Forecasting the United States 2024 Presidential Election based on Polling Data*

Model Predicts Donald Trump as Winner

Parth Samant

November 3, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

[TO DO : MENTION THAT THE PRESIDENT IS DECLARED BASED ON HAVING 270 ELECTORAL VOTES. THERE ARE ALSO 538 TOTAL VOTES.] With the 2024 U.S. Presidential Election approaching, there becomes increasing attention and speculation as to who may win. The two candidates, Kamala Harris and Donald Trump, represent opposing ideologies and visions of how a country should be run. Thus, the winner of this election could play a large factor in how the United States (and the world) operates for the next 4 or more years. This paper aims to use polling data, by state, to forecast the winner of the electoral college – and thus the presidency.

Overview paragraph

[TO DO]

Estimand paragraph

The estimand (i.e. what we are estimating), is whether Trump or Harris will win more electoral seats (and thus be the U.S. President). For each state, the support of harris and trump is estimated and the state's electoral votes will go to whoever has more support based on polling.

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2....

*Code and data are available at: https://github.com/samantparth/USA-2024-General-Election-Prediction.

2 Data

2.1 Overview

The original dataset used for this paper was sourced from the website FiveThirtyEight (FiveThirtyEight 2024). The dataset is a compilation of political polls by a variety of pollsters that is put in a standardized format.

The dataset organizes the polls in a way such that each row corresponds to a poll that measures a specific candidate's support. It also includes variables related to the reliability/accuracy of pollster's as determined by FiveThirtyEight. It also summarizes key characteristics of each political poll, such as the state polled, types of voters sampled, and the polls methodology.

The statistical programming language R (R Core Team 2023), the R package tidyverse (Wickham et al. 2019) were used to perform data cleaning and analysis on this original dataset. Data cleaning was performed by filtering for pollsters with high accuracy/reliability (numeric_grade >= 2.8) and for polls that began no earlier than 21 July 2024 (when Joe Biden Dropped out). Polls that did not deal with Trump nor Harris's support were also filtered out, as this papers aim involves comparing the support of Harris and Trump.

Data cleaning involved separating the dataset into two smaller subsets, based on if the polls are related to either Trump or Harris's support. A new variable for both datasets was also created that calculates the number of respondents who supported each candidate. This was calculated by multiplying the percentage of support by each polls sample size. I will refer to this as **Harris/Trump Analysis Data**, or simply **Analysis Data** if both datasets are being talked about.

The dataset (found in `data/analysis_data/state_prediction_analysis_data.parquet`) predicts the winner of each state based on estimated support of both Trump and Harris. This is found under `scripts/05-model_data.R`. I will refer to this as **State Prediction Data**

2.2 Measurement

Measurement of this dataset first begins with understanding that there is some "true" overall support of both Harris and Trump for each state. The 538 dataset consists of a compilation of pollster's who attempted to estimate overall support for a candidate.

The technique done to achieve this statistic largely depends on the specific poll. However, there are some similarities as pollsters often begin with a technique called 'stratified sampling'. This is where pollsters divide a population based on characteristics such as age, state, and their voter status. Then, groups within these characteristics are randomly sampled from, with the aim to provide a better representation of the voting population. Participants are often, though not always, polled by phone (as indicated by the large amount of "Live Phone" polls in the data set).

Another popular alternative of estimating a candidate's support is through an online panel, where participants are asked on their support through an online polling website.

The results from these polls serve as a tool for estimating overall support of a political candidate. However, individual polls still may have error contained within them. Many models, such as the one in this paper (in **?@sec-model**), then aggregate these polling results with an aim to estimate a candidate's support even more accurately.

2.3 Outcome/Predictor Variables

The decision to divide this section between outcome variables for Analysis Data and state prediction data is based on the idea that predictor variables are meant to predict/estimate a given phenomenon (outcome variable). However, the presence of both a state prediction dataset and an analysis dataset make it more complicated.

This is because the state prediction dataset is an extrapolation of the model used for the analysis data. It is essentially summarizing the predicted support for candidates in each state to make it easier to predict an electoral college winner. More specifically, it predicts **num_harris** for a sample size of 100 in a given state.

Thus, the state prediction dataset also uses the same predictor variables as the analysis data to make an estimate of who will win which state.

2.3.1 Analysis Data Outcome Variables

For Harris(and Trump) Data:

num_harris(or num_trump) : the amount of people that support Harris (Trump) for a given poll, based on pct × sample_size.

pct : the percentage of people who support Harris(Trump) for a given poll. This is not technically an outcome variable as it is not used in the model, but num_harris(or num_trump) is used in a way that mimics pct. However, it can help for exploratory data analysis and understanding overall support of each candidate - as is shown in **?@sec-exploratory**.

2.3.1.1 more on num_harris,num_trump, and pct {#sec-exploratory}:

Based on Figure 1 and Figure 2, it is evident that Kamala Harris has a slightly higher mean support (47.5% vs 46.7% respectively). However, as the president is not selected based on overall support (but rather electoral college), analysing the state prediction dataset may be more useful for a more accurate estimand. This analysis is shown in the next section, Section 2.3.2.

