Forecasting the 2024 US Presidential Election: A Poll-Based Approach*

My subtitle if needed

Xinxiang Gao Ariel Xing John Zhang

October 21, 2024

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

Table of contents

*Code and data are available at: https://github.com/xgao28/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.

```r
data <- read.csv("../data/02-analysis_data/cleaned_president_polls.csv")
skim(data)
```

Table 1: Data summary

Name	data
Number of rows	6674
Number of columns	29
Column type frequency:	
character	15
logical	5
numeric	9
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
pollster	0	1.00	3	47	0	80	0
sponsors	3254	0.51	3	94	0	146	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
methodology	330	0.95	3	54	0	31	0
state	3055	0.54	4	14	0	50	0
start_date	0	1.00	6	8	0	543	0
end_date	0	1.00	6	8	0	519	0
sponsor_candidate	6664	0.00	10	16	0	4	0
sponsor_candidate_party	6664	0.00	3	3	0	3	0
population	0	1.00	1	2	0	4	0
population_full	0	1.00	1	2	0	4	0
notes	6580	0.01	9	82	0	15	0
partisan	6427	0.04	3	3	0	3	0
party	0	1.00	3	3	0	9	0
answer	0	1.00	4	11	0	46	0
candidate_name	0	1.00	7	25	0	47	0

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
tracking	6674	0.00	NaN	:
internal	5930	0.11	0.01	FAL: 734, TRU: 10
ranked_choice_reallocated	0	1.00	0.00	FAL: 6674
ranked_choice_round	6674	0.00	NaN	:
hypothetical	0	1.00	0.79	TRU: 5244, FAL: 1430

