My title*

My subtitle if needed

First author Yongqi Liu Yuxuan Wei

October 22, 2024

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

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
library(ggplot2)
analysis_data <- read_csv(here::here("data/02-analysis_data/cleaned_US_voting.csv"))
```

```
Rows: 1951 Columns: 10
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (5): pollster_rating_name, methodology, state, candidate_name, populati...
dbl  (3): numeric_grade, sample_size, percent
date (2): start_date, end_date

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

*Code and data are available at: [https://github.com/wyx827/2024USpresidentialelection.git].

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 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023)…. Our data (Toronto Shelter & Support Services 2024)…. Following Alexander (2023), we consider…

Overview text

2.2 Measurement

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

2.3 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.

2.4 Predictor variables

Add graphs, tables and text.

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

3 Model

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

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

3.1 Logistic Regression Model Build up

To predict whether Donald Trump would win over Kamala Harris based on polling data, we constructed a logistic regression model using a subset of 1,000 randomly sampled observations from our cleaned dataset. The dependent variable, trump_win, is a binary outcome representing whether Trump was predicted to win in a given state (1 for Trump win, 0 for Harris win). The independent variables included in the model are pollster_rating_name, state, population_group, numeric_grade, sample_size, and methodology.

We used the logit link function given that our outcome variable is binary. The logistic regression model estimates the log odds of Trump winning as a linear function of these predictors. Each predictor variable was selected based on its relevance to polling outcomes:

- `pollster_rating_name captures the credibility of the polling firm, as rated by independe`
- `state accounts for geographic variations in voter preferences.`
- `population_group represents different demographic categories.`
- `numeric_grade reflects the poll's quality rating.`
- `sample_size measures the size of the poll's sample, which can impact the poll's reliabil`
- `methodology captures how the poll was conducted (e.g., online, phone-based, etc.).`

We applied a Bayesian approach to estimate the logistic regression model using Stan, with weakly informative priors (Normal distribution with mean 0 and standard deviation 2.5) for both the coefficients and the intercept. These priors reflect our assumption that the effect of each predictor is centered around 0, with a moderate level of uncertainty. By incorporating state and demographic information along with pollster quality and methodology, we aim to capture the variability in polling predictions across different regions and polling practices. The use of Bayesian inference allows us to quantify uncertainty in our predictions more effectively.

We use weakly informative Normal priors for the intercept ( ) and coefficients (_i), centered around zero with a standard deviation of 2.5. These priors were chosen to allow the data to influence the predictions while still reflecting reasonable uncertainty. Specifically:

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

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

```
logistic_model <- readRDS(here::here(("models/logistic_model.rds")))
```

4 Diagnostic plots for the MLR model

par(mfrow = c(2, 2)) plot(mlr_harris_model) par(mfrow = c(1, 1)) # Reset plot layout

4.0.1 Model justification

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

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

5 Results

Our results are summarized in …

```
library(rstanarm)
```

```
Loading required package: Rcpp
```

```
This is rstanarm version 2.32.1
```

```
- See https://mc-stan.org/rstanarm/articles/priors for changes to default priors!

- Default priors may change, so it's safest to specify priors, even if equivalent to the defa

- For execution on a local, multicore CPU with excess RAM we recommend calling

  options(mc.cores = parallel::detectCores())
```

```
first_model <-
  readRDS(file = (here::here("models/logistic_model.rds")))
```

6 Discussion

6.1 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.

6.2 Second discussion point

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

6.3 Third discussion point

6.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Pollster Methodology Overview and Evaluation

A.1 Overview of SurveyUSA

SurveyUSA is a privately held opinion research company that operates nationwide, across all 50 U.S. states. Since its founding, the company has conducted over 40,000 research projects, serving a client base of 400 organizations, including media outlets, corporations, non-profits, government agencies, and academic institutions. Known for its expertise in localized opinion research, SurveyUSA focuses on gathering data at the city, county, and regional levels. The company offers timely, cost-effective surveys tailored to meet specific client needs, distinguishing itself from larger global firms.

A.2 Population, Frame, and Sample

- Target Population: U.S. citizens eligible to vote in the 2024 presidential election.
- Sample Frame: U.S. households with either home telephones or access to devices such as phones or tablets.
- Sample Size: Sample sizes vary across different polls. For the 2024 U.S. presidential election cycle, SurveyUSA conducted 49 polls, with sample sizes ranging from 507 to 2,330 for registered voters or likely voters. The average sample size for these polls is approximately 1,045 households.

A.3 Recruitment

SurveyUSA employs a mixed-method approach to recruitment, including online panels, telephone calls, and a text-to-web method. Some respondents are recruited through Random Digit Dialing (RDD) using telephone samples purchased from Aristotle, while others, who do not use home telephones, are invited to complete the survey on an electronic device such as a phone or tablet. Respondents from non-probability online panels are selected randomly by Cint/Lucid Holdings LLC.

A.4 Sampling approach and Trade-offs

SurveyUSA uses a blend of probability and non-probability sampling methods. Some respondents are drawn from non-probability online panels, while others are recruited using probability-based telephone sampling. Responses are weighted based on the latest U.S.

