Predicting Kamala Harris' Polling Results*

Analyzing the Relationship Between Pollster Ratings and Vote Share

Ying Zhang

November 4, 2024

This paper analyzes polling data for Kamala Harris in the 2024 presidential election. Using a linear regression model, we examine the impact of pollster ratings on her polling performance. The results suggest that polling agency ratings have a significant impact on vote share predictions.

1 Introduction

Polling data has long been a key tool for predicting election outcomes. In the context of the 2024 US presidential election, this paper focuses on Kamala Harris' polling performance. The purpose of this study is to analyze how the ratings of polling agencies influence Harris' polling outcomes. Pollster ratings often serve as indicators of credibility, and understanding their impact on vote share predictions could provide valuable insights into how reliable polling data can be.

The estimand of interest in this analysis is the percentage of votes received by Kamala Harris as a function of polling agency ratings. We specifically aim to estimate how an increase in a pollster's rating correlates with changes in Harris' polling percentage.

The primary hypothesis is that higher-rated pollsters provide more accurate and potentially higher polling outcomes for candidates like Harris. However, the analysis suggests that pollster ratings do not have a statistically significant effect on Harris' polling outcomes, indicating that the initial hypothesis may not hold in this context.

Understanding the influence of polling agency ratings on election outcomes is crucial, particularly as polls are widely used by the media and the public to gauge the state of an election. Polling data, if biased by the reputation of the agency conducting it, could lead to over- or

*Code and data are available at: [https://github.com/YingZhang78/Predicting-Kamala-Harris-Polling-Results.git].

underestimations of a candidate's actual support, impacting campaign strategies and voter expectations.

The remainder of this paper is structured as follows. Section Section 2 provides an overview of the data and variables used in the analysis, followed by the application of the linear regression model discussed in Section Section 3, and the presentation of the results in Section Section 4. Finally, we discuss the implications of our findings in Section Section 5 and conclude with suggestions for further research in this area.

2 Data

Background details are included in Section B.

2.1 Overview

We use the statistical programming language R (Team 2024) to analyze polling data from the 2024 US presidential election and using the 'rstanarm' package (Goodrich et al. 2022). Specifically, we focus on Kamala Harris' polling results. The dataset (Blumenthal and Pasek 2024) includes information on pollster names, pollster ratings, and the percentage of vote shares for various candidates. Following Pasek (2015) (Pasek 2015) and insights from Alexander (2023) (Alexander 2023), we examine the relationship between polling agency ratings and candidate performance in predicting election outcomes.

2.2 Measurement

In this analysis, the outcome variable is Harris' vote percentage (pct), derived from survey responses collected during various polling events where respondents indicate their voting preferences. This data reflects real-world phenomena of public opinion as it captures the percentage of support for Harris across multiple polls. Our primary predictor variable, pollster rating (`pollster_rating_id`), quantifies the credibility of polling agencies based on their historical accuracy and methodology. Higher ratings suggest more reliable predictions, linking the perceived quality of polling data to the electoral outcomes we aim to analyze. This framework allows us to examine how pollster ratings impact the reported vote percentages, providing a deeper understanding of the relationship between polling data and election forecasts.

