Forecasting the 2024 US General Election*

Analyzing pollster data

Sakura Hu Luka Tosic

October 30, 2024

This paper uses the poll of polls method to create a general linear model to predict the outcome of the 2024 US general election.We find that current polling predicts _____ as the winner of the election. Third sentence. A limitation of the model is that it does not account for the electoral college in predicting the result, and is thus based only on the popular vote.

1 Introduction

You can and should cross-reference sections and sub-sections. We use R Core Team (2023) and Wickham et al. (2019).

The US presidental election …

#Estimand

Lorem Ipsum # lorem ipsum is just a placeholder in case it's unfamiliar

2 Data

Our data is the Project 538 Presidential General Election polls dataset (current cycle) from October 17th. It contains roughly 16000 observations of polls with ___ unique polls with unque poll id from ___ pollsters with unique pollster id. The methodology of each poll was as an online panel, an app panel, a live phone call, … , and/or a probability panel. Each poll has been given an associated pollscore, numeric grade and transparency score. We also know the duration of the poll by start and end date, as well as the time of publication and the number of particpants. The candidate sponsoring each poll is given as well as the candidate endorsed

*Code and data are available at: https://github.com/LukaTosic09/Forecasting-2024-Election

by each poll. Of course, each poll has a predicted outcome which is captured in the data by the

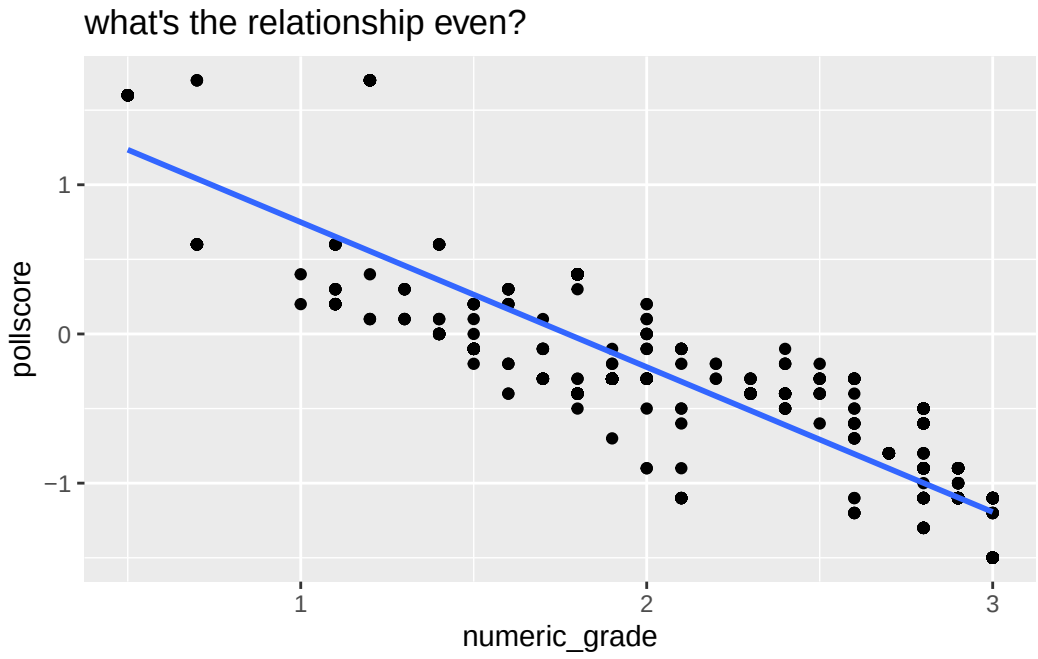
Each poll has an associated pollster, and each pollster is assigned a numeric grade (0.5 to 3.0) and pollscore (-1 to 1) corresponding to the quality and reliability of the pollster and a transparency score (0 to 10) corresponding to the level of transparency the pollster provides in regards to their methodology. Furthemore, each poll has an associated methodology.

