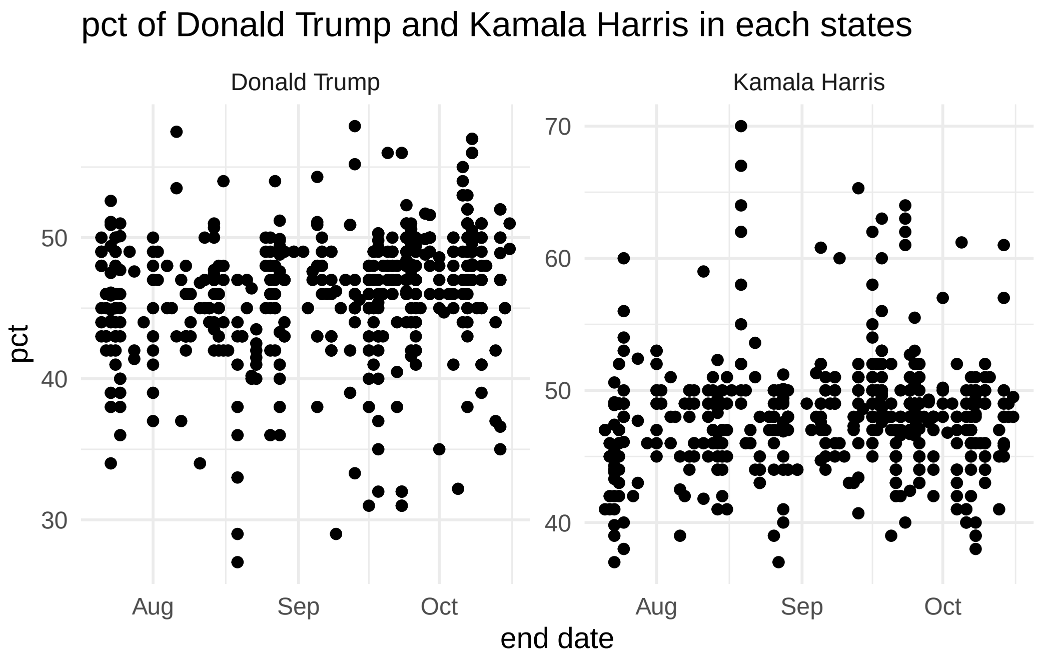
# Forecasting the 2024 US Presidential Election*

## A Poll-of-Polls Approach Using Generalized Linear Modeling

Xinqi Yue Yawen Tan Duanyi Su

November 3, 2024

This paper builds a predictive model to forecast the 2024 US presidential election using a "poll-of-polls" approach, analyzing high-quality polls for Kamala Harris and Donald Trump. The model incorporates data on polling methodology, state trends, and candidate support. Our findings indicate that xxxxxxxx. This study contributes to a deeper understanding of electoral forecasting, helping us interpret aggregated polling data to anticipate election results accurately.



*Code and data are available at: https://github.com/xinqiyue/2024_US_Election_Prediction.

# 1 Introduction

The 2024 U.S. presidential election is generating significant public interest, with recent polling reflecting shifts in candidate support across various demographic and regional groups. Polling data, while insightful, presents challenges due to inherent biases, variability in polling methodology, and regional influences on voter behavior. This study aims to address these challenges using a "poll-of-polls" approach, aggregating high-quality national and state polls to forecast voter support for Kamala Harris and Donald Trump. By applying Bayesian modeling, this analysis incorporates both the temporal dynamics of candidate support and state-specific trends.

In this paper, we explore the following questions: How is support for each candidate likely to evolve as the election approaches? How do regional differences influence overall voter sentiment? This paper contributes to electoral forecasting literature by synthesizing aggregated polling data to provide a nuanced view of candidate support trends.

## 1.1 Estimand

Our primary estimand is the percentage of voter support for Donald Trump and Kamala Harris on election day. We estimate this support percentage both at the national level and by state, incorporating time and state-specific effects.