Table 1: Summary Statistics of Harris Outcome Variables

pct_harris_mean	pct___harris_sd	num_harris_mean	num_harris_sd
47.54053	3.730503	711.225	2078.148

Figure 1: Harris Outcome Variables Summary Statistics

Table 2: Summary Statistics of Trump Outcome Variables

pct_trump_mean	pct___trump_sd	num_trump_mean	num_trump_sd
46.78821	3.94343	687.2139	1875.914

Figure 2: Trump Outcome Variables Summary Statistics

2.3.2 State Prediction Data Outcome Variables

electoral_votes : the number of electoral votes that each state has.

harris_pred : the predicted percent support of Harris in a given state based on the model.

trump_pred : the predicted percent support of Trump in a given state based on the model.

win : a binary variable where 1 represents Harris winning a state and 0 represents Trump winning a state. This is based on whether the predicted percent support for Harris(**harris_pred**) is higher than the predicted support for Trump(**trump_pred**).

electoral_votes_harris: the number of electoral votes Kamala Harris will receive assuming she has won a given state. Otherwise, this is 0. For each state, this is win × electoral_votes.

The decision to not mention much about these variables in this section is because a summary of these variables is better provided in Section 4, as summarizing these variables also means revealing the prediction results of this model.

2.4 Predictor Variables

pollster : the organization that was behind the poll.

methodology : the method used to conduct the poll

state : the U.S. state in which the poll was conducted in (or focused on)

end_date_num : a continuous variable from 0 to 1, corresponding to how recent the poll was. Values close to 0 indicate polls conducted closer to July 21st, and values closer to 1 are polls closer to October 28th. This is simply a transformation done on the date the poll was finished.

Since these are mostly categorical variables, I believe that graphing these predictor variables is not very useful in understanding them. However, I believe that a table is more useful as it can show which values each predictor can take on. Thus,

Table 3: Unique Pollsters

Table 3: Pollsters in Dataset

x
Data Orbital
Emerson
Ipsos
Siena/NYT
SurveyUSA
CNN/SSRS
YouGov
CES / YouGov
Suffolk
Marist
Monmouth
Quinnipiac
SurveyUSA/High Point University
U. North Florida
Beacon/Shaw
University of Massachusetts Lowell/YouGov
Marquette Law School
The Washington Post
Christopher Newport U.
YouGov/Center for Working Class Politics
McCourtney Institute/YouGov
Muhlenberg
MassINC Polling Group
Selzer
YouGov Blue

Table 3 suggests that there are quite a few pollsters that are present in the cleaned dataset.

Table 4: Unique Methodologies

Table 4: Methodologies Used in Dataset

x
Live Phone/Text-to-Web
IVR/Online Panel
Probability Panel
Live Phone
Online Panel
Online Panel/Text-to-Web
IVR/Online Panel/Text-to-Web
Live Phone/Online Panel/Text-to-Web
Live Phone/Text-to-Web/Email/Mail-to-Web/Mail-to-Phone
Live Phone/Online Panel/Text
Live Phone/Email

Unlike Table 3, Table 4 only has a handful of overall methodologies. Thus, most high quality pollsters stick to a handful of polling methods.

Table 5: Unique Methodologies

Table 5: Individiual States Polled

x
Arizona
Michigan
National
Texas
Nebraska
Nebraska CD-2
North Carolina
Nevada
Montana
Florida
Georgia
Minnesota
New Hampshire
Ohio
Pennsylvania
Virginia

x
Wisconsin
South Dakota
Maryland
California
Washington
Massachusetts
New York
New Mexico
Connecticut
Rhode Island
Missouri
Indiana
Iowa

Although it is arguably the most significant predictor variable (as it may have the biggest effect on Harris and Trumps popularity), Table 5 does not have all the 50 states present. Thus, this presents challenges with how the electoral college will be predicted under this framework. However, my solution to this is found in Section 4.

For the final predictor variable used(**end_date_num**), plotting this variable alone may not be very useful, as it is relative to the dataset. However, plotting percent support based on the date may offer more insight on potential trends in overall support. This is shown in Figure 3.

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

Based on Figure 3, it is hard to determine a trend based on the data points alone. However, the trends, shown from the line, indicate that the overall difference in support between Trump and Harris seem to be narrowing. Thus, having an end_date_num closer to one may be beneficial for Trump's estimated support in the model.

3 Model

The strategy behind this model is based on first modelling the percent support of both Harris and Trump given a specific state. After this is done, **I want to factor in the complications that come from the electoral college**. It is possible to estimate the winner of the election based on national polling data and not pay much attention to state, and it is in fact a lot easier. However, for better or for worse, the winner of the presidential nomination is not based

