2024 USA Election Prediction*

Ruiyang Wang

October 19, 2024

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

1 Introduction

The 2024 U.S. Presidential election is approaching, and understanding the factors that influence polling results is crucial for predicting the election outcome. This paper uses polling data to predict the support percentages for the top 5 candidates leading up to November 5th, 2024. We first explore the relationship between sample size and support percentage using simple linear regression (SLR). We then employ multiple linear regression (MLR) to model the influence of various polling factors and culminate in a prediction graph for each candidate's support. Our goal is to provide an evidence-based prediction of the likely election result.

1.1 Estimand

Our primary estimand is the predicted support percentage for each candidate based on polling data. We aim to quantify how polling factors such as sample size, poll scores, and transparency scores affect the predicted support percentage.

1.2 Structure of the Paper

The remainder of this paper is structured as follows: - Section Section 2 describes the dataset and key variables. - Section Section 3 presents the simple linear regression (SLR) model used to explore the effect of sample size on support percentage. - Section Section 4 presents the multiple linear regression (MLR) model used for predictions. - Section Section 5 provides the key predictions leading up to the election.

*Code and data are available at: [https://github.com/Ruiyang-Wang/STA304-Paper2.git).

2 Data

2.1 Overview

The data used for this analysis includes polling results for the top 5 candidates in the 2024 U.S. Presidential election. Key variables include: - **Sample Size**: The number of respondents in each poll. - **Support Percentage (pct)**: The percentage of respondents supporting a particular candidate. - **Poll Score**: A measure of the reliability of the pollster. - **Transparency Score**: An indicator of how transparent the pollster is about their polling methods.

2.2 Measurement

Each polling result is recorded with the above variables, allowing us to investigate the relationship between polling factors and the support percentage of each candidate. This analysis focuses on the top 5 most frequently polled candidates.

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 dependent variable in both the SLR and MLR analyses.

Some of our data is from FiveThirtyEight (2024)

3 Simple Linear Regression (SLR): Verification of Sample Size Effect

3.1 Overview

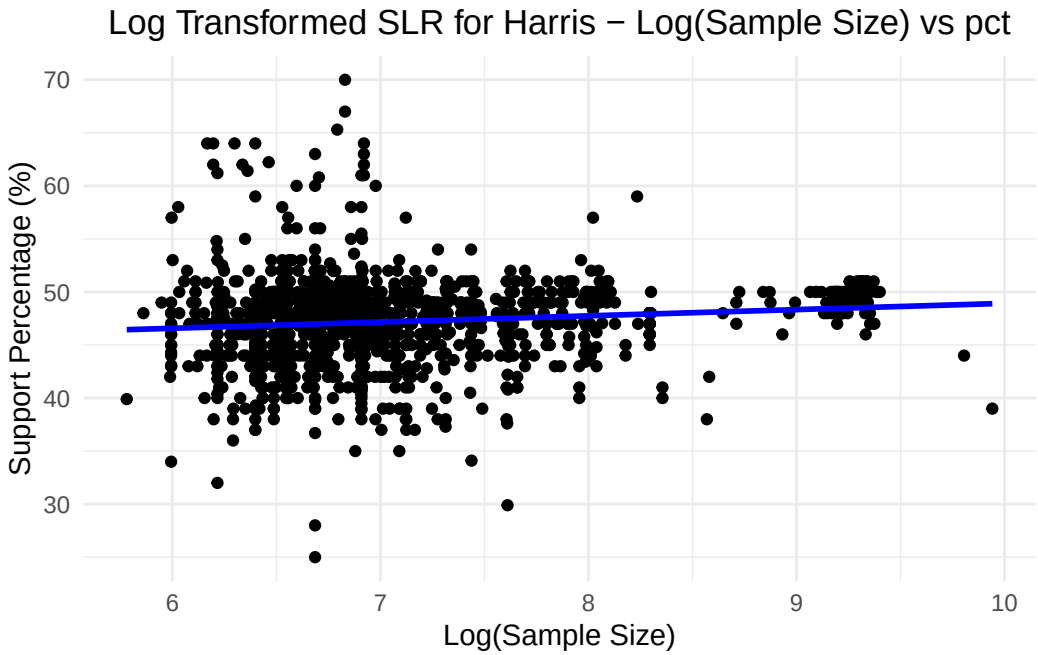
In this section, we verify whether sample size has a significant impact on the support percentage of the candidates. Simple linear regression (SLR) is used to examine the relationship between the sample size and the percentage of support for each of the top 5 candidates.