Results paragraph

Why it matters paragraph

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

# 2 Data

## 2.1 Data Overview

The dataset encompasses polling data focused on two candidates, Donald Trump and Kamala Harris. Data has been filtered to include only results from pollsters with a numeric grade above 2.5 to maintain quality and consistency. This selection allows us to focus on higher-rated sources, which may enhance the reliability of predictions. Each record includes information on the pollster, their grading, polling methodology, transparency score, and specific polling results for each candidate.

The analysis utilizes FiveThirtyEight's dataset of national presidential general election polls (FiveThirtyEight 2024). Following the approach outlined in (Alexander 2023), we aim to predict the election outcome based on this polling data. The analyses were conducted in R R Core Team (2023), with support from several packages. The `tidyverse` packages (Wickham et al. 2019) were used in the process of data simulation, testing beforehand. After the original raw data was downloaded by using `tidyverse` package (Wickham et al. 2019), data cleaning process was done by using `tidyverse` package (Wickham et al. 2019), `lubridate` package (Grolemund and Wickham 2011), and `arrow` package (Richardson et al. 2024). Then, models were constructed using `tidyverse` package (Wickham et al. 2019), `lubridate` package (Grolemund and Wickham 2011), `rstanarm` (Goodrich et al. 2022) package, and `splines` package (R Core Team 2024). The model results are then presented by `modelsummary` (Arel-Bundock 2022) package, and graphs were made with `ggplot2` package (Wickham 2016).

## 2.2 Data Clean

We clean the raw data to produce the analysis dataset using the tools listed in the Section 2.1, providing high-quality data for subsequent election analysis. Firstly, we standardize column names to lowercase and remove special characters. Then, we retain only the fields relevant to the election analysis, such as `state`, `end_date`, `sample_size`, `candidate_name`, and `pct` (percentage support), simplifying the data structure by removing unnecessary columns. Next, we remove rows with missing values and replace empty values in the `state` column with "National," indicating these records pertain to nationwide polls. Additionally, we convert the `end_date` column to date format, ensuring accuracy in date-based filtering. Then, we filter the data to retain only records that meet specific criteria: a `numeric_grade` of 2.5 or higher, a `candidate_name` of either "Kamala Harris" or "Donald Trump" (focusing on these two candidates' support levels), an `end_date` on or after July 21, 2024 (focusing on recent polling when Biden withdrawal from the election), and a non-empty `end_date` (filtering out records with missing dates).

The cleaned data is displayed in Table 1, which contains a total of 5 variables:

Table 1: First 6 entries of Analysis Dataset

| State | End date | Sample Size | Canadidate Name | PCT |
|---|---|---|---|---|
| National | 2024-10-16 | 1000 | Kamala Harris | 49.5 |
| National | 2024-10-16 | 1000 | Donald Trump | 49.2 |
| Arizona | 2024-10-16 | 1435 | Kamala Harris | 48.0 |
| Arizona | 2024-10-16 | 1435 | Donald Trump | 51.0 |
| National | 2024-10-15 | 1457 | Kamala Harris | 48.0 |
| National | 2024-10-15 | 1457 | Donald Trump | 45.0 |

## 2.3 Predictors Explanation

- **state**: The U.S. state where the poll was conducted, which allows for state-by-state analysis and comparison of support levels.

- **end date**: The date the poll concluded, marking the end of data collection for that specific poll.
- **sample size**: The number of respondents in the poll, indicating the scope and potential statistical reliability of the results.
- **candidate_name**: The name of the candidate being polled, which in this data set focuses on Donald Trump and Kamala Harris.

## 2.4 Outcome Exaplanation

- **pct**: The support percentage each candidate received in the poll, which serves as the outcome variable for analysis.

## 2.5 Explanation of Other Variables Used

- **pollster**: The organization that conducted the poll, providing information on who gathered the data.
- **numeric grade**: A quality score assigned to the pollster, with higher values indicating greater reliability. Only pollsters with a score above 2.5 are included to enhance accuracy.
- **pollscore**: A specific rating of the individual poll, reflecting additional factors that impact the poll's quality and reliability.
- **methodology**: The method used by the pollster to conduct the poll, which may affect the reliability and interpretation of the results.
- **transparency score**: A measure of how openly the pollster reports their methodology and results. Higher scores suggest greater transparency and reliability.

