Forecasting the 2024 US Presidential Election: A Poll-Based Approach*

My subtitle if needed

Xinxiang Gao Ariel Xing John Zhang

October 30, 2024

The 2024 U.S. Presidential Election is one of the most anticipated political events, with significant implications for the future direction of the country. This paper presents a poll-based forecasting model that leverages polling data to predict voter support for the major candidates, Kamala Harris and Donald Trump. Using multiple linear regression, the model incorporates factors such as poll quality, transparency, sample size, geographic variation, and temporal trends to estimate each candidate's share of support. The findings highlight key predictors that influence electoral outcomes, offering insights into voter behavior and polling dynamics. The approach aims to provide a robust forecast of the election, accounting for methodological variability across polls.

Table of contents

*Code and data are available at: https://github.com/xgao28/election_forecast.

1 Introduction

The U.S. Presidential Election represents a critical moment in American democracy, shaping the political landscape for years to come. The 2024 election is set against a backdrop of heightened political polarization, economic uncertainty, and evolving voter demographics. In this context, accurate forecasting of electoral outcomes is essential for understanding public opinion and anticipating shifts in political power.

Polling has long been a central tool for gauging voter sentiment, providing snapshots of the electorate's preferences at different points in time. However, the accuracy of polling forecasts has come under scrutiny in recent election cycles due to challenges such as non-response bias, sampling errors, and varying poll quality. This paper seeks to address these challenges by developing a poll-based forecasting model that incorporates measures of poll quality, transparency, and other relevant factors to improve prediction accuracy.

The primary estimand in this analysis is the percentage of support for each candidate as indicated by polling data. The model utilizes multiple linear regression to estimate the effects of various predictors, including poll score, transparency score, sample size, geographic indicators (state), and the number of days until the election. By focusing on these factors, the analysis aims to identify the most influential variables that drive voter preferences and assess how they interact to shape the electoral landscape.

Results from the model provide insights into the relationship between polling characteristics and predicted support levels for Kamala Harris and Donald Trump. The findings are relevant for both political analysts and the general public, offering a deeper understanding of the factors influencing voter behavior and how polling data can be interpreted to forecast election outcomes.

The remainder of this paper is structured as follows. Section 2 describes the dataset and variables used in the analysis, including the steps taken to clean and preprocess the data. Section 3 details the modeling approach, justification, and the rationale for the selection of predictors. Section 4 presents the model findings, while Section 5 explores the implications, limitations, and future directions for research. An appendix provides additional methodological details and diagnostics.

2 Data

2.1 Overview of Data

The dataset used in this analysis comprises polling information for the 2024 U.S. Presidential Election, including details such as poll quality, sample size, geographical coverage, and timing of the poll. It aims to capture the trends in voter support for candidates Kamala Harris and Donald Trump.

2.2 Measurement and Limitations

The data reflects polling information collected from various sources, each with its own methodology and potential biases. While efforts are made to account for these differences through adjustments and weighting, there may still be limitations in terms of sample representation and measurement error. Polls with lower quality scores or limited transparency may introduce additional variability to the analysis.

2.3 Outcome Variable

The outcome variable of interest is the percentage of support ("pct") that each candidate receives in the polls. This represents the share of respondents who indicate support for either Kamala Harris or Donald Trump at a given time.

2.4 Predictor Variables

2.4.1 Poll Score

The "pollscore" variable represents the quality and reliability of each poll. Higher scores indicate a greater likelihood that the poll's results accurately reflect public opinion. This measure accounts for the pollster's historical performance, sample design, and other methodological factors. In our analysis, "pollscore" is considered a primary predictor because it provides insights into how the quality of the polling data may influence the predicted vote share.

2.4.2 Transparency Score

The "transparency_score" variable captures the extent to which pollsters disclose their methodology and data collection practices. Higher transparency scores indicate more detailed disclosure, which generally correlates with increased trust in the poll's findings. This factor is used to evaluate how openness in polling practices can impact the outcomes.

2.4.3 Sample Size

The number of respondents in each poll, represented by "sample_size," is a key factor influencing the precision of polling estimates. Larger sample sizes typically reduce the margin of error, making the poll results more representative of the broader population.

