Analysing the 2024 U.S. Presidential Election*

Spotlight on North Carolina's Role as a Battleground State

David Qi Haobo Ren Xinrui Xie

November 4, 2024

This paper examines the relationship between polling results and economic conditions to forecast the 2024 U.S. presidential election, with results analyzed for both swing states and at a national level. Findings suggest that Republicans have a modest chance (6.6%) of winning, with North Carolina emerging as the pivotal swing state. The study may offer guidance for campaign strategies and market forcasts. While providing understanding on changes of swing states in U.S. politics and how Florida is no longer a swing state.

Table of contents

*Code and data are available at: https://github.com/HaoboRrrr/USA_Election

1 Introduction

Josh Pasek have discussed how to the use of Aggregation and Predictive Modeling on polling can increase prediction effectiveness(Pasek 2015). Following this principle, we will make use of predictive modeling to create a model that aim to predict the results of the 2024 United States presidential election. This paper uses polling data gathered by FiveThirtyEight(FiveThirtyEight 2024b) with Dow Jones industrial average data(S &P Dow Jones Indices LLC 2024) as an indicator of the economy to train a simple linear regression model to predict the result of election in each swing state.

The model aims to predict the result of a poll that ends on November 5^{th}, 2024, we believe this will reflect the result of the upcoming election on that day, the result of the election may be assumed to be close to the mean of these polls. And by this process we derive the result of the election. After the result is gained, we computed the electoral votes for the state, and used Monte Carlo simulation to derive winning possibility of each party.

We conclude that Donald Trump have 6.6% probability of winning this election. From our prediction, we highlight North Carolina as the most important battleground state. The state has a predicted 14.2% probability to be won by Trump, and will give him 16 electoral votes. Along with the votes that we predict he will almost certainly gain, this pushes him one step (9 electoral votes) to victory. This provide powerful guidence to how campaign and fund raising strategies should be targeted on the election day. And provides a forecast on one of the most important influences in economy in the next several months, which will be especially helpful in planning market stragities.

The scripts in this paper are written in R(R Core Team 2023), using mainly tidyverse(Wickham et al. 2019) The table in our study are generated by knitr (Xie 2014), and graph with ggplot2(Wickham 2016). Data input and output were done using arrow(Richardson et al. 2024), and curl(Ooms 2024). Data cleaning made use of janitor(Firke 2023).

The rest of the paper is structured as follows: in Data section we will discuss the various aspects of the data we use, how they were gathered and the variable we shall use. In the Model section we will discuss details of our model, why it is justified and the coefficient summaries. In the Results section, we present the predictions of the model. In the Discussion section we will discuss the interpretation to the predictions, and what we may learn from the coefficients for the model, we will also look as possible weakness and future improvement for the model.

2 Data

We will make use of polling data organized by FiveThirtyEight(FiveThirtyEight 2024b), collected from polling made by various sources before the election. As well as the Dow Jones industrial average data(S &P Dow Jones Indices LLC 2024).

2.1 Overview For Polling data

Fist we take a look at the the polling data.

The data set contain 17208 observations of polls, for our propose, we choose only to include 1176 entries, the justification for this data selection is given in the Justification section.

2.2 Data Measurement

The data was organized by FiveThirtyEight (2024), collected form polls made by various institutions, they each use a distinct methodology. While FiveThirtyEight checks their methodology, and include the polling data if it meets their standard. An example of the methodology of one poll is included in Appendix 1 It is a reliable and comprehensive source of US election polls so we have choose this data set.

2.3 Justification For Poll Selection

The goal of our study is to predict the result of US presidential election, it is known that the only reasonable winner will be form the Democratic Party, or the Republican Party. Which can be seen form their overwhelmingly high support rate in the polling results and historical results. So we will only include the polling results for Kamala Harris and Donald Trump, the candidate for the two parties.

Another observation coming form past experience that is not so obvious is that the result of presidential elections are often only determined by the result in "swing states".(Schultz and Hecht 2017; Aldrich et al. 2023) Following the definition given by Schultz and Hecht, which have given having below 5% difference between the two parties as one criteria, we have examined our data and have identified the swing states from the data set, past experience, and outside sources.

Table 1 Shows the absolute value of the difference in average support rate from all the polls. If we take all states such the absolute difference is less than 5% to be swing states, similar to the method of less than 5% difference between two parties in the last election. We will see they coincide except that our list includes Maine CD-2 and have not included Florida(Federal Election Commission 2021). This list, when Florida is added, also include the swing state

listed by popular media(BBC News 2024; Dorn 2024; ABC News 2024) and polling platform(FiveThirtyEight 2024a) before this upcoming election.

Also note that some states only have very limited polls, which will make modeling these states very difficult and inaccurate, the swing states we choose allow us to have sufficient polls for each state.

The list of swing states we use are: North Carolina, Pennsylvania, Nevada, Georgia, Arizona, Michigan, Wisconsin, Maine's 2nd congressional district, and Florida. We will also include national polls to help with modeling.

The data have also contained ratings for each poll based on their methodology and transparency. As a summary made to understand the data, Table 1 is made by selecting all the polling, however, for the rest of this study, we will only use polls that are rated 2.5 or above by FiveThirtyEight to ensure quality of data.

