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.

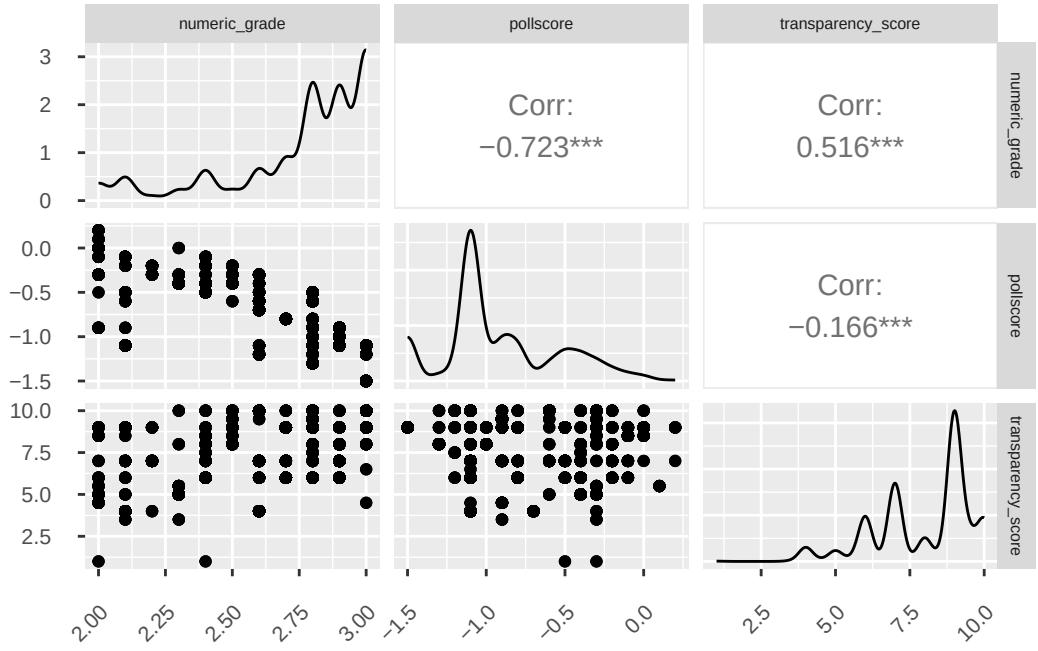
2.4.4 State

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.

