Forecasting the US Presidential Election: A Poll-of-Polls Approach Using Linear Models By Using Data of Pollsters From Five Thirty Eight In 2024*

Ziqi Zhu Yuanchen Miao Claire Chang

November 2, 2024

This paper presents a multiple linear regression model for forecasting the outcome of the 2024 U.S. Presidential Election. Using nationwide polling data, we predict the percentage of support for U.S. presidential candidates based on key factors such as sample size, pollster quality, and timing. Our model aggregates these predictions to simulate potential election outcomes. The results reveal how specific poll attributes affect the accuracy of forecasts, offering valuable insights for predicting the likelihood of various electoral scenarios.

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 Five Thirty Eight(FiveThirtyEight, n.d.), the website gathered survey data from different pollsters of 2024 US president election, We use the statistical programming **language R** (R Core Team 2023), and packages **lubridate**(Grolemund and Wickham 2011), **dplyr**(Wickham et al. 2023), **tidyverse**(Wickham et al. 2019), **model summary**(Arel-Bundock 2022) to process the data. Also, we use **tibble** (Müller and Wickham 2023) to create the table for simulate table and save it as csv files. Graphs are created by using ggplot2 in from package **tidyverse**(Wickham et al. 2019)Following **Telling stories with data**(Data 2024), we consider using multiple linear model to predict and forecasting the result of US president election in 2024. We are creating different graphs and models for the candidates from different parties in general state and combine all models and graphs together to create a new scatter plot to predict and forecasting the US president election in 2024.

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.

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 trend of support percentage for different parties in 2024 Presidential election as much as possible. In the following section we briefly describe the models we used to investigate. Background details and diagnostics are included in Appendix B.

3.1 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 find in Section 2.3, and Section 2.4

3.1.1 Result

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

We first check the residual vs. fitted plot as shown in Appendix B, 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 REP Section B.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, which we can see blue line which is Democratic party has a steeper slope and start to take the lead between day 100 and day 200.

3.2 Bayesian Model set-up

The Bayesian model is build from the MLR models from above section. With the inspiration from the example R code, we introduce the random effects for Pollsters, and we use a logistic link to predict the probability of support, modeling count data which is in binomial distribution. We choose to use a default prior and enable autoscale, the formula is shown below

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

	DEM	REP	IND	GRE	LIB
(Intercept)	37.15	50.18	−4.45	3.01	4.36
	(1.46)	(1.42)	(3.47)	(0.77)	(1.41)
log(sample_size)	0.41	−0.25	2.26	−0.14	−0.21
	(0.16)	(0.16)	(0.43)	(0.10)	(0.17)
pollscore	−0.71	−3.33	−1.09	−0.04	−0.86
	(0.35)	(0.34)	(0.55)	(0.12)	(0.26)
numeric_grade	0.12	−3.55	−1.73	−0.07	−1.04
	(0.45)	(0.44)	(0.77)	(0.16)	(0.32)
transparency_score	−0.32	0.06	0.12	0.00	0.07
	(0.06)	(0.06)	(0.14)	(0.03)	(0.05)
days_since_start	0.03	0.01	−0.03	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Num.Obs.	1348	1348	636	307	126
R2	0.395	0.150	0.204	0.136	0.109
R2 Adj.	0.392	0.147	0.198	0.122	0.072
AIC	7320.7	7249.2	3792.0	619.1	313.9
BIC	7357.1	7285.7	3823.2	645.2	333.8
Log.Lik.	−3653.347	−3617.616	−1888.994	−302.568	−149.968
RMSE	3.64	3.54	4.72	0.65	0.80