3.2 Model Overview

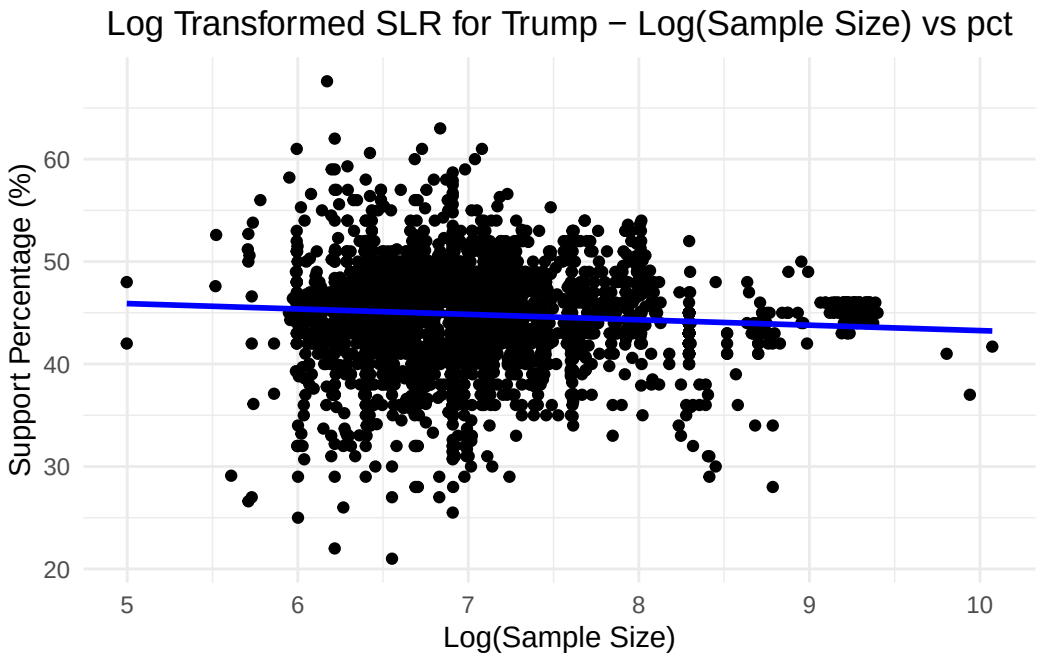
In the SLR model, the **dependent variable** is the support percentage (pct), and the **independent variable** is the sample size. A log-transformation is applied to the sample size to account for its wide range across different polls.

