Predicting the 2024 US Presidential Election: A Polling-Based Forecast*

My subtitle if needed

Tianrui Fu Yiyue Deng

October 30, 2024

This paper forecasts the 2024 US Presidential election outcome using polling data from [insert pollster name]. By applying simple and multiple linear regression models, we analyze the effect of polling factors, including sample size, poll score, and transparency, on support percentages for key candidates. Our findings suggest significant relationships between these variables, providing an evidence-based approach to predicting election outcomes. We further explore methodological strengths and weaknesses and propose an ideal polling survey methodology. This work highlights the potential for data-driven insights into political forecasting.

Table of contents

[*] Code and data are available at: https://github.com/RohanAlexander/starter_folder.

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section A....

2 Data

2.1 Overview

We use the statistical programming language R Core Team (2023) to complete this analysis. Our data about the latest polling outcomes is from the website[@]. It has 52 variables and 15891 observations in this data set, including pollster, poll score, etc.

Table 1: The example of chosen variable datset

pollster	pollscore	numeric_grade	transparency_score	end_date	pct
Marist	-0.9	2.9	7	2024-08-04	48.0
Emerson	-1.1	2.9	7	2024-07-23	47.5

Beacon/Shaw	-1.1	2.8	9	2024-09-24	46.0
AtlasIntel	-0.8	2.7	6	2024-10-17	49.9
Quinnipiac	-0.5	2.8	9	2024-09-22	48.0
Marquette Law School	-1.1	3.0	10	2024-08-01	47.0
YouGov	-1.1	3.0	9	2024-08-31	49.0
Emerson	-1.1	2.9	7	2024-09-28	49.9
Marist	-0.9	2.9	7	2024-09-05	48.0
MassINC Polling Group	-0.8	2.8	7	2024-09-18	40.0

2.2 Cleaning Data

We have cleaned the data set and the detailed procedure see from the appendix.

2.3 Measurement

After cleaning the data set, we have get about 38 variables and 492 observations about the polling outcomes for Donald Trump. In order to analysis if the different pollster can influence the final result of U.S. president election, we choose some related variable to build the Bayesian model. The related variables has the numeric grade, transparency score, poll score, pollster, the end date and percentage. The explain of each used variable please see in the (**appendix-variable?**).

2.4 Outcome variables

The Figure 1 shows the distribution of polling data over time, with each color representing a different polling organization. We can see that as time progresses from August to October, there is a steady increase in the amount of polling data, peaking around early October. This suggests that multiple organizations have contributed to polling data on a consistent basis over this period.

In the Figure 2, it reveals a central peak around the 45-50% range, indicating that most of the values for this variable are concentrated in this area. The shape of the plot suggests a right-skewed distribution, with relatively few data points falling below 40% or above 50%. This gives an overview of the central tendency and variability of the polling percentages.