Importantly for our modeling, ...

```
ggplot(real_data, aes(x=numeric_grade, y=pollscore)) + geom_point() + geom_smooth(formula = y
```

```
Warning: Removed 211 rows containing non-finite outside the scale range
(`stat_smooth()`).
```

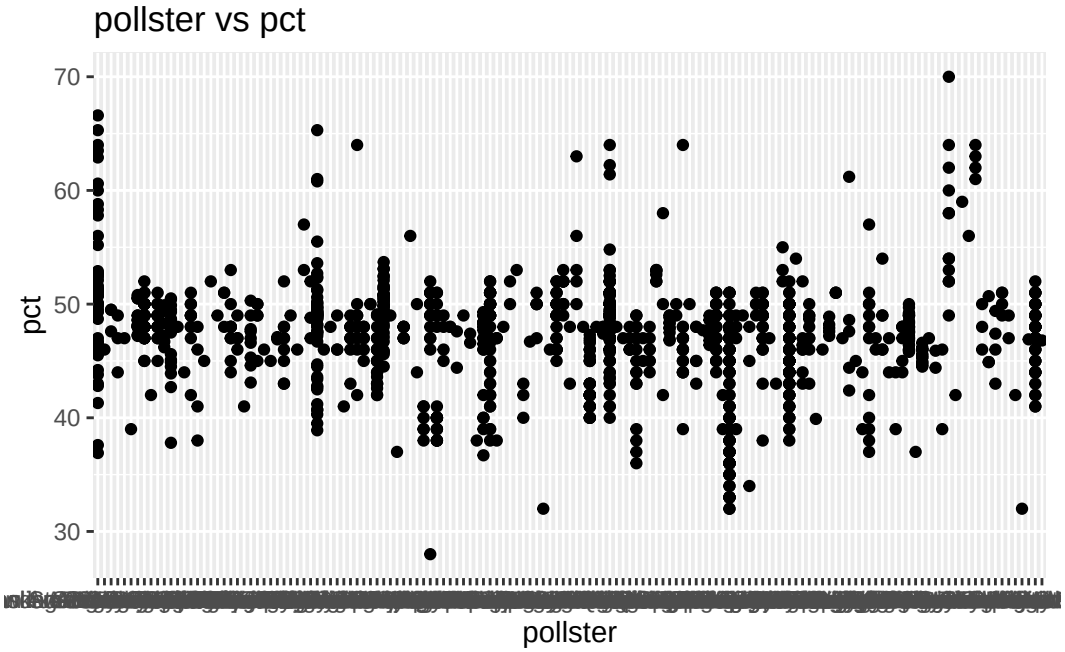
```
Warning: Removed 211 rows containing missing values or values outside the scale range
(`geom_point()`).
```



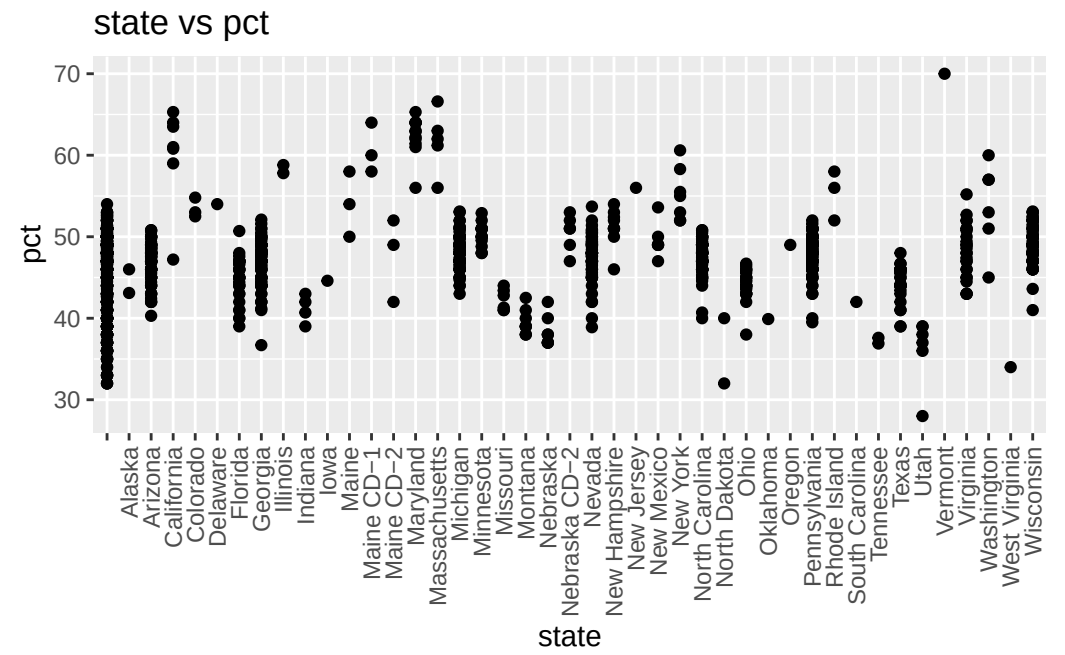
Talk more about it.