Table 1: Comparison of average polling result for two candiates

State	Number of Polls	Average Support Rate(Harris)	Average Support Rate(Trump)	Absolute difference
North Carolina	136	47.35831	47.86676	0.5084559
Pennsylvania	196	47.94173	47.11251	0.8292224
Nevada	96	47.44635	46.57792	0.8684375
Georgia	128	47.15758	48.04609	0.8885156
Arizona	129	46.80628	47.78860	0.9823256
Michigan	159	47.89094	46.37331	1.5176305
Wisconsin	157	48.62229	46.50686	2.1154376
National	632	47.89134	45.65974	2.2316049
Maine CD-2	10	44.30000	47.30000	3.0000000
Iowa	2	43.80000	49.15000	5.3500000
Florida	57	44.51491	50.28881	5.7739013
Minnesota	22	49.63091	43.33667	6.2942424
Texas	43	44.43023	50.90767	6.4774419
Virginia	27	49.26037	42.75852	6.5018519
New Hampshire	21	50.74286	43.81818	6.9246753
New Mexico	11	49.05455	41.78182	7.2727273
Oregon	3	49.00000	41.66667	7.3333333
Ohio	30	44.32367	51.67200	7.3483333
Kansas	2	41.30000	49.05000	7.7500000

Table 1: Comparison of average polling result for two candiates

State	Number of Polls	Average Support Rate(Harris)	Average Support Rate(Trump)	Absolute difference
Nebraska CD-1	1	43.00000	51.00000	8.0000000
Alaska	7	43.58571	51.71429	8.1285714
Nebraska CD-2	11	50.81818	42.09091	8.7272727
Maine	7	52.00000	41.28571	10.7142857
South Carolina	7	41.84857	54.36286	12.5142857
Colorado	6	54.76667	41.70000	13.0666667
Missouri	7	42.21429	55.44286	13.2285714
New Jersey	6	53.75000	39.95000	13.8000000
Indiana	5	41.40000	55.92000	14.5200000
Arkansas	1	40.00000	55.00000	15.0000000
Connecticut	1	53.00000	37.00000	16.0000000
Nebraska	14	39.00000	55.21429	16.2142857
New York	13	55.93846	39.43846	16.5000000
Illinois	3	58.33333	41.66667	16.6666667
Montana	13	38.91538	55.94615	17.0307692
Rhode Island	5	54.40000	37.20000	17.2000000
Delaware	2	54.80000	36.45000	18.3500000
North Dakota	2	36.00000	54.50000	18.5000000
Washington	7	56.15714	36.84286	19.3142857
Utah	10	34.70000	55.30000	20.6000000
Tennessee	5	36.56000	59.44000	22.8800000
Maine CD-1	7	59.14286	35.14286	24.0000000
California	18	60.17222	35.02273	25.1494949
South Dakota	2	34.75000	60.50000	25.7500000
West Virginia	1	34.00000	61.00000	27.0000000
Massachusetts	8	61.91250	33.06250	28.8500000
Maryland	17	62.02000	33.08647	28.9335294
Oklahoma	4	34.00000	63.47500	29.4750000
Vermont	2	68.50000	28.00000	40.5000000

Table 1: Comparison of average polling result for two candiates

State	Number of Polls	Average Support Rate(Harris)	Average Support Rate(Trump)	Absolute difference
Nebraska CD-3	1	25.00000	70.00000	45.0000000

2.4 Varibles

This section explains the important variables relevant to our analysis: Percent, End Date, state, numeric grade, Candidate Name and sample size.

2.4.1 Percent

Percent (`pct`) is the percentage support rate learned by this poll. This is the main estimated we are targeted to model, the justification of using this as estimand is given in model section.

2.4.2 End Date

End Date(`end_date`) is a time variable that specify the date that the poll ends.

2.4.3 State

State (`state`) is a categorical variable representing each U.S. state or Congressional district that is considered a separate entity in the U.S. presidential election. Each poll reflect the result of a state, if the state is not specified, the poll is national.

2.4.4 Numeric grade

Numeric grade (`numeric_grade`) variable is a measurement of the quality of the poll given by FiveThirtyEight.It is a numeric score from 3 to 0. In this study, we will only use polls with Numeric grade higher than 2.5.

2.4.5 Candidate Name

Candidate Name (`candidate_name`) is a categorical variable that specify the name of candidate whose the support rate this poll measure.

2.4.6 Sample Size

Figure 1 is a plot showing the sample size(`sample_size`) of the data we use for modeling. The sample size of each poll is 432 minimum and 78247 maximum, the mean sample size is 1497. Observe that the sample size is extremely small comparing to the entire population of a state, this observation is important in how we predict using our model.