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

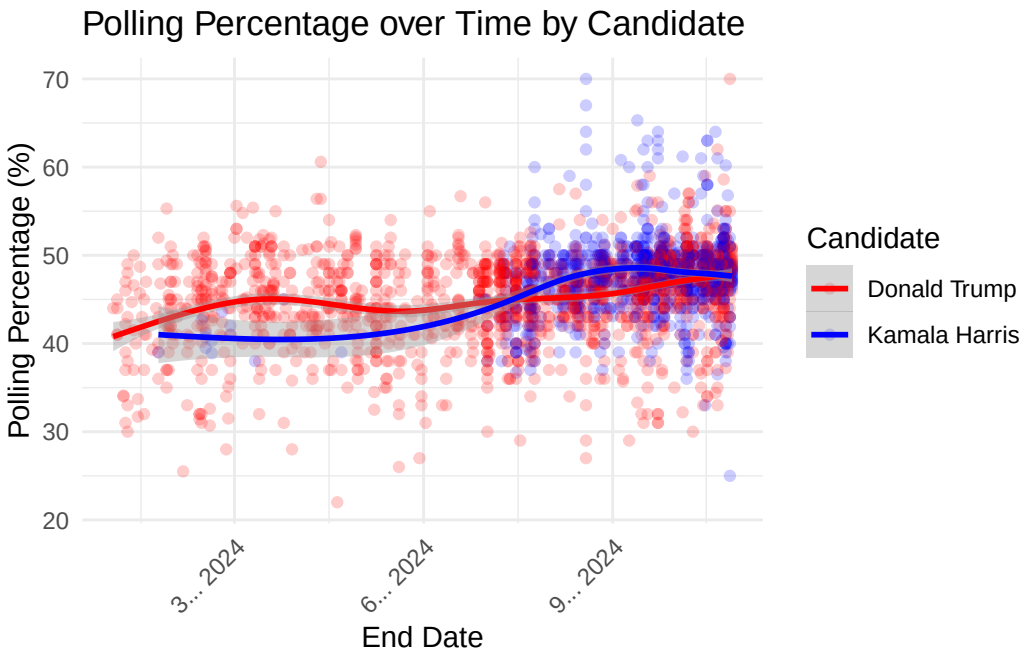


Figure 1

term	estimate	std.error	statistic	p.value	candidate
(Intercept)	48.0203353	0.1564742	306.889758	0	Kamala Harris
days_towards_election	-0.0061445	0.0009513	-6.459117	0	Kamala Harris
(Intercept)	45.9150331	0.1508145	304.447172	0	Donald Trump
days_towards_election	-0.0050979	0.0004077	-12.504363	0	Donald Trump

term	estimate	std.error	statistic	p.value	state	candidate
(Intercept)	48.8683887	0.4964579	98.434100	0.0000000	Nevada	Kamala Harris
days_towards_election	-0.0245932	0.0044407	-5.538186	0.0000041	Nevada	Kamala Harris
(Intercept)	47.1557040	0.4119580	114.467273	0.0000000	Arizona	Kamala Harris
days_towards_election	-0.0148567	0.0044182	-3.362581	0.0014397	Arizona	Kamala Harris
(Intercept)	49.3876638	0.2831917	174.396559	0.0000000	Wisconsin	Kamala Harris
days_towards_election	-0.0106782	0.0031918	-3.345491	0.0013152	Wisconsin	Kamala Harris

term	estimate	std.error	statistic	p.value	state	candidate
(Intercept)	48.3172446	0.3599062	134.249533	0.0000000	Michigan	Kamala Harris
days_towards_election	-0.0082338	0.0042324	-1.945408	0.0563383	Michigan	Kamala Harris
(Intercept)	48.8541682	0.3084645	158.378585	0.0000000	Pennsylvania	Kamala Harris
days_towards_election	-0.0203378	0.0036433	-5.582246	0.0000002	Pennsylvania	Kamala Harris
(Intercept)	48.9052297	0.3143316	155.584842	0.0000000	North Carolina	Kamala Harris
days_towards_election	-0.0257650	0.0054034	-4.768261	0.0000123	North Carolina	Kamala Harris
(Intercept)	47.6561673	0.4826396	98.740692	0.0000000	Georgia	Kamala Harris
days_towards_election	-0.0158840	0.0050312	-3.157076	0.0026291	Georgia	Kamala Harris
(Intercept)	47.3602317	0.4823642	98.183564	0.0000000	Nevada	Donald Trump
days_towards_election	-0.0039210	0.0015864	-2.471624	0.0157184	Nevada	Donald Trump
(Intercept)	48.7931474	0.4420177	110.387327	0.0000000	Arizona	Donald Trump
days_towards_election	-0.0091857	0.0016679	-5.507426	0.0000003	Arizona	Donald Trump
(Intercept)	47.2334451	0.3697600	127.740834	0.0000000	Wisconsin	Donald Trump
days_towards_election	-0.0076254	0.0017102	-4.458713	0.0000177	Wisconsin	Donald Trump
(Intercept)	47.2516714	0.3797487	124.428784	0.0000000	Michigan	Donald Trump
days_towards_election	-0.0078627	0.0016016	-4.909402	0.0000029	Michigan	Donald Trump
(Intercept)	47.0385962	0.3313679	141.952776	0.0000000	Pennsylvania	Donald Trump
days_towards_election	-0.0077522	0.0013854	-5.595815	0.0000001	Pennsylvania	Donald Trump
(Intercept)	47.8171166	0.2786974	171.573592	0.0000000	North Carolina	Donald Trump
days_towards_election	-0.0030779	0.0011512	-2.673558	0.0088500	North Carolina	Donald Trump

term	estimate	std.error	statistic	p.value	state	candidate
(Intercept)	48.3358474	0.3775705	128.018088	0.0000000	Georgia	Donald Trump
days_towards_election	-0.0038228	0.0013054	-2.928405	0.0041637	Georgia	Donald Trump

state	Kamala Harris	Donald Trump	difference
Nevada	48.86839	47.36023	1.5081570
Arizona	47.15570	48.79315	-1.6374434
Wisconsin	49.38766	47.23345	2.1542187
Michigan	48.31724	47.25167	1.0655732
Pennsylvania	48.85417	47.03860	1.8155720
North Carolina	48.90523	47.81712	1.0881130
Georgia	47.65617	48.33585	-0.6796801

