Polling the Polls:
Assessing 2024 Election Poll Accuracy with Multiple Linear Regression

Abstract:
Presidential election polls have severely underperformed in the previous two elections, but remain an integral piece of the U.S. political machine. The goal of this research is to investigate the performance of and factors influencing 2024 general election poll accuracy using multiple linear regression. Our analysis considers a sample of 1933 polls conducted from July 21, 2024, to Election Day, using data collected by FiveThirtyEight. Our main variables of interest were: *poll accuracy* (response), the region in which the poll was conducted, poll *methodology*, *sample size*, and *time difference*, among others. We found evidence of a positive relationship between *poll accuracy* and log(*sample size*), as well as a positive relationship between *swing state* and *poll accuracy*, all with a significance threshold of $\alpha = 0.05$, in our final model. These results make intuitive sense and are in accordance with our initial hypotheses. An F-test yielded that our model provides a better fit than an intercept-only model ($p < 2.2 * 10^{-16}$), with an R^2_{adj} of .1486. Our analysis is useful at evaluating the important predictors in predicting poll accuracy, but future models that include more robust estimation techniques or regularizations (e.g., LASSO) could provide more accurate predictions.

I. Introduction and Background:

Polls and other political surveys influence nearly every facet of the American electoral system — from campaign strategy and donor behavior to post-election analysis and, most importantly, voter decision-making. Given their importance, it is no surprise that perhaps the second-most pressing question voters ask during election season is: "Will the polls be accurate this time?"

In 2016, state-level polls largely oversampled Hillary Clinton supporters, thereby misleading public expectations – when turnout among rural, Republican-leaning voters exceeded forecasts, it became a decisive factor in Donald Trump's victory (Gayman, 2024). Four years later, the polls correctly predicted Joe Biden's win but were still inaccurate in regards to his margin of victory; a Scientific American study found that the margin of Biden's victory was overestimated in 93% of polls (Parshall, 2024). In response to the inaccuracy, many pollsters adjusted their methodologies leading up to the 2024 election cycle. Rather than relying primarily on demographic weighting, these new approaches placed great emphasis on respondents' past voting behavior (Paz, 2024).

Our research aims to assess how successful the 2024 polls were in predicting the presidential election outcome. We are similarly interested in understanding which characteristics were the most important predictors of poll accuracy. In regards to the latter, we first hypothesized that polls taken closer to the election would be more accurate than those taken further away from the election. Our second hypothesis was that polls conducted in swing states would better predict the popular vote than non-swing state polls. We thirdly expected that polls with larger sample sizes would outperform those with smaller sample sizes, due to the decreased variability risk. Lastly, we anticipated an interaction between polling methodology and poll sample size, as we expected that the relationship between a poll's accuracy and its methodology would vary based on the poll's sample size.

II. Data and Exploratory Analysis

a. Data and Variables

Our dataset is from FiveThirtyEight (538), an independent aggregator of datasets pertaining to sports, science, politics, and other fields. The original dataset included 18,095 observations, with each representing the percent share of the vote for a singular candidate from an individual poll in an individual state conducted in the 2024 election cycle (FiveThirtyEight, 2018). Our final dataset was narrowed down to exclude polls taken prior to Vice President Harris' acceptance of the Democratic nomination on July 21, 2024, along with cases whose candidate name was not Donald Trump – we selected Trump as our point of comparison due to the fact that he won the election and popular vote. Another 403 polls were removed due to missing values, and our final dataset thus consisted of 1933 polls taken at both the national and state levels with fifty-three initial total variables. Our variables of interest included *poll accuracy*, *polling methodology*, *region*, *partisanship*, *transparency score* (index), *sample size*, *time difference*, *swing state*, *voter population*, and *pollscore* (index). The majority of our variables were collected by 538 with the key exception of our numerical response variable, *poll accuracy*, and two other categorical variables mentioned later, *swing state* and *time difference*. *Poll accuracy* was constructed by subtracting the true proportion of the popular vote that President-Elect Donald Trump won (49.91%) from the share of the vote Trump was projected to win in each poll. The variables *pollscore* and *transparency score* were constructed by 538 – *pollscore* (shortened version of Predictive Optimization of Latent skill Level in Surveys, Considering Overall Record, Empirically) is a numerical variable which captures a poll's error and bias, whereas *transparency score* is a continuous variable which captures how transparent a poll is based on the amount of information it discloses about its polls. *Pollscore* ranges from -1.5 to 1.7, with negative values indicating less error and bias, and *transparency score* ranges from 1 to 10, with a 1 being the least transparent and a 10 being the most transparent.

