Forecasting the US Election 2024*

##TODO: result here

Maryam Ansari Amy Jin Maggie Zhang

November 2, 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 2024)). 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) for plots, gt(Iannone et al. 2024) for tables, tidyr(Wickham and Henry 2023), arrow(Richardson and Labs 2023) for parquet, here(Müller 2023), rstanarm(Goodrich et al. 2023), broom(Robinson et al. 2023), loo(Vehtari et al. 2023), lubridate (Grolemund and Wickham 2023), while tibble (Wickham and Müller 2023) helped simplify data frame management. The testthat package (Wickham et al. 2023) was essential for unit testing and ensuring code reliability, and we employed styler (Walthert and Meyer 2023) for reformatting and maintaining a consistent code style.

2.2 Variables

The dataset comprises several key variables of interest, including but not limited to,

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

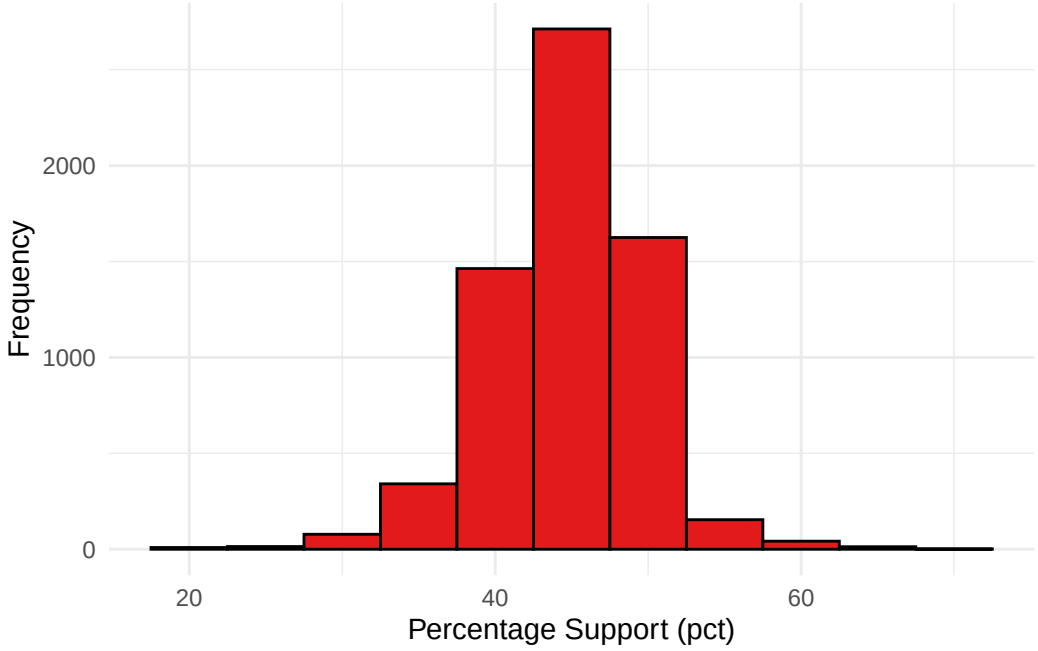
population: specifying the voting group described (e.g., likely voters)

candidate_name: the names of the candidates in the poll (e.g., Kamala Harris, Donald Trump)

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 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: Histogram of Trump Support Percentage (pct)



2.2.1 Response variables:

The main response variable of our analysis is the percentage support (pct) for Donald Trump and Kamala Harris shown in each poll. This number tells us what percentage of people said they would vote for each candidate when the survey was taken. We use it as the basis of our electoral prediction as it helps us understand how people feel about the candidates.

TODO: descibe

2.2.2 Predictor variables:

Our model takes into account various variables; their names and descriptions are,

Pollster: The name of the polling organization that conducted the poll (e.g., YouGov, RMG Research).

State: The U.S. state where the poll was conducted.

Sample Size: The total number of respondents that participated in the poll. End Date: The date the poll ended.

These variables were selected and kept due to their evident relationship with the response variable and the significance of it. More details of the relationships and model are provided in later sections.

Table 2: 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

2.2.3 Data Manipulation and cleaning:

The data was cleaned, filtered and mutated to better align with our prediction model. The software and packages used have been highlighted in Section 2.1. The key steps of our cleaning process were:

Filtered out lower quality polls by only keeping polls that met a set numeric grade threshold (numeric_grade >= 2.5). Filtered out pollsters that had less than 5 entries. Filtered out data that was collected before July 21st, 2024. Added a variable num_trump to count the number of votes for Donald Trump using pct. For data entries where the state was missing, they were categorised as "National", to standardize the data.

The dataset had several other variables that we did not include in our analysis. These variables were either majorly missing values or not informative enough to be included.

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 table (Table 3) and bar chart Figure 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 3: Summary Statistics for Support Percentages Across Pollster Reliability Levels