Table 4: Predicted Electoral Votes for Each Candidate

Candidate	Solid_State_Votes	Predicted_Swing_State_Votes	Total_Predicted_Votes
Harris	226	66	292
Trump	219	27	246

3.1.1 Model Formulations

3.1.1.1 1. Baseline Model (Winner-Take-All National Model)

The baseline model aggregates national polling data to predict a winner, with all electoral votes awarded to the candidate with the higher national polling percentage.

1. **National Polling Projections**
   Let:

   - $P_H(t)$: Harris's national polling percentage at t days towards the election.
   - $P_T(t)$: Trump's national polling percentage at t days towards the election.

   We model each candidate's polling as:

$$P_H(t) = \alpha_H + \beta_H t$$

$$P_T(t) = \alpha_T + \beta_T t$$

   where α_H and α_T are the intercepts (predicted percentages on election day, $t = 0$), and β_H and β_T are the slopes.

2. **Election Day Prediction**
   On election day:
$$P_H(0) = \alpha_H, \quad P_T(0) = \alpha_T$$

3. **Winner-Take-All Outcome**
   The candidate with the higher predicted national percentage wins all electoral votes:
$$W = \begin{cases} \text{Harris wins (all electoral votes)} & \text{if } \alpha_H > \alpha_T \\ \text{Trump wins (all electoral votes)} & \text{otherwise} \end{cases}$$

3.1.1.2 2. Primary Model (Swing State-Based Model)

This model forecasts electoral outcomes by projecting state-level polling results for each candidate, combining swing state outcomes with votes from solid states.

1. **Swing State Polling Projections**
   For each swing state $s \in S$, let:

   - $P_{H_s}(t)$: Harris's polling percentage in state s at t days before the election.
   - $P_{T_s}(t)$: Trump's polling percentage in state s at t days before the election.

   We model each swing state polling as:
$$P_{H_s}(t) = \alpha_{H_s} + \beta_{H_s} t$$
$$P_{T_s}(t) = \alpha_{T_s} + \beta_{T_s} t$$

   where α_{H_s} and α_{T_s} are intercepts (predicted percentages at $t = 0$) and β_{H_s}, β_{T_s} are slopes.

2. **Predicting Swing State Winners**
   On election day ($t = 0$), each candidate's predicted support in state $s$is:
$$P_{H_s}(0) = \alpha_{H_s}, \quad P_{T_s}(0) = \alpha_{T_s}$$

   Define W_s as an indicator for the swing state winner:
$$W_s = \begin{cases} 1 & \text{if } \alpha_{H_s} > \alpha_{T_s} \\ 0 & \text{otherwise} \end{cases}$$

3. **Electoral Vote Aggregation**
   Let EV_H and EV_T represent total electoral votes for Harris and Trump:
$$EV_H = \sum_{s \in S} W_s \cdot EV_s + \sum_{s \in S_H} EV_s$$
$$EV_T = \sum_{s \in S} (1 - W_s) \cdot EV_s + \sum_{s \in S_T} EV_s$$

4. **Final Election Outcome**
   The candidate surpassing 270 electoral votes wins:

$$W = \begin{cases} \text{Harris wins} & \text{if } EV_H \geq 270 \\ \text{Trump wins} & \text{if } EV_T \geq 270 \end{cases}$$

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 widely recognized survey and market research company that operates one of the largest global online panels, providing insights on public opinion across various topics, from politics to consumer behavior. As an online survey platform, YouGov employs a nonprobability sampling methodology, relying on a self-recruited panel and extensive demographic data to approximate representativeness. This methodology offers several advantages, including cost-efficiency, rapid data collection, and the ability to adapt to specific research needs. However, like all methodologies, it comes with its own set of trade-offs, particularly in terms of representativeness and potential selection biases.

