Predicting the 2024 US Presidential Election with a Model-Based Forecast*

Using Generalized Linear Models to Predict Election Outcomes

Yuanyi (Leo) Liu Dezhen Chen Ziyuan Shen

October 27, 2024

In this paper, we use a logistic regression model based on "poll-of-polls" data to predict the winner of the 2024 U.S. presidential election. By combining multiple polls, the model smooths out short-term fluctuations in surveys to make more reliable estimates of voter preferences. The findings show early trends in the leading candidates and reflect public opinion and its evolution through the election period. This prediction improves understanding of electoral trends and helps to explain the changing political scene more clearly.

1 Introduction

The presidential election has a significant impact on national and global policies, which makes election forecasting a valuable tool for understanding political outcomes. However, individual polls often contain biases and short-term fluctuations that can obscure long-term trends. To deal with these challenges, we use a 'poll-of-polls' approach that combines multiple polls to provide more stable predictions of voter preferences over time. In this paper, we use national polling data from the period leading up to the 2024 U.S. presidential election to generate a logistic model that predicts the winner while capturing changes in public opinion throughout the election period.

The model estimates the probability of a candidate winning an election based on aggregated polling data. Logistic regression is used for binary outcomes such as election results, which allows us to model the likelihood of each candidate winning as new data becomes available. By taking into account changes in opinion polls and tracking trends in public opinion over time, this model improves the accuracy of predictions.

*Code and data are available at: Forecasting the 2024 US Presidential Election.

Our analysis identifies key trends among the major candidates and tracks changes in public sentiment throughout the campaign. The results suggest that aggregated polls can reduce some uncertainty by smoothing short-term fluctuations between surveys, but uncertainty cannot be eliminated. A sudden change in public opinion can still change the path of an election campaign, which highlights the importance of continuous polling monitoring.

A reliable election forecast can influence public expectations, and campaign strategies, and increase the visibility of reporting on election trends. An accurate forecast can also increase voter participation by providing individuals with a clearer perspective of the electoral landscape, and help them participate more effectively in the process of democracy. Prediction models like ours not only help with resource allocation for campaigns but also support the wider public by providing them with a clearer understanding of political dynamics.

The structure of the paper is organized as follows: following Section 1, Section 2 presents the data collection and cleaning process, along with an overview of the variables used in the analysis. Section 3 introduces the forecasting models, explaining why the selected models are suitable for predicting election outcomes based on aggregated polling data. Then Section 4 presents the main findings, including detailed crime trends for each neighborhood and year. Finally, Section 5 provides the results, highlighting key trends and predictions. Eventuallt, Section 5 concludes with a discussion of the findings, evaluating the reliability of the forecasts and identifying potential limitations of the models.

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023) to process and analyze polling data for the 2024 U.S. Presidential election. The dataset used for this analysis was obtained from the FiveThirtyEight 2024 U.S. Presidential Election Polls (Ryan Best 2024). It consists of polling data for the 2024 general election, covering various polling organizations and methodologies. Following methodologies discussed by "Telling Stories with Data" (Alexander 2023), we forecast election outcomes using the "poll of polls" method, which aggregates results from multiple polls to reduce bias and provide a more accurate representation of voter sentiment.

The dataset includes 15,891 rows and 52 columns, covering various pollster attributes such as pollster name, state, methodology, and polling results. To ensure the reliability of our analysis, we filter the data to include only polls with a numeric grade of 2.5 or higher, which represents high-quality, reputable pollsters. This filtering allows us to focus on polls that follow rigorous standards and have demonstrated transparency and accuracy.

Additionally, we limit the dataset to polls conducted after July 15, 2024, when Donald Trump officially announced his campaign, ensuring that the data reflects the most current public

sentiment. Other similar datasets from prior elections, such as data from previous general election cycles, could have been used for comparison. However, given that this analysis focuses on the upcoming 2024 election and the shift in public opinion, the most relevant data is specific to the current cycle.

2.2 Measurement

Polling data is a measurement of public sentiment captured through various survey methods. Polling organizations collect responses from individuals representing different segments of the population and ask them about their voting preferences. These responses are then weighted to reflect a more accurate representation of the electorate, based on factors like age, gender, race, and geographic region.