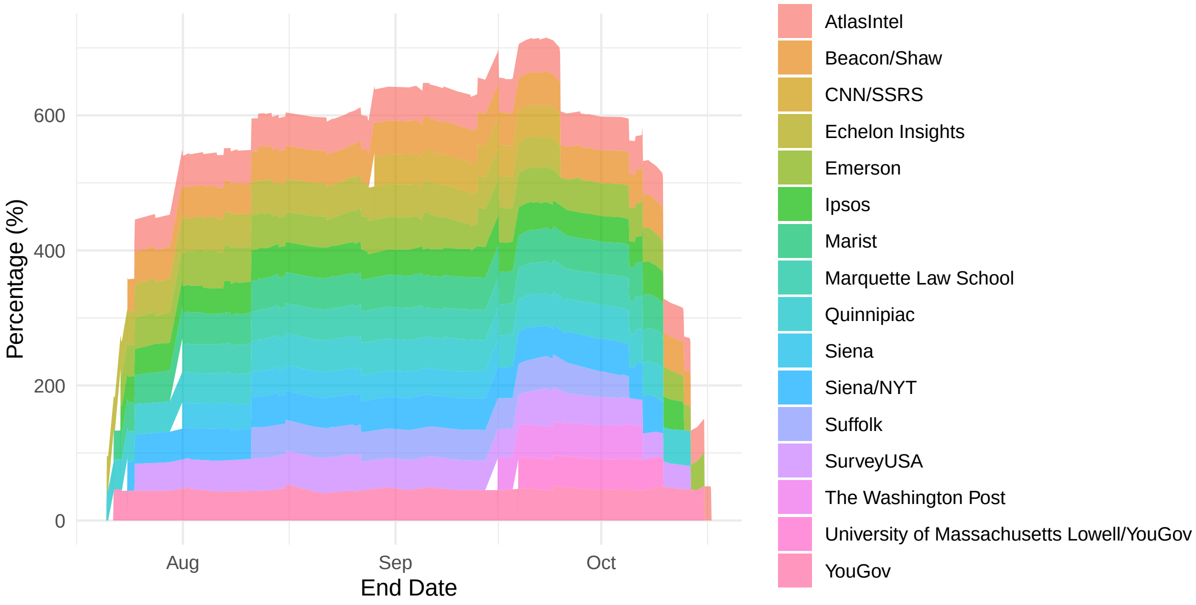

Figure 1: Percentage of Each Pollster Over Time

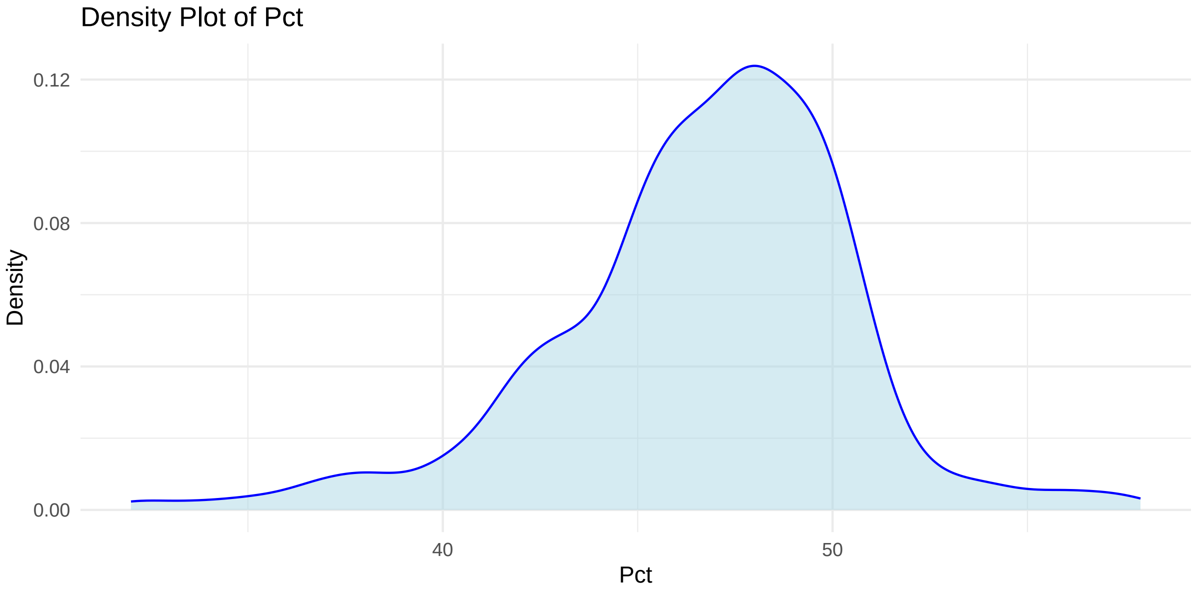

Figure 2: Density plot of pct

2.5 Predictor variables

In Figure 3, the predictor variable is the count of polls conducted by each pollster. The bar chart displays the number of polls released by different polling organizations, ordered from highest to lowest frequency. The bars range in color from light blue to dark blue, indicating the frequency of polls conducted by each organization. Siena/NYT and YouGov are the most active pollsters, with 94 and 59 polls conducted, respectively. Emerson and Beacon/Shaw also have high polling frequencies, with 58 and 46 polls. In contrast, several pollsters, like Christopher Newport University and Data Orbital, conducted only one poll, indicating their minimal activity in comparison.