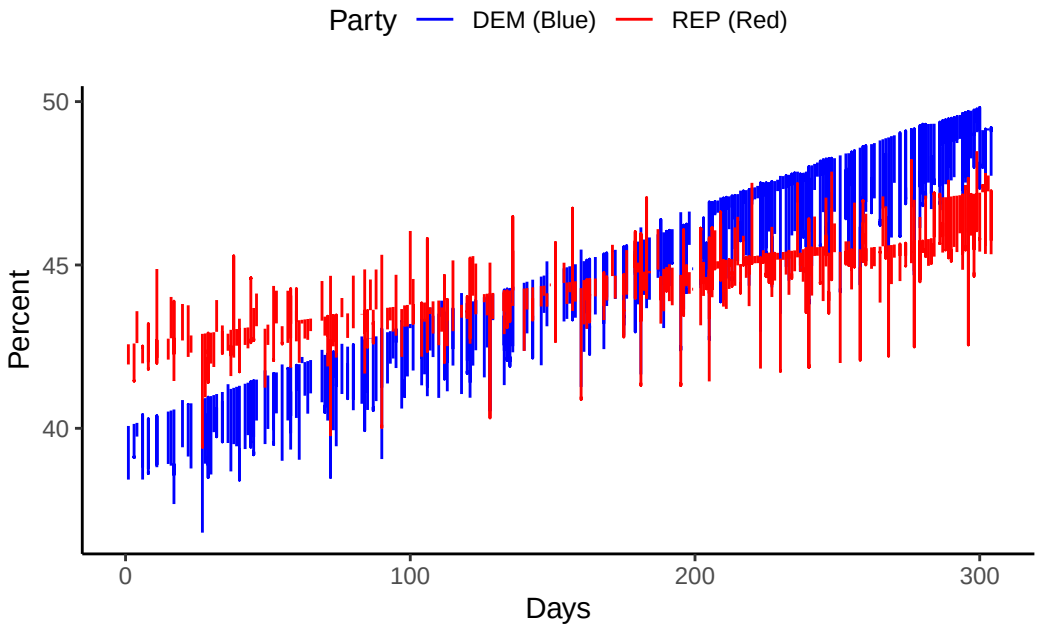

Figure 1: Prediction with MLR models for DEM and REP party

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \log(\text{sample\_size}_i) + \beta_2\, \text{pollscore}_i + \beta_3\, \text{numeric\_grade}_i \tag{3}$$

$$+ \beta_4\, \text{transparency\_score}_i + \beta_5\, \text{days\_since\_start}_i + \alpha_j \tag{4}$$

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

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

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

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

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

$$\beta_4 \sim \text{Normal}(0, 2.5) \tag{10}$$

$$\beta_5 \sim \text{Normal}(0, 2.5) \tag{11}$$

$$\sigma_{\text{pollster}} \sim \text{Exponential}(1) \tag{12}$$

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.

Table 2

	(1)
(Intercept)	−0.67
log(sample_size)	0.04
pollscore	−0.10
numeric_grade	−0.07
transparency_score	0.00
days_since_start	0.00
Sigma[pollster × (Intercept),(Intercept)]	0.01
Num.Obs.	1348
ICC	0.9
Log.Lik.	−11 224.203
ELPD	−11 500.4
ELPD s.e.	326.4
LOOIC	23 000.9
LOOIC s.e.	652.8
WAIC	23 018.0
RMSE	0.03

3.2.1 Result

Summary of Bayesian model is hsown in Table 2, and the ppcheck is in appendix Figure 5.

4 Prediction

To get a prediction by the Bayesian model we have, we first spline fit the model, and then create a data frame and generate posterior predictions by the new data frame.

4.1 Spline Fit for Bayesian model set-up

We run the model in R (R Core Team 2023) using the `rstanarm` package of (Goodrich et al. 2022). We use priors Normal$(0, 5)$ to allow for more flexibility of predictors effects with "autoscale = TRUE". Then we crate a data frame like Table 3, we set the quality scale of predict data to the best to perform a poll that has high quality. the result is shown as Section 4.2

Table 3

	pollscore	sample_size	numeric_grade	transparency_score	days_since_start	pollster
1	-1.5	1200	3	10	0.000000	TIPP
2	-1.5	1200	3	10	3.070707	TIPP
3	-1.5	1200	3	10	6.141414	TIPP
4	-1.5	1200	3	10	9.212121	TIPP
5	-1.5	1200	3	10	12.282828	TIPP
6	-1.5	1200	3	10	15.353535	TIPP

4.2 Result

The prediction graph Figure 2 contains prediction of spline fit model for two parties, blue line represents the DEM party's predicted percentage of support, and the red line represents the REP party's predicted percentage of support over time (measured in days since the start of 2024). We can tell that the predicted support of DEM party is taking the lead by aroud 2 percent until the last day of the dataset.

5 Discussion

5.1 Crossing point in the prediction

From the prediction of spline fit model Figure 2, There is a clear increase in the support for the DEM party starting around day 200, while the REP party's support remains more stable with only a slight increase over time. The increase in support for the DEM party in the predictions could be attributed to Harris's rise as the Democratic candidate after Biden's exit from the race On July 21, 2024, which is about 230 days from start of 2024.

