My title*

My subtitle if needed

Jiwon Choi Kevin Roe

November 1, 2024

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

Table of contents

*Code and data are available at: https://github.com/jwonc4602/2024_US_Election_Forecast.

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

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

2 Data

We have used a poll-of-polls of the upcoming US presidential election dataset obtained from FiveThirtyEight (FiveThirtyEight, n.d.). Data was collected and analyzed using R statistical programming software (R Core Team 2023), with additional packages like tidyverse (Wickham et al. 2019), rstanarm (Goodrich et al. 2022) knitr (Xie 2020), here (Müller 2020), and many others for support.

2.1 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

2.2 Outcome variables

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

Some of our data is of penguins (**?@fig-bills**), from Horst, Hill, and Gorman (2020).

Talk more about it.

And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

Talk way more about it.

2.3 Predictor variables

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

3 Model

For our analysis, we employ a Bayesian generalized linear model (GLM) to forecast the popular vote percentage for Kamala Harris. This approach allows us to capture polling characteristics and account for known variations between pollsters, population types, and the recency of polling data. By incorporating these factors, we aim to model a nuanced estimate of Harris's projected popular vote.

The first step in our process involved selecting a reliable dataset for model development. Here, we utilized high-quality national polling data gathered after Harris's campaign announcement. We filtered the dataset to include polls with a numeric grade of 2.7 or above, ensuring data quality, and focused only on polls conducted after July 21, 2024, the date of Harris's declaration of candidacy. This filter helps limit bias from older polls that may not reflect the current voter sentiment.

The GLM is specified as follows:

$$Y_i = \beta_0 + \beta_1 x_{pollster_i} + \beta_2 x_{population_i} + \beta_3 x_{recentpoll_i} + \epsilon_i \tag{1}$$

In equation 1, each β represents a coefficient determined through regression analysis. The variables chosen for this project are pollster, population type, and recency of the poll. Each predictor variable was carefully selected based on its significance in polling analysis and its correlation with voting trends. The identity of the polling organization is critical to our model, as each pollster may exhibit unique biases. Including Pollster as a fixed effect allows us to account for these variations without introducing unnecessary complexity. Population Type distinguishes among different groups surveyed (e.g., voters, likely voters) and helps in capturing generalizability. Categorizing polls as either recent or not ensures that more recent polls, which are better predictors of voting behavior closer to the election, receive appropriate emphasis. Y_i denotes the predicted popular vote percentage for Kamala Harris in the i-th poll. ϵ_i is the Gaussian-distributed error term, accounting for residual variation in the model.

The selection of predictors in this model is based on observable patterns in the data, such as the consistent tendencies of certain pollsters and the impact of population type on polling outcomes. To enhance the model's robustness, Bayesian priors were applied, introducing regularization and incorporating plausible ranges grounded in previous election data and polling analysis. For the coefficient priors β, a normal distribution with a mean of 0 and a scale of 2.5 (autoscaled)

was chosen to provide flexibility while mitigating overfitting. Similarly, the intercept uses a normal prior with a mean of 0 and scale of 2.5 to stabilize model estimates. For the error term (sigma), an exponential prior with a rate of 1 was selected to constrain the residuals, aligning with Gaussian assumptions. These priors offer a balance between model flexibility and constraint, informed by established trends in polling data and comparable electoral forecasts.

The model was implemented in R (R Core Team 2023) using the `rstanarm` package, which offers an accessible interface for Bayesian generalized linear models (GLMs), allowing specification of priors and customization of model parameters. Once the logistic regression model for predicting the popular vote is developed, we apply the `predict()` function in R (R Core Team 2023) to generate popular vote percentage predictions for each candidate, using filtered national data. These predictions are then added to the dataset as a new variable (`predicted_pct`) for further analysis. The results are saved in a CSV file (`popular_vote_predictions.csv`), enabling an in-depth examination of potential outcomes in the popular vote. This approach facilitates understanding of vote distributions at a national level and provides a foundation for forecasting electoral outcomes based on demographic and polling data.

4 Results

To assess model reliability, we examined several key diagnostics. Convergence metrics, such as Rhat values, were close to 1 for all parameters, indicating strong convergence. Additionally, the effective sample size (n_eff) was high across parameters, suggesting low autocorrelation and contributing to model stability. For validation, we conducted out-of-sample testing and calculated the Root Mean Square Error (RMSE) to assess predictive accuracy. This test data, which includes polls not used in training, provides an unbiased estimate of model performance.

The model operates under the assumption that residuals follow a Gaussian distribution, though this may not entirely capture extreme polling variances. Additionally, pollster effects are treated as fixed, which, while limiting the model's ability to reflect dynamic shifts in polling methodologies, simplifies the model and reduces the risk of overfitting. Alternative specifications, including a hierarchical model with random intercepts for pollsters, were considered. However, the added complexity of this approach did not significantly improve predictive accuracy. The chosen GLM specification thus offered a balanced approach, optimizing both interpretability and performance, making it the preferred model for our study.

