Forcasting US Election 2024*

##TODO: result here

Maryam Ansari Amy Jin Maggie Zhang

October 22, 2024

first sentence: specify the general area of the paper and encourage the reader; second sentence: specify the dataset and methods at a general level; third sentence: specify the headline result; and a fourth sentence about implications

1 Introduction

As of October-November 2024, the U.S. presidential election is almost at its final stage, with all candidates actively campaigning across numerous states. The presidential election will happen on November 5, 2024 ("2024, n.d.). The two major parties, Democrats and Republicans, are in particular focus. Former President Donald Trump represents the Republican party, and President Joe Biden initially led re-election for the Democrats. However, Biden has dropped out, and Kamala Harris has taken over as the Democratic nominee (Galva 2024), who is the first Black woman nominee. Harris is the Both candidates are focusing their efforts on securing votes by addressing their aspect on social issues. pollsters and election forecasting with data evidence are important in predicting election outcomes. Polling can highlight underlying or emerging issues, as well as reveal voter preferences. The polling outcomes allow campaigns of different parties to strategically target specific demographics. For example in 2012, pollsters and data modellers such as Nate Silver have all used survey research and statistical models to successfully predict the result of Barack Obama's victory in the presidential election with consistent and reliable forecasts (Blumenthal 2017).

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

*Code and data are available at: https://github.com/aj3616/Forcasting-US-Elections.

2 Polling Data

2.1 Overview

We use the polling data to forecast the potential outcomes of the 2024 U.S. presidential election between Kamala Harris and Donald Trump.The dataset was obtained from https://projects.fivethirtyeight.com/polls/president-general/2024/national/ or (FiveThirtyEight's "Poll of Polls" for the 2024 U.S. Presidential election (**fivethirtyeight?**)). It provides a comprehensive view of voter preferences through aggregated results from numerous national polls conducted by various polling organizations. The dataset was simulated, cleaned, analyzed, and tested using the R programming language (R Core Team 2023), tidyverse (Wickham et al. 2019), knitr (Xie 2014), ggplot2 (Wickham 2016), gt(Iannone et al. 2024).

2.2 Variables

The dataset comprises several key variables of interest, including poll_id, a unique identifier for each poll conducted; pollster, which indicates the organization conducting the poll; sample_size, representing the total number of respondents; population, specifying the voting group described (e.g., likely voters); candidate_name, the names of the candidates in the poll (e.g., Kamala Harris, Donald Trump); and pct, the percentage of the vote or support received by each candidate. These variables allow us to explore various dimensions of polling data, such as trends in voter support across different states and the influence of pollster reliability on polling outcomes. The Table 1 below provides a snapshot of the polling data, displaying the first ten entries. This includes the pollster names, sample sizes, and the percentage of support for each candidate, which can reveal patterns in public opinion and help identify how different organizations may report varying levels of support for Kamala Harris and Donald Trump.

Table 1: Sample sizes and support percentages for Kamala Harris and Donald Trump

Polling Data Snapshot: 2024 U.S. Presidential Election

Poll ID	Pollster	Sample Size	Population	Candidate	Support (%)
460	SurveyUSA	558	lv	Kamala Harris	40
460	SurveyUSA	558	lv	Donald Trump	56
940	Lake Research	600	lv	Kamala Harris	42
940	Lake Research	600	lv	Donald Trump	52
1347	Cygnal	400	lv	Kamala Harris	43
1347	Cygnal	400	lv	Donald Trump	53
1775	GQR	500	lv	Kamala Harris	50
1775	GQR	500	lv	Donald Trump	46

