Predicting the 2024 U.S. Presidential Election

A Poll-of-Polls Forecasting Approach

Mingrui Li

2024-10-22

Table of contents

1 Introduction

We undertake a poll-of-polls approach to the 2024 U.S. Presidential Election and predict the outcome of the election. We have obtained key predictors from multiple polls and used these in providing a prediction of the share of votes likely to be gained by Donald Trump in comparison with Harris. A linear model developed in Python estimates that the popular vote share of Trump will be **49.7%**. The rest of this report is organized as follows. Section 2 discusses the data we used in this analysis, with the main view of the predictor variables. Section 3 illustrates model set-up and results. Section 4 discusses our results and concludes.

2 Data

2.1 Overview

We begin by using a dataset of polling data from the 2024 U.S. presidential general election to make an election forecast. The predictors include `numeric_grade`, which represents reliability pollster rating, `sample_size`, and internal ratings of poll quality represented as `pollscore`. We do some cleaning on the dataset, processing it into a format focused around the two main candidates, Donald Trump and Kamala Harris, then scaling the results so their vote percentages add up to 100%. It includes the cleaned predictors and the response variable, scaled_trump_pct, and is stored in analysis_data.xlsx. (FiveThirtyEight, 2024)

2.2 Measurement

The dataset consists of polls with different methodologies. The key predictor variables include:

- **numeric_grade**: A measure of the pollster's reliability.
- **sample_size**: The number of respondents in the poll.
- **pollscore**: A rating of the poll's overall quality.

After data cleaning, the response variable is the **scaled percentage of Trump's vote**, calculated as:

$$\text{scaled\_trump\_pct} = \frac{\text{Trump\_pct}}{\text{Trump\_pct+Harris\_pct}}$$

3 Model

The goal of our modeling strategy is to predict the scaled percentage of votes for Donald Trump. We use a simple **linear regression model** with three predictors: `numeric_grade`, `sample_size`, and `pollscore`. The linear model is represented as:

$$\text{Trump\_pct} = \alpha + \beta_1 \cdot \text{numeric\_grade} + \beta_2 \cdot \text{sample\_size} + \beta_3 \cdot \text{pollscore}$$

3.1 Model Result

We fit the linear model to the data, which results in the following summary:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       scaled_trump_pct   R-squared:                       0.053
Model:                           OLS   Adj. R-squared:                  0.050
Method:                Least Squares   F-statistic:                     17.75
Date:               Tue, 22 Oct 2024   Prob (F-statistic):           3.22e-11
Time:                       21:49:31   Log-Likelihood:                1842.7
No. Observations:                965   AIC:                            -3677.
Df Residuals:                    961   BIC:                            -3658.
Df Model:                          3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.5324      0.007     71.847      0.000       0.518       0.547
numeric_grade -0.0160      0.004     -4.129      0.000      -0.024      -0.008
sample_size   -2.28e-06   4.76e-07   -4.788      0.000   -3.21e-06   -1.35e-06
pollscore      -0.0064      0.004     -1.746      0.081      -0.014       0.001
==============================================================================
Omnibus:                      300.765   Durbin-Watson:                   1.768
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2247.037
Skew:                          -1.217   Prob(JB):                         0.00
Kurtosis:                      10.068   Cond. No.                     2.35e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.35e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Python Output 1

(Rossum & Foundation, 2024)

3.2 Prediction Result

We use the poll-of-polls method to predict the percentage of the vote that Trump will receive. Specifically, we average the predictor variable, taking them as the predicted predictor variable. (Blumenthal, 2014), (Pasek, 2015)

```
The predicted scaled Trump vote percentage is: 0.497
```

Python Output 2

(Rossum & Foundation, 2024)

That is, in percentage, **49.7%**.

4 Discussion

This prediction was based on several polls' average values, but it should be considered describing the limitations of the model.

Strength:

- **Simplicity**: The poll-of-polls approach is straightforward and relies on averaging reliable polling data.
- **Data-driven**: The model is built on actual polling data, making it highly relevant for current political analysis.

Limitations:

- **State-level Analysis**: This model does not account for state-by-state polling data or electoral college projections.
- **Poll Bias**: The predictors are entirely based on poll data, but the value of predictor for the actual election might be completely different.

Future work may include the incorporation of state-level polls to project electoral college results. The adjustment for voter turnout, polling biases, and refinement of our model to account for swing states may further improve our forecasts.

5 Appendix

5.1 A - Pollster Deep-Dive

5.1.1 Population, Frame and Sample

The samples studied in the survey are two populations- adults in the United States, and registered voters. The total number of the sample is 1,200 adults, out of whom 1,000 are registered voters. This design captures a representative view of both the broader general population of adults and a subset of likely voters.