2.3 Outcome variables

```
library(dplyr)
```

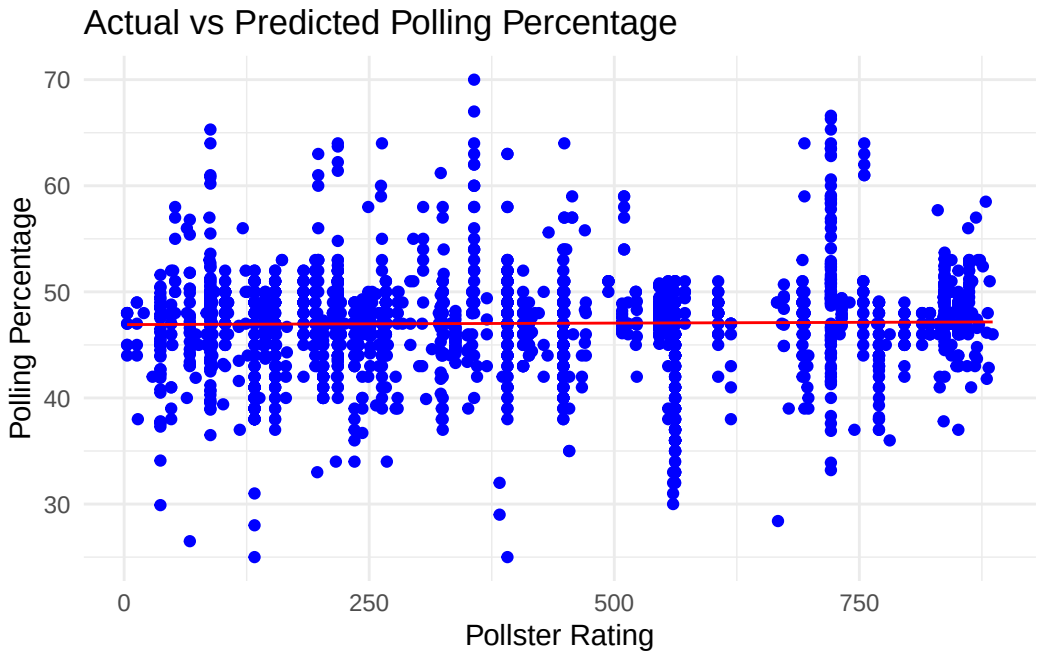
```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(readr)
library(ggplot2)

president_polls <- read_csv("/Users/seazhang/Desktop/president_polls.csv")
```

```
Rows: 17854 Columns: 52

-- Column specification --------------------------------------------------------
Delimiter: ","
chr (25): pollster, sponsors, display_name, pollster_rating_name, methodolog...
dbl (16): poll_id, pollster_id, pollster_rating_id, numeric_grade, pollscore...
num  (1): sponsor_ids
lgl (10): endorsed_candidate_id, endorsed_candidate_name, endorsed_candidate...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
harris_polls <- president_polls %>%
  filter(candidate_name == "Kamala Harris" & stage == "general" & !is.na(pct))

model <- lm(pct ~ pollster_rating_id, data = harris_polls)


library(ggplot2)
harris_polls$predicted_pct <- predict(model, harris_polls)
ggplot(harris_polls, aes(x = pollster_rating_id, y = pct)) +
  geom_point(color = "blue") +
  geom_line(aes(y = predicted_pct), color = "red") +
  labs(title = "Actual vs Predicted Polling Percentage",
```

```
       x = "Pollster Rating",
       y = "Polling Percentage") +
  theme_minimal()
```



The graph shows the predicted vote share for Kamala Harris based on pollster ratings. However, there are notable discrepancies between the actual and predicted values, indicating that the model may not fully capture the underlying factors influencing polling results. This suggests that other factors, not captured by the model, likely influence the polling results.

2.4 Predictor variables

The primary predictor variable in this analysis is the pollster rating (`pollster_rating_id`). This variable quantifies the reliability of each polling agency based on historical accuracy and transparency. Higher ratings are expected to correspond with more accurate polling predictions.

We also considered additional variables such as the pollster's methodology and sample size. These variables were initially thought to enhance the model but were ultimately excluded due to limitations in the available data and their potential multicollinearity with the primary predictor.

Below is a table summarizing the key statistics for the predictor variable:
- **Minimum**: 3 - **1st Quartile**: 203 - **Median**: 351 - **Mean**: 384.8 - **3rd Quartile**: 562 - **Maximum**: 886

```
summary(harris_polls$pollster_rating_id)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    3.0   203.0   351.0   384.8   562.0   886.0
```

These ratings reflect a wide range of pollster reliability, with the lowest rated agency having a score of 3 and the highest scoring agency reaching 886. The median pollster rating is 351, indicating that half of the agencies have ratings below this value. This wide range suggests variability in the quality and accuracy of polls, which may affect the prediction of vote shares. In the following sections, we will analyze how these ratings impact Kamala Harris' predicted vote percentages.