For this dataset, the poll results reflect the percentage of respondents who support a particular candidate. Different polling methods—such as live phone interviews, online panels, and mixed methodologies—capture this information. Each polling organization applies its own methodology, which can influence the results. For example, polls that use live interviews may experience different respondent behavior than those that use online surveys. More details on polling methodologies can be found in the ABC News methodology explanation.

The key measurement process involves converting real-world voter preferences into a dataset of percentages that reflect support for a candidate at a specific point in time. Polling organizations typically provide this data at a state or national level, which allows for both localized and broad interpretations of voter sentiment.

2.3 Outcome variables

2.3.1 Percent Support for Donald Trump (pct)

The primary outcome variable is the percentage (`pct`) of respondents who indicated support for Donald Trump in the 2024 general election. This percentage is calculated based on the total number of respondents in each poll divide by `sample_size`. The data includes observations from various states, providing both national and state-specific measures of Trump's support.

Figure 1 shows Donald Trump's polling percentages over time from July to October 2024. Each black dot represents an individual poll conducted on a specific date, indicating the percentage of voters who supported Trump in that poll. The blue line represents a trendline, smoothing out the individual poll results to show the general trajectory of Trump's support over time.

Initially, Trump's polling percentage appears stable with small fluctuations around 45%. However, as the campaign progresses into October, there is a slight upward trend in Trump's polling percentage, suggesting that his support increased as the election neared.

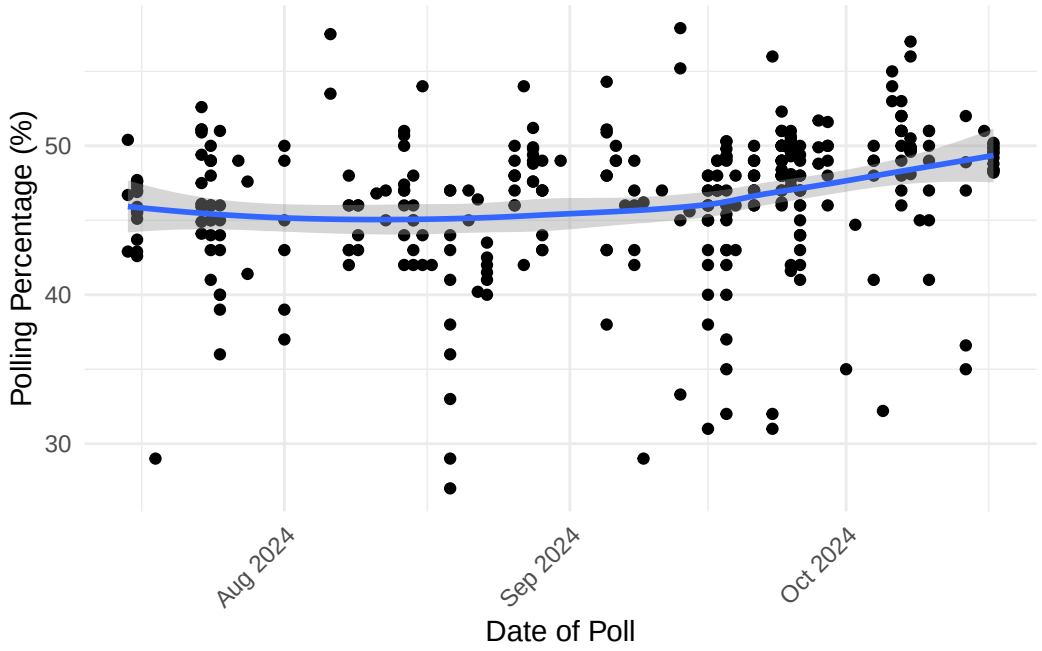

Figure 1: Donald Trump's polling percentages from July to October 2024. Each dot indicates the polling percentage for Trump on a specific date, while the blue line shows the trend in Trump's support.

The shaded region around the blue line represents the confidence interval, indicating the range within which the true polling percentage is likely to fall, accounting for sampling variation.

2.4 Predictor variables

2.4.1 Polling Methodology (methodology)

The `methodology` variable describes the method each pollster used to collect data. It includes approaches such as live phone interviews, online surveys, and mixed methods. For polls that used multiple collection methods (e.g., phone and online), we labeled them as "Mixed Voting" for simplicity. The methodology is important because different approaches can introduce varying degrees of bias, influencing the final reported percentages.