Univariate and Multivariate Exploratory Data Analysis

We then conducted univariate and multivariate exploratory data analysis on our variables of interest. The distribution of the response variable *poll accuracy* (**Appendix B**) is unimodal and nearly symmetric, with a mean poll accuracy of -3.24 and a standard deviation of 4.21 (**Appendix C**) , reflecting a relatively large spread around the center. Along with the mean, the median of *poll accuracy* is also negative (-2.91), which implies that the polls in our dataset underestimated the share of the popular vote President-Elect Donald Trump would receive in the election. The distribution of *sample size* (**Appendix D**) demonstrates a heavy right skew with three noticeably large outliers – in order to eliminate the right skew, we applied a Log_{10} transformation to the variable prior to modeling.

The scatterplot matrix of our continuous variables reveals weak, negative linear relationships between *poll accuracy* and the three continuous predictor variables (**Appendix F**). *Pollscore* appears to have the strongest correlation with *poll accuracy* (r = -0.099), indicating that polls with decreased *pollscore* values are associated with greater poll accuracies. The scatterplots also suggest a concern regarding multicollinearity between *pollscore* and *transparency score*, which is consistent with the fact that the two variables are both constructed by 538.

The side-by-side boxplots observing the relationships between *poll accuracy* and the five categorical variables contain a common theme across all categories for all variables: the median *poll accuracy* is negative, indicating that there is no category in our dataset for which over half of the polls did not underestimate President-Elect Trump's share of the vote (**Appendix G**). Moreover, there are particularly discernible differences across the categories of *population*, *swing state*, and *time difference*. Polls conducted among likely voters, polls conducted in swing states, and polls conducted closer to the election ($\leq$ 1 month) all outperformed their counterparts in median *poll accuracy*. Polls conducted in the Southern region of the United States, along with those employing a "mixed" methodology technique, contained the median *poll accuracy* values closest to 0, which would indicate the most accurate prediction to two decimal places.

To explore the potential interaction between *sample size* and *poll methodology*, we plotted *poll accuracy* versus *sample size* across different polling methods. The plots do not suggest the viability of an interaction term due to the relatively parallel lines and overlap of points, indicating a lack of explicit subdivisions across the *methodology* categories (**Appendix H**). As mentioned in the introduction, the interaction was formally tested to determine whether or not it should be included in the final model.

III. Model and Results

a. Analytic Methods

We investigated the use of a multivariate normal error linear regression model to predict *poll accuracy*, our response variable, with all other variables (excluding *partisanship* due to many missing values) as predictors of a poll's accuracy. We also included the interaction term between *sample size* and *polling methodology* due to its relevance to the initial hypotheses. We then determined how to model the several categorical variables in our model. The "state" variable was recoded into a variable describing which *region* the poll was taken in, with states split up into "Northeast," "Midwest," "South," and "West" as given by the U.S. Census' regional distinctions (U.S. Census Bureau, 1984). Polls that took a nationwide sample produced the fifth "national" category – the creation of the *region* variable was decided upon to develop a more stable model, as opposed to having 51 total categories for all of the states. We similarly combined the vast number of categories for polling *methodology* into four larger categories of "Online," "Phone," "Mixed," and "Other" due to imbalanced observations across the various categories, and utilized the same reasoning to combine the *population* categories into "Likely Voters" and "Other." Moreover, we developed our own *swing state* binary indicator variable based on the 7 states that were widely deemed pivotal to the presidential election across the media (North Carolina, Pennsylvania, Michigan, Wisconsin, Nevada, Arizona, and Georgia). In a similar vein, we added *time difference* as a categorical variable with categories of "≤1 month", "1-2 months", "2-3 months", and ">3 months" representing the difference in time between the poll's end date and the general election date of November 5, 2024.

The initial model's diagnostic plots demonstrated a violation of both the constant variance and normally-distributed error assumptions for a multivariate normal error linear regression model, so a Box-Cox transformation technique was implemented (**Appendix I**). The suggested transformation of $Y = Y^{1.5}$ was unsuccessful in improving the diagnostics, and following a series of other unsuccessful power transformations, we looked into converting our response variable from an index to a proportion, so as to apply the commonly-utilized logit or log-odds transformation of $log(p/1 - p)$. The logit-transformed poll accuracy demonstrated an improvement in model assumptions – most notably, although there remains significant deviation from the line present in the QQ-plot, the distribution of residuals appears normal in nature, albeit with thicker and larger tails (**Appendix J**). As a result, we proceeded to statistical inference with the logit transformation, remaining cautious regarding the validity of the inferential analysis.