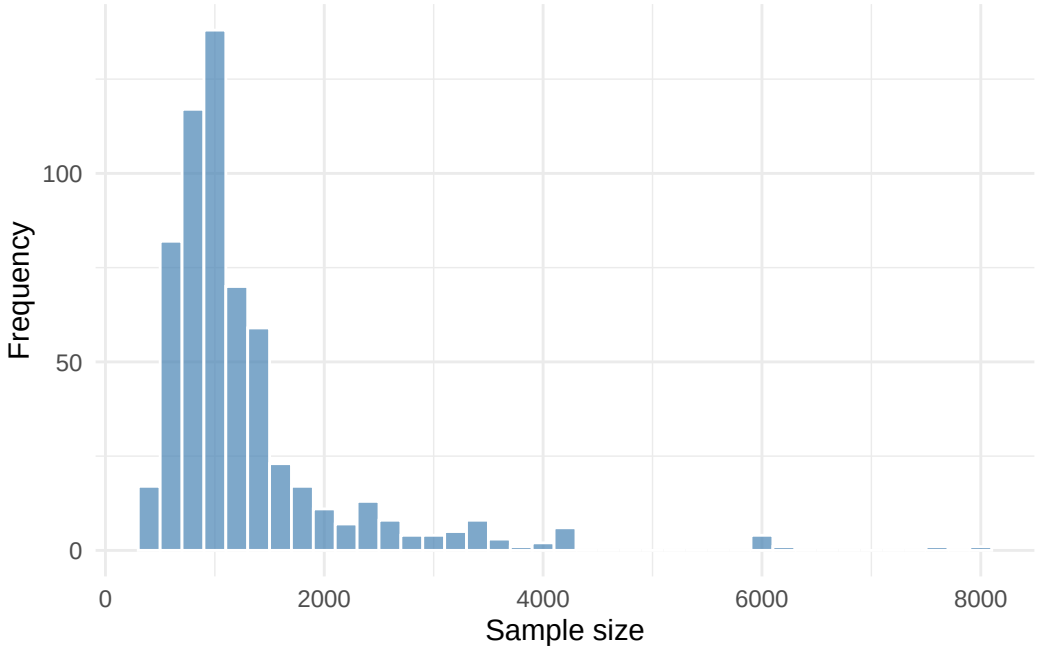

Figure 1: Sample Sizes of the Polls We Use

2.5 Overview Dow Jones Data

The Dow Jones Industrial Average(DJIA) is a renowned economic indicator that launched in 1896(S&P Dow Jones Indices 2024). It is used long in economical analysis and is considered "one of the analyst's oldest friend and one of his most useful tool"(Milne 1966). The index is still being used in economic studies today and is considered an important indicator of the economic of the United States(Arendas, Malacka, and Schwarzova 2018).

The data includes average DJIA data for each weekday, there are no data for weekends and certain holidays that DJIA do not operate. For the propose of this analysis, we have filled the data by assigning these dates with DJIA values on the closest date before. The data after filling is a numerical value corresponding to each day.

2.6 DJIA Data Measurement

The data publisher(S&P Dow Jones Indices LLC) have full transparency and reproducible of their methodology(Milne 1966), DJIA is a weighted average of stock price of 30 firm that are considered important to all aspect of economy except transportation and utilities. The firm and corresponding weights are fully transparent. As these firms play a key role in the U.S. economy and the presidential election, Dow Jones Industrial Average will be a useful predictor for electoral results.

It is also a daily updated, easy to access data that is not subject to change. Some other important indicators of economy, such as United States Nonfarm Payrolls, only updates monthly and previous data may be corrected in the future(U.S. Bureau of Labor Statistics (n.d.)).

3 Model

We will use simple linear regression model to predict the general polling result of Harris and Trump. The mode will be implemented using R(R Core Team 2023).

3.1 Model set-up

For each poll, we assumed they are being sampled similarly.
Let the support rate for Harris in a single poll that end at day t be denoted by H_t. And we denote the Dow Jones Industrial Average at day t by D_t.Denote State by s, and the corresponding regression coefficient by β_s Then we will use linear regression model:

$$H_t = \beta_0 + \beta_1 D_{t-7} + \beta_2 D_{t-14} + \beta_3 D_{t-28} + \beta_4 D_{t-60} + \beta_5 D_{t-150} + \beta_6 D_{t-180} + \beta_s + \epsilon_t$$

Where ϵ_t is an independent, normally distributed error term.
$\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6$ are regression coefficients.
β_s is the corresponding regression coefficient to state s.
We will predict the election support rate of Harris to be $mean(H_t)$

For Donald Trump, we use a similar model, let the support rate in a single poll for Trump at day t be denoted by T_t. The the linear regression model can be written as:

$$T_t = \alpha_0 + \alpha_1 D_{t-7} + \alpha_2 D_{t-14} + \alpha_3 D_{t-28} + \alpha_4 D_{t-60} + \alpha_5 D_{t-150} + \alpha_6 D_{t-180} + \alpha_s + e_t$$

Where e_t is an independent, normally distributed error term.
$\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6$ are regression coefficients.
α_s is the corresponding regression coefficient to state s.
We will predict the election support rate of Donald Trump to be $mean(T_t)$

3.2 Model justification

In deciding this model, we have made many decisions. In this section we justify the importat choices we made during modeling.

3.2.1 Use of DJIA

The use of economic data to predict result in presidential election can be supported by various studies that shows the result and voting pattern in presidential elections are highly correlated to the economy(Aldrich et al. 2023). And DJIA is an important and wide used indicator of the economy. We hypothesize that DJIA will have a linear relationship with the support rate. We expect the effect economy has on polling results to have a delayed effect, which is the reason for including lagged terms in linear regression.

An alternative approach is to use changes in the DJIA to reflect economic fluctuations. However, we choose to use the absolute DJIA values, as inputing missing days for change data is difficult to justify, and the absolute economic status cannot be reflected by the change in one day. Our hypothesis is that a strong economy favors the ruling party in elections, as switching parties could risk economic shifts and potentially lower living standards. Here, the absolute economic level serves as an indicator, reflecting the overall standard of living.

3.2.2 Use of mean to predict election

We may treat each poll as an observation for the population of support rate, what our model does is that it predicts the Percent('pct') variable in a single poll. Since the sample size of is extremely small comparing to the population of each state, which is 1497 people on average comparing to at least 700 thousand people. And by statistic around 60% people will participate in each election(Aldrich et al. 2023), we may consider the election result to be very close to the population mean by the weak law of large numbers. So by considering the mean of the predicted polls, we may predict the election result.