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

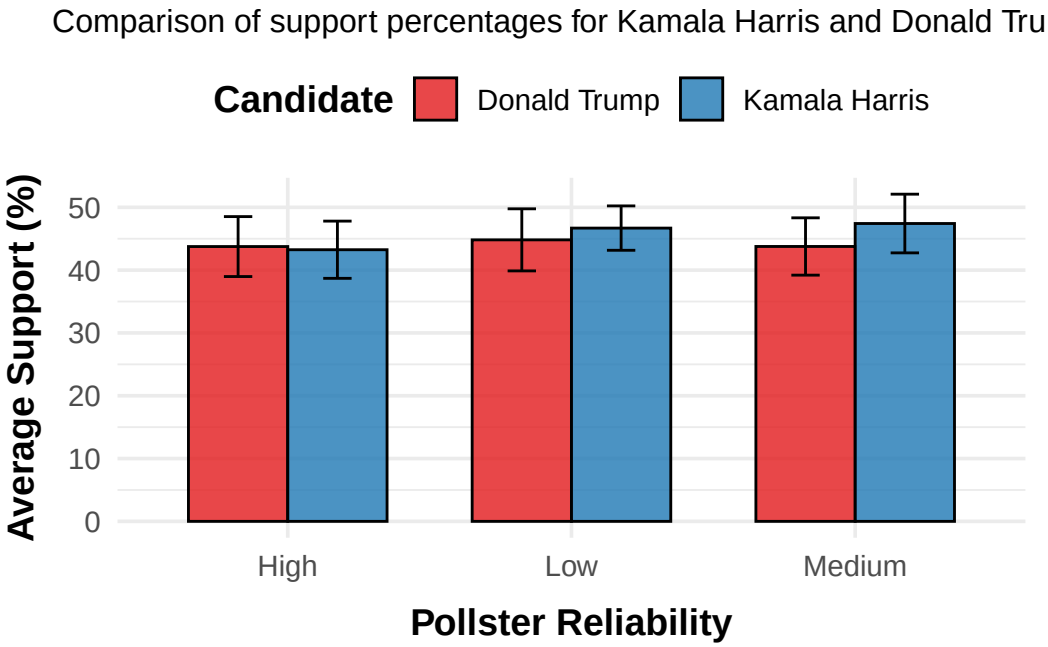


Figure 1: Average Support Percentage by Pollster Reliability

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 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 4 Discussion: We observed potential biases in support based on the polling method, providing insights into which methodologies might offer more reliable forecasts.

Table 4: Table of Methodologies comparing the Mean Support Percentage for Donald Trump and Kamala Harris with Difference (Kamala Harris - Donald Trump)

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 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 2 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.

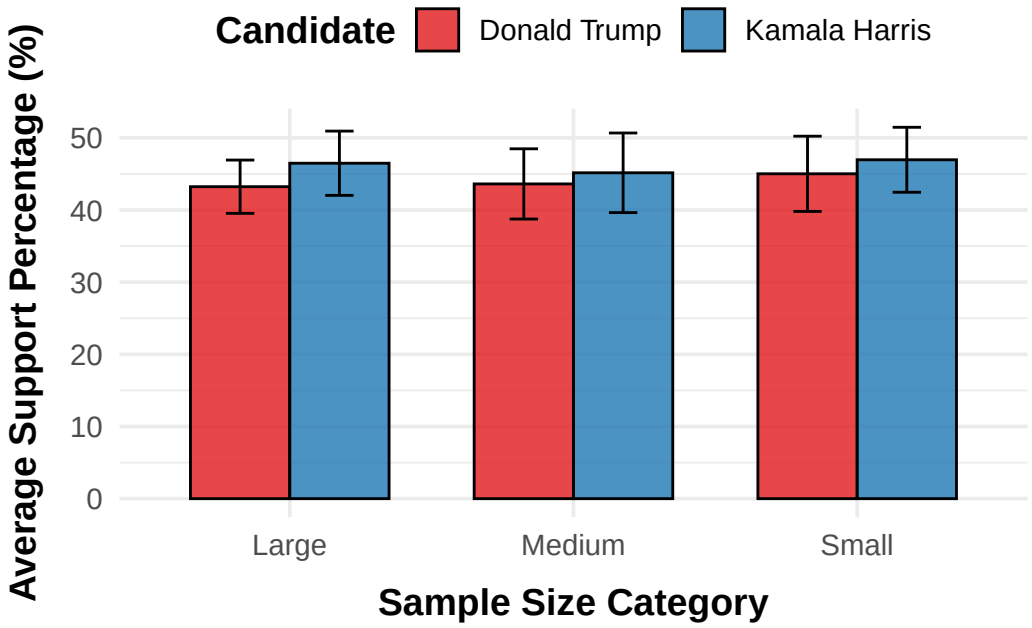


Figure 2: Candidate Vote Percentages by Sample Size Category

2.3.4 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 3. Discussion: This analysis reveals potential biases in partisan-sponsored polls and assesses the objectivity of different polling organizations.

2.4 Measurement and Limitations

This dataset comes with several important measurement and limitations:

Polling Accuracy: While efforts are made to prioritize high-quality polls through numeric ratings, differences in polling methods can still introduce some degree of bias. Pollster ratings enhance reliability but cannot fully eliminate methodological inconsistencies.

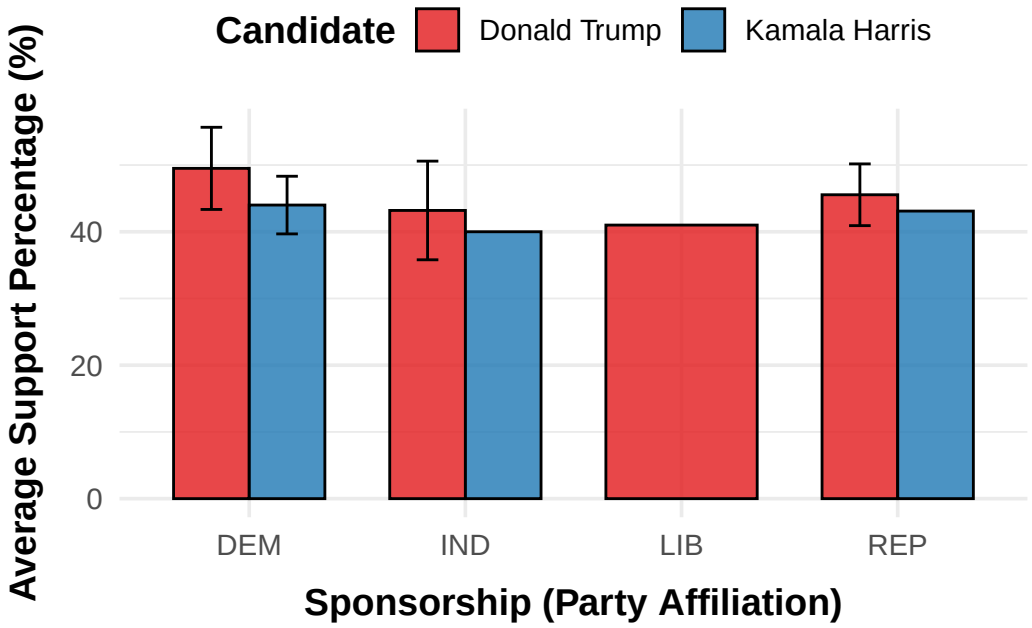


Figure 3: Vote Percentages by Partisan Sponsorship

Time Sensitivity: The dataset captures voter preferences at particular points rather than over time, providing only a static view. This limitation means that shifts in voter sentiment throughout the campaign are not represented.