Variable type: numeric

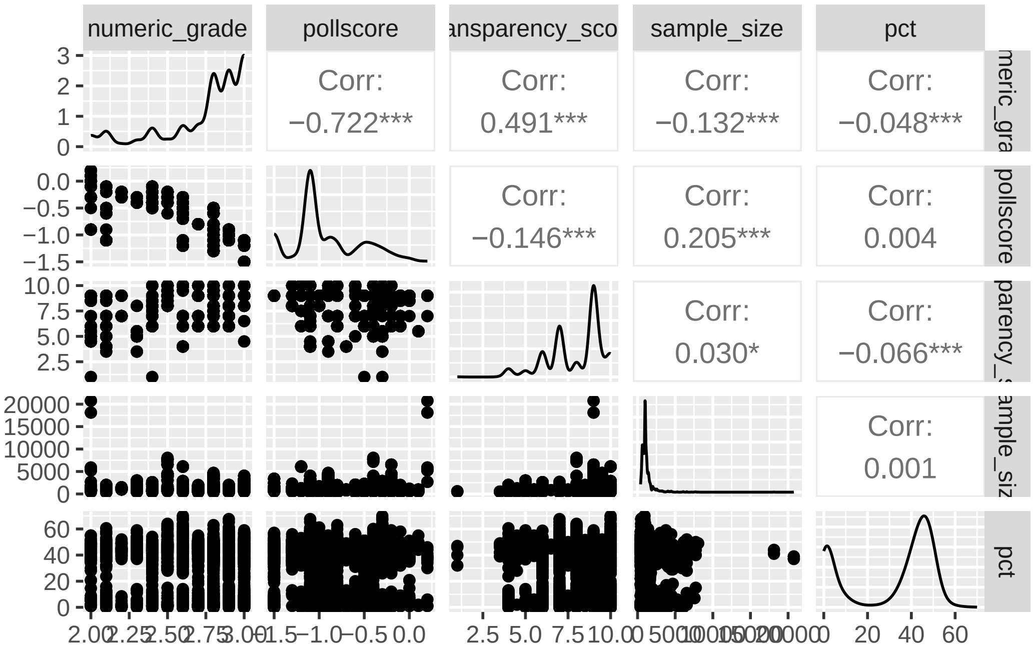
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
numeric_grade	0	1.00	2.75	0.28	2.0	2.7	2.8	3.0	3.0	
pollscore	0	1.00	-0.91	0.37	-1.5	-1.1	-1.1	-0.7	0.2	
transparency_score	1267	0.81	7.97	1.57	1.0	7.0	9.0	9.0	10.0	
question_id	0	1.00	193222.04	16214.43	140769.0	184828.0	197997.0	205623.0	212464.0	
sample_size	62	0.99	1193.65	882.21	320.0	800.0	1004.0	1257.0	20762.0	
source	6583	0.01	538.00	0.00	538.0	538.0	538.0	538.0	538.0	
race_id	0	1.00	8872.96	51.00	8749.0	8839.0	8905.0	8914.0	8914.0	
pct	0	1.00	32.99	18.61	0.0	12.0	42.0	46.0	70.0	
days_towards_election	0	1.00	277.86	252.85	26.0	91.0	191.0	368.0	1299.0	

```r
colnames(data)
```

```
 [1] "pollster"                   "sponsors"
 [3] "numeric_grade"              "pollscore"
 [5] "methodology"                "transparency_score"
 [7] "state"                      "start_date"
 [9] "end_date"                   "sponsor_candidate"
[11] "sponsor_candidate_party"    "question_id"
[13] "sample_size"                "population"
[15] "population_full"            "tracking"
[17] "notes"                      "source"
[19] "internal"                   "partisan"
[21] "race_id"                    "ranked_choice_reallocated"
[23] "ranked_choice_round"        "hypothetical"
[25] "party"                      "answer"
[27] "candidate_name"             "pct"
[29] "days_towards_election"
```

```r
numeric_data <- data %>%
  select(where(is.numeric)) %>%
  select(numeric_grade, pollscore, transparency_score, sample_size, pct) %>%
  filter(complete.cases(.))


# Create a pair plot
ggpairs(numeric_data)
```



```r
model <- data %>%
  filter(numeric_grade >= 2.0, candidate_name == "Kamala Harris") %>%
  lm(pct ~ numeric_grade + days_towards_election, data = .)

# Get the summary of the model
summary(model)
```

```
Call:
lm(formula = pct ~ numeric_grade + days_towards_election, data = .)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2329  -2.2173   0.2222   1.8971  22.3638

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           50.2956083  1.4261392  35.267  < 2e-16 ***
numeric_grade         -0.8399643  0.5184618  -1.620    0.106
days_towards_election -0.0060959  0.0009632  -6.329 4.43e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.134 on 696 degrees of freedom
Multiple R-squared:  0.0581,    Adjusted R-squared:  0.05539
F-statistic: 21.46 on 2 and 696 DF,  p-value: 9.003e-10
```

"pollster"
"sponsors"
"numeric_grade"
"pollscore"
"methodology"
"transparency_score"
"state"
"start_date"
"end_date"
"sponsor_candidate"
"sponsor_candidate_party"
"question_id"
"sample_size"
"population", "population_full", "tracking", "notes"
"source"
"internal"
"partisan"
"race_id"
"ranked_choice_reallocated" "ranked_choice_round"
"hypothetical"
"party"
"answer"
"candidate_name"
"pct"

1.1 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.

2 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 B.

2.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$
\begin{align}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) \tag{1}\\
\mu_i &= \alpha + \beta_i + \gamma_i \tag{2}\\
\alpha &\sim \text{Normal}(0, 2.5) \tag{3}\\
\beta &\sim \text{Normal}(0, 2.5) \tag{4}\\
\gamma &\sim \text{Normal}(0, 2.5) \tag{5}\\
\sigma &\sim \text{Exponential}(1) \tag{6}
\end{align}
$$

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`.

2.1.1 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.

3 Results

Our results are summarized in **?@tbl-modelresults**.

4 Discussion

4.1 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.