## 2.6 Measurement

The goal of this measurement is to turn individual voter opinions into a reliable estimate of electoral college outcomes by analyzing polling data. This data, sourced from FiveThirtyEight (2024)—a trusted platform known for high standards—includes only polls that meet rigorous quality criteria, ensuring broad representation of likely U.S. voters. Polls provide essential details, such as the pollster's identity, survey dates, sample size, and methodology, covering data collection mode, demographic weighting, and adjustments to create a representative sample. These quality standards, while robust, are based on historical practices, which may overlook recent methodological shifts or emerging biases.

## 2.7 Limitation

Polling has inherent limitations. It provides snapshots of voter sentiment at specific times rather than ongoing updates, potentially missing rapid shifts in public opinion, especially near Election Day. Adjustments are made for recency, but participation and response biases—like self-selection and social desirability bias—can still impact accuracy by creating gaps between expressed opinions and actual voting behavior. Additionally, regional polling disparities, with battleground states polled more frequently than "safe" states, can lead to imbalances in representation, highlighting the challenge of achieving uniformly accurate forecasts across the nation.

## 2.8 Data visualization

### 2.8.1 Exploring the Relationship Between State and PCT

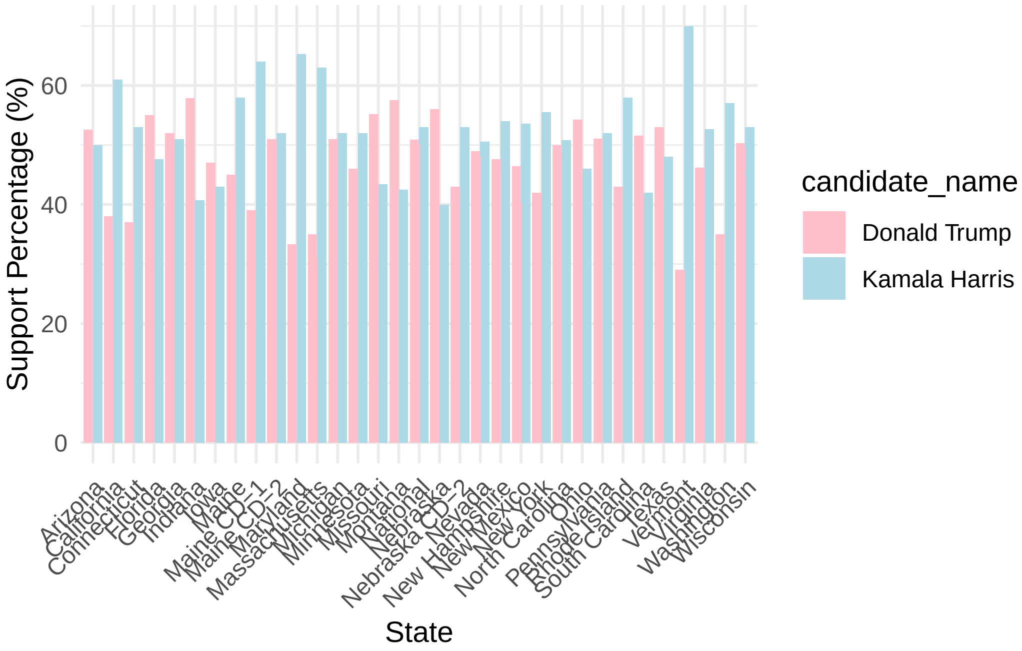


Figure 1: Support for Candidates by State

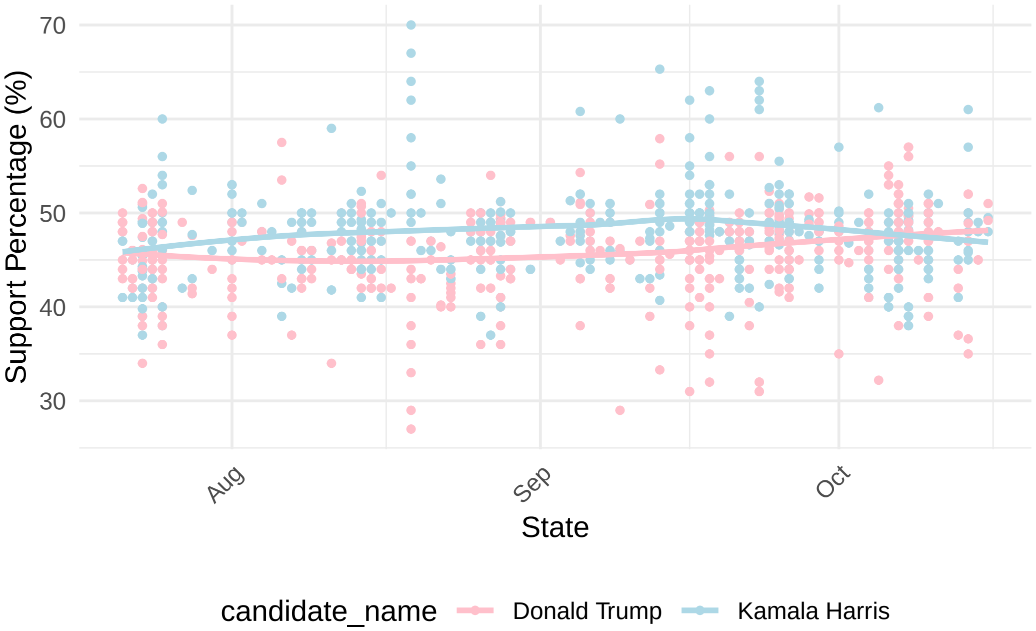


Figure 2: Trends in Candidate Support Over Time

### 2.8.2 Exploring the Relationship Between End Date and PCT

# 3 Model

In this analysis, we use a Bayesian linear regression model to estimate the percentage of support (`pct`) for each candidate in the upcoming US presidential election. The model includes key predictors such as polling end date (`end_date`), candidate (`candidate_name`), sample size (`sample_size`), and state (`state`). Each predictor is selected based on its relevance to the variation in candidate support, as discussed in the data section.

