2024 USA Presidential Election Prediction*

Ruiyang Wang

November 1, 2024

This article analyzes the polling data from the 2024 U.S. presidential election to estimate the likely winner and understand the factors that influence candidate support. Using simple and multiple linear regression models, we examined the effect of variables such as sample size, poll scores, and transparency on the percentage of support. The analysis showed that while the effect of sample size was limited, highly transparent polls were associated with an increase in candidate support, highlighting the impact of poll quality on public perception. Based on our model, because of the continued upward trend in support, we expect Donald Trump to be the likely front-runner. The paper also highlights the potential and limitations of statistical modeling in election prediction, enhancing our understanding of how voting methods shape public opinion in high-stakes elections.

1 Introduction

The 2024 U.S. presidential election has attracted the world-wide attention, from economic uncertainty to social policy and global affairs. As the election proceeds, polling data has become a key tool for understanding public sentiment and predicting potential outcomes. However, it is challenging to accurately interpret the data due to a variety of factors that can influence poll results. For example, sample size, pollster methodology, and the demographics of respondents. This article aims to provide an analysis of polling data(FiveThirtyEight (2024)) only based on statistical methods.

To construct this analysis, we first explored how the sample size affected the support of the top five candidates in this election - Joe Biden, Kamala Harris, Donald Trump, Robert Kennedy Jr., and Jill Stein. Given that some candidates, notably Trump and Harris, have larger voter bases, it is necessary to determine whether changes in sample size are correlated with fluctuations in candidate support. Using sample sizes for simple linear regression (SLR) and logarithmic transformations to address model assumptions violations, we confirm that while

---

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

sample sizes do affect support percentages, the bias will not strong when we apply it to the multiple linear regression model.

After the simple regression analysis, we move on to build multiple linear regression (MLR) models for each of top five candidate. Combining multiple predictor variables, these MLR models allow us to predict support percentage over time, and a visual representation of these projections shows trends throughout the election period.

1.1 Estimand

The estimates in our MLR analysis are regression coefficients for each predictor - poll score, logarithmic conversion sample size, and transparency score. These coefficients represent the estimated impact of each predictor on the percentage of support after the square root conversion, while the other predictors remain constant.

Coefficient for poll score: This estimand quantifies the effect of poll scores on approval ratings, controlling for sample size and transparency. The positive coefficient suggests that high-quality polls (indicated by higher poll scores) are associated with higher support for candidates, possibly due to more accurate methodology.

Coefficient for Log-transformed sample size: This estimand captures the effect of sample size (in its log-converted form) on percentage of support, controlling for poll scores and transparency. The logarithmic transformation accounts for less violations for statistical model assumptions. The positive coefficient indicates that a larger sample size tends to produce higher candidate support, possibly due to reduced sampling error.

Coefficient for transparency score: This estimand represents the effect of transparency onoverall support rating, controlling for poll scores and sample size. A positive coefficient means that polls with higher transparency scores tend to report higher levels of support, suggesting that greater methodological transparency is associated with higher candidate support.

These estimands provide insight into how each predictor independently affects the percentage of support after conversion. By examining these coefficients, we can better understand the role of poll quality, sample size, and transparency in shaping reported levels of voter support, and thus gain a fuller picture of the factors that influence public sentiment represented in poll data.

1.2 Structure of the Paper

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

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023)to generate analysis. Our data (FiveThirtyEight 2024) provides a detailed look at voter support trends for the candidates in 2024 US election. By exploring this data, we aim to gain insights into how sample size and polling quality impact reported support percentages.To enhance our approach, we follow principles from Telling Stories with Data (Alexander 2023), which emphasizes clear and effective communication of data insights.We employ a variety of visualization techniques and statistical models to ensure that our findings are accessible and well supported. With these approaches, we highlight the nuances of poll data analysis, particularly in terms of sample size effects and model accuracy, ultimately helping to get a clearer picture of voter sentiment as we approach the 2024 election.

2.2 Measurement

The variables in this dataset are designed to measure specific aspects of public opinion and voting quality, allowing us to accurately analyze voter support trends. The main variable, the pct, represents the support rate of each candidate and serves as an indicator of public opinion at the time of each poll. This measurement assumes that the sample population accurately reflects the broader voting population, and that a higher percentage indicates a candidate is more popular.

To ensure accuracy, the dataset includes sample_size, which affects the representativeness of the poll. A larger sample size usually results in a more reliable percentage of support, which minimizes the margin of error. We performed a logarithmic transformation of sample_size in our analysis to stabilize the variance and satisfy the assumptions of the regression model.

Research indicates that transparency in survey methodologies significantly enhances the credibility and trustworthiness of poll results(GeoPoll (2022)). Variables related to quality, such as transparency_score and pollscore, give insights into the reliability of each poll. These scores measure the transparency with which each poll makes its methodology public, with higher scores indicating greater reliability. This transparency is important because polls that employ high-quality methodologies are less likely to introduce bias into our analysis.

However, we acknowledge the potential limitations of these measurements. Differences in survey methods, regional bias in sampling, and variability in public opinion affect the validity and reliability of these variables. Despite these challenges, the dataset provides a comprehensive view of trends in public opinion, allowing us to explore sample size effects and candidate support with reasonable confidence.

