Forecasting US Presidency Election Results for 2024[*]

Ruizi Liu Yuechen Zhang Bruce Zhang

October 22, 2024

This paper forecasts the 2024 U.S. Presidential election with a focus on Donald Trump's support using polling data from FiveThirtyEight. We applied linear and Bayesian models to analyze data from pollsters like YouGov and Siena/NYT. Results show Trump's support remains stable between 40% and 50% across national and state polls from August to October 2024. Higher-quality pollsters provided more consistent predictions. While the models offer insights, limitations include the assumption of linearity and exclusion of other candidates. Overall, Trump's support is increasing, but further monitoring is needed for precise forecasts.

1 Introduction

Predicting the outcome of the United States presidential election is an important and complex task, and polling data is a key indicator of voter sentiment and election trends. Aggregating data from multiple polls allows for more reliable analyses of voter preferences. However, the accuracy of these predictions is affected by several factors, including the methods used by different pollsters, how they handle sampling, and the timing of the survey. Despite the wide availability of polling data, making reliable predictions remains a challenging responsibility due to these uncertain variations.

This paper aims to predict the winner of the upcoming 2024 US presidential election through building linear models by using data from FiveThirtyEight's president election polls which including data collected form different pollsters such as YouGov and RMG Researchd. By focusing on trends in polling data, this analysis seeks to identify patterns in voter support and how they vary across key factors such as geography, sampling methodology, and non-response bias. Additionally, we will delve into the methods of a prominent pollster to assess how their polling methodology affects the accuracy of a candidate's projected support.

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

The results show that candidate support has remained stable and increased in states over time. Specific polling trends from our selected pollsters show similar results.

2 Data

2.1 Overview

In this paper, we use the statistical programming R (R Core Team (2023)) to analyze poll results of the 2024 US presidential election results. The data for these results are obtained from FiveThirtyEight ((**citefivethirtyeight?**)), which is a platform operated by ABC news that consolidates poll results from different pollsters. The data presents variables such as percentage support per candidate, source state/region of data, method of data collection, and many other key pieces of information that can be vital in forecasting the presidential election results.

2.2 Measurement

The method of data collection or the key measurement methods of the data is by pollsters including YouGov, ActiVote, Pew, Research Co., and many more. These pollsters distributed polls through an array of methodologies, including but not limited to online panels, live phone calls, text, and email. The target was eligible voters in the United States. Those that were surveyed were asked about their party of preference and candidate of choice. The results showed a candidate supported of the poll and a percentage of support found. Consolidated, the results form a representative picture of the general support pattern for each candidate across the country, which can be validly used to forecast the presidential election results.

2.3 Outcome variables

The dependent variables or the variables of interest in this case are the percentage support and the candidate supported. This includes variables classified in the data set "pct" and "answer". Other variables such as "candidate_name" and "candidate_id" are also informative of the respondent's choice but are redundant for the analysis. The independent variables include many factors. This is because many different aspects can come into play when determining the presidential choice. Here, we are trying to look at what variables have an impact on the dependent variables. The independent variables are as follows. The "pollster" variable defines the specific company or organization that implemented the poll. Along with the "numeric_grade", which defines the quality of a pollster based on historical accuracy and reliability as per FiveThirtyEight, this information can suggest the relationship between the poll and the candidate support outcome. The "methodology" can show the trends of different polling methods in relation with candidate preference results. The "state" variable defines the state of which the poll was conducted in, which can indicate the candidate supported per state. The “state_date” and “end_date” variables can indicate the time the poll was implemented and the time span it was live. It may be able to demonstrate trends of change in the election preferences over time. Controlled variables include those that we would like to keep constant while other independent factors are being examined. In this case, we only want to forecast the outcomes of one candidate, Donald Trump. Therefore, we will control the “name” variable to only be “Trump” and eliminate other candidates when performing our analysis. Failure to do this may lead to the percentage of other candidates influencing Trump’s predicted outcomes.