Census estimates for age, gender, ethnicity, and region, ensuring alignment with the target population. Questions and answer choices are rotated to reduce order bias, recency effects, and latency effects.

- **Advantages:**
  The diverse sampling approach not only ensures a broad range of opinions is captured but also complements probability-based sampling, which accurately reflects the overall population. Furthermore, reweighting the data according to U.S. Census demographics strengthens the credibility of the results by ensuring demographic accuracy. Additionally, rotating questions and answer choices helps mitigate bias, further improving the reliability of the data. Finally, the use of online surveys offers a cost-effective solution for efficient data collection.

- **Disadvantages:**
  Phone-based data collection tends to be time-consuming and can be affected by interviewer effects during telephone interviews. Additionally, challenges like non-response issues, such as busy signals or refusals to participate, can hinder the effectiveness of the data collection process.

A.5 Non-response Handling

In cases of non-response, SurveyUSA attempts follow-up calls if interviews are interrupted by answering machines or busy signals. Weighting is applied to adjust for non-response bias, although this doesn't completely eliminate challenges posed by unreachable or unwilling participants.

A.6 Questionnaire Evaluation

- **Positive Aspects:** A logical flow between questions facilitates easy navigation for respondents throughout the survey, while simple wording promotes inclusivity by enabling individuals from diverse backgrounds to comprehend the questions. Furthermore, all questions are directly relevant to analyzing the 2024 U.S. presidential election, and providing predefined response options simplifies the choices for participants.

- **Negative Aspects:** Static options for party affiliation and ideology may fail to capture the nuances of respondents' political beliefs. These rigid categories could oversimplify complex political identities.

A.7 Summary Evaluation

SurveyUSA's methodology reflects a balanced approach, leveraging various sampling approach and method to reach a representative sample. While its blend of probability and non-probability methods has strengths, such as cost-effectiveness and broad reach, it faces challenges related to telephone interview logistics, potential interviewer bias, and the limitations of fixed questionnaire options. Nevertheless, the inclusion of data weighting and question rotation adds credibility to its results, making SurveyUSA a reliable pollster for localized opinion research.

B Appendix B: Idealized Methodology and Survey

B.1 Objective and Overview

The goal of this survey methodology is to accurately forecast the outcome of the U.S. presidential election by collecting high-quality, representative data from a diverse set of respondents across the country. With a budget of $100,000, this methodology incorporates sophisticated sampling techniques, robust respondent recruitment strategies, and rigorous data validation protocols. The approach is designed to maximize accuracy, reduce bias, and account for various demographic, geographic, and political factors that influence voting behavior.

B.1.1 Core Objectives:

1. Obtain a representative sample of the U.S. electorate.
2. Ensure data quality through rigorous validation.
3. Leverage statistical modeling and poll aggregation for an accurate prediction.

B.2 1. Sampling Strategy

The sampling strategy is designed to ensure that the survey reaches a broad, representative section of the voting population. To achieve this, we will use **stratified random sampling** combined with **quota sampling** for key demographics. This ensures that each important subgroup within the population is adequately represented.

B.2.1 Stratification Variables:

- **Age Groups**: 18-29, 30-44, 45-64, 65+
- **Gender**: Male, Female, Non-binary/Other

- **Race/Ethnicity**: White, Black, Hispanic/Latino, Asian, Indigenous, Other
- **Education Level**: No high school, High school graduate, College graduate, Post-graduate
- **Income Bracket**: <$30,000, $30,000-$60,000, $60,000-$100,000, >$100,000
- **Geographic Region**: Northeast, Midwest, South, West

Sample Size:
A total of **10,000 respondents** will be surveyed, providing a margin of error of approximately ±1% at a 95% confidence level. This sample size will allow for detailed subgroup analysis (e.g., by state, demographic group), yielding statistically robust predictions.

Weighting:
We will apply post-stratification weights to adjust for any oversampling or undersampling of specific demographic groups. For example, younger voters or underrepresented minorities will be weighted to reflect their true proportions in the voting population.

B.3 2. Recruitment Strategy

B.3.1 Recruitment Channels:

To maximize respondent diversity and ensure accurate sampling, the survey will employ **multi-channel recruitment**:

- **Digital Advertisements**: Targeted ads on platforms like Facebook, Instagram, and Google will be used to recruit respondents based on their demographic profiles (age, gender, location, political interest). Custom audience features will be utilized to reach specific demographic and geographic groups.
- **Email Outreach**: If permissible, we will access voter registration databases and send email invitations to registered voters. This will allow us to target specific voter demographics that are harder to reach via digital ads.
- **Partnerships with Civic Organizations**: Partnering with non-profits and civic organizations that engage diverse communities (e.g., minority voter outreach programs) will further boost respondent diversity.
- **Incentives**: To increase response rates, each participant will be entered into a lottery with a chance to win a $100 gift card, encouraging broader participation.

B.4 3. Data Validation and Quality Assurance

Maintaining data integrity and ensuring high-quality responses are critical to the accuracy of the election forecast. Therefore, several measures will be put in place to validate responses and reduce noise in the dataset.