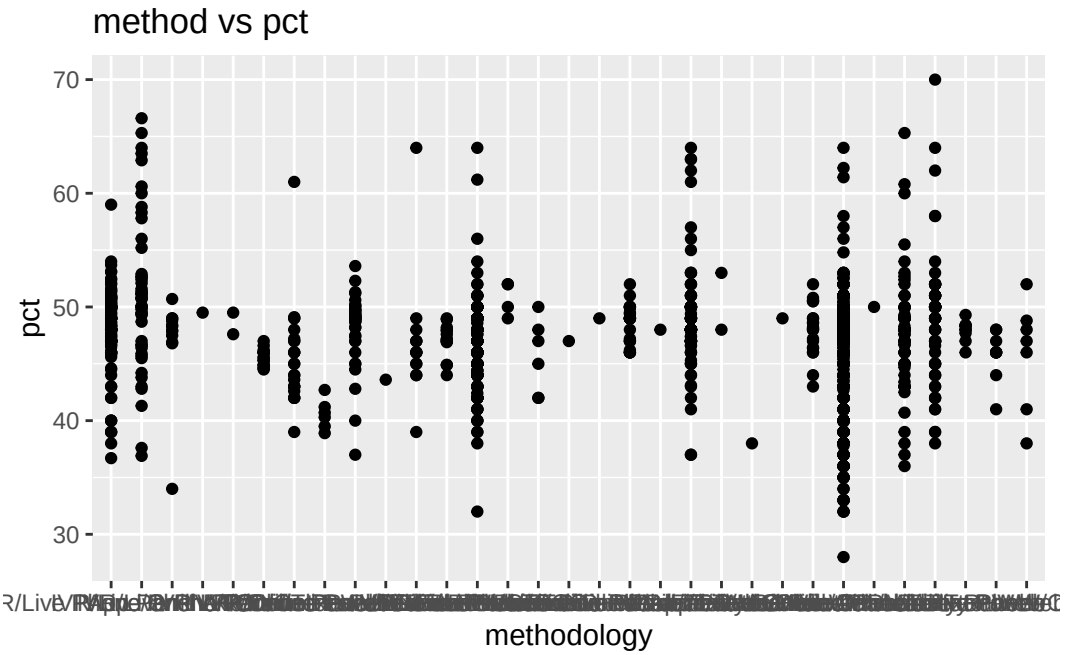
And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

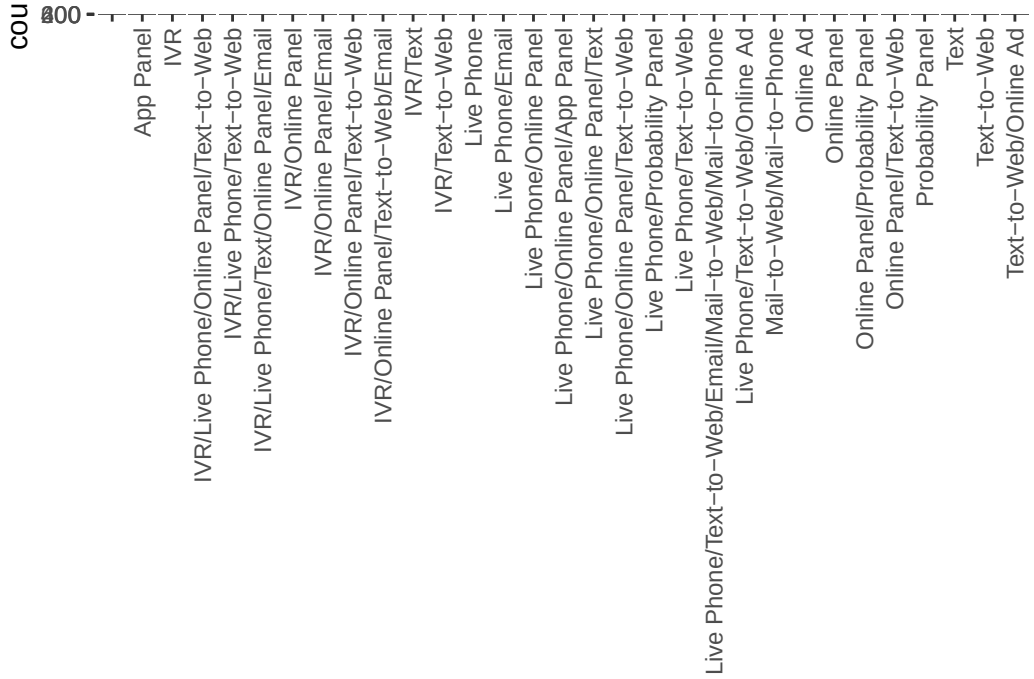
::: {#tbl-predictor vs response plots .cell height='20' width='6'} ::: {.cell-output-display}