This appendix takes a closer look at YouGov's sampling approach, panel recruitment methods, handling of non-response, and overall survey methodology. It also includes a reflection on the user experience, examining how elements like the onboarding process and incentive structure can shape participant engagement and data quality. By analyzing the strengths and weaknesses of YouGov's approach, this paper aims to provide a comprehensive understanding of how YouGov balances methodological rigor with the practical realities of online polling, highlighting both the reliability and the limitations of its data.

A.1 Population, Frame, and Sample

For YouGov surveys, the population often includes U.S. adults or specific subgroups, such as registered voters or other demographic or political segments. The frame from which YouGov draws its samples is its proprietary online panel, consisting of self-recruited individuals who provide extensive demographic information upon joining. This frame is nonprobabilistic and is designed to capture a wide demographic, with adjustments to enhance representativeness through post-survey weighting (YouGov).

The sample is selected from this online panel based on survey requirements. For general surveys representing the U.S. adult population, YouGov applies demographic matching and stratified sampling to recruit individuals whose characteristics resemble the target population. For more targeted surveys, such as those involving younger adults or specific voter segments, YouGov filters panelists based on demographic and political characteristics, ensuring that the survey sample aligns with the study's focus (YouGov).

A.2 Sample Recruitment

YouGov's panelists are recruited through online advertisements, partnerships with various websites, and organic sign-ups. This open online recruitment allows any adult within the

United States with internet access to join the panel, promoting inclusivity in terms of access and accessibility (YouGov). Once part of the panel, participants voluntarily complete surveys for points that can be exchanged for rewards, making monetary compensation a moderate incentive but not the sole motivator for participation. This approach is well-suited to engage a broad and diverse sample, although it does introduce limitations by excluding those without internet access.

Because the recruitment is continuous and participants often stay on the panel long-term, YouGov can maintain and monitor demographic balance, update panelist information over time, and reduce redundancy in questions for returning members. However, relying solely on online recruitment may introduce a bias toward individuals more inclined toward online engagement, who may not fully represent groups with lower internet usage, such as older adults or those from lower-income backgrounds (YouGov).

A.3 Sampling Approach and Trade-offs

YouGov employs a nonprobability sampling approach, meaning that not every individual in the population has an equal chance of selection. Within its panel, YouGov uses stratified sampling, drawing subsets of respondents based on demographic targets (e.g., age, race, gender, and region) to match the sample as closely as possible to the desired population structure.

A.3.1 Trade-offs of the Nonprobability Approach

- **Strengths**: Nonprobability sampling allows YouGov to gather data rapidly and cost-effectively. By using stratified sampling and post-survey weighting, the company aims to approximate the representativeness of probability-based samples while maintaining flexibility in addressing specific research goals and client needs (YouGov).
- **Limitations**: Because nonprobability samples lack random selection, they do not guarantee true representativeness, which can introduce unknown biases. YouGov addresses this partially by applying demographic weights, but nonprobability sampling's inherent limitations mean results should be interpreted as estimates rather than fully precise reflections of the target population (Huffington Post).

A.4 Handling Non-Response

Non-response is a common issue in survey research, particularly with online surveys where participants may decline to participate, partially complete surveys, or provide inconsistent answers. YouGov tackles non-response bias through quality control processes that exclude unreliable responses. These controls involve detecting respondents who complete surveys too quickly, answer inconsistently, or fail attention checks (YouGov).

In addition, YouGov adjusts for non-response by weighting the final sample to reflect the target population's demographic characteristics more accurately. This adjustment helps address imbalances resulting from non-response, enhancing the data's generalizability. However, non-response bias remains a concern, especially as weighting cannot entirely account for individuals not present in the sample or for differences in survey engagement across demographics (YouGov).