Placeholder for a table showing the distribution of polling methodologies across different states.

2.4.2 State (state)

The `state` variable identifies whether the poll is state-specific or national. In state-specific polls, the data reflects localized voter preferences, while national polls aggregate opinions across the entire country. For this analysis, we ignore national polls and focus entirely on state polls because electoral outcomes are determined on a state-by-state basis in the U.S. election system.

2.4.3 Polling Score and Transparency (numeric_grade and transparency_score)

The `numeric_grade` variable is a rating assigned to each polling organization, indicating the reliability of the poll. Pollsters with higher numeric grades are considered more reliable and consistent in their methodology (e.g. higher than 2.5). Additionally, the `transparency_score` measures how openly a pollster shares their methodology and data, which further affects the trustworthiness of their results.

2.4.4 Sample Size (sample_size)

The `sample_size` variable represents the total number of respondents surveyed in each poll. A larger sample size generally leads to more reliable and precise estimates of voter sentiment, as it reduces the margin of error. In contrast, smaller sample sizes can result in greater variability and less confidence in the results.

2.4.5 Poll Duration (duration)

The `duration` variable is a constructed variable that represents how long Donald Trump has been in the 2024 presidential campaign, measured as the number of days from his official campaign announcement on July 15, 2024 until the `end_date` of each poll. This variable helps quantify how much time has passed since Trump officially entered the race and allows us to examine how his support has evolved over the course of his campaign.

In summary, the dataset offers a rich variety of variables that capture voter sentiment, pollster reliability, and electoral dynamics. By carefully filtering and analyzing this data, we can build a robust model to forecast the outcome of the 2024 US Presidential election.

3 Model

Our modeling approach aims to quantify the relationship between polling characteristics and the probability of Harris winning the election. We use a generalized linear model (GLM) to evaluate how different factors—such as poll type, polling organization, poll score, and transparency score—impact Harris' support rate. The model is implemented using the stan_glm function, with a Gaussian distribution applied to describe the variability in support percentages.

In our analysis, we filtered the data to include only polls from organizations with a poll score above 3, ensuring the reliability of our data sources. We selected predictors including poll_type, pollster, pollscore, and transparency_score. Since the response variable (support percentage) is continuous, we utilized a Gaussian linear regression model.

The model assumes that the distribution of support percentage, given the polling characteristics, follows a normal distribution. This assumption makes the parameter estimation more straightforward and is commonly used in practice. We assigned normal priors with a mean of 0 and a variance of 2.5 for the model coefficients, which helps to maintain reasonable uncertainty while avoiding overfitting. Additionally, an exponential prior was included to handle extra variation in the model, contributing to the stability of the estimated results.

3.1 Model set-up

Define (y_i) as the predicted support percentage for Harris in the (i)-th poll. The predictor variables include (x_1, x_2, x_3, x_4, x_5, x_6), corresponding to poll type (State or National), pollster identity (e.g., Marquette Law School, Siena/NYT), poll score, transparency score, state, and pollster organization respectively. The model can be described by the following equations:

$$y_i \mid \mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 \tag{2}$$
$$\beta_0 \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta_j \sim \text{Normal}(0, 2.5), \quad j = 1, 2, 3, 4, 5, 6 \tag{4}$$
$$\sigma \sim \text{Exponential}(1) \tag{5}$$

The model is executed in R using the `rstanarm` package. Default priors from `rstanarm` are used, with the priors set to have a mean of zero and a moderate standard deviation to ensure a reasonable level of regularization.

3.1.1 Model justification

Considering the historical polling data and what we know about the political landscape, we believe that different aspects of polls will have a notable impact on Harris' predicted support in the upcoming election. Specifically, polls from McCourtney Institute/YouGov, YouGov/Center for Working Class Politics, higher poll scores, and greater transparency tend to have a positive influence on Harris' support. In contrast, polls from Siena/NYT, The Washington Post, YouGov, and YouGov Blue generally show a negative impact on her predicted support. Our purpose is to determine how much each polling characteristic affects voter support, thus helping us confirm or reconsider our assumptions about voter preferences.