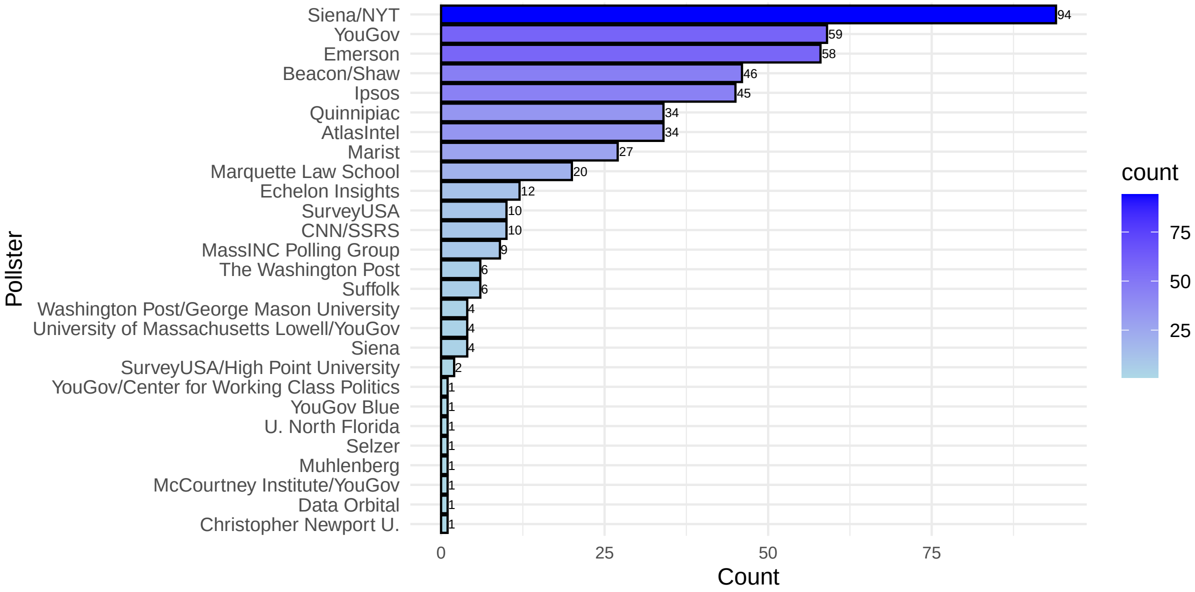

Figure 3: Frequency of each pollster

In Figure 4, the predictor variable is the average poll score for each pollster over time, from August to October. This heat map visualizes the poll scores of each organization on different dates, with color intensity indicating the score levels—darker colors represent lower scores. The pollsters with the most consistently low average poll scores (darker colors) include Selzer, Siena/NYT, and Marquette Law School. These pollsters frequently show results lower than others over time, suggesting they may consistently lean toward one direction in their results. In contrast, pollsters like YouGov Center for Working Class Politics and YouGov Blue display lighter colors, indicating relatively higher scores across their polls.