A.5 Strengths and Weaknesses of YouGov's Methodology

A.5.1 Strengths

- **Efficient Data Collection**: YouGov's nonprobability online sampling allows for efficient, cost-effective data collection, enabling timely survey responses that can address rapidly changing social and political topics.
- **Targeted Panel Management**: YouGov's proprietary panel enables consistent, direct contact with respondents, allowing for long-term engagement and improved tracking of demographic and opinion changes without needing repetitive data entry.
- **Rigorous Quality Controls**: Extensive quality checks, including IP verification, email authentication, and response validity checks, help YouGov maintain high data integrity. Respondents who fail quality checks are excluded, helping to reduce noise from unreliable data.
- **Adjustments for Representativeness**: While nonprobability sampling has limitations, YouGov partially mitigates these by applying sophisticated weighting based on benchmarks from sources like the U.S. Census. This enhances the sample's demographic alignment with the general population (YouGov).

A.5.2 Weaknesses

- **Nonprobability Sampling Limitations**: The lack of true random sampling introduces potential bias, as certain population segments may be underrepresented. For example, people without internet access or those less inclined toward online engagement are systematically excluded, which could affect representativeness.
- **Potential Selection Bias**: Although incentives are modest, they may still attract individuals who are already more engaged or motivated to participate in surveys, which could influence results. The weighting procedures help but may not fully eliminate biases if certain groups are consistently underrepresented.
- **Limitations of Online-Only Surveys**: Although internet penetration in the U.S. is high, reliance on online-only surveys may still leave out individuals less comfortable with or able to access online technology, skewing results slightly.

- **Questionnaire Design Constraints**: While YouGov's questionnaire design is generally sound, with randomized question orders and multimedia inclusion for clarity, online surveys can limit response depth compared to in-person or phone surveys. Responses are constrained by pre-set answer options, potentially oversimplifying complex views and reducing respondent nuance (Huffington Post).

A.6 Reflection

We created an account on YouGov to gain real experience with the platform, as illustrated in Figure 2 and here is my reflection on the onboarding process and user experience.

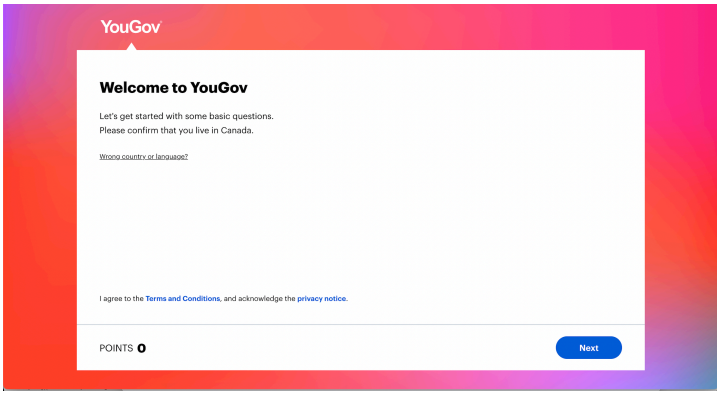


Figure 2: YouGov Welcome Page

The YouGov survey onboarding process is designed to be clear and straightforward, making it easy for new users to sign up and begin participating. The questions asked during onboarding are simple and direct, covering basic demographic details without overwhelming the participant. The first question of the first survey we took at YouGov is shown at Figure 3. This streamlined approach likely helps increase engagement and reduces drop-off rates, ensuring that users can quickly understand what is required of them. The layout and visuals are accessible, contributing to a positive first impression and making it more likely that new panelists will continue engaging with future surveys.

To enhance data integrity, YouGov includes verification steps such as confirming users' email addresses through a code. This extra step not only ensures that participants are real and unique individuals but also helps prevent fraudulent or duplicate accounts, as shown in Figure 4, which can otherwise skew survey results. By verifying email addresses, YouGov can maintain a higher quality of data and build a more reliable panel, knowing that each participant is committed enough to complete these security measures. This verification process is a straightforward yet effective method for improving data quality from the outset.