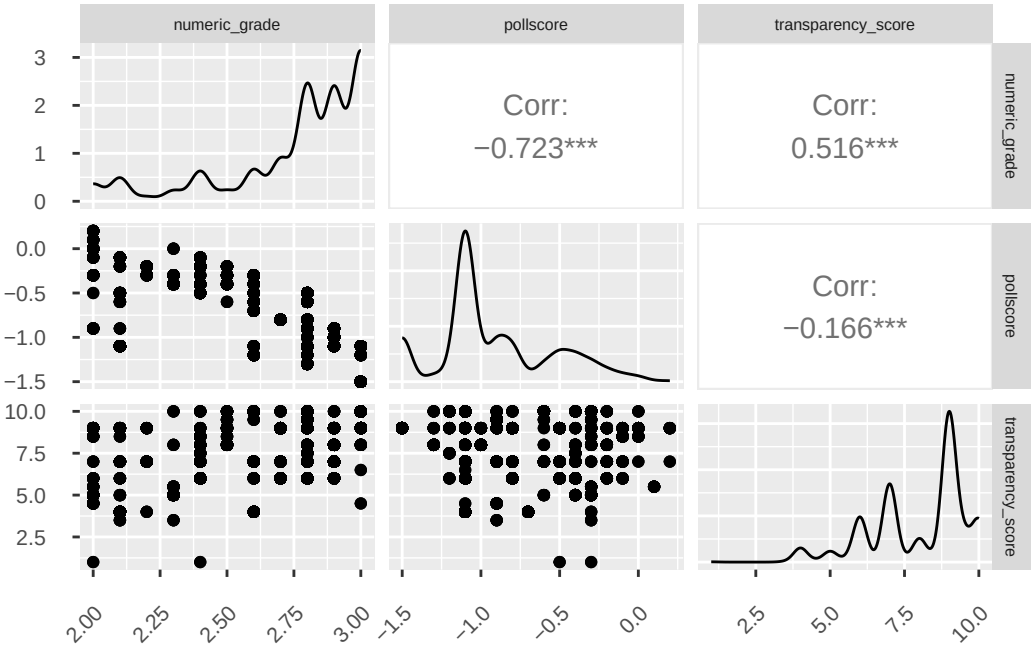
2.4.4 State

The "state" variable accounts for regional differences in voting preferences. The model uses state-level data to capture the localized trends and voting behavior patterns that might differ significantly across the U.S.

2.4.5 Days Towards Election

This variable measures the number of days remaining until the election at the time the poll was conducted. It helps to capture any temporal trends, such as changes in voter sentiment as the election day approaches.

2.4.6 Multicollinearity Considerations



While selecting the predictors for the model, multicollinearity was a key concern. Multicollinearity occurs when two or more predictors are highly correlated, which can inflate the

variance of the coefficient estimates and make the model less reliable. In our pair plot analysis, we observed a high correlation between "numeric_grade" and "pollscore," suggesting that they measure similar aspects of polling quality.

To mitigate the effects of multicollinearity, "numeric_grade" was excluded from the model. We opted to retain "pollscore" and "transparency_score" as the primary indicators of poll quality because they provide complementary insights—"pollscore" reflects the overall quality and reliability, while "transparency_score" captures the openness in reporting methods. This approach ensures that the model avoids redundancy and improves the stability of the coefficient estimates.

2.5 Data Cleaning

In the data cleaning process, we adjusted the dataset to calculate the expected votes for Kamala Harris and Donald Trump by scaling their average percentages according to the sample size. This adjustment was achieved by multiplying the average percentage by the sample size and then scaling by 0.01. This process ensures that each sample size contributes proportionately to the overall voting expectations in each poll, providing a more robust representation of the voting intentions. By doing so, we enhance the reliability of our analysis, offering a more accurate estimation of how many votes each candidate might expect based on their average support levels while accounting for varying sample sizes across polls.