## 3.1 Model set-up

In this Bayesian framework, support level are modeled as normally distributed and influenced by several predictors: end date, candidate name, sample size, and state. The variable $y_i$ denotes the percentage of votes for a candidate in a specific poll, with $\beta_i$ representing the candidate effect, $\gamma_i$ reflecting the influence of sample size, and $\delta_i$ corresponding to the state effect. The intercept $\alpha$ indicates the baseline poll result, while each $\beta_i$ coefficient quantifies the influence of its associated predictor on the vote percentage. Additionally, the variable $\text{state}_j$ captures the effects attributed to different states, and $\text{end\_date}_i$ models temporal trends.

The model can be mathematically expressed as follows:

$$
\begin{align}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) \tag{1}\\
\mu_i &= \alpha + \beta_1 \cdot \text{end\_date}_i + \beta_2 \cdot \text{candidate\_name}_i + \beta_3 \cdot \text{sample\_size}_i + \sum_{j=1}^{M} \gamma_j \cdot \text{state}_j \tag{2}\\
\alpha &\sim \text{Normal}(50, 10) \tag{3}\\
\beta_1, \beta_2, \beta_3 &\sim \text{Normal}(0, 2.5) \tag{4}\\
\gamma_j &\sim \text{Normal}(0, 2.5) \tag{5}\\
\sigma &\sim \text{Exponential}(1) \tag{6}
\end{align}
$$

Priors are defined as follows. For the intercept, a `Normal(50, 10)` distribution is utilized, reflecting a belief that the baseline support percentage is centered around 50% with moderate variability. This choice indicates a moderately informative prior that captures plausible ranges for the intercept. Each regression coefficient is assigned a `Normal(0, 2.5)` prior, which is weakly informative, allowing flexibility in parameter estimation while minimizing bias. These priors are selected to strike a balance between allowing the data to inform estimates and providing regularization to prevent extreme values and overfitting. The chosen priors are grounded in reasonable expectations regarding support ranges, aiming to enhance the robustness of the model.