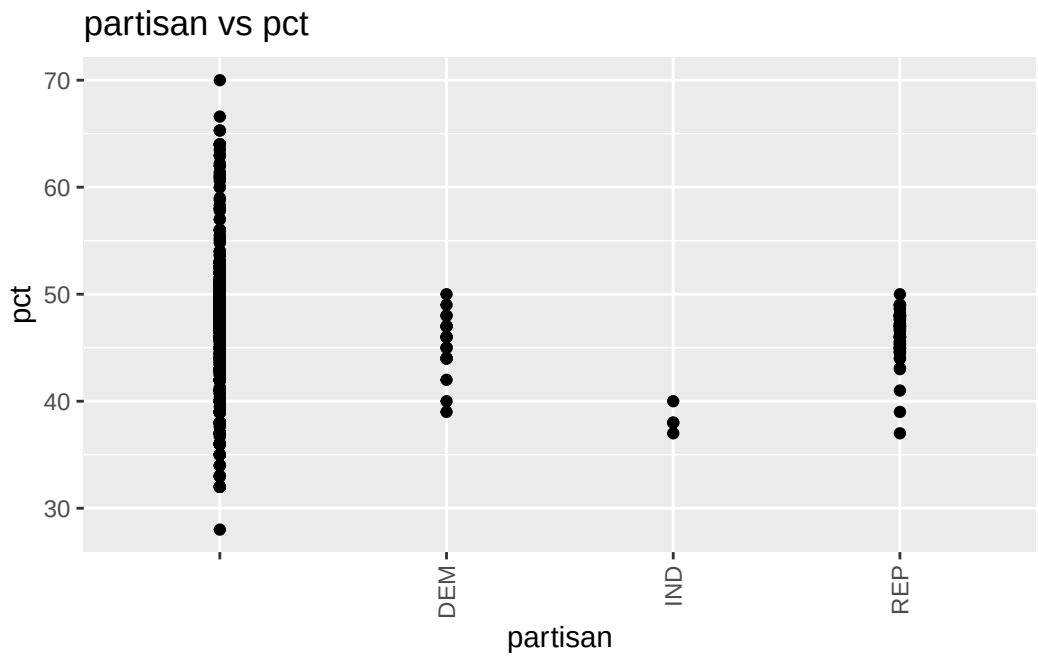


:::