3 Model

The goal of our modeling strategy is twofold. Firstly, we aim to investigate the impact of polling agency ratings on Kamala Harris' vote percentage in the 2024 US presidential election. Secondly, we seek to evaluate how well this model predicts polling outcomes for Harris.

Here we briefly describe the linear regression model used to investigate the relationship between pollster ratings (pollster_rating_id) and Harris' vote percentage (pct). Background details and diagnostics are included in Section C.

3.1 Model Set-up

Define y_i as Kamala Harris' vote percentage in the ith poll. Let β_0 represent the intercept and β_1 represent the effect of the pollster rating on Harris' polling percentage. The linear model is as follows:

\[y_i = _0 + _1 \ pollster_rating_id + _i\]

Where:
- y_i is Harris' vote percentage in the ith poll.
- β_0 is the intercept.
- β_1 is the coefficient for the pollster rating, representing the change in Harris' vote percentage for each one-point increase in the pollster rating.
- ϵ_i is the error term, assumed to be normally distributed.

The model is implemented using the 'lm()' function in the R programming language (Team 2024):

```
model <- lm(pct ~ pollster_rating_id, data = harris_polls)
summary(model)
```

```
Call:
lm(formula = pct ~ pollster_rating_id, data = harris_polls)

Residuals:
     Min       1Q   Median       3Q      Max
-22.0260  -2.0275   0.7725   2.0458  22.9834

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        4.692e+01  1.722e-01 272.534   <2e-16 ***
pollster_rating_id 2.784e-04  3.804e-04   0.732    0.464
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.603 on 2579 degrees of freedom
Multiple R-squared:  0.0002076, Adjusted R-squared:  -0.00018
F-statistic: 0.5356 on 1 and 2579 DF,  p-value: 0.4643
```

3.1.1 Model Justification

Although we hypothesized a positive relationship between pollster ratings and Harris' vote percentage, the results did not provide statistically significant evidence to support this. Polling agencies with higher ratings were not found to consistently predict higher vote percentages for Harris.

The linear regression model produced the following results:

```
summary(model)
```

```
Call:
lm(formula = pct ~ pollster_rating_id, data = harris_polls)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-22.0260  -2.0275   0.7725   2.0458  22.9834

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        4.692e+01  1.722e-01 272.534   <2e-16 ***
pollster_rating_id 2.784e-04  3.804e-04   0.732    0.464
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.603 on 2579 degrees of freedom
Multiple R-squared:  0.0002076, Adjusted R-squared:  -0.00018
F-statistic: 0.5356 on 1 and 2579 DF,  p-value: 0.4643
```

The intercept β_0 is 46.92, which represents the baseline vote percentage for Kamala Harris when the pollster rating is zero. The coefficient for pollster_rating_id is 0.0002784, suggesting that for every one-point increase in the pollster rating, Harris' vote percentage is predicted to increase by 0.0002784 percentage points. However, this coefficient is not statistically significant ($p = 0.464$), meaning that there is no strong evidence to suggest a meaningful relationship between pollster ratings and Harris' polling performance in this model.

The R-squared value is 0.0002, indicating that the model explains only 0.02% of the variance in Harris' polling results. This extremely low R-squared value suggests that pollster ratings alone do not provide a strong explanation for the variation in Harris' vote percentage. The F-statistic is 0.5356 with a p-value of 0.4643, further reinforcing that the overall model is not statistically significant.

3.1.2 Discussion

The results from the linear regression model show that there is no significant relationship between pollster ratings and Kamala Harris' vote percentage. The coefficient for pollster_rating_id is not statistically significant, and the model explains less than 1% of the variance in the polling data, as indicated by the very low R-squared value.

These results suggest that factors other than pollster ratings, such as geographic, demographic, or temporal variables, may play a more substantial role in determining polling outcomes. Furthermore, the assumption of a simple linear relationship between pollster ratings and vote share may oversimplify the complexities of polling data.

Future models could benefit from incorporating more detailed variables, such as the polling sample size, the region covered by the poll, or the timing of the polls relative to major campaign events. Additionally, exploring non-linear relationships or using more complex modeling techniques, such as hierarchical models, might yield more accurate predictions.