3.3 Model Overview

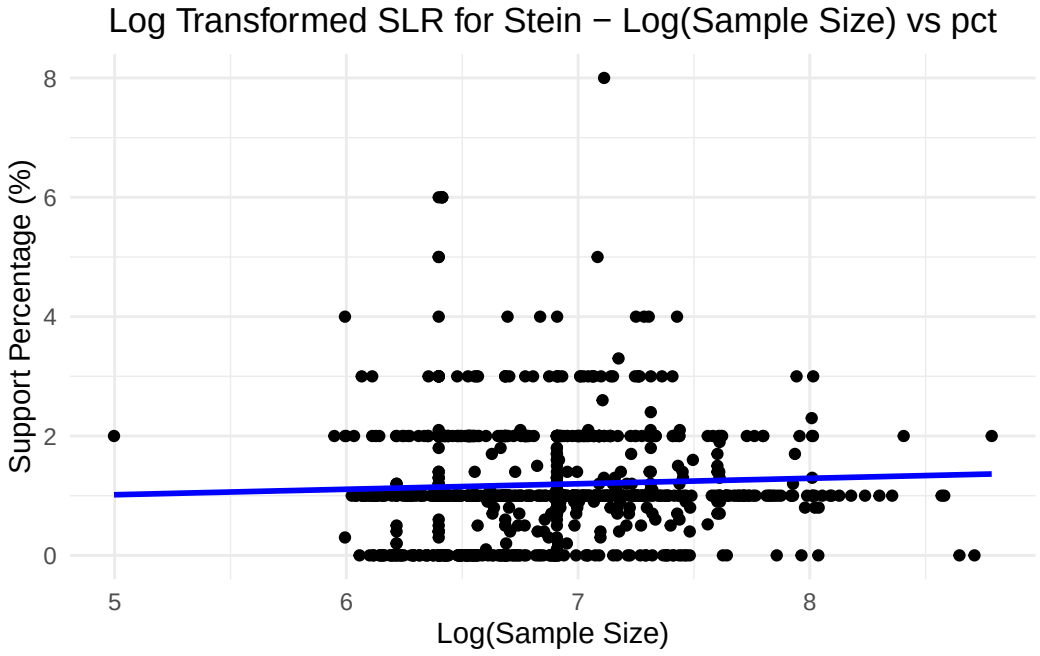
```
`geom_smooth()` using formula = 'y ~ x'
```



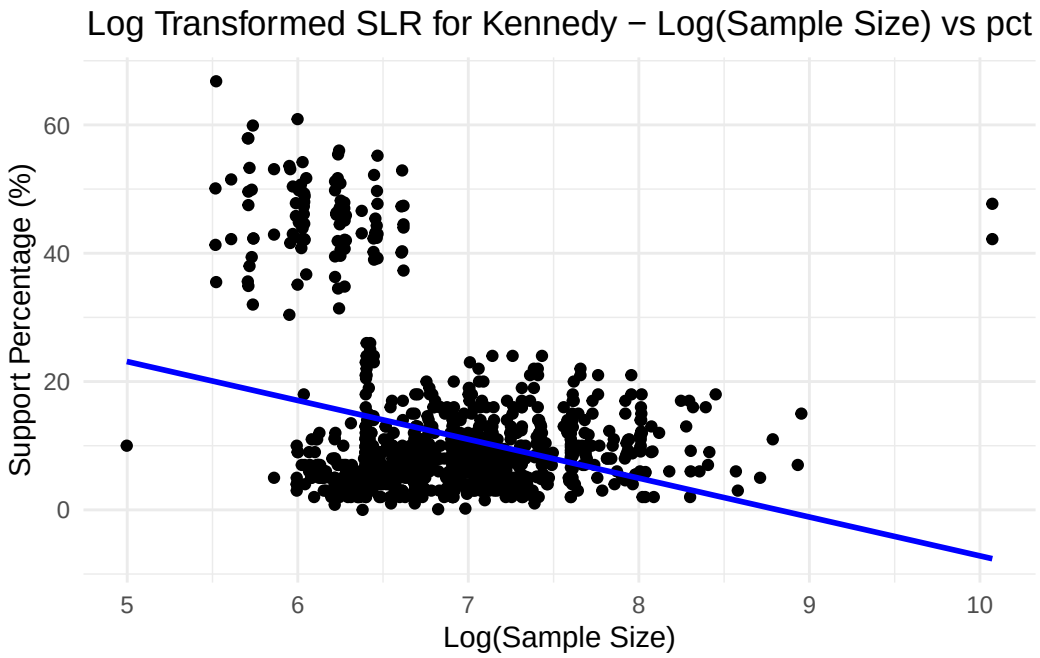

```
`geom_smooth()` using formula = 'y ~ x'
```