Data privacy is also well-considered in YouGov's onboarding, with clear options allowing users to control how their information is shared. For instance, participants can opt in or out of having

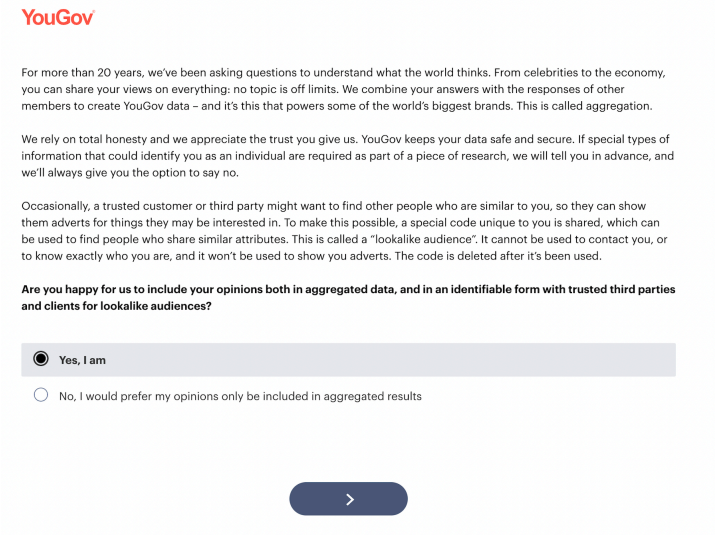


Figure 3: YouGov Survey

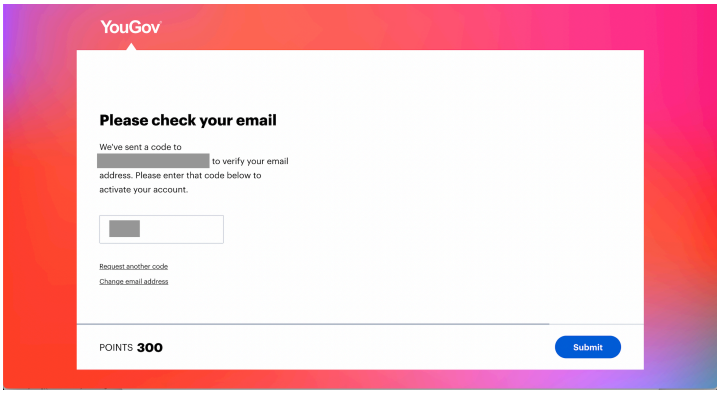


Figure 4: YouGov Email Verification

their responses shared in identifiable formats with trusted third parties. Figure 5 includes an example of how YouGov is letting users to control their information privacy. This transparency around data usage builds trust, showing participants that their privacy is valued and that they have agency over their data. The option to decide on data sharing likely encourages long-term participation, as users are reassured that their information will be handled responsibly.

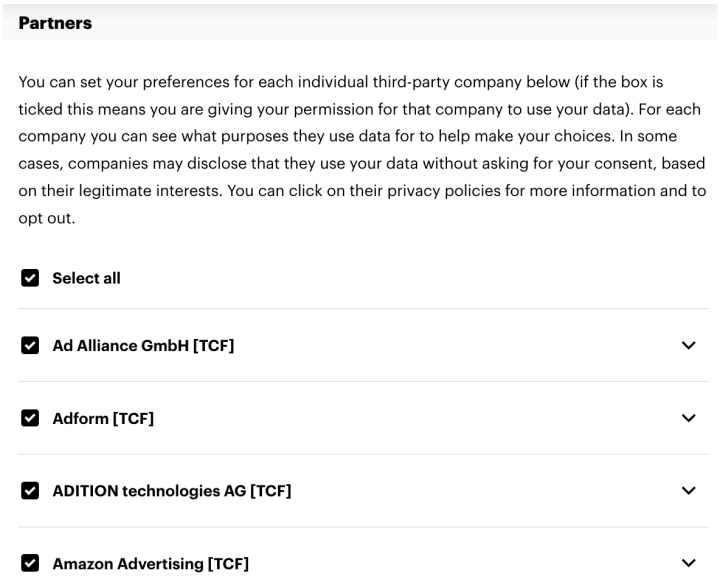


Figure 5: YouGov Data Privacy

However, the onboarding interface in Figure 6 has a gamified feel, awarding points for each step completed. This reward system, while motivating, could encourage participants to treat the surveys more like a task for monetary gain than an opportunity to provide genuine responses. The point-based incentive structure may attract users who are primarily interested in earning rewards, which risks impacting data quality if participants prioritize speed over thoughtful answers. This setup introduces the possibility that some users are engaging more for the rewards than for contributing to research, which could affect the reliability of the responses YouGov collects.