Geographical Coverage: The dataset includes both national and state-level polling, but coverage is uneven. High-interest states, like swing states, are polled more frequently, while data from other regions may be sparse, potentially affecting regional analysis.

Participation Bias: Even with strict rules for polling, certain biases remain possible, such as selection bias in survey participation and response bias due to social desirability factors.

3 Statistical Model and Analysis

3.1 Model Overview

The goal of our modeling is to predict which candidate from Donald Trump and Kamala Harris will win the 2024 US election. We will firstly, predict support for the two candidates using the tailored cleaned data in the 2024 US presidential election using aggregated polling data, and secondly, to select a model that balances complexity and interpretability while accurately capturing the key trends and effects. This section describes the statistical models used, justifies the modeling choices, and discusses the assumptions, limitations, and validation

strategies. Background details, including diagnostics and additional model exploration, are available in Appendix C.

We utilized both traditional linear models and Bayesian models. The Bayesian approach, in particular, allows us to incorporate prior beliefs and better quantify uncertainty, which is crucial given the inherent variability in polling data.

3.1.0.1 Model Specifications

We constructed four models of increasing complexity:

Model 1: A linear regression where the dependent variable,

pct, the percentage of respondents supporting each candidate, is modeled as a function of the date the poll ended,

end_date, to capture temporal trends.

Model 2: An extension of Model 1 that includes a categorical variable,

pollster, to account for systematic differences between polling organizations.

Model 3: Adds a binary indicator,

is_national, to Model 2 to differentiate between national and state-specific polls.

Model 4: A Bayesian model implemented using the `rstanarm` package, incorporating the same predictors as Model 2 but allowing for more robust estimation under uncertainty.

3.1.1 We mathematically define these models as follows:

3.1.1.1 Model 1

$$y_i = \beta_0 + \beta_1 \cdot \text{end\_date}_i + \epsilon_i, \quad \epsilon_i \sim \text{Normal}(0, \sigma^2)$$

3.1.1.2 Model 2

$$y_i = \beta_0 + \beta_1 \cdot \text{end\_date}_i + \sum_{j=1}^{J} \gamma_j \cdot \text{pollster}_{ij} + \epsilon_i, \quad \epsilon_i \sim \text{Normal}(0, \sigma^2)$$

3.1.1.3 Model 3

$$y_i = \beta_0 + \beta_1 \cdot \text{end\_date}_i + \sum_{j=1}^{J} \gamma_j \cdot \text{pollster}_{ij} + \delta \cdot \text{is\_national}_i + \epsilon_i, \quad \epsilon_i \sim \text{Normal}(0, \sigma^2)$$

3.1.1.4 Model 4 (Bayesian)

$$y_i \mid \mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma), \quad \mu_i = \beta_0 + \beta_1 \cdot \text{end\_date}_i + \sum_{j=1}^{J} \gamma_j \cdot \text{pollster}_{ij} + \delta \cdot \text{is\_national}_i$$

Priors for Model 4:

$$\beta_0, \beta_1, \gamma_j, \delta \sim \text{Normal}(0, 5), \quad \sigma \sim \text{Exponential}(1)$$

3.1.1.5 Variables and Justification

1. **Dependent Variable**: (y_i) (support percentage, `pct`).
2. **Predictors**:
    - (end_date): Captures time trends, which are essential in understanding shifts in public opinion.
    - (pollster): Accounts for differences in polling methodologies and potential biases.
    - (is_national): Differentiates between polls conducted at the national level and those specific to individual states, as national polls tend to exhibit different characteristics and sample compositions compared to state polls.

These modeling choices reflect insights from the data section, where we observed systematic variations in support percentages over time, across pollsters, and between national and state polls.

3.1.1.6 Model Justification

The modeling approach was designed to balance interpretability and predictive power:

Model 1 serves as a baseline, capturing the temporal trend.

Model 2 addresses variability between pollsters, which is crucial given differences in methodologies and sample recruitment.

Model 3 adds another layer of nuance by accounting for the national vs. state-specific distinction.

Model 4 employs Bayesian inference, allowing us to incorporate uncertainty and use prior distributions for more robust predictions.

The Bayesian approach is particularly useful given the variability in polling data. Using `rstanarm`, we employ default weakly informative priors, which are appropriate given the lack of strong prior information. These priors help regularize the estimates and prevent overfitting.

3.1.1.7 Assumptions, Limitations, and Circumstances

1. **Assumptions**:
    - Linear relationships between predictors and the outcome.
    - Independence of observations, which may be violated if polls from the same organization are correlated.
    - Normality of residuals in linear models.
2. **Limitations**:
    - The models do not account for potential interactions between variables or non-linear effects, which could improve predictive accuracy.
    - The Bayesian model's convergence can be sensitive to the choice of priors and the data's variability.
3. **Circumstances**:
    - The models may be less appropriate in highly volatile electoral environments where sudden shifts in public opinion occur.
    - The temporal model assumes a relatively smooth trend, which may not hold in the presence of significant events influencing voter sentiment.

3.1.1.8 Software and Model Validation

The models were implemented in R using the `tidyverse` and `rstanarm` packages. We assessed model performance using:

AIC/BIC: For model selection based on goodness-of-fit and penalization for complexity.

Cross-validation: (To be detailed further in the appendix) to evaluate out-of-sample performance.

Residual Diagnostics: To check for violations of model assumptions.

Model Convergence: Monitored using diagnostics provided by `rstanarm` for the Bayesian model.

The results and details are in Section C.

3.1.1.9 Strengths, Weaknesses, and Final Model Choice

- **Strengths**: The models are interpretable and straightforward, with the Bayesian approach providing a robust framework for uncertainty quantification.
- **Weaknesses**: Potential oversimplification and assumptions of linearity. Future iterations could explore non-linear effects or hierarchical models.
- **Final Choice**: We selected **Model 2** (linear model with `end_date` and `pollster`) for its balance of interpretability and predictive performance. The Bayesian model serves as a robustness check, with results presented in the appendix.

Our final model selection and analysis provide a reliable and transparent prediction framework for Kamala Harris's support in the 2024 US election.

4 Results

Our results are summarized in Table 5.

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.