3 Methodology

##Data Collection

We processed 2024 US presidential election polling data from (**fivethirtyeight?**) by using R (R Core Team 2023). The data includes results for different regional population and time periods. Missing values were processed and post-survey weighting was used to adjust for potential bias.

3.1 Model selection

We tested a variety of linear models for predicting candidate vote share using predictors such as start, end dates, states, and pollsters. Models included simple linear regression models.

3.2 Linear models.

We built a generalized linear model to estimate Trump’s support based on a few key variables selected. By looking at the size and significance of the coefficients (beta values) associated with each variable, we can see which factors are most influential in shaping Trump’s electoral support. From there, we can further assess poll bias and track trends over time. ##Assumptions and Limitations

We assume that there is a linear relationship between the independent variables (e.g., pollster, state, sample size, poll date) and the outcome (Trump’s polling percentage) in a linear model. This means that we assumed that each factor had a constant effect on Trump’s poll numbers, but the reality of dynamic polling does not follow a strictly linear pattern.

##Software and tools

We use the statistical programming language R (R Core Team 2023) and the following packages: tidyverse, dplyr, readr, ggplot2, janitor, lubridate, broom, modelsummary, rstanarm, and splines.

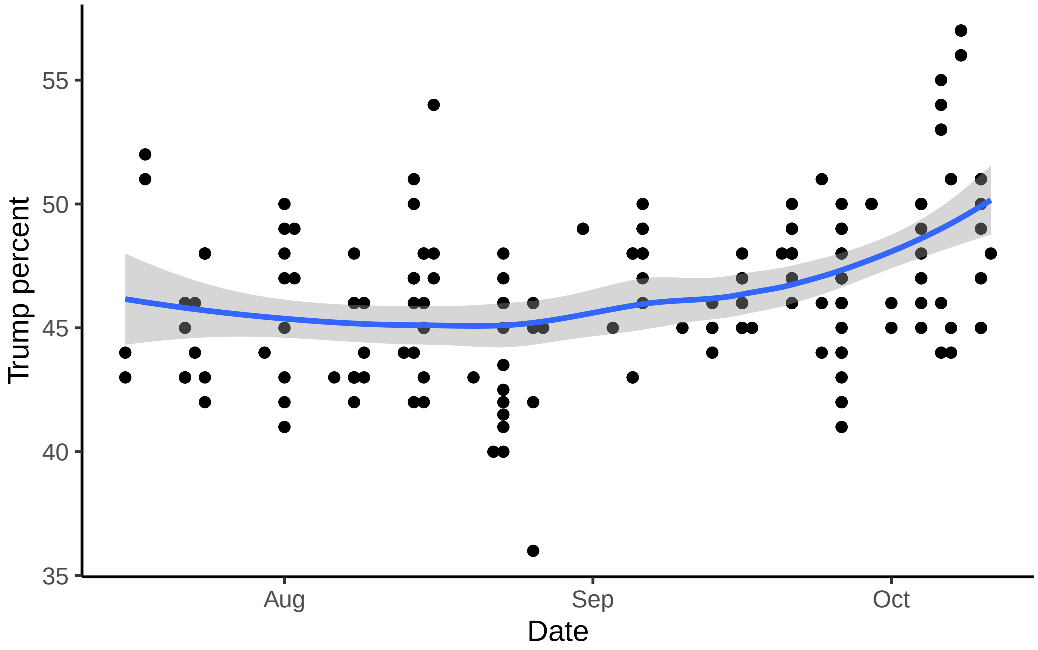


Figure 1: Bills of penguins

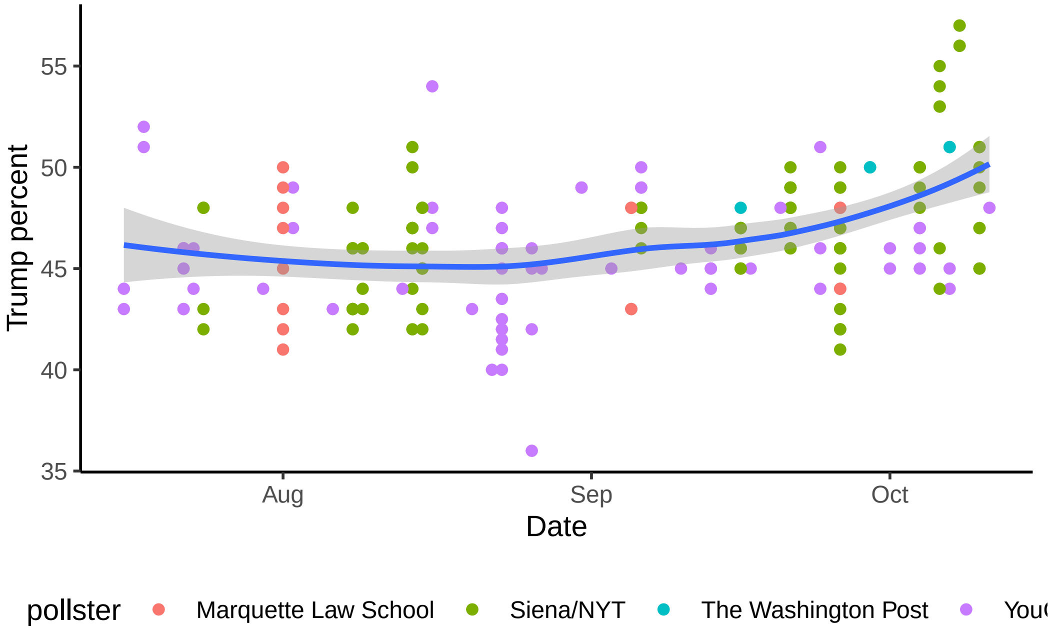


Figure 2: Bills of penguins

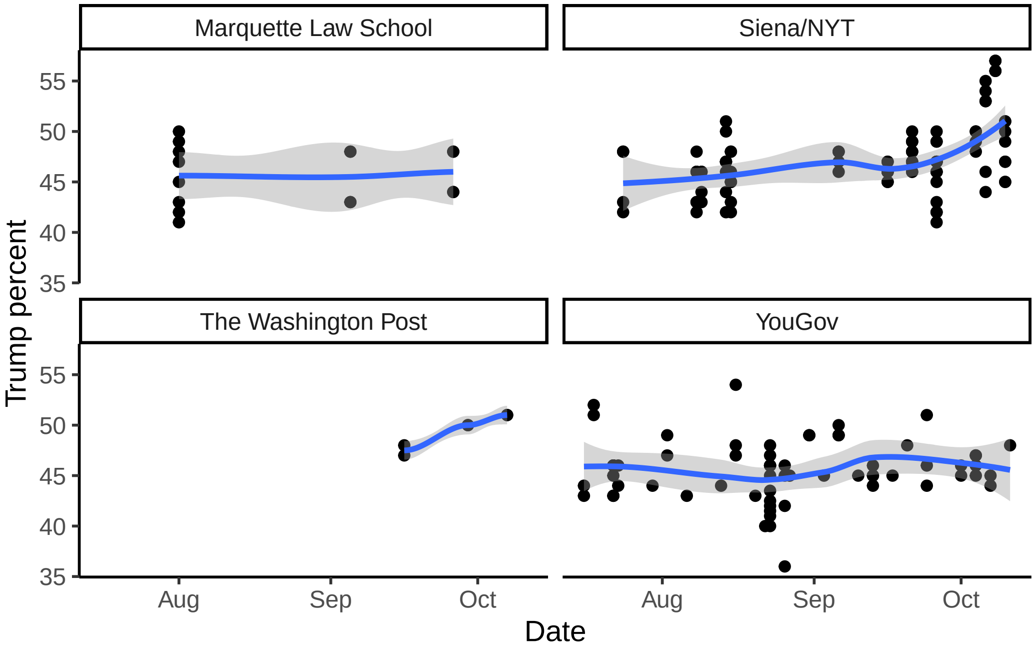


Figure 3: Bills of penguins

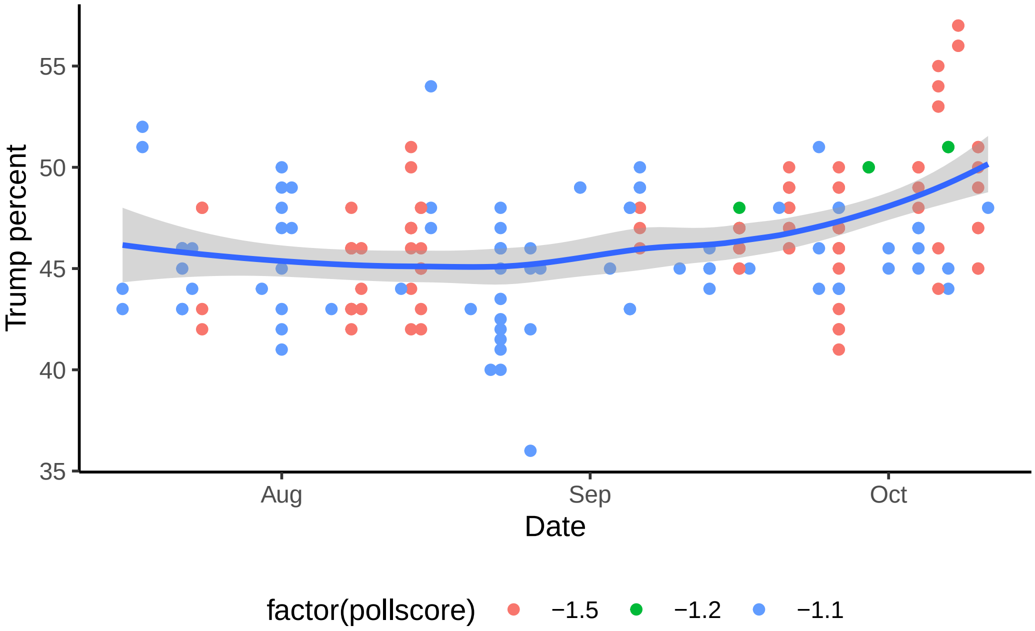


Figure 4: Bills of penguins

Talk more about it.

And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

Talk way more about it.

3.3 Predictor variables

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

4 Model

4.0.1 Model justification

5 Results

Our results are summarized in **?@tbl-modelresults**.