294	McLaughlin	600	lv	Donald Trump	41
294	McLaughlin	600	lv	Donald Trump	44

2.3 Summary statistics & Relationships

2.3.1 Pollster Reliability and Election Outcome

The relationship between polling organization reliability, as indicated by the pollscore, and the percentage of support (pct) for each candidate was examined. It is anticipated that more reliable pollsters will yield more accurate predictions. Consequently, polls were categorized into three tiers based on their pollscore: high reliability (pollscore greater than 0), medium reliability (pollscore between -1 and 0), and low reliability (pollscore less than -1). By comparing the average support percentages (pct) for Kamala Harris and Donald Trump across these tiers, this analysis sought to identify whether more reliable pollsters produce different outcomes than those deemed less reliable. Mean pct values were calculated for each category, and standard deviation was included to illustrate variability within the poll results. The tableTable 2 and bar chartFigure 1 showing the average `pct` for each candidate across different levels of pollster reliability. Discussion: This analysis highlights whether pollsters with higher reliability scores offer more accurate predictions and if their estimates favor one candidate over the other.

Table 2: Mean and Standard Deviation of Support Percentages (pct) for Kamala Harris and Donald Trump Across Pollster Reliability Levels

Polling Data Summary: Pollster Reliability and Support Percentages

Candidate	Pollster Reliability	Mean Support (%)	Standard Deviation
Donald Trump	High	43.75	4.77
Donald Trump	Low	44.82	4.94
Donald Trump	Medium	43.76	4.56
Kamala Harris	High	43.25	4.55
Kamala Harris	Low	46.69	3.53
Kamala Harris	Medium	47.42	4.67

2.3.2 Impact of Methodology on Poll Results

Polling methodology is a critical factor that influences the results of any survey. Polls were categorized according to their methodology (e.g., Online Panel, Phone Interview) to evaluate how different polling methods might affect the percentage of support (pct) for each candidate. For each polling methodology, the average pct for Kamala Harris and Donald Trump

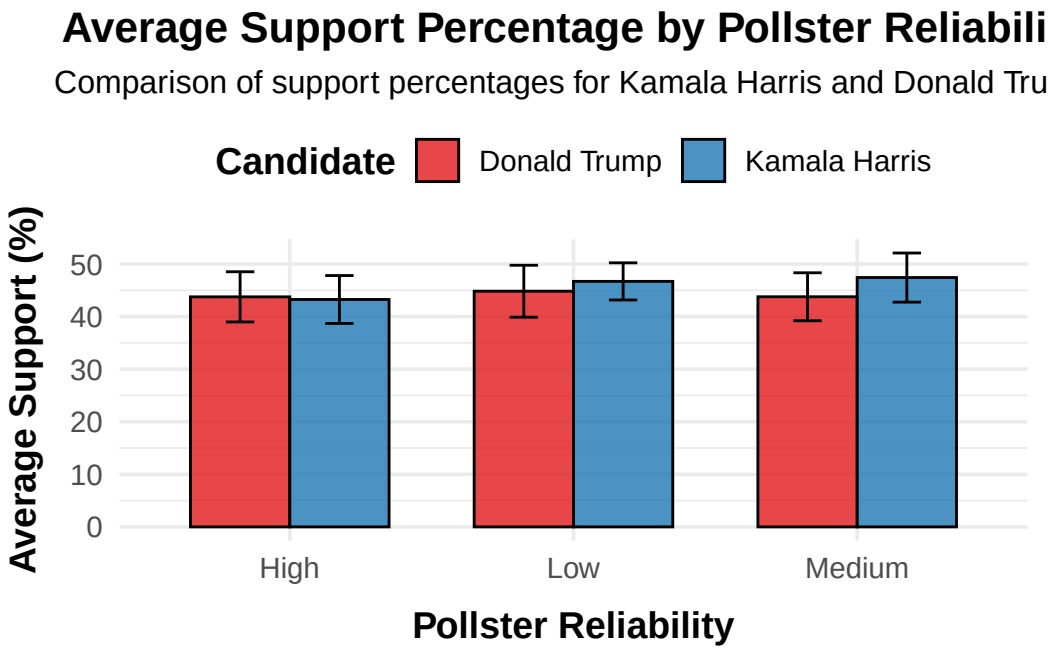

Figure 1: Graph of Dates Categorized by Year and Presented by Each Months

was calculated and compared. This analysis aimed to determine whether specific methodologies consistently resulted in higher or lower support for either candidate. Comparison of candidate vote percentages by polling methodology is demonstrated by Table 3 Discussion: We observed potential biases in support based on the polling method, providing insights into which methodologies might offer more reliable forecasts.

