to compute a forecast. The forecasted percentages of support for each candidate are shown in Figure 4 below.

With Figure 4, it appears as the presidential election will be a tight race as the forecasts of percentage of support for Harris and Trump in Week 13 are very similar. In the final week leading up to the election (October 27, 2024 to November 2, 2024), Harris is expected to receive more support while Trump is expected to receive less support in Week 13. The exact forecasts of the model are as follows:

- This model forecasts that the percentage of support for Harris in the week leading up to the election will be 48.03%.
- On the other hand, the model forecasts that the percentage of support for Trump in the week leading up to the election will be 47.98%.

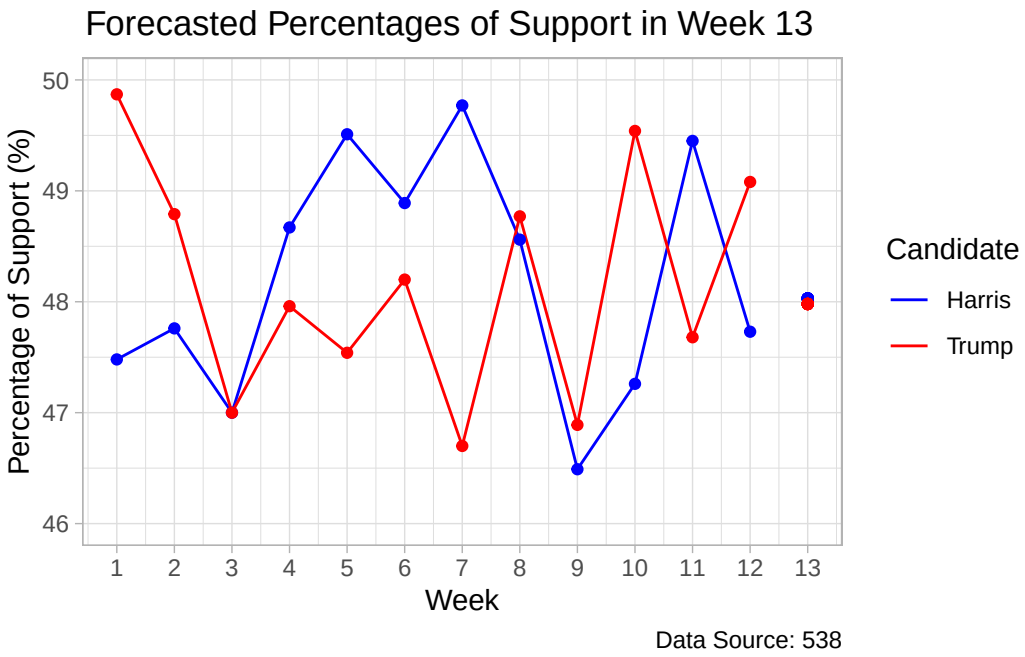

Figure 4: Presenting the Forecasted Percentages of Support for Harris and Trump in Week 13

5 Discussion

As the 2024 US election approaches, the present analysis forecasts the percentage of support both presidential candidates will receive in the final week leading up to election day on Tuesday November 5, 2024. Through the use of seasonal indexes and a regression model, this analysis takes in to account how voter opinions or preferences change over time to generate more precise forecasts. With this, an understanding of how support for Harris and Trump has changed over time is provided in addition to the model's forecasts.

This analysis finds that while support for both Harris and Trump fluctuate over the course of the 12 weeks defined by the model, overall support for Harris appears to increase over time. In contrast, despite having a high percentage of support in Week 1, Trump is found to have received less support in Week 12. This pattern indicates that the percentages of support for Harris and Trump are converging as time goes on, implying that the election will be a tight race that may need to be determined by the swing states. Additionally, the forecasts computed by the model show that Harris will receive greater support than Trump during the week before the election (Week 13). However, as the difference in support that Harris and Trump will receive during this time is extremely small (0.05%), this result also gives rise to the notion that the 2024 election will continue to be a close race until election day.

As briefly mentioned in Section 3.2, a limitation of this analysis is that it assumes that polls

are unbiased to pool them together. By not considering the effects of methodology, vote population (e.g. registered voters vs. likely voter), different states, and poll quality, the model may be amplifying the biases of different polls as it pools them together (Jackman 2005). Another limitation of this model is that it considers 12 weeks of data to forecast. Though this is a sufficient amount of data for the model, a stronger representation of the change in candidate support over time as well as more robust forecasts may be produced with a larger quantity of data.