Then, the complete model, summarizing the components, can be expressed as:

$$
y_i \sim \text{Normal}\left(\alpha + \beta_1 \cdot \text{end\_date}_i + \beta_2 \cdot \text{candidate\_name}_i + \beta_3 \cdot \text{sample\_size}_i + \sum_{j=1}^{M} \gamma_j \cdot \text{state}_j, \sigma^2\right)
$$

To further clarify the model:

- **Formula (1)** shows that the response variable $y_i$ is normally distributed, with $\mu_i$ as the expected value and $\sigma$ representing standard deviation.
- **Formula (2)** illustrates that $\mu_i$ is derived from the intercept $\alpha$ and the predictors.
- **Formula (3)** details the basis functions applied to the state variable, allowing for a nuanced relationship between states and vote percentages. The prior distributions are:
- **Formula (4)** specifies that $\alpha$ has a normal distribution with mean 50 and variance $10^2$, indicating our expectation of average vote percentage.
- **Formula (5)** outlines the priors for $\gamma_k$, which are normal distributions centered at 0 with variance $2.5^2$, reflecting the belief in potentially small effects from state variables.
- **Formula (6)** assigns an exponential distribution to the standard deviation $\sigma$, suggesting that smaller variance is more likely while allowing for broader variations.

## 3.2 Assumptions of the Bayesian Models

1. **Linearity**: The model assumes a linear relationship between independent variables (predictors) and the dependent variable (outcome, `pct`). The expected outcome is represented as a linear combination of the predictors. Nonlinear effects, if present, may not be fully captured, potentially affecting the model's fit.
2. **Normality of Errors**: It is assumed that the residuals (the differences between observed and predicted values) follow a normal distribution. This assumption is crucial for making valid inferences; violations can lead to incorrect conclusions and unreliable credible intervals. Diagnostic checks and posterior predictive checks are used to assess model fit and adherence to assumptions.
3. **Homoscedasticity**: The variance of the residuals is expected to be constant across all combinations of the predictors. If residuals display heteroscedasticity (non-constant variance), it can result in inefficient estimates and biased interpretations of the modeled relationships.
4. **Independence of Errors**: The model assumes that errors are independent of one another, meaning the error term for any single observation does not affect the error term for another observation. However, polling data often includes repeated measures from the same pollster, which may introduce dependence among observations.
5. **Additivity**: Predictors are assumed to additively influence `pct`, meaning interactions or non-additive effects are not considered. Future versions of the model might benefit from testing interactions, such as between state and candidate.
6. **Prior Distributions**: Bayesian regression requires the specification of prior distributions for the model parameters (coefficients). The selection of these priors is significant, as they can impact the resulting posterior distributions, particularly when data is limited.
7. **Parameter Estimation**: In Bayesian linear regression, parameters are estimated using a posterior distribution derived from both the likelihood of the observed data and the prior distributions. This framework allows for uncertainty quantification and more flexible inference.

## 3.3 Software and Validation

The model is implemented using the `rstanarm` package in R, which allows for efficient Bayesian inference with automatic convergence diagnostics. Model fit and convergence were monitored through trace plots and effective sample sizes. To further ensure robustness, we performed posterior predictive checks using `pp_check(model_bayes)` to evaluate model fit and identify any systematic deviations.

Table 2: 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 |

Overall, this Bayesian model provides a balanced approach, capturing key effects while maintaining interpretability and flexibility, making it suitable for predicting candidate support percentages in the election context.

### 3.3.1 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 $\theta$.

## 4 Results

Our results are summarized in Table 2.

# 5 Discussion

## 5.1 Overview of the Paper

This paper presents a comprehensive analysis of polling data to forecast support for candidates Kamala Harris and Donald Trump in the upcoming U.S. presidential election. By employing a Bayesian linear regression model, the study investigates how various factors such as polling date, candidate identity, sample size, and geographic location influence voter support. The findings contribute to understanding the dynamics of electoral support and the effectiveness of polling methodologies in predicting election outcomes.

## 5.2 Insights About the World

One significant insight derived from this analysis is the importance of temporal dynamics in voter support. The model highlights that the timing of polling can substantially impact the reported support for candidates. This finding suggests that voters' preferences are not static; rather, they fluctuate based on current events, campaign strategies, and media coverage. Such insights underscore the need for political strategists and candidates to consider timing when conducting polls and planning campaign efforts.

