# Forecasting the US Presidential Election: A Poll-of-Polls Approach Using Linear Models*

Ziqi Zhu Yuanchen Miao author3

October 21, 2024

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

## Table of contents



*Code and data are available at: https://github.com/zzq20010617/2024_USelection_prediction

# 1 Introduction

This paper examines the development of a multiple linear regression (MLR) model to predict the percentage of support (pct) for U.S. presidential candidates based on polling data. The data includes a range of predictors, such as sample size, pollster quality factors, and time-related factors. The focus of the analysis is on nationwide polling data, aggregated from different pollsters, to create a robust model for forecasting election outcomes. The goal is to provide a clearer understanding of how various poll attributes influence polling results and to derive insights that can predict election outcomes.

The primary estimand in this study is the expected percentage of support (pct) for a U.S. presidential candidate, given polling data and relevant predictor variables like rating of the pollster and sample size of the poll.

Our Results shows that

This analysis is useful in understanding how various polling factors contribute to predicting election results. By modeling the relationship between polling result and factors, the study enhances the ability to forecast election outcomes based on public opinion data. This model provides a practical tool for researchers, political strategists, and analysts to assess the reliability of polling data and its implications for elections.

# 2 Data

## 2.1 Overview

Our data is download from (FiveThirtyEight, n.d.), the website gathered survey data from different pollsters, We use the statistical programming language R (R Core Team 2023), and

packages (Grolemund and Wickham 2011), (Wickham et al. 2023), (Wickham et al. 2019), to process the data. Following (**tellingstories?**), we consider…

## 2.2 Measurement

The dataset is about public opinions about U.S. presidential candidates, which are captured through polling.

## 2.3 Outcome variables

Our primary outcome variable is **support percentage (pct)**, which represents the percentage of respondents who support each candidate. This variable is modeled as the response variable in the MLR analyses.

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

### 2.4.1 Sample Size

Sample size is an important predictor because it influences the reliability of a poll. Larger sample sizes tend to reduce sampling error, providing more accurate reflections of voter sentiment. In our MLR model, the sample size is log-transformed to account for diminishing returns—larger polls do not necessarily offer proportionally better accuracy.

### 2.4.2 Poll Score

The poll score is a measure of the error and bias we can attribute to a pollster, negative number is better.

### 2.4.3 Numeric Grade

This variable is an aggregate score of the poll based on a numeric scale. It serves as an indicator of the pollster's historical performance and the methodology used. 3 is the maximum and some pollsters have no rating.

#### 2.4.4 Transparency Score

Transparency score measures how openly a pollster reports their methodology and results. A higher transparency score suggests that the pollster has disclosed key details about how the poll was conducted, improving trust in the poll results.

#### 2.4.5 Days Since Start

This variable represents the number of days since the beginning of the election polling period. It helps capture the dynamic nature of voter preferences over time, accounting for shifts in public opinion as the election date nears.

# 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 B.

## 3.1 Model set-up

Define $y_i$ as the number of seconds that the plane remained aloft. Then $\beta_i$ is the wing width and $\gamma_i$ 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 (**rstanarm?**). We use the default priors from `rstanarm`.

Table 1: Explanatory models of flight time based on wing width and wing length

#### 3.1.1 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 $\theta$.

## 4 Results

Our results are summarized in Table 1.

## 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 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…

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

## 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…

Checking the convergence of the MCMC algo-
rithm

## References


FiveThirtyEight. n.d. "FiveThirtyEight Presidential General Election Polls - National (2024)." https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

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.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.