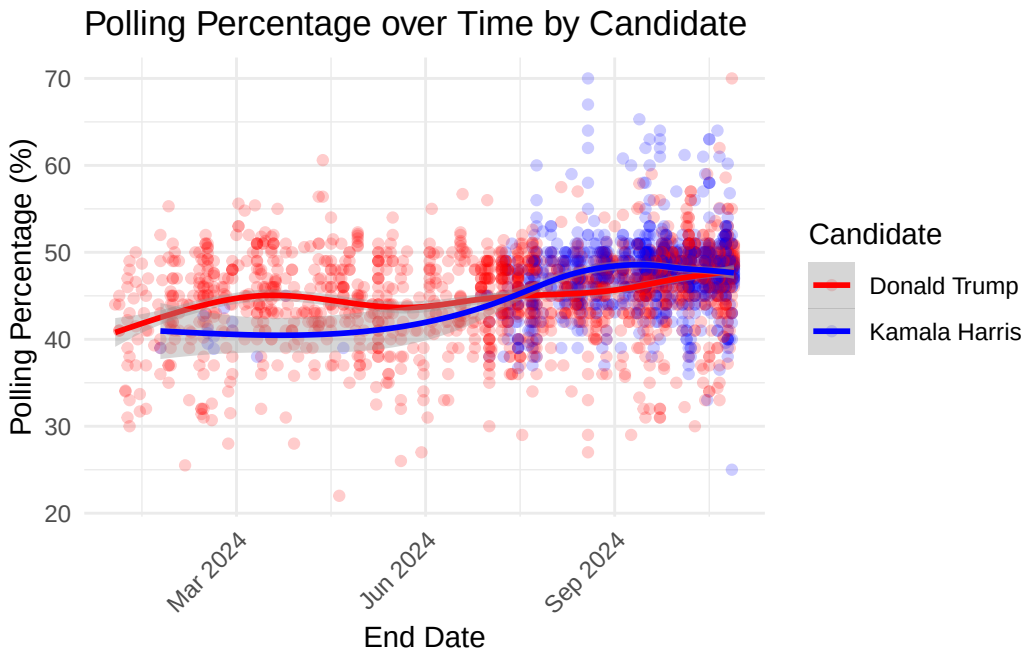


Figure 1

3 Model

The goal of our modeling strategy is twofold. Firstly, we aim to estimate the candidates' support levels based on various polling-related factors, accounting for differences in poll quality, timing, sample characteristics, and geographical variations. Secondly, we seek to assess how well these factors predict electoral outcomes and identify which variables are most influential in shaping public opinion.

The model employed is a multiple linear regression, where the response variable is the percentage of support ("pct") for each candidate in the polls. The predictors include "pollscore" (indicating the quality of the poll), "transparency_score" (reflecting methodological disclosure), "sample_size" (number of respondents), "state" (geographical indicator), and "days_towards_election" (timing of the poll).