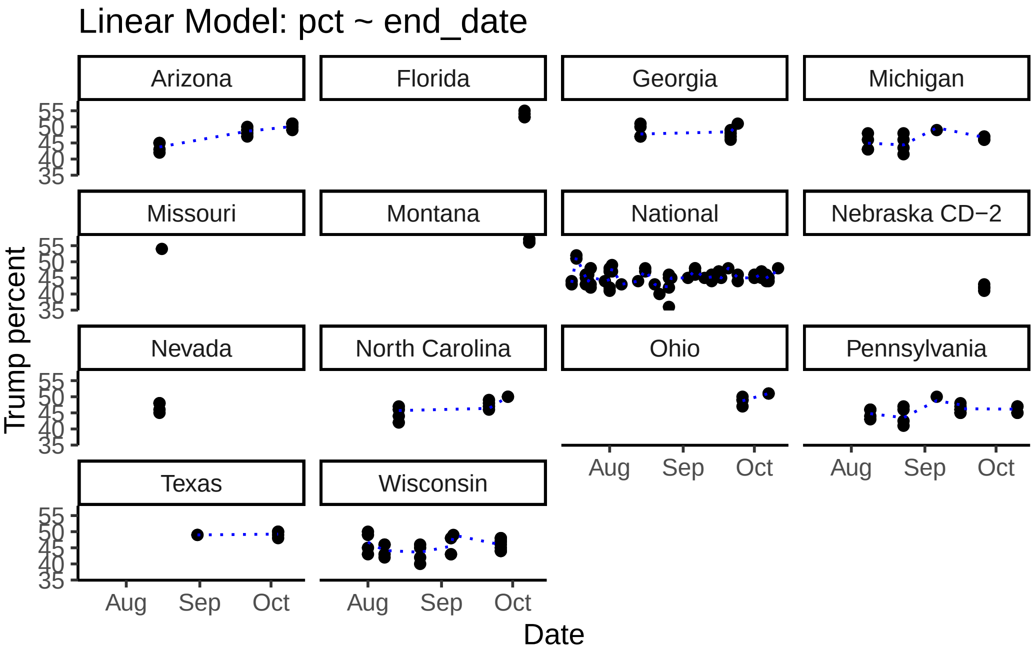
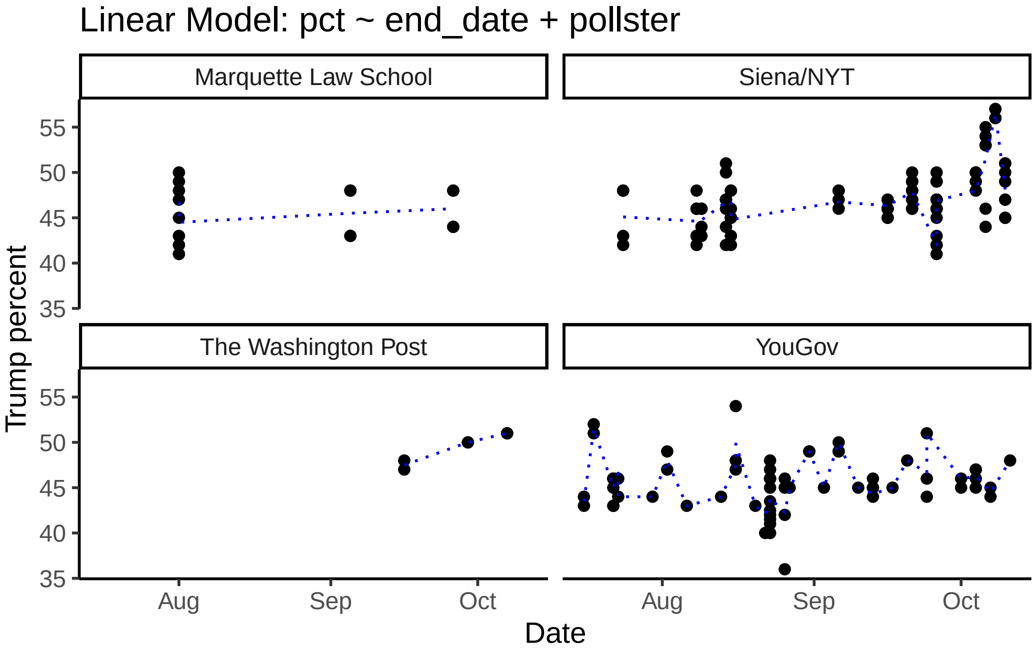


Table 1: **?(caption)**



6 Discussion

6.1 Overview

In this study, we aimed to predict the outcome of the 2024 U.S. Presidential election by analyzing polling data, with a particular focus on Donald Trump's support across different states and pollsters. Our analysis utilized polling data sourced from FiveThirtyEight, with variables such as candidate support percentages, polling methodologies, and the geographical distribution of respondents. The data was analyzed using both linear regression models and Bayesian methods, and our results indicate several key trends.

6.2 Key Findings

The data suggests that Trump's support has remained relatively stable, with polling percentages generally falling between 40% and 50% from August to October 2024. These trends were consistent across both national and state-level polls, though slight variations were observed depending on the pollster and the region in question. Pollsters with higher ratings, as determined by their numeric grade, generally showed more consistent results in predicting Trump's support levels, supporting the hypothesis that higher-quality polling organizations provide more reliable insights into voter behavior.

Additionally, the time factor (i.e., the start and end dates of polls) appeared to have a notable influence on the predictions. This aligns with the understanding that public opinion is dynamic and can shift throughout the election cycle based on external events, candidate campaigns, and media coverage. However, the state-specific analysis showed a smaller degree of fluctuation, with several key states consistently reporting stable support levels for Trump.