Prior to inference, we addressed the earlier multicollinearity concern regarding *transparency score* and *pollscore* through an analysis of their generalized variance inflation factors. Neither of the variables had GVIF values high enough to warrant consideration of dropping from the model (**Appendix K**). Moreover, we performed outlier detection in the form of a Cook's distance plot – the plot raised concerns regarding roughly one hundred of the observations (**Appendix L**), but given that 538 only adds polls to their dataset which pass a stringent set of criteria, we were certain that there were no data collection errors, and thus decided to keep all observations (FiveThirtyEight, 2018).

b. Final Model and Results

The final model included all previously-mentioned predictor variables, but the interaction term was excluded due to a partial F-test indicating that it was insignificant at $\alpha = 0.05$, corroborated through the interaction term plot. All relevant coefficients, estimates, p-values, and 95% confidence intervals are presented in the table in **Appendix M**. The overall regression F-test conducted with 17 and 1915 degrees of freedom yields an F-statistic of 23.46, with a corresponding p-value of $< 2 * 10^{-16}$, meaning that our final model is more effective at predicting *poll accuracy* than an intercept-only model, and the regression is significant. The final model explains 14.86% of the variance in *poll accuracy* when adjusting for the number of predictors, as indicated by the R_{adj}^2 value of 0.1486. Our model outperformed the suggested model from best subset selection (no two-way interactions considered) and both-direction stepwise regression selection resulted in our suggested model.

The model contains many significant adjusted effects – in particular, in response to initial hypotheses, it appears as though polls taken in swing states ($p = 5.06 * 10^{-11}$) and polls taken closest to the election (≤1 month, several p-values all $< 3.12 * 10^{-9}$) are more accurate when holding other variables constant, as was hypothesized. Through a two-sided t-test for slope, we conclude that *sample size* is an insignificant predictor of *poll accuracy* ($p = 0.0747$) when holding other variables constant. However, a one-sided t-test with $H_A\text{: } \beta_{Log(SampleSize)} > 0$ results in a p-value of 0.0374, meaning that at a significance level of $\alpha = 0.05$, we verify the initial hypothesis that polls with larger sample sizes are associated with greater accuracy when adjusting for other variables. We also find that nationally-conducted polls were outperformed by polls conducted in the South ($p = 0.0607$), but drastically underperformed by polls conducted in the Northeast ($p = 8.89 * 10^{-8}$), which is in line with President-Elect Trump's victory and his robust support in the South.

We also consider a prediction estimate for a poll conducted online among a sample of 1000 likely voters in Pennsylvania, with less than a month until the election date and a mean pollscore and transparency score of -0.48 and 6.42 respectively. We estimate that on average, a poll with these characteristics is associated with a *poll accuracy* of -0.137 log units. On the other hand, we estimate that a poll taken in New York with all other characteristics remaining the same is associated with a *poll accuracy* of -0.204 log units on average. The stark difference in accuracy between these two predictions demonstrates the significance of polls being taken in swing states; we estimate that when adjusting for all other variables, polls taken in swing states are 0.0671 log units more accurate on average than those not taken in swing states. In a similar vein, a poll taken among a population of voters 2-3 months before the election date (with all other characteristics remaining the same as the previous hypothetical poll) has an estimated *poll accuracy* of -0.286 log units on average. This prediction emphasizes the effect of *time difference*; we estimate that when holding all other variables constant, polls conducted 2-3 months before the election are 0.0819 log units less accurate than their counterparts taken less than 1 month before the election on average. This effect is the largest across all variables, thereby indicating that proximity to the date of the election is a significant predictor of a poll's accuracy.

IV. Discussion

We conclude through a multivariate normal error linear regression model that polls once more underestimated the share of the vote that President-Elect Donald Trump would receive in the election. We also conclude that a poll's location (particularly whether it is in a swing state or not), time to election, pollscore, transparency score, and population surveyed are all significant predictors of a poll's accuracy when holding other variables constant. With a regression F-test p-value of $< 2 * 10^{-16}$ and an R_{adj}^2 value of 0.1486, we conclude that the regression is significant and explains 14.86% of the variability in logit-transformed *poll accuracy* in our dataset.

Future research could address the limitations of our work, such as our concerns with assumptions, which could be alleviated with a more robust model such as the LAD estimator or Huber Loss function. Other regression model selection techniques, such as LASSO, may also help reinforce our decision-making process or allow us to consider a different model that could help explain more of the variability in poll accuracy. Furthermore, future analysis could include the statistical sampling method of the polls (i.e. stratified sampling, cluster sampling, etc.), as well as an analysis of the polls' demographic data to determine how representative the samples are of the United States' population at large (i.e. underrepresentation of women/men, African-Americans/Hispanics, etc.). It must also be remembered that as a feature of being observational studies, all polls suffer from forms of bias that can be aggregated as *total survey error* (Shirani-Mehr et al., 2018). By tackling these challenges, future research can contribute to more reliable and nuanced insights into the factors affecting poll performance and accuracy.

