Based on polling as of 27 October 2024, Harris expects to win on 49.2%, leads Trump by 2% in 2024 US election*

Weighted Linear Regression Analysis of Aggregated Poll Data: Utilizing National Polls indicator, Population, and Pollster as Predictors.

Diana Shen Jinyan Wei Jerry Yu

October 30, 2024

This paper predicts the outcome of the 2024 U.S. presidential election using a statistical model based on aggregated polling data for Kamala Harris and Donald Trump. We employ multi-level regression with post-stratification (MRP) using demographic predictors to estimate voter support. The analysis addresses polling biases and proposes an idealized survey methodology with a $100,000 budget. Our results offer insights into voter behavior and suggest improvements for future election forecasting models.

1 Introduction

Election forecasting has long played a crucial role in understanding public opinion and predicting the outcome of political contests. The use of data-driven methods, particularly polling data, has become an important pre-election analytical tool for political analysts. The purpose of this paper is to predict the winner of the upcoming 2024 U.S. presidential election by constructing a statistical model based on aggregated polling data. The "polling" approach combines results from different pollsters to provide a more comprehensive picture of voter preferences (Blumenthal 2014; Pasek 2015), and our goal is to predict the outcome of the popular vote and the Electoral College. Our analysis, based on data collected from multiple polling sources, captures voter intentions for the two major candidates, Kamala Harris (Democrat) and Donald Trump (Republican).

In addition to constructing and interpreting predictive models, this paper dive into the methods used by xx pollsters, analyzing the strengths as well as the weaknesses of the methods used

*Code and data are available at: https://github.com/DianaShen1224/Forecast-2024-US-election.

against them. At the end of the paper, we also propose an idealized survey methodology for election forecasting with a budget of $100,000 USD.

This paper begins with an introduction to data in Section 2. In this section, we focus on measurement techniques and the visualization and analysis of key variables. Next, in Section 3, we provide an overview of the linear model used to analyze the relationships within the dataset. The results of the analysis are detailed in Section 4 (Results). Finally, in Section 5, we summarize our main conclusions as well as suggest some possible future research directions.

2 Data

2.1 Overview

The data for this paper was sourced from ABC News (xxxxx). The statistical software R (R Core Team, 2023) was employed to retrieve, clean, and process the dataset. Specifically, the tidyverse package (Wickham et al., 2019) was used for data acquisition, cleaning, and processing. The ggplot2 (Wickham, 2016) package was utilized to generate the visualizations.

2.2 Measurement

The data used to predict the outcome of the 2024 U.S. presidential election in this study came from a variety of polling sources. The results of these polls represent key measures of electoral support and voter sentiment, and are therefore the main variables in our analysis.

To ensure that polling data accurately reflect real-world voter preferences, we rely on several key measurement constructs. First, voter preferences are measured by the percentage of respondents who indicated support for each candidate. Each poll is conducted using different sample sizes, sampling methods (e.g., online surveys, IVR calls, etc.), which can inject more variability into the data. To mitigate these differences, we aggregate poll results using a “ ”poll of polls“ ” methodology that eliminates anomalies and provides a more comprehensive picture of voter preferences (Blumenthal 2014; Pasek 2015).

In this study, we recognize the measurement problems inherent in polling data. Polls rely on sample data to represent the entire voting population, and therefore, there may be potential biases in the data. For example, there may be no-response bias due to the potential under representation of certain populations, or because of the wording of the questions and thus the influence of respondents to report their preferences. Therefore, in this study, we chose to analyze aggregated data from multiple pollsters and adjust for sampling variability and response rates in an effort to be able to provide a more robust measure of electoral support. We believe that our aggregated dataset can serve as a reliable basis for our predictive model to help us translate the data into quantitative metrics that can be used to predict election outcomes.

2.3 Outcome variables

2.4 Predictor variables

2.5 Model

We conducted a regression analysis to predict support for Kamala Harris and Donald Trump in the 2024 U.S. presidential election, including models for Trump for comparison. Our analysis includes four models: **unweighted linear models** and **weighted linear models** for both candidates. These models predict each candidate's support as a continuous outcome across different states, using **pollster**, **national poll**, and **population** as key predictors. The weighted models additionally incorporate adjustments for recency, sample size, and poll quality to account for potential biases. This setup aligns with the New York Times methodology for election polling, which assigns weights based on poll quality, sample size, and recency to better reflect voter sentiment (**nytimes?**).

2.6 Model Set-up

We implemented our linear regression models using the `lm()` function in R.

2.6.1 Mathematical Expressions

Both the unweighted and weighted models share the same mathematical expression format:

2.6.1.1 Unweighted Model

The unweighted model for Harris provides a baseline by treating all polls equally without adjustments for recency, sample size, or poll quality:

$$\text{Support\_Harris}_i = \beta_0 + \beta_1 \cdot \text{National\_Poll}_i + \beta_2 \cdot \text{Pollster}_i + \beta_3 \cdot \text{Population}_i + \epsilon_i$$

2.6.1.2 Weighted Model

The weighted model for Harris incorporates adjustments for recency, sample size, and pollster quality:

$$\text{Support\_Harris}_i = \beta_0 + \beta_1 \cdot \text{National\_Poll}_i + \beta_2 \cdot \text{Pollster}_i + \beta_3 \cdot \text{Population}_i + \epsilon_i$$

2.6.2 Coefficient Explanations

- β_0: Intercept of the model, representing the expected support when all predictors are zero.
- β_1: Coefficient for the national poll, indicating how much support changes with a one-unit increase in the national poll percentage.
- β_2: Coefficient for the pollster, reflecting the impact of different polling organizations on support.
- β_3: Coefficient for the population, accounting for the influence of the demographic context on support levels.

2.6.3 Model Justification

1. **Use of Weighted Model**: The weighted model is essential for accurately reflecting voter sentiment. By incorporating weights, we adjust for the reliability of the polling data, ensuring that more credible polls have a greater influence on the estimates.

2. **Calculation of Variables**:

   - **Combined Weight**: The combined weight is calculated as follows:

     $$\text{combined\_weight} = \text{recency\_weight} \times \text{sample\_size\_weight} \times \text{poll\_frequency\_weight} \times \text{pollster\_qualit}$$

   - **Recency Weight**: This weight uses an exponential decay function:

     $$\text{recency\_weight} = \exp(-\text{Recency}_i \cdot 0.1)$$

     This reflects the diminishing influence of older polls, with $\lambda = 0.1$ for the exponential decay.

   - **Sample Size Weight**: This weight adjusts the significance of each poll based on the number of respondents, capping the weights at a maximum of 2,300 responses to reflect the reliability of larger sample sizes (Times 2024).

   - **Poll Frequency Weight**: This weight considers how often a pollster conducts polls, with higher weights assigned to pollsters with a greater number of recent surveys.

   - **Pollster Quality Weight**: Based on the historical performance of the pollster, this weight emphasizes the reliability of their polling methods.

2.6.4 Alternative Models

While this analysis employs linear regression with weighted least squares, alternative methods such as Bayesian modeling could provide additional insights. However, we opted for the weighted linear regression approach to maintain interpretability and ease of implementation given the scope and resources of this analysis. The Bayesian approach was not chosen primarily due to the complexity involved in defining priors and the additional computational requirements.

