Forecasting the US Presidential Election: A Poll-of-Polls Approach Using Linear Models*

Ziqi Zhu Yuanchen Miao Claire Chang

October 22, 2024

1) what was done, 2) what was found, and 3) why this matters. 4) sentence.

Table of contents

*Code and data are available at: https://github.com/zzq20010617/2024_USelection_prediction

1 Introduction

This paper examines the development of a multiple linear regression (MLR) model to predict the percentage of support (pct) for U.S. presidential candidates based on polling data. The data includes a range of predictors, such as sample size, pollster quality factors, and time-related factors. The focus of the analysis is on nationwide polling data, aggregated from different pollsters, to create a robust model for forecasting election outcomes. The goal is to provide a clearer understanding of how various poll attributes influence polling results and to derive insights that can predict election outcomes.

The primary estimand in this study is the expected percentage of support (pct) for a U.S. presidential candidate, given polling data and relevant predictor variables like rating of the pollster and sample size of the poll.

Our Results shows that

This analysis is useful in understanding how various polling factors contribute to predicting election results. By modeling the relationship between polling result and factors, the study enhances the ability to forecast election outcomes based on public opinion data. This model provides a practical tool for researchers, political strategists, and analysts to assess the reliability of polling data and its implications for elections.

2 Data

2.1 Overview

Our data is download from (FiveThirtyEight, n.d.), the website gathered survey data from different pollsters, We use the statistical programming language R (R Core Team 2023), and packages (Grolemund and Wickham 2011), (Wickham et al. 2023), (Wickham et al. 2019), to process the data. Following (**tellingstories?**), we consider…

2.2 Measurement

The dataset is about public opinions about U.S. presidential candidates, which are captured through polling.

2.3 Outcome variables

Our primary outcome variable is **support percentage (pct)**, which represents the percentage of respondents who support each candidate. This variable is modeled as the response variable in the MLR analyse.

Talk more about it.

And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

Talk way more about it.

2.4 Predictor variables

2.4.1 Sample Size

Sample size is an important predictor because it influences the reliability of a poll. Larger sample sizes tend to reduce sampling error, providing more accurate reflections of voter sentiment. In our MLR model, the sample size is log-transformed to account for diminishing returns—larger polls do not necessarily offer proportionally better accuracy.

2.4.2 Poll Score

The poll score is a measure of the error and bias we can attribute to a pollster, negative number is better.

2.4.3 Numeric Grade

This variable is an aggregate score of the poll based on a numeric scale. It serves as an indicator of the pollster's historical performance and the methodology used. 3 is the maximum and some pollsters have no rating.

2.4.4 Transparency Score

Transparency score measures how openly a pollster reports their methodology and results. A higher transparency score suggests that the pollster has disclosed key details about how the poll was conducted, improving trust in the poll results.

2.4.5 Days Since Start

This variable represents the number of days since the beginning of the election polling period. It helps capture the dynamic nature of voter preferences over time, accounting for shifts in public opinion as the election date nears.

3 Models

The goal of our modelling strategy is try to capture the. Firstly,... In the following section we briefly describe the models we used to investigate. Background details and diagnostics are included in Appendix D.

3.1 SLR Model set-up

3.2 MLR Model set-up

We first filtered our data by different parties, as the single linear model but include other 3 parties(GRE, IND, and LIB, Conservative only has two entries in cleaned nationwide data and has 0 pct so its been removed), and fit linear model for each of them. The formula for Multiple linear regression is as follow

$$\text{pct}_i = \beta_0 + \beta_1 \log(\text{sample\_size}_i) + \beta_2 \,\text{pollscore}_i + \beta_3 \,\text{numeric\_grade}_i \tag{1}$$

$$+ \beta_4 \,\text{transparency\_score}_i + \beta_5 \,\text{days\_since\_start}_i + \epsilon_i \tag{2}$$

Response Variable is pct. Predictors are log(sample_size), pollscore, numeric_grade, transparency_score, and days_since_start, detail can be found in Section 2.3, and Section 2.4