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

B.1 Methodology

This part of the appendix provides a detailed methodology and survey design to forecast the 2024 U.S. presidential election. Our survey focuses on swing states: Arizona, Georgia, Michi-

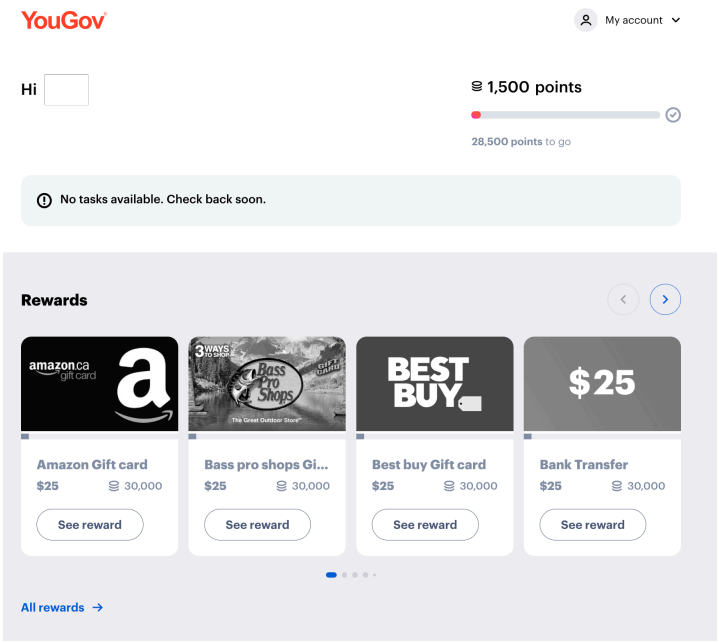


Figure 6: YouGov Interface

gan, Nevada, North Carolina, Pennsylvania, and Wisconsin due to the results of other states being certain. The survey targets different age groups using tailored recruitment strategies. Below, we break down the methodology, budget, recruitment strategies, and survey design to ensure representative participation.

B.1.1 Sampling Approach

- Total Sample Size: 7,000 respondents (1,000 participants per state).
- Target Population: Eligible voters from the following swing states: Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, Wisconsin.

B.1.1.1 Sampling Breakdown by Age Group:

- 18-24 years: 150 participants per state (total 1,050) – reached through university students.
- 25-60 years: 700 participants across states (total 4,900) – reached via online ads and survey platforms.
- 60+ years: 150 participants per state (total 1,050) – reached through phone calls.
- 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

B.1.2.1 18-24 Years (University Students):

- University Collaboration: Partner with academic institutions in each swing state to send survey invitations via email on behalf of the project.
- Rationale: Universities provide direct access to students, improving recruitment efficiency and response rates.
- Incentive: Each respondent receives a $10 e-gift card for a Walmart or Costco purchase.

B.1.2.2 25-60 Years (General Population):

- Online Ads and Survey Platforms: Use Facebook, Instagram, and platforms like Prolific to recruit participants from swing states.
- Budget Allocation: Allocate $10,000 for online ads to reach residents aged 25-60.
- Incentive: Participants receive e-gift cards upon survey completion.

B.1.2.3 60+ Years (Senior Citizens):

- Phone Surveys: Conduct live calls to reach seniors, considering that older people are not comfortable using electronic devices. Phone numbers will be sourced from voter registration databases (where legally accessible) or senior community networks.
- Incentive: Seniors can choose a physical Walmart or Costco gift card, which will be sent by mail.
- Recruitment Channels:
  - Complement with **social media ads** on Facebook, Instagram, and LinkedIn to reach underrepresented groups.
  - **Incentives**: Offer **$10 gift cards** to increase participation and engagement.

B.1.3 Trade-off

B.1.3.1 Data Representativeness vs. Control over Data Collection

- Issue: Although age quotas ensure participation from different age groups (18-24, 25-60, 60+), it is not feasible to set strict limits for race and income. Without perfectly matched samples for race and income, bias may arise in the survey results.
- Solution:
  - Weighting Strategy: Calculate population proportions for race and income within each state using Census data. Assign weights to groups based on their representation to mitigate bias.