5.2 Weaknesses and next steps

Weaknesses and next steps should also be included.

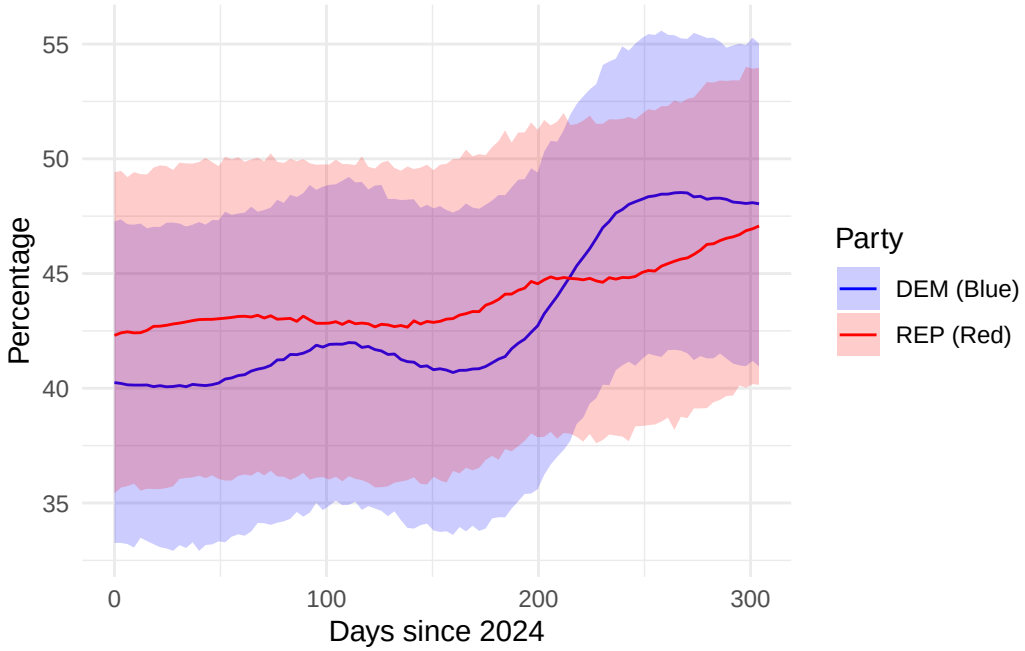

Figure 2: Poll Percentage over Time with Spline Fit for DEM and REP party

A Appendix

A.1 Idealized Methodology

To forecast the 2024 U.S. presidential election with a budget of $100,000, this methodology combines stratified sampling, multimodal recruitment, and aggregation techniques that leverage the strengths of multiple data sources. The approach employs stratified random sampling with targeted oversampling of key subgroups, like younger and minority voters, to address their frequent underrepresentation in polling. By segmenting eligible U.S. voters by demographics—age, gender, race, party affiliation, region, and urban/rural residency—the method ensures broad representational coverage, aiming to capture diverse voting patterns accurately (PewResearchCenter 2019) . Partnering with a voter database provider gives us access to a detailed list of registered voters, organized by demographics, to improve the accuracy of our sample. This approach also builds on public trust, as Americans generally trust polls from news organizations (43%) more than those from websites that combine multiple polls (30%) (Pasek 2015).

Respondents are recruited through multiple channels to reduce potential biases associated with any single mode. The multimodal approach includes Interactive Voice Response (IVR) to reach older and rural populations (30%), targeted online surveys to capture difficult-to-reach demographics (40%), text surveys directed at younger voters (20%), and live calls in key

battleground states (10%). This mix helps address "coverage error" by compensating for the limitations associated with each mode (Blumenthal 2014).

Survey data is collected via a short, 5–7 minute questionnaire with 10 core questions covering demographics, voting intention, likelihood of voting, and key issues. A shorter survey length and mix of direct and indirect questions help reduce Social Desirability Bias, while randomizing question order minimizes positional bias (). Given that any individual survey is likely to suffer from random and systematic errors—sampling errors, coverage errors, and response biases—aggregation across multiple polls can mitigate these sources of error, improving the accuracy of forecasts by averaging out errors specific to individual polls (Blumenthal 2014).