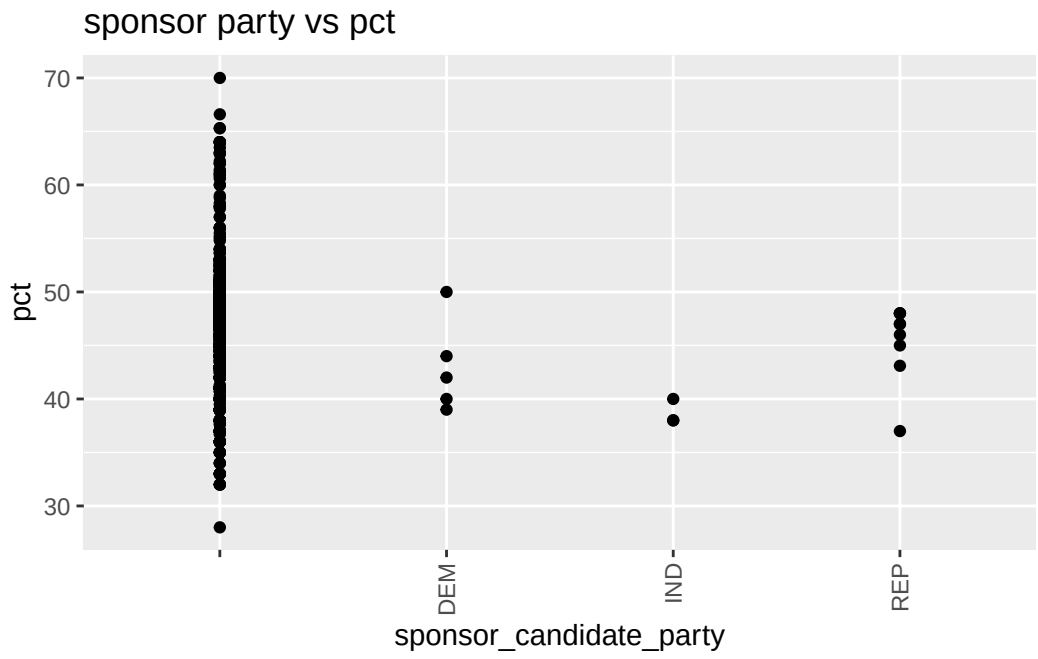

method vs pct
pct
70
60
50
40
30
methodology


cou
App Panel
IVR
IVR/Live Phone/Online Panel/Text−to−Web
IVR/Live Phone/Text−to−Web
IVR/Live Phone/Text/Online Panel/Email
IVR/Online Panel
IVR/Online Panel/Email
IVR/Online Panel/Text−to−Web
IVR/Online Panel/Text−to−Web/Email
IVR/Text
IVR/Text−to−Web
Live Phone
Live Phone/Email
Live Phone/Online Panel
Live Phone/Online Panel/App Panel
Live Phone/Online Panel/Text
Live Phone/Online Panel/Text−to−Web
Live Phone/Probability Panel
Live Phone/Text−to−Web
Live Phone/Text−to−Web/Email/Mail−to−Web/Mail−to−Phone
Live Phone/Text−to−Web/Online Ad
Mail−to−Web/Mail−to−Phone
Online Ad
Online Panel
Online Panel/Probability Panel
Online Panel/Text−to−Web
Probability Panel
Text
Text−to−Web
Text−to−Web/Online Ad



:::