B.1.3.2 Efficient Recruitment vs. Potential Selection Bias

- Issue: Limiting the 18-24 age group to university students may introduce selection bias. University students often differ from their non-student peers in socioeconomic background and possibly political perspectives.
- Impact: This selection bias may affect the generalizability of findings, as university students' perspectives might differ from all young adults.

B.1.3.3 Effective Senior Outreach vs. Higher Costs and Response Bias

- Issue: Using live phone operators to reach seniors increases costs, and some seniors may remain unreachable by phone.
- Impact: Higher costs may reduce resources for other age groups. Response bias may arise if unreachable seniors have different views, affecting the survey's representativeness.

B.1.3.4 Incentivizing Participation vs. Economic Bias in Sample

- Issue: Offering $10 gift cards as an incentive could attract participants more likely to need the incentive, potentially introducing economic bias.
- Impact: Overrepresentation of lower-income participants could skew results, affecting insights if income level correlates with political perspectives.

B.1.4 Survey Implementation and Design

In designing this survey, I referenced insights from the article on survey methods from (**AnnualReviewofEconomics?**), particularly on the advantages of online surveys over in-person, phone, or mail surveys. The article highlighted that online surveys provide flexibility, allowing respondents to complete them at their convenience, which minimizes selection bias related to work schedules or availability. This flexibility is especially beneficial for students and working-age individuals, who may find it challenging to respond during traditional hours. The article also emphasized how mobile technology can increase participation by reaching populations that are otherwise hard to engage, such as younger respondents, frequent movers, or those in remote areas. In line with these insights, I designed this survey with an online format for younger and working-age participants, maximizing accessibility and reducing drop-out rates. Additionally, I applied the article's suggestion of keeping surveys concise to prevent fatigue and disengagement, which helps maintain completion rates and improve data quality. The use of digital incentives, as noted in the article, also broadens the appeal and ensures we reach a diverse participant base across income levels, providing flexibility in rewards to cater to different motivations. This approach helped shape a balanced methodology that optimizes participation across demographic groups and aligns with best practices in survey design.

- Survey Platform: Google Forms
  - **Link**: https://forms.gle/2MGYeZavDsCNuWZ1A
  - 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.1.5 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.1.6 Poll Aggregation and Reporting

- Poll Aggregation:
  - 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.1.7 Budget Allocation

Expense	**Estimated Cost**
Participant Incentives	$70,000
Online Ads and Panel Provider Fees	$10,000
Phone Survey Salaries	$8,000
Google Forms (Platform)	Free
Data Validation & Analysis	$10,000
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:

1. What is your age?
   - 18-24
   - 25-39
   - 40-60
   - 60+

2. What is your gender?
    - Male
    - Female
    - Other: _____
3. What is your race/ethnicity?
    - White
    - Black
    - Hispanic or Latino
    - Asian
    - Indigenous
    - Other: _____
4. What is your highest level of education?
    - Less than high school
    - High school diploma
    - Some college
    - Bachelor's degree
    - Graduate degree or higher
5. What is your annual household income?
    - Less than $25,000
    - $25,000 - $49,999
    - $50,000 - $99,999
    - $100,000 or more
6. Which state do you currently reside in?

7. Are you a registered voter?
    - Yes
    - No
    - Not sure
8. How likely are you to vote in the 2024 presidential election?
    - 1 (Definitely will not vote)
    - 2
    - 3
    - 4
    - 5 (Definitely will vote)
9. If the 2024 election were held today, who would you vote for?
    - Kamala Harris (Democrat)
    - Donald Trump (Republican)
    - Undecided
    - Other: _____
10. What is the most important issue for you in this election?

- The economy
- Healthcare
- Immigration
- Climate change
- Social Security and Medicare
- Foreign policy

11. In your opinion, how will the majority in your state vote in 2024?

- Democrat
- Republican
- Too close to predict

12. Do you have any additional comments or suggestions?

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

https://forms.gle/2MGYeZavDsCNuWZ1A

C Model details

C.1 Posterior predictive check

C.2 Diagnostics

D References