References

FiveThirtyEight. (2018, June 28). *National : President: general election : 2024 Polls*. FiveThirtyEight. https://projects.fivethirtyeight.com/polls/president-general/2024/national/

Gayman, D. (2024). *Election polling methods constantly changing, improving | Nebraska Today*. University of Nebraska-Lincoln. https://news.unl.edu/article/election-polling-methods-constantly-changing-improving

Parshall, A. (2024, October 31). *Why Election Polling Has Become Less Reliable*. Scientific American. https://www.scientificamerican.com/article/why-election-polling-has-become-less-reliable/

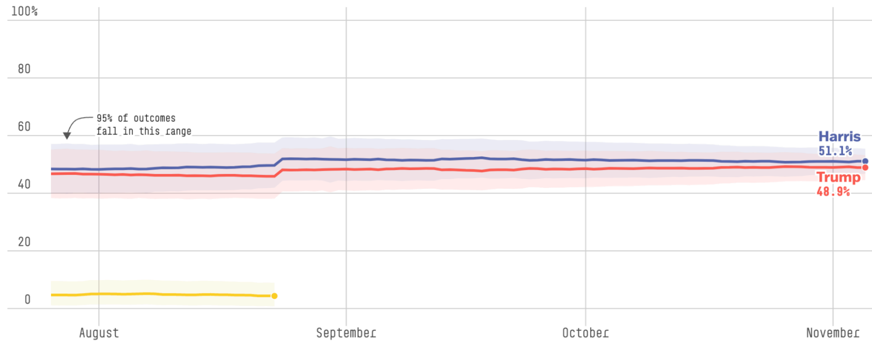
Paz, C. (2024, September 10). *Can we trust the polls this year?* Vox; Vox. https://www.vox.com/2024-elections/370649/trust-polls-2016-2020-election-2024-pollster-polling-miss

Shirani-Mehr, H., Rothschild, D., Goel, S., & Gelman, A. (2018). Disentangling Bias and Variance in Election Polls. *Journal of the American Statistical Association*, *113*(522), 607–614. https://doi.org/10.1080/01621459.2018.1448823

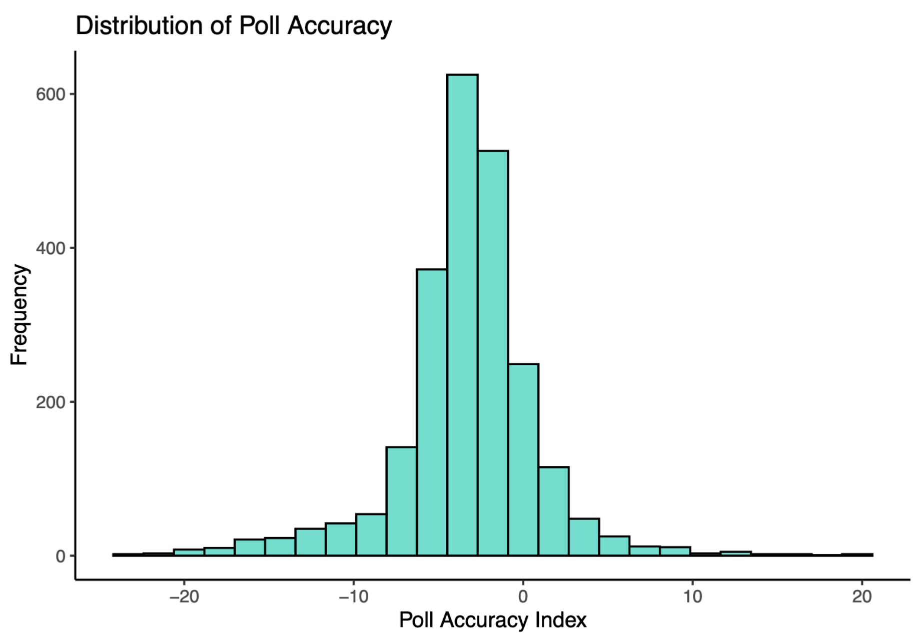
U.S. Census Bureau. (1984). *Census Regions and Divisions of the United States* . https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

Appendix

All code for the purpose of this project was written, and executed, in RStudio (version 4.2.2).