3.2.3 Using different model for Harris and Trump

There exist the possibility for using one model with a categorical variable, however we find the voting behavior of the two groups of people may be vastly different and cannot be summarized by a simple shift in the regression line. The use of interaction terms may capture this difference, however will make the model more complicated and difficult to write out.

3.2.4 Using the same model for each state

As states in United States, we expect each state to share some similarity with other states, as well as with the national data. So we have made state a categorical variable in regression. The main difference with candidates here is that we have data on national polls, which have direct correlation with the polling result of each state rather than just hypothesized connection, to make use of those data, we include them by using a categorical for state.

3.2.5 Linerity and Independence Assumption

The justification for these assumptions are in Appendix 3

3.2.6 Reason for choosing these lag intervals

Selection process is explained in Appendix 3, we have explicitly made two model useing the same set of variables so comparison is possible.

3.2.7 Including only swing states

As we stated above, the election is determined almost solely by swing states. And many other states do not have enough data for sound prediction. Thus our model only include the data for swing states and nation polls.

3.3 Model summary

Table 2 and Table 3 are the model coefficients fitted for Harris and Trump. The model for Harris have adjusted R-squared 0.1781, and model for trump have adjusted R-squared 0.3195.

Note that DJIA_lag_s is the variable D_{t-s}

Table 2: Model summary for Harris

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-54.0660685	18.9315602	-2.8558697	0.0044412
DJIA_lag_7	-0.0005393	0.0001708	-3.1574568	0.0016718
DJIA_lag_14	0.0009283	0.0001693	5.4828319	0.0000001
DJIA_lag_28	-0.0000371	0.0002097	-0.1770339	0.8595418
DJIA_lag_60	0.0002896	0.0002533	1.1432876	0.2533774
DJIA_lag_150	0.0006820	0.0002549	2.6757187	0.0076615

Table 2: Model summary for Harris

	Estimate	Std. Error	t value	Pr(>|t|)
DJIA_lag_180	0.0012440	0.0002292	5.4271579	0.0000001
stateFlorida	-2.6907785	0.6921252	-3.8877050	0.0001125
stateGeorgia	0.4974847	0.4858331	1.0239827	0.3062582
stateMaine CD-2	1.3305681	1.1063681	1.2026450	0.2295905
stateMichigan	1.2979068	0.4580112	2.8337884	0.0047555
stateNational	0.9957287	0.3838082	2.5943391	0.0097100
stateNevada	1.2069144	0.5768089	2.0923989	0.0368245
stateNorth Carolina	1.1388302	0.4508729	2.5258343	0.0117999
statePennsylvania	1.3898126	0.4305804	3.2277660	0.0013160
stateWisconsin	2.2287666	0.4458997	4.9983581	0.0000008

Table 3: Model summary for Trump

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-16.6158282	28.7174566	-0.5785968	0.5630802
DJIA_lag_7	-0.0001238	0.0001770	-0.6993244	0.4846226
DJIA_lag_14	0.0009952	0.0001851	5.3770647	0.0000001
DJIA_lag_28	-0.0005880	0.0003009	-1.9541033	0.0511571
DJIA_lag_60	0.0007728	0.0003244	2.3822249	0.0175216
DJIA_lag_90	-0.0004417	0.0002670	-1.6545396	0.0985453
DJIA_lag_120	-0.0000762	0.0003335	-0.2285582	0.8192909
DJIA_lag_150	0.0006650	0.0002804	2.3719288	0.0180127
DJIA_lag_180	0.0004529	0.0002778	1.6302158	0.1035851
stateFlorida	1.9891740	0.7142195	2.7851017	0.0055215
stateGeorgia	-0.3390713	0.4988264	-0.6797381	0.4969347
stateMaine CD-2	-1.0424738	1.1378168	-0.9162053	0.3599304
stateMichigan	-2.3541752	0.4702566	-5.0061505	0.0000007
stateNational	-3.5854099	0.3979204	-9.0103691	0.0000000
stateNevada	-1.5426254	0.5925286	-2.6034617	0.0094590
stateNorth Carolina	-1.2692506	0.4636078	-2.7377683	0.0063708
statePennsylvania	-2.0492433	0.4430287	-4.6255319	0.0000046
stateWisconsin	-2.4290600	0.4578286	-5.3056101	0.0000002

4 Results

4.1 Model prediction

Table 4 shows the predicted result for the election in the swing states as well as the national result. However, this table have not considered confidence intervals. Note that the reason for using confidence interval instead of prediction intervals is that as explained in Use of mean to predict election section, the actual result of the election is expected to be close to the population mean. For instance in North Carolina, the 95% confidence interval for Trump is $[48.22990, 49.91708]$, while for Harris it is $[48.90044, 50.54273]$, they have overlapped, which means that we are not very certain with the result. To gain a better understanding of the possibility of a person winning in each state and the nation, we use Monte Carlo simulation to get the possibility. The last column of Table 4 is made using this method and give the possibility we predict for Trump to win each state.