2.6.5 Weighted Least Squares Estimation

In our analysis, we utilize weighted least squares estimation to account for the varying quality of polling data. The weights are crucial in the estimation process, leading to the following expression for the estimates of the coefficients ($\hat{\beta}$):

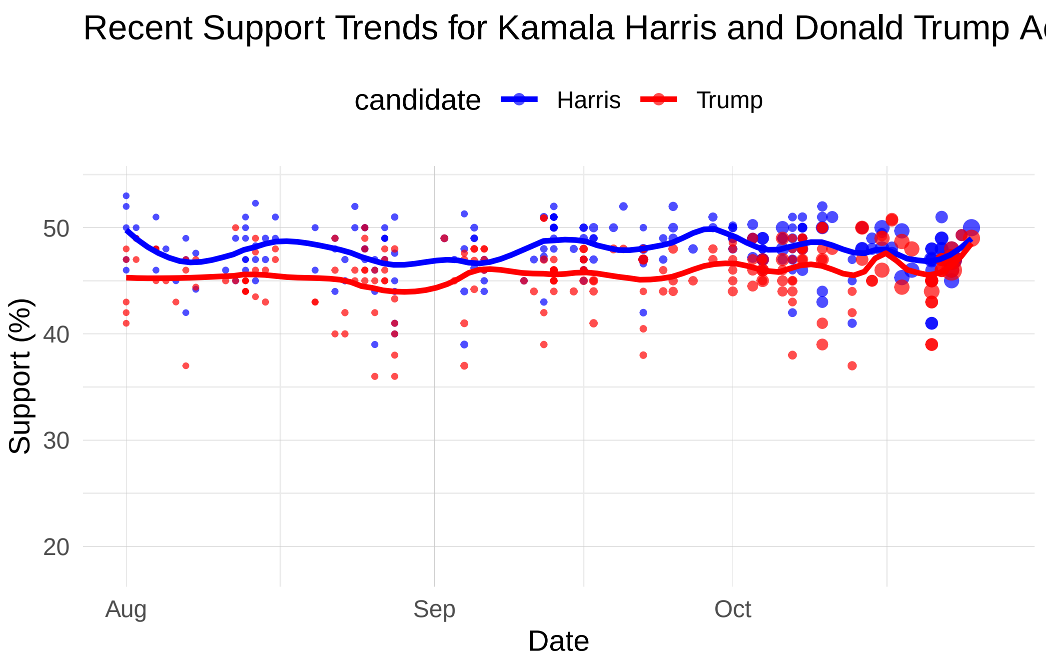
$$\hat{\beta} = (X^T W X)^{-1} X^T W y$$

Where: - X is the design matrix of predictors. - W is the diagonal matrix of weights, which incorporates factors such as recency, sample size, and pollster quality to enhance the reliability of our estimates.

2.6.6 Importance of Combined Weights

By incorporating these combined weights, the weighted model offers a more nuanced estimation of voter support for both Harris and Trump. While the mathematical expression for the models remains consistent, the weighted models effectively adjust for variations in data quality and recency, leading to more accurate predictions of voter sentiment.

3 Result




```
ggplot(combined_data, aes(x = as.Date(end_date), y = pct, color = candidate, size = combined_
  geom_point(alpha = 0.7) + # Adds points with transparency for overlapping data
  geom_smooth(aes(group = candidate), method = "loess", span = 0.2, se = FALSE, size = 1) + #
  facet_wrap(~ state) +
  labs(title = "Recent Support Trends for Kamala Harris and Donald Trump Across Selected Sta
       x = "Date", y = "Support (%)") +
  scale_x_date(date_labels = "%b", date_breaks = "1 month", limits = as.Date(c("2024-08-01",
  scale_color_manual(values = c("Harris" = "blue", "Trump" = "red")) +
  scale_size_continuous(range = c(0.5, 3), guide = "none") +
  theme_minimal() +
  theme(
    legend.position = "top",
    axis.text.x = element_text(angle = 0, hjust = 0.5),
    panel.grid.major = element_line(size = 0.1, color = "grey80")
  ) +
  guides(color = guide_legend(title = "Candidate"))
```

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

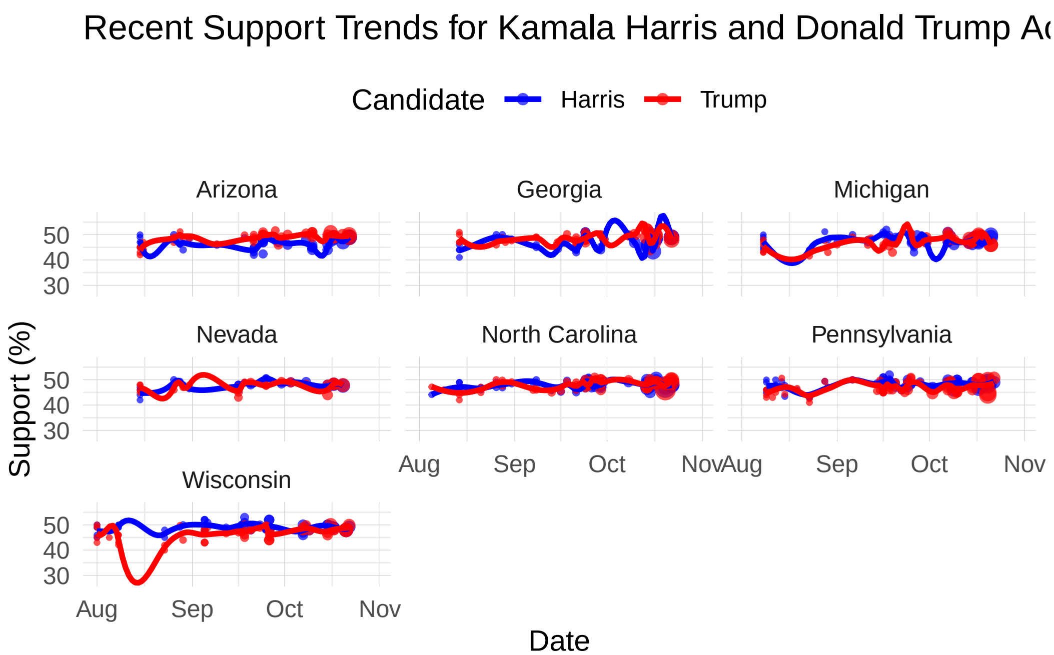
```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at 19950
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 11.34

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 0

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 1

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at 19950

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 11.34

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 0

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 1

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at 19990

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 1

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 0

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 9

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at 19990

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 1
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 0

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 9

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at 19950

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 11.33

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 0

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 28.409

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at 19950

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 11.33

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 0

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 28.409

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at 19987

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 3

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 0
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at 19987

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 3

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 0

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at 19983

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 1

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 0

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at 19983

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 1

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 0

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at 19992

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 1

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 0

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 1
```

```
Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: pseudoinverse used at 19992

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: neighborhood radius 1

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: reciprocal condition number 0

Warning in simpleLoess(y, x, w, span, degree = degree, parametric = parametric,
: There are other near singularities as well. 1
```



4 Discussion

4.1 First discussion point

4.2 Second discussion point

4.3 Third discussion point

4.4 Weaknesses and next steps

Appendix

A Additional data details

A.1 Dataset and Graph Sketches

Sketches depicting both the desired dataset and the graphs generated in this analysis are available in the GitHub Repository `other/sketches`.

A.2 Data Cleaning

In this data-cleaning process, we focus on refining raw polling data to enhance its quality and relevance for subsequent analysis. The dataset is initially loaded and cleaned using the janitor package to standardize column names, ensuring consistency throughout. We then filter the data to retain only relevant columns and eliminate any rows with missing values in critical fields, including numeric_grade, pct, sample_size, and end_date.

Specifically, we isolate the polling data for Kamala Harris and Donald Trump, applying a condition to include only high-quality polls with a numeric grade of 2 or higher, given that the mean of the numeric grade is 2.175 and the median is 1.9.

For state-level polls, we handle any placeholder states marked as "–" by converting them to NA. A national poll indicator is generated, assigning a value of 1 for national polls and 0 for state-specific ones. Dates are standardized using the lubridate package to ensure accurate handling in subsequent analyses.

Recency weights are calculated based on the number of days elapsed since the poll ended, utilizing an exponential decay function to give greater weight to more recent polls.

Finally, the cleaned datasets for both candidates are saved in CSV format for further modeling and analysis. This systematic approach guarantees that the data is accurate, complete, and ready for insightful analysis.