Table 1: Explanatory models of pct based on pollscore, sample size, and time

	DEM	REP	GRE	IND	LIB
(Intercept)	37.33	50.12	3.34	−5.76	4.13
	(1.67)	(1.62)	(0.87)	(3.68)	(1.70)
log(sample_size)	0.38	−0.23	−0.19	2.42	−0.12
	(0.18)	(0.18)	(0.11)	(0.45)	(0.21)
pollscore	−0.72	−3.33	0.02	−1.16	−1.00
	(0.38)	(0.37)	(0.13)	(0.58)	(0.30)
numeric_grade	0.20	−3.55	−0.01	−1.80	−1.07
	(0.49)	(0.47)	(0.17)	(0.81)	(0.39)
transparency_score	−0.34	0.07	0.00	0.13	0.01
	(0.07)	(0.06)	(0.03)	(0.14)	(0.06)
days_since_start	0.03	0.01	0.00	−0.03	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Num.Obs.	1235	1235	270	603	105
R2	0.357	0.121	0.155	0.179	0.133
R2 Adj.	0.355	0.117	0.139	0.172	0.090
AIC	6780.2	6711.8	552.1	3623.3	269.7
BIC	6816.0	6747.6	577.3	3654.1	288.3
Log.Lik.	−3383.085	−3348.901	−269.070	−1804.655	−127.852
RMSE	3.74	3.64	0.66	4.83	0.82

3.2.1 Result

Our results for the MLR model are summarized in Table 1.

We first check the residual vs. fitted plot as shown in Appendix D, and see no obvious violation of assumptions. According to the summary, DEM and REP models are based on the largest number of observations (both 1235). Smaller parties have fewer observations, which may lead to less robust models. The r^2 shows that the model explain the highest amount of variance on Democratic Party than others with value 0.357, and other parties less than 0.2. Based on the summary of models for DEM and REPTable 2, we are not sure if those predictor that measure quality of polls are significant or not, so we keep them for now in Bayesian model. Prediction of MLR models is shown in Figure 1

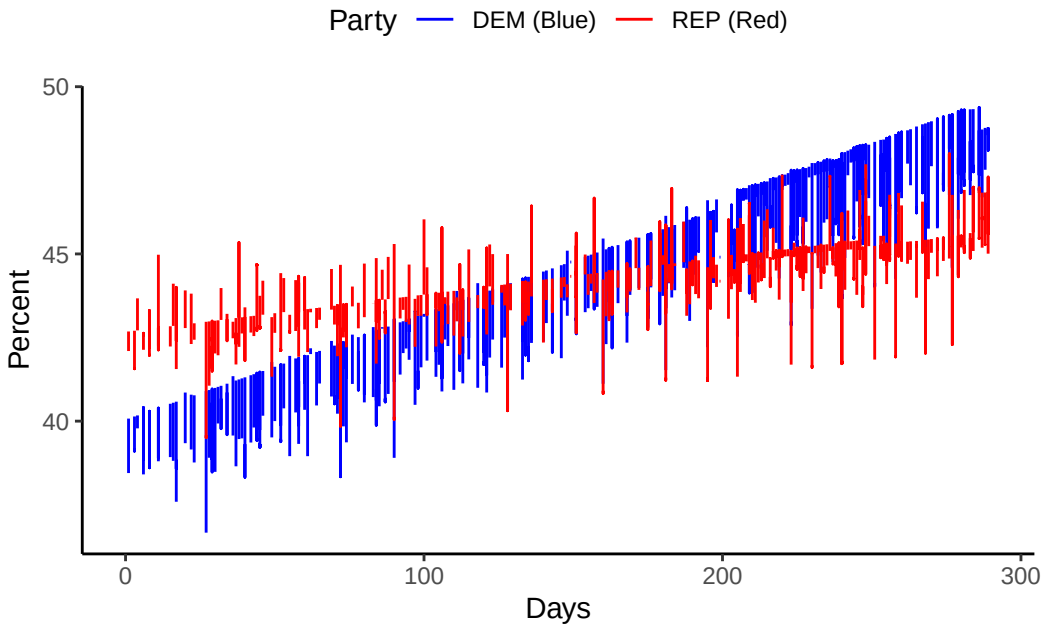


Figure 1: Prediction with MLR models for DEM and REP party

3.3 Bayesian Model set-up

Our model is build from a transformed version of the MLR models from above section, bu y_i represent the number of individuals in the sample that support the Democratic party (this corresponds to num_party). The response is modeled as binomial: $y_i \sim \text{Binomial}(\text{sample\_size}_i, p_i)$, where p_i is the probability that an individual in poll i supports the Democratic party, and sample_size_i is the total number of individuals surveyed in poll i.