Table Table 1 presents the estimated coefficients for the predictors in our GLM model. These coefficients fit into the GLM equation, allowing us to interpret the impact of each predictor on Harris's predicted vote percentage. Positive values indicate a higher predicted percentage, while negative values indicate a decrease. Key predictors, such as specific pollsters and population types, show distinct effects on the forecasted outcomes. This table is produced using kable from knitr for enhanced readability.

Table 1: Coefficients from the GLM Model

term	estimate	std.error	conf.low	conf.high
(Intercept)	43.81	1.27	41.78	45.88
pollsterBeacon/Shaw	0.87	1.56	-1.65	3.41
pollsterCES / YouGov	4.76	2.28	1.03	8.56
pollsterCNN/SSRS	-0.33	1.76	-3.22	2.58
pollsterEchelon Insights	1.27	1.55	-1.32	3.84
pollsterEmerson	2.58	1.57	0.06	5.17
pollsterIpsos	0.17	1.12	-1.67	2.01
pollsterMarist	1.14	1.41	-1.15	3.50
pollsterMarquette Law School	1.06	1.42	-1.33	3.44
pollsterMcCourtney Institute/YouGov	0.72	3.07	-4.27	5.76
pollsterMonmouth	-0.08	2.25	-3.78	3.72
pollsterQuinnipiac	-0.38	1.55	-2.97	2.23
pollsterSiena/NYT	-1.49	1.20	-3.50	0.49
pollsterSuffolk	-0.14	1.93	-3.23	3.02
pollsterSurveyUSA	0.80	3.04	-4.24	5.84
pollsterYouGov	0.04	1.09	-1.78	1.84
populationVoters	3.35	0.72	2.16	4.55
recent_pollRecent	0.69	0.46	-0.04	1.42

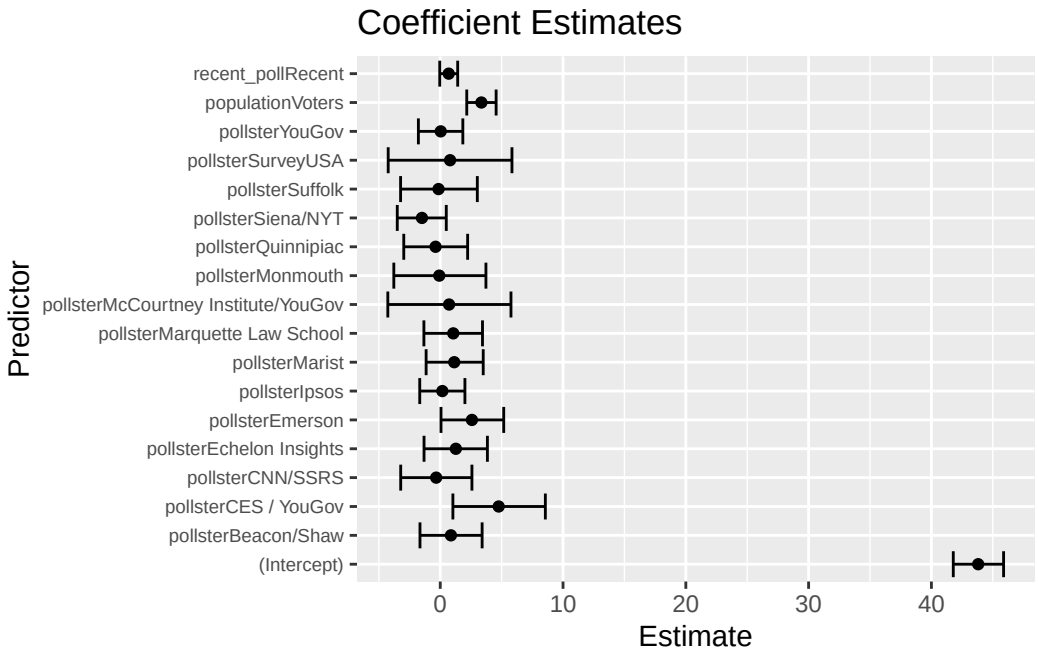

Figure 1: Coefficient Estimates for Predictors

Figure Figure 1 visually represents the model coefficients, with error bars indicating the confidence interval for each estimate. Positive coefficients suggest that specific pollsters or demographic characteristics increase Harris's predicted vote share. For example, the 'CES / YouGov' pollster shows a significant positive impact, while 'Siena/NYT' has a slight negative effect. These error bars help contextualize the reliability of each predictor.

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 Implications

5.5 Weaknesses and Next Steps

Weaknesses and next steps should also be included.

A YouGov Methodology Analysis

YouGov employs a panel-based methodology for conducting online surveys, recruiting participants who voluntarily join their platform. Their surveys span topics like politics, policy opinions, and voter behavior. In the context of election polling, YouGov applies Multilevel Regression and Post-Stratification (MRP) models to predict outcomes by blending respondent data with external data like voter files (YouGov 2024).

The target population for YouGov's surveys typically consists of U.S. adults, with specific samples drawn for each survey depending on the topic. YouGov's sample frame consists of individuals who have opted into their online panel. To ensure representativeness, YouGov sets demographic quotas, stratifying by variables such as age, gender, and political affiliation. Additionally, YouGov monitors and adjusts their sample post-survey using MRP to correct for demographic imbalances and potential sampling bias.