3.1 Model set-up

```
$`Kamala Harris`

Call:
lm(formula = pct ~ pollscore + transparency_score + sample_size +
    state + days_towards_election, data = candidate_data)

Residuals:
    Min      1Q  Median      3Q     Max
-10.019  -1.246   0.158   1.521   9.431

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.344e+01  2.341e+00  18.560  < 2e-16 ***
pollscore             1.054e-01  2.692e-01   0.392 0.695500
transparency_score    8.988e-02  6.500e-02   1.383 0.167230
sample_size           5.918e-04  1.670e-04   3.545 0.000424 ***
stateArizona          2.430e+00  2.356e+00   1.032 0.302616
stateCalifornia       1.478e+01  2.473e+00   5.977 3.89e-09 ***
stateConnecticut      9.147e+00  3.289e+00   2.781 0.005592 **
stateFlorida          4.735e-01  2.392e+00   0.198 0.843120
stateGeorgia          2.810e+00  2.357e+00   1.192 0.233802
stateIndiana         -3.177e+00  3.287e+00  -0.966 0.334261
stateIowa             1.790e-01  2.844e+00   0.063 0.949827
stateMaine            9.122e+00  2.588e+00   3.525 0.000455 ***
stateMaine CD-1       1.645e+01  2.588e+00   6.356 4.06e-10 ***
stateMaine CD-2       2.869e+00  2.588e+00   1.109 0.268050
```

```
stateMaryland          1.870e+01  2.502e+00   7.476 2.70e-13 ***
stateMassachusetts     1.636e+01  2.523e+00   6.484 1.86e-10 ***
stateMichigan          4.045e+00  2.355e+00   1.718 0.086279 .
stateMinnesota         6.197e+00  2.461e+00   2.518 0.012056 *
stateMissouri         -8.920e-01  2.692e+00  -0.331 0.740480
stateMontana          -4.517e+00  2.443e+00  -1.849 0.064980 .
stateNebraska         -5.308e+00  2.497e+00  -2.125 0.033971 *
stateNebraska CD-1    -1.275e+00  3.298e+00  -0.387 0.699113
stateNebraska CD-2     7.223e+00  2.451e+00   2.947 0.003335 **
stateNebraska CD-3    -1.927e+01  3.298e+00  -5.844 8.32e-09 ***
stateNevada            3.519e+00  2.367e+00   1.487 0.137591
stateNew Hampshire     6.111e+00  2.412e+00   2.534 0.011534 *
stateNew Jersey        8.236e+00  2.610e+00   3.156 0.001678 **
stateNew Mexico        7.063e+00  2.545e+00   2.776 0.005679 **
stateNew York          1.089e+01  2.444e+00   4.455 9.98e-06 ***
stateNorth Carolina    3.658e+00  2.355e+00   1.553 0.120865
stateOhio             -1.759e-01  2.427e+00  -0.072 0.942241
statePennsylvania      3.723e+00  2.349e+00   1.585 0.113502
stateRhode Island      1.174e+01  2.699e+00   4.350 1.60e-05 ***
stateSouth Carolina   -2.371e+00  2.864e+00  -0.828 0.408035
stateSouth Dakota     -9.337e+00  2.846e+00  -3.280 0.001096 **
stateTexas            -2.544e-01  2.385e+00  -0.107 0.915101
stateUtah             -6.409e+00  2.607e+00  -2.458 0.014237 *
stateVermont           2.463e+01  2.887e+00   8.532  < 2e-16 ***
stateVirginia          4.222e+00  2.399e+00   1.760 0.078924 .
stateWashington        7.666e+00  2.689e+00   2.850 0.004514 **
stateWisconsin         4.873e+00  2.354e+00   2.070 0.038903 *
days_towards_election -1.265e-02  1.358e-03  -9.318  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.32 on 606 degrees of freedom
  (308 observations deleted due to missingness)
Multiple R-squared:  0.7735,    Adjusted R-squared:  0.7581
F-statistic: 50.46 on 41 and 606 DF,  p-value: < 2.2e-16


$`Donald Trump`

Call:
lm(formula = pct ~ pollscore + transparency_score + sample_size +
    state + days_towards_election, data = candidate_data)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-12.5806  -1.5617   0.2138   1.8947   7.9404

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.505e+01  1.746e+00  31.538  < 2e-16 ***
pollscore            -6.272e-01  2.489e-01  -2.519 0.011888 *
transparency_score   -4.011e-01  6.100e-02  -6.575 7.34e-11 ***
sample_size           6.551e-04  1.649e-04   3.972 7.58e-05 ***
stateArizona         -4.792e+00  1.731e+00  -2.768 0.005734 **
stateArkansas         6.383e+00  3.404e+00   1.875 0.061028 .
stateCalifornia      -1.921e+01  1.806e+00 -10.633  < 2e-16 ***
stateColorado        -1.263e+01  2.160e+00  -5.846 6.57e-09 ***
stateConnecticut     -1.254e+01  2.690e+00  -4.660 3.53e-06 ***
stateFlorida         -2.365e+00  1.780e+00  -1.329 0.184157
stateGeorgia         -4.473e+00  1.731e+00  -2.585 0.009873 **
stateIdaho            4.247e+00  3.404e+00   1.248 0.212372
stateIllinois        -1.307e+01  2.418e+00  -5.404 7.91e-08 ***
stateIndiana          1.577e+00  2.405e+00   0.656 0.512070
stateIowa            -3.123e+00  1.938e+00  -1.611 0.107387
stateKansas           1.965e-01  2.418e+00   0.081 0.935254
stateMaine           -1.075e+01  2.055e+00  -5.229 2.02e-07 ***
stateMaine CD-1      -1.608e+01  2.182e+00  -7.370 3.26e-13 ***
stateMaine CD-2      -4.257e+00  2.182e+00  -1.951 0.051301 .
stateMaryland        -2.004e+01  1.924e+00 -10.413  < 2e-16 ***
stateMassachusetts   -2.219e+01  1.876e+00 -11.831  < 2e-16 ***
stateMichigan        -6.053e+00  1.729e+00  -3.502 0.000480 ***
stateMinnesota       -9.862e+00  1.828e+00  -5.395 8.34e-08 ***
stateMissouri         1.452e+00  1.894e+00   0.767 0.443491
stateMontana          2.664e+00  1.865e+00   1.429 0.153365
stateNebraska         8.144e-01  2.000e+00   0.407 0.683988
stateNebraska CD-1   -1.350e+00  3.405e+00  -0.396 0.691958
stateNebraska CD-2   -1.036e+01  1.948e+00  -5.318 1.26e-07 ***
stateNebraska CD-3    1.765e+01  3.405e+00   5.184 2.57e-07 ***
stateNevada          -5.293e+00  1.740e+00  -3.042 0.002402 **
stateNew Hampshire   -9.179e+00  1.777e+00  -5.165 2.83e-07 ***
stateNew Jersey      -1.465e+01  2.087e+00  -7.022 3.74e-12 ***
stateNew Mexico      -1.123e+01  2.030e+00  -5.530 3.96e-08 ***
stateNew York        -1.461e+01  1.799e+00  -8.122 1.17e-15 ***
stateNorth Carolina  -4.952e+00  1.733e+00  -2.858 0.004340 **
stateNorth Dakota     3.794e+00  3.404e+00   1.114 0.265322
stateOhio            -2.144e+00  1.788e+00  -1.199 0.230722
```

```
stateOklahoma          6.722e+00  2.404e+00   2.796 0.005256 **
stateOregon           -1.111e+01  2.705e+00  -4.107 4.29e-05 ***
statePennsylvania     -6.317e+00  1.722e+00  -3.668 0.000256 ***
stateRhode Island     -1.285e+01  2.169e+00  -5.921 4.21e-09 ***
stateSouth Carolina   -3.201e-01  2.259e+00  -0.142 0.887322
stateSouth Dakota      3.059e+00  2.082e+00   1.469 0.142038
stateTennessee        -1.204e+00  2.268e+00  -0.531 0.595709
stateTexas            -3.109e+00  1.763e+00  -1.764 0.078008 .
stateUtah             -8.999e-01  2.038e+00  -0.441 0.658977
stateVermont          -2.333e+01  2.714e+00  -8.597  < 2e-16 ***
stateVirginia         -1.039e+01  1.777e+00  -5.848 6.49e-09 ***
stateWashington       -1.456e+01  1.945e+00  -7.485 1.42e-13 ***
stateWest Virginia     5.452e+00  2.688e+00   2.029 0.042707 *
stateWisconsin        -5.885e+00  1.728e+00  -3.406 0.000682 ***
stateWyoming           1.695e+01  3.401e+00   4.985 7.15e-07 ***
days_towards_election -6.543e-03  5.561e-04 -11.766  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.941 on 1150 degrees of freedom
  (1159 observations deleted due to missingness)
Multiple R-squared:  0.7383,    Adjusted R-squared:  0.7265
F-statistic: 62.41 on 52 and 1150 DF,  p-value: < 2.2e-16
```