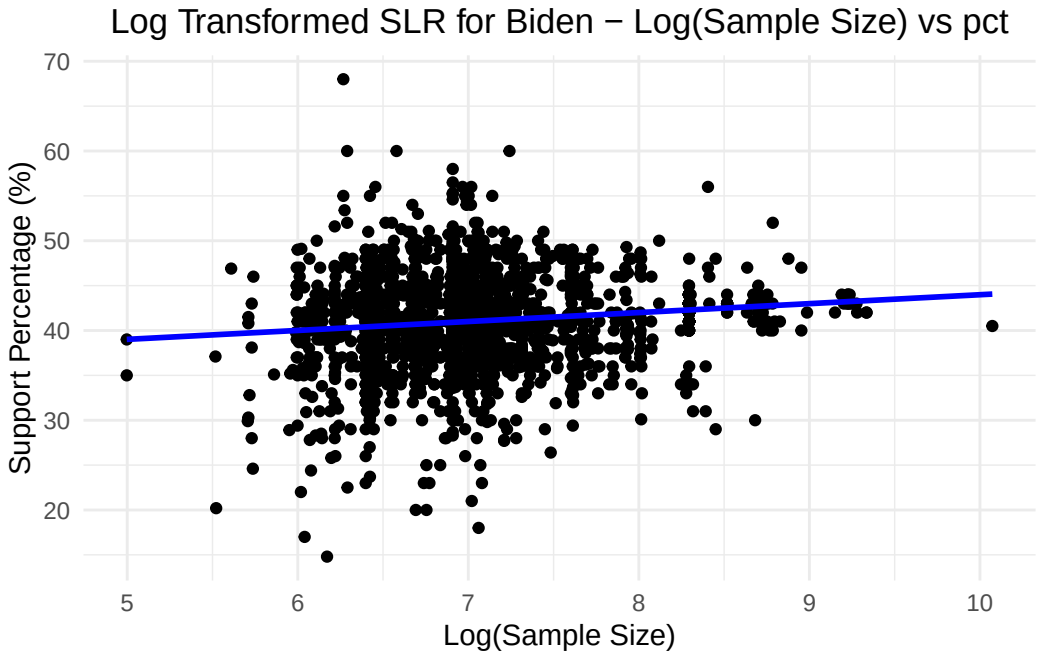
`geom_smooth()` using formula = 'y ~ x'



`geom_smooth()` using formula = 'y ~ x'

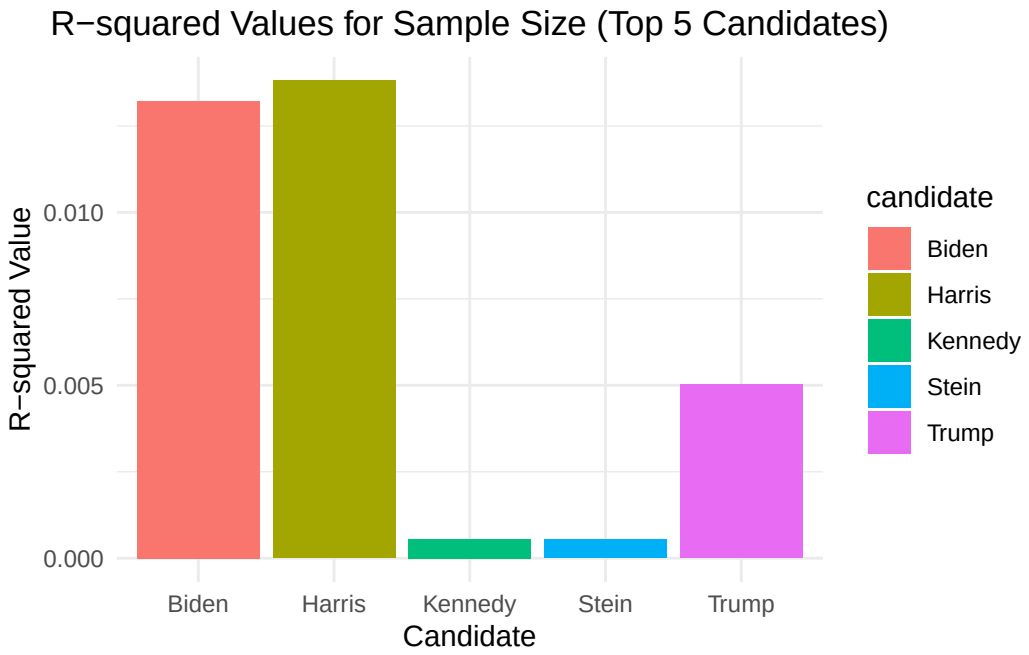



```
`geom_smooth()` using formula = 'y ~ x'
```



