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

My subtitle if needed

Xinxiang Gao Ariel Xing John Zhang

October 19, 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.

```
data <- read.csv("../data/01-raw_data/president_polls.csv")
skim(data)
```

Table 1: Data summary

Name	data
Number of rows	15423
Number of columns	52
_____	
Column type frequency:	
character	30
logical	6
numeric	16
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
pollster	0	1	3	63	0	219	0
sponsor_ids	0	1	0	38	7994	262	0
sponsors	0	1	0	155	7994	262	0
display_name	0	1	3	125	0	219	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
pollster_rating_name	0	1	3	90	0	202	0
methodology	0	1	0	54	973	50	0
state	0	1	0	14	7386	54	0
start_date	0	1	6	8	0	868	0
end_date	0	1	6	8	0	867	0
sponsor_candidate	0	1	0	21	15094	18	0
sponsor_candidate_party	0	1	0	3	15094	5	0
population	0	1	1	2	0	4	0
population_full	0	1	1	2	0	4	0
created_at	0	1	12	14	0	1911	0
notes	0	1	0	82	15171	29	0
url	0	1	0	277	2	1902	0
url_article	0	1	0	277	3322	1376	0
url_topline	0	1	0	310	9390	810	0
url_crosstab	0	1	0	252	7270	1072	0
internal	0	1	0	5	13081	3	0
partisan	0	1	0	3	14067	5	0
office_type	0	1	14	14	0	1	0
election_date	0	1	7	7	0	1	0
stage	0	1	7	7	0	1	0
nationwide_batch	0	1	5	5	0	1	0
ranked_choice_reallocated	0	1	4	5	0	2	0
hypothetical	0	1	4	5	0	2	0
party	0	1	3	3	0	9	0
answer	0	1	4	11	0	59	0
candidate_name	0	1	7	25	0	60	0

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
endorsed_candidate_id	15423	0.0	NaN	:
endorsed_candidate_name	15423	0.0	NaN	:
endorsed_candidate_party	15423	0.0	NaN	:
subpopulation	15423	0.0	NaN	:
tracking	13923	0.1	1	TRU: 1500
seat_name	15423	0.0	NaN	:

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
poll_id	0	1.00	85868.11	2559.50	74681.0	84533.0	86842.0	87688.0	88665.0	
pollster_id	0	1.00	1224.89	495.50	26.0	1075.0	1302.0	1597.0	1901.0	
pollster_rating_id	0	1.00	353.06	227.27	3.0	195.0	267.0	546.0	870.0	
numeric_grade	1843	0.88	2.17	0.66	0.5	1.8	1.9	2.8	3.0	
pollscore	1829	0.88	-0.38	0.71	-1.5	-1.1	-0.3	0.0	1.7	
transparency_score	3209	0.79	6.28	2.57	0.0	4.0	7.0	9.0	10.0	
sponsor_candidate_id	15094	0.02	28390.16	5845.18	16651.0	31042.0	31042.0	31042.0	37463.0	
question_id	0	1.00	193494.68	8520.37	140691.0	184734.5	198460.0	205407.0	212464.0	
sample_size	132	0.99	1608.92	1807.94	111.0	703.0	1006.0	1563.0	26230.0	
source	15222	0.01	538.00	0.00	538.0	538.0	538.0	538.0	538.0	
race_id	0	1.00	8872.69	52.44	8749.0	8839.0	8905.0	8914.0	8914.0	
cycle	0	1.00	2024.00	0.00	2024.0	2024.0	2024.0	2024.0	2024.0	
seat_number	0	1.00	0.00	0.00	0.0	0.0	0.0	0.0	0.0	
ranked_choice_round	15404	0.00	2.58	1.35	1.0	1.5	2.0	3.5	5.0	
candidate_id	0	1.00	21102.75	6081.51	16638.0	16651.0	19368.0	30966.0	37473.0	
pct	0	1.00	33.70	17.89	0.0	27.0	42.0	46.0	70.0	

```
colnames(data)
```

```
 [1] "poll_id"                   "pollster_id"
 [3] "pollster"                  "sponsor_ids"
 [5] "sponsors"                  "display_name"
 [7] "pollster_rating_id"        "pollster_rating_name"
 [9] "numeric_grade"             "pollscore"
[11] "methodology"               "transparency_score"
[13] "state"                     "start_date"
[15] "end_date"                  "sponsor_candidate_id"
[17] "sponsor_candidate"         "sponsor_candidate_party"
[19] "endorsed_candidate_id"     "endorsed_candidate_name"
[21] "endorsed_candidate_party"  "question_id"
[23] "sample_size"               "population"
[25] "subpopulation"             "population_full"
[27] "tracking"                  "created_at"
[29] "notes"                     "url"
[31] "url_article"               "url_topline"
[33] "url_crosstab"              "source"
[35] "internal"                  "partisan"
[37] "race_id"                   "cycle"
[39] "office_type"               "seat_number"
[41] "seat_name"                 "election_date"
```

```
[43] "stage"                      "nationwide_batch"
[45] "ranked_choice_reallocated" "ranked_choice_round"
[47] "hypothetical"               "party"
[49] "answer"                     "candidate_id"
[51] "candidate_name"             "pct"
```