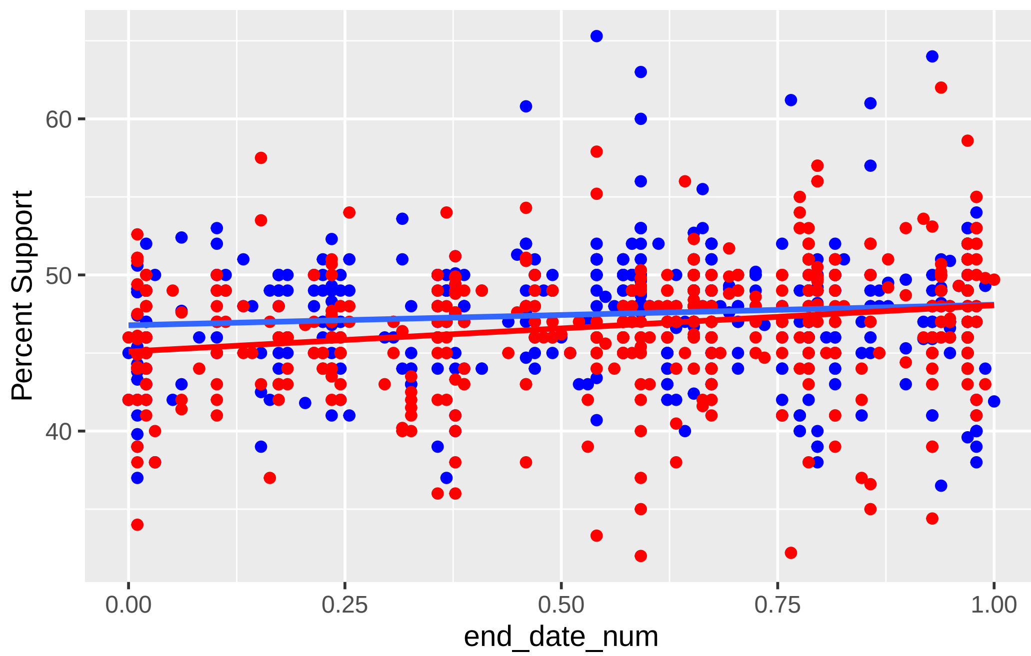

Figure 3: Trump (Red) and Harris's (Blue) Percent Support Based On end_date_num Variable

on who has more overall support in the country - it is, instead, based on whether one wins the Electoral College.

Background details and diagnostics of this model are included in Appendix D.

3.1 Model set-up

The generalized linear model used in this analysis follows a logistic regression model which helps us predict a discrete outcome. In this case, we are trying to predict the 'true' support of each candidate. We use the following model:

$$\text{logit (p} \mid \text{state)} = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot \text{I(state)} + \beta_3 \cdot \text{I(pollster)} + \beta_4 \cdot \text{I(methodology)} \quad (1)$$

Where: - p represents the probability of the candidate (either Trump or Harris) winning the poll - β_0 represents the intercept. - β_1 represents the effect of end_date_num on the candidate's support - x_1 represents the end_date_num variable. - β_2 represents the estimated change in support of a candidate given that the poll was conducted in a certain state. (where I(state) denotes an indicator variable that is equal to 1 when it is a specific state and 0, otherwise. -

β_3 represents the estimated change in support for the candidate given the poll was conducted by a certain agency. I(pollster) is an indicator variable. - β_4 represents the estimated change in support when a specific methodology is used. I(methodology) is an indicator variable.

Additionally, since this is a Bayesian model, we use the prior $\beta \sim$ Normal $(0, 2.5)$ for each coefficient in the model.

This model is then used to predict support for a given candidate in a specific state. The model is then ran in R (R Core Team 2023) using the `stan_glmer` function of rstanarm package ((**stanarm?**)). We use the prior mentioned above, as well as the model, to generate a posterior. More on the posterior distribution is found in (Section D.1). ## Model justification

This model can be justified because of the way that the electoral college works as well as how Bayesian binomial logistic regerssion is parameterized. Putting categorical variables as coefficients (instead of predictor variables) helps run the bayesian model code.

'end_date_num' was also considered to be an important variable in the model as support of both candidates can be heavily dependent on the date in which the poll has been conducted. end_date_num also provides for a more accurate estimation for the support of both candidates closer to the election date (November 5th).

The justification for the priors following $Normal(0, 2.5)$ is since we are not sure about how a certain variable may negatively or positively affect support for each candidate. However, since we don't want to assume that each prior has the same amount of effect, a variance is added. A normal distribution was also used as we want a relatively non-informative prior, as the true effect of each of the variables is unknown.

3.2 Model Assumptions

- **Accurate Polling Data**: since the model is based on the polling data, incorrect polling data will result in an inaccurate model.
- **Linear relationship** : There should exist a linear relationship between the predictor variables and the logit transformation outcome variable.
- **Independent Observations** : The model assumes that each polling observation is independent from one another
- **No Multicollinearity** : The model assumes that the predictor variables are not too highly correlated with one another.

3.3 Model Limitations

[LIMITATIONS ABOUT PRIORS, AND HOW IT DOESN'T HAVE EVERY STATE]

Since this model requires accurate polling data, it is very much possible that there may be a bias in the polling data that was unaccounted for. For example, nonresponse bias may not be properly accounted for, where certain supporters of a candidate may not be willing to give their opinion. Biased polling data could mean an inaccurate model.

This model also has a limitation in predicting many states, as many states do not have polls on Harris or Trump's popularity for the dataset when filtering out polls. However, these are all states that tend to almost always vote one way (for example, there are no polls specifically involving Mississippi, even though Mississippi will likely give their electoral votes to Trump).

Another limitation (does not affect the predicted final results), is that a few states do not cast all their electoral votes based on who wins. Maine and Nebraska do not have a winner-take-all system. However, this tends to balance out as Maine tends to give 1 electoral vote to Republicans, while Nebraska gives 1 electoral vote to Democrats.

3.4 Model Validation

Additional diagnostics can be found in **?@sec-diagnostics**.

To first validate the model, the function `modelsummary()` from the modelsummary package is used (Arel-Bundock (2022)). This summarizes the models and is shown in Table 6 for both Trump and Harris.

```
Warning:
`modelsummary` uses the `performance` package to extract goodness-of-fit
statistics from models of this class. You can specify the statistics you wish
to compute by supplying a `metrics` argument to `modelsummary`, which will then
push it forward to `performance`. Acceptable values are: "all", "common",
"none", or a character vector of metrics names. For example: `modelsummary(mod,
metrics = c("RMSE", "R2")` Note that some metrics are computationally
expensive. See `?performance::performance` for details.
 This warning appears once per session.
```

Where Sigma[(factor) × (Intercept), (Intercept)] denotes how much support varies depending on that variable.