We used a Gaussian linear regression model to examine the connection between polling characteristics and Harris' predicted support percentage. The support percentage (y_i) is a continuous measure, and its distribution aligns well with the assumptions of Gaussian regression. This method allows us to investigate how several factors—like poll type, pollster score, and transparency level—affect the level of support for Harris. By numerically encoding these predictors, we can better understand the influences on polling outcomes, ultimately offering a clearer picture of Harris' potential performance in the upcoming election.

Further justification for this model is based on its ability to demonstrate how different polling organizations and their characteristics affect Harris' predicted support percentage. The intercept value represents the baseline level of support when all other predictors are set to zero, providing a reference point. Each coefficient then reflects the expected change in support, depending on factors like poll type, pollster identity, poll scores, and transparency measures. Positive coefficients, such as those for McCourtney Institute/YouGov and YouGov/Center for Working Class Politics, suggest increased support for Harris, while negative coefficients, such as those for Siena/NYT, The Washington Post, YouGov, and YouGov Blue, indicate a decrease in support.

The relatively modest R^2 value reflects that the model only captures part of the variation in Harris' support, which is not unexpected given the unpredictable nature of polling data and the

Table 1: Summary of the Linear Regression Model

	first_model
(Intercept)	43.387
	(4.986)
numeric_grade	1.859
	(1.624)
as.factor(methodology)Live Phone	−1.280
	(2.374)
as.factor(methodology)Mixed Voting	0.012
	(2.329)
as.factor(methodology)Online Ad	−0.948
	(2.399)
as.factor(methodology)Online Panel	−2.321
	(2.407)
as.factor(methodology)Probability Panel	−1.492
	(2.423)
duration	0.030
	(0.006)
sample_size	0.002
	(0.001)
transparency_score	−0.323
	(0.160)
stateCalifornia	−16.090
	(1.530)
stateConnecticut	−12.659
	(2.081)
stateFlorida	2.113
	(0.858)
stateGeorgia	0.021
	(0.569)
stateIowa	−0.751
	(2.097)
stateMaine	−5.214
	(1.369)
stateMaine CD-1	−7.540
	(2.173)
stateMaine CD-2	1.640
	(1.390)
stateMaryland	−17.786
	(1.612)
stateMassachusetts	−16.475
	(1.120)
stateMichigan	−1.478
	(0.627)
stateMinnesota	−4.143
	(0.979)
stateMissouri	5.423
	(2.075)
stateMontana	6.082
	(1.486)
stateNebraska CD-2	−6.064
	(1.096)

complex dynamics of voter behavior. Despite this, including key polling attributes allows us to examine how different factors contribute to voter support, helping us understand the sources of variation in predicted support levels. The choice of Gaussian linear regression is appropriate in this case, as it effectively models the continuous outcome variable and provides a clear framework for interpreting the relationship between polling features and voter preferences.

4 Results

Our results are summarized in Table 2.

5 Discussion

5.1 First discussion point

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 Second discussion point

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

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Table 2: Explanatory models of flight time based on wing width and wing length

	First model
(Intercept)	43.387
	(4.986)
numeric_grade	1.859
	(1.624)
as.factor(methodology)Live Phone	−1.280
	(2.374)
as.factor(methodology)Mixed Voting	0.012
	(2.329)
as.factor(methodology)Online Ad	−0.948
	(2.399)
as.factor(methodology)Online Panel	−2.321
	(2.407)
as.factor(methodology)Probability Panel	−1.492
	(2.423)
duration	0.030
	(0.006)
sample_size	0.002
	(0.001)
transparency_score	−0.323
	(0.160)
stateCalifornia	−16.090
	(1.530)
stateConnecticut	−12.659
	(2.081)
stateFlorida	2.113
	(0.858)
stateGeorgia	0.021
	(0.569)
stateIowa	−0.751
	(2.097)
stateMaine	−5.214
	(1.369)
stateMaine CD-1	−7.540
	(2.173)
stateMaine CD-2	1.640
	(1.390)
stateMaryland	−17.786
	(1.612)
stateMassachusetts	−16.475
	(1.120)
stateMichigan	−1.478
	(0.627)
stateMinnesota	−4.143
	(0.979)
stateMissouri	5.423
	(2.075)
stateMontana	6.082
	(1.486)
stateNebraska CD-2	−6.064
	(1.096)

Appendix

A Pollster Methodology Overview and Evaluation

A.1 Overview of Siena/NYT

Siena College Research Institute (SCRI) and The New York Times launched a collaborative political poll in July 2013. Siena/NYT was known for its rigorous methodology and accuracy, which provided its first real-time midterm election poll in 2018.In 2024, FiveThirtyEight named it the most accurate pollster in the United States. The polling methodology emphasizes live phone interviews and representative sampling and focuses on transparency and statistical accuracy to provide a deep understanding of voter sentiment.