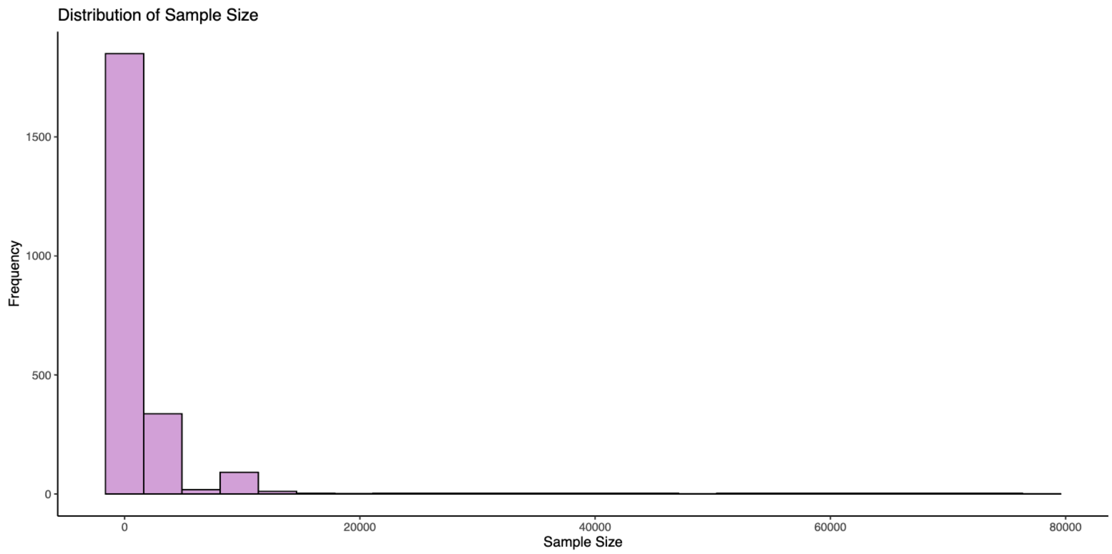
Appendix A: A graphical representation of 538's weighted average of the predicted popular vote breakdown from August 2024 to November 2024. There was no point in the buildup to the election at which 538 projected that President-Elect Donald Trump would win the popular vote, further highlighting polls' inaccuracies.



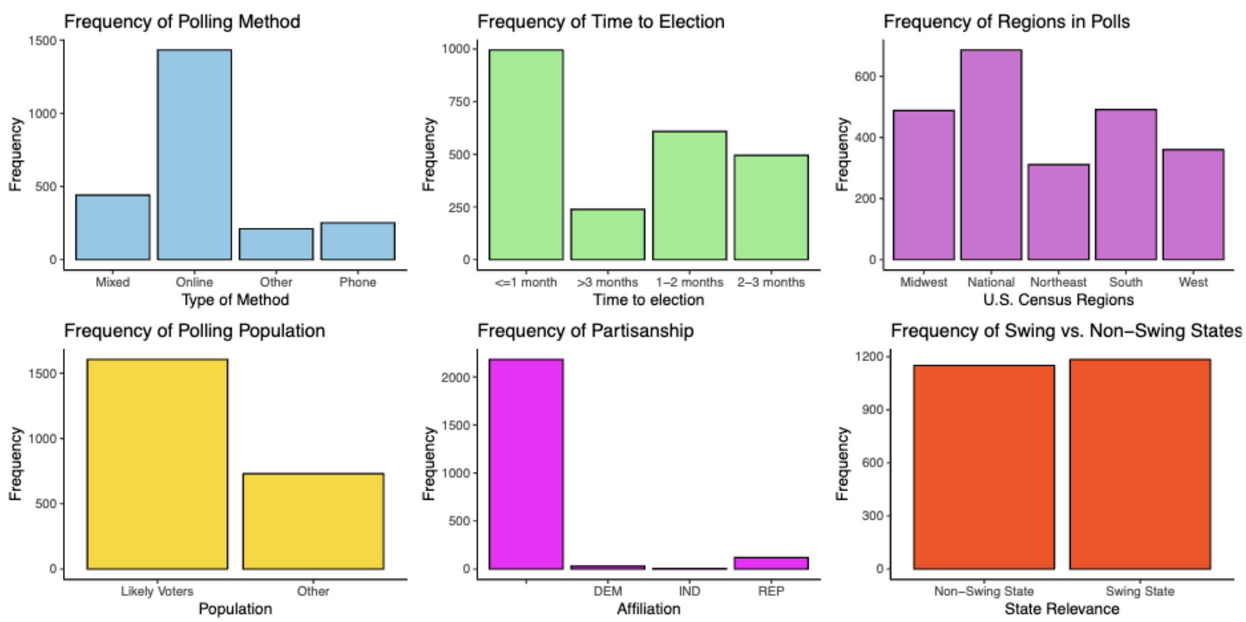
Appendix B: Histogram displaying the distribution of initial response variable, poll accuracy.