For data validation and quality control, screening questions confirm attentiveness, demographic cross-checks validate responses against voter registration data, and post-survey weighting adjusts for demographic imbalances. Weighting responses to match population demographics ensures that key groups are proportionally represented, while data cleaning removes low-quality responses to maintain data integrity (Blumenthal 2014). To further address biases, surveys are conducted anonymously in formats like online and SMS. Weighting adjustments are also made to address non-response bias, enhancing accuracy by reflecting the true population distribution. We will keep the process and result transparent by posting on our website for the public to view.

Weekly aggregation serves as a "poll of polls" that balances fluctuations across samples and smooths out errors. By incorporating recent data more heavily, this rolling average adapts to shifts in voter sentiment leading up to the election. Bayesian modeling, applied to smooth sample variations and incorporate prior election trends, stabilizes the forecast further. Poll aggregation, which is commonly employed in election analyses, effectively combines estimates from various samples, reducing random error and discounting non-universal biases (Blumenthal 2014). Given the variability in survey methodologies—such as IVR surveys missing cell-only populations or web surveys excluding offline individuals—aggregation allows errors in one type of survey to offset those in another. Weighting each survey according to sample size and precision further refines the accuracy of the aggregated forecast.

A.1.1 Budget Allocation:

- **Sample Acquisition**: Partnership with a voter database provider to access a list of registered voters across all states $10,000
- **Interactive Voice Response (IVR)**: Automated voice surveys for older and rural demographics $20,000
- **Online Surveys**: Targeted ads and opt-in digital surveys to capture harder-to-reach groups (40% of sample) $25,000
- **Text Messaging**: SMS-based surveys to engage younger demographics (20% of sample) $10,000

- **Live Phone Calls**: Calls in battleground states to reach underrepresented groups (10% of sample) $20,000
- **Data Validation & Cleaning**: Screening questions, demographic verification, and removal of low-quality responses to maintain data integrity $5,000
- **Data Analysis & Forecast Modeling** Bayesian modeling, poll aggregation, and analysis of survey trends to produce election forecasts $10,000

A.2 Idealized Survey

The proposed survey questionnarie design is in the following link: https://forms.gle/Zm5Kfj3kL58gwCz38

A.2.1 Survey Copy

1. **What is your age??**
    - 18-29
    - 30-44
    - 45-64
    - 65+
2. **What is your gender identity?**
    - Female
    - Male
    - Another Gender Identity
3. **Which of the following best describes your race/ethnicity?**
    - White
    - Black/African American
    - Hispanic/Latino
    - Asian
    - Native American
    - Prefer not to say
    - Other
4. **What is your highest level of education?**
    - High school or less
    - College
    - Bachelor's degree
    - Graduate degree
    - Prefer not to say

5. **In which U.S. region do you currently reside?**
    - Midwest
    - Northeast
    - South
    - West
6. **Are you registered to vote in the 2024 presidential election?**
    - Yes
    - No
7. **2024 Which candidate do you plan to vote for?**
    - Kamala Harris
    - Donald Trump
    - Undecided
    - Other
8. **2020 Which candidate did you vote for?**
    - Joe Biden
    - Donald Trump
    - Did not vote
    - Prefer not to say
    - Other
9. **What is the most important issue influencing your vote?**
    - Economy
    - Healthcare
    - Immigration
    - Climate Change
    - Social Justice
    - National Security
    - Other
10. **On a scale of 1-5, how likely are you to vote in the 2024 U.S. presidential election?**

- 1 being Not Likely to Vote
- 5 being Likely to Vote

A.3 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, live phone calls, text messages, emails, digital dial back interface and online targeted opt-in digital survey platforms (The Trafalgar Group 2024). Trafalgar places particular emphasis on reducing Social Desirability Bias, designing short questionnaires and using nontraditional question formats to help participants feel more comfortable expressing their true preferences, especially on sensitive topics (The Trafalgar Group 2024). They also adjust for non-response by weighting the sample, ensuring it reflects the broader population's demographic composition. While this helps to reduce bias from underrepresented groups, reliance on post-survey adjustments can introduce new biases, especially if response rates vary significantly among demographic groups.