6.3 Model Performance and Limitations

While the linear regression models provided a reasonable approximation of Trump's polling performance, the assumption of linearity introduced some limitations. Polling data often exhibits nonlinear patterns due to complex voter dynamics, and our analysis did not fully capture potential interactions between variables such as state, methodology, and time. Additionally, while we focused on Trump, the exclusion of other candidates in the model could lead to oversimplifications. Poll results for other candidates could have indirect effects on Trump's support, which were not accounted for in this analysis.

The Bayesian models added an additional layer of robustness, allowing us to estimate uncertainty more effectively. These models provided credible intervals for Trump's polling percentage, giving us a better understanding of the potential range of outcomes. The posterior predictive checks suggested a good model fit overall, though further refinement might be needed to address potential biases from underrepresented demographics or under-sampled regions.

6.4 Weaknesses and Next Steps

One of the main weaknesses of this study is the reliance on polling data alone. Polls are prone to various forms of bias, including non-response bias and sampling errors. Additionally, the changing nature of campaigns, external events, and voter turnout can lead to discrepancies between polling predictions and actual election outcomes. Another limitation is the focus on national-level polling; while state-level analysis was conducted, more granular regional data might provide deeper insights into key battleground areas.

Future research should explore the inclusion of additional variables such as media coverage, campaign spending, and real-time voter engagement metrics to complement the polling data. Furthermore, expanding the analysis to include dynamic models that account for time-varying effects could provide a more accurate forecast of electoral outcomes.

7 Conclusion

This paper examines polling data for the 2024 U.S. Presidential election, with a focus on trends in candidate support and the factors affecting polling outcomes. Using data compiled by FiveThirtyEight from sources such as YouGov, Siena/NYT, Marquette Law School, and

The Washington Post, we modeled polling percentages over various time periods, pollsters, and states. The analysis was done specifically for the support of Donald Trump. To account for variations in the data and estimate uncertainty in the predictions, a linear regression model was employed. The results showed that the support for Donald Trump was stable and increasing over time from early August to late October. The state-specific trends all show the support for Donald Trump at around 40-50% support rate, which is high considering that there are 5 candidates currently in the running as of October of 2024. The poll-specific trends out of the pollsters that were graded at 3 or higher by FiveThirtyEight showed similar results where Trump's support was around 40-50%. Overall, this indicates stable support for Trump that's higher and still on the increasing trend. With the current situation, it is likely that Trump prevails in the presidency election. However, it is important to continue to monitor each candidates' campaigns and potential turn of events in support rates to provide the most accurate forecast possible.

Appendix

.1 Appendix A – Idealized Methodology and Survey

.1.1 Overview

YouGov is an internationally recognized polling organization known for its expertise in political forecasting. It is particularly noted for its use of the "poll-of-polls" approach, which aggregates multiple surveys to create a more accurate picture of public opinion. Relying heavily on online panel polling, YouGov provides timely, cost-effective data, particularly useful in the context of U.S. presidential elections. Below is an examination of the key components of YouGov's polling techniques.

.1.2 Sampling

YouGov's polling targets U.S. likely voters, identified based on eligibility and past voting behavior, voter registration status, and intent to vote in the upcoming election. The firm uses a vast online panel, consisting of millions of U.S. residents recruited through various digital channels. This panel is strategically segmented and weighted to match the U.S. population's key demographic characteristics such as age, gender, race, education, and geographic region. Participants for YouGov's polls are recruited through a variety of methods, including targeted online advertisements, partnerships with affiliate websites, and incentive-based surveys where respondents earn rewards for participation. While these methods help build a large and diverse panel, there is an inherent risk of self-selection bias, as individuals who voluntarily join the panel may not be fully representative of the broader population. YouGov uses a stratified sampling method, establishing quotas for demographic categories like age, race, and region, and applies post-stratification weighting to ensure the sample more closely reflects the overall population.

.1.3 Advantages and Disadvantages

An advantage of YouGov is cost and speed. Online panels are less expensive and faster to implement than traditional methods like phone surveys. Another advantage is targeting efficiency. They target specific groups, such as likely voters, that can be reached more easily and efficiently. A disadvantage is the non-probability sampling method used. This is because not all individuals have an equal chance of being selected, which introduces potential bias. Another disadvantage is internet accessibility. Having the surveys online leads to the issue where individuals without internet access, particularly those from older or lower-income demographics, may be underrepresented, which could affect the results.