A notable aspect of Table 6 is how much variability in support there is depending on the state. Additionally, the support for Trump seems to increase, on average, as the date gets closer to October 28th 2024 (as indicated by end_date_num).

Another notable statistic from Table 6 is the **RMSE** (root mean square error), which measures the average difference between the model's predicted values and the actual observed values of the dataset. In other words, this is the standard deviation for the residuals (i.e., the error between the predicted and observed values). An RMSE of 0.02 suggests that the predicted percent support of both Harris and Trump is off from the actual percent support by 0.02,

Table 6: Support of Models that predict Harris and Trump's support

	Harris Support Model	Trump Support Model
(Intercept)	−0.094	−0.226
end_date_num	0.074	0.100
Sigma[state × (Intercept),(Intercept)]	0.070	0.080
Sigma[pollster × (Intercept),(Intercept)]	0.002	0.002
Sigma[methodology × (Intercept),(Intercept)]	0.001	0.001
Num.Obs.	511	519
ICC	1.0	1.0
Log.Lik.	−2602.452	−2574.702
ELPD	−2668.3	−2641.4
ELPD s.e.	84.9	80.6
LOOIC	5336.6	5282.8
LOOIC s.e.	169.7	161.2
WAIC	5330.2	5274.3
RMSE	0.02	0.02

which is relatively close. The next plot, Figure 4, is a visual that can help understand this small RMSE.

To further validate the model, the function `fitted(model_har)` was run, which predicts the percent support for the actual Harris data. We can see, roughly, how accurate this model is by plotting this. The plot is shown in Figure 4. The same was done for the data on Trump's support predictions, which is shown in **?@sec-diagnostics**.

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

4 Results

The results are separated in three sections: results from the model in Section 4.1, results from "missing" states in Section 4.2, and a final prediction (Section 4.3). This is necessary as the Section 4.1 include predictions for states that have polling on Harris and Trump, allowing the model to accurately predict a winner. Section 4.1 involves predicting states not included in the model (which are called "missing" states for simplicity), and Section 4.3 provides the overall final prediction.

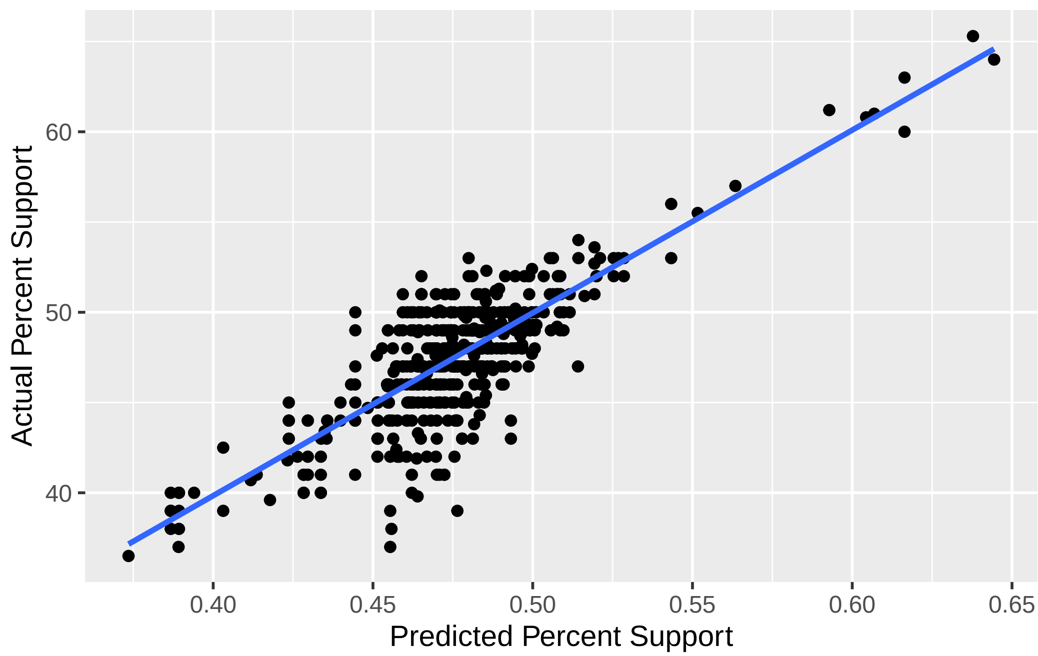

Figure 4: Predicted vs Actual Percent Support for Harris

As mentioned in the introduction, states need a total

4.1 Results from Model

After running this model regarding predicted support, we can make predictions of support based on the state and put the subsequent votes gained by Harris (as mentioned in Section 2.3.2). The results of each state with available polling data is shown in Table 7.

```
[1] "Harris Receives 208 out of 377 modelled electoral votes"
```