Table 4: Prediction for Each state

State	Prediction Trump	Prediction Harris	Predicted Result	Trump's Win Rate
North Carolina	49.07349	49.72159	Harris win	0.14164
Pennsylvania	48.29182	49.97257	Harris win	0.00123
Georgia	49.98242	49.08024	Trump Win	0.91569
Arizona	50.31073	48.58276	Trump Win	0.99599
Nevada	48.76209	49.78967	Harris win	0.09500
Michigan	47.96037	49.88066	Harris win	0.00080
Wisconsin	47.88471	50.81152	Harris win	0.00000
Maine CD-2	49.17846	49.91332	Harris win	0.32732
Florida	52.33229	45.89198	Trump Win	1.00000
National	46.79477	49.57848	Harris win	0.00000

Note that the United States uses Electoral College system to determine the winner of the election. Given this model, we may predict using Monte Carlo simulation the win rate of each candidate. Assume that each state except for swing states voted for the the party they voted for in 2020, this gives Democratic party 226 electoral votes and The republican 188 electoral votes. Using our model, we find that republican wins 66199 out of 1000000 simulation with no draws, mean electoral votes for democratic is 284.765953 and mean electoral votes for republican is 253.234047. The estimated win rate for republican is 0.066199. If we follow our prediction with out considering possibility, Harris will win by 293 versus 245 electoral votes.

5 Discussion

Our studies have presented a forecast of the U.S. 2024 presidential election

5.1 Interpretation of Model Coefficients

First we note that unlike our initial instinct, the votes of the two parties are not always mutually exclusive. In the case of β_6 and α_6(DJIA_lag_180), the two candidates both have positively correlated support rate. This may indicate that people are less likely to select an independent candidate when the economic is good, as people will be less likely to seek change under such circumstance. However in the case of states, we see that every β_s have opposite sign from α_s.

We also see there is a difference in the R-squares, this may tell us that economy is a better predictor for support rate of Trump than Harris, possibly indicating that Trump has a campaign more focused on economy. β_6 is almost 2 times as large as α_6, which aligns with our hypothesis that good long term economy will make the election favor the ruling party.

5.2 Comparson with 2020

Looking back at 2020, based on our prediction, the difference was that Harris is predicted to take North Carolina and the entire Maine, While Trump is predicted to win Arizona and Georgia, which makes the republican win 10 more votes than the 2020 election. Or 13 more votes after considering the slight changes to number of electoral votes of each state. However the Democratic party is in a great advantage and will not be defeated unless Trump wins in North Carolina. Which by our prediction is difficult but not impossible.

5.3 Impact of Each Swing state

The modeling have highlighted the impact swing states has on presidential election. And also suggest the definition of swing state is subject to change. Florida may be considered a swing state from the election results in 2020, however the polling result and model analysis have shown that Trump will certainly win this state. We also need to highlight the importance and closeness of the election in North Carolina. As a swing state with 16 electoral votes, we predict Republican Party to have a reasonable probability to win North Carolina, and North Carolina will be deceive in determining the election results.

5.4 Weaknesses and next steps

Our studies have relied solely on DJIA data and have not considered other factors that also has great influence to the result of the presidential election. Although as Pasek(2015) says, it is never possible to model all the factors, it is even unrealistic to model all the important factors since each year's election is unique in it's way. We still may improve the analysis by giving consideration to more variables. For instance, we may do a weighted regression by using the numerical scores as weights, or consider other measurements of economy that covers the transportation and utility aspect of economy that DJIA did not cover.

Although we have claimed that non-swing states will almost certainly vote for the party they supported, bringing in their data still may help our. However we have to note having all the states will greatly increase the complexity of the model, and prediction for many of the states will not be trustworthy due to scarcity of relevant data.

In the modeling, we have view polling as taking one sample form the population, while election as taking 60% of the data form the population. However, polls may be biased by various factors.For example non-response bias or moderacy bias(Stantcheva 2023). A closer and more careful investigation by looking at the difference between poll results an actual election outcomes is required to understand the relationship between polls and actual election.

A Appendix 1: Polling methodology

A.1 Overview of Morning Consult

Morning Consult is a globally renowned enterprise technology company that specializes in providing intelligent data to support leaders' decision-making processes. Morning Consult has established itself as a key player in the field of artificial intelligence and decision-making consulting. The company offers a suite of services that include market research, brand analysis, consumer behavior studies, and more. So that the company can optimize decision-making processes and provide clients with a competitive edge(Consult 2024b)

A.2 Features of the Sample

- Population: The survey aims at all the American citizens who is eligible to vote in the 2024 presidential election in all the parties.
- Sampling Frame: A list of the American citizen's email address and telephone number.
- Sample Size: The polling samples showed a time range from 2011 to 2024. During each election period, the sample size varies a lot. The sample size ranges from 111 to 78247 across the voting period.

A.3 Sample Recruitment

The company used many sampling methods, such as online panel, live phone call, text, emails and a mix of these approaches. During the recruitment, the data was collected by contacting landlines via Interactive Voice Response. The sampling teams made phone calls and sent text message to randomly selected phone numbers, and collect their responses. The online team collect votes from the active users in the websites and other online panel of voters provided by research marketplace such as CINT. However, some of the polling results come from a survey from long ago, which means that it might be incorrect or biased(Consult 2024a).

A.4 Sample Approach

Both probability and non-probability sampling methods are involved in the sample collection. For the people surveyed via phone calls and text messages, the main sampling method is Simple Random Sample (SRS). While for those sampled via online panel, the main approach is convenience sample.

The Morning Consult company used so-called "tracking polls" (Consult 2024a), interviews from a certain period are incorporated into future polls, getting reweighted with different samples until they are too old and dropped from the analysis.

A.5 Strengths and Weakness

Strengths: Since both probability and non-probability sampling methods are used, the pollster can cover a wider range of people surveyed. This means that the sample result can better represent the whole population. Also, the reweighting mechanism can guarantee different methods can have different influence on the final result.