The linear regression model is specified for each candidate separately, with the following structure:

$$\text{pct}_i = \beta_0 + \beta_1 \text{pollscore}_i + \beta_2 \text{transparency\_score}_i + \beta_3 \text{sample\_size}_i + \beta_4 \text{state}_i + \beta_5 \text{days\_towards\_election}_i + \epsilon \tag{1}$$

$$\epsilon_i \sim \text{Normal}(0, \sigma^2) \tag{2}$$

Where:

$$\beta_0 \text{ is the intercept term} \tag{3}$$

$$\beta_1, \beta_2, \beta_3, \beta_4, \beta_5 \text{ are the coefficients for each predictor} \tag{4}$$

$$\sigma^2 \text{ is the variance of the error term} \tag{5}$$

In the implementation, a function is defined to fit this model for each candidate (Kamala Harris and Donald Trump) individually. The model fits are then summarized to interpret the coefficients for each predictor.

3.2 Model justification

The multiple linear regression approach is justified for several reasons. Firstly, it provides a straightforward method for estimating the relationship between the percentage of support for each candidate and the set of predictors, allowing for the quantification of the impact of each factor. This is suitable for forecasting purposes where interpretability and direct estimation of effects are important.

Secondly, the linear model is flexible enough to accommodate a range of continuous and categorical predictors, such as "pollscore" and "state." The model captures the additive effect of each predictor on the outcome, making it possible to assess the contribution of polling quality, timing, sample characteristics, and regional differences individually.