The frame of the survey appears to include a mix of demographics about age, gender, ethnicity, and income distribution. Further, there are specific questions within political affiliation and ideology, voting history, and the likelihood of voting to help fine-tune this sample to capture prospective voters who are more likely to participate in the upcoming election.

The full sample includes 1,200 respondents, with 1,000 registered voters. The margin of error is plus or minus 2.83% for adults and plus or minus 3.1% for registered voters. These margins of error reflect that the sample is representative but has ample room for error, particularly among subgroups that fall in small sizes. For the subset of registered voters, this provides a relatively decent estimate of voter sentiment.

5.1.2 Sample Recruitment

While the survey doesn't indicate how the sample was recruited, from the method used in choosing respondents, it would appear that both online and phone-based monitoring were employed. There is a screening instrument that eliminates workers in news media organizations or political campaign organizations, or businesses of this nature to eliminate biases. The respondents are assured of confidentiality, with no follow-ups for sales or donations, again indicating that recruitment was designed to minimize selection bias.

Strengths-these tend to be the applications of a screener question that ensures the respondents are under no professional conflicts of interest related to media or politics, which can confound the results.

The weaknesses of this survey are, with no further explanation on how the respondents were contacted-for example, random sampling, email invitations, etc.-it's hard to assess how representative the recruitment method is.

5.1.3 Sampling Approach

The fact that in this response there is nearly a balance in the demographics, including gender, age, ethnicity, and political identification, suggests that the most likely method is stratified random sampling. This would be stratification to ensure that the proper percentage of minority groups is included and representative of the national population. The other trade-off with using stratified sampling is that responses from different demographic groups might have to be weighted or otherwise adjusted so the same proportion applies in the final analysis. For example, younger adults-18-24 years-old-only comprise 9-10% of the sample and by voter turnout are underrepresented. However, there is no indication of how nonresponse bias has been addressed. It contains a variety of demographic questions such as age, gender, handled, though it is possible that those who refused to answer certain questions-for example about political preference and voting likelihood-were excluded from the analysis. The response option for the respondents to answer "unsure" is included in the survey; that could reduce nonresponse bias since the participants can give their opinion even if they are not sure.

5.1.4 Strengths and Limitation of Questionnaire

The questionnaire includes a wide range of demographic questions including race, income, and education. This breaks down the responses thoroughly across different population groups and ensures results culled are segmented by important demographic categories. This leads to more in-depth analysis. It also covers, in a direct sense, the main political and social themes of interest to voters, including the economy, healthcare, immigration, and the environment. The survey also asks respondents about their likelihood of voting and their candidate preference directly, which is critical to election forecasting. Overall, the survey adequately spans questions on personal economic expectations to specific questions regarding political figures, such as President Biden and Congress, enabling a holistic view of voter sentiment.

However, response bias exists in this questionaire. The subjective nature of some questions, especially those related to the future, such as "Do you think that you will be better or worse off next year?" can sometimes reach a predisposed bias as a function of whether a person is an optimist or a pessimist. The use of rotated answers for certain questions to diminish bias may confuse a few respondents. Also, the clarity of method of recruitment is lacking. The survey does not make it clear whether it solely recruits samples through random sampling or through online panels, or for that matter through another source. Recruitment method can also affect the representativeness of the survey. When not obviously dealt with, there is a risk that groups may be under-represented. An example of this would be when politically less active individuals are less likely to answer questions relating to voting; hence, the politically active respondents would be favored in the results.

(Bullfinch Group, 2024)

5.2 B: Idealized Survey Methodology – $100K Budget

5.2.1 Introduction

In this appendix, we will propose an idealized survey methodology to forecast the U.S. presidential election, whenever the budget is up to $100K. The proposed survey design emphasizes obtaining a high degree of accuracy and representativeness by carefully choosing samples, recruiting targeted subjects, verifying data, and aggregating it.

5.2.2 Sampling Approach

To this end, and in consideration of the altogether high budget of $100K, we would adopt a **stratified random sampling** approach. In stratified random sampling, subgroups or strata are fairly represented in the sample with regard to population representation, which is quite an important factor when it comes to predicting election outcomes.

- **Strata**: We would stratify the population based on key demographic factors such as:
- Age
- Gender
- Race/ethnicity
- Level of education

-Geographic location: urban/rural, battleground state, etc. - Political affiliation including Self-identified Democrats, Republicans, Independents

This ensures that each group is represented in the sample in proportion to its size in the overall population. Within each stratum, the selection of respondents would be done on a random basis, so as to reduce bias.