$$y_i | p_i \sim \text{Binomial}(\text{sample\_size}_i, p_i) \tag{3}$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \log(\text{sample\_size}_i) + \beta_2\, \text{pollscore}_i + \beta_3\, \text{numeric\_grade}_i + \alpha_j \tag{4}$$

$$\alpha_j \sim \text{Normal}(0, \sigma_{\text{pollster}}) \tag{5}$$

$$\beta_0 \sim \text{Normal}(0, 2.5) \tag{6}$$

$$\beta_1 \sim \text{Normal}(0, 2.5) \tag{7}$$

$$\beta_2 \sim \text{Normal}(0, 2.5) \tag{8}$$

$$\beta_3 \sim \text{Normal}(0, 2.5) \tag{9}$$

$$\sigma_{\text{pollster}} \sim \text{Exponential}(1) \tag{10}$$

3.4 Spline Model set-up

We run the model in R (R Core Team 2023) using the `rstanarm` package of (**rstanarm?**). We first priors as default with "autoscale = TRUE",

4 Prediction

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.

Appendix

A Idealized methodology

B Idealized Survey

C Trafalgar group's methodology overview and evaluation

The Trafalgar Group conducts polls ranging from major political campaigns to marketing surveys. The organization's 0.7 pollster rating indicates moderate accuracy and reliability compared to other pollsters. The population consists of all eligible voters in the U.S., while the sampling frame includes registered voters across states, segmented by demographics like age, gender, party affiliation, and ethnicity. Trafalgar Group typically samples likely voters, focusing on those most likely to participate in upcoming elections. Trafalgar Group recruits its sample using a mix of interactive voice response (IVR), live phone calls, text messages, online panels, and email surveys. Trafalgar Group employs stratified random sampling to ensure proportional representation of subgroups like political party, gender, and region. This method improves accuracy by reflecting the electorate's demographic and political makeup but may risk over-stratification, giving smaller voter groups disproportionate influence and potentially skewing results. Trafalgar Group handles non-response by using weighting to adjust the sample, ensuring respondent demographics match population parameters. This method reduces bias by accounting for underrepresented groups, though heavy reliance on post-survey adjustments may introduce new biases, particularly when there are large discrepancies in response rates across demographics.