B Model details

B.1 Assumption check for MLR models

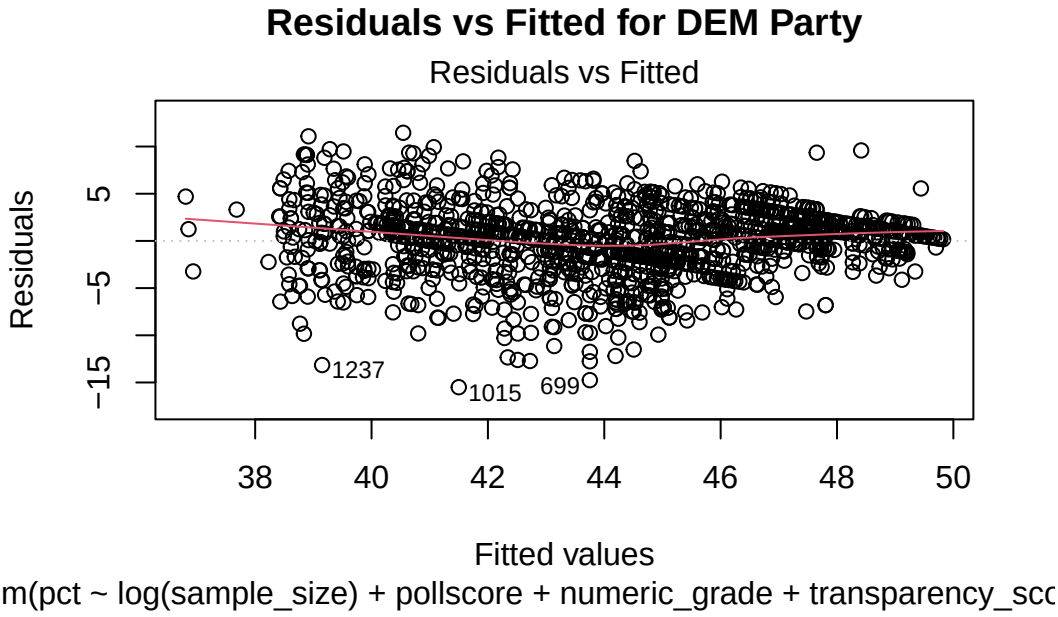

Figure 3: Residuals vs Fitted for DEM Party

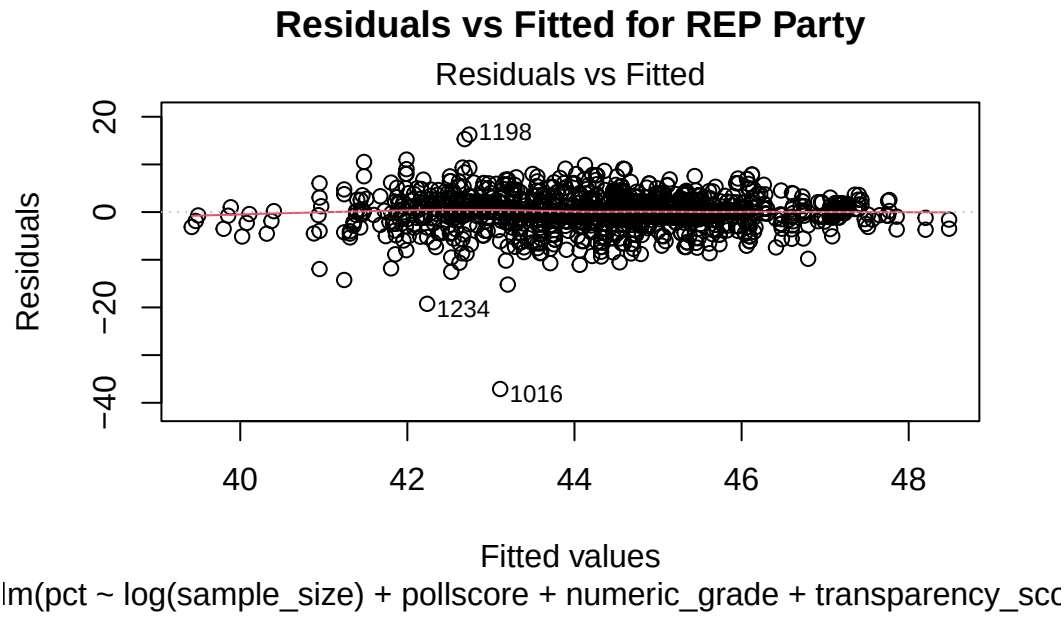

Figure 4: Residuals vs Fitted for REP Party

B.2 Significant check for MLR models

see Table 4, and Table 5

Table 4: Summary of DEM 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
-15.4994  -1.8873   0.4338   1.9556  11.4563