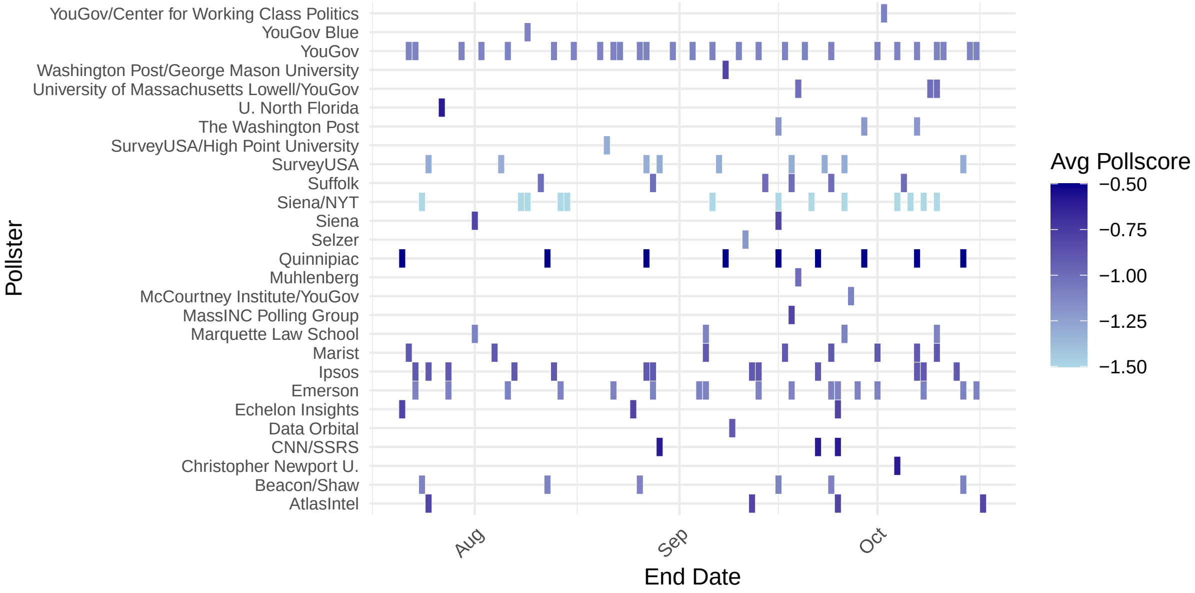

Figure 4: Average Pollscore by Pollster and End Date

In Figure 5, there is a clear positive relationship between transparency score and numeric grade. Pollsters with higher transparency scores, such as those scoring around 10, tend to achieve higher numeric grades, close to 3.0. Conversely, pollsters with lower transparency scores tend to have lower numeric grades, closer to 2.7. This pattern suggests that greater transparency is associated with better overall pollster ratings.

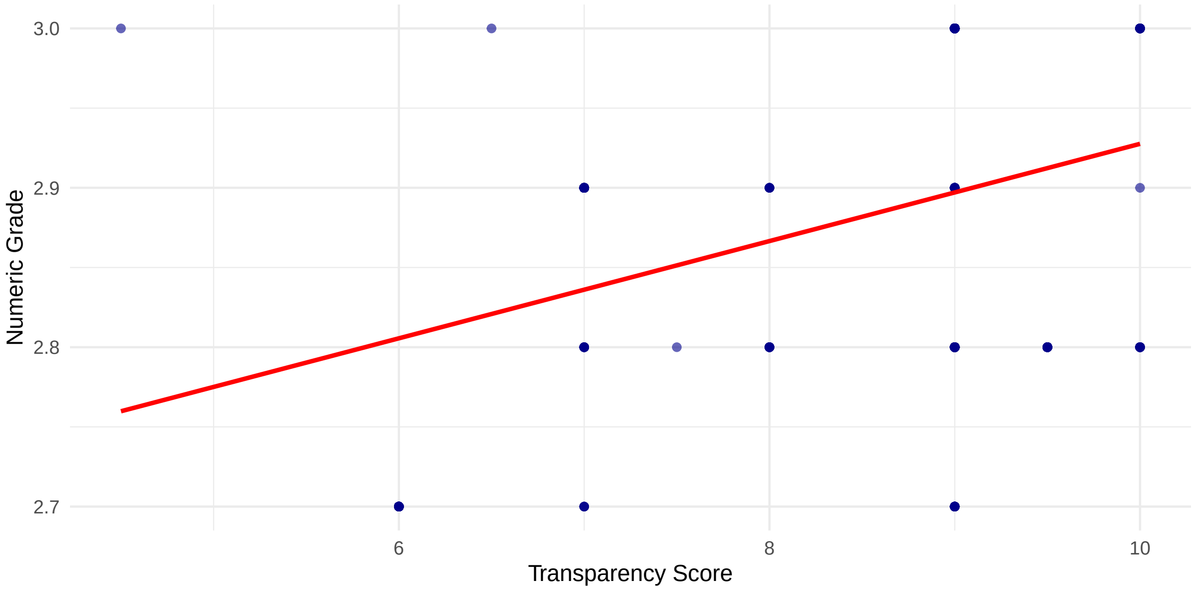

Figure 5: Scatter Plot of transparency score and numeric grade

3 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix C.

