Forecasting the 2024 US Presidential Election: A Poll-of-Polls Approach to Predict the Outcome[*]

Utilizing Poll Aggregation and Statistical Modeling to Project a Clear Path for the Presidential Race

Ruiying Li   Lorina Yang

October 21, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

The ongoing U.S. presidential election is a major focus in the media, finance, and politics, making it essential to provide reliable forecasts and understand the factors influencing varying levels of candidate support. This paper aims to give reliable prediction of election outcome by building a multiple linear regression (MLR) model to forecast the percentage of voter support for top 5 candidates in the 2024 U.S. presidential election.

1.1 Paper Structure

The remainder of this paper follows a structure Section 2,Section 3 about MLR model,Section 4,Section 5,Section A about pollster methodology,Section B about our methodology on survey.

[*] Code and data are available at: https://github.com/Lorina-Y/US-Presidential-election.git.

1.2 Estimand paragraph

This paper's estimand is the predicted supported percentage, we estimate it by using predictors such as sample size, party,pollster,poll quality score, transparency score and candidat name. These predictors allow for a more nuanced understanding of how various elements affect a candidate's polling percentage.

1.3 Results paragraph

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023)…. Our data (Toronto Shelter & Support Services 2024)…. Following Alexander (2023), we consider… Here are variables used in the paper: pct: This is the response variable, representing the percentage of support a candidate received in a poll. sample_size: The size of the sample surveyed in each poll. This can be a predictor as larger samples typically yield more reliable poll results. party: The party affiliation of the candidate (e.g., Democrat or Republican). pollster: The name of the polling organization that conducted the poll. answer: The candidate's name (e.g., Harris, Trump). In this paper, we only study the top 5 candidate most frequently appeared in answer column. transparency_score: A numeric score representing the transparency of the poll's methodology. pollscore: A score evaluating the quality of the poll, likely based on the pollster's historical accuracy. state: A location name represent states in the U.S.

The 10 observation of dataset we use is displayed below:

```
# A tibble: 6 x 8
    pct sample_size party pollster     answer transparency_score pollscore state
  <dbl>       <dbl> <chr> <chr>        <chr>               <dbl>     <dbl> <chr>
1  43.3        1084 DEM   Trafalgar G~ Harris                  4       0.6 Penn~
2  46.3        1084 REP   Trafalgar G~ Trump                   4       0.6 Penn~
3  46           691 DEM   Redfield & ~ Harris                  9       0.4 Ariz~
4  49           691 REP   Redfield & ~ Trump                   9       0.4 Ariz~
5   1           691 GRE   Redfield & ~ Stein                   9       0.4 Ariz~
6  45          1275 DEM   Redfield & ~ Harris                  9       0.4 Flor~
```

2.2 Measurement

The method involves translating public opinion polling into measurable data for analysis. Polls capture the percentage of support for each candidate, recorded as pct, along with factors like sample_size, transparency_score, and pollscore, which reflect the quality and methodology of each poll. These variables represent the underlying polling practices and voter sentiment, allowing us to analyze how real-world opinions are measured and recorded in the dataset to forecast election outcomes.

2.3 Outcome variables

The paper predict an outcome variables of PCT, which represent the percentage of support a candidate received in a poll, allow use to predict which candidate get most support.

Add graphs, tables and text. Use sub-sub-headings for each outcome variable or update the subheading to be singular.

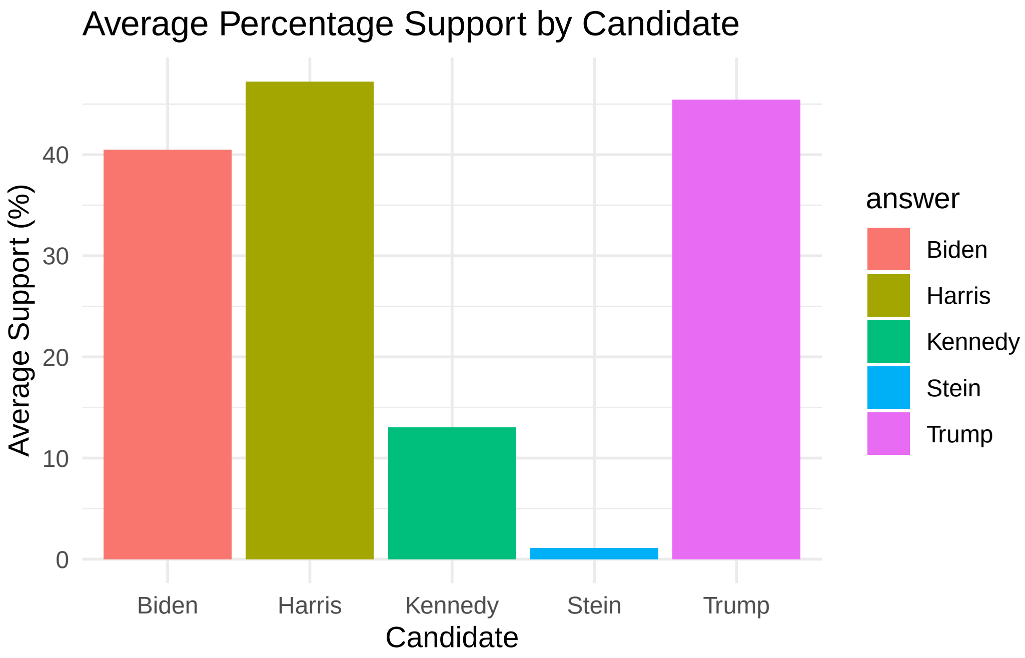

Figure 1: Average Percentage Support(pct) by top 6 Candidate

2.4 Predictor variables

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

```
 DEM  GRE  IND  REP 
2038  668  713 1996 
```

3 Model

The goal of our modelling strategy is use multiple linear regression to model the relationship between response variable pct, and 5 predictors variables including sample_size, party, transparency_score, pollscore. MLR allows us to estimate the effect of each predictor on the response while controlling for other factors in the model.