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 Verification of Sample Size Effect

3.1 Introduction

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.
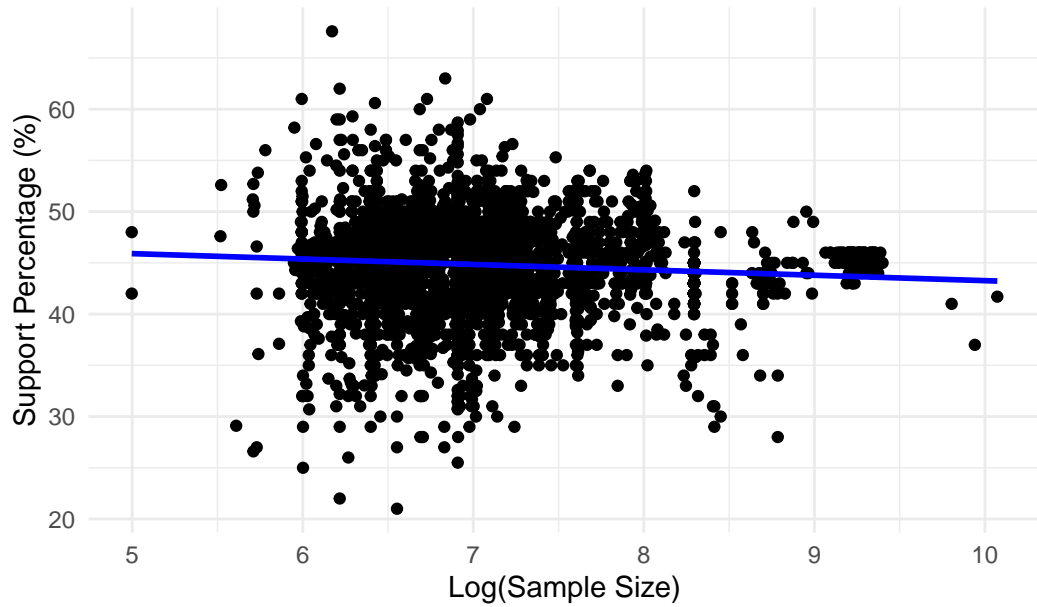
3.2 Model Overview