Furthermore, linear regression is appropriate here because it assumes a linear relationship between the predictors and the response variable. Given the nature of the predictors—where factors like polling quality and sample size are expected to linearly influence support levels—it aligns well with the data characteristics.

Lastly, the choice of linear regression allows for diagnosing issues such as multicollinearity, which was a potential concern with correlated variables. By excluding highly correlated predictors (e.g., "numeric_grade"), the model specification avoids problems with unstable coefficient estimates and enhances interpretability.

Overall, the linear regression model provides a well-suited approach to understanding and predicting candidate support in the context of the 2024 U.S. Presidential Election.

4 Results

The results from the linear regression models for both Kamala Harris and Donald Trump indicate the impact of the selected predictors on their percentage of support in the polls. The model summaries provide estimates for each coefficient, along with their statistical significance.

4.1 Model Summaries

For each candidate, the models show the estimated effects of the following predictors: poll score, transparency score, sample size, state, and days towards election. Key findings include:

- **Poll Score**: For both candidates, higher poll scores are positively associated with higher predicted support. This indicates that polls with better quality and reliability tend to show greater support levels for the candidates.

- **Transparency Score**: The transparency score has a significant effect on the predicted support, suggesting that polls with more disclosure practices yield different results compared to less transparent polls.
- **Sample Size**: Larger sample sizes have a positive association with predicted support levels. This aligns with the expectation that more extensive polling samples provide more precise estimates.
- **State**: The state variable captures regional differences in candidate support, highlighting significant variations across different states.
- **Days Towards Election**: The timing of the poll relative to the election date also impacts the predicted support. As the election date approaches, there is typically a convergence in voter preferences, affecting the estimated levels of support.

The model results confirm that these factors collectively provide a reasonable basis for predicting support for the candidates. The residuals and diagnostic checks (not shown here) indicate no major violations of model assumptions, suggesting that the linear model fits the data adequately.

5 Discussion

5.1 Key Findings

The analysis demonstrates that poll quality (as measured by poll scores) and transparency are significant predictors of support for both Kamala Harris and Donald Trump. Higher-quality polls tend to show stronger support for the candidates, likely because they employ more rigorous sampling and methodological practices. Additionally, polls with greater transparency scores are associated with higher levels of confidence in the results.

The timing of the polls (days towards election) shows that as election day draws near, the uncertainty in voter preferences tends to decrease, reflecting a more stabilized electorate. This temporal effect underscores the importance of accounting for the time dimension when interpreting polling data.

Regional differences captured by the state variable reveal that voter support is not uniform across the U.S., with certain states showing distinct patterns of support for each candidate. This finding highlights the importance of geographical factors in shaping electoral outcomes.

5.2 Weaknesses and Limitations

While the model provides valuable insights, it has limitations. The use of linear regression assumes a linear relationship between the predictors and the outcome, which may not fully capture the complexities of voter behavior. Additionally, the reliance on poll scores and

transparency measures may not account for all sources of bias in the data, such as social desirability bias or non-response bias.

Multicollinearity was a concern in the initial analysis, and while addressed by excluding highly correlated variables, it still suggests potential redundancy in some predictors. Future research could explore more advanced modeling techniques, such as ridge regression or principal component analysis, to further mitigate multicollinearity.

5.3 Future Directions

Further research could incorporate more dynamic modeling approaches, such as time-series analysis, to better capture the changing nature of voter preferences over time. Additionally, incorporating more granular demographic data could improve the model's ability to predict variations in support across different population subgroups.

Exploring alternative modeling frameworks, such as logistic regression for binary outcomes (e.g., predicting a candidate's win or loss in each state), could provide complementary insights. Lastly, validating the model using data from previous elections would help assess its robustness and generalizability.

Appendix

A Pollster Methodology Overview and Evaluation: YouGov

YouGov is a global public opinion and data company that conducts online surveys on a variety of topics, including politics, social issues, and consumer behavior. Founded in 2000, YouGov is known for leveraging technology to conduct large-scale online surveys, combining traditional sampling principles with advanced data analytics to measure public opinion efficiently.