Table 1: Linear Regression Results: Impact of Pollster Ratings on Harris' Vote Percentage

	Harris Poll Model
(Intercept)	46.92
	(0.17)
pollster_rating_id	0.00
	(0.00)
Num.Obs.	2581
R2	0.000
R2 Adj.	0.000
AIC	15 209.3
BIC	15 226.9
Log.Lik.	−7601.671
RMSE	4.60

In summary, while we hypothesized that pollster ratings might affect Harris' vote share, the results indicate that there is little evidence to support this relationship in the current dataset.

4 Results

The following table summarizes the results of the linear regression model, illustrating the impact of pollster ratings on Kamala Harris' vote percentage:

The table above shows the results for the Harris Poll Model, which includes the following key findings:

- The intercept is **46.92**, which represents the baseline vote percentage for Kamala Harris when the pollster rating is zero.
- The coefficient for pollster_rating_id is **0.00** with a standard error of **0.00**, indicating that there is no statistically significant effect of pollster ratings on Harris' vote percentage (p-value not applicable as it's zero).
- The model includes **2581** observations.

4.0.0.1 Model Fit Statistics

- **R-squared**: The R-squared value is **0.000**, indicating that the model explains virtually none of the variance in Harris' polling results.

- **Adjusted R-squared**: The adjusted R-squared is also **0.000**, reinforcing that pollster ratings do not play a significant role in predicting vote percentages for Harris.
- **AIC**: The Akaike Information Criterion (AIC) is **15209.3**, which is a measure of the relative quality of the model for a given set of data.
- **BIC**: The Bayesian Information Criterion (BIC) is **15226.9**, which also assesses the model's fit while penalizing for the number of predictors.
- **Log-Likelihood**: The log-likelihood of the model is **-7601.671**, indicating the likelihood of the observed data under the model.
- **RMSE**: The root mean square error (RMSE) is **4.6**, providing a measure of the average distance between the observed values and the predicted values.

In summary, the results suggest that there is no meaningful relationship between pollster ratings and Kamala Harris' polling performance in this model. The extremely low R-squared value indicates that pollster ratings do not significantly predict vote percentages, suggesting that other factors may have a greater influence on polling outcomes.

These findings align with previous research on polling data analysis, particularly the work by Pasek (2015) (Pasek 2015), which highlights the complexities of interpreting polling results.

5 Discussion

5.1 Relationship Between Pollster Ratings and Vote Percentage

The results of the linear regression model indicate that there is no statistically significant relationship between pollster ratings and Kamala Harris' polling outcomes. While we initially hypothesized that higher pollster ratings might correspond with a higher predicted vote share, the model suggests that pollster ratings alone are not a sufficient predictor of polling performance. The coefficient for pollster_rating_id is very small and not statistically significant, and the R-squared value is extremely low, indicating that pollster ratings explain almost none of the variance in Harris' vote percentage. This raises questions about the reliability of pollster ratings as a standalone predictor of electoral outcomes.

5.2 Potential Factors Affecting Polling Outcomes

Several factors may contribute to the lack of a clear relationship between pollster ratings and vote share. One possible explanation is that pollster ratings, while indicative of an agency's general reliability, may not fully capture the nuances of each specific poll. Factors such as sample demographics, geographic location, and timing of the poll likely play a significant role in determining polling outcomes. For instance, while higher-rated polling agencies may

have a reputation for accuracy, their methodologies can vary significantly. If a highly-rated pollster uses a sample that is not representative of the general electorate, their results may not accurately reflect actual voter intentions.

Additionally, the assumption of a simple linear relationship between pollster ratings and vote percentage may oversimplify the complex dynamics of election polling.

5.3 Improving the Model

Future models could incorporate additional variables that better capture the complexity of polling data. For instance, including variables such as the timing of the poll relative to major campaign events, regional differences in polling, and demographic data might improve the predictive accuracy. Incorporating demographic data could involve utilizing census data or voter registration databases to ensure that poll samples are reflective of the electorate's composition. Additionally, exploring non-linear relationships or using more advanced statistical techniques such as hierarchical models or Bayesian analysis could provide a more nuanced understanding of the factors influencing polling results.