Table 3: Table of Methodologies comparing the Mean Support Percentage for Donald Trump and Kamala Harris with Difference (Kamala Harris - Donald Trump)

Polling Data Summary: Methodology and Support Percentages

Mean Support Percentages (pct) for Kamala Harris and Donald Trump by Polling Methodology Category

Polling Methodology	Trump(%)	Harris(%)	Difference
App Panel	46.58	49.89	3.30
Email	45.04	46.36	1.31
IVR	45.59	45.75	0.16
Live Phone	44.53	47.30	2.77
Mail-to-Phone	41.10	49.00	7.90
Mail-to-Web	42.27	49.00	6.73
Online Ad	46.00	46.99	0.99

Online Panel	43.98	45.59	1.61
Probability Panel	40.50	48.89	8.38
Text	45.41	46.12	0.71
Text-to-Web	45.30	48.09	2.79

2.3.3 State-by-State Polling Trends

Given the importance of state-level polling data in U.S. presidential elections, this analysis focused on key battleground states (e.g., Florida, Pennsylvania, Wisconsin). The polling results for Kamala Harris and Donald Trump were examined by state (state), and the average pct was calculated for each swing state. These trends were visualized to assess whether one candidate consistently received more support in these critical states, as such patterns could significantly influence the final election outcome. A state-by-state breakdown of the polling percentages for each candidate, with a focus on swing states is shown by Figure 2. Discussion: The results highlight key states where the election might be decided, helping to pinpoint potential trends favoring one candidate.

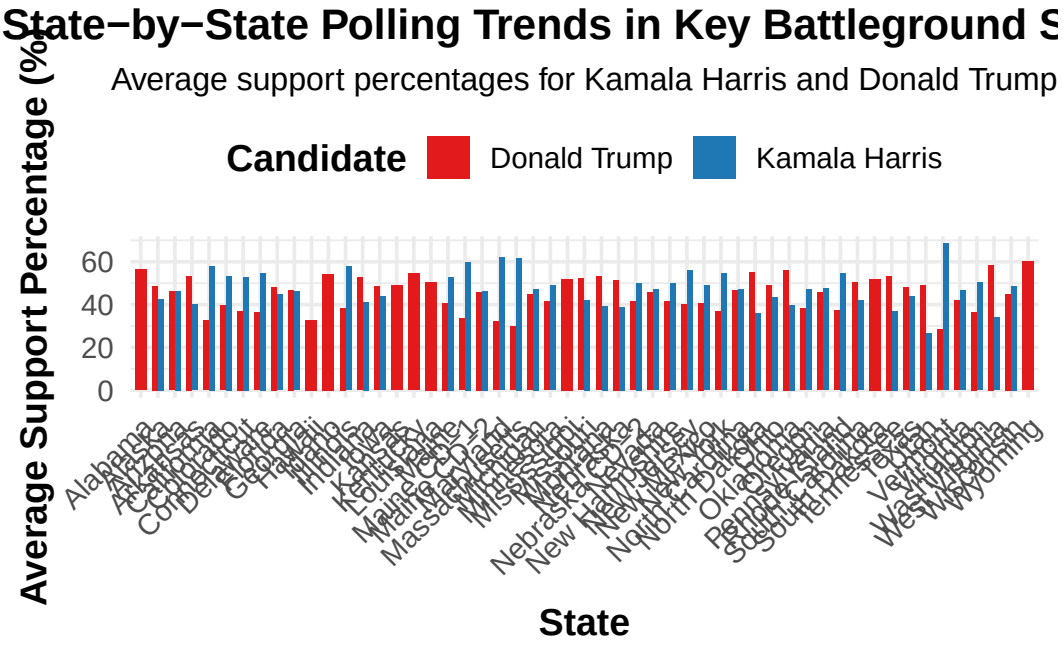

Figure 2: Graph of Dates Categorized by Year and Presented by Each Months

2.3.4 Sample Size and Poll Accuracy

The impact of sample size (sample_size) on the accuracy of polling results was examined. Larger sample sizes are typically considered more reliable, prompting the categorization of