pollster

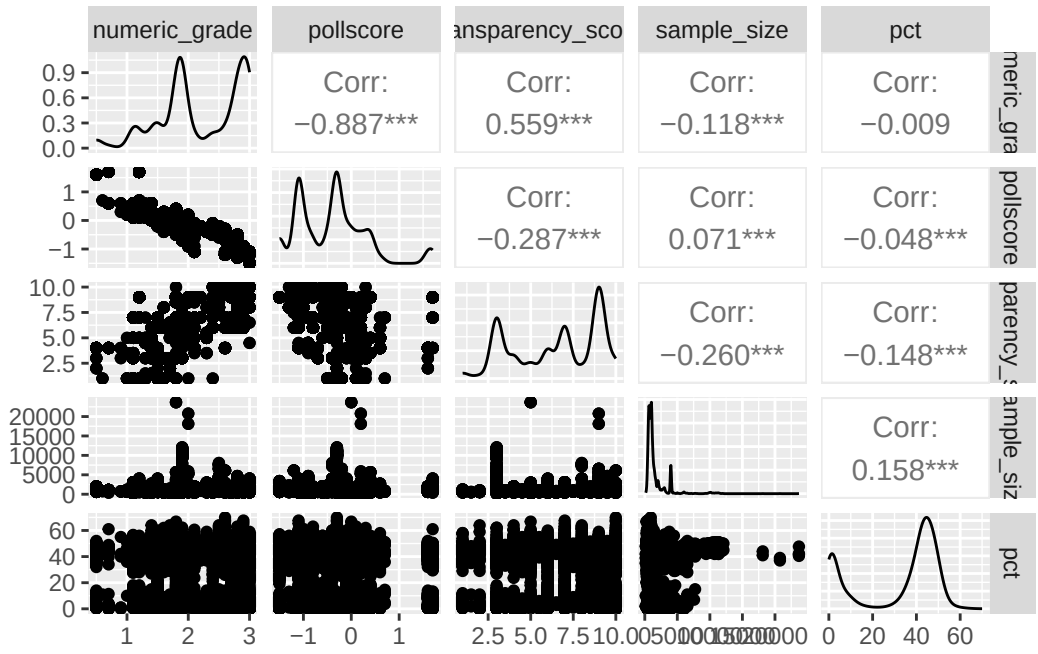
pollscore, numeric_grade

sponsor_candidate, sponsor_candidate_party

sample_size

```
data_a <- data %>%
  select(end_date) %>%
  mutate(end_date = as.Date(end_date, format="%m/%d/%y"))

# Define the election date
election_date <- as.Date("11/5/24", format="%m/%d/%y")

# Calculate the days towards the election date
data_a <- data_a %>%
  mutate(days_towards_election = as.numeric(difftime(election_date, end_date, units = "days")

data$days_towards_election = data_a$days_towards_election
```

```
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)
```




```
data %>%
  select(numeric_grade, pollscore) %>%
  filter(complete.cases(.)) %>%
  cor()
```

```
              numeric_grade  pollscore
numeric_grade     1.0000000 -0.8880981
pollscore        -0.8880981  1.0000000
```

```
sort(unique(data$sample_size)) %>%
  summary()
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  111.0   788.5  1203.0  2245.2  2014.0 26230.0
```

```
data %>%
  filter(numeric_grade >= 2.0) %>%
  skim()
```