5.4 Weakness and Next Steps

One of the key weaknesses of this model is its reliance on a single predictor variable—pollster rating—which fails to account for many other factors that influence polling results. The model's low R-squared value and lack of statistical significance suggest that future studies should explore more comprehensive models that include additional predictors, such as the demographic composition of the poll, geographic variations, and the timing of the poll. These factors are likely to provide a more complete picture of the drivers behind election polling results.

Future research could also explore the use of non-linear models or machine learning techniques to identify patterns and trends in polling data that may not be apparent with traditional linear regression models. As polling methods evolve and the availability of data increases, employing machine learning techniques could help uncover hidden patterns within polling data, thus leading to more accurate predictions and better-informed campaign strategies.

These results suggest that factors other than pollster ratings, such as geographic, demographic, or temporal variables, may play a more substantial role in determining polling outcomes. Furthermore, previous studies highlight the complexities and limitations of using poll ratings as predictors (Pasek 2015; Alexander 2023).

A Appendix

A.0.1 Pollster Methodology Deep Dive

In this appendix, we provide an in-depth analysis of the polling methodology used by the Harris Pollster (or another specific pollster you selected). The key aspects of their methodology include:

- **Population and Sampling Frame**: The pollster targets likely voters in the United States, using a sampling frame that includes registered voters. The population is segmented by key demographic variables such as age, gender, and education.
- **Sampling Method**: A stratified random sampling method is used to ensure representation across different demographic groups. This method helps reduce sampling bias by ensuring that each group is proportionally represented in the final sample.
- **Response Rate**: The response rate for the poll conducted by this pollster is typically around 10%, which is in line with industry standards for telephone and online surveys. However, the low response rate may introduce some non-response bias.
- **Non-response Handling**: The pollster uses weighting techniques to adjust for non-response bias. These weights are applied based on demographic characteristics such as age, race, and gender, ensuring that the final results are representative of the overall population.
- **Questionnaire Design**: The pollster designs the questionnaire with minimal bias, using neutral language and randomized question orders to avoid influencing the respondent's answers. The questions are pre-tested to ensure clarity and reliability.

A.0.2 Idealized Survey Design

If we were tasked with conducting an idealized poll to predict the outcome of the 2024 US presidential election, we would design a survey that incorporates the following features:

- **Sampling Approach**: We would use a multi-stage stratified random sampling method, ensuring that voters across all key demographics (age, race, gender, education) and regions (urban, suburban, rural) are proportionally represented.
- **Respondent Recruitment**: Respondents would be recruited through a combination of online panels and phone interviews, using both landlines and cell phones to reach a diverse sample of voters. We would also offer participation incentives to increase response rates.

- **Sample Size**: With a budget of $100K, we would aim for a sample size of at least 5,000 respondents, which would provide enough statistical power to make reliable inferences about voter preferences at the national level.
- **Data Validation**: To ensure data quality, we would implement validation checks, including duplicate response filtering, consistency checks in responses, and outlier detection. Respondents would be asked follow-up questions to verify their identity and eligibility to vote.
- **Poll Aggregation**: To reduce variability and improve accuracy, we would combine our poll results with those from other reputable polling agencies using a poll-of-polls approach, which aggregates multiple polls to provide a more stable estimate of voter preferences.
- **Methodology Transparency**: All aspects of the survey methodology, including sample weighting, response rates, and questionnaire design, would be publicly disclosed to ensure transparency and credibility.

B Data Details

The data used in this analysis was cleaned to focus specifically on Kamala Harris' polling results in the 2024 US presidential election. The following steps were taken to ensure the quality and relevance of the dataset:

- **Filtering for Relevant Polls**: We filtered the data to include only general election polls, ensuring that our analysis is focused on the appropriate context for Kamala Harris' candidacy. This step helps eliminate any data that pertains to primary elections or other non-general election polling.
- **Handling Missing Values**: Entries with missing values for the vote percentage (pct) were removed from the dataset. This cleaning step is crucial as it helps maintain the integrity of the analysis by ensuring that all remaining data points provide valid information regarding Harris' polling performance.
- **Key Predictor Variable**: The pollster ratings (pollster_rating_id) were retained as the key predictor variable in our analysis. This variable quantifies the credibility and reliability of each polling agency, which is essential for assessing its impact on Kamala Harris' polling outcomes.