In the SLR model, the **dependent variable** is the support percentage (pct), and the **independent variable** is the sample size. A log-transformation is applied to the sample size to account for its wide range across different polls.

3.4 Result



For each candidate, we analyze whether sample size significantly influences their support percentage. The results will guide us in assessing whether sample size should be further considered in the multiple linear regression (MLR) models.

4 Multiple Linear Regression (MLR): Predicting Support Percentages

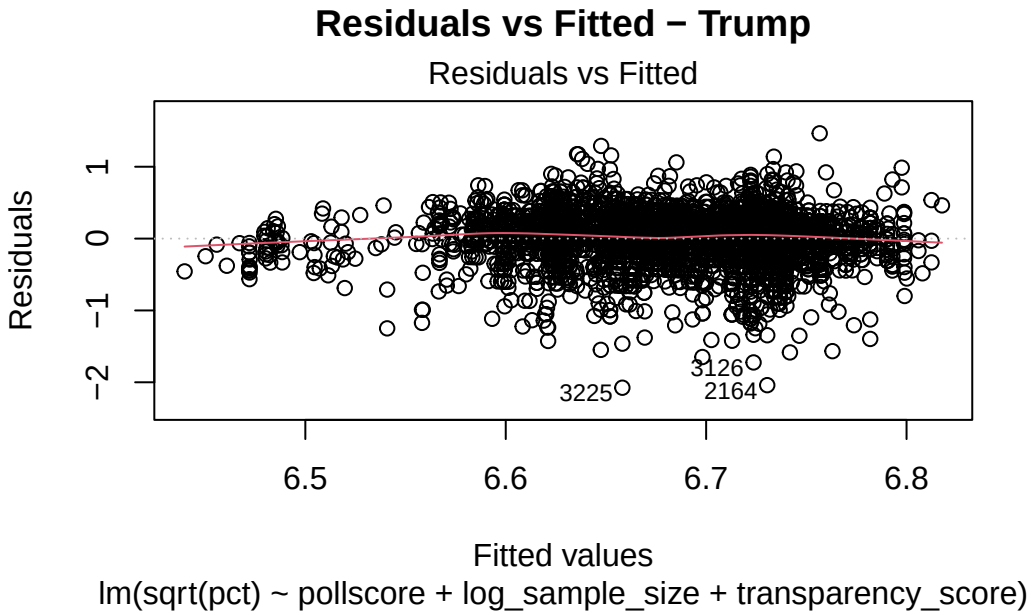
4.1 Overview

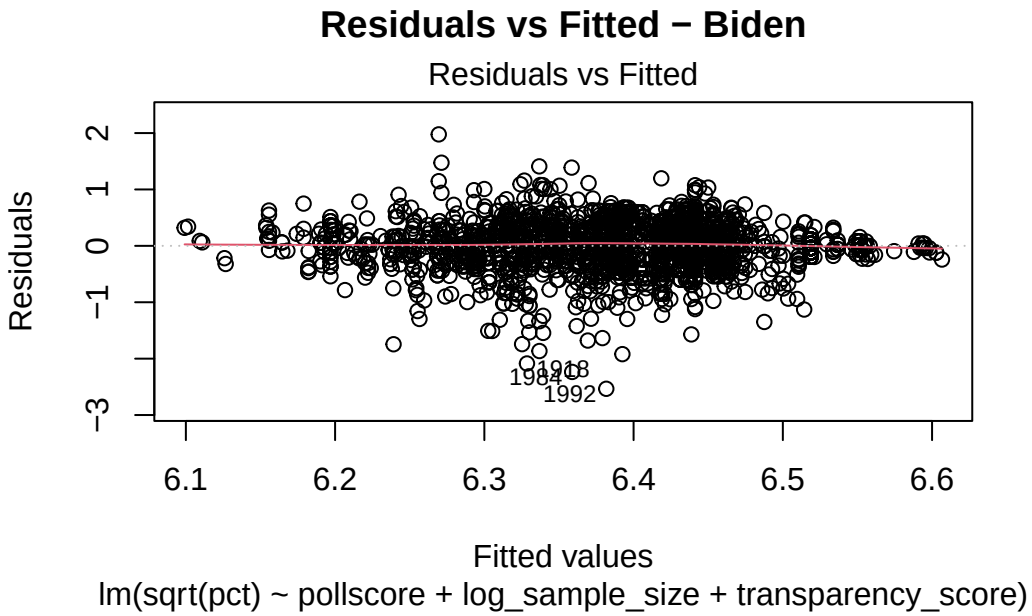
After examining the effect of sample size on support percentage through SLR, we now introduce the multiple linear regression (MLR) model. This model incorporates additional variables, such as poll score and transparency score, to predict the support percentage for each candidate.