3.1 Model set-up

Define y_i as the percentage. Then β_i represents the numeric grade, γ_i represents the transparency score, and δ_i represents the pollscore.

$$
\begin{aligned}
y_i|\mu_i,\sigma &\sim \text{Normal}(\mu_i,\sigma) && (1)\\
\mu_i &= \alpha + \beta_1 \cdot \text{numeric\_grade}_i + \beta_2 \cdot \text{transparency\_score}_i + \beta_3 \cdot \text{pollscore}_i && (2)\\
&\quad + \sum_{j=1}^{N} \gamma_j \cdot \text{pollster}_j + \delta_i \cdot \text{end\_date}_i && (3)\\
\alpha &\sim \text{Normal}(50,10) && (4)\\
\beta_1,\beta_2,\beta_3 &\sim \text{Normal}(0,5) && (5)\\
\gamma_j &\sim \text{Normal}(0,5) && (6)\\
\sigma &\sim \text{Exponential}(1) && (7)
\end{aligned}
$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in Table 2.

Table 2: Explanatory models of flight time based on wing width and wing length

	Model Summary	
	Model by glm	Model by bayes
(Intercept)	−524.927*** [−773.481, −276.373]	−590.212 [−814.191, −354.733]
numeric_grade	−14.386 [−43.487, 14.716]	−3.356 [−11.443, 5.059]
transparency_score	0.903 [−0.348, 2.153]	−0.164 [−0.881, 0.569]
pollscore	−3.028 [−32.532, 26.476]	1.121 [−5.151, 7.339]
Beacon/Shaw	−2.833 [−12.913, 7.246]	0.976 [−1.509, 3.389]
Christopher Newport U.	−9.186* [−18.242, −0.130]	−5.294 [−10.778, 0.269]
CNN/SSRS	−4.208 [−10.838, 2.423]	−1.343 [−4.986, 2.341]
Data Orbital	−3.602 [−12.297, 5.093]	−0.426 [−6.114, 5.093]
Echelon Insights	−3.313 [−7.738, 1.113]	0.389 [−2.466, 3.257]
Emerson	1.010 [−6.052, 8.072]	1.561 [−0.559, 3.788]
Ipsos	−7.483** [−13.082, −1.883]	−3.927 [−6.355, −1.584]
Marist	0.162 [−4.587, 4.911]	1.111 [−0.987, 3.216]
Marquette Law School	−3.213 [−12.696, 6.270]	−0.561 [−3.603, 2.499]
MassINC Polling Group	−7.280*** [−10.601, −3.958]	−6.566 [−8.857, −4.172]
McCourtney Institute/YouGov	−1.354 [−10.616, 7.908]	−1.745 [−7.385, 3.817]
Muhlenberg	−2.040 [−10.915, 6.834]	0.196 [−5.405, 5.867]
Quinnipiac	−1.513 [−10.660, 7.633]	−0.189 [−3.441, 3.117]
Selzer	−2.952 [−15.951, 10.046]	−0.220 [−5.949, 5.492]
Siena	−9.684*** [−13.295, −6.072]	−7.344 [−10.509, −4.019]
Siena/NYT	−2.174 [−20.461, 16.113]	0.976 [−3.002, 5.133]
Suffolk	−4.753 [−10.985, 1.479]	−3.290 [−6.325, −0.369]
SurveyUSA	−5.810 [−20.672, 9.053]	−2.163 [−5.651, 1.268]
SurveyUSA/High Point University	−5.876 [−23.153, 11.401]	−0.224 [−5.003, 4.990]
The Washington Post	0.576 [−10.255, 11.407]	2.315 [−1.070, 5.912]
University of Massachusetts Lowell/YouGov	−6.065 [−13.331, 1.201]	−3.502 [−7.054, 0.165]
Washington Post/George Mason University	−8.943*** [−14.000, −3.886]	−4.783 [−8.412, −1.163]
YouGov	−2.338 [−10.874, 6.199]	−0.764 [−3.385, 1.935]
YouGov Blue	−1.128 [−11.743, 9.488]	0.335 [−5.326, 5.727]
End Date	0.030*** [0.019, 0.042]	0.032 [0.021, 0.044]
U. North Florida		1.990 [−3.834, 7.845]
YouGov/Center for Working Class Politics		−2.360 [−8.276, 3.399]
Num.Obs.	492	492
R2	0.332	0.318
R2 Adj.		0.258
AIC	2575.9	
BIC	2701.8	
Log.Lik.	−1257.932	−1264.371
ELPD		−1289.0
ELPD s.e.		25.3
LOOIC		2577.9
LOOIC s.e.		50.6
WAIC		2575.6
RMSE	3.12	3.38