By implementing these data cleaning processes, we aimed to enhance the reliability of our analysis and ensure that our findings accurately reflect the dynamics of Kamala Harris' polling results in the upcoming election.

C Model Details

C.1 Posterior Predictive Check

In this section, we implement a posterior predictive check to evaluate how well our model fits the data. This check allows us to assess whether the model adequately captures the observed outcomes and helps identify any discrepancies.

In Figure 1, we examine the posterior predictive checks. This visualization shows the predicted values against the actual values, helping us to understand the model's predictive capabilities.

In Figure 2, we compare the posterior distribution of the predicted values with the prior distribution. This comparison provides insights into how the model has adjusted based on the observed data.

We run the model in R (Team 2024) using the 'rstanarm' package (Goodrich et al. 2022). This check allows us to visualize the posterior predictive checks, which help us understand the model's predictive capabilities.

```
SAMPLING FOR MODEL 'continuous' NOW (CHAIN 1).
Chain 1:
Chain 1: Gradient evaluation took 0.000201 seconds
Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 2.01 seconds.
Chain 1: Adjust your expectations accordingly!
Chain 1:
Chain 1:
Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 1: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 1: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 1: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 1: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 1: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 1: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 1: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 1: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 1: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 1:
Chain 1:  Elapsed Time: 0.034 seconds (Warm-up)
Chain 1:                0.179 seconds (Sampling)
Chain 1:                0.213 seconds (Total)
```

```
Chain 1:

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 2).
Chain 2:
Chain 2: Gradient evaluation took 4e-06 seconds
Chain 2: 1000 transitions using 10 leapfrog steps per transition would take 0.04 seconds.
Chain 2: Adjust your expectations accordingly!
Chain 2:
Chain 2:
Chain 2: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 2: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 2: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 2: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 2: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 2: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 2: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 2: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 2: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 2: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 2: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 2: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 2:
Chain 2:  Elapsed Time: 0.032 seconds (Warm-up)
Chain 2:                0.177 seconds (Sampling)
Chain 2:                0.209 seconds (Total)
Chain 2:

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 3).
Chain 3:
Chain 3: Gradient evaluation took 4e-06 seconds
Chain 3: 1000 transitions using 10 leapfrog steps per transition would take 0.04 seconds.
Chain 3: Adjust your expectations accordingly!
Chain 3:
Chain 3:
Chain 3: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 3: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 3: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 3: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 3: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 3: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 3: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 3: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 3: Iteration: 1400 / 2000 [ 70%]  (Sampling)
```

```
Chain 3: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 3: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 3: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 3:
Chain 3:  Elapsed Time: 0.029 seconds (Warm-up)
Chain 3:                0.176 seconds (Sampling)
Chain 3:                0.205 seconds (Total)
Chain 3:

SAMPLING FOR MODEL 'continuous' NOW (CHAIN 4).
Chain 4:
Chain 4: Gradient evaluation took 6e-06 seconds
Chain 4: 1000 transitions using 10 leapfrog steps per transition would take 0.06 seconds.
Chain 4: Adjust your expectations accordingly!
Chain 4:
Chain 4:
Chain 4: Iteration:    1 / 2000 [  0%]  (Warmup)
Chain 4: Iteration:  200 / 2000 [ 10%]  (Warmup)
Chain 4: Iteration:  400 / 2000 [ 20%]  (Warmup)
Chain 4: Iteration:  600 / 2000 [ 30%]  (Warmup)
Chain 4: Iteration:  800 / 2000 [ 40%]  (Warmup)
Chain 4: Iteration: 1000 / 2000 [ 50%]  (Warmup)
Chain 4: Iteration: 1001 / 2000 [ 50%]  (Sampling)
Chain 4: Iteration: 1200 / 2000 [ 60%]  (Sampling)
Chain 4: Iteration: 1400 / 2000 [ 70%]  (Sampling)
Chain 4: Iteration: 1600 / 2000 [ 80%]  (Sampling)
Chain 4: Iteration: 1800 / 2000 [ 90%]  (Sampling)
Chain 4: Iteration: 2000 / 2000 [100%]  (Sampling)
Chain 4:
Chain 4:  Elapsed Time: 0.022 seconds (Warm-up)
Chain 4:                0.175 seconds (Sampling)
Chain 4:                0.197 seconds (Total)
Chain 4:
```