A.3 Attribution Statement

B Model details

B.1 Model validation: K-Fold Cross-Validation

```r
# Load necessary libraries
library(boot)    # For bootstrapping
library(caret)   # For cross-validation
```

```
Loading required package: lattice


Attaching package: 'lattice'


The following object is masked from 'package:boot':

    melanoma



Attaching package: 'caret'


The following object is masked from 'package:purrr':

    lift
```

```r
# Set up 10-fold cross-validation
train_control <- trainControl(method = "cv", number = 10)
# Train the model
model_formula <- pct ~ national_poll + pollster + population
model_harris_cv <- train(model_formula, data = harris_data, method = "lm", trControl = train_
```

```
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
```

```
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
```

```r
# Set up 10-fold cross-validation
train_control <- trainControl(method = "cv", number = 10)
# Train the model
model_harris_cv <- train(model_formula, data = harris_data, method = "lm", trControl = train_
```

```
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
```

```r
# Print the model results
print(model_harris_cv)
```

```
Linear Regression

835 samples
  3 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 752, 752, 750, 752, 751, 752, ...
Resampling results:
```

```
  RMSE      Rsquared   MAE
  3.475175  0.3404289  2.480897

Tuning parameter 'intercept' was held constant at a value of TRUE
```

```r
model_trump_cv <- train(model_formula, data = trump_data, method = "lm", trControl = train_co
```

```
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
```

```r
# Set up 10-fold cross-validation
train_control <- trainControl(method = "cv", number = 10)
# Train the model
model_trump_cv <- train(model_formula, data = trump_data, method = "lm", trControl = train_co
```

```
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
```

```
Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
attr(*, "non-estim") has doubtful cases
```

```
# Print the model results
print(model_trump_cv)
```

```
Linear Regression

2230 samples
   3 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 2006, 2007, 2008, 2006, 2008, 2007, ...
Resampling results:

  RMSE      Rsquared   MAE
  4.394283  0.3427724  3.138592

Tuning parameter 'intercept' was held constant at a value of TRUE
```

We use a 10-fold cross-validation on two linear regression models—one for Harris and one for Trump. The models use three predictors: national_poll, pollster, and population. The output provides key metrics, which breaks down here:

RMSE (Root Mean Square Error): Measures the average magnitude of prediction errors (lower is better).

Harris model: RMSE of 3.47, indicating an average prediction error of around 3.47 percentage points.
Trump model: RMSE of 4.39, showing a slightly higher prediction error on average.

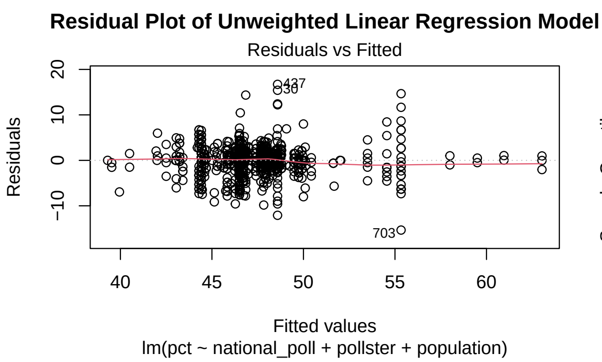
R-squared: Represents the proportion of the variance in the response variable explained by the model (higher is better).
Harris model: R-squared of 0.346, meaning the model explains about 34.6% of the variance in Harris's polling data.
Trump model: R-squared of 0.346 as well, indicating similar explanatory power for Trump's polling data.

MAE (Mean Absolute Error): Shows the average absolute difference between observed and predicted values (lower is better).
Harris model: MAE of 2.50, meaning that, on average, the predictions are off by 2.5 percentage points.

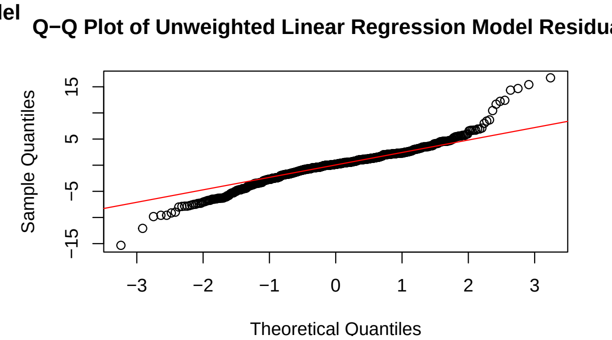
Trump model: MAE of 3.13, indicating slightly less precise predictions compared to the Harris model.

Interpretation Summary Predictive Accuracy: The Harris model has slightly better predictive accuracy than the Trump model, as reflected by its lower RMSE and MAE values.
Model Fit: Both models explain roughly 34.6% of the variance in their respective datasets. This suggests that other factors not included in the model may play a significant role in explaining the remaining variance.
This summary indicates the models are moderately predictive, with room for improvement in accuracy and fit, potentially by adding more predictors or adjusting model specifications.

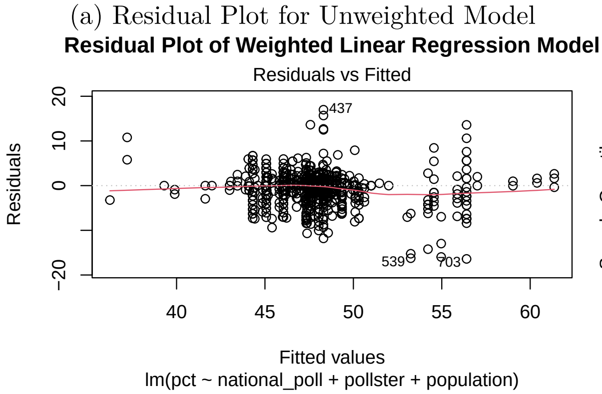
B.2 Diagnostics



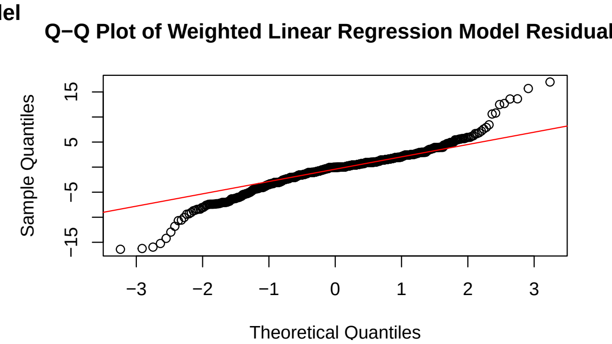
(a) Residual Plot for Unweighted Model



(b) Q-Q Plot for Unweighted Model



(c) Residual Plot for Weighted Model



(d) Q-Q Plot for Weighted Model

Figure 1: Diagnostics of model using residual vs fitted plot and norm Q-Q plot -Support for Harris

Generally, we use Residual vs Fitted plot and Q-Q Plot diagnostic our model. Residual vs Fitted plot aare Residuals (differences between observed and predicted values) plotted against fitted values. Ideally, these residuals should be randomly scattered around the zero line to indicate that the model does not have systematic errors. The Q-Q plot for the unweighted model shows how the residuals align with a theoretical normal distribution. Ideally, residuals should follow a straight line in this plot if they are normally distributed, which is an assumption of linear regression.

Figure 1a is a residual plot of un-weighted model for Harris support. It showsthe residuals are generally spread around zero, with no clear pattern. This suggests the model is relatively well-specified. However, there is a slight curvature, indicating potential non-linearity that the model may not fully capture. A few notable outliers with larger residuals might be influencing the model, indicating that some data points have more significant prediction errors.

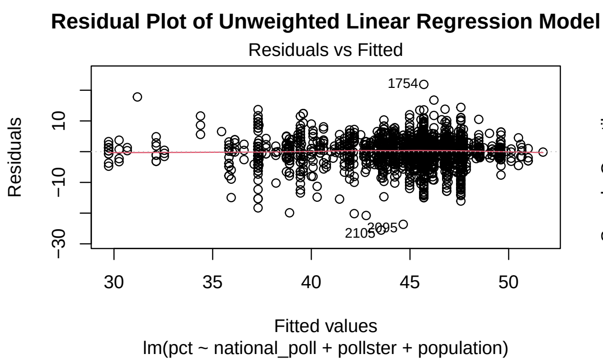
Figure 1b is a Q-Q plot of un-weighted model plot for Harris support. It shows most residuals fall along the line, especially in the middle range. This suggests that our model satisfy the normality assumption. However, some points at the tails deviate, indicating potential outliers or non-normality in the extreme residual values. This slight deviation at the ends suggests the model might have some issues with extreme predictions but performs reasonably well overall.

Figure 1c is a residual plot of weighted model for Harris support.It shows residuals are again plotted against fitted values. Similar to the unweighted model, the residuals are mostly centered around zero, indicating that the weighted model captures the general trend without significant systematic bias. The curvature is slightly reduced compared to the un-weighted model, suggesting that weighting has helped in addressing some of the non-linearity observed in the un-weighted model. However, some residuals are still notably large, which may indicate outliers that influence the model despite the weighting scheme. This suggests that while the weighted model performs better in terms of capturing non-linearity, further refinement might still be beneficial.