Table 5: 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

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 Appendix B: Idealized Survey Methodology and Survey Design

A.1 Methodology Overview

We will be developing a prediction model and conducting a survey targeting eligible voters in the United States for the 2024 presidential election. The study will observe key demographic groups influencing voting behaviour, such as age, ethnicity, income, education, and political affiliation (Leighley and Nagler 2013). Employing a combined stratified and quota sampling approach, we will ensure representation across these demographics, with a particular emphasis on swing states to enhance the accuracy of our predictions (FitzGerald 2024). Recruitment will utilize both online and offline methods, including community outreach and random digit dialling, while incentives will be offered to boost participation. Data validation will involve screening questions and consistency checks to ensure reliable responses. Finally, we will implement hybrid modelling techniques to aggregate survey data and apply statistical models for more accurate electoral predictions, correcting for any biases through weighting methods (Pasek 2015) (Wlezien and Erikson 2002). A copy of our designed survey will be attached.

A.2 Target Population

Target Population: The target population for our prediction model and survey includes all eligible voters across the United States participating in the upcoming 2024 presidential election. Population context study: We will observe specific key groups within the population that will affect people's voting actions. These include age groups, racial and ethnic, educational levels geographic regions, political affiliation, etc. (Leighley and Nagler 2013).

A.3 Sampling Approach

We will combine Stratified Sampling and Quota Sampling. Stratification: Stratified sampling involves dividing the population into distinct subgroups (strata) and then randomly sampling from each stratum (Thompson, 2012). This will ensure that each demographic group is represented in the sample, allowing for a more informative analysis of voter behaviour within each group.

A.3.1 Quota Sampling within Strata:

To improve representation, quota sampling is used within each stratum. This will involve collecting data until the predetermined quotas for each stratum are met (Thompson, 2012). This approach will ensure that the sample is more representative of the broader population, better to make a generalization.

A.3.2 Idefitiying Strata:

The target population consists of eligible voters in the U.S., which can be further divided into strata based on key demographic characteristics and grouping factors identified in the previous stage. The criteria for defining these strata include age, ethnicity, income level, education, and geographic location (Leighley and Nagler 2013). For example, we can have: Age: 18-29, 30-44, 45-59, 60+ Income: Low-income (<$40,000), Middle-income ($40,000-$100,000), High-income (>$100,000) Political Affiliation: Democrat, Republican, Independent, Other Set Quotas: Within the defined strata, we will establish quotas for each subgroup to ensure adequate representation based on key demographic features. These quotas will be determined using recent census data and voter registration records, in line with methodologies from reliable national polling (Keeter et al., 2016). A well-designed quota system will enhance balanced representation across the defined strata. For example, we can implement: Age: 18-29: 15% (180 respondents) 30-44: 25% (300 respondents) 45-59: 30% (360 respondents) 60+: 30% (360 respondents) Swing States: An important planning consideration is our focus on swing states to enhance the precision and efficiency of our election forecast. Swing states, characterized by competitive races, significantly influence the final outcome. In the 2024 election, these states could realistically be won by either Democrat Kamala Harris or Republican Donald Trump. Key swing states include Arizona, Georgia, Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin, where both major campaigns are actively targeting undecided voters (FitzGerald 2024).

To achieve this, we will:

Increase Quota Allocation for Swing States: Allocate a higher portion of the sample size to respondents from swing states to ensure these areas are more heavily represented. Adjust Stratification and Quotas by Region: Implement customized quota settings within each swing state to capture its unique demographics. This approach allows us to predict the voting patterns of distinct groups, increasing the likelihood of capturing accurate outcomes. For example: In Michigan, we will allocate more resources and further divide our efforts by ethnicity. The state has a significant Arab-American population, and the attitudes of both major parties toward this group will likely influence their decisions, which can greatly impact the overall outcome of the state (FitzGerald 2024).

In contrast, states like Hawaii will be excluded from our analysis, as it has consistently leaned Democratic since its admission in 1959, indicating a solid Democratic stronghold ("Who Will Win Hawaii?: The Hill and DDHQ" 2024).

A.3.3 Respondent Recruitment

A.3.3.1 Online Recruitment Methods

With advancements in technology, online recruitment has become increasingly cost-efficient. This approach will direct potential respondents to an online survey platform where they can easily participate in the study.

Email Outreach: Collaborations with community organizations and advocacy groups will facilitate email outreach to their members.

Online Panels: Utilization of established online survey panels will possibly make recruitment efficient.

Social Media Advertising: Targeted advertisements will be utilized on platforms such as Facebook, Twitter, and Instagram to reach specific demographic groups based on age, location, and political interests. While this approach effectively captures the desired participant groups, we must remain cautious of potential sampling bias.

A.3.3.2 Offline Recruitment Methods

Phone Outreach - Random Digit Dialing (RDD): We will employ a random-digit dialling method to recruit respondents for phone interviews. This approach allows us to reach individuals who may not be engaged online, particularly older demographics and those in rural areas. However, this traditional recruitment format will constitute only a small portion of our overall efforts, as advancements in technology have made online recruitment significantly more efficient. We plan to use phone outreach as a coplelement when online recruitment does not yield sufficient participants. This approach helps avoid inclusion errors by reaching populations without internet access or those who are not engaged in online communities. However, we face challenges such as increasing non-response rates due to the public's growing reluctance to participate in phone surveys and the rising use of technologies that screen unsolicited calls (Wang et al. 2015).

A.3.4 Incentives for Participation

Singer and Ye (Singer and Ye 2013) discuss incentives in surveys can significantly enhance response rates and reduce nonresponse bias, leading to more accurate data collection. However, it is essential to strike a balance to avoid low-quality data resulting from participants motivated solely by the incentive; excessively large rewards can skew results. Monetary Compensation: Respondents will have the opportunity to receive a monetary reward of approximately $20 for doing the survey. Gift Cards: Alternatively, we can offer gift cards to popular retailers or online platforms as incentives. Additionally, we may collaborate with funding resources willing to provide funds, enabling us to distribute gift cards while simultaneously promoting their brand.

A.3.5 Extra Effort on Specific Demographics:

We will keep monitoring the demographic composition of respondents throughout the data collection process to ensure that specific groups meet their quotas. If certain demographics are underrepresented, we will implement targeted outreach efforts, such as collaborating with organizations that can help raise awareness of the survey within those communities. This approach will enhance participation from these groups. If the collected sample still does not align with the desired proportions indicated by the population, we will perform post-stratification adjustments (weighting) to correct any imbalances.