polls into small, medium, and large groups based on sample size. For each group, the average percentage of support (pct) for Kamala Harris and Donald Trump was calculated, and variability within each sample size category was analyzed to evaluate the reliability of the results. A chartFigure 3 comparing candidate vote percentages by sample size category. Discussion: The results suggest whether larger sample sizes produce more accurate and reliable forecasts, helping us to understand the potential limitations of smaller polls.

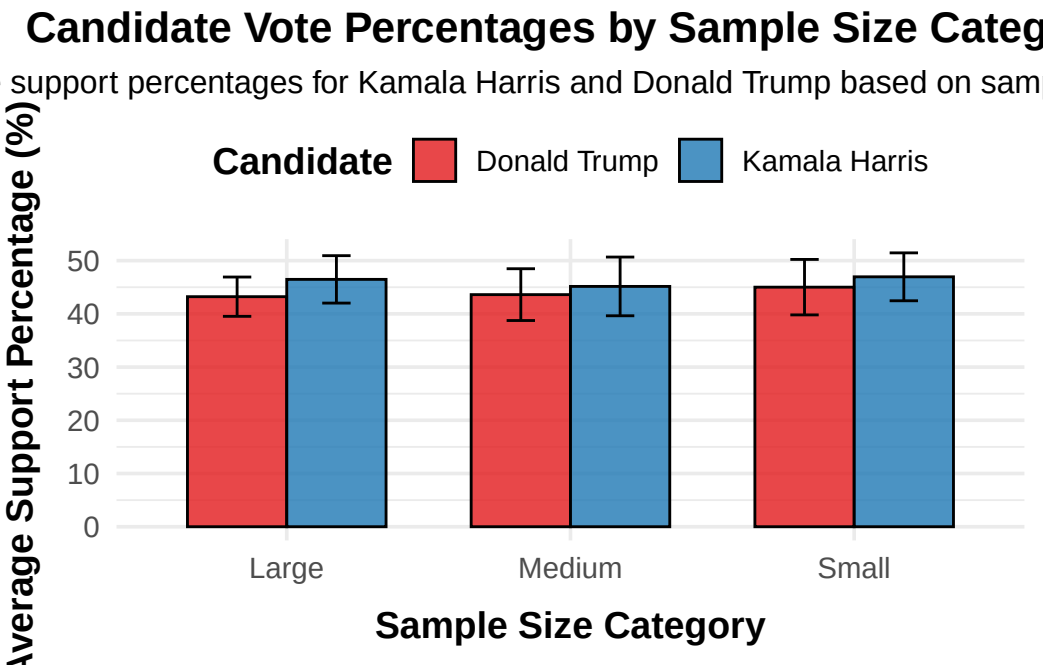

Figure 3: Graph of Dates Categorized by Year and Presented by Each Months

Alternate Analysis Data 5: Sponsorship and Bias in Polling Finally, the potential bias of polls sponsored by partisan organizations was assessed to determine whether such sponsorship influenced support for the associated candidate. The vote percentages (pct) for Kamala Harris and Donald Trump were compared in polls sponsored by Democratic-leaning and Republican-leaning organizations (sponsor_candidate_party). By calculating the average pct for each candidate within partisan-sponsored polls, this analysis explored whether partisan sponsorship resulted in a systematic overestimation of support for a particular candidate. A comparison of the polling results in partisan-sponsored polls, broken down by party affiliation is shown by Figure 4. Discussion: This analysis reveals potential biases in partisan-sponsored polls and assesses the objectivity of different polling organizations.