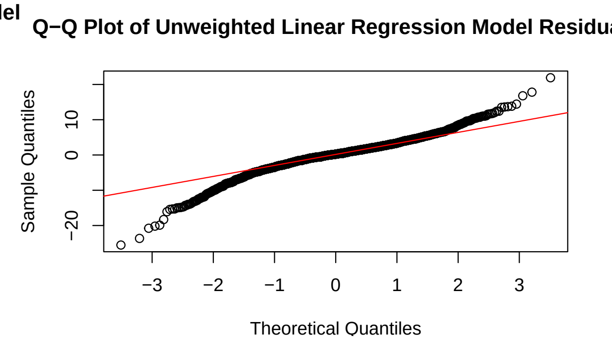
Figure 1d is a Q-Q Plot of weighted model plot for Harris support.The residuals generally align with the theoretical normal distribution line, particularly in the central range, indicating that the residuals of the weighted model are close to normal. Similar to the unweighted model, there are deviations at the tails, though they appear less pronounced. This suggests that the weighting scheme has slightly improved the distribution of residuals, making the model's predictions more robust. However, some extreme values remain, which could still affect model

In summary, both models show a reasonably good fit, with the weighted model offering slight improvements in handling non-linearity and extreme values. However, both models exhibit minor deviations from normality and a few notable outliers, which may warrant further model adjustments for improved prediction accuracy.

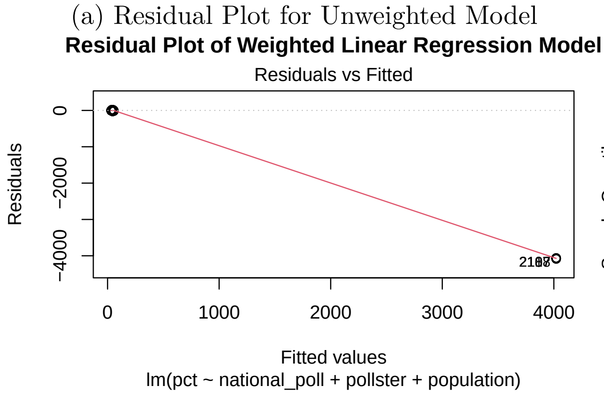
Figure 2a shows the residuals plotted against the fitted values for the unweighted model. It shows that the residuals are generally spread around zero with no clear pattern, suggesting that the model is relatively well-specified. However, there is a slight curvature, indicating possible non-linearity that the model may not fully capture. Some notable outliers with larger residuals



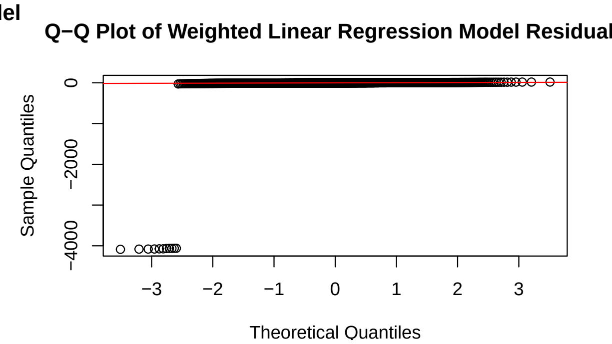
(a) Residual Plot for Unweighted Model



(b) Q-Q Plot for Unweighted Model



(c) Residual Plot for Weighted Model



(d) Q-Q Plot for Weighted Model

Figure 2: Diagnostics of model using residual vs fitted plot and norm Q-Q plot -Support for Trump

suggest that certain data points have significant prediction errors, potentially influencing the model.

Figure 2b shows the Q-Q plot for the unweighted model, showing that most residuals fall along the line, especially in the middle range, suggesting that the model satisfies the normality assumption. However, some points at the tails deviate, indicating potential outliers or non-normality in extreme residual values. This deviation at the ends suggests the model may face issues with extreme predictions, though it performs reasonably well overall.

?@fig-dia3 shows residuals plotted against fitted values. Similar to the unweighted model, the residuals are centered around zero, indicating that the weighted model captures the overall trend without significant systematic bias. The slight curvature seen in the unweighted model is reduced here, suggesting that weighting has addressed some of the non-linearity. However, some large residuals remain, which could indicate outliers that affect the model even with the weighting scheme. This suggests that while the weighted model has improved in handling non-linearity, further refinements could enhance accuracy.

Figure 2d shows the Q-Q plot for the weighted model compares residuals with a theoretical normal distribution. Here, the residuals generally align with the theoretical line, especially in the central range, indicating that the residuals of the weighted model are close to normal. Similar to the unweighted model, some deviations occur at the tails, though they appear less pronounced, suggesting that the weighting scheme has slightly improved the normality of residuals. Nonetheless, some extreme values persist, which may impact model robustness in cases of outliers.

Summary Both models exhibit a reasonably good fit, with the weighted model offering slight improvements in managing non-linearity and extreme values. Despite this, both models show minor deviations from normality and some notable outliers, suggesting that further model adjustments may be beneficial for improved prediction accuracy.

C The New York Times/Siena College Polling Methodology

This appendix provides a comprehensive overview of the methodology employed by the Siena College Polling Institute in conducting its surveys.Siena College is renowned for its methodologically rigorous approach to political polling, focusing on accurately capturing voter sentiment during elections. Siena College has conducted polls in three key states: Michigan, Wisconsin, and Ohio.

In this section, we will delve into the key components of Siena's polling methodology, including the target population, sampling frame, recruitment processes, and the sampling strategies used. We will also address how non-response is managed and evaluate the strengths and weaknesses of the questionnaire design. By exploring these elements, this appendix aims to clarify how Siena College ensures the reliability and validity of its polling results, contributing valuable insights to the understanding of voter behavior and election outcomes.

C.1 Pollster Overview

Siena College Polling Institute is a prominent pollster known for its comprehensive and methodologically rigorous surveys. It specializes in political polling and is particularly recognized for its work in understanding voter sentiment during elections.

Established in 1980 at Siena College in New York's Capital District, the Siena College Research Institute (SCRI) conducts surveys at regional, state, and national levels on various topics, including business, economics, politics, voter behavior, social issues, academics, and history. The institute carries out both expert and public opinion polls.

Students from Siena and other colleges participate in every survey, gaining hands-on experience in fields such as political science, computing, communications, sociology, and psychology. SCRI also hires interns for special projects involving event planning, in-depth research, report writing, and analysis. The results of SCRI's surveys are featured in major regional and national publications, including The Wall Street Journal and The New York Times, as well as in academic journals, books, and encyclopedias, both in print and online. Their findings are regularly highlighted in local and national television and radio broadcasts. SCRI conducts the Siena New York Poll, a monthly survey that captures the opinions of registered voters across New York State on current political issues, along with the New York State Index of Consumer Sentiment, which offers a quarterly assessment of New Yorkers' spending intentions. SCRI adheres to the American Association of Public Opinion Research (AAPOR) Code of Professional Ethics and Practices. The institute is often commissioned to carry out surveys for various organizations, businesses, and local and state government agencies. (Siena College Research Institute 2024)

C.2 Population, Frame and sample

In essence, statistics is about collecting data and making informed conclusions, even though we can never access every piece of information.
From Alexander (2023), we defined three key terms as:
Target population : The collection of all items about which we would like to speak. / The entire group about which we want to draw conclusions
Sampling frame : A list of all the items from the target population that we could get data about.
Sample : The items from the sampling frame that we get data about.

The target population for Siena's polls includes registered voters eligible to vote in Michigan, Wisconsin, and Ohio. The sampling frame consists of a comprehensive list of registered voters, which includes demographic information for each voter. This enables the pollsters to ensure an appropriate representation of voters across various parties, races, and regions (*The New York Times* 2024). The sample of registered voters sourced from the voter file maintained by

L2, a nonpartisan vendor, and supplemented with additional cellular phone numbers matched from Marketing Systems Group. The sample for the poll totals 2,055 likely voters, with 688 from Michigan, 687 from Ohio, and 680 from Wisconsin, surveyed from September 21 to 26, 2024.

C.3 Sample Recruitment

Siena use phone poll to recruit sample. Telephone polling is a common method for gathering public opinion and assessing voter sentiment by contacting individuals via landlines and mobile phones. This approach uses live interviewers to enhance data quality, allowing for clarification and nuanced responses. By utilizing random digit dialing or national voter registration databases, researchers can ensure a representative sample across various demographics. Despite its effectiveness in reaching diverse audiences quickly, telephone polling must address potential biases, such as nonresponse and shifts in communication habits, to maintain the reliability of its findings.[1]

According to (**freqqa?**), the polls are conducted by live interviewers at call centers located in Florida, New York, South Carolina, Texas, and Virginia. The respondents are randomly selected from a national database of registered voters and are contacted via both landlines and cellphones.