A.3.6 Data Validation

Screening Questions: We will implement initial screening questions to ensure the data we collect reflects reality. This will include verifying participants' voting eligibility by confirming their citizenship and age. Consistency Checks: Drawing from YouGov's questionnaire methodology, we will include multiple questions in various formats to assess a single point, such as voting intention. By identifying inconsistent responses in this manner, we can enhance the accuracy of our data. Participants with heavily inconsistent answers will be noted, and if discrepancies are significant, their responses will be excluded from the final analysis.

A.3.7 Poll Aggregation and Modeling with Hybrid Approaches

Adopting Hybrid Models for Election Forecasting: According to Pasek (Pasek 2015), surveys have become essential tools for forecasting outcomes with three primary strategies for pooling survey data: aggregation, predictive modelling, and hybrid models. From his study, the hybrid approach has the most potential to accurately pool election information and make predictions. It combines both aggregation and prediction. Here, we will be using a traditional hybrid model, which aggregates survey data and uses statistical models to predict.

A.3.7.1 Survey Aggregation:

This involves pooling results from multiple surveys, weekly rolling averages will be used to smooth out random fluctuations and limit random error by reducing the uncertainty of the estimates (Pasek 2015).

A.3.7.2 Predictive Modeling:

Our models might include variables such as demographic information, past election results, and economic indicators. The goal is to build a model that use statistical techniques to analyze these factors and identify patterns which help to forecast potential electoral outcomes (Pasek 2015). Some specific poll-based forecasting models can be used to mitigate the tendency

for survey results to overreact to campaign events by accounting for how polls are expected to relate to voter behaviour and discounting temporary shifts (Wlezien and Erikson 2002). Integration: The most effective aspect of our hybrid model is its ability to integrate both aggregated survey data and results from predictive models. This allows us to leverage individual survey results alongside predictive elements such as trends and patterns, ultimately refining our predictive outcomes. By doing so, the model not only generates meaningful predictions but also enhances our understanding of the interplay among various electoral forces (Pasek 2015).

A.3.7.3 Weighting:

Despite our best efforts to ensure the sample accurately reflects the population during the sampling stage, non-response bias and responses being discarded during consistency checks may still result in an imperfect representation of the population. Similar to methodologies used by YouGov, we will apply weights to adjust responses based on demographic representation within the voting population. This approach helps correct for any over- or under-sampled groups.

A.3.7.4 Trade-off:

By implementing a complex hybrid model that requires careful weighting adjustments, we are able to ensure that each demographic group's influence aligns with its actual proportion in reality; making the conclusion more easily generated by the population. At the same time, as with all complex models, this brings more challenges in data analysis, making it hard to interpret the result. This also requires more time and financial investment

A.3.8 Survey Questions and Design

A.3.8.1 Online surveys:

We will be using online surveys because of the advantages offered by technology and expanding internet access across regions. Online surveys provide flexibility and convenience, enabling us to customize questions to fit specific contexts, order, and coverage of questions. It is also cost-effective, especially when reaching large, geographically dispersed samples, as in our case. Additionally, they facilitate real-time data monitoring to ensure data quality, while reducing the administrative and logistical costs associated with traditional survey methods (Evans and Mathur 2005).

A.3.8.2 Question Structure:

The survey design will be designed to ensure all questions remain with clarity, neutrality, and inclusivity. which helps avoid introducing systematic bias. We will avoid leading language that could skew responses predictably. By reducing ambiguity, we enhance response quality and reliability, thereby reducing measurement error and ensuring our data represents a broad spectrum of opinions. Question Type: The survey will primarily use Likert scales and multiple-choice formats to capture insights on voting intentions and demographics, both crucial for understanding potential election outcomes. These are decided to gather high-quality responses with high usability in prediction.

A.3.8.3 Demographic:

To accurately reflect real-world diversity, the survey will include demographic questions, such as those on race, ethnicity, and education level, with an inclusive range of options. This minimizes sampling bias by offering choices that accommodate diverse identities, thus helping participants feel represented and reducing the oversimplification of our sample. For example, in the gender question, we'll include options like "Woman," "Man," "Transgender," "Non-binary/non-conforming," "Prefer not to answer," and "Other" (with a text box to specify). This approach allows participants to self-identify accurately, reducing both sampling bias and response variance across different groups.

A.3.8.4 Voting intention:

This section will assess participants' likelihood of voting, designed to minimize social desirability bias, a type of response bias where the participant might change their answer to appear more socially responsible and maintain themselves in better self-pictures (Grimm 2010). We will be using a Likert scale to enable nuanced responses without implying a "correct" answer. This way,with less biased data, we can capture a range of genuine attitudes (such as: "On a scale of 1–5, how likely are you to vote in the upcoming election?")

A.3.8.5 Political Preference:

In this section, we delve into participants' existing preferences for specific political parties. To reduce variance from inconsistent answers, we will ask multiple questions about party preference, including questions on past voting behaviour. By phrasing similar questions in varied ways, we can confirm consistency in responses (such as: "Which political party do you currently support?" and "Which party did you vote for in the last election?")

A.3.8.6 Social Issues:

This section aims to explore participants' viewpoints on influential issues in this presidential election, such as economic policy, healthcare, and immigration (Nadeem 2024). Questions will be designed free from framing bias, this is where specific wording might unduly influence responses (Entman 2007). Neutral wording helps capture participants' genuine opinions on issues most likely to impact their voting behaviour.

A.3.9 Budget Allocation Breakdown: Total of $100K

Survey Design and Implementation (30% - $30,000): Questionnaire development, pilot testing, and data collection. Sampling and Recruitment (25% - $25,000): Stratified sampling, participant incentives, and outreach. Data Validation and Cleaning (15% - $15,000): Screening and data preparation. Data Analysis and Modeling (20% - $20,000): Software licenses, model development, and bias correction. Reporting and Dissemination (10% - $10,000): Report and present to the public. Contingency Fund (5% - $5,000): For unexpected costs.