	Min	Mean	SD	Median	IQR	Max
Poll Accuracy	-22.91	-3.24	4.21	-2.91	4.00	20.09
Sample Size	301.00	1637.81	2829.82	976.00	678.00	78247.00
Pollscore	-1.50	-0.48	0.65	-0.40	0.90	1.70
Transparency Score	0.00	6.42	2.68	7.00	5.00	10.00

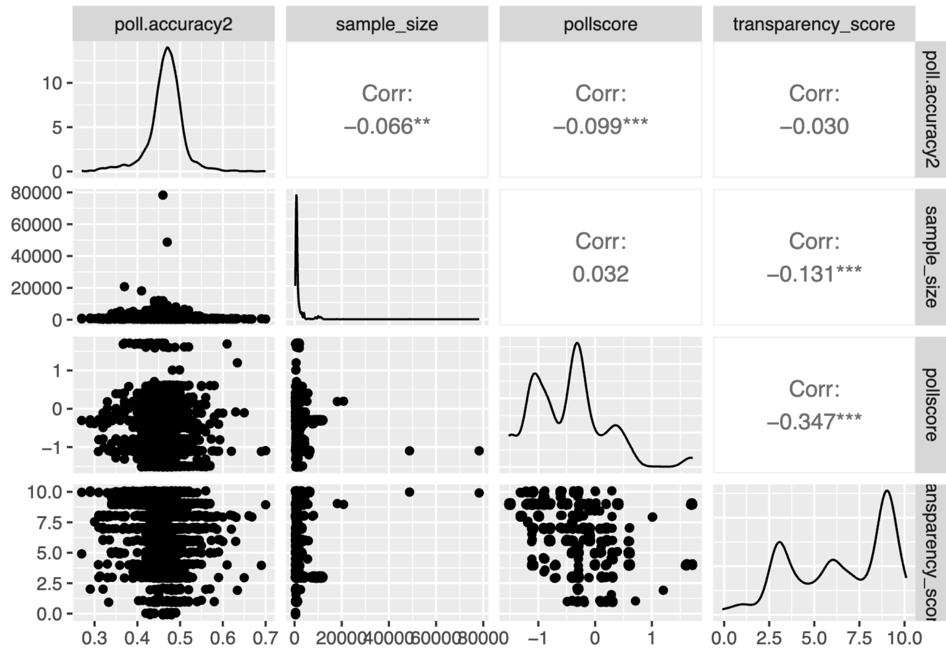
Appendix C: Numerical summary of the continuous variables of interest. Poll accuracy has a negative mean and median, indicating an underestimation of President Trump's performance.



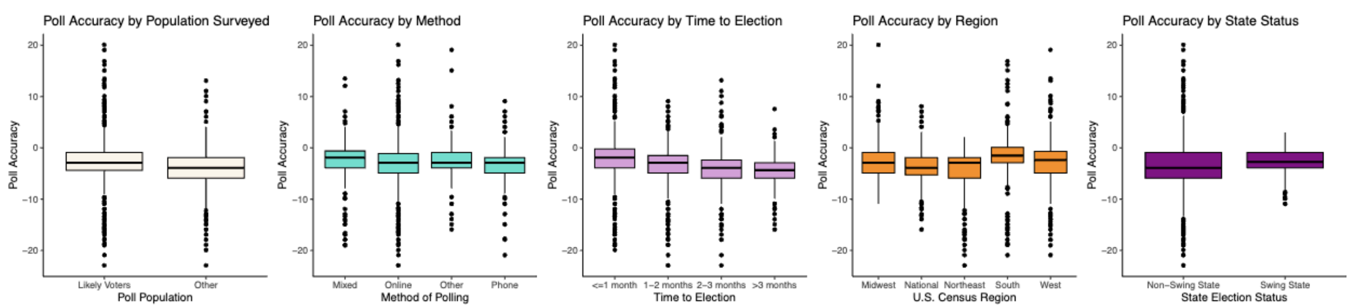
Appendix D: Histogram displaying the distribution of sample size, a predictor variable. There is a severe right skew with several large outliers, resulting in a $\log_{10}$ transformation.



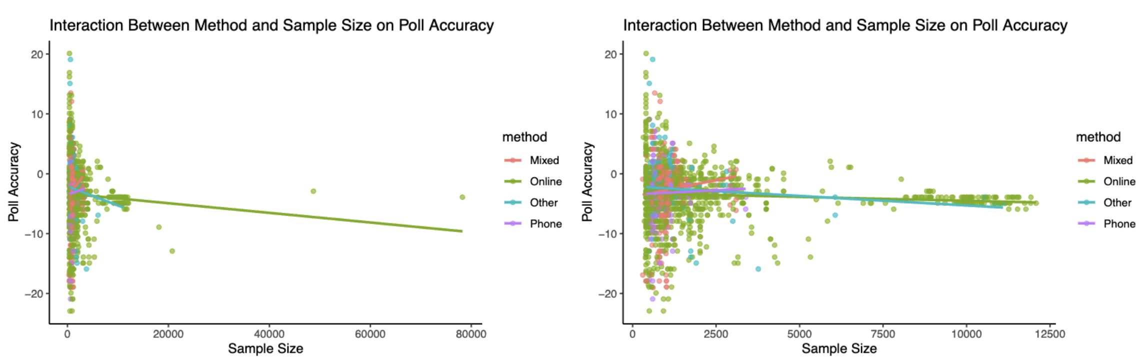
Appendix E: Bar plots depicting the distribution of each of the categorical variables of interest. Partisanship contains a vast majority of missing values, so it was dropped from the model.