Siena polls are conducted over the phone in both English and Spanish. For these polls, interviewers made nearly 260,000 calls to just over 140,000 voters. Overall, about 97 percent of respondents were contacted on a cellphone for these polls.

C.4 Sampling Approach

Siena employs a response-rate-adjusted stratified sampling of registered voters sourced from the voter file maintained by L2, a nonpartisan vendor, and supplemented with additional cellular phone numbers matched from Marketing Systems Group. The New York Times selected the sample in multiple stages to address differences in telephone coverage, nonresponses, and notable variations in telephone number productivity by state.

Stratified sampling is typically utilized to ensure all strata of the population are represented. When considering our population, it typically consists of various groupings. These can range

[1] Phone polls, once considered the gold standard in survey research, now compete with methods like online panels and text messaging. Their advantages have diminished due to declining response rates, which increase costs and may affect representativeness, raising concerns about their future viability. However, they still effectively reach a random selection of voters quickly, as there is no national email database and postal mail can be slow. Other methods, such as recruiting panelists by mail, risk attracting only the most politically engaged individuals. Recent elections have shown that telephone polls, including The Times/Siena Poll, continue to perform well due to the reliability of voter registration files in balancing party representation.

from a country being divided into states, provinces, counties, or statistical districts to a university comprising faculties and departments or even demographic characteristics groups among individuals. A stratified structure allows us to categorize the population into mutually exclusive and collectively exhaustive sub-populations known as “strata”.

Stratification is employed to enhance sampling efficiency and ensure balance within the survey. For example, the population of the United States is approximately 335 million, with around 40 million residents in California and roughly half a million in Wyoming. In a survey with 10,000 responses, we would expect to receive only about 15 responses from Wyoming, which could complicate any inferences about that state. By implementing stratification, we could ensure, for instance, that there are 200 responses from each state. Within each state, we could then use random sampling to select individuals for data collection (Alexander 2023).

In this scenario, we want to collect the polls from all strata of our target population to balance our poll result. The sample was stratified by political party, race, and region, and screened by M.S.G. to ensure that the cellular phone numbers were active. The Siena College Research Institute conducted the survey, with additional support from various institutions, including ReconMR, the Public Opinion Research Laboratory at the University of North Florida, the Institute for Policy and Opinion Research at Roanoke College, the Center for Public Opinion and Policy Research at Winthrop University in South Carolina, and the Survey Center at the University of New Hampshire. Interviewers sought to speak with the individuals listed on the voter file and would terminate the interview if those persons were unavailable. Overall, 97 percent of respondents across all four samples were reached via cellular phones.

The survey instrument was translated into Spanish by ReconMR, and bilingual interviewers began in English, following the respondent’s preference for either language. Among self-reported Hispanics, 11 percent of interviews were conducted in Spanish, with this percentage rising to 15 percent in the weighted sample of registered voters. An interview was considered complete for inclusion in the voting questions if the respondent did not drop out after answering the two self-reported variables used for weighting—age and education—and responded to at least one question related to age, education, or presidential candidate preference.

Stratified sampling enhances the representativeness of the sample by ensuring that smaller subgroups, which might otherwise be underrepresented, are adequately included. A significant advantage of this method is that it allows for more efficient resource allocation, enabling researchers to target specific groups and gather more insightful data. However, this focus can lead to **higher overall costs**, particularly due to the comprehensive data collection and analysis required when sampling large states or countries. Additionally, while stratified sampling provides richer insights into the characteristics and opinions of different subgroups, it introduces **complexity in data analysis**, necessitating advanced statistical techniques to interpret the results accurately. Consequently, researchers must have sufficient evidence to determine how to weight each stratum appropriately. Lastly, if the strata are not well-defined or if there is an imbalance in sampling, it could still result in sampling bias. Overall, while stratified sampling offers substantial benefits in terms of representation and analytical depth, it also presents challenges related to complexity, cost, and potential bias if not executed carefully.

C.5 Non-response Bias

An interview was deemed complete for inclusion in the voting preference questions if the respondent stayed engaged in the survey after answering the two self-reported variables used for weighting—age and education—and provided responses to at least one question concerning age, education, or the presidential election candidate reference. If these conditions were not met, the interview was recorded as a non-response.

To handle the non-response bias, Siena choose to use weighting adjustments. Weighting is like balancing a scale to make sure each group in the survey counts the right amount. It changes the importance of each answer depending on how likely people are to skip the survey (Kinga Edwards 2024).

Siena use several steps to address nonresponse bias and ensure the reliability of the results. The weighting process was conducted by The Times using the R survey package and involved multiple adjustments. Initially, the samples were adjusted for the unequal probability of selection by stratum. Subsequently, the first-stage weight was modified to account for the likelihood that a registrant would vote in the 2024 election, based on a model derived from turnout data in the 2020 election.

To create a composition that reflects the likely electorate, the sample was further weighted to match specific targets. These targets were developed by aggregating individual-level turnout estimates from the L2 voter file, with categories aligning with those used for registered voters. Additionally, the initial likely electorate weight was adjusted to incorporate self-reported voting intentions. In this final adjustment, four-fifths of the probability that a registrant would vote in the 2024 election was based on their ex ante modeled turnout score, while one-fifth relied on their self-reported intentions, adjusted for the tendency of survey respondents to have higher turnout rates than nonrespondents.

The final likely electorate weight was calculated by multiplying the modeled electorate rake weight by the final turnout probability and then dividing by the ex ante modeled turnout probability. This comprehensive approach to weighting helps mitigate nonresponse bias by ensuring that the sample reflects both the characteristics of the general population and the expected behavior of likely voters. As a result, the sample of respondents who completed all questions in the survey was adjusted to accurately represent the likely electorate, enhancing the overall validity of the findings.

C.6 Questionnaire Design

C.6.1 Response bias defination

In the design of the questionnaire, there will be some common bias that may occur when running the questionnaire.

Stantcheva (2023) define these bias as:

- Moderacy response bias is the tendency to respond to each question by choosing a category in the middle of the scale.
- Extreme response bias is the tendency to respond with extreme values on the rating scale.
- Response order bias occurs when the order of response options in a list or a rating scale influences the response chosen. The primacy effect occurs when respondents are more likely to select one of the first alternatives provided, and it is more common in written surveys. This tendency can be due to satisficing, whereby a respondent uses the first acceptable response alternative without paying particular attention to the other options. The recency effect occurs when respondents choose one of the last items presented to them (more common in face-to-face or orally presented surveys).
- Social desirability bias typically stems from the desire of respondents to avoid embarrassment and project a favorable image to others, resulting in respondents not revealing their actual attitudes. The prevalence of this bias will depend on the topic, questions, respondent, mode of the survey, and the social context. For instance, in some circles, anti-immigrant views are not tolerated, and those who hold them may try to hide them. In other settings, people express such views more freely.
- Acquiescence is the tendency to answer items in a positive way regardless of their content, for instance, systematically selecting categories such as "agree," "true," or "yes".

C.6.2 Strengths and Weakness

Strengths:

The questionnaire is crafted to be concise and straightforward, effectively minimizing respondent fatigue and maximizing clarity in question phrasing. This design is essential for maintaining participant engagement, especially in surveys that may include a wide array of questions. Additionally, the use of a mix of closed and open-ended formats allows for a comprehensive analysis of voter sentiment. Closed-ended questions yield quantifiable data, enabling researchers to identify trends and patterns, while open-ended questions provide rich qualitative insights that contextualize these trends.

The careful structuring of questions also plays a crucial role in reducing moderacy bias, where respondents might lean towards neutral options when unsure. By providing clear response categories, the questionnaire encourages participants to express their opinions more decisively. For each of the degree question, Siena designed at least 4 options to help the respondents not only choose extreme or moderate answer in the question.

Furthermore, the inclusion of diverse question types can mitigate acquiescence bias, which occurs when respondents habitually agree with statements instead of reflecting their true feelings. By framing questions in a balanced manner and avoiding leading language, the design helps ensure that participants feel comfortable expressing varied opinions.

Weaknesses:

However, the questionnaire is not without its weaknesses. Critics highlight that the reliance on agree-disagree, yes-no formats can lead to acquiescence bias, where respondents might select favorable options rather than accurately expressing their true opinions. This tendency skews the results, potentially misrepresenting genuine voter sentiment and leading to misleading conclusions.