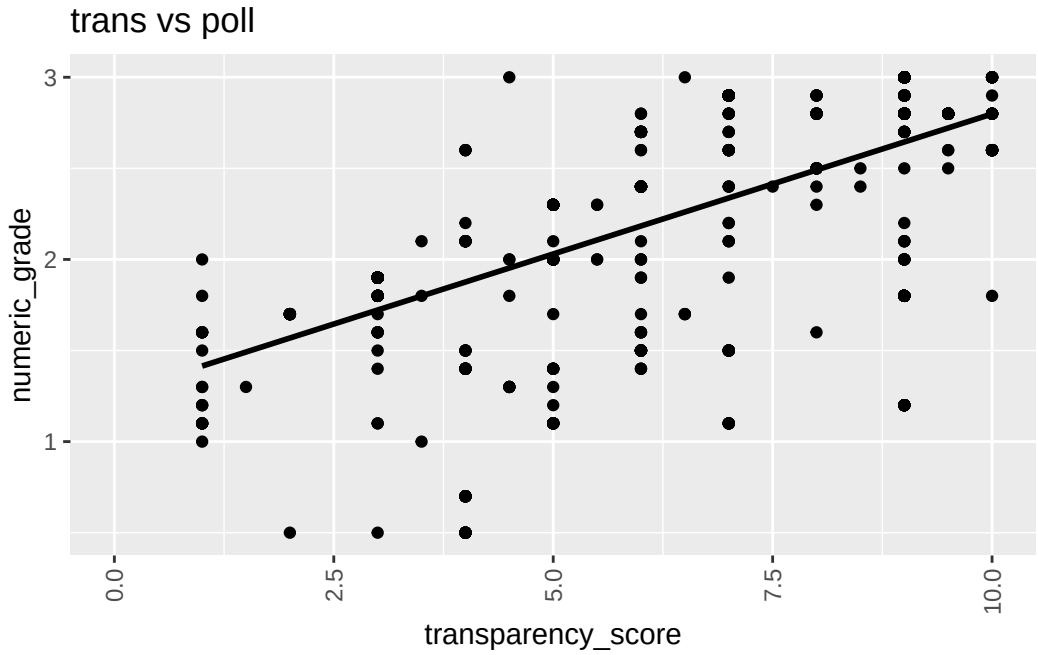
```
ggplot(real_data, aes(x=sponsor_candidate_party, y=pct)) + geom_point() + geom_smooth(formula
```



```
ggplot(real_data, aes(x=transparency_score, y=numeric_grade)) + geom_point() + geom_smooth(f
```

```
Warning: Removed 336 rows containing non-finite outside the scale range
(`stat_smooth()`).
```

```
Warning: Removed 336 rows containing missing values or values outside the scale range
(`geom_point()`).
```




```
test_model <- lm(pct ~ end_date + state, data = quality_data)
test_model
```

```
Call:
lm(formula = pct ~ end_date + state, data = quality_data)

Coefficients:
        (Intercept)     end_date10/11/24     end_date10/15/24
          4.850e+01            2.500e+00            7.051e-15
   end_date10/16/24      end_date10/2/24      end_date10/4/24
         -5.000e-01           -4.950e+00           -1.250e+00
```

```
    end_date10/6/24      end_date10/7/24      end_date2/27/23
         -5.000e-01           -5.000e-01           -6.500e+00
    end_date6/27/22       end_date7/1/24      end_date7/18/24
         -6.000e+00           -3.500e+00           -5.000e-01
    end_date7/22/24      end_date7/24/24       end_date8/1/22
         -2.500e+00           -2.000e+00           -5.000e+00
    end_date8/16/24       end_date8/2/24      end_date8/22/24
          2.500e+00            1.500e+00           -4.500e+00
    end_date8/26/24      end_date8/27/24      end_date8/31/24
         -2.000e+00           -1.500e+00           -4.500e+00
    end_date9/10/24      end_date9/13/24      end_date9/16/24
         -3.500e+00            7.500e-01           -2.750e+00
    end_date9/17/24      end_date9/20/24      end_date9/24/24
          5.000e-01            3.500e+00           -5.000e-01
    end_date9/29/24       end_date9/3/24       end_date9/6/24
         -1.000e+00           -1.500e+00           -2.750e+00
       stateArizona         stateGeorgia        stateMichigan
                 NA            1.000e+00            4.250e+00
stateNorth Carolina             stateOhio    statePennsylvania
                 NA           -3.500e+00            3.250e+00
        stateTexas
                 NA
```

Talk way more about it.

3 Measurement

Lorem Ipsum

4 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 C.

4.1 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`.

4.1.1 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.

5 Results

Our results are summarized in Table 1.

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.

Table 1: Lorem Ipsum

	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

6.2 Second discussion point

6.3 Third discussion point

6.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix 1

Lorem Ipsum

A Appendix 2 {Idealized Survey}

We are provided a budget of $100'000 to forecast the 2024 presidential election. First we must define our key terms. We are looking to forecast the winner of the US general election, so our key parameter of interest for survey participants is the candidate they would support and/or vote for. There are other secondary parameters to discuss later. Our target population is registered voters in the US as they would be the ones informing our parameter of interest, with our sampling frame being registered voters that we can reach with our survey. The sample then is the registered voters who end up taking the survey. We must also define a sample size that is both realistically achievable, and adequately large to make a meaningful prediction of the election winner. A sample size in the range of 2500-6000 is within the realm of possibility and would be sufficient.

#2.1 Sampling

The first task is to develop a sample from a sampling frame through some sample method(s). We would make use of our large budget in order to go about a probabilistic sampling strategy, specifically a stratified sampling approach with additional simple random sampling within the given strata. We would stratify along state, age, education, sex, income, and self-identified political affiliation. The motivation behind these strata is that we know these factors create bias; note that democrats are more likely to answer surveys [], as are those with higher educational attainment []. Additionally, certain battleground states as they are referred to like Georgia or Pennsylvania are important to gather data on as they are considered the most important states in deciding the election []; another reason is that the US population is not split evenly among states, and even in a large poll we would expect only a small, and less reliable amount of responses from certain states with smaller populations.