2.4 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset. A thorough discussion of measurement, relating to the dataset, is provided in the data section.

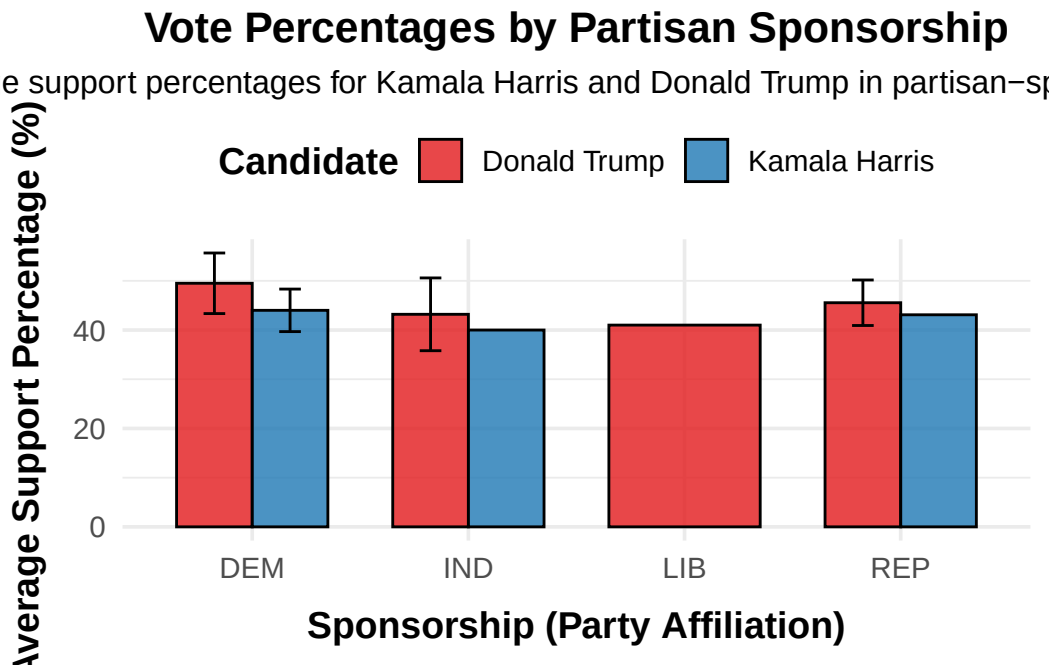

Figure 4: Graph of Dates Categorized by Year and Presented by Each Months

Please ensure that you explain how we went from some phenomena in the world that happened to an entry in the dataset that you are interested in.

2.5 Outcome variables

Add graphs, tables and text. Use sub-sub-headings for each outcome variable or update the subheading to be singular.

Some of our data is of penguins **?@fig-bills**, from (**palmerpenguins?**).

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.

2.6 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.

3 Model

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

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

The model should be nicely written out, well-explained, justified, and appropriate.

Detail the statistical model used, defining and explaining each aspect and its importance, and ensure that variables are well-defined and correspond with those discussed in the data section.

The model should have an appropriate balance of complexity—neither overly simplistic nor unnecessarily complicated—and be justified as suitable for the situation.

3.1 Modeling decisions

Explain how decisions made in modeling reflect the aspects discussed in the data section, including why specific features are included (e.g., why use age rather than age-groups, treating province effects as levels, categorizing gender, etc?).

3.2 Mathematical notations

Present the model using appropriate mathematical notation supplemented with plain English explanations, defining every component. If applicable, define sensible priors for Bayesian models.

3.3 Assumptions, limitations, circumstances

Clearly discuss the underlying assumptions, potential limitations, and circumstances where the model may not be appropriate.

