Forecasting the 2024 US Presidential Election*

Tianning He Julia Lee Shuangyuan Yang

November 4, 2024

1 Introduction

A paragraph for the broader context and motivation for the analysis + the gap that we want to address

A paragraph for that details what the analysis aims to do (i.e. its objective) + how

A paragraph for what was found and why findings are important (i.e. their implications)

A paragraph for setting up the rest of the paper

2 Data

2.1 Presidential General Election Polls Data

A detailed description of the poll data (e.g. what the data shows, its variables, the date in which the data was downloaded + tables/graphs showing what the data looks like)

To simulate, test, download, and clean the Presidential General Election Polls data, 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), and `knitr` (Xie 2015).

(**pollster methodology and idealized methodology sections can be found in the Appendix - Section A**)

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

2.2 Analysis Data

A detailed description of the clean poll data that was used within the analysis (e.g. what the data shows, why certain variables were chosen + tables/graphs showing what the data looks like)

2.3 An Account on Measurement

A investigation into how voters' opinions or stances were translated into percentage of votes (i.e. a value) within the data

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 (CITE) 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 show strong intentions to vote on election day (CITE). 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 1. 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 1: Weeks Defined by the Model

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

Table 1: Weeks Defined by the Model

Week	Dates
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 1. An example outcome is shown below for Kamala Harris (Table 2).

Table 2: 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 2) 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 2, and then divide 1000, 619, 661, and 693 from the first four rows of Table 2 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 2, 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 2.

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 3).

Table 3: 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 3 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 4.

Table 4: 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 5.

Table 5: 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 5, 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 Percentage of Votes over time across Pollsters

This initial analysis aims to see how the percentage of votes for the Democratic Presidential candidate, Harris, has changed over time for various high-quality pollsters. (*a better, more detailed description will be added*)

5 Discussion

5.1 Summary

A paragraph that summarizes what was done in the analysis and a brief overview of the main findings

5.2 Implications

A paragraph about what the main findings imply about the election - why they are relevant

5.3 Limitations

A paragraph about the limitations of the model - shortcomings due to the decisions that were made in the model

5.4 Future Directions

A paragraph about what can be done in the future

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

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

Wickham, Hadley. 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.