3.1 Model set-up

Define y_i as the pct- percentage support a candidate get. Then β_i is the sample_size, pollscore, transparency score, party.

We run the model in R (R Core Team 2023).

3.1.1 Model justification

We expect a positive relationship between the pct and sample size,pollscore and transparency score. And the party variable may show significant effects for certain categories (such as Democrat or Republican), we expect Candidates from major parties may have higher support compared to those from minor parties.

4 Results

Our results are summarized in Table 1.

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

Residuals:
```

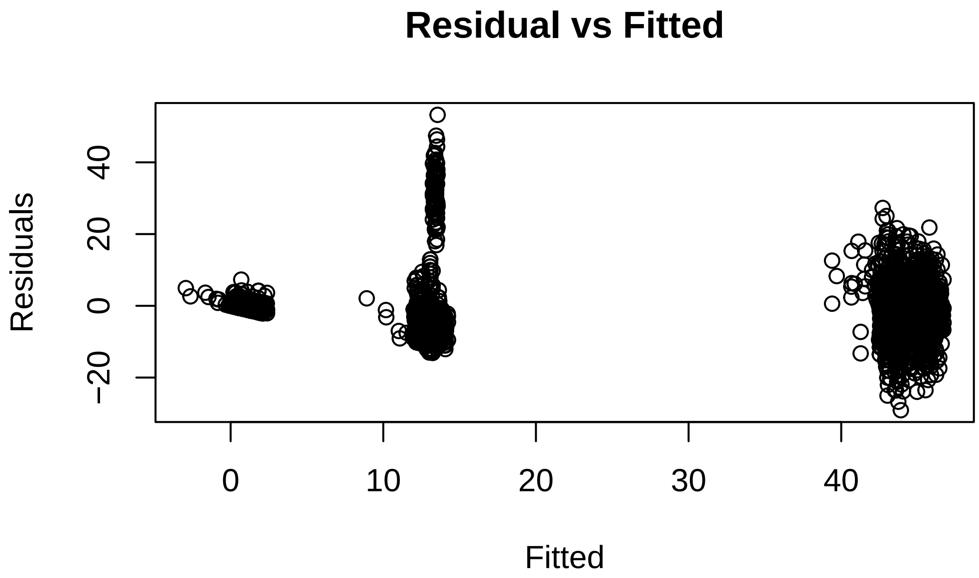

Figure 2: Average Percentage Support(pct) by top 6 Candidate

```
    Min      1Q  Median      3Q     Max 
-29.105  -3.800  -0.044   3.106  53.242 

Coefficients:
                    Estimate Std. Error  t value Pr(>|t|)    
(Intercept)        4.522e+01  3.341e-01  135.367  < 2e-16 ***
sample_size       -6.202e-04  2.286e-04   -2.713  0.00669 ** 
partyGRE          -4.234e+01  3.172e-01 -133.480  < 2e-16 ***
partyIND          -3.049e+01  3.104e-01  -98.207  < 2e-16 ***
partyREP           1.869e+00  2.235e-01    8.364  < 2e-16 ***
transparency_score -2.046e-01  4.189e-02   -4.885 1.07e-06 ***
pollscore         -3.733e-01  1.613e-01   -2.314  0.02068 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.096 on 5408 degrees of freedom
Multiple R-squared:  0.8446,    Adjusted R-squared:  0.8444 
F-statistic:  4897 on 6 and 5408 DF,  p-value: < 2.2e-16
```

Conclusion In conclusion, the multiple linear regression model explains 84.44% of the variability in candidate support, with an Adjusted R-squared of 0.8444. Key findings show that

Table 1: Explanatory models of flight time based on wing width and wing length

	First model
(Intercept)	45.22
	(0.33)
sample_size	0.00
	(0.00)
partyGRE	−42.34
	(0.32)
partyIND	−30.49
	(0.31)
partyREP	1.87
	(0.22)
transparency_score	−0.20
	(0.04)
pollscore	−0.37
	(0.16)
Num.Obs.	5415
R2	0.845
R2 Adj.	0.844
AIC	36 597.8
BIC	36 650.5
Log.Lik.	−18 290.877
RMSE	7.09

larger sample sizes slightly reduce support percentages, while party affiliation plays a significant role, with Green and Independent candidates receiving notably less support compared to Republicans. Both transparency score and poll quality score negatively affect support, suggesting more credible polls may report lower percentages. Despite a few outliers, the model is highly significant, providing valuable insights into factors influencing candidate support.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

A Appendix1

B Appendix2

C Additional data details

D Model details

D.1 Posterior predictive check

Examining how the model fits, and is affected by, the data

D.2 Diagnostics

Checking the convergence of the MCMC algorithm

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

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

Toronto Shelter & Support Services. 2024. *Deaths of Shelter Residents*. https://open.toronto.ca/dataset/deaths-of-shelter-residents/.