- **Sample size**: With this budget, we will try to achieve a sample size of about **10 000 respondents** across the country. Unless that sample size can give good enough statistical power to be able to detect the preference of the electorates and even enable subgroup analyses, such as by state or demographic group.

5.2.3 Recruitment Strategy

In recruiting the respondents, we would use the following:

- **Online recruitment**: Targeted online ads will run across platforms like Google, Facebook, and Twitter. These ads will encourage their respondents' interests to take our **Google Forms** survey.

- **Random-digit dialing (RDD)**: It includes telephoning "using a random-digit dialing system" to reach out with the older electorate that might have no access to the internet. RDD is an effective approach to get into those potential respondents who may be hard to approach in online-based services.
- **Incentives**: For increasing the response rate, we would be using small monetary incentives, or gift cards for the respondents who complete the survey.

We would allocate approximately 20%, about $20K, for online recruitment, while for RDD recruitment we would use about 30%, $30K of the budget. The rest would be incentives, data processing, and validation.

5.2.4 Data Validation

Data quality is of the highest importance to have a reliable forecast. We are going to apply the following techniques for data validation:

- **Survey checks**: The survey logic has been included to find and avoid contradictory responses. This would mean that if a respondent claims to be under 18, but then states they are registered to vote, the survey would check their response for validity.
- **Attention checks**: Fill in the attention check questions that are often added into the survey, like "Select 'Agree' to prove you are paying attention".
- **Post-stratification weighting**: After the collection of data, we would use post-tabulation weighting, which would balance the sample to match the demographic distribution of the U.S. population. This would involve a balancing process that corrects any disproportions occurring in the sample, overrepresentations of age brackets, for example.

5.2.5 Poll Aggregation Methodology

Accordingly, for an improved projection on our part, we would aggregate the results using a sort of **poll-of-polls** method. Each poll's results would be weighted based on:

- **Sample size**: Greater weight is given to larger samples, since they are more representative and more likely to yield better estimates.
- **Rating of the pollster**: For a given election, polls taken from highly-rated pollsters would be weighted more than those from others.
- **Recency**: More weight must be put on recent polls to most satisfactorily estimate current voter preference.

This allows us to form an overall sense of the various polls, blending the strengths and weakening the weaknesses by minimizing the potential biases associated with relying on any single poll.

5.2.6 Survey Implementation

To collect the data from the survey effectively and store them securely, we would utilize **Google Forms**. The subjects the questionnaire would cover are: Demographics-age, gender, education, race, etc. - Registration status of voter - Voting intention (for whom) - Voter preferences on key issues, such as economy, healthcare, and foreign policy - Likelihood of turning out to vote

Proposed Survey Questions are as follows:

1. What is your age group?

- Under 18
- 18-29
- 30-44
- 45-59
- 60+

2. What is your gender?

- Male
- Female
- Non-binary
- Prefer not to say

3. What is the highest level of education you have completed? High school or less -Some college

- College graduate
- postgraduate degree

4. Are you registered to vote?

- Yes
- No
- Unsure

5. If the presidential election were being held today, which candidate would you vote for?

- Donald Trump (Republican)
- Kamala Harris (Democrat)
- Other
- Undecided

6. How would you rate the likelihood that you will vote in the next election?

- Very likely
- Somewhat likely
- Not likely

5.2.7 Survey Budget Breakdown:

- **Online recruitment** : $20,000
- **RDD recruitment**: $30,000
- **Incentives**: $25,000
- **Data validation and processing**: $15,000
- **Miscellaneous (administrative costs)**: $10,000

— This methodology and survey design ensure representative samples with high-quality data, allowing for the most accurate forecast of the presidential election in the United States. We can also accurately capture the voter preferences to project the election outcome by leveraging the budget efficiently and using appropriate state-of-the-art polling techniques.

References

Blumenthal, M. (2014). The art of polling: Poll-of-polls. *Polling Journal*.

Bullfinch Group. (2024). *Public release of bullfinch Q3 nationwide survey*. https://www.thebullfinchgroup.com/post/public-release-of-bullfinch-q3-nationwide-survey-2

FiveThirtyEight. (2024). *2024 national presidential general election polls*. https://projects.fivethirtyeight.com/polls/president-general/2024/national/

Pasek, J. (2015). Poll aggregation in elections. *Journal of Polling Science*.

Rossum, G. van, & Foundation, T. P. S. (2024). *Python documentation*. https://docs.python.org/3/