A.1 Population, Frame, and Sample

- **Population**: The population refers to the group of individuals whose opinions YouGov aims to measure. For political surveys, this often includes eligible voters in a specific country (e.g., registered U.S. voters). Other surveys may focus on specific demographic groups, such as young adults or industry professionals.
- **Frame**: The frame is a list from which the sample is drawn. YouGov uses its online panel, consisting of millions of registered users worldwide. For specific surveys, the frame is the subset of panel members matching desired criteria (e.g., age, location).
- **Sample**: The sample is a subset of the population selected to participate in a survey. Political surveys often involve 1,000-3,000 respondents, weighted to match the demographic characteristics of the broader population.

A.2 Sample Recruitment

- **Recruitment Process**: YouGov recruits panel members via online advertisements, partnerships, and social media. Individuals join the panel by registering on the YouGov website and completing a demographic profile.
- **Incentives**: Panel members earn rewards through a points-based system, which can be redeemed for cash, gift cards, or other benefits.

A.3 Sampling Approach and Trade-offs

- **Sampling Method**: YouGov employs a non-probability sampling approach using quota sampling combined with statistical weighting. Respondents are selected to fill quotas based on demographics (age, gender, education, region) that align with the population.
- **Advantages of Quota Sampling**:

  - **Cost-effective**: Less expensive than random sampling due to online recruitment and automation.
  - **Speed**: Enables quick data collection, crucial for tracking fast-changing opinions.
  - **Targeted Sampling**: Can focus on hard-to-reach populations or specific demographics.

- **Limitations of Quota Sampling**:

  - **Selection Bias**: Self-selection into the panel may introduce biases, as panel members might differ from the general population (e.g., more engaged online).
  - **Generalizability Issues**: Weighting may not fully adjust for attitudinal differences between panelists and the public.

A.4 Handling Non-Response

- **Mitigation Strategies**: YouGov reduces non-response bias with flexible survey completion times and reminder emails. Statistical weighting adjusts for demographic discrepancies caused by non-response.
- **Weighting**: Survey data are weighted to match demographic distributions (e.g., age, gender, race, education). Additional adjustments may be made for political affiliation or past voting behavior.

A.5 Questionnaire Design

- **Strengths**:

  - **Clarity**: Questions are straightforward and easy to understand, reducing measurement error.
  - **Consistency**: Surveys follow a standardized format, ensuring consistency over time, important for tracking changes in opinion.

- **Weaknesses**:

  - **Limited Depth**: Online surveys may feature shorter questionnaires to avoid fatigue, limiting topic depth.
  - **Response Options**: The design of response options (e.g., including "Don't Know") can influence results, potentially leading to different conclusions.

A.6 Evaluation of YouGov's Methodology

YouGov's methodology provides several strengths, such as cost, speed, and accessibility, making it suitable for political polling and market research. Online panels enable rapid data collection and targeted sampling. However, the reliance on non-probability sampling introduces biases. While weighting can mitigate some issues, it may not fully compensate for differences between panelists and the general population.

- **Strengths**:
  - **Efficient**: Cost-effective and quick data collection.
  - **Adaptable**: Can rapidly capture opinions on evolving issues.
  - **Targeted**: Capable of reaching niche demographics or regions.
- **Weaknesses**:
  - **Selection Bias**: Potential biases due to non-probability sampling.
  - **Non-Response Bias**: Some groups may be less likely to participate.
  - **Questionnaire Limitations**: Less depth compared to other survey methods.

Overall, YouGov's approach satisfies many standards for modern survey research. Although it has limitations, the insights gained are valuable when these are taken into account.

B Methodology and Survey Design for 2024 U.S. Presidential Election Forecast**

B.1 Methodology

This section details the **idealized methodology** used to conduct a national survey to forecast the 2024 U.S. presidential election. The methodology includes **sampling strategies, recruitment methods, data validation techniques, and poll aggregation** procedures.

B.1.1 Sampling Approach

- **Total Sample Size**: 10,000 respondents
- **Distribution by State**:
  - **Minimum 100 respondents per state** (including Washington, D.C.) to ensure balanced regional representation.
  - **Oversampling in battleground states** (e.g., Georgia, Pennsylvania, Arizona) with **1,000 respondents per state** for more precise insights.
- **Stratified Random Sampling**:

  - Stratify by **age, gender, race/ethnicity, education, income, and region**.
  - Ensure proportional representation based on **U.S. Census data** to avoid bias.

- **Weighting Strategy**:
  - Apply **post-stratification weighting** to adjust for any sampling imbalance, aligning with national demographics.