Weakness: According to the data collected, majority of the pollster are done by the convenience sample, which is a non-probability sampling. Thus the actual surveyed population may be biased from the target population, and the result may be less representative. Additionally, the survey showed some overlapping time period, which means that some opinions of the interviewers may be counted more than once.

A.6 Non-response Handling

Just like all the surveys, non-response bias is the biggest challenge met during the pollster. This is also the greatest factor that may affect the effectiveness of the survey. In all these sampling recruitment and sampling approaches, a portion of the respondents would fail to answer the survey questions as required for various reasons (such as refusing to participate, being unreachable, or being unable to understand the survey questions).

Based on the online research, Morning Consult company usually adopt these measures to minimize the effect of the non-response bias, such as increasing the survey response rate, following up with non-respondents, and using a broader sample for the survey. Additionally, by comparing and analyzing the characteristics of respondents and non-respondents and supplementing the data with other relevant sources, researchers can further assess the influence of non-response bias on the study results.

According to the collected data, we can also see that all the sample recruitment methods request the use of the cell phone. This automatically left out the people who do not use cell phones or who do not use Internet, such as some old people. But this type of non-response bias is still unsolved.

A.7 Comments of the Questionnaire

The good thing about the questionnaires is that most of the questions are direct and easy to understand. So that this can avoid the bias due to misunderstanding of the questions. Secondly, the questionnaire is short, which makes more people willing to answer without causing waste of time.

On the other hand, the questionnaire requests the interviewers to identify their party, which may cause people not willing to vote for the candidates from the other party, even though they actually want to. This will also cause a slight bias in the response.

B Appendix 2: Ideal methodology and survey

B.1 Objective

The survey aims to capture voter intention and mood across important demographic and geographic divisions in order to predict the result of the US presidential election. This survey's methodology employs stratify sampling, effective respondent recruitment strategy and data validation techniques to ensure the accuracy of prediction.

B.2 Sampling Approach

This survey employs Stratified Random Sampling, which involves divide the target population(In this case, the people who are eligible to vote in US) into subgroups based on their demographics characteristic (Stantcheva, n.d.). To simulate the national electorate, stratify the sample according to age, gender, race/ethnicity, education level, urban/rural domicile, and region. Also refer to the US Census data to determine the proportion of each strata, for example: if the census data shows there are 20% US citizen are between the age of 20 to 30, then there should be 20% of respondent in this age range. We aim for a starting sample of five thousand people since there were budget constraint.

B.3 Respondent Recruitment

Partner with an online panel provider to find respondents that match each stratum's specifications (Stantcheva, n.d.). Specify quotas to ensure that, for instance, you wish to have 18% of respondents from the Midwest or 25% aging between 30 - 40. Then randomly select respondent within each strata. To promote greater response rates and retention, provide a small monetary reward, such as $5 for each completed survey. To improve inclusivity and take into consideration differences in internet availability, use both phone and online surveys.

B.4 Data Validation

To keep distinct replies between responses, make sure respondents don't participate more than once each wave, and highlight straight-lining, inconsistent replies, and other low-effort answers for data quality purposes (Horn et al. 1997).

B.5 Budget Expenditure

- $50k - $60k for panel provider
- $25k for incentives
- $10k for Telephone survey services
- $5k for data analysis

B.6 Survey Structure

B.6.1 Introduction:

First of all, thank you for taking the time to participate in this survey.

This survey aims to precisely capture voter intention and mood across important demographic and geographic divisions in order to predict the result of the US presidential election.

Please note:

- Your responses will remain confidential.
- Please answer the question honestly.
- Complete the survey will receive a reward for $5 - $10.

If you have concerns or questions, reach out to haobo.ren@mail.utoronto.ca (Haobo Ren)

B.6.2 Screener Section: Screening out people who are eligible to vote in each state.

- Are you currently registered to vote?
  - Yes
  - No
- Which state are you currently registered to vote in?
  - (Drop down box)

B.6.3 Presidential Ballot Section: Ask participants whether they would vote, who they would vote for, and how confident they are about that.

- Do you plan to vote in the upcoming Presidential election?
  - Yes, I plan to vote in person on election day
  - Yes, I plan to vote in person early
  - Yes, I plan to vote by mail
  - No, I do not plan to vote
- If the election were held today, who would you vote for?
  - Democrat Kamala Harris
  - Republican Donald Trump
  - A third party / Independent candidate
  - Unsure
- How certain are you about your choice
  - Scale 1 to 5

B.6.4 Demographics Section: This section collects detailed demographic characteristics of the participants.

- Your gender identification
  - Female
  - Male
  - Other
- Age Range
  - 18 to 24
  - 25 to 29
  - 30 to 39
  - 40 to 49
  - 50 to 59
  - 60 or older
- Which of the following best describes your race or ethnicity
  - White or Caucasian
  - Black or African American
  - Hispanic
  - Asian or Pacific Islander
  - Native American
  - Other

- How would you describe your educational level
  - High School or less
  - College Graduate
  - Post-graduate degree
  - Unsure
- Employment Status
  - Full time
  - Part time
  - Unemployed
- What's your estimated household income?
  - Under $30,000
  - $30k - $49,999
  - $50k - $99,999
  - $100k - $199,999
  - $200k - $249,999
  - $250k or more
  - Unsure
- Do you consider yourself to be lesbian, gay, bisexual, transgender, queer or questioning?
  - Yes
  - No
  - Unsure
  - Prefer not to answer
- Would you best describe the neighborhood or area you live in as
  - Urban
  - Suburban
  - Rural
  - Unsure
- What is your religious background
  - Religious
  - Spiritual
  - Not religious or spiritual

Link to the survey: https://forms.gle/spJjASENRQ6nKK1E6

C Appendix 3: Model diagnostics

Figure 2 and Figure 3 are the model diagnostics for the two models. We can see that the residue only have limited deviation from normality and did not display an obvious pattern, which justifies the assumption of linearity and independence.