A.2 Target Population, Frame, and Sample

- Target Population: Registered voters or likely voters in the United States, especially voters in major battleground states during an election in 2024.
- Sample Frame: U.S. registered voters based on a national voter file maintained by L2, with a cell phone number that matches the voter file.
- Sample Size: Sample sizes vary across different polls. The poll in late October 2024 surveyed over 2,500 voters, and 2,097 of them completed the full survey. The margin of error for likely voters is about ± 2.2 percentage points.

A.3 Sample Recruitment

The sample was conducted by live telephone interviews in both English and Spanish. Over 98% of respondents were contacted by telephone.

A.4 Sampling Approach and Trade-offs

Siena/NYT used a stratified sampling method. Polls were stratified by demographic characteristics such as political party, race and region. This improves accuracy but adds cost and complexity. It also makes it challenging to reach certain populations such as young voters.

Advantages of stratified sampling:

- Generates a representative sample that captures the diverse opinions of different voter populations.
- Helps to ensure that under-represented groups are given appropriate weighting, and reduces potential bias.
- Provides a more accurate reflection of the wider population by stratifying on key characteristics.

Disadvantages of stratified sampling:

- Increased cost and time to complete polls, especially for live telephone interviews.
- There are challenges in reaching specific demographic groups, such as younger voters who may be less likely to participate in the poll.
- Reliance on weighting adjustments to correct for response discrepancies, which does not fully eliminate residual bias.

A.5 Non-response Handling

Non-response was a key problem for Siena/NYT. The poll used weighted adjustments for demographic characteristics such as age, race, gender, education, and different political parties to address non-response bias. The weighting adjustment was used to ensure that the sample closely matches the larger population and compensates for differences in response rates across demographic groups. However, non-response can still present residual bias even with weighting, especially when certain demographic groups have been less likely to respond to survey calls.

A.6 Questionnaire Evaluation

Strengths:

- Minimize bias through the use of straightforward and non-leading questions.
- Transparency was maintained by publishing the questions verbatim to allow for public scrutiny.

- Bilingual interviewers ensured that respondents could participate in their preferred language (English or Spanish), which increased the participation of Spanish-speaking voters and improved the quality of respondents' answers.

Weaknesses:

- The length of the poll can lead to respondent fatigue, which affects the quality of the data and increases dropout rates.

- It can be challenging to limit calls to less than 15 minutes, which can lead to incomplete or rushed responses, thus affecting the overall reliability of the data.

A.7 Conclusion

The Siena/NYT poll uses stratified sampling and live phone interviews to reach a representative sample of voters. While the use of composite weighting and bilingual interviews provides advantages such as transparency and reduced bias, challenges remain including high costs and the potential for bias due to non-response. Despite these drawbacks, Siena/NYT's commitment to transparency and rigorous methodology makes it a reliable source for understanding voter sentiment and opinion.

B Idealized Methodology and Survey

B.1 Overview

The objective of this survey is to predict voter sentiment in the upcoming U.S. presidential election, by using a budget of $100,000 to collect accurate and varied data on voting intentions, candidate favourability, and key issues from a representative sample of registered or likely voters across the United States. The methodology is designed to maximize accuracy, minimize bias and take into account a variety of demographic, geographic and political factors that influence voting behaviour.

B.2 Sampling Approach

The sampling approach is designed to increase coverage and ensure that a representative voting population is covered. This will be achieved by using a combination of stratified random sampling and quota sampling of key demographics.

B.2.1 Stratification Variables

- Age: 18-29, 30-44, 45-64, 65+
- Gender: Male and Female
- Race: White, Black /African American, Hispanic/Latino, Asian, Native and Other
- Education level: High School, Some College, Bachelor's Degree, Graduate Degree, Other
- Geographic region: Northeast, Midwest, South, West
- Political Party: Democratic, Republican, Independent, Other

B.2.2 Sample Size

The target sample size was 10,000 respondents, which are stratified by the Stratification Variables described above. This stratification ensures that all major voter groups are proportionately represented. 10,000 samples have a margin of error of about $\pm 1\%$ and a confidence level of 95%, which is appropriate for predicting elections.