The specifics of the stratification along the various strata will depend on facts about distribution of population, educational attainment, age, sex, and more among registered voters. Once such statistics are collected likely through census data, we can specify the quantities we hope to see in our sample. ie if we have a sample of size n, and a population of registered voters in a given state of x, we hope to see in our data (x*n)/y where y is the number of registered voters in th US.

#2.2 Survey Method

The survey method we will employ is an online panel of questions in the form of a Google form. These forms are secure and anonymous, and provide a simple way to store the results we receive. Additionally, the format is low cost to produce and takes no additional labour to

maintain unless issues arise, in contrast with a live phone that must be either directly manned or monitored. Another advantage is the transparency provided, users can read and reread instructions, and return to questions they are unsure of easily. Keeping the form short, up to 15 minutes can help stave of disengagement from possible participants.

The maintaining of the survey would take some paid labour. The survey would run for 1 to 2 weeks over the course of 3 months prior to the election as a way to collect data that captures shifting trends in public opinion. Doing this would require us to note when the survey was submitted.

#2.3 Recruitment

Distributing the survey will require multiple things. We must choose how we will distribute the survey. To keep people who are not registered voters from answering, there must be a way for participants to verify that they are registered voters. We propose they register with an email to receive the survey, and they would only receive the survey if their information matches a registered voter in the US, this system would have to be automated and not save sensitive information in order to maintain anonymity. Of course, we'd need to somehow access information on registered voters in the US. Such a system would of course take a portion of our budget to implement

In order to incentivize participation, we could allocate part of the budget to providing a incentive of monetary value for participants. Something like a small giftcard or voucher that is guaranteed upon survey submission.

To reach potential participants, a campaign to send advertise the survey would be launched. We could pay businesses to send a promotional email to users of their service, particularly of services used by demographics we are trying to account for in our strata. A powerful tool for advertising this survey would social media, especially since in the modern day, social media are used by a very wide range of people from the most partisan Republicans and Democrats to the median individual. We can advertise physically as well, more so in states that we are hoping to account for per our stratification.

#2.4 Data Validation

Once the data is collected, it would have to be parsed and cleaned. We would want to know how many questions were unanswered, and to produce basic summary statistics on things like the demographics and state of participants in order to gain an initial understanding of our data. Doing this while the survey is running can tell us if we need to push it further in certain parts of the nation or to certain groups with the hope of amassing more responses in those categories.

As survey cycles continue, we would clean the data to remove non-sensical or uninterpretable responses as well as automated or repeated responses.

The next part of validation is applying stratification weights to ensure our data conforms to the stratification criteria we outlined above. In other words, apply post-stratification weights to

align survey responses with Census-based demographics, using our previously outlined strata as variables. Use iterative proportional fitting (raking) to adjust for sample imbalances.

#2.5 Poll Aggregation and Forecasting

We would aggregate weekly poll results, adjusting for recent responses with higher weights []. Next we would employ Bayesian updating as in our paper to forecast trends based on historical election and polling data, which accounts for typical shifts and volatility in key regions.

We would also have to account for a margin of error and integrate relevant data (e.g., voter registration numbers, early voting patterns) into final our models to refine predictions, particularly in aforementioned battleground states.

#2.6 Survey Example

A complete example survey can be found here: https://forms.gle/tkFYQruLWuwCMLAm9

The questions provided are: Are you currently a registered voter? What is the current state you are registered to vote in? How would you describe your political leanings? Indicate your age in years Indicate your sex What is your highest level of educational attainment? What was your 2023 taxable income amount? (can prefer not to say) If the general election were held tomorrow, what party would you vote for? How Likely are you to vote in this election? If you voted in the 2020 general election, who did you vote for? How well do you feel your most preferred candidate's policies address your concerns and most important issues as a voter What are your most top 3 most important issues when deciding who to vote for? How confident are you in the outcome of this 2024 election?

B Additional data details

C Model details

C.1 Posterior predictive check

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/.

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.