3.4 Software & model validation

Mention the software used to implement the model, and provide evidence of model validation and checking—such as out-of-sample testing, RMSE calculations, test/training splits, or sensitivity analyses—while addressing model convergence, diagnostics, and any alternative models or variants considered,

3.5 Strenth and weeknesses and final choice

including their strengths and weaknesses and the rationale for the final model choice.

3.6 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of (**rstanarm?**). We use the default priors from `rstanarm`.

3.6.1 Model justification

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

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

Table 4: Explanatory models of flight time based on wing width and wing length

	First model
(Intercept)	1.12
	(1.70)
length	0.01
	(0.01)
width	−0.01
	(0.02)
Num.Obs.	19
R2	0.320
R2 Adj.	0.019
Log.Lik.	−18.128
ELPD	−21.6
ELPD s.e.	2.1
LOOIC	43.2
LOOIC s.e.	4.3
WAIC	42.7
RMSE	0.60

4 Results

Our results are summarized in Table 4.

Results will likely require summary statistics, tables, graphs, images, and possibly statistical analysis or maps. There should also be text associated with all these aspects.

Show the reader the results by plotting them where possible. Talk about them. Explain them. That said, this section should strictly relay results.

Regression tables must not contain stars.

5 Discussion

5.1 What is done in this paper?

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

5.2 What is something that we learn about the world?

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

5.3 What is another thing that we learn about the world?

5.4 Weaknesses and next steps

What are some weaknesses of what was done? What is left to learn or how should we proceed in the future?

Weaknesses and next steps should also be included.

Appendix 1

YouGov has been a popular polling organization in predicting U.S. elections for a long time, they have experience predicting US elections in the past. One of the key reasons for us to select YouGov to dive into it is their innovative use of an advanced statistical model known as Multilevel Regression with Post-stratification (MRP). This model allows YouGov to have detailed predictions by using survey data with demographic background information, ensuring the representativeness of broader populations.

Moreover, FiveThirtyEight found that YouGov did not exhibit a strong house effect during the 2016 U.S. election, meaning they did not show a bias toward any political party(Natesilver 2016), this further built up their credibility. Their model is intended to predict the votes of everyone in the US national voter file. YouGov's MRP model used to predict support for each candidate in the US 2024 presidential election has three parts. Firstly, they will use people's responses in the survey to estimate the likelihood of voting. From here, calculate their probability of voting for a specific party if people will vote. Combined and reweight responses, to give predictions for candidates (Bailey and Rivers 2024).

.1 Frame and sample

YouGov's frame of the sample is people in the YouGov panel which use a nonprobability sampling. Everyone can join the YouGov panel, and YouGov will recruit participants (the sample) to the survey with a rigours process. YouGov recruits American adults through various advertising methods and partnerships. They also offer surveys in multiple languages.

By completing surveys, participants earn points which can be exchanged for monetary rewards or vendor equivalents ("Methodology.", n.d.). YouGov will invite a representative set of panellists from their panel to take the survey. YouGov also have the resource to link participants to TargetSmart's voter file, so they can ensure they are verified registered voters. They include precinct-level vote data of voters to make their sample more representative by improving the geographic representation of underrepresented areas (Bailey and Rivers 2024). Other information used to invite survey participants who match the characteristics of the population of interest includes government data and other information collected from respondents when they join our panel like age, gender, race, education, etc ("Methodology.", n.d.).

Most of the participants in their presidential election survey have decided to vote while a small proportion are undecided. Thus, the answer they get is only what people plan to do instead of committed actions. So, the survey result and built model are best to reflect the current stage of the race, not perfect for prediction in the future. People can even change their decision to vote or not over time, which adds more changes over time (Bailey and Rivers 2024). This will affect YouGov's model based on their methodology, and YouGov is reflecting these changes with an updated model corresponding.