B.3 Recruitment Strategy

The recruitment strategy was separated into two main approaches to ensure a diverse and representative sample:

Online Survey panels:

- Work with established survey panels such as Lucid or YouGov.
- These panels have access to large representative databases of voters.
- Recruiting through panels ensures quality and reliability, as they have built-in validation and quality control processes.

Social Media Advertising:

- Use social media platforms such as Facebook, Instagram and LinkedIn to reach potential respondents.

- Target groups that are underrepresented in traditional panels, including young voters and minority groups.

- Social media ads will be customized with demographic information to increase reach and participation.

Incentives: Offer gift card lottery prizes to all participants to encourage participation and ensure a high response rate.

B.4 Data validation and quality control

To maintain the integrity of the data, we will take the following measures:

- **Attention Checks:** At least two 'attention check' questions will be included in the survey to identify inattention or robots.

- **Duplicate Detection:** Use email or IP-based verification to prevent duplicate responses. Ensure that each respondent provides only one response set to keep the dataset unique.

- **Response Time Monitoring:** Track the time it takes respondents to complete surveys. Responses with apparently faster-than-average completion times are labelled as potentially low-quality.

- **Data Consistency Check:** Analyses the consistency of responses, especially between related questions. Inconsistent answers will be labelled for review.

- **Panel partner validation:** For respondents recruited through an online survey panel, rely on the panel provider's in-built validation systems. These systems use various quality checks, such as identity verification and response consistency, which ensure high-quality data from panel participants.

B.5 Poll aggregation and forecasting

B.5.1 Poll aggregation

- 'Poll of polls' approach: Survey data will be combined with data from some reliable polling sources such as Lucid, YouGov and others. This approach aggregates the results of multiple reputable polls to improve the forecast reliability by reducing bias and individual polling errors.

- Weighting adjustments: Polls are weighted according to sample quality, size, repeatability, and demographic match. Demographic mismatches are adjusted using raking to ensure that the aggregated data accurately represents the U.S. population.

B.5.2 Forecasting Method

- Bayesian modelling: Aggregate polling data is combined with historical electoral trends and demographic information using Bayesian statistical models. This approach allows for dynamic updating as new data becomes available, which improves accuracy.

- Uncertainty Quantification: Includes confidence intervals to express uncertainty in forecasts for a more detailed interpretation.

B.6 Budget Allocation

- Survey Panel Recruitment: $50,000

- Social Media Advertising: $50,000

- Incentives(lottery prizes): $15,000
- Data Cleaning and Validation: $5,000

- Poll Aggregation and Modeling: $10,000

B.7 Survey Implementation

The survey will be available via Google Forms to ensure broad accessibility and easy data collection. The link to the survey will be distributed via email invitations to panellists and targeted social media adverts. To ensure maximum reach and diversity, multiple channels will be used, including Facebook, Instagram, and other social media platforms.

Distribution Plan:

- **Panel invitations:** Members of established survey panels (e.g. Lucid, YouGov) will receive email invitations to participate in the survey.
- **Social media activities:** Targeted adverts will be placed on social media platforms to reach under-represented groups, especially young voters and national minority groups.
- **Follow-up reminders:** An automated reminder email will be sent to participants who have not completed the survey to maximize response rates.
- **Survey Monitoring:** Response rates will be closely monitored throughout the data collection period and sampling methods will be adjusted if certain population quotas are not reached.
- **Accessibility:** The survey will be mobile-friendly, which will allow respondents to complete the survey on any device.

B.8 Survey Structure

C Additional data details

D Model details

D.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

Examining how the model fits, and is affected
by, the data

Figure 2: **?(caption)**

D.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

Checking the convergence of the MCMC algorithm

Figure 3: **?(caption)**

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

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

Ryan Best, Aaron Bycoffe. 2024. "National: President: General Election: 2024 Polls." *FiveThirtyEight*. https://projects.fivethirtyeight.com/polls/president-general/2024/national/.