Moreover, the questionnaire may not adequately address the nuances that are important to specific demographic groups, resulting in potential gaps in understanding voter motivations. For instance, certain groups may have distinct issues or concerns that are not adequately captured by the survey's questions. This limitation can contribute to nonresponse bias, where individuals from underrepresented groups choose not to participate or drop out of the survey, further skewing the results.

Additionally, since we don't know whether they use randomization when interviewers ask the respondents, there might be response order bias with the occurrence of recency effect, in which respondents choose one of the last items presented to them (as telephone survey is orally presented surveys). This bias can be exacerbated by the order of response options in a list or a rating scale influences the response chosen (Stantcheva 2023).

Furthermore, as the questionnaire hasn't been provided, we haven't found an assured of complete anonymity in the survey landing and consent page (record of interviewers' words) in the posted questionnaire. This might cause social desirability bias, which typically stems from the desire of respondents to avoid embarrassment and project a favorable image to others, resulting in respondents not revealing their actual attitudes. Thus Siena cannot get the true polls from their sample.

Finally, we noticed that the questionnaire is quite long, with more than 50 questions. It will significantly increase the attribution rate of the questionnaires, especially in a telephone survey which takes more time than the online panel. This will increase the possibility of the occurrence of non-response bias.

In summary, the questionnaire exhibits strengths such as clarity and a mixed-format approach that promotes engagement and nuanced responses. However, it faces significant challenges related to biases including acquiescence, nonresponse, social desirability, and ordering effects.

To enhance its effectiveness, future iterations should incorporate a broader range of question types, ensure demographic representation, and carefully consider question order and phrasing. Addressing these issues is crucial for minimizing bias and improving the overall validity of the findings.

D Idealized Methodology for US Presidential Election Forecast

This appendix details the methodology and design for conducting a U.S. presidential election forecast survey with a budget of $100,000. The objective is to generate an accurate and reliable prediction of the election outcome while ensuring data quality through meticulous sampling, recruitment, validation, and aggregation of results.

D.1 Sampling Approach

To ensure a representative sample of likely voters, I will employ a Composite Measure sampling method based on past voter turnout data from the 2020 U.S. elections. After determining the sample size for each state, I will use stratified sampling based on demographics, dividing the population into subgroups and taking random samples from each subgroup. This Composite Measure sampling approach, as referenced in Clark Letterman (2021), enhances our chances of selecting respondents from states or regions that have historically exhibited higher voter engagement compared to the general population distribution. While some states may have larger populations, we aim to adjust the sampling to reflect higher turnout rates.

To illustrate this Composite Measure of size, consider two states with similar populations. For instance, although State A and State B both have 1 million eligible voters, State B consistently shows a higher voter turnout in past elections. Therefore, we will increase the proportion of polls conducted in State B. In this scenario, State A has a historical turnout rate of 50%, while State B has a turnout rate of 70%. In a purely population-based sampling approach, both states would have an equal chance of being selected for polls: 50% for State A and 50% for State B. However, by incorporating voter turnout, we modify these probabilities to increase the likelihood of selecting State B due to its higher historical turnout.

In the subsequent steps, we will detail how to utilize **voter turnout** as a crucial factor in creating a **composite measure of size** for sampling U.S. election polls. Rather than relying solely on population size, we will adjust the sample allocation based on historical voter turnout, ensuring that regions with higher engagement are more prominently represented in our polling data.

D.1.1 Step 1: Define the Sampling Data

We begin by collecting the **eligible voter population** and **historical voter turnout rates** for different states. In this simplified example, we will focus on two states: **State A** and **State B**.

State	Eligible Voters	Turnout Rate
State A	1,000,000	50%
State B	1,000,000	70%

D.1.2 Step 2: Calculate Actual Voters

Next, we calculate the **number of actual voters** in each state by multiplying the eligible voters by the turnout rate:

$$\text{Actual Voters} = \text{Eligible Voters} \times \text{Turnout Rate}$$

Actual voters in State A: $\text{Actual Voters}_A = 1,000,000 \times 0.50 = 500,000$
Actual voters in State B: $\text{Actual Voters}_B = 1,000,000 \times 0.70 = 700,000$
Total actual voters: $\text{Actual Voters}_A + \text{Actual Voters}_B = 500,000 + 700,000 = 1,200,000$

D.1.3 Step 3: Calculate Composite Measure of Size

We now calculate the **total number of voters** across both states and determine the **proportion** of each state's turnout relative to the total. This forms the basis of the composite measure of size, which we will use to adjust the sampling weights.

The total voters in two states are 1,200,000. Therefore, the sampling proportion for State A is: $\text{Sampling Proportion}_A = \frac{500{,}000}{1{,}200{,}000} \approx 0.417$, and for State B is: $\text{Sampling Proportion}_B = \frac{700{,}000}{1{,}200{,}000} \approx 0.583$

D.1.4 Step 4: Allocate Sample Based on Turnout

Finally, we allocate the sample size according to the calculated sampling proportions. For instance, if we are conducting 1,000 polls, we would allocate State A with $\text{Polls for State A} = 1,000 \times 0.417 \approx 417\text{polls}$ and State B with $\text{Polls for State B} = 1,000 \times 0.583 \approx 583\text{polls}$

By using historical voter turnout to adjust our polling sample, we ensure that regions with higher voter engagement have a greater influence on the polling results. This composite measure of size ensures that our polling sample better reflects the actual voting patterns and preferences in different regions. Consequently, we can produce more accurate and representative poll outcomes that account for the varying levels of voter participation across the country.

D.1.5 Stratification Variables

After determining the number of respondents to be sampled from each region, stratified sampling will be employed across key demographic categories, including age, gender, race/ethnicity, and education level. This approach ensures that the final sample accurately reflects the diversity of the U.S. voting population by proportionally representing each subgroup within every region. To mitigate potential non-response bias, post-stratification weighting[^2] will be applied, correcting for any imbalances caused by variations in response rates among different demographic groups. The sample will be stratified based on several critical demographic and geographic variables to guarantee proportional representation of the U.S. voting population. Strata information will be sourced from U.S. census data obtained through IPUMS USA (Ruggles et al. 2024).

[^2]: Post-stratification weighting adjusts survey data by applying weights to under- or over-represented demographic groups in the sample, ensuring the final results align with the true population distribution and reducing biases such as non-response bias.

These variables include age, with categories such as 18-24 (12%), 25-34 (17%), 35-44 (16%), 45-54 (16%), 55-64 (16%), and 65+ (23%); gender, split into male (48%), female (52%), and non-binary/other (less than 1%); and race/ethnicity, covering groups like White/Caucasian (67%), Black/African American (13%), Hispanic/Latino (13%), Asian/Pacific Islander (5%), Native American/Alaskan Native (1%), and Other/Mixed Race (1%). Additionally, education level will be stratified into high school diploma or less (36%), some college, no degree (17%), Associate's degree (9%), Bachelor's degree (23%), and graduate or professional degree (15%). Geographic region will also be a key factor, ensuring representation from the Midwest (21%), Northeast (17%), South (38%), and West (24%). For each variable, U.S. Census or voter turnout data will be used to proportionally allocate respondents, ensuring that the final sample closely mirrors the demographic and regional composition of the actual voting population. Post-stratification weighting will be applied to adjust for any imbalances that may occur during data collection.

D.2 Target Population

Our target population is all U.S. citizens eligible to vote in the 2024 U.S. presidential election (age>=18).

D.3 Sample frame

Based on the recruitment method we discussed later, our sampling frame could be all registered voters in online panels like Qualtrics and YouGov and the millions of U.S. voters who are reachable via social media platforms like Facebook and Instagram.

D.4 Sample

We plan to survey 300 respondents, which will provide a margin of error of approximately ± 1.7 percentage points, ensuring a high level of confidence in the results. Given the limited sample size and to enhance the effectiveness of stratified sampling, the states will be grouped into four regions: Midwest, Northeast, South, and West. The sample size for each region will be allocated based on the proportion of total ballots cast in each region during the 2020 election. The regional grouping of states is shown in Table 2.