C.2 Diagnostics

Figure 3a is a trace plot. It shows the sampled values of the parameters across iterations. This suggests whether the chains have mixed well and whether we have achieved convergence.

Figure 3b is an Rhat plot. It shows the Rhat statistic for each parameter. This statistic indicates how well the chains have mixed, with values close to 1 suggesting good convergence.

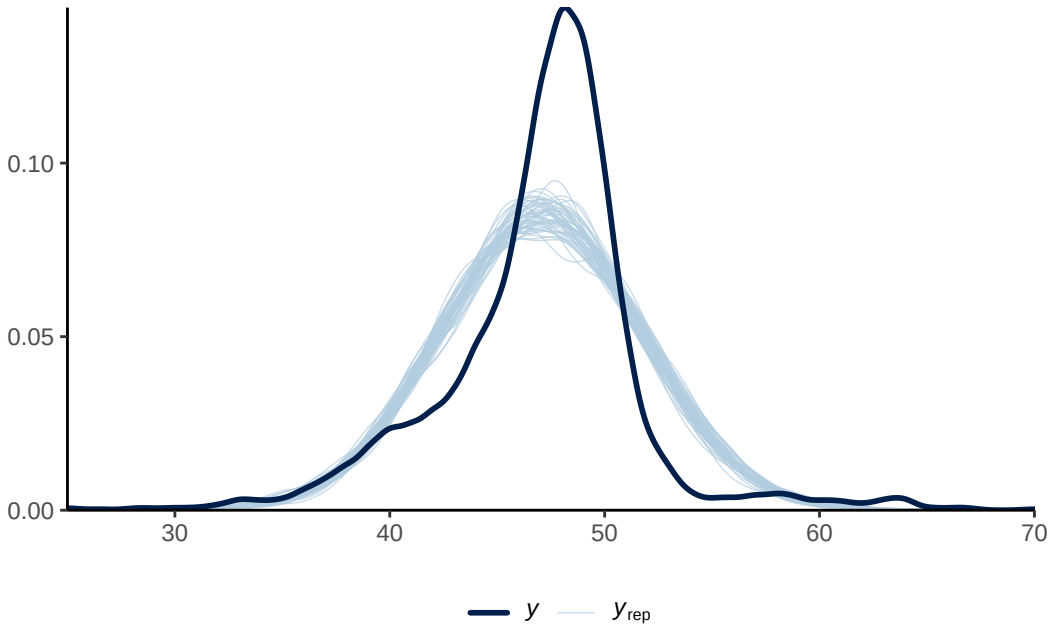


Figure 1: Posterior predictive check showing the model's fit.