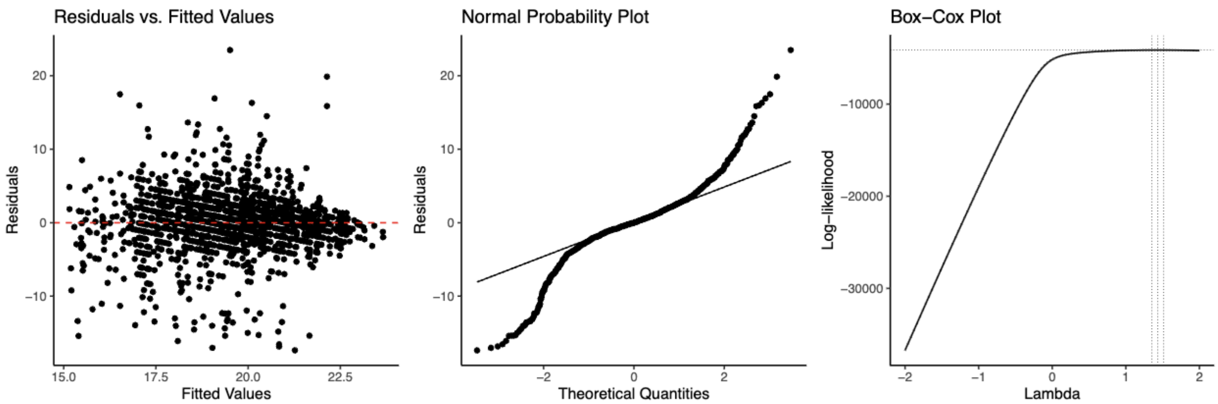
Appendix F: Scatterplot matrix of the four relevant continuous variables, with jitter applied to improve interpretability. Poll score appears to contain the strongest correlation to poll accuracy (r = -0.099). Transparency score and poll score demonstrate multicollinearity concerns (r = -0.347).



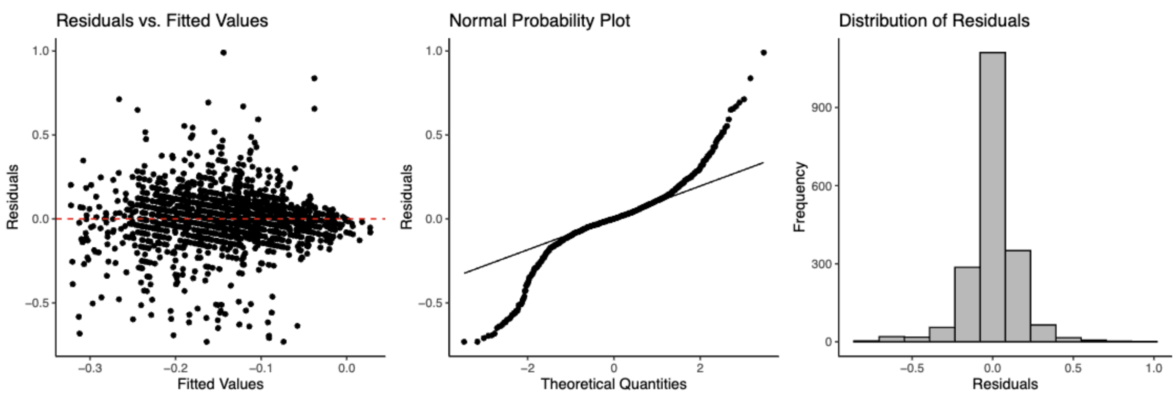
Appendix G: Box plots depicting relationship between poll accuracy and categorical variables of interest. Polls in swing states, the Southern region, and closer to the election date show greater accuracy, but the median across all categories for all categorical variables is negative.



Appendix H: Interaction term plots with large sample size outliers removed on the right for greater interpretability. No interaction appears present between sample size and methodology.