Table 7: Models Prediction of State Winner

Table 7: Models Prediction of State Winner

state	harris_pred	trump_pred	win	electoral_votes_harris
Montana	40.4465	56.5010	0	0
New Hampshire	50.6055	44.9985	1	4
Pennsylvania	49.5140	47.7100	1	19
North Carolina	48.5255	48.5350	0	0

state	harris_pred	trump_pred	win	electoral_votes_harris
Wisconsin	49.6465	47.9830	1	10
South Dakota	36.6715	59.3500	0	0
Georgia	47.4220	49.5635	0	0
Arizona	47.4320	49.7030	0	0
Maryland	63.3475	33.3095	1	10
Texas	44.9220	51.3630	0	0
Michigan	49.0605	47.2800	1	15
Florida	44.7590	52.2345	0	0
California	60.1315	36.7890	1	54
Washington	57.3995	36.7405	1	12
Nevada	49.0755	47.7495	1	6
Ohio	44.4430	51.6545	0	0
Massachusetts	60.7145	35.0715	1	11
Virginia	51.1840	44.5090	1	13
Minnesota	50.9195	44.9920	1	10
New York	54.4935	40.9915	1	28
Nebraska	40.4205	54.5365	0	0
New Mexico	52.3780	43.2375	1	5
Connecticut	51.7520	39.0445	1	7
Rhode Island	53.4735	43.2835	1	4
Missouri	42.9460	55.0925	0	0
Indiana	40.7335	56.6905	0	0
Iowa	44.9570	48.0285	0	0

From this data, we can see how large the difference in the predicted support is between the two candidates, We can also see how many electoral seats Kamala Harris gains in the electoral college based on these results. **She also wins 208 out of 377 electoral votes by states that were modelled (270 required to win President). Thus, Trump wins 169 of these votes.**

However, not only are there some missing states, but it also does not provide an overview of predictions at a glimpse. These two issues are fixed in Section 4.2; which predicts the remainder of the states, and Section 4.3; which maps the subsequent results.

4.2 Results from "Missing States"

The presence of states that have not been polled (individually) complicates the prediction of a winner based on electoral college. Fortunately, none of these states are swing states and all tend to be strong Republican or Democrat states, making it easier to predict the winner.

The missing states include Alabama, Alaska, Arkansas, Colorado, Delaware, Hawaii, Idaho, Illinois, Kansas, Kentucky, Louisiana, Maine, Mississippi, New Jersey, North Dakota, Oklahoma, Oregon, South Carolina, Tennessee, Utah, Vermont, West Virginia, and Wyoming.

Using general American political knowledge as well as FiveThirtyEight (FiveThirtyEight (2024)), we predict the winner (where win = 1 if Harris wins the state) of the missing states in Table 8.

```
[1] "Harris Receives 68 out of 161 'missing' electoral votes"
```

Table 8: Prediction of Winner For States Not Polled (win=1 Indicates Harris Win)

Table 8: Prediction of Winner For States Not Polled

state	win	electoral_votes_harris
Alabama	0	0
Alaska	0	0
Arkansas	0	0
Colorado	1	10
Delaware	1	3
Hawaii	1	4
Idaho	0	0
Illinois	1	19
Kansas	0	0
Kentucky	0	0
Louisiana	0	0
Maine	1	4
Mississippi	0	0
New Jersey	1	14
North Dakota	0	0
Oklahoma	0	0
Oregon	1	8
South Carolina	0	0
Tennessee	0	0
Utah	0	0
Vermont	1	3
West Virginia	0	0
Wyoming	0	0
District of Columbia	1	3

A notable aspect is that aside from a few exceptions (such as Illinois and New Jersey), most states without polling data on Harris and Trump are relatively smaller states.

Harris also wins 68 out of 161 electoral votes for states that were not individually polled, where Trump wins 93 of them.

4.3 Prediction of Winner and Visualization

Based on the results results from Section 4.1 and **?@sec-missingresults**, we can conclude that Harris wins 208 electoral votes from the model and 68 electoral votes from 'missing' states. Trump wins 169 electoral votes from the model and 93 from the 'missing' states.

Thus, Harris receives 276 electoral votes and Trump receives 262, making Harris the predicted winner of the Presidential Election based on this model.

The mapped results of this prediction are shown in Figure 5.

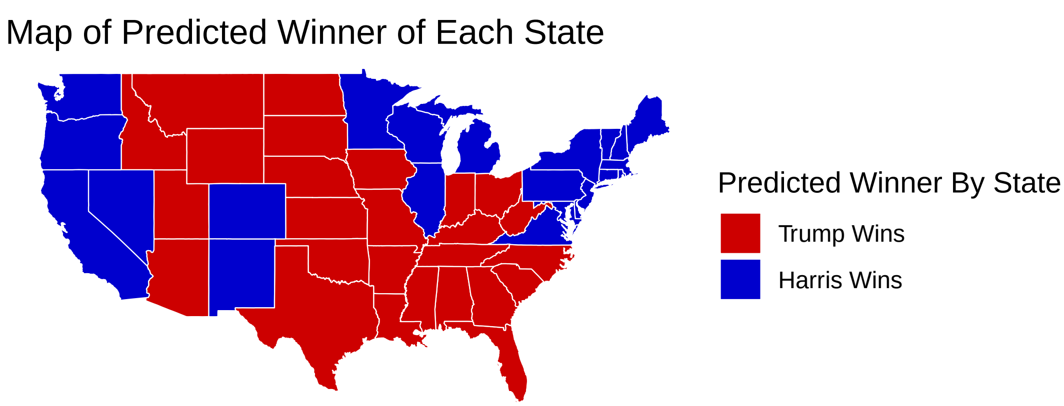

Figure 5: Map of Winner of Each State

A notable aspect of the winners by state, geography wise, is that states that are predicted to "go blue" (i.e., vote for Harris) tend to be concentrated in regions across the country.