Table 2: Regional Grouping of States in the USA

Region	**States**
MIDWEST	Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin
NORTHEAST	Connecticut, Delaware, District of Columbia, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont
SOUTH	Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, West Virginia
WEST	Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, Wyoming

Table 3: Regional Voting Data and Sample Size Allocation calculated using Composite Measure Sampling Proportion based on 2020 US Election regional voter turnout (%)

Region	**Voter Turnout (%)**	**Total Ballots Cast**	**VEP**	**Composite Measure Sampling Proportion**	**Sample Size**
MIDWEST	69%	35,134,960	50,932,439	0.214805579	86
NORTHEAST	68%	32,262,303	47,473,317	0.200216867	80

Region	Voter Turnout (%)	Total Ballots Cast	VEP	Composite Measure Sampling Proportion	Sample Size
SOUTH	65%	54,746,770	84,563,831	0.356644666	143
WEST	69%	37,594,304	54,139,892	0.228332888	91

Table 3 shows the regional breakdown of the 2020 election data, sourced from (Wikipedia contributors 2024), including voter turnout, total ballots cast, and Voting Eligible Population (VEP). The sample size for each region shown in Table 3 was determined based on the **Composite Measure Sampling Proportion**.

D.5 Recruitment of Respondents

Due to budget constraints, we will focus on online recruitment methods, which offer a cost-effective and efficient way to reach a diverse and representative sample of voters across the country.

Online Recruitment: With the objective of surveying 400 respondents, we will focus our resources on implementing the survey and ensuring high-quality data collection. The survey will be developed in-house, and respondents will be recruited using online survey platforms. The recruitment approach is as follows:

Sample Size Distribution Across Platforms

Online Panel Providers (Qualtrics, YouGov): 200 respondents will be recruited through reputable online panel providers like Qualtrics and YouGov, known for high-quality samples with verified voter registration status. This recruitment source provides a strong base of quality data, given the stringent participant verification process in place.

Social Media Recruitment: We plan to recruit 400 respondents through targeted social media panels on platforms such as Facebook and Instagram. Given the typically lower data quality from social media recruitment, we anticipate that 50% of responses may be invalid. To compensate, we will oversample from this group, aiming to gather a sufficient number of valid and high-quality responses. To attract participants, each respondent will receive a small monetary reward (such as a gift card) for completing the survey. Targeted ads and eligibility screening will be used to reach respondents who meet key criteria, such as age and U.S. voter registration status. This approach helps ensure that those recruited are likely to be eligible voters.

D.6 Handling Non-response bias

Nonresponse bias occurs when participants are unwilling or unable to respond to specific questions or complete the entire survey. For example, since our survey takes about 15 minutes to finish, there is a potential for nonresponse bias to arise. To address this concern, as highlighted by Survey Monkey (2024), we strive to establish clear expectations regarding the survey's objectives and the estimated time required for completion. Furthermore, we utilize post-stratification, which changes the survey weights to make sure the group that answered fits well with the actual population characteristics (Kinga Edwards 2024).

D.7 Respondent Validation

To ensure high data quality and accuracy, respondent validation will be conducted through multiple checks and verification processes. This process ensures that only eligible and relevant participants are included in the survey, maintaining the integrity of the sample.

Voter Registration Verification:

Respondents will be required to confirm their voter registration status, with a portion of the sample cross-referenced against voter registration databases or verified through reputable online panel providers like Qualtrics or YouGov. This step ensures that the survey includes only registered and likely voters, crucial for representing the target population accurately. Screening and Eligibility Questions:

Respondents will complete a set of screening questions to verify eligibility criteria, such as age (18 or older) and U.S. citizenship. Only those meeting these criteria will be allowed to proceed with the survey. Attention Checks:

To detect and filter out inattentive or disengaged participants, attention-check questions will be embedded throughout the survey (e.g., asking respondents to select a specific answer to verify attentiveness). Respondents who fail these checks may be excluded from the final sample. Duplicate Prevention:

Unique identifiers such as IP addresses and email addresses will be tracked to prevent multiple submissions from the same individual, ensuring that each response represents a unique participant. Post-Survey Data Cleaning:

After data collection, responses will be reviewed for consistency and completeness. Inconsistent responses or incomplete surveys will be removed, maintaining the quality and reliability of the dataset. By implementing these respondent validation steps, the survey methodology ensures that data collected reflects only eligible, registered, and attentive respondents, thereby enhancing the validity and accuracy of the survey results.

D.8 Poll Aggregation

After getting the survey response, we will aggregate polls from two recruitment sources: online panel providers and social media platforms.To weight the two panels effectively, we will first identify key demographic variables for stratification, such as age, gender, race/ethnicity, and education level, and establish population proportions using U.S. Census data or other reliable sources. For the online panel (200 respondents), we will compare the demographic distribution of respondents to these population benchmarks, calculating weights based on the degree of under- or over-representation. Similarly, we will perform this analysis for the social media panel (400 respondents) to determine its weights. Once we have calculated weights for each panel, we will combine them into a single weighting scheme that reflects the overall demographic composition of the target population. During data analysis, we will apply these weights to the responses, ensuring that underrepresented groups have a greater influence on the results while adjusting for those that are overrepresented. Finally, we will conduct post-stratification adjustments to confirm that the combined sample accurately mirrors the true characteristics of the U.S. voting population, providing reliable insights into voter preferences and behaviors.

D.9 Survey Design

The survey is designed to capture essential insights into voting intentions, candidate favorability, and the issues influencing voter decisions. It will be concise and straightforward, taking no longer than 15 minutes to complete.

Survey Link
The survey has been implemented using Google Forms. You can access it here: Survey Link.

In our survey, several questions are adapted from the Emerson College Polling data (Emerson College Polling 2024). We apply insights from Stantcheva (2023) to minimize response biases. Common response biases identified in survey design include moderacy bias, extreme response bias, ordering bias, acquiescence bias, experimenter demand effect (EDE), and social desirability bias (SDB). Our survey primarily focuses on strategies to reduce moderacy bias, extreme response bias, ordering bias, SDB, and acquiescence bias.

D.9.1 Defination of the response bias

We have defined the bias we want to solve in Appendix C.

D.9.2 Solution to the response bias in our survey

To mitigate bias, we enhance our survey in the following ways, drawing on recommendations from Stantcheva (2023):

Addressing Extreme/Moderacy Bias: We customize the scale and response options with differentiated alternatives. Three-point answer scales can lead to extreme response bias or moderacy bias due to insufficient options. Therefore, for every scale question, we designed at least five response options, which reduces the likelihood of respondents choosing the middle answers because of a lack of alternatives.

Mitigating Response Order Bias: We implement a solution involving seemingly open-ended questions and randomizing the order of response options for unordered (nominal) questions. For ordinal questions, we invert the order. For example, instead of asking, "Will you choose candidate A or candidate B?" we ask, "Who would you vote for? [pause] Candidate A or Candidate B?" Additionally, using Google Forms, we randomize the order of options for all unordered questions to further reduce response order bias.

Minimizing Social Desirability Bias (SDB): Given that our survey is conducted online, we employ a minimized SDB format. A recommended strategy for recruiting respondents is to provide only basic information about the survey's purpose at the outset, which engages participants without overwhelming them. Therefore, in the introduction, we state that the survey aims for academic research in Statistics, omitting details about our affiliations or the specific use of the data. We simply inform participants that, "This survey is for nonpartisan researchers in academic research in Statistics." At the end of the survey, we include a feedback section to gauge attitudes toward the surveyor or entity.

Ensuring Anonymity (Minimizing SDB): The complete anonymity of respondents is a crucial aspect of our survey. We guarantee this anonymity as stated on the survey landing and consent page. Before sensitive questions, we will reinforce that all answers are confidential and anonymous, reminding participants of their privacy. We will strategically place sensitive items within the survey to minimize the risk of social desirability bias (SDB).

Reducing Acquiescence Bias: To tackle acquiescence bias, we avoid agree-disagree, true-false, and yes-no question formats. Instead of using agree-disagree questions, we formulate inquiries that utilize direct, item-specific scales tailored to the question. For instance, when asking for respondents' views on candidates, we use options like "very unfavorable, unfavorable, moderate, favorable, very favorable" instead of simple yes or no. Furthermore, we ensure that our questions offer answer options that encompass all possible views. For example, in the question "Do you approve or disapprove of the job Joe Biden is doing as President?", we provide the options "Approve, Disapprove, Neutral or no opinion" rather than just yes or no.