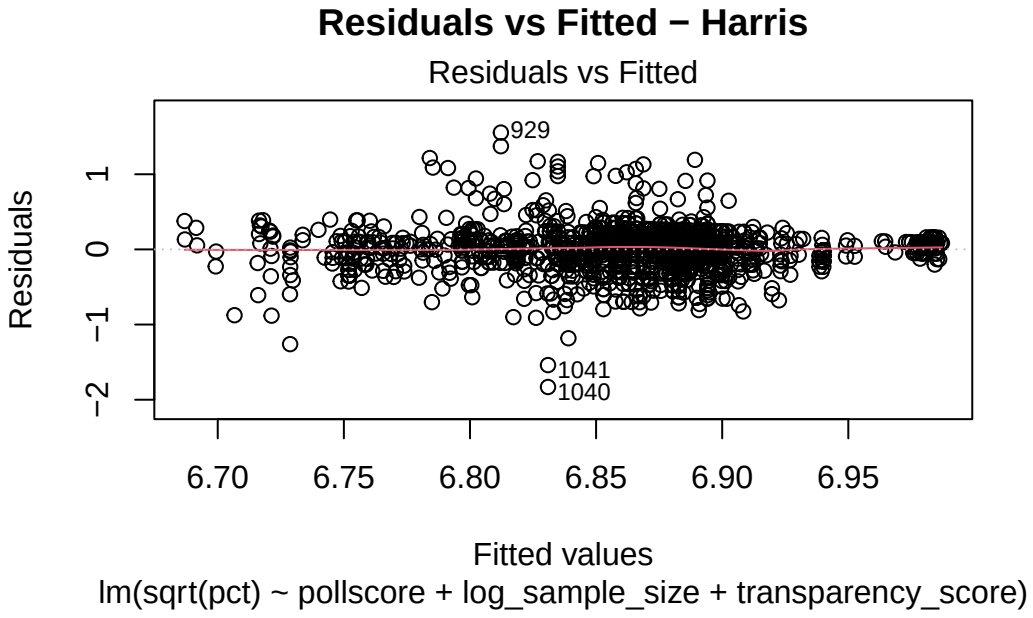
4.2 Model Overview & Residual Analysis

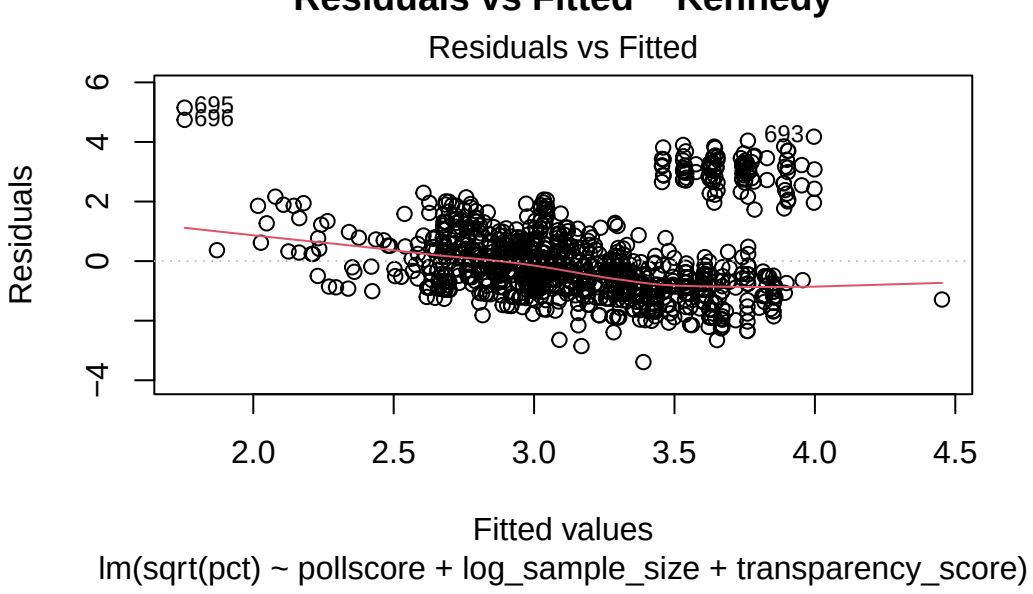
The MLR model includes: - **Dependent Variable**: Support Percentage (sqrt-transformed) - **Independent