Given these limitations, moving forward, this analysis should be adapted to incorporate a way to account for variation across polls. With a model that considers the effects of methodology, vote population, states, and poll quality along with seasonal indexes, more precise and less biased estimates of the percentage of support for candidates during a given time period could be made. As the findings of this analysis point to a very close race that can be determined by the swing states, a next step for the model could be to forecast the percentage of support for each candidate within each swing state to present a more detailed prediction of the election. Lastly, it may also be helpful to adopt a more sophisticated method to pool the polls, such as pooling the polls by considering the margin of error of each poll (Jackman 2005). A more enhanced method may allow for better precision, and in turn, provide a more reliable way to forecast.

A Appendix

A.1 Pollster Methodology

We selected YouGov, a polling organization, and discussed its survey methodology and its main features, strengths, and weaknesses. From the data obtained, the population of YouGov surveys is American voters, especially citizens who are eligible to vote. YouGov's framework is usually participants who voluntarily register and participate in surveys regularly. These panel members express their opinions in the form of online questionnaires. The sample is a part of YouGov's online panel. In the data, we see some specific stratification information, such as political parties (DEM, REP, etc.), which indicates that YouGov may use stratified sampling to ensure the diversity of the sample. YouGov's sample is recruited through a voluntary online panel. Users can actively register to become panel members and accept survey invitations at any time. This recruitment method is non-random, but the cost is relatively low and the number of people is large.

YouGov uses stratified sampling, which stratifies respondents according to demographic variables, such as age, gender, political party, etc., to ensure that each subgroup in the sample is fully represented. Stratified sampling can ensure that different groups (such as party supporters, different age groups, etc.) are properly represented, avoid a single group dominating, improve estimation accuracy, and reduce sampling errors. However, if the stratification criteria are not properly chosen or there are large individual differences within the strata, stratified sampling may increase sampling errors. It may be more time-consuming to design and implement than simple random sampling, and more resources are required to determine the stratification and sampling scheme, especially when the population size is large.

YouGov uses weighting to adjust when dealing with non-response issues. When some people do not respond or the response rate of certain groups is low, YouGov will weight the responses of these groups according to demographic data to ensure that the final survey results can more accurately reflect the overall situation. This can help correct the bias caused by the low response rate of certain groups and make the results more representative. YouGov's questionnaire is answered online, which can quickly obtain a large amount of data, and the population (people who answer the questionnaire) is distributed in various places, which improves flexibility and efficiency. The questionnaire also covers a variety of candidates and political parties, and distinguishes the support rates of different political parties. The content of the questionnaire can be changed according to different groups to ensure that the survey questions are relevant to the background of the respondents. At the same time, there are some potential problems with the questionnaire. First, since the questionnaire is answered online, some people may not answer the questionnaire seriously, which may affect the accuracy of the questionnaire. Secondly, since it is voluntary to participate in the questionnaire, some groups may be under-represented or over-represented, which will also lead to biased survey results.

A.2 Idealized Pollster Methodology

Survey Form Link: https://forms.gle/S4cyiZNej46zfxq29

A.2.1 Introduction

In this appendix, I present a detailed survey methodology designed to predict the outcome of the upcoming U.S. presidential election. The design leverages a $100K budget and focuses on achieving a representative, accurate, and methodologically robust sample. The survey will use mixed-mode recruitment (in person, phone, online, and SMS), with a sample size of 3,000 respondents. Detailed weighting adjustments and validation strategies will ensure the integrity of the data, while aggregation with other polls will provide a more accurate forecast.

A.2.2 Sampling Approach

A.2.2.1 Target Population

The survey targets the U.S. voting-age population, defined as U.S. citizens aged 18 and older. The target population includes both already voted voters and those who plan to vote.

A.2.2.2 Sample Size and Confidence

As for sample size, a total of 3,000 respondents will be surveyed. This provides a margin of error of $\pm 2\%$ at a 95% confidence level, ensuring reliable predictions at the national level. As for the confidence, the larger the sample size, the smaller the margin of error. Given the $100K budget, this is an optimal balance between cost and statistical reliability.

A.2.2.3 Stratified Sampling Approach

To ensure representatives, the sample will be stratified across key demographic factors like their age group, gender, race, education level, job type, income level, house situation, political party, living state and so on. Specifically, for those Swing states (such as Nevada, Arizona, Wisconsin, Michigan, Pennsylvania, North Carolina, and Georgia) will be over-sampled to ensure an accurate prediction in these battleground regions, where small shifts in voter behavior can heavily influence the election outcome. For example, instead of targeting only 8% of the sample in swing states (proportional to the population), we might over-sample to 20%.