YouGov's panel is recruited through various online methods, including digital advertisements and direct outreach. This recruitment strategy leads to a convenience sample, meaning participants self-select to join the panel, which can introduce selection bias. Although YouGov attempts to offset this bias using quotas and post-stratification weighting, the sample still primarily consists of individuals comfortable with online surveys. This method may underrepresent certain demographics, such as older adults or individuals with limited internet access.

YouGov's use of non-probability sampling—drawing respondents from a convenience sample—provides cost efficiency and quick data collection. However, the trade-off lies in the introduction of potential biases due to self-selection. While YouGov's MRP model corrects some biases by leveraging population-level data for small samples, the non-random nature of the sample may still skew results. For instance, respondents who engage in online surveys may systematically differ from those who avoid them, such as individuals with lower levels of social trust.

Non-response bias is a significant concern in YouGov's methodology. Individuals who opt out of surveys may systematically differ from those who participate. YouGov employs several strategies to mitigate this issue, including **post-survey weighting**, which adjusts the survey data to reflect population characteristics, and **imputation techniques**, which estimate missing responses based on patterns in the available data. Despite these efforts, non-response bias can still affect the accuracy of predictions, particularly when certain demographic groups (e.g., less-educated or low-trust individuals) are less likely to participate.

YouGov places emphasis on clear and concise survey design, reducing participant fatigue and improving data quality. Their surveys generally avoid overly complex or leading questions, which could skew results. However, question wording remains a critical area, as even subtle changes in phrasing can influence responses. For instance, YouGov's research shows how question wording can significantly affect voter perception and reported support for policies.

Additionally, longer surveys can lead to increased dropout rates, potentially introducing non-response bias. To combat this, YouGov often incorporates survey logic, dynamically adjusting questions based on prior responses to keep participants engaged.

YouGov's panel-based methodology, combined with advanced statistical techniques like MRP, offers a robust approach to polling. Their methods effectively address many challenges associated with non-random sampling and non-response bias. However, limitations persist, particularly regarding the biases inherent in self-selection and question framing. Overall, while YouGov's methodology is well-suited to large-scale surveys, particularly in politics, careful consideration of their sampling and non-response strategies is necessary to fully interpret their findings.

B Idealized Methodology

Because the US operates on the Electoral College system, where state-affiliation matters, I would use Stratified Random Sampling to ensure each state is represented proportionally in the survey, improving accuracy and representativeness. My target population would be voter-eligible citizens in the United States who are above 18 years old. My sampling frame will use voter registration databases that are stratified by key demographics (age, race, gender, geographic region, income, and education level). Finally for my potential sample size, I will aim for around 10,000 respondents to ensure a reasonable margin of error, but I expect the actual sample to be less.

I would primarily recruit respondents using online methods such as surveys on social media platforms like Facebook and Instagram to reach younger audiences and use interactive voice response systems on the phone to reach older or less tech-savvy demographics.

To ensure the data aggregated are valid, I would make sure to cross-check responses with voter registration databases and implement various logic checks within the survey to detect inconsistent answers. For example, if a survey respondent claims to have already voted but stated they're unlikely to vote makes the data response unreliable.

The biggest concern with online surveys is to handle non-response bias. To mitigate this, I will make sure to over-sample underrepresented groups and use multiple attempts to contact individuals online or run weekly polls. By running frequent surveys, I will make sure to aggregate results using a moving average in the summary results dashboard to smooth out short-term fluctuations. I will also make sure to develop checks online to prevent someone from filling out the form more than once or track the different people we called to ensure we do not have duplicate entries. Finally, because the United States does not run elections using a popular vote, I will weight the survey results to reflect the US voting population based on US Census data and voter turnout estimates, while also factoring demographic factors such as race, gender, and age.

[NTD: ADD MORE DETAILS ABOUT METHODOLOGY] To allocate the $100,000 budget, I would budget $20,000 for a survey platform subscription to handle a large amount of data. I would put $50,000 into advertising and recruitment for targeted digital ads and phone surveys. I will also use $20,000 for data analysts and poll aggregation services and $10,000 for re-contact surveys, contingency costs, and other miscellaneous costs. However, I believe $100,000 is on the lower end to ensure a high-quality poll and will most likely cost more. Regardless, most of the costs will come from collecting the data, and I have budgeted for this reality (ADD CITATION).

[NTD: ADD THIS AS ANOTHER SECTION] Please find the proposed question list for the online survey attached here (INSERT URL).

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…

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

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…

Checking the convergence of the MCMC algo-
rithm

References

FiveThirtyEight. n.d. "Latest Polls - FiveThirtyEight." https://projects.fivethirtyeight.com/polls/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *Palmerpenguins: Palmer Archipelago (Antarctica) Penguin Data*. https://doi.org/10.5281/zenodo.3960218.

Müller, Kirill. 2020. *here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2020. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

YouGov. 2024. "How YouGov's MRP Model Works for the 2024 Presidential and Congressional Elections." https://today.yougov.com/politics/articles/50587-how-yougov-mrp-model-works-2024-presidential-congressional-elections-polling-methodology.