In the SLR model, the **dependent variable** is the support percentage (pct), and the **predictor variable** is the sample size. A log-transformation is applied to the sample size to account for the less violation for statistical model assumptions.
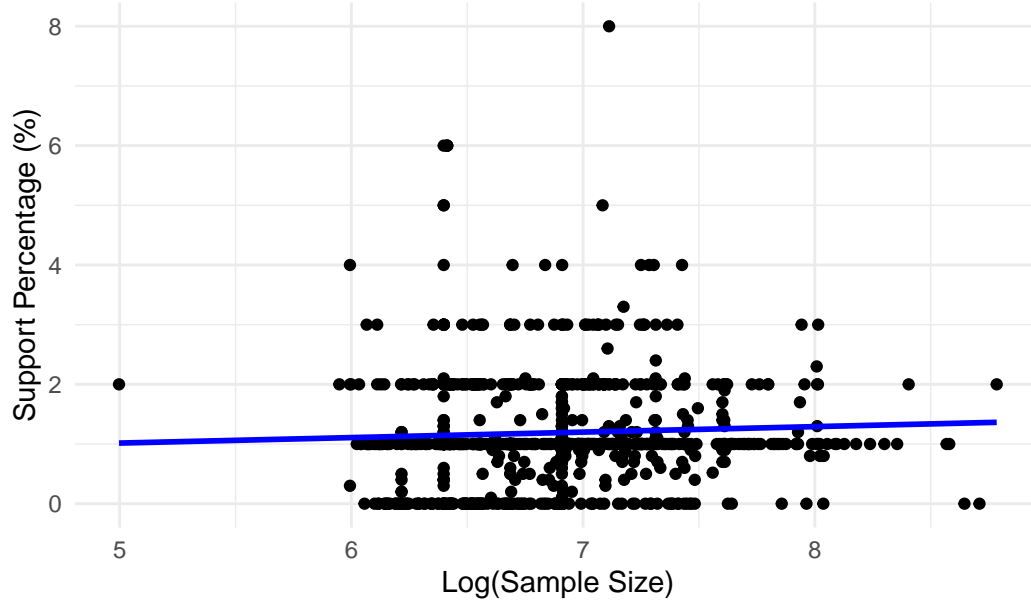
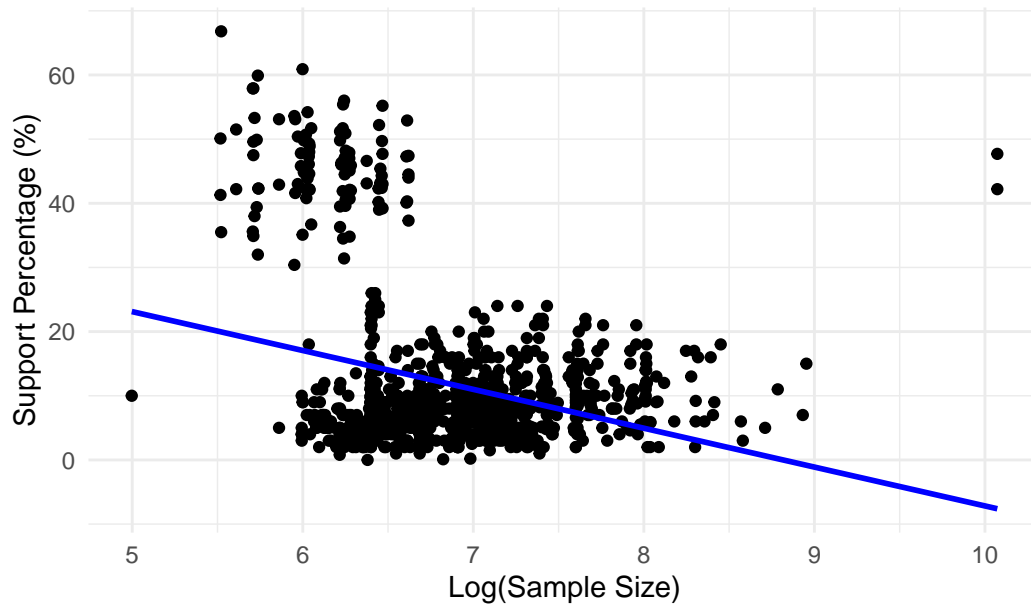## Log Transformed SLR for Harris – Log(Sample Size) vs pct
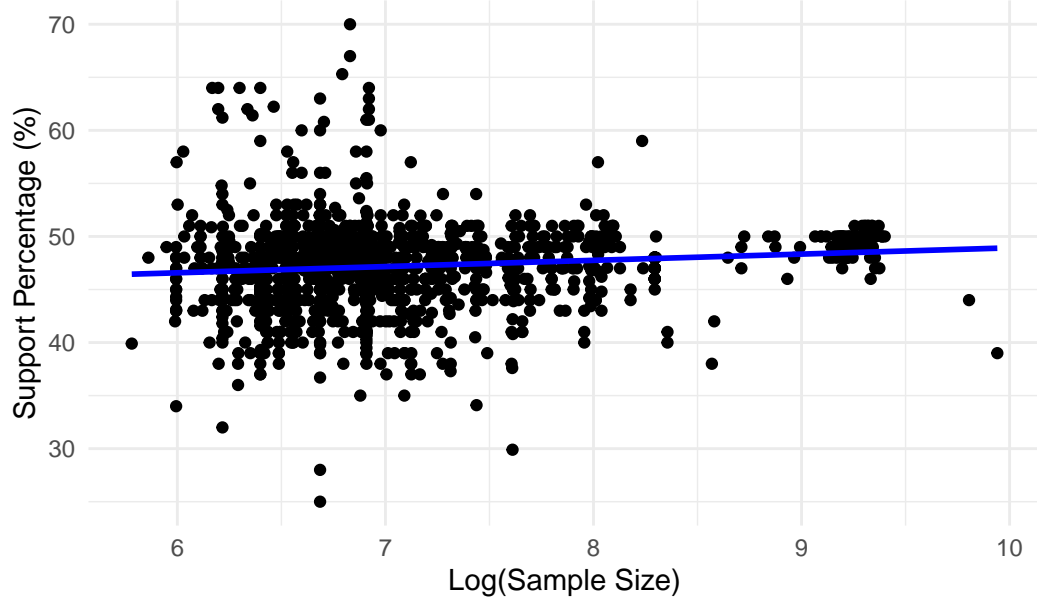
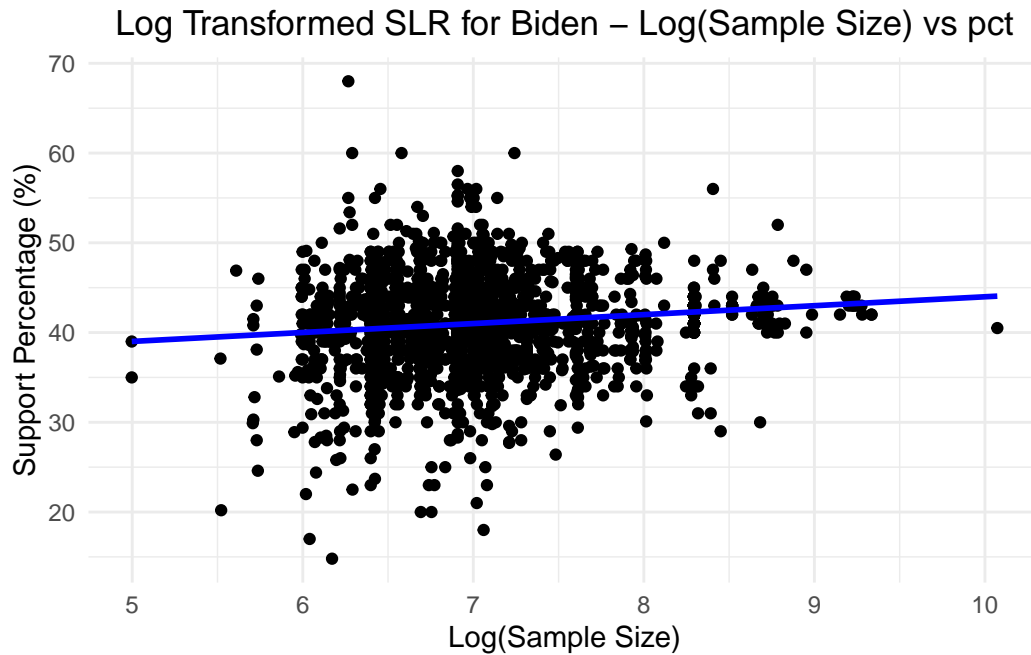## Log Transformed SLR for Trump – Log(Sample Size) vs pct