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       37.146364   1.460673  25.431  < 2e-16 ***
log(sample_size)   0.409551   0.159944   2.561   0.0106 *
pollscore         -0.707759   0.349179  -2.027   0.0429 *
numeric_grade      0.122160   0.448527   0.272   0.7854
transparency_score -0.315843   0.059923  -5.271 1.58e-07 ***
days_since_start   0.031532   0.001149  27.448  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.645 on 1342 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.3946,    Adjusted R-squared:  0.3923
F-statistic: 174.9 on 5 and 1342 DF,  p-value: < 2.2e-16
```

Table 5: Summary of 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
-37.111  -1.287  -0.072   1.798  16.258

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        50.183139   1.422464  35.279  < 2e-16 ***
log(sample_size)   -0.247843   0.155760  -1.591    0.112
pollscore          -3.325025   0.340045  -9.778  < 2e-16 ***
numeric_grade      -3.548531   0.436794  -8.124 1.01e-15 ***
transparency_score  0.059077   0.058356   1.012    0.312
days_since_start    0.012280   0.001119  10.977  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.55 on 1342 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.1505,    Adjusted R-squared:  0.1473
F-statistic: 47.54 on 5 and 1342 DF,  p-value: < 2.2e-16
```

B.3 Posterior predictive check

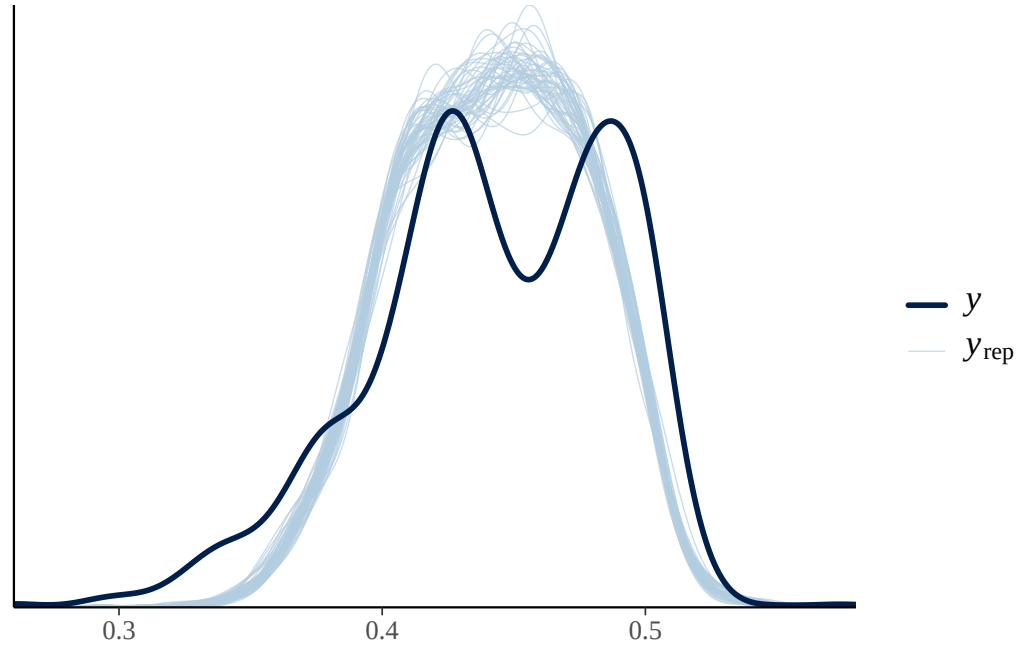

Figure 5: Posterior prediction check for Bayesian model

B.4 Diagnostics

Checking the convergence of the MCMC algorithm

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Blumenthal, Mark. 2014. "Polls, Forecasts, and Aggregators" 47 (2): 297–300. https://doi.org/10.1017/S1049096514000055.

Data, Telling Stories with. 2024. "Telling Stories with Data." https://tellingstorieswithdata.com/.

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

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

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/.

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

Pasek, Josh. 2015. "Predicting Elections: Considering Tools to Pool the Polls." *Public Opinion Quarterly*. https://academic.oup.com/poq/article/79/2/594/2277466?login=true#84746216.

PewResearchCenter. 2019. "A Field Guide to Polling: Election 2020 Edition." *Pew Research Center*. https://www.pewresearch.org/methods/2019/11/19/a-field-guide-to-polling-election-2020-edition/#how-can-you-tell-a-good-poll-from-a-bad-one.

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

The Trafalgar Group. 2024. "Polling Methodology." https://www.thetrafalgargroup.org/polling-methodology/.

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.