D Model details

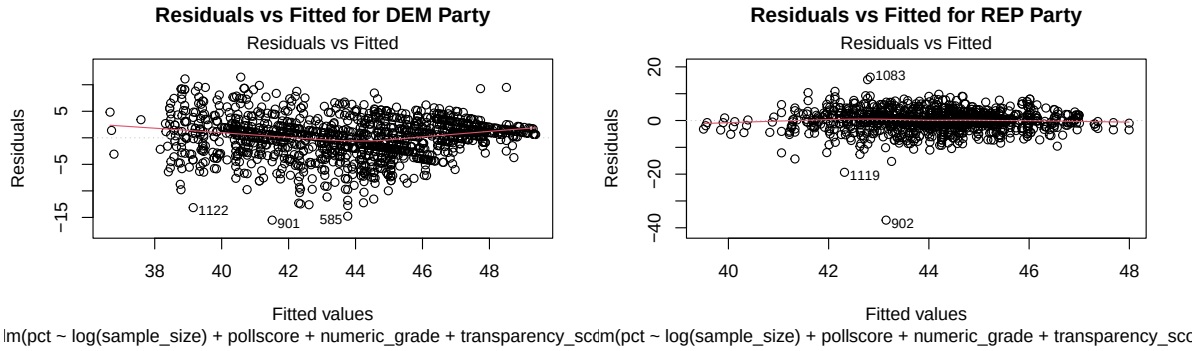
D.1 Assumption check for MLR models

D.2 Significant check for MLR models

D.3 Posterior predictive check

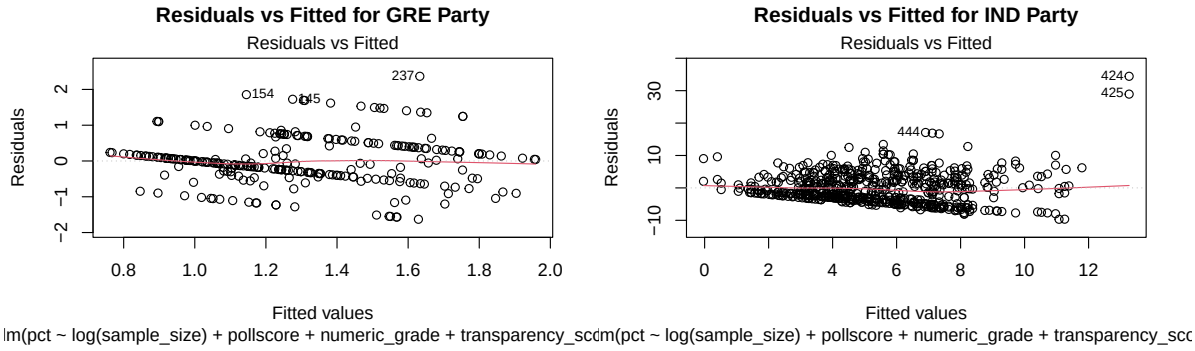
In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...



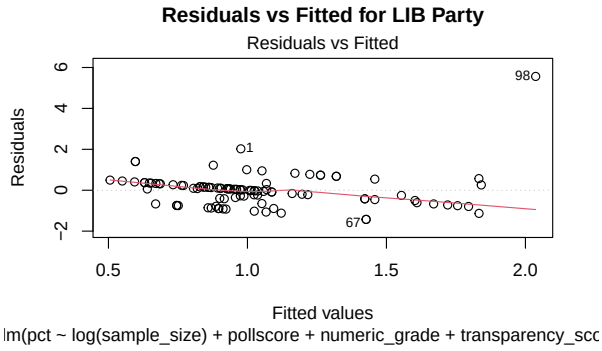
(a)

(b)



(c)

(d)



(e)

Figure 2: Examining how the model fits, and is affected by, the data

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

Table 2: Summary of DEM and REP data fit by MLR model

```
Call:
lm(formula = pct ~ log(sample_size) + pollscore + numeric_grade +
    transparency_score + days_since_start, data = party_data)

Residuals:
     Min       1Q   Median       3Q      Max
-15.5091  -2.0181   0.4498   2.1124  11.4283

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        37.33222    1.66922  22.365  < 2e-16 ***
log(sample_size)    0.37656    0.18292   2.059   0.0397 *
pollscore          -0.72219    0.37612  -1.920   0.0551 .
numeric_grade       0.19976    0.48656   0.411   0.6815
transparency_score -0.33589    0.06612  -5.080 4.35e-07 ***
days_since_start    0.03163    0.00132  23.971  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.754 on 1229 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.3572,    Adjusted R-squared:  0.3546
F-statistic: 136.6 on 5 and 1229 DF,  p-value: < 2.2e-16


Call:
lm(formula = pct ~ log(sample_size) + pollscore + numeric_grade +
    transparency_score + days_since_start, data = party_data)

Residuals:
    Min      1Q  Median      3Q     Max
-37.146  -1.297  -0.054   1.861  16.171

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        50.124690   1.623651  30.872  < 2e-16 ***
log(sample_size)   -0.230845   0.177923  -1.297    0.195
pollscore          -3.333315   0.365853  -9.111  < 2e-16 ***
numeric_grade      -3.547745   0.473277  -7.496 1.25e-13 ***
transparency_score  0.066665   0.064311   1.037    0.300
days_since_start    0.011144   0.001284   8.682  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
Residual standard error: 3.652 on 1229 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.1206,    Adjusted R-squared:  0.1171
F-statistic: 33.72 on 5 and 1229 DF,  p-value: < 2.2e-16
```

D.4 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

Checking the convergence of the MCMC algorithm

References

FiveThirtyEight. n.d. "FiveThirtyEight Presidential General Election Polls - National (2024)." https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

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

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.