A.3.10 Google Forms Survey Link

Include a link to your Google Forms survey here. Survey Link Copy of the Survey Questions

B Additional data details

C Model details

C.1 Model Validation and Evaluation

Our goal in developing predictive models for the 2024 US presidential election is to ensure that our models are not only accurate but also robust and generalizable. To achieve this, we implement a comprehensive model validation process. This section describes our approach to model validation, including criteria used to compare models, diagnostic techniques employed to evaluate model assumptions, and specific methods to ensure the reliability of our Bayesian model.

We validate our models using several strategies: Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for model selection, cross-validation to evaluate out-of-sample performance, and diagnostic plots to assess model assumptions. Additionally, we perform posterior predictive checks and convergence diagnostics for the Bayesian model.

C.1.0.1 Software and Implementation

The models were implemented in R using the `tidyverse` and `rstanarm` packages. The validation process was executed using tools from these packages and additional packages, such as `broom` and `loo`, for model diagnostics.

C.1.1 Model Validation Process

C.1.1.1 1. AIC and BIC for Model Selection

To compare the linear models' goodness-of-fit while penalizing complexity, we use AIC and BIC. These metrics help identify the most appropriate model for the data, balancing explanatory power and parsimony. The results are presented in (Table 6):

Table 6: Model Comparison Table: AIC and BIC Values for Model Selection

Model	AIC	BIC
model_date	2,242.24	2,254.22
model_date_pollster	2,111.27	2,183.16
model_date_pollster_national	2,112.00	2,187.89

Results: Model 2 has the lowest AIC and BIC, suggesting it offers the best balance between goodness-of-fit and complexity. This model includes both time trends and categorical effects for pollsters and national versus state-specific polls.

C.1.1.2 2. Cross-Validation for Out-of-Sample Performance

We use leave-one-out cross-validation (LOO) for the Bayesian model to assess predictive performance. The `loo` package computes the expected log predictive density (ELPD), which indicates how well the model generalizes to new data. Table (Table 7) shows the result.

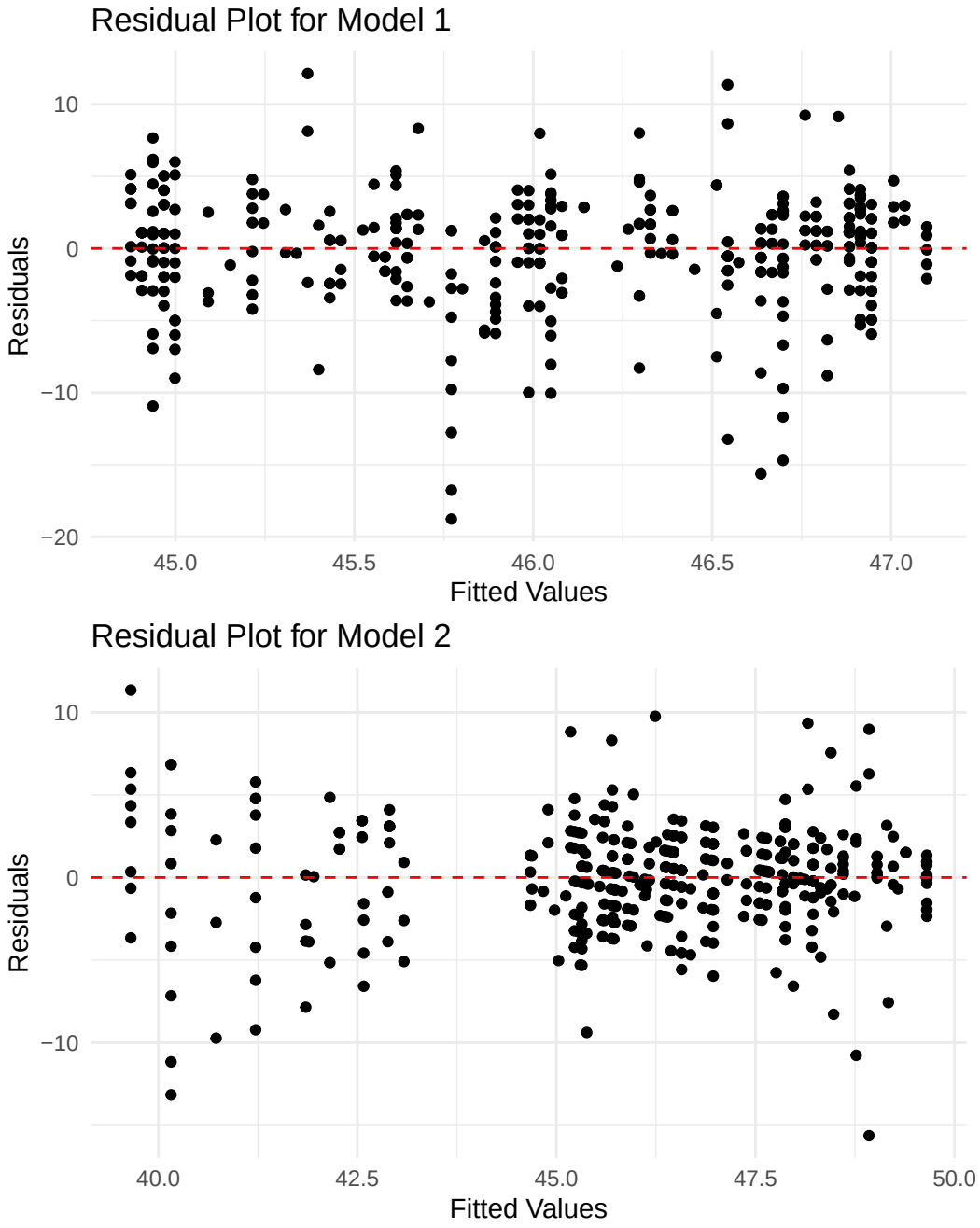
Table 7: LOO Summary for Bayesian Model LOOCV Result