Variables**: Poll Score, Log-transformed Sample Size, and Transparency Score.

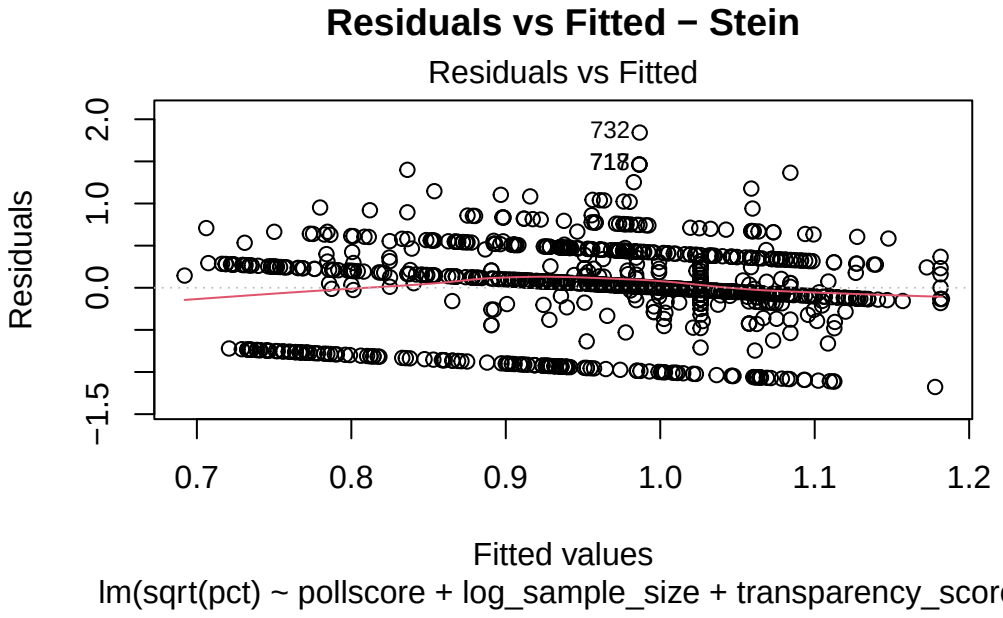
The residual analysis is used to evaluate the model's fit, helping us assess the quality of the predictions.