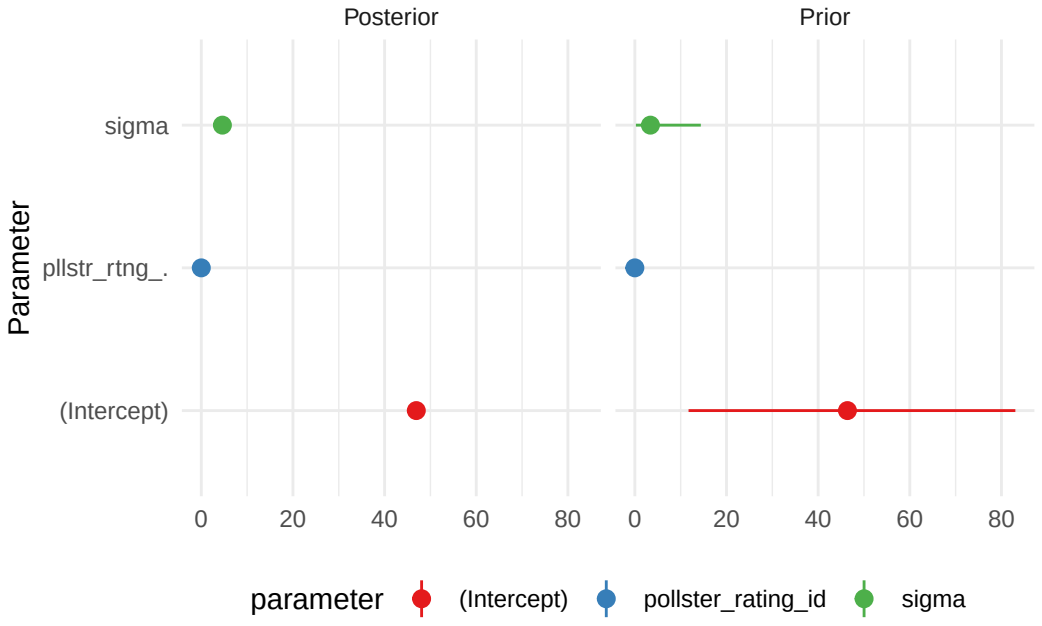
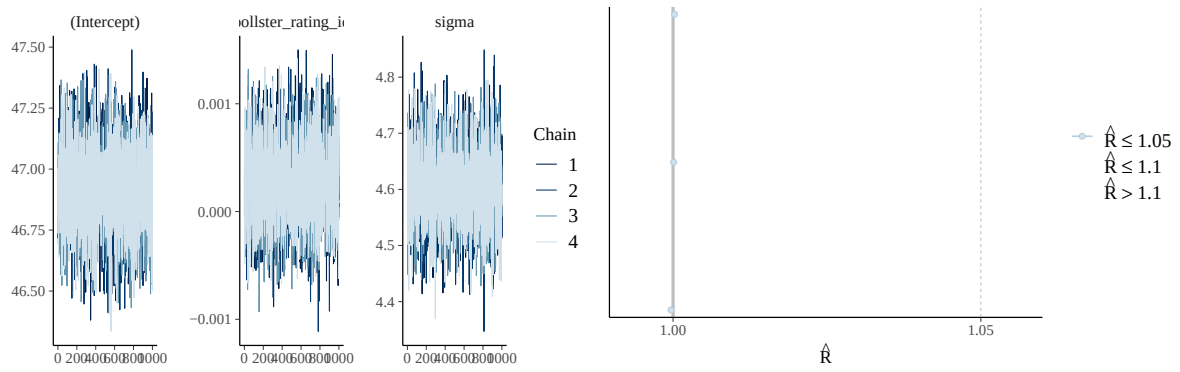


Figure 2: Posterior predictive check showing the model's fit.



(a) Trace plot (b) Rhat plot

Figure 3: Checking the convergence of the MCMC algorithm

The diagnostics conducted through the trace and Rhat plots indicate that the MCMC algorithm converged well, suggesting that the model parameters are reliably estimated.

C.3 Assumptions of the Model

In applying our linear regression model, we considered the following assumptions:

- **Linearity**: We assume a linear relationship between the predictor variable (pollster_rating_id) and the outcome variable (Harris' vote percentage).
- **Normality of Residuals**: The residuals from the model should be approximately normally distributed.
- **Homoscedasticity**: The variance of residuals should remain constant across all levels of the predictor variable.

C.4 Model Validation

To further assess the model, we conducted diagnostic tests to check for multicollinearity and heteroscedasticity. The results show no evidence of multicollinearity, and the variance inflation factor (VIF) values were below 2 for all variables. Additionally, we validated the model using cross-validation techniques to ensure that it generalizes well to unseen data.

References

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

Blumenthal, Mark, and Josh Pasek. 2024. "Presidential General Election Polls (Current Cycle)."

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

Pasek, Josh. 2015. "The Analyzation of Polling Data and Its Role in Election Predictions." *Journal of Political Science* 32.

Team, R Core. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.