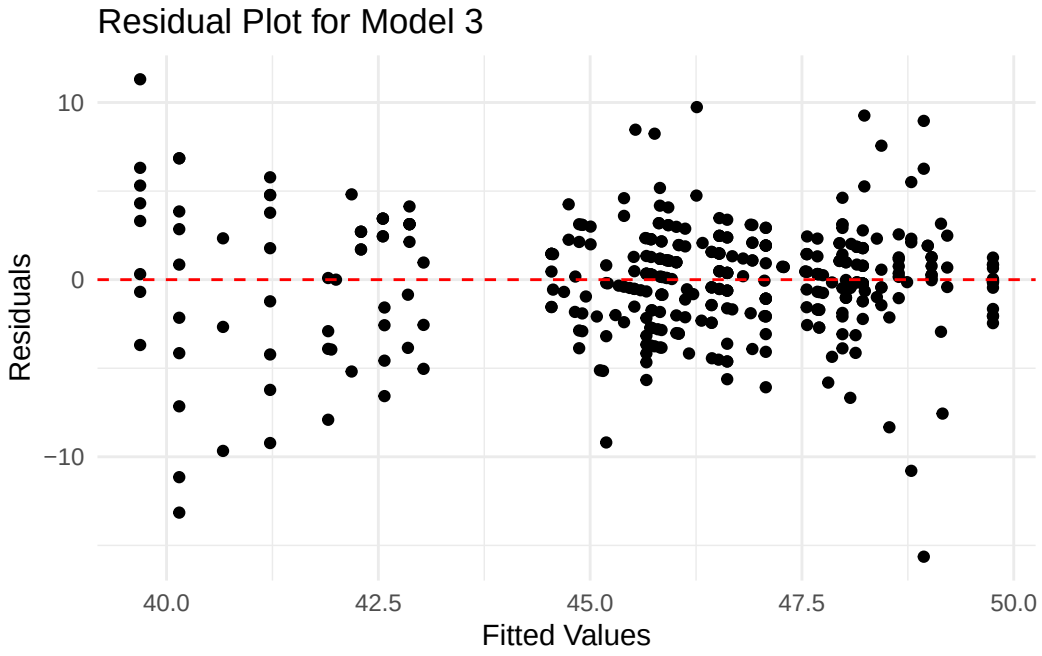
Metric	Value
LOOIC	2,117.29
SE(LOOIC)	45.42
p_loo	21.21
elpd_loo	-1,058.64
SE(elpd_loo)	22.71

Cross-Validation Results: The LOO estimate for the Bayesian model shows that the model has good predictive performance, supporting its use for forecasting election outcomes.

C.1.1.3 3. Residual Diagnostics for Linear Models

We generate residual plots to check for violations of assumptions such as homoscedasticity (constant variance of residuals) and linearity.





- **Model 1 (Simple Time Trend)**: The residuals show a pattern, suggesting potential model misfit.
- **Model 2 (Time Trend + Pollster)**: The residuals are more evenly spread, indicating an improved model fit.
- **Model 3 (Time Trend + Pollster + National/State)**: The residuals show no obvious patterns, suggesting that the model assumptions are reasonably met.

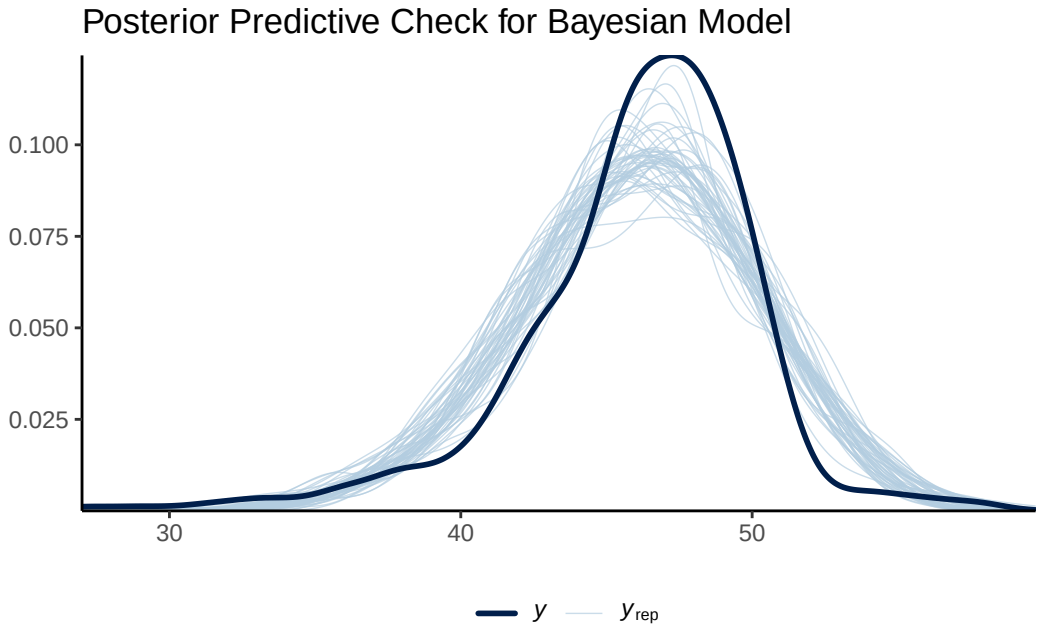
Interpretation: Model 2's residuals suggest that the model adequately captures the data's underlying structure, with no major violations of the linear model assumptions.

C.1.1.4 4. Bayesian Model Diagnostics

For the Bayesian model, we perform the following diagnostics:

C.1.1.4.1 a. Posterior Predictive Check

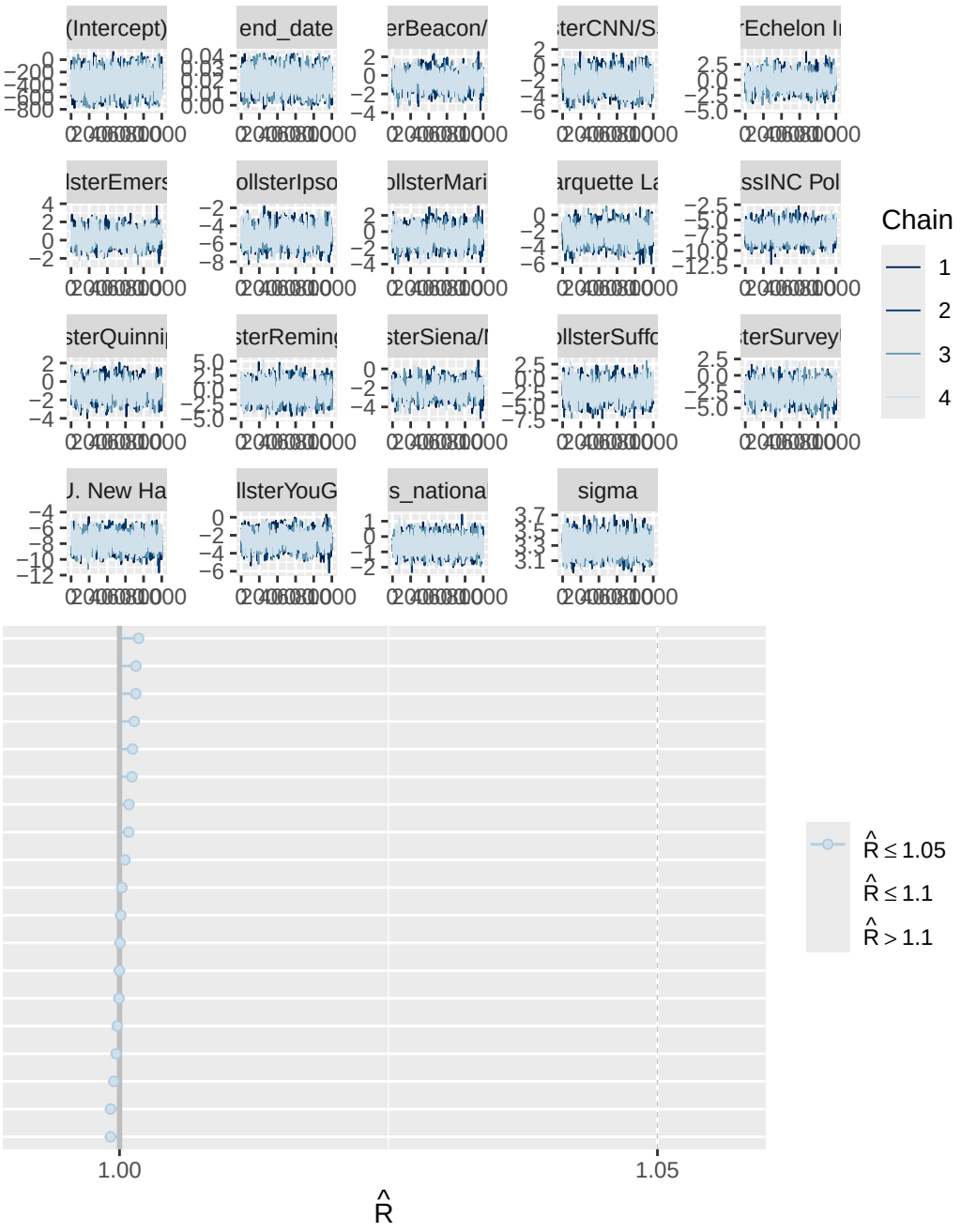
We use `pp_check()` from `rstanarm` to visualize how well the model's predictions align with the observed data.



Interpretation: The predictive distribution closely matches the observed data, indicating a good model fit.

C.1.1.4.2 b. Convergence Diagnostics

We assess MCMC convergence using trace plots and Rhat values: - **Trace Plots**: Show that the chains mix well and converge to a common distribution, indicating no issues with convergence. - **Rhat Values**: All Rhat values are close to 1, further confirming convergence.



Conclusion: The Bayesian model's diagnostics confirm that the model converges well, and the posterior samples are reliable.

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).

Entman, Robert M. 2007. "Framing Bias: Media in the Distribution of Power." *Journal of Communication* 57 (1): 163–73. https://doi.org/10.1111/j.1460-2466.2006.00336.x.

Evans, Joel R., and Anil Mathur. 2005. "The Value of Online Surveys." *Internet Research* 15 (2): 195–219.

FitzGerald, J. 2024. "Seven Swing States Set to Decide the 2024 US Election." *BBC News*. https://www.bbc.com/news/articles/c511pyn3xw3o.

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

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

Goodrich, Ben, Jonah Gabry, Imad Ali, Sam Brilleman, and Rok Češnovar. 2023. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm/.

Grimm, Pamela. 2010. "Social Desirability Bias." In *Wiley International Encyclopedia of Marketing*. Wiley. https://doi.org/10.1002/9781444316568.wiem02057.

Grolemund, Garrett, and Hadley Wickham. 2023. *Lubridate: Make Dealing with Dates a Little Easier*. https://CRAN.R-project.org/package=lubridate.

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.

Leighley, Jan E., and Jonathan Nagler. 2013. *Who Votes Now?: Demographics, Issues, Inequality, and Turnout in the United States*. Princeton University Press.

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

Müller, Kirill. 2023. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Nadeem, R. 2024. "Issues and the 2024 Election." *Pew Research Center*. https://www.pewresearch.org/politics/2024/09/09/issues-and-the-2024-election/.

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

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, and Ursa Labs. 2023. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Robinson, David, Alex Hayes, Simon Couch, Nicholas Tierney, and Max Kuhn. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Singer, Eleanor, and Cong Ye. 2013. "The Use and Effects of Incentives in Surveys." *The ANNALS of the American Academy of Political and Social Science* 645 (1): 112–41.

Vehtari, Aki, Jonah Gabry, Yuling Yao, Andrew Gelman, Mans Magnusson, and Paul-Christian Bürkner. 2023. *Loo: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models*. https://mc-stan.org/loo/.

Walthert, Lorenz, and Nicolas Meyer. 2023. *Styler: Non-Invasive Pretty Printing of r Code*. https://CRAN.R-project.org/package=styler.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. "Forecasting Elections with Non-Representative Polls." *International Journal of Forecasting* 31 (3): 980–91.

"Who Will Win Hawaii?: The Hill and DDHQ." 2024. *The Hill*. https://elections2024.thehill.com/forecast/2024/president/hawaii.

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.

Wickham, Hadley, and Lionel Henry. 2023. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

Wickham, Hadley, Jim Hester, Lionel Henry, and Winston Chang. 2023. *Testthat: Unit Testing for r*. https://CRAN.R-project.org/package=testthat.

Wickham, Hadley, and Kirill Müller. 2023. *Tibble: Simple Data Frames*. https://CRAN.R-project.org/package=tibble.

Wlezien, Christopher, and Robert S. Erikson. 2002. "The Timeline of Presidential Election Campaigns." *The Journal of Politics* 64 (4): 969–93. https://doi.org/10.1111/1468-2508.00159.

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.