Another critical takeaway is the varying levels of support for candidates across different states. The analysis reveals regional disparities in voter sentiment, indicating that geographic factors significantly influence political preferences. This highlights the importance of localized campaign strategies and targeted outreach, as what resonates in one state may not hold true in another. Understanding these regional differences can enhance candidates' ability to connect with voters and tailor their messaging effectively.

## 5.3 Limitations of the Study

Despite its contributions, the study is not without weaknesses. One notable limitation is the reliance on existing polling data, which may introduce biases if the data collection methods are not standardized across pollsters. Variations in methodology, sample demographics, and question phrasing can lead to discrepancies in reported support levels. Additionally, the model assumes a linear relationship between predictors and outcomes, which may oversimplify the complexities of voter behavior. A more nuanced approach that incorporates non-linear effects or interactions between variables could provide deeper insights into voter dynamics.

### 5.4 Future Directions

Looking ahead, several avenues for further research emerge from this study. First, it would be valuable to explore the impact of specific events, such as debates or major news stories, on voter sentiment. Incorporating real-time data and sentiment analysis from social media could enrich the understanding of how public opinion shifts in response to external influences. Furthermore, expanding the model to include demographic factors—such as age, income, and education—could yield a more comprehensive picture of voter support and the diverse motivations behind electoral choices.

In conclusion, while this paper offers valuable insights into electoral dynamics, it also highlights the complexities of voter behavior and the need for continued exploration in this field. Future research should strive to incorporate a broader range of variables and methodologies to enhance the predictive accuracy of electoral forecasts and better inform political strategy.

# Appendix

## A Methodology Analysis of Texas Politics Project Poll

### A.1 Introduction

This section provides a comprehensive examination of the methodology utilized by YouGov for the October 2024 Texas Statewide Survey, emphasizing the reliability of the organization and its methods, along with a critical analysis of the survey's strengths, weaknesses, and potential improvements.

### A.2 Population, Frame, and Sample

- **Population**: The survey aimed to represent Texas registered voters, defined as individuals who meet the eligibility requirements to vote in Texas, including age, residency, and registration status. This demographic is critical given the state's political landscape and upcoming elections.
- **Frame**: The sampling frame is the specific population from which a sample is drawn, constructed from diverse data sources to ensure validity. Key components included:
  - **American Community Survey (ACS)**: This provides demographic information that reflects the socio-economic characteristics of Texas residents, essential for establishing a representative sampling frame.
  - **Public voter registration records**: These ensure that the survey captures a valid sample of registered voters, allowing for accurate political representation.
  - **2020 Current Population Survey (CPS)** and **National Election Pool (NEP) exit polls**: These sources offer insights into voting behaviors and preferences, enriching the sampling frame with relevant context and historical data.
- **Sample Size**: The initial sample comprised 1,338 respondents, which was refined to 1,200 for the final dataset through a process known as matching. This process enhances the demographic representativeness of the sample across key characteristics such as gender, age, race, and education.

### A.3 Sample Recruitment

- **Recruitment Methods**: YouGov employs a proprietary opt-in survey panel that includes approximately 1.5 million U.S. residents. Key recruitment methods include:

    - **Web Advertising Campaigns**: Targeted ads based on keyword searches increase the likelihood of engaging relevant respondents, leveraging digital platforms for broad reach.
    - **Permission-Based Email Campaigns**: This method allows for direct outreach to potential participants who have expressed interest in surveys, ensuring a more willing and engaged respondent pool.
    - **Telephone-to-Web and Mail-to-Web Recruitment**: These methods broaden the outreach, ensuring access to respondents who may not engage online, thus increasing the overall sample diversity.

- **Diversity and Representation**: The multifaceted recruitment strategy is designed to mitigate biases and enhance the demographic diversity of the panel. This is essential for capturing a representative sample in a politically diverse state like Texas.

## A.4 Sampling Approach and Trade-offs

- **Two-Stage Sampling Approach**:

    1. **Target Sample Selection**: A probability sample is drawn from the defined target population, ensuring it accurately reflects the demographic composition of Texas. Probability sampling means that each individual in the population has a known, non-zero chance of being selected.
    2. **Matched Sample Selection**: Each target sample member is matched to one or more respondents from the YouGov panel, utilizing a set of variables available in consumer and voter databases. Matching involves finding respondents in the panel who share key demographic and behavioral traits with the target sample.