A.2.2.4 Bias

Firstly, relying too much on online survey may skew results towards who are younger and more internet-savvy respondents. To solve this, a portion of the budget will be dedicated to reaching older and rural voters via phone surveys and in person survey. Besides, We will solve non-response bias by offering incentives.

A.2.3 Recruitment Strategy

A.2.3.1 Mixed-Mode Recruitment

The recruitment strategy uses a mix of recruitment channels to ensure diverse participation across demographic groups.

1. Phone Recruitment (Random Digit Dialing - RDD) or in person survey: Budget: $30,000 Goal: Target older voters, particularly those 65+ and rural populations, who are less likely to respond to online surveys. Method: RDD will include both land-lines and mobile numbers to maximize reach, especially among older voters. Response Rate: Assuming a 10-15% response rate, we expect to recruit around 1,000 respondents via phone interviews.
2. Online Panel Recruitment: Budget: $40,000 Goal: Capture younger, more tech-savvy respondents (ages 18-44) and urban populations who are more likely to participate in online surveys. Method: Use reputable online panels such as YouGov or Ipsos. These panels provide access to a large pool of respondents pre-screened for voter eligibility. Response Rate: With a budget of $20,000, we expect to recruit about 1,500 respondents from these panels.
3. Text-to-Web Invitations (SMS Surveys): Budget: $15,000 Goal: Reach respondents through mobile-friendly surveys, targeting younger voters (18-34) and those who prefer mobile interaction. Method: Send SMS invitations with a link to the online survey (via Google Forms), targeting respondents in both urban and rural areas. Response Rate: We expect to recruit 500 respondents via SMS links.

A.2.3.2 Incentives for Participation

To improve response rates, we will offer $5 digital gift cards as an incentive to complete the survey. Budget: $15,000 for approximately 3,000 respondents. This will particularly help increase participation among hard-to-reach groups, such as low-income individuals and rural populations.

A.2.4 Data Validation

A.2.4.1 Weighting Adjustments

Post-stratification weighting will be used to adjust the sample to reflect the actual U.S. voting population. This ensures that underrepresented groups (e.g., younger voters, minorities) are appropriately represented in the final analysis. Weights will be calculated based on age, race, gender, education, and some other factors by using Census data as a benchmark.

A.2.4.2 Screening Questions

The survey will include key screening questions to ensure eligibility:

"Are you a U.S. citizen eligible to vote in the ongoing U.S. presidential election?" "Are you at least 18 years old, making you eligible to vote in the ongoing U.S. presidential election?" "Have you already voted in the 2024 U.S. presidential election?" Respondents who do not meet these criteria will be excluded from the analysis.

A.2.4.3 Fraud Detection

To ensure high-quality responses:

1. Consistency Checks: Use validation questions to ensure the consistency of answers. For example, responses on party affiliation and voting intention will be cross-checked to identify inconsistencies.
2. Re-contacting: Randomly re-contact a subset of respondents to verify their initial responses.

References

FiveThirtyEight. 2024. "538 - Election Polls, Polictics, and Analysis." https://projects.fivethirtyeight.com/polls/.

Jackman, Simon. 2005. "Pooling the polls over an election campaign." *Australian Journal of Political Science* 40 (4): 499--517. https://doi.org/10.1080/10361140500302472.

Müller, Kirill, and Jennifer Bryan. 2020. *here: A Simpler Way to Find Your Files*. https://here.r-lib.org/.

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

Radcliffe, Mary, and G. Elliott Morris. 2023. "538's polls policy and FAQs." https://abcnews.go.com/538/538s-polls-policy-faqs/story?id=104489193.

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

Wickham, Hadley. 2011. *testthat: Get Started with Testing*. *The R Journal*. Vol. 3. https://journal.r-project.org/archive/2011-1/RJournal_2011-1_Wickham.pdf.

———. 2016. "ggplot2: Elegant Graphics for Data Analysis." https://ggplot2.tidyverse.org.

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.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. "dplyr: A Grammar of Data Manipulation." https://cran.r-project.org/web/packages/dplyr/index.html.

Xie, Yihui. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2019. "TinyTeX: A Lightweight, Cross-Platform, and Easy-to-Maintain LaTeX Distribution Based on TeX Live." *TUGboat* 40 (1): 30–32. https://tug.org/TUGboat/Contents/contents40-1.html.