Residuals vs Fitted − Harris
Residuals vs Fitted
Residuals
929
1041
1040
6.70
6.75
6.80
6.85
6.90
6.95
Fitted values
lm(sqrt(pct) ~ pollscore + log_sample_size + transparency_score)


Residuals vs Fitted − Kennedy
Residuals vs Fitted
Residuals
695
696
693
2.0
2.5
3.0
3.5
4.0
4.5
Fitted values
lm(sqrt(pct) ~ pollscore + log_sample_size + transparency_score)



We examine residuals vs. fitted plots to assess the fit of the MLR models. This step ensures that the model assumptions are met and that the MLR models adequately predict the support percentages based on polling data.

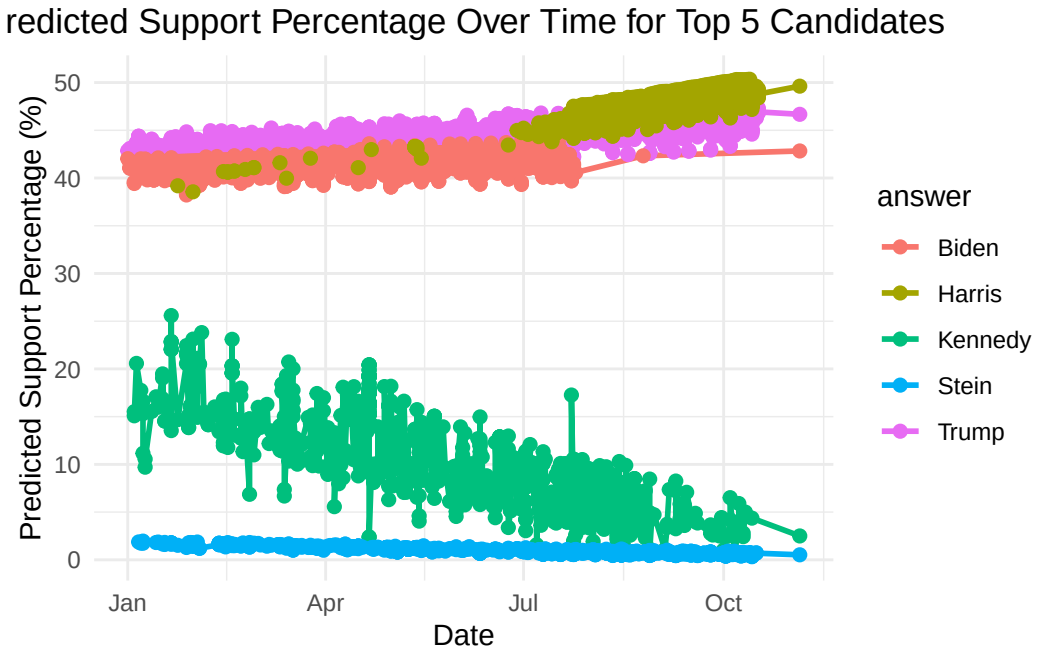
5 Prediction of Support Over Time

5.1 Model Overview

In this section, we present the predicted support percentages for each candidate using the MLR model. We focus on predictions leading up to November 5th, 2024, the election day. The model incorporates poll score, sample size, and transparency score, as well as the date of the poll, to predict future support trends.

5.2 Predictions

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



For each candidate, we project their likely support percentage over time, visualizing how their polling support is expected to evolve as the election date approaches. These predictions offer insights into which candidate is likely to gain the most support based on the trends observed in the polling data.

6 Conclusion

This paper presents a data-driven prediction of the 2024 U.S. Presidential election outcome. Through simple linear regression (SLR), we verified that sample size does influence support percentages for some candidates. Then, using multiple linear regression (MLR), we predicted the support percentages for the top 5 candidates up to November 5th, 2024. The MLR models show that poll score and transparency score are important factors in predicting support percentages. Future research could explore additional polling factors or include regional breakdowns to further refine the predictions.

FiveThirtyEight. 2024. "2024 Wisconsin Presidential General Election Polls." https://projects.fivethirtyeight.com/polls/president-general/2024/wisconsin/.