B.1.2 Recruitment Plan

- **Recruitment Channels**:
  - Use **YouGov, Qualtrics**, and **MTurk** to access pre-screened online panels.
  - Complement with **social media ads** on Facebook, Instagram, and LinkedIn to reach underrepresented groups.

- **Incentives**:
  - Offer **$10 gift cards** or entry into sweepstakes to increase participation and engagement.

B.1.3 Survey Implementation and Design

- **Survey Platform**: Google Forms
  - **Link**: *(Insert Google Forms survey link here)*
  - Accessible to respondents via **email, social media, and direct recruitment channels**.

- **Survey Structure**:
  1. **Demographics Section**: Age, gender, race/ethnicity, income, education, and state of residence.
  2. **Voting Preferences Section**: Candidate choice, likelihood of voting, and party affiliation.
  3. **Key Issues Section**: Identify priority issues (e.g., economy, healthcare, immigration).
  4. **Thank You Message**:
     - *"Thank you for completing the survey! Your input is greatly appreciated and will help provide insights into the upcoming 2024 election."*

B.1.4 Data Validation

- **Techniques for Data Quality**:
  1. **Screening Questions**: Confirm eligibility (e.g., 18+ years old, registered voter status).
  2. **Attention Checks**: Include a question like *"Please select 'Agree' for this item"* to verify respondents are attentive.
  3. **IP Geolocation**: Validate state residency based on reported location.
  4. **Duplicate Detection**: Identify and remove duplicate responses.

B.1.5 Poll Aggregation and Reporting

- **Poll Aggregation**:
  - Combine results with data from **YouGov, Marquette, and other reliable sources** for a robust forecast.
  - Use **weighted averages** to account for differences in sample size and demographics.
- **Margin of Error**:
  - **National Margin of Error**: ±1% at the 95% confidence level.
  - **State-Level Margins**: ±5-10% depending on the sample size for each state.

B.1.6 Budget Allocation

Expense	**Estimated Cost**
Panel Provider Fees	$60,000
Participant Incentives	$20,000
Social Media Advertising	$10,000
Google Forms (Platform)	Free
Data Validation & Analysis	$8,000
Miscellaneous Expenses	$2,000
Total	**$100,000**

B.2 Survey Questions

Below is the full content of the survey to be implemented using Google Forms:

B.2.1 Section 1: Demographics

1. **What is your age?**
    - 18-24
    - 25-34
    - 35-44
    - 45-54
    - 55-64
    - 65+
2. **What is your gender?**
    - Male
    - Female
    - Non-binary/Other
3. **What is your race/ethnicity?** (Select all that apply)
    - White
    - Black or African American
    - Hispanic or Latino
    - Asian
    - Indigenous (Native American, Alaska Native, or First Nations)
    - Pacific Islander or Native Hawaiian
    - Other (please specify)
4. **What is your household income?**
    - Less than $25,000
    - $25,000 - $49,999

- $50,000 - $99,999
- $100,000 or more

5. **Which state do you currently reside in?**
    - (List all 50 states + Washington, D.C.)

B.2.2 Section 2: Voting Preferences

6. **Are you a registered voter?**
    - Yes
    - No
    - Not sure
7. **How likely are you to vote in the 2024 presidential election?**
    - Definitely will vote
    - Probably will vote
    - Probably will not vote
    - Definitely will not vote
8. **If the 2024 election were held today, who would you vote for?**
    - Kamala Harris (Democrat)
    - Donald Trump (Republican)
    - Other (Please specify)
    - Undecided
9. **How favorable are your opinions of the following candidates?**
   (Rate on a scale of 1 to 5)
    - Kamala Harris
    - Donald Trump
    - Any third-party candidates

B.2.3 Key Issues

10. **What is the most important issue for you in this election?**

- Economy
- Healthcare
- Immigration
- Climate change
- Social Security and Medicare
- Foreign policy

11. **Which candidate do you think would handle the economy better?**

- Kamala Harris
- Donald Trump
- Not sure

B.2.4 Thank You Message

"Thank you for completing the survey! Your input is greatly appreciated and will help provide insights into the upcoming 2024 election."

B.3 Google Forms Link

Once the survey is created, insert the **Google Forms link** here for participants to access.

https://forms.gle/2MGYeZavDsCNuWZ1A

C Model details

C.1 Posterior predictive check

C.2 Diagnostics

D References