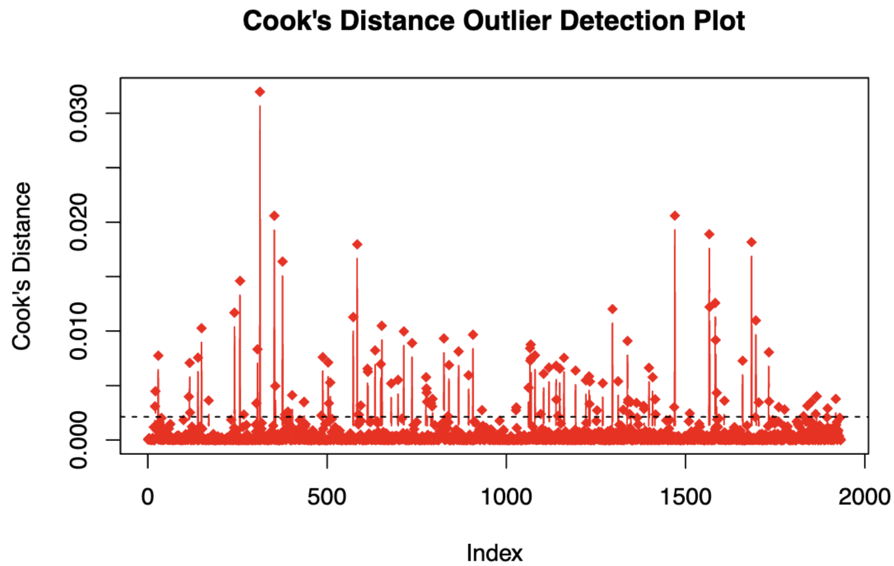
Appendix I: Initial model diagnostics, violating constant variance and normally distributed errors assumptions. Box-Cox plot on the right suggests power transformation of $Y^{1.5}$.



Appendix J: Transformed model diagnostic plots show slight improvement. The histogram of the residuals is fairly normal, but contains much thicker tails.

	Variable	GVIF	Degrees of Freedom	Adjusted GVIF
1	Population	1.116	1	1.056
2	Method	1.340	3	1.050
3	Region	2.809	4	1.138
4	Time Difference	1.126	3	1.020
5	Swing State	1.892	1	1.375
6	Pollscore	1.327	1	1.152
7	Transparency Score	1.225	1	1.107
8	Log10(Sample Size)	1.788	1	1.337

Appendix K: Generalized Variance Inflation Factor (GVIF) of the variables in the final model demonstrate no strong concern for multicollinearity.



Appendix L: Cook's Distance plot indicates many observations above the threshold for influence.

	Estimates	Standard Error	Test Statistic	P-Value	95% Confidence Intervals
(Intercept)	-0.207	0.0514	-4.036	5.66e-05	(-0.308, -0.107)
I(Likely Voters)	0.0330	0.00799	4.124	3.88e-05	(0.0173, 0.0486)
I(Mixed)	0.0207	0.00942	2.202	0.0278	(0.00227, 0.0392)
I(Other)	0.0577	0.0160	3.599	0.000327	(0.0263, 0.0892)
I(Phone)	0.000778	0.0118	0.066	0.947	(-0.0223, 0.0239)
I(Midwest)	-0.0100	0.0144	-0.693	0.488	(-0.0384, 0.0183)
I(Northeast)	-0.0803	0.0150	-5.369	8.89e-08	(-0.110, -0.0510)
I(South)	0.0251	0.0134	1.876	0.0607	(-0.00113, 0.0513)
I(West)	0.00395	0.0152	0.260	0.795	(-0.0258, 0.0337)
I(>3 Months)	-0.0742	0.0125	-5.953	3.12e-09	(-0.0987, -0.0498)
I(1-2 Months)	-0.0593	0.00879	-6.740	2.08e-11	(-0.0765, -0.0420)
I(2-3 Months)	-0.0819	0.00970	-8.441	< 2e-16	(-0.101, -0.0629)
I(Swing State)	0.0671	0.00964	6.959	4.68e-12	(0.0482, 0.0860)
Pollscore	-0.0261	0.00623	-4.185	2.99e-05	(-0.0383, -0.0139)
Transparency Score	-0.00603	0.00146	-4.128	3.81e-05	(-0.00889, -0.00316)
Log10(Sample Size)	0.0256	0.0143	1.783	0.0747	(-0.00255, 0.0537)

Appendix M: A table demonstrating the estimated effects, standard errors, t-test statistic values, p-values, and 95% confidence intervals for all of the variables in the final model. Although not mentioned in the paper, we note that 538's pollscore and transparency score both demonstrate significant adjusted effects which are contradictory in nature: an increase in pollscore (which represents more error/bias) is associated with a decrease in transformed poll accuracy, whereas an increase in transparency score (which represents greater poll transparency) is associated with a decrease in transformed poll accuracy when adjusting for other variables. These results suggest that increased poll transparency is not a predictor of greater poll accuracy, whereas decreased margin of error and bias are. The population which a poll surveys is also a significant adjusted effect; polls which survey "likely voters" are associated with 0.0330 log units greater poll accuracy than polls which do not when holding all other variables constant.