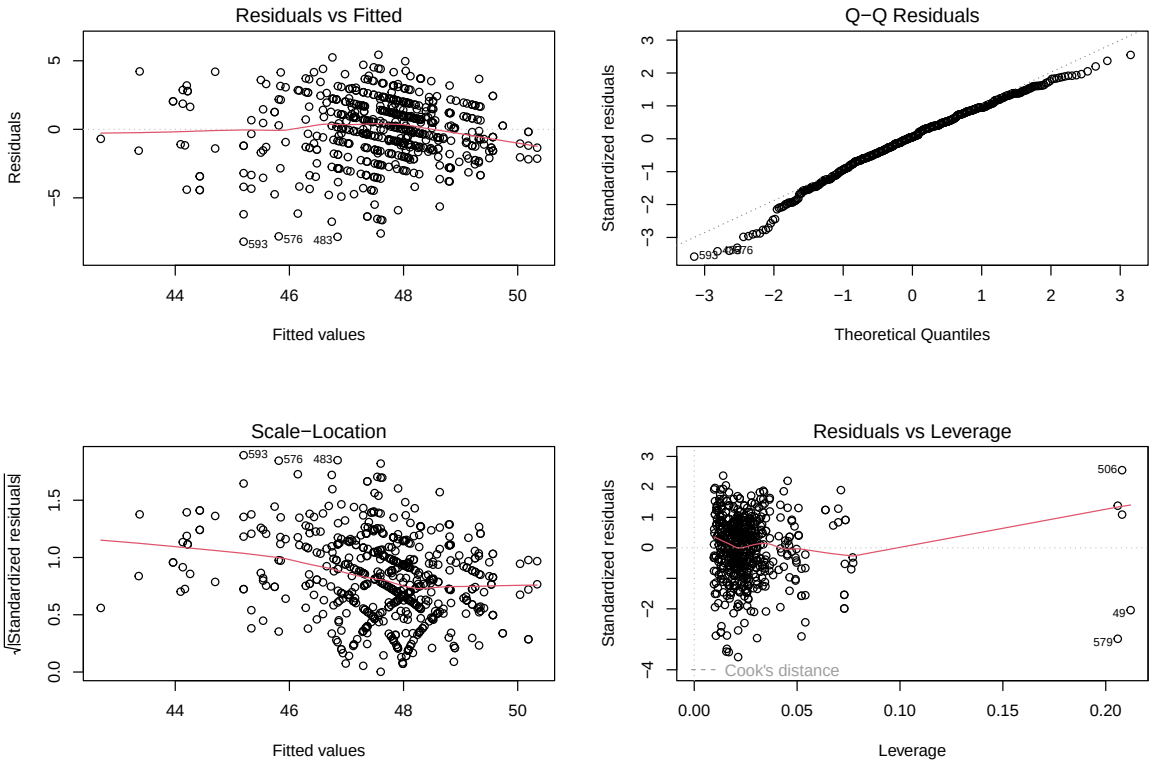

Figure 2: Model diagnostics for Harris model

C.1 Selection

Our model selection is based on AIC(Akaike information criterion) and interpretation of data.

The best model in terms of AIC in selecting was "DJIA_lag_7 + DJIA_lag_14 + DJIA_lag_90 + DJIA_lag_120 + DJIA_lag_150 + DJIA_lag_180 + state" for Harris and "DJIA_lag_14 + DJIA_lag_28 + DJIA_lag_60 + DJIA_lag_90 + DJIA_lag_150 + DJIA_lag_180 + state" For Trump. However, we want the two model to have the same set