4.2 Second discussion point

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

4.3 Third discussion point

4.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Pollster Methodology Overview and Evaluation: YouGov

YouGov is a global public opinion and data company that conducts online surveys on a variety of topics, including politics, social issues, and consumer behavior. Founded in 2000, YouGov is known for leveraging technology to conduct large-scale online surveys, combining traditional sampling principles with advanced data analytics to measure public opinion efficiently.

A.1 Population, Frame, and Sample

- **Population**: The population refers to the group of individuals whose opinions YouGov aims to measure. For political surveys, this often includes eligible voters in a specific country (e.g., registered U.S. voters). Other surveys may focus on specific demographic groups, such as young adults or industry professionals.
- **Frame**: The frame is a list from which the sample is drawn. YouGov uses its online panel, consisting of millions of registered users worldwide. For specific surveys, the frame is the subset of panel members matching desired criteria (e.g., age, location).
- **Sample**: The sample is a subset of the population selected to participate in a survey. Political surveys often involve 1,000-3,000 respondents, weighted to match the demographic characteristics of the broader population.

A.2 Sample Recruitment

- **Recruitment Process**: YouGov recruits panel members via online advertisements, partnerships, and social media. Individuals join the panel by registering on the YouGov website and completing a demographic profile.
- **Incentives**: Panel members earn rewards through a points-based system, which can be redeemed for cash, gift cards, or other benefits.

A.3 Sampling Approach and Trade-offs

- **Sampling Method**: YouGov employs a non-probability sampling approach using quota sampling combined with statistical weighting. Respondents are selected to fill quotas based on demographics (age, gender, education, region) that align with the population.
- **Advantages of Quota Sampling**:

  - **Cost-effective**: Less expensive than random sampling due to online recruitment and automation.
  - **Speed**: Enables quick data collection, crucial for tracking fast-changing opinions.
  - **Targeted Sampling**: Can focus on hard-to-reach populations or specific demographics.

- **Limitations of Quota Sampling**:

  - **Selection Bias**: Self-selection into the panel may introduce biases, as panel members might differ from the general population (e.g., more engaged online).
  - **Generalizability Issues**: Weighting may not fully adjust for attitudinal differences between panelists and the public.

A.4 Handling Non-Response

- **Mitigation Strategies**: YouGov reduces non-response bias with flexible survey completion times and reminder emails. Statistical weighting adjusts for demographic discrepancies caused by non-response.

- **Weighting**: Survey data are weighted to match demographic distributions (e.g., age, gender, race, education). Additional adjustments may be made for political affiliation or past voting behavior.

A.5 Questionnaire Design

- **Strengths**:

  - **Clarity**: Questions are straightforward and easy to understand, reducing measurement error.
  - **Consistency**: Surveys follow a standardized format, ensuring consistency over time, important for tracking changes in opinion.

- **Weaknesses**:

  - **Limited Depth**: Online surveys may feature shorter questionnaires to avoid fatigue, limiting topic depth.
  - **Response Options**: The design of response options (e.g., including “Don’t Know”) can influence results, potentially leading to different conclusions.

A.6 Evaluation of YouGov's Methodology

YouGov's methodology provides several strengths, such as cost, speed, and accessibility, making it suitable for political polling and market research. Online panels enable rapid data collection and targeted sampling. However, the reliance on non-probability sampling introduces biases. While weighting can mitigate some issues, it may not fully compensate for differences between panelists and the general population.

- **Strengths**:
  - **Efficient**: Cost-effective and quick data collection.
  - **Adaptable**: Can rapidly capture opinions on evolving issues.
  - **Targeted**: Capable of reaching niche demographics or regions.
- **Weaknesses**:
  - **Selection Bias**: Potential biases due to non-probability sampling.
  - **Non-Response Bias**: Some groups may be less likely to participate.
  - **Questionnaire Limitations**: Less depth compared to other survey methods.

Overall, YouGov's approach satisfies many standards for modern survey research. Although it has limitations, the insights gained are valuable when these are taken into account.

B Model details

B.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…

B.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

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

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