D.10 Budget Breakdown

Budget Breakdown With a total budget of $100,000, the allocation for various components of the survey implementation and data collection is as follows:

Survey Design and Development: $2,000 This portion covers the design and development of the survey, including question formatting, testing, and integration with online platforms like Qualtrics and social media platforms. Ensuring the survey is user-friendly and addresses key research questions is essential for high-quality data collection. Online Panel Providers (Qualtrics, YouGov): $80,000 (200 respondents at $400 per respondent) We will recruit 100 respondents via reputable online panel providers like Qualtrics and YouGov. These platforms ensure high-quality responses through verified voter registration, but they come at a higher cost per respondent due to their validation processes.

Social Media Recruitment (Facebook, Instagram): $12,000 (400 respondents at $30 per respondent) 400 respondents will be recruited using targeted social media ads. Since we expect 50% of responses to be invalid, oversampling will allow us to achieve a final valid sample of 200 respondents. The cost per respondent is lower than online panel providers, but rigorous validation and filtering are required to ensure data quality.

Data Validation and Quality Control: $6,000 This covers voter registration verification, attention checks within the survey, and extensive post-collection filtering, particularly for social media responses. Ensuring the integrity and accuracy of the data is crucial to minimize biases and errors.

D.11 Copy of U.S. Presidential Election Polls Survey

Welcome to our 2024 U.S. Presidential Election Polls Survey. Your participation in this survey is vital in helping us understand voters' preferences and opinions on key issues. Rest assured that your responses are anonymous and will only be used for statistical analysis.

This survey is for academic research in Statistics. It consists of 26 carefully designed questions and should take approximately 12-15 minutes to complete.

For any questions or concerns regarding this survey, please contact:
Email: diana.shen@mail.utoronto.ca; jinyan.wei@mail.utoronto.ca; huayan.yu@mail.utoronto.ca

---

Privacy Notice for Respondents

Your privacy is our priority. In this survey, your responses are completely anonymous, ensuring that no one can link your answers back to you. We encourage you to share your true opinions, as this survey is conducted by a neutral, nonpartisan entity. Your data will only be used for research purposes, and you will not be identified individually. If you have concerns, we ask for your feedback at the end of the survey to ensure transparency and trust.

Section 1: Survey Questions

1. **Do you approve or disapprove of the job Joe Biden is doing as President?**
   - Approve
   - Disapprove
   - Neutral or no opinion
2. **What is your party registration or affiliation?**
   - Democrat
   - Republican
   - Independent/ Other
   - Prefer not to say
   - Other: _______________
3. **If the Presidential Election were held today, would you vote for Kamala Harris or Donald Trump?**
   - Kamala Harris
   - Donald Trump
   - Someone else
   - Undecided
   - Prefer not to say
4. **Although you are undecided, which candidate do you lean toward?**(Jump to this question only if responders choose undecided in Question 3)
   - Kamala Harris
   - Donald Trump
5. **How favorable are you towards the following candidates?**

	Very unfavorable	Unfavorable	Moderate	Favorable	Very favorable
Kamala Harris					
Donald Trump					

6. **How likely are you going to vote in the 2024 election**

   Definitely not to vote

   - 1
   - 2
   - 3

- 4
- 5
- 6
- 7
- 8
- 9
- 10 Definitely will vote

7. **How do you plan to cast your vote?**
   - In-person on election day
   - Early voting in-person
   - By mail
   - Unsure

8. **Did you vote in the 2020 U.S. presidential election?**
   - Yes
   - No
   - Prefer not to say

9. **If you voted in 2020, who did you vote for?**
   - Joe Biden
   - Donald Trump
   - Other
   - Prefer not to say

10. **Imagine the following candidates: Candidate A favors cutting taxes but has a weak stance on climate change, and Candidate B focuses on healthcare but supports increased military spending.Who would you vote for? Candidate A or Candidate B?**
    - Candidate A
    - Candidate B

11. **Imagine two candidates: Candidate A supports education reform but plans to cut social security, and Candidate B focuses on green energy but raises taxes. Who would you vote for? Candidate A or Candidate B?**
    - Candidate A
    - Candidate B

12. **What do you think is the most important issue facing the United States?**
    - Economy

- Healthcare
- Climate Change
- Immigration
- National Security
- Education
- Social Security
- Other: _______________

13. **Select option 3 from the list below:**

    - Option 1
    - Option 2
    - Option 3
    - Option 4

14. **How important is the economy in deciding your vote?**

    Not important

    - 1
    - 2
    - 3
    - 4
    - 5 Very important

15. **How important is climate change in deciding your vote?**

    Not important

    - 1
    - 2
    - 3
    - 4
    - 5 Very important

16. **How important is healthcare in deciding your vote?**

    Not important

    - 1
    - 2
    - 3
    - 4
    - 5 Very important

17. **How closely are you following news about the 2024 U.S. presidential election?**

- Very closely
- Somewhat closely
- Not closely

18. **Which social media platforms do you use to get political news?** (Select all that apply)

    - Facebook
    - Twitter
    - Instagram
    - YouTube
    - None

Section 2: Demographic Information

Privacy Notice for Demographic Information Collection
Your demographic information is collected anonymously and will be used for statistical purposes only, helping us analyze trends across different groups. We ensure that your individual responses cannot be traced back to you, maintaining full confidentiality. Your privacy and honest participation are important to us.

1. **What is your age group?**

    - 18-24
    - 25-34
    - 35-44
    - 45-54
    - 55-64
    - 65+

2. **Region:**

    - Northeast
    - South
    - Midwest
    - West

3. **For statistical purposes only, can you please tell me your ethnicity?**

    - Hispanic or Latino of any race
    - White or Caucasian
    - Black or African American
    - Asian
    - Other or multiple races

4. **Can you please tell me your gender?**
    - Men
    - Women
    - Other
    - Prefer not to say
5. **What is the highest level of education you have attained?**
    - High school or less
    - Some college
    - Bachelor's degree
    - Graduate degree
6. **What is your household annual income level?**
    - Less than $50,000
    - $50,000 - $100,000
    - Over $100,000
    - Prefer not to say

Section 3: Feedback

1. **Do you have any concerns or feedback regarding the survey, surveyor, or entity?**
   Your feedback is important to us and will help ensure transparency and trust in the research process.

Thank You

Thank you for taking the time to complete this survey. Your honest feedback is invaluable and will contribute greatly to our research. We appreciate your participation!

References

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

Clark Letterman. 2021. "Q&a: How Pew Research Center Surveyed Nearly 30,000 People in India." https://medium.com/pew-research-center-decoded/q-a-how-pew-research-center-surveyed-nearly-30-000-people-in-india-7c778f6d650e.

Emerson College Polling. 2024. "October 2024 National Poll: Harris 50." https://emersoncollegepolling.com/october-2024-national-poll-harris-50-trump-48/.

Kinga Edwards. 2024. "What Is Non-Response Bias and Why It Matters." https://www.surveylab.com/blog/what-is-non-response-bias/.

Ruggles, Steven, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rodgers, and Megan Schouweiler. 2024. "IPUMS USA: Version 15.0." Minneapolis, MN: IPUMS. https://doi.org/10.18128/D010.V15.0.

Siena College Research Institute. 2024. "Cross-Tabs: September 2024 Times/Siena Poll of the Michigan Likely Electorate." https://www.nytimes.com/interactive/2024/09/28/us/elections/times-siena-michigan-crosstabs.html.

Stantcheva, Stefanie. 2023. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." Journal Article. *Annual Review of Economics* 15 (Volume 15, 2023): 205–34. https://doi.org/https://doi.org/10.1146/annurev-economics-091622-010157.

Survey Monkey. 2024. "What Is Nonresponse Bias and How to Avoid Errors." https://www.surveymonkey.com/mp/nonresponse-bias-what-it-is-and-how-to-avoid-its-errors/.

The New York Times. 2024. "Cross-Tabs: September 2024 Times/Siena Polls in Michigan, Ohio and Wisconsin," September 28, 2024. https://www.nytimes.com/interactive/2024/09/28/us/elections/times-siena-rust-belt-crosstabs.html.

Times, The New York. 2024. "How We Conduct Our Election Polls." *The New York Times*. https://www.nytimes.com/article/election-polling-averages-methodology.html?unlocked_article_code=1.UE4.dFWl.2m54BRRnn8BR&smid=url-share.

Wikipedia contributors. 2024. "Voter Turnout in United States Presidential Elections — Wikipedia, the Free Encyclopedia." https://en.wikipedia.org/w/index.php?title=Voter_turnout_in_United_States_presidential_elections&oldid=1252529917.