Log Transformed SLR for Stein – Log(Sample Size) vs pct

Log Transformed SLR for Kennedy – Log(Sample Size) vs pct

Log Transformed SLR for Biden – Log(Sample Size) vs pct

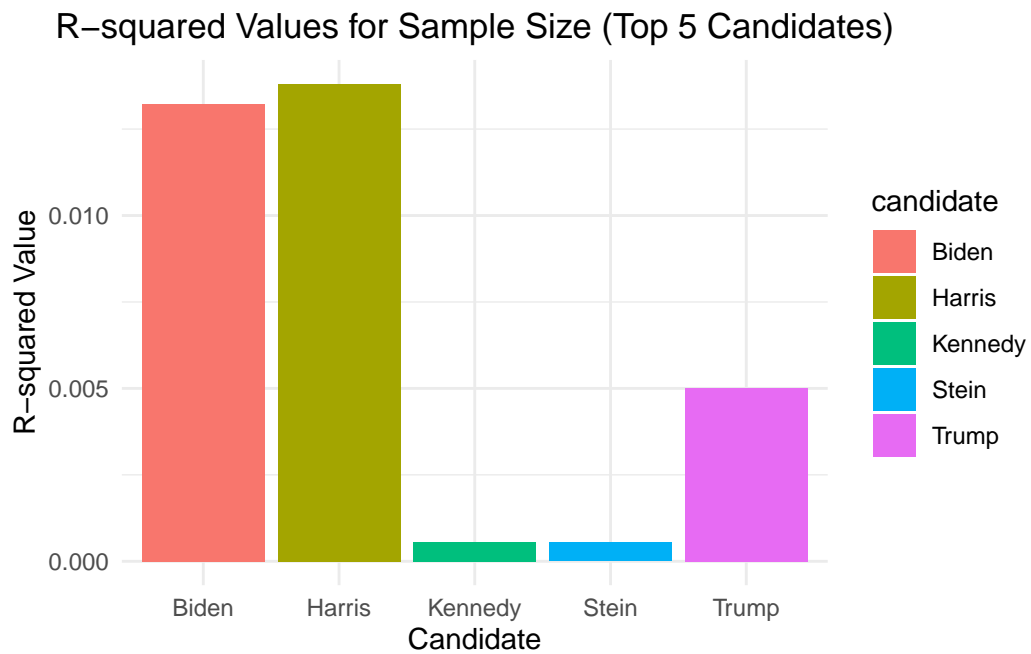## 3.3 Result

R−squared Values for Sample Size (Top 5 Candidates)

The graph shows the R-squared value for the simple linear regression model that examines the relationship between the log-transformed sample size and the percentage of support for

the top five candidates. As the chart shows, all of the candidates have low R-squared values, with Biden and Harris having the highest R-squared values, both around 0.04. Low R-squared values across all five candidates suggest that the sample size alone has limited power to explain the percentage of support. This finding suggests that sample size may have less bias when we add it into our further multiple linear regression model.

4 Regression Analysis of Key Predictors

4.1 Introduction

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

4.2 Model Overview & Residual Analysis

Those 5 MLR models include:
- **Outcome Variable**: Support Percentage (square root transformed)
- **Predictor Variables**: Poll Score, Log-transformed Sample Size, and Transparency Score.

These models are designed to evaluate the statistical relationship between outcomes and predictor variables, providing insight into the relative impact of each predictor on different candidate datasets. As shown below, the coefficients for each model reveal their statistical significance, highlighting patterns that may help explain changes in candidate support.

```
### Coefficients for Model 1
                    Estimate  Std. Error    t value      Pr(>|t|)
(Intercept)       7.09653546 0.063779076 111.267455 0.000000e+00
pollscore        -0.06503249 0.009299695  -6.992970 3.208904e-12
log_sample_size  -0.04673309 0.008183427  -5.710699 1.218827e-08
transparency_score -0.01757029 0.002489835  -7.056810 2.044311e-12


### Coefficients for Model 2
                    Estimate  Std. Error   t value      Pr(>|t|)
(Intercept)       5.85450463 0.099438282 58.875762 0.000000e+00
pollscore        -0.10407955 0.013709468 -7.591800 4.640748e-14
log_sample_size   0.08074150 0.012545870  6.435704 1.505422e-10
transparency_score -0.01223066 0.003759518 -3.253252 1.158288e-03
```

```
### Coefficients for Model 3
                       Estimate Std. Error    t value     Pr(>|t|)
(Intercept)          6.596174650 0.07859785 83.923097 0.000000e+00
pollscore           -0.063082028 0.01191356 -5.294977 1.381901e-07
log_sample_size      0.042564277 0.01022819  4.161465 3.356051e-05
transparency_score  -0.009354258 0.00312811 -2.990386 2.835223e-03


### Coefficients for Model 4
                       Estimate Std. Error    t value     Pr(>|t|)
(Intercept)          7.10864426 0.47841073 14.858873 1.300388e-45
pollscore            0.13506373 0.05310677  2.543249 1.112010e-02
log_sample_size     -0.49241707 0.07153946 -6.883154 9.865868e-12
transparency_score  -0.07865696 0.01786561 -4.402701 1.174374e-05


### Coefficients for Model 5
                       Estimate Std. Error    t value     Pr(>|t|)
(Intercept)          0.23448155 0.23594811  0.9937844 3.205847e-01
pollscore           -0.08192730 0.02241477 -3.6550584 2.714285e-04
log_sample_size      0.13548041 0.03477032  3.8964382 1.045681e-04
transparency_score  -0.03349813 0.00697395 -4.8033220 1.816474e-06
```
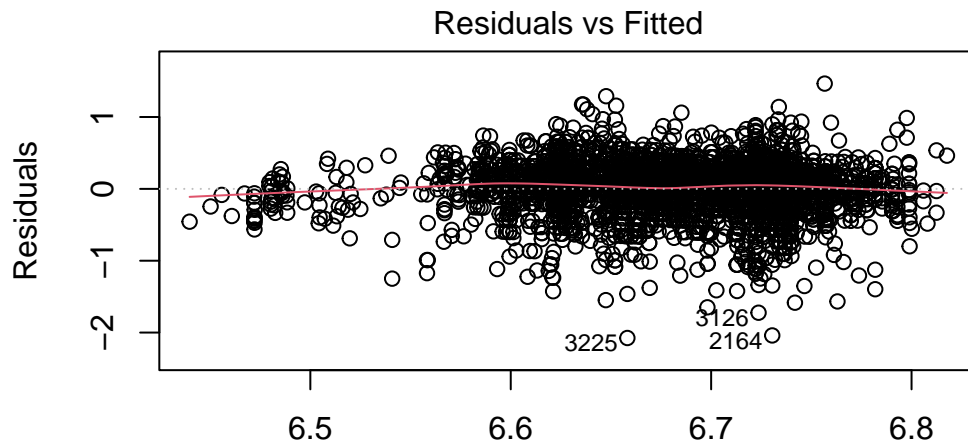
We observe that for each top 5 candidates, pollscore, log_sample_size, and transpanrency_score are all statistically significant to the response variable. Thus, we might proceed on further Residual vs. Fitted graph constructing.