+ p \num{< 0.1}, * p \num{< 0.05}, ** p \num{< 0.01}, *** p \num{< 0.001}

[a] This table shows the regression models with custom variable names.

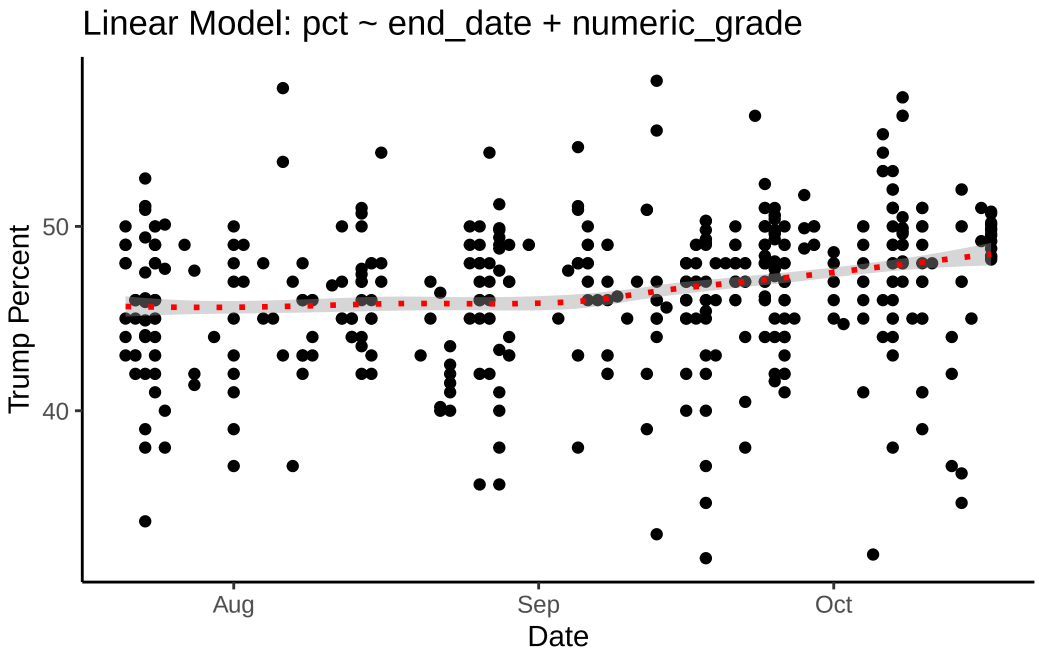

Figure 6: ddd

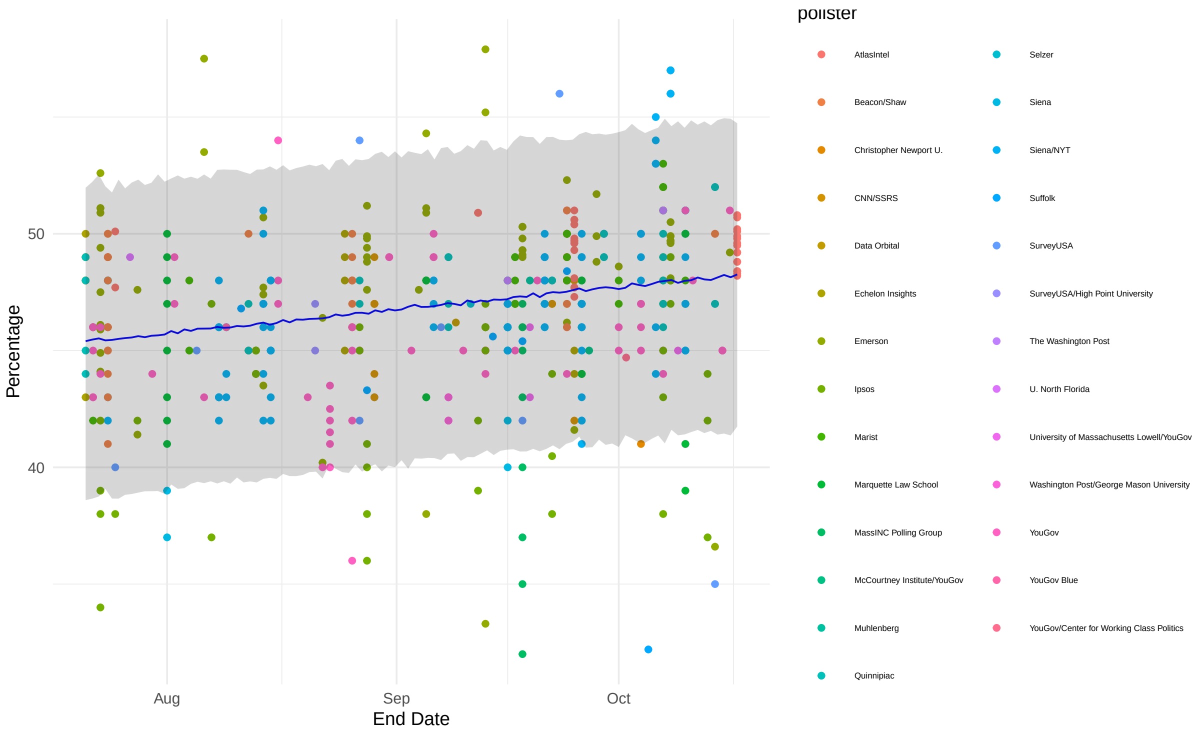

Figure 7: Poll Percentage over Time with Bayesian Fit

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

A Appendix

A.1 Clean data

It starts by reading the dataset and standardizing column names. Next, it removes irrelevant columns like sponsor_ids, sponsors, and etc and filters for high-quality polls (with a rating of 2.7 or higher) that specifically track Trump's support. National polls missing state information are labeled as "National," and dates are formatted and filtered to include only those on or after July 21, 2024 (when Trump declared his candidacy). Finally, the percentage supporting Trump is converted to an actual count for further modeling. ## Variable details We have used the variable numeric_grade, transparency_score, pollster, poll score, end_date and pct in the building of Bayesian model. Among these variables in the data set after analysis, the numeric_grade is the numeric rating given to the pollster to indicate their quality or reliability from 2.7 to 3.0. The transparency_score is the grade for how transparent a pollster is, calculated based on how much information it discloses about its polls and weighted by recency from 4.5 to 10.0. The pollster is the name of the polling organization that conducted the poll included Marquette Law School, CNN/SSRS and etc. The poll score is the numeric value representing the score or reliability of the pollster in question from -1.5 to -0.5 and negative numbers of poll score are better. The end_date is the date of polling ends. The pct is the percentage of the vote or support that the candidate received in the poll and keep integer.

B Additional data details

C Model details

C.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...

Examining how the model fits, and is affected
by, the data

C.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows... This suggests...

?@fig-stanareyouokay-2 is a Rhat plot. It shows... This suggests...

Checking the convergence of the MCMC algorithm

References

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

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.