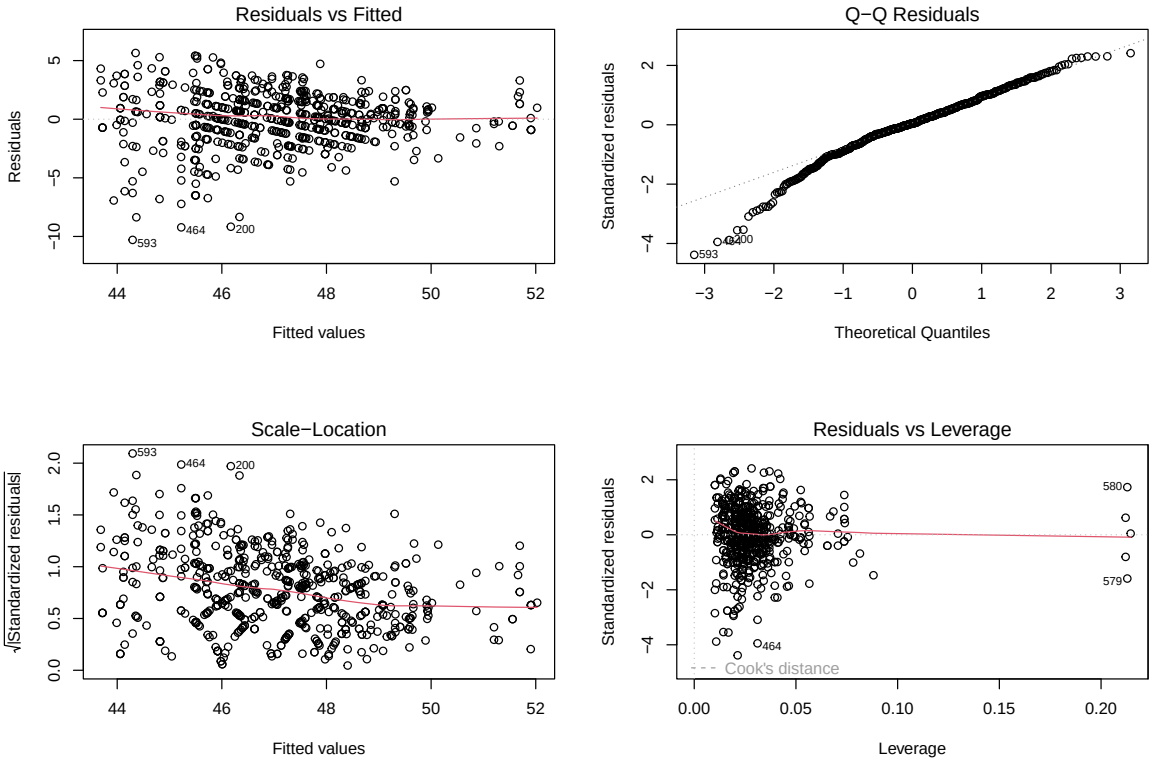

Figure 3: Model diagnostics for Trump model

of predictors, which may increase interpretability of the model. So we have selected a model that include relevant terms for both candidates.

References

ABC News. 2024. "Trump Leads in Swing State Polls, Tied with Biden Nationally." https://abcnews.go.com/538/trump-leads-swing-state-polls-tied-biden-nationally/story?id=109506070.

Aldrich, J. H., J. L. Carson, B. T. Gomez, and J. L. Merolla. 2023. *Change and Continuity in the 2020 and 2022 Elections*. Rowman & Littlefield.

Arendas, P., V. Malacka, and M. Schwarzova. 2018. "A Closer Look at the Halloween Effect: The Case of the Dow Jones Industrial Average." *International Journal of Financial Studies* 6 (2): 1–12. https://doi.org/10.3390/ijfs6020042.

BBC News. 2024. "Election 2024: Swing State Polls Near Tie in Blue Wall as Trump Commands Arizona and Harris Leads Nevada." https://www.bbc.com/news/articles/c511pyn3xw3o.

Consult, Morning. 2024a. "2024 Election State-Level Polls." https://pro.morningconsult.com/trackers/2024-election-state-polls.

———. 2024b. "Imagine Having Foresight to Back Every Decision. Welcome to Decision Intelligence." https://morningconsult.com/?gclid=CjwKCAiAgbiQBh.

Dorn, Sara. 2024. "Election 2024: Swing State Polls Near Tie in Blue Wall as Trump Commands Arizona and Harris Leads Nevada." https://www.forbes.com/sites/saradorn/2024/11/03/election-2024-swing-state-polls-near-tie-in-blue-wall-as-trump-commands-arizona-and-harris-leads-nevada-updated/.

Federal Election Commission. 2021. "Federal Elections 2020: Election Results for the u.s. President, the u.s. Senate, and the u.s. House of Representatives." https://www.fec.gov/resources/cms-content/documents/federalelections2020.pdf.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

FiveThirtyEight. 2024a. "2024 Swing the Election." https://projects.fivethirtyeight.com/2024-swing-the-election/.

———. 2024b. "Dataset: US Presidential General Election Polls." https://projects.fivethirtyeight.com/polls/data/president_polls.csv.

Horn, Werner, Silvia Miksch, Gerhilde Egghart, Christian Popow, and Franz Paky. 1997. "Effective Data Validation of High-Frequency Data: Time-Point-, Time-Interval-, and Trend-Based Methods." *Computers in Biology and Medicine* 27 (5): 389–409.

Milne, R. D. 1966. "The Dow-Jones Industrial Average Re-Examined." *Financial Analysts Journal* 22 (6): 83–88. https://doi.org/10.2469/faj.v22.n6.83.

Ooms, Jeroen. 2024. *Curl: A Modern and Flexible Web Client for r*. https://CRAN.R-project.org/package=curl.

Pasek, Josh. 2015. "Predicting Elections:considering Tools to Pool the Polls." *Public Opinion Quarterly* 79 (2): 594–619. https://doi.org/10.1093/poq/nfu060.

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

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan

Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

S &P Dow Jones Indices LLC. 2024. "Dow Jones Industrial Average [DJIA]." https://fred.stlouisfed.org/series/DJIA.

Schultz, David A., and Stacey Hunter Hecht. 2017. *Presidential Swing States: Why Only Ten Matter*. Lexington Books.

S&P Dow Jones Indices. 2024. "Dow Jones Industrial Average (DJIA) Index Overview." https://www.spglobal.com/spdji/en/indices/equity/dow-jones-industrial-average/#overview.

Stantcheva, Stefanie. 2023. "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible." *Annual Review of Economics* 15 (1): 205–34. https://doi.org/10.1146/annurev-economics-091622-010157.

———. n.d. "Online Appendix for 'How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible'."

U.S. Bureau of Labor Statistics. n.d. "The Employment Situation - [11, 2024]." https://www.bls.gov/news.release/empsit.nr0.htm.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.