Residuals vs Fitted – Trump
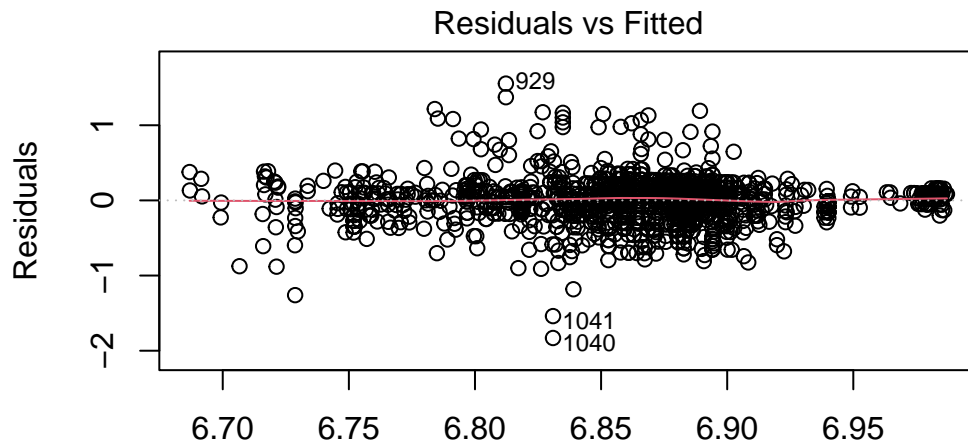
Residuals vs Fitted

Fitted values
lm(sqrt(pct) ~ pollscore + log_sample_size + transparency_score)

Residuals vs Fitted – Biden

Residuals vs Fitted
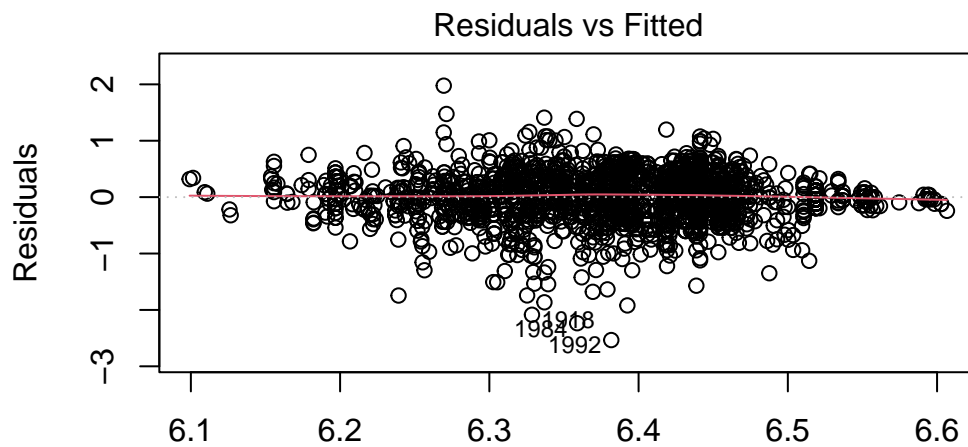
Fitted values
lm(sqrt(pct) ~ pollscore + log_sample_size + transparency_score)

Residuals vs Fitted – Harris

Residuals vs Fitted

Fitted values
lm(sqrt(pct) ~ pollscore + log_sample_size + transparency_score)

Residuals vs Fitted – Kennedy

Residuals vs Fitted

Fitted values
lm(sqrt(pct) ~ pollscore + log_sample_size + transparency_score)