5 Discussion

REMOVE : What is done in this paper? What is something that we learn about the world?

5.1 A Narrow Margin of Victory

REMOVE {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.}

As mentioned in @, this model predicts that Harris will win 276, while Trump will win 262. This is quite a small margin of victory, suggesting that there is likely a large amount of uncertainty associated with these predictions. This is further supported by many of these model predictions - North Carolina, in this model, was won by Trump by a difference of less than 0.01%. Although North Carolina does not tend to have margins this small (suggesting a potential error in the model), it highlights an interesting feature of the electoral college in practice.

History has shown that these fractions of percentage points can single-handedly determine entire elections. For example, the 2000 U.S. Presidential Election was won by Former President George W. Bush due to having a margin of 537 over his competitor Al Gore. This minuscule difference, especially considering the over 5 million Floridian votes, shows how tight elections can be especially in such a large country.

The tightness of this election is further supported by other polling aggregators and polling organisations such as Nate Silver's FiveThirtyEight (FiveThirtyEight (2024)) and the New York Times (**REMEMBER TO CITE POLLING DATA**)

The nature of how small margins can determine the winner and genuine uncertainty about voter outcome also suggests that the results of the model be taken with a 'grain of salt'.

5.2 Other Notable Aspects of Results

Another notable aspect of the results of this model is that states that tend to go blue (support Harris) are concentrated in certain regions of the country, whereas states that go red (support Trump) are all connected geographically and tend to be more in the centre of the country. This also tends to correlate with states that tend to have a larger urban population. For example, New York State and California are predicted to be in strong support for Harris, which are both states with a significant urban population. This is supported from polls that show that voters in urban areas tend to be Democrat much more frequently (https://www.pewresearch.org/social-trends/2018/05/22/urban-suburban-and-rural-residents-views-on-key-social-and-political-issues/).

5.3 Weaknesses and next steps

This data does not apply a popular feature of political pollsters - weighing of individuals based on their characteristics. This corresponds to a polling respondents answer being weighted more (or less) based on how much groups are over or underrepresented in the survey. For example, if a survey has less male respondents, then respondents that identify as male would have their answer be more influencial on the prediction. It thus aims to properly represent the voting population.

https://goodauthority.org/news/pollsters-are-weighting-surveys-differently-in-2024/

This is because of a phenomenon known as nonresponse bias, where some groups are more likely to respond to polls than others. This can correspond to biased data, where support for a candidate can be systemically higher or lower than the polls report. This may be the explanation of how this model predicted North Carolina won by Trump with a margin of 0.01%, where Pennsylvania is won by Harris with a margin of about 2%. This is in contrast to many other pollsters who predict North Carolina of voting more in favour of Trump than Pennsylvania (FiveThirtyEight (2024)). Thus, a next step for a future model could be including this weighing into percent support for both candidates.

Another weakness of this data is the presence of 'missing states'. As mentioned before, this is where high-quality polling data on both Harris and Trump is not present. Although I do believe this overall does not affect the election predictions, it does undermine how winning any said state is never guaranteed for either candidate.

Another possible weakness relating to the data is the potential violation assumptions that were required for logistic regression. In particular, I believe the assumption of a linear relationship has the highest chance of being violated – especially for the relationship between end_date_num and candidate support. It is quite likely that there is not a linear relationship between these variables, where the support for each candidate can highly fluctuate depending on the day/month in a way that is turbulent and not suggestive of an overall linear trend.

Another limitation of the models result is due to the nature of political polling data itself (which this model relies on for accuracy). It is true that polls tend to sample certain 'populations', including groups such as likely voters or registered voters. However, this does not change the fact that certain ages and incomes tend to have different voting patterns and thus turnout. Thus, certain demographics may affect the election more (or less) than expected despite all identifying as a likely or registered voter. (https://ourworldindata.org/grapher/voter-turnout-rate-by-age-usa)

Furthermore, many polls, including the *New York Times/Siena College* poll, tend to have an extremely low respondent rate of around 1 percent. In other words, around 1 percent of people contacted for the poll actually responded. This fact alone can make the data biased in tremendous (although unknowable) ways (https://www.scientificamerican.com/article/why-election-polling-has-become-less-reliable/).

Appendix

A Polling Methodology for YouGov

REMOVE : (The deep dive provides a thorough understanding of how something goes from being a person's opinion to part of a result for this pollster. Provide a thorough overview and evaluation of the pollster's methodology, and sampling approach, highlighting both its strengths and limitations.)

A.1 Overview of How Survey is Conducted

Surveys are conducted online and can be taken on either a phone, tablet, or computer, where respondents can respond anywhere and at a time of their choice. YouGov also operates their own panel (i.e., group of people chosen for the poll) and any new panel member is required to provide demographic information. YouGov also employs a form of longitudinal sampling to track changes of views and behaviours over time. Furthermore, a method of non-probability sampling is used where a panel is made to have a representative sample of the voter base (YouGov (2024)).

A.2 Recruitment/Sampling Approach

Any adult living in the United States is eligible to join the panel. However, as mentioned before, choosing a panel is highly strategic and is based on whether you would serve as a good representation of the voter base. To ensure that participants of different backgrounds can properly understand and respond to the survey, surveys are also offered in different languages such as Spanish (YouGov (2024)).