.1.4 Survey Design

YouGov's surveys are designed to be clear and consistent, often pre-tested on smaller groups to ensure accuracy. The questions are regularly updated to stay relevant to current political events. However, the use of online surveys limits the depth of questions, often favoring simpler, multiple-choice formats. This simplicity can lead to superficial responses, especially if participants experience survey fatigue due to frequent polling requests. Additionally, individuals without internet access are excluded, which may reduce the representativeness of certain demographic groups.

.1.5 Conclusion

YouGov's use of online panels offers significant advantages in terms of cost, speed, and flexibility, making it a valuable tool for political forecasting. While it effectively uses stratification and weighting to mitigate some biases, challenges such as self-selection, non-response bias, and the exclusion of individuals without internet access persist. Despite these limitations, YouGov remains a key player in the field of modern political polling, providing valuable insights for political analysis and forecasting.

.2 Appendix B – Idealized Methodology and Survey

.2.1 Overview

This appendix explains an idealized methodology for the data collection phase to predict the US presidential election results. Current pollster methodologies mainly include panels, live phone calls, interactive voice responses (IVR), and text-to-web. Many characteristics of a poll can determine the quality of the poll, including but not limited to timing, suitability of methodology for the target, scope of distribution, and many more (citation). The idealized methodology described below considers these aspects in the context of predicting election results and shapes the poll in a way that best suits this purpose.

.2.2 Sample Frame and Sample

An idealized methodology, given a large budget of $100K, would aim to recruit participants that are representative of the entire nation. A way of recruiting participants that is most likely to gain responses that is most representative of the nation is stratified sampling by state. Eligible voters will be stratified based on the state they would be voting from. A weighted random sample would be conducted, meaning that a specific number of participants will be randomly selected from each state based on the population size of the state relative to the country. The optimal sample size for a total number of voters of more than 160M (based on previous numbers of total voters) is around 2500 people (citation for number of voters; citation for optimal number of voters). We will aim for 3000 people to account for potential incomplete responses, no responses, or other issues with specific respondents. Based on the populations of each state, a specific number out of the 3000 total respondents will be allocated to the state. The respondents will be selected at random and contacted by email and in person. The respondents will be given a survey, which will be described in detail in the next few paragraphs. The responses will be gathered and analyzed to generate a forecast for the election results of the current election cycle.

.2.3 Polling Methodology

The ideal polling method to reach the amount of people we aim to reach would be an online survey-style questionnaire. This method would allow us to reach a large amount of people in a short amount of time and have quantifiable data that's easy to analyze. This can also allow us to allocate most of the budget to sample selection to ensure that the sample is representative. The ideal survey will be a set of questions regarding the political positions, preferences for each candidate, tendency to vote a specific candidate, and potentially other questions. Each question will be worded in a way that is easy to understand. Each question will be asked in a neutral way to ensure that ideas aren't primed for the respondents. The order of the questions is also important. It is difficult to balance primacy, recency, and moderacy effects for each participant because only one can be first and one can be last. A way to control for this is to scramble to order of questions on parties and candidates within stratified samples. For example, for two participants living in California, one would be asked about their thoughts on the Democratic party first while the other would be asked about the Republican party first. Similar order manipulation can be done for the candidates. Another aspect that can be added to the surveys is demographic information. Although this may not be directly related to forecasting the election, it could offer interesting statistical patterns that can be analyzed for other purposes in the future. These questions can include "what is your ethnicity?", "what is your self-identified gender identity?", and many more.

.2.4 Survey

Based on the description of an idealized methodology and survey, a survey has been constructed to abide to these conditions. This has been attached in the link and sample figures below. https://forms.gle/WpFwHWs5YtPzGRTm9

A Additional data details

B Model details

B.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

Figure 5: **?(caption)**

B.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

Figure 6: **?(caption)**

References

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