Residuals vs Fitted – Stein

Residuals vs Fitted
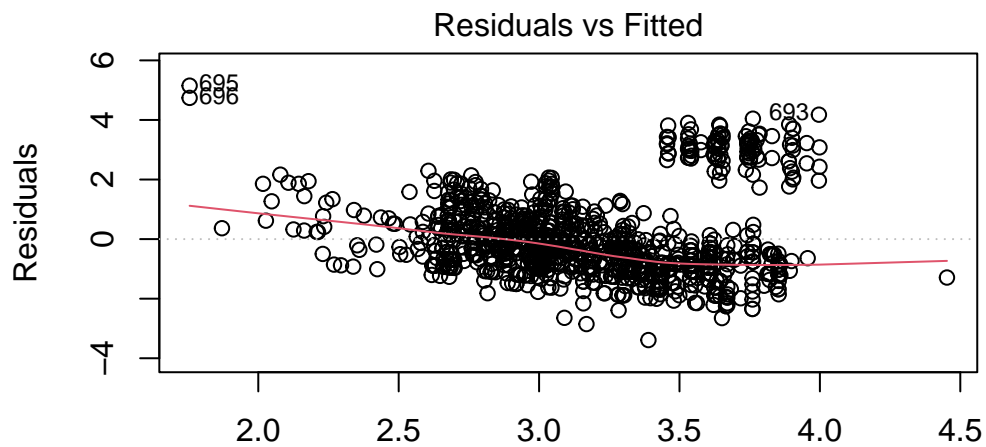
Fitted values
lm(sqrt(pct) ~ pollscore + log_sample_size + transparency_score)
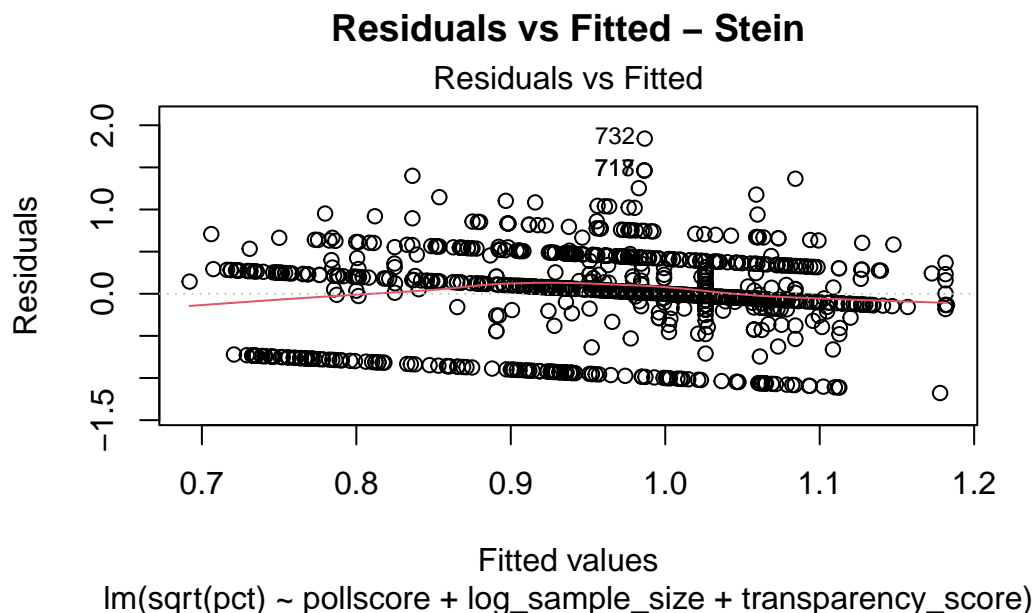
For Trump, Biden and Harris's graph, the residuals are generally centered around zero, and there is no clear pattern, indicating that the model fits well. The distribution of residuals is relatively consistent among the fitted values, indicating that the model satisfies the mean square error hypothesis without any obvious signs of heteroscedasticity. While there are a few standout points, these outliers do not appear to significantly affect the distribution of residuals or the fit of the model.

Kennedy's data present a significant challenge to our model assumptions. Kennedy's original data showed serious violations in multiple assumptions.Among all transoformation tested, logarithmic conversions of sample size and taking the square root of the outcome variable proved to be the most effective in reducing hypothesis violations. However, after these adjustments, the Kennedy model's residuals still show a clear downward trend, suggesting that the linear assumption may be violated because the linear model may not fully capture the relationship between the predictor and the outcome. In addition, there is evidence of heteroscedasticity, with residuals showing larger variances at lower fit values, indicating that the error variance is not constant. In addition, there do exsit 2 main clusters of points. Together, these issues suggest that the Kennedy model may violate the assumptions of linear, constant variance, and uncorrelated errors, and future studies need to be further modified to adequately address these errors.

Stein shows a residual plot with a slight downward trend, which suggests that the model may benefit from different specifications or transformations to improve the fit. The distribution of residuals varies between the fitted values, especially at the low end, which indicates slight heteroscedasticity .Outliers stand out in the plot and can affect the performance of the model. Examining these points further provides insight into their impact on residuals and potential adjustments to the model.

12

Overall, the MLR model seems to fit Biden, Harris, and Trump. However, it requires an improved or alternative approach to Kennedy and potentially Stein. Kennedy's data requires further data transformations. Future studies may consider more advanced techniques or alternative conversions to better address the observed violations.
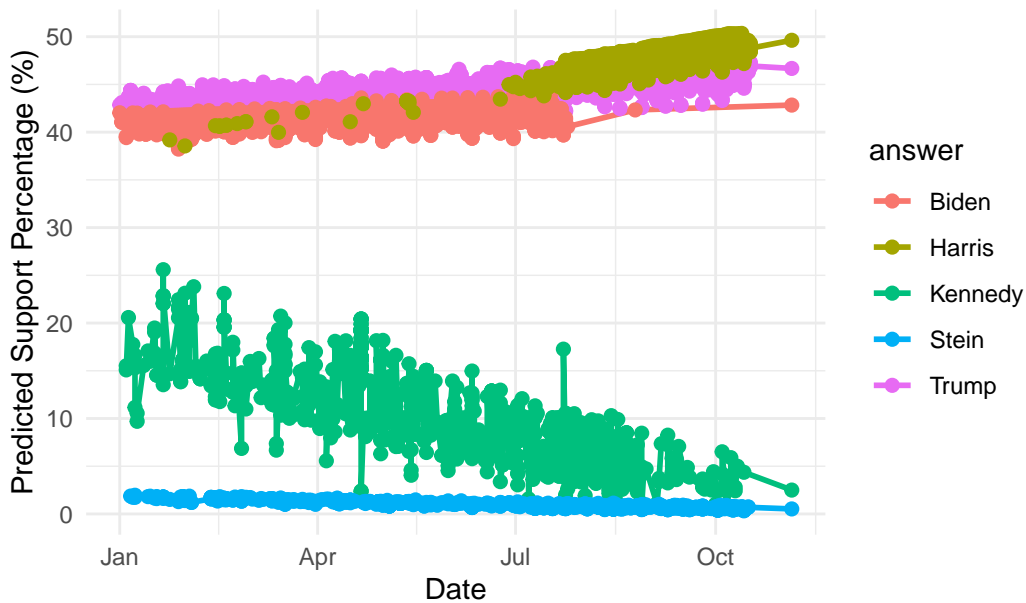
5 Prediction of Support Over Time