Additionally, panel members are recruited through many sources, including through advertising and partnerships with a diverse set of websites. Monetary incentives are provided upon finishing the survey ((**yougouv?**)).

A.2.1 Determining Who is Chosen for Panel

To determine who is chosen for the panels, YouGov starts with first deciding which group they are trying to estimate (it is usually the opinion of all U.S. Adults/Citizens). However, it is sometimes groups such as registered voters (YouGov (2024)).

Then, based on the group of interest, demographic research is done on the characteristics of this population. From these characteristics, the sample is chosen in a way that (as a whole) is representative of the population of interest. YouGov aims to recruit roughly 1-2,000 respondents.

A.3 Weighing of Responses

As mentioned in Section 5.3, YouGov also employs the popular technique of weighing based on their samples. To re-iterate, weighing give more or less weight to a respondent based on their characteristics (such as age or gender). The panel will thus have a certain percent of the sample with these characteristics. YouGov then compares the characteristics of the sample with the characteristics of the population of interest to adjust the weights of responses (YouGov (2024)).

They also use weighing to estimate the vote based on post-stratification weighing. This is where adjusting the weights of the responses is also used in the model for estimating the votes (YouGov (2024)).

A.4 Checks to Ensure Data Reliability

YouGov claims to ensure a many checks to ensure reliability. Not only are panelists required to verify their e-mail addresses, but another way they ensure response reliability is by looking at features of how the survey was filled out. This includes the time it took to complete the poll and how consistent the answers given were. Based on these characteristics, some responses will be removed from the final sample, and respondents who repeatedly do this are removed from the panel altogether (YouGov (2024)).

Furthermore, YouGov also measures response reliability by comparing the responses to highly predictable information about the panelists. They also check the locations of the respondents devices to detect misrepresentation based on location.

A.5 Strengths and Weaknesses of YouGov's Methodology

A.5.1 Strengths

A strength of this method of non-probability sampling is that it is likely a cheaper and faster approach to be representative of the entire population. Many underrepresented groups, such as minorities, may be harder to reach based on probability sampling. Probability sampling would likely need a higher sample size to be more representative.

I believe that another strength is the way they perform checks on the polling data itself. Detecting misrepresentation based on location data and poor response data is vital to ensuring that the data is less biased.

Furthermore, weighing allows YouGov to have a more representative sample, as some groups tend to not be as willing to participate in polls.

A.5.2 Weaknesses

A weakness of this method is that non-probabiilty sampling may not properly be representative of the voter base as a whole. Certain groups having a higher likelihood of being selected may make it more difficult to eliminate bias from the poll, as it puts more pressure on the weighing system to be more accurate.

Another weakness is that it may be difficult to predict the voting patterns of hundreds of millions of Americans based on a sample of just 1-2,000. Furthermore, as the U.S. relies on an electoral college system, whether Harris or Trump is more popular overall may not represent who will win the electoral college.

A final weakness of this method of polling is that it assumes that the voting patterns of Americans can be represented via the Americans that have an internet connection. In order words, an internet connection is required to complete the survey. Although technically true that most Americans have an internet connection, complications from an online survey may add to some misrepresentation. In practice, this may include elderly voters or certain communities having more difficulty with properly completing the survey.

B Idealised Methodology and Survey for U.S. General Election Polling

B.1 Overview

In this section of the appendix, we provide an idealised methodology for a poll that forecasts the U.S. General election. In other words, the 'ideal' way a poll would be conducted in a way that aims for maximal accuracy. This idealised survey employs many popular techniques used by various pollsters (such as post-stratification weighing) as well as including a diverse set of recruitment strategies to have a wider reach (and thus a more representative sample).

B.2 Sampling Approach

First, the sample will only filter for those who identify as likely voters, as the general population does not properly represent those who are more likely to vote (and thus affect the election). After this, A method of sampling called stratified sampling will be employed to control for characteristics such as gender, age, ethnic/racial background, education, state, and income (more information in Section B.2.1). Stratified sampling minimizes the phenomenon of certain groups responding more than other groups.

To properly represent the voting patterns based on these characteristics, census data and voter registration records will be analysed to determine which groups make up a given portion of the electorate.

Polling via telephone will be employed, where telephones would call at a random time when most are awake (around 8am-10pm). Calling at a certain time would be minimized, as it may unintentionally select for certain people who need to wake up (or stay awake) which may affect the data.

As mentioned by Section 5.3, many polls that employ telephone polling tend to have quite a low response rate- around 1%. Thus, Around 200,000 people will be contacted to ensure a relatively desirable sample size of 2,000.

B.2.1 Stratification of Characteristics

The Poll will divide the populace based on these characteristics:

Gender: Man, Woman, Non-binary/Other, Prefer not to Respond

Age: 18-30, 31-44, 45-64, 65+

Racial/Ethnic Background (if Multiracial, choose multiple): White, Black/African American, Asian, Pacific Islander, Native American, Other

With subsequent clarifications:

"Do you identify as Hispanic/Latino?": Yes/No)

Education: No high school, high school diploma, Bachelor's Degree or Equivalent, Associate's Degree, Advanced Degree (such as a Master's degree or PhD), Other

State: Respondent types which state they reside in, with invalid states being removed.

Income: <$35 000, $35 000-$60 000, $60 000-$100 000, >$100 000

Having the respondent fill out this demographic data helps when weighing the data to match the electorate.