- **Strengths**: This methodology allows for the creation of a sample that closely mirrors the target population's characteristics, theoretically increasing the external validity of the survey results—external validity refers to the extent to which findings can be generalized to a broader context beyond the study sample.

- **Weaknesses**:

    - **Non-Random Selection**: The reliance on an opt-in panel introduces potential biases associated with self-selection. Respondents who choose to participate may differ significantly from those who do not, potentially affecting the survey's internal validity (the degree to which the survey accurately measures what it intends to measure).
    - **Matching Limitations**: While proximity matching aims to find the closest respondents, it may not account for all variables that could influence responses, leading to unobserved biases. Unobserved biases occur when factors influencing survey responses are not measured or controlled for in the analysis.

## A.5 Handling Non-Response

- **Weighting Strategy**: To address non-response bias, YouGov implemented a robust weighting methodology:
  - **Propensity Score Adjustment**: This involves estimating the likelihood of response based on demographic characteristics, which are then used to weight the matched cases back to the target population. Propensity score weighting is a statistical technique that adjusts for differences in observed characteristics between respondents and non-respondents.
- **Challenges**: The success of this strategy hinges on accurate assumptions regarding non-response characteristics, which can be difficult to ascertain. If the reasons for non-response are systematically related to the survey topics or demographics, this could undermine the validity of the results, highlighting the importance of addressing potential non-response bias.

## A.6 Questionnaire Evaluation

- **Question Design**: The YouGov questionnaire included a series of structured questions aimed at eliciting clear responses regarding voter preferences and opinions on key issues. Structured questions are those that offer fixed response options, facilitating easier analysis.
- **Pros**:
  - **Clarity and Relevance**: Questions were designed to be straightforward, addressing current political concerns likely to engage respondents effectively.
  - **Comprehensive Coverage**: The questionnaire covered a wide array of topics pertinent to the election, providing rich data for analysis and ensuring relevant insights into voter sentiment.
- **Cons**:
  - **Length and Complexity**: The extensive nature of the questionnaire could lead to respondent fatigue, increasing the likelihood of incomplete or rushed responses, which may affect data quality. Respondent fatigue occurs when participants become tired or disengaged from a survey, leading to lower-quality data.
  - **Framing and Bias**: The manner in which questions are framed can introduce bias. Questions that lack neutral wording may lead to skewed responses based on how they are interpreted by participants, potentially distorting the data collected.

### A.7 Recommendations for Improvement

1. **Enhanced Recruitment Strategies**: Consider integrating more diverse recruitment channels, such as partnerships with community organizations, to reach underrepresented groups who may not be captured through standard methods.
2. **Improved Matching Techniques**: Exploring advanced statistical techniques or machine learning approaches for matching respondents could enhance the accuracy of the matched sample.
3. **Shortening the Questionnaire**: Streamlining the questionnaire to focus on the most critical issues could reduce respondent fatigue and improve engagement, ensuring higher quality data collection.
4. **Pilot Testing**: Implementing pilot surveys to test question clarity and structure could help identify potential biases or misunderstandings before full deployment, ensuring that the questions effectively capture the desired information.
5. **Transparency in Weighting Methods**: Clearly communicating the assumptions and methods used in weighting would enhance transparency and allow for a more critical evaluation of the findings, fostering trust in the results.

### A.8 Conclusion

In summary, YouGov's methodology for the October 2024 Texas Statewide Survey demonstrates several strengths, particularly in its diverse recruitment strategies and thorough weighting processes. However, inherent challenges associated with opt-in panels, matching techniques, and questionnaire design warrant careful consideration. By addressing these weaknesses and implementing the suggested improvements, the reliability and validity of future surveys can be enhanced, ultimately providing richer insights into voter behavior and preferences.

## B Additional data details

## C Model details

### C.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…

## C.2 Diagnostics

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

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

## References

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

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

FiveThirtyEight. 2024. "Presidential General Election Polls (Current Cycle)." https://projects.fivethirtyeight.com/polls/data/president_polls.csv.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

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

———. 2024. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

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.