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

redicted Support Percentage Over Time for Top 5 Candidates

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

In this prediction, we exclude Kennedy and Biden. As the residual analysis(Section 4) shows, Kennedy's data has serious violations on linearity, constant variance, and uncorrelated error. These problems implies that this prediction model did not fit Kennedy, making it difficult to draw reliable conclusions from estimates. Biden, on the other hand, was excluded from the election due to his recent exit, so his support percentage data has nothing to do with the remaining results of the 2024 US election. His leave might lead to a reallocation of his supporters, potentially affecting the dynamics among the other candidates.

As the focus narrows to Harris, Stein and Trump, Harris' final estimated support is slightly higher than Trump's. However, several factors lead us to estimate Trump is likely the front-runner. While Harris's estimated percentage is slightly higher, Trump's overall trend across the timeline shows a steady upward trend, suggesting that increasing trend is building as the election nears. The continued growth in support suggests that Trump may continue to attract more voters, further solidifying his base.

Moreover, Trump's strengthening trend positions him well to benefit from the reallocation of Biden's supporters, who may now seek to align themselves with a new candidate after Biden's exit.

Harris' supports do show only a slight upward trend, but lacking the same dynamic growth and timeline length as Trump's. Their final score gap is also likely to be within the range of natural fluctuations, making Trump's momentum a more reliable indicator of his potential advantage.

7 Discussion

This paper uses statistical modeling to study the dynamics of poll support for the top 5 candidates in the 2024 U.S. president election. By using simple linear regression (SLR) and multiple linear regression (MLR) models, we aim to understand the relationship between vote quality measures such as vote score, sample size, and transparency score and voter support percentage. Our findings first suggest that sample size has a limited impact on support, so that we might proceed to further analysis. High-quality polls (as measured by transparency and poll scores) tend to show higher levels of support for candidates. This trend underscores the importance of methodological rigor, as transparent and well-executed polls are more likely to be perceived as reliable, and might influence respondents' preferences.

In addition, the MLR model captured important trends in favor of Joe Biden, Kamala Harris, and Donald Trump, showing power in predicting their rankings as Election Day approached. For example, Biden and Harris benefited from higher support percentage in transparent polls, suggesting that polling methods influence public trust. This relationship between vote quality

and voter sentiment highlights the potential for statistical models to provide valuable insights into election predictions.

While the model provides a solid basis for predicting election outcomes based on statistical indicators, it also exposes the limitations of purely statistical methods. Real-world factors such as party loyalty, voting patterns of specific demographics, and recent changes in candidate rankings remain unresolved in our analysis. Future improvements could involve integrating qualitative data and simulations to account for the reallocation of voters after a candidate drops out, creating a more comprehensive and adaptable predictive model.

8 Weaknesses and next steps

A key limitation of this paper is Biden's unexpected exit from the election, which challenges the accuracy of our model's predictions. As a well-known Democratic candidate, Biden's exit could mean that his supporters will reallocate their votes, and many will turn to Harris because she belongs to the same party. This redistribution of support skews the data, potentially inflating Harris's expected support in a way that models assuming a static voter base cannot accurately capture.

Another major challenge was encountered in applying the model to Kennedy's supporting data, which revealed substantial problems. Despite various transformations, including logarithm and square root adjustments, the residual of Kennedy's support showed a clear downward trend, suggesting a violation of the model's linear assumptions. The presence of heteroscedasticity further suggests that Kennedy's base of support was more variable and unpredictable than other candidates.

Overall, this paper only relies on purely statistical methods, which do not fully account for real-world factors. The model operates under the assumption that all polls are fair and representative; However, changes in poll methodology, question wording, and sample demographics can introduce biases that affect the accuracy of support predictions. In addition, our analysis does not account for external influences, such as major election campaigns, media coverage, or changes in public sentiment, which are critical to shaping voter behavior.

To address these weaknesses, future research should consider alternative models to better capture complex voter dynamics. Advanced statistical techniques can also help control for bias and demographic factors in a particular poll. Moreover, a hybrid model that combines statistical analysis with real-world data, might provide a more complete picture of what drives voter behavior in elections.

A Appendix

B Pollster Discussion

We choose Morning Consult as our focus. Morning Consult is a pollster known for its consistent approach to gathering public opinion data, especially during election periods. This discussion delves into Morning Consult's methodology, including sampling, recruitment, and strengths and weaknesses, based on the information available in the dataset.

B.1 Population, Frame, and Sample

Morning Consult's polls are designed to be representative of the general voting population, with a particular focus on reaching a broad range of likely voters(Morning Consult (2024)). Sample sizes in the data range widely, ranging from around 490 people to more than 1,000. This variability reflects an attempt to balance feasibility and accuracy, as larger sample sizes generally produce more reliable results, although they are more resource-intensive.

B.2 Sample Recruitment

Although the dataset does not explicitly detail the recruitment methods used, Morning Consult employs online panels to gather data, digitally reaching a large and diverse group of respondents. This approach enables effective data collection, but can introduce biases related to Internet accessibility and self-selection, as individuals who choose to participate in online surveys may not be fully representative of the wider population.

B.3 Sampling Approach and Trade-Offs

Morning Consult primarily uses a quota sampling method, ensuring that various demographics within the population are represented in its sample based on population distribution. This method helps balance the sample to reflect key demographics such as age, gender, race, and political affiliation, but it does not provide the randomness of probability sampling, which can affect the generality of the results. In addition, quota sampling may not fully capture less common or rapidly changing views among voters.