Table 5: Data summary

Name	Piped data
Number of rows	6674
Number of columns	53
_____	
Column type frequency:	
character	30
logical	6
numeric	17
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
pollster	0	1	3	47	0	80	0
sponsor_ids	0	1	0	38	3254	147	0
sponsors	0	1	0	94	3254	147	0
display_name	0	1	3	103	0	80	0
pollster_rating_name	0	1	3	90	0	71	0
methodology	0	1	0	54	330	32	0
state	0	1	0	14	3055	51	0
start_date	0	1	6	8	0	543	0
end_date	0	1	6	8	0	519	0
sponsor_candidate	0	1	0	16	6664	5	0
sponsor_candidate_party	0	1	0	3	6664	4	0
population	0	1	1	2	0	4	0
population_full	0	1	1	2	0	4	0
created_at	0	1	12	14	0	971	0
notes	0	1	0	82	6580	16	0
url	0	1	0	277	2	1086	0
url_article	0	1	0	277	1203	819	0
url_topline	0	1	0	310	3684	473	0
url_crosstab	0	1	0	245	2020	712	0
internal	0	1	0	5	5930	3	0
partisan	0	1	0	3	6427	4	0
office_type	0	1	14	14	0	1	0
election_date	0	1	7	7	0	1	0
stage	0	1	7	7	0	1	0
nationwide_batch	0	1	5	5	0	1	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ranked_choice_reallocated	0	1	5	5	0	1	0
hypothetical	0	1	4	5	0	2	0
party	0	1	3	3	0	9	0
answer	0	1	4	11	0	46	0
candidate_name	0	1	7	25	0	47	0

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
endorsed_candidate_id	6674	0	NaN	:
endorsed_candidate_name	6674	0	NaN	:
endorsed_candidate_party	6674	0	NaN	:
subpopulation	6674	0	NaN	:
tracking	6674	0	NaN	:
seat_name	6674	0	NaN	:

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
poll_id	0	1.00	85808.60	2698.47	74706.0	84545.0	86771.0	87719.00	88665.0	
pollster_id	0	1.00	1017.26	480.91	26.0	568.0	1102.0	1424.00	1901.0	
pollster_rating_id	0	1.00	290.13	171.31	3.0	106.0	279.0	407.00	861.0	
numeric_grade	0	1.00	2.75	0.28	2.0	2.7	2.8	3.00	3.0	
pollscore	0	1.00	-0.91	0.37	-1.5	-1.1	-1.1	-0.70	0.2	
transparency_score	1267	0.81	7.97	1.57	1.0	7.0	9.0	9.00	10.0	
sponsor_candidate_id	6674	0.00	32984.00	2874.94	31066.0	31081.0	31302.5	35727.75	37144.0	
question_id	0	1.00	193222.01	16214.43	140769.0	184828.0	197997.0	205623.00	212464.0	
sample_size	62	0.99	1193.65	882.21	320.0	800.0	1004.0	1257.00	20762.0	
source	6583	0.01	538.00	0.00	538.0	538.0	538.0	538.00	538.0	
race_id	0	1.00	8872.96	51.00	8749.0	8839.0	8905.0	8914.00	8914.0	
cycle	0	1.00	2024.00	0.00	2024.0	2024.0	2024.0	2024.00	2024.0	
seat_number	0	1.00	0.00	0.00	0.0	0.0	0.0	0.00	0.0	
ranked_choice_round	6674	0.00	NaN	NA	NA	NA	NA	NA	NA	
candidate_id	0	1.00	21347.98	6209.19	16638.0	16651.0	19368.0	31042.00	37473.0	
pct	0	1.00	32.99	18.61	0.0	12.0	42.0	46.00	70.0	
days_towards_election	0	1.00	277.86	252.85	26.0	91.0	191.0	368.00	1299.0	

1.1 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

1.2 Measurement

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

1.3 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.

1.4 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.

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.

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

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 Additional data details

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

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

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.

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

Toronto Shelter & Support Services. 2024. *Deaths of Shelter Residents*. https://open.toronto.ca/dataset/deaths-of-shelter-residents/.