.2 Survey and After-Survey Process

YouGov's surveys are all online, with any device and at any time the respondent prefers. Questions asked include who they will likely vote for, as well as their likelihood to vote in the actual election and other questions. The same almost sample is being used in each survey. In this way, participants are tracked over time, so YouGov can study their shifts as the campaign progresses. For example, they have discovered stability in voters' candidate preferences for 2024 (Bailey and Rivers 2024). YouGov ensures the reliability and validity of its surveys through high standards and transparent reporting. Participants' privacy is protected by giving them control over the usage of their data such as allowing requests for data corrections and opting out of cookies. They aggregate responses in reporting to protect participants' identities. Something good about their questionnaire is that participants will have the choice of "prefer not to say" and skipping the question ("Methodology.", n.d.).

Another noticeable good point of their questionnaire is that it includes a broad range of topics including political, economic, environmental, and education issues, at the same time capturing people's potential engagement with voting by questions such as their likelihood of voting and method of voting. For the presidential election, YouGov has an unusually large sample compared to other opinion polling, with nearly 100,000 interviews in total contributing to their estimate model. They don't only continue with this model; after September, they will update with re-interviewed responses from over 20,000 people (Bailey and Rivers 2024). They build long-term relationships with their panellists.

For the presidential election, YouGov's U.S. panellists have been surveyed regularly since December 2023, and these panellists do not only get interviewed once but instated, and engaged in monthly or quarterly re-interviews (Bailey and Rivers 2024). But besides their effort to gather a large sample and actively include underrepresented regions by selective sampling, there are still areas that are hard to reach and lacking data groups that only have a small sample, such as people in small states and younger voters in larger states. Thus, after completing surveys, YouGov employs a weighting process to adjust respondents' influence based on their demographic characteristics and presidential vote (Bailey and Rivers 2024). In helps to develop the inclusiveness of the data gathered,

.3 Quality of Data Gathered

To ensure accuracy, YouGov applies a strategy of monitoring, testing, and refinement. They have a team with techniques to catch unreliable respondents. They also apply consistency checks to people's responses to avoid fraudulent responses. People who do not meet response quality standards will be removed from the final sample ("Methodology.", n.d.). For example, in their poll from October 12-15, 1,869 started the survey initially. 145 were deleted due to break-offs, 100 more were removed for data quality purposes, reasons including short interview completion time, and failed attention check, failed consistent checks etc. So the reporting is based on the remaining 1,624 respondents.

.4 Weakness

Their surveys are completely online, and therefore limited to individuals with internet access. Also, there could be additional factors that could influence the results beyond the reported margin of error. These include how they are phrasing questions and respondent bias, all of which can introduce potential errors in the survey outcomes.

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

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 algo-
rithm

References

"2024. n.d. "Presidential Election Calendar - 270towin." 270toWin. Com." *Www* 270towin.com/2024-presidential-election-calendar.

Bailey, Delia, and Douglas Rivers. 2024. "How YouGov's MRP Model Works for the 2024 u. S. Presidential and Congressional Elections." *YouGov* today.yougov.com/politics/articles/50587-how-yougov-mrp-model-works-2024-presidential-congressional-elections-polling-methodology (September).

Blumenthal, Mark. 2017. "After Obama, Models and Survey Science Won the Day." *HuffPost* www.huffpost.com/entry/2012-poll-accuracy-obama-models-survey_n_2087117 (December).

Galva, Alejandro A. Alonso. 2024. "The President Has Dropped Out of the Race. What's Next?" *Colorado Public Radio* 2024 (July): 07/23.

Iannone, Richard, Joe Cheng, Barret Schloerke, Ellis Hughes, Alexandra Lauer, JooYoung Seo, Ken Brevoort, and Olivier Roy. 2024. *Gt: Easily Create Presentation-Ready Display Tables*. https://CRAN.R-project.org/package=gt.

"Methodology.". n.d. *YouGov*. today.yougov.com/about/panel-methodology.

Natesilver. 2016. "Election Update: Leave the LA Times Poll Alone!"

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. 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.