B.4 Handling of Non-Response

The dataset does not specify Morning Consult's approach to non-responses; However, due to the immediate participation of participants, online surveys generally see fewer non-responses than telephone or mail surveys. That said, if certain demographic groups are less likely to complete the survey, Morning Consult may still face challenges from non-response bias, which could skethe results despite balancing demographics with quotas.

B.5 Questionnaire Designs

Morning Consult's questionnaire design is generally consistent, focusing on standard polling indicators such as candidate support and party affiliation. The structured format helps keep the polls clear and comparable. However, there are limitations: online questionnaires sometimes lack the interactive depth provided by interviewee led surveys, and respondents may misinterpret questions without an opportunity for clarification. In addition, using sorted choice redistribution, a feature in some of Morning Consult's polls, allows for a more nuanced view of candidate preferences in multi-candidate scenarios, although this can complicate data interpretation for those unfamiliar with the system.

B.6 Overall Strengths & Weaknesses

Morning Consult's online methodology provides quick and cost-effective access to a broad sample, enabling frequent updates to poll data. Its use of quota sampling also helps ensure a balanced population, making polls more reflective of the general population. In addition, Morning Consult scored 3 on transparency in this dataset, indicating a moderate openness to methodology that is valuable in accurately interpreting the results.

However, reliance on online samples can introduce biases associated with Internet access and self-selection. The lack of probability sampling reduces the generality of survey results, non-response processing remains unclear, and data quality can be compromised if certain groups systematically avoid participation. The format of the questionnaire, while consistent, can also limit engagement because online surveys may not promote the level of understanding of respondents as much as face-to-face or telephone surveys.

C Idealized Methodology and Survey Design

With a budget of $100,000 dedicated to predicting the U.S. presidential election, we will implement a comprehensive survey methodology that prioritizes representation, data integrity, and forecast accuracy. The method includes deep sampling method, respondent recruitment strategy, data validation protocol and poll aggregation technique. In addition, we will utilize digital survey platforms, such as Google Forms, to collect feedback in an effective and inclusive manner.

C.1 Sampling Approach

The ideal sampling method for this survey is stratified random sampling. This approach ensures that we capture key demographic variables, such as age, gender, race, geographic location, and political affiliation, in proportion to their presence in the U.S. electorate. By dividing the population into these classes and sampling within each group, we enhance the representativeness of the sample. This approach will allow us to create a data set that accurately reflects the diversity of the voting population, reduce bias, and make the findings more broadly applicable to a wider range of voters.

C.2 Respondent Recruitment

In order to recruit diverse respondents, we will use a hybrid model that combines online and telephone recruitment. Online recruitment will target a broad audience through social media platforms, news sites and email outreach, using targeted advertising to reach specific populations that have been underrepresented in previous surveys. For people who are unlikely to respond online, such as the elderly or those with limited Internet access, we conduct telephone interviews. This approach ensures that we reach a wide range of voters and minimizes potential non-response bias.

C.3 Data Validation

Data validation is critical to maintaining the accuracy and reliability of survey results. Our data verification protocol will involve cross-referencing demographic information with external databases to verify the authenticity of responses and flag duplicates or inconsistencies. In addition, we will incorporate response checks, such as attention-checking questions, into the survey to identify and filter out unattended or automated responses. To further improve the quality of the data, we conduct pilot tests and adjust the survey design based on the weaknesses identified.

C.4 Poll Aggregation

Given the dynamic nature of election forecasting, poll aggregation will be an integral part of our methodology. The aggregation of results from multiple rounds of polls conducted in the weeks leading up to the election allows us to track changes in public sentiment over time. Each poll cycle provides the latest insights, allowing us to observe trends and momentum changes among the candidates. We will apply a weighted average to the aggregated results, assigning more weight to recent polls and responses from demographic groups that have historically had high voter turnout. This approach helps eliminate variability in the data and provides more stable and reliable predictions.

C.5 Survey Implementation

Using Google Forms as a survey platform provides a cost-effective and easily accessible way to manage surveys with large audiences. The platform's integration with Google Sheets and other analytics tools also facilitates effective data collection, organization, and analysis. As described in the Recruitment section, survey links will be shared across multiple channels to maximize reach. In addition, we will set up regular survey reminders to encourage participation and improve response rates. A link to a sample Google Forms survey will be included in the appendix, along with a full copy of the survey, for greater transparency.

C.6 Questionnaire Design

Research indicates that concise surveys with varied question formats can enhance respondent engagement and data ((**qualitypewresearch2022?**)). As a result, the survey will consist of several sections aimed at understanding voter intentions and demographics. Key questions include candidates' preferences, likelihood of voting, and views on key issues affecting their choices. Demographic questions will cover age, gender, race, income level, and political affiliation, allowing us to break down the data and get a nuanced understanding of voter groups. The survey will be concise to reduce dropout rates and incorporate multiple choice and Likert scale questions to keep respondents engaged and facilitate statistical analysis.

C.7 Budget Allocation

A budget of $100,000 will be allocated to all aspects of the investigation process. About 40 percent will be spent on recruitment and rewards for respondents, while higher grants will go to hard-to-reach groups. Another 30% will be used for data processing and validation to ensure the highest data quality. The remaining budget will be used to support the vote aggregation process, ongoing analysis and reporting of results.

References

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

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

GeoPoll. 2022. "The Importance of Data Transparency in Survey Research." https://www.geopoll.com/blog/data-transparency/.

Morning Consult. 2024. "Morning Consult." https://morningconsult.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/.