B.3 Recruitment of Respondents

B.3.1 Diverse Recruitment Strategy

A diverse recruitment strategy will be employed to reach out to voters of many different backgrounds and characteristics. This includes:

Telephone Polls: This method of recruitment will be used to help especially those who are not very familiar with technology. The benefit of using telephone polls is that phone calls tend to be more intuitive than online polling. Telephone polls can also help us properly stratify based on location (from phone area codes). It can additionally help with harder-to-reach populations, such as those in rural areas who may not have as much access to the internet.

Advertising. In addition to telephone polls, advertisements on a broad range of websites will be used to recruit new members of the panel. Advertisements on many websites allow the poll to reach a more diverse group of voters. These ads could also be targeted based on website to reach voters with characteristics that may be harder to reach. For example, if Hispanic/Latino Americans become harder to reach, then advertising may be put on a website with a known Hispanic/Latino American audience.

Online Survey: In addition to telephone polls and advertisements, online surveys also may be used. An online panel of voters, in addition to telephone polls, can help us gain a more comprehensive view of candidate popularity.

B.3.2 Incentives

Incentives will be provided for those who decide to complete the survey. This is deemed a noteworthy method of recruiting by reputable pollsters such as the American Association for Public Opinion Research and YouGov (Public Opinion Research (n.d.);YouGov (2024)). These incentives may include:

-**Gift Cards**: Upon completion, respondents will receive a $10 gift card.

-**Sweepstakes**: If respondent so chooses, their name will be put in a pool of other names and 10 winners will receive $100 gift cards.

B.3.3 Weighing

We can first research which groups tend to have a higher amount of nonresponse bias based on available data to know where to initially have heavier advertising. This data will also be used to determine the initial weighing. Respondents will then answer questions in Section B.2.1, which helps us to determine which demographic groups tend to be underrepresented in our polls specifically.

Based on this data, the method of post-stratification weighing will be used to determine which groups in our polls are underrepresented. More advertising may be used to further target harder-to-reach groups. Additionally, gift cards of higher value may be provided to those who belong to those groups.

B.4 Data Validation

To ensure the validity of polls, various techniques will be used. Some of these techniques have also been used by YouGov (YouGov (2024)), including assessing time taken to complete poll and ensuring their IP address roughly matches their claimed location (which is accessible to websites). These techniques include:

- **Pilot Studies**: pilot studies will be employed on a smaller sample to ensure that the data does not have any clear flaws. This may include abnormal responses based on a sample's characteristics as well as the clearness of the questions in the poll. Respondents in these pilot survey's will also be asked if there is any room for improvement.
- **Data Cleaning**: after the respondents have submitted their answers, extensive data cleaning and analysis will be performed to remove any anomalies. This may include multiple responses from the same individual or answers that are formatted in a different way. This helps to properly analyse the responses. Furthermore, respondents that added a US state that does not exist will have their answer removed.

-**Assessing Time Taken**: Similar to a technique used by YouGov (YouGov (2024)), the poll will count how much time it takes for reach respondent to fill out the questionnaire. Responses that have taken an unusually short amount of time (<2 minutes) are assumed to not be genuine data and are removed from the dataset.

-**Ensuring Accurate Location**: Also similar to a YouGov (YouGov (2024)) technique, this poll will validate that the respondent was honest about their based on analysing their IP. Respondents whose claimed location is notably different than their IP address location (e.g. a different state) will be removed. Although it is possible that they may be travelling, we ask that respondents do not take the poll while outside of their state.

B.5 Survey Design

The survey will be structured as follows:

(options regarding characteristics can be found in Section B.2.1)

1. "Are you a U.S. citizen and intending to vote in this upcoming election?" (Answers that say no will be discarded)
2. "What gender do you identify with?"
3. "What is your age?"
4. "What is your race/ethnicity (if multi-racial, choose multiple)" 4a. "Do you identify as Hispanic/Latino?"
5. "What is your education level?"
6. "What state do you reside in?"
7. "What is your individual income?"
8. "Which candidate of the U.S. Presidential Election do you plan to vote for?"

- Kamala Harris
- Donald Trump

- Other (Please Specify)

This survey has been implemented through Google Forms through the hyperlink below:

Google Form of Poll

B.6 Budget Allocation

The \$100 000 given for this survey will be allocated as follows:

Incentives for Finishing Survey: \$21 000 - 2 000 × \$\10 (gift card) = \$20 000 - 10 × \$100 (sweepstakes) = \$1 000

Marketing and Advertising: \$39 000

Hiring of Statisticians and Other Staff: \$20 000 - Statisticians (help with data analysis and cleaning): \$10 000 - Other Staff (including staff that call phone numbers): \$10 000

Survey: \$10 000 - Pilot Study: \$5 000 - Maintaining Forms: \$5 000

Unanticipated Issues Funds: \$10 000

C Additional data details

D Model details

D.1 Posterior predictive check

References

Arel-Bundock. 2022. "Modelsummary: Data and Model Summaries in r." *Journal of Statistical Software*. https://doi.org/10.18637/jss.v103.i01.

FiveThirtyEight. 2024. *FiveThirtyEight: 2024 US Presidential Election Polls*. https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

Public Opinion Research, American Association for. n.d. "Best Practices for Survey Research." https://aapor.org/standards-and-ethics/best-practices/#1668112232459-8b1678f0-d862.

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

YouGov. 2024. "Methodology." https://today.yougov.com/about/panel-methodology/.