B.4.1 Data Validation Protocols:

1. **Real-time Captcha Verification**: This will prevent automated bots from submitting responses.
2. **Email/Phone Verification**: Respondents will verify their email or phone number to ensure authenticity. This ensures that each respondent only participates once.
3. **Time on Task Monitoring**: The survey platform will monitor the time respondents spend on each question. Responses completed suspiciously quickly (e.g., below 30% of the average completion time) will be flagged for review or exclusion.
4. **Voter Registration Cross-Check**: If feasible, respondents will be cross-referenced with voter registration records to ensure they are eligible to vote in the upcoming election.
5. **Response Audits**: Randomly selected respondents will be contacted to verify the accuracy of their responses, ensuring integrity in the dataset.

B.5 4. Poll Aggregation and Data Analysis

B.5.1 Poll Aggregation:

This survey will be combined with results from reputable polling firms (e.g., YouGov, Ipsos, Gallup) to strengthen our forecast through a **poll-of-polls** approach.

- **Weighting by Methodology and Recency**: Poll results will be weighted based on the rigor of their methodology (e.g., online vs. phone surveys, sample size) and the recency of the poll. Recent, methodologically sound polls will receive more weight in the aggregation process.
- **Handling Bias and Variability**: Aggregated results will adjust for pollster house effects (biases in methodology) and variability between polls, ensuring that no single poll dominates the prediction.

B.5.2 Modeling Approach:

We will implement **Bayesian hierarchical models** to account for variability across different states, demographics, and regions. This will allow us to model the popular vote and potentially translate it into **Electoral College predictions**.

B.6 5. Budget Allocation

- **Respondent Recruitment (Targeted ads, outreach)**: $70,000
- **Incentives (e.g., lottery prizes)**: $10,000
- **Survey Platform (Google Forms, Qualtrics subscription)**: $5,000
- **Data Validation Tools**: $5,000
- **Poll Aggregation & Analysis Software**: $10,000

B.7 6. Survey Implementation

The survey will be implemented via **Google Forms**, which offers a cost-effective platform for data collection. The link to the live survey can be found here: Google Form Survey. A copy of the questions is provided below.

B.7.1 Survey Structure:

1. **Introduction**: Thank you for taking part in this survey aimed at predicting the outcome of the 2024 US Presidential election. Your insights are valuable to our research.

Please note: - **All responses will be kept strictly confidential.** - **Your participation is entirely voluntary.** - **We kindly request that you answer all questions honestly and to the best of your knowledge.** - **The survey is estimated to take approximately 10 minutes to complete.** If you have any inquiries or concerns regarding this survey, please don't hesitate to contact the research team at shaw.wei@mail.utoronto.ca.

Your contribution to this study is greatly appreciated! Each participant will be entered into a lottery with a chance to win a $100 gift card!

2. **Section 1: Eligibility Screening**: Are you a U.S. citizen?

- Yes
- No [If No, end survey]

Will you be 18 or older by Election Day (November 5, 2024)? - Yes - No [If No, end survey]

Are you registered to vote in the United States? - Yes - No - Not sure - Plan to register before the election

3. **Section 2: Demographic Information**: What is your age group?

- 18-29
- 30-44
- 45-64
- 65 or older
- Prefer not to say

What is your gender? - Male - Female - Non-binary/Other - Prefer not to say

What is your race/ethnicity? (Select all that apply) - White - Black or African American - Hispanic or Latino - Asian - American Indian or Alaska Native - Native Hawaiian or Pacific Islander - Prefer not to say - Other: [Short text answer]

What is your highest level of education completed? - No high school - High school graduate or equivalent - Some college, no degree - Bachelor's degree - Graduate or professional degree - Prefer not to say

What was your total household income in 2023? - Less than $30,000 - $30,000 - $59,999 - $60,000 - $99,999 - $100,000 - $149,999 - $150,000 or more - Prefer not to say

In which region of the United States do you currently reside? - Northeast (ME, NH, VT, MA, RI, CT, NY, NJ, PA) - Midwest (OH, IN, IL, MI, WI, MN, IA, MO, ND, SD, NE, KS) - South (DE, MD, DC, VA, WV, NC, SC, GA, FL, KY, TN, AL, MS, AR, LA, OK, TX) - West (MT, ID, WY, CO, NM, AZ, UT, NV, WA, OR, CA, AK, HI)

4. **Section 3: Political Views and Voting Intentions**: How likely are you to vote in the 2024 Presidential election?

- Definitely will vote
- Probably will vote
- Might or might not vote
- Probably will not vote
- Definitely will not vote

Generally speaking, do you usually think of yourself as a: - Democrat - Republican - Independent - Prefer not to say - Other: [Short text answer]

If the 2024 Presidential election were held today, who would you vote for? - Kamala Harris (Democrat) - Donald Trump (Republican) - Undecided - Prefer not to say - Other: [Short text answer]

How certain are you about your choice? - Very certain - Somewhat certain - Not very certain - Not at all certain - Prefer not to say

Which THREE issues are most important to you in deciding your vote? (Select exactly three) - Economy and jobs - Healthcare - Immigration - Climate change - National security - Education