The 2024 U.S. Presidential Election Forecast

Summary of technical model details

Haowei Fan, Fangning Zhang, Shaotong Li

January 10, 2025

Overview

The purpose of this document is to explain the details used in the linear model.This linear model is used to forecast the U.S. 2024 president election by predicting the wining rate outcomes of polls based on several features of polls. Broadly, the model can be described as a Linear hierachical model:

1. A model to capture the association between the outcome (the support or winning rate of candidate) and the input variables (poll reliability score, quality-related factors, and etc)

2.Each independent variables in this model is consider distributed normally, so as the intercept.

Overview of the input values for prediction

The values are selected based on the significant value of variables of each candidate.The significant values of variables are chosen based on the their distribution.

For numeric variables, we chose either the mean or the median as the representative feature value, depending on the data's distribution. If the data distribution was approximately symmetric (low skewness), we used the mean, as it best represents the central tendency

For highly skewed data, we used the median to avoid the influence of outliers, providing a more robust measure of central tendency.

For categorical variables, we selected the mode (the most frequent category) as the feature value. The mode is often the best representative of categorical data, as it reflects the most common category and thus the main trend within the data.

For Boolean variables, we identified whether TRUE or FALSE was more frequent and used the most common value as the feature value. This approach ensures that the feature value represents the predominant condition in the data, rather than relying on a percentage or proportion.

Here are the chosen significant values for Trump and Harris:

Table 1: Candidate Feature Summary

	Candidate Feature Summary	
Variable	Donald Trump	Kamala Harris
numeric_grade	2.15	2.2
pollscore	-0.368	-0.4
methodology	level3	level3
transparency_score	6.17	6.37
sample_size	1003	1000
ranked_choice_reallocated	FALSE	FALSE
hypothetical	TRUE	FALSE
duration	3	3
population	rv	lv

Variables explained and outcomes estimated

These are the independent variables used in the model:

- α_i: Poll reliability score
- β_i: Poll quality score (numeric grade)
- γ_i: Poll transparency score
- δ_i: Poll duration
- θ_i: Sample size
- κ_i: Population type surveyed
- λ_i: Indicator of whether the poll is hypothetical
- μ_i: Poll methodology type
- ρ_i: Rank-choice reallocation usage

This is the dependent variable, or the outcome of this model:

y_i : Winning rate (percentage) for a candidate in the 2024 U.S. presidential election polls

Model

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_1 \times \phi_i + \beta_2 \times \beta_i + \beta_3 \times \gamma_i + \beta_4 \times \delta_i \tag{2}$$
$$+ \beta_5 \times \theta_i + \beta_6 \times \kappa_i + \beta_7 \times \lambda_i + \beta_8 \times \mu_i + \beta_9 \times \rho_i \tag{3}$$
$$\alpha \sim \text{Normal}(0, 10) \tag{4}$$
$$\beta_j \sim \text{Normal}(0, 2.5) \quad \text{for each } j = 1, \ldots, 9 \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

This model combines a normal likelihood with hierarchical priors, creating a Bayesian linear regression framework. The priors on the coefficients and intercept provide regularization, helping to avoid overfitting by constraining the values of α_i and β_j coefficients. The prior on σ helps control the model's uncertainty about the data by penalizing very high variance. This setup allows for flexibility in estimating μ_i while accounting for the effects of the predictors and controlling for the noise in the data.

Hirerachical structure

The hierarchical structure in this model is established through the use of priors on both the intercept α and the coefficients β_j, which represent the relationship between predictors and the outcome variable y_i. Each coefficient β_j has a normal prior centered around zero with a moderate spread (standard deviation of 2.5), allowing for some variability in the effect of each predictor while constraining extreme values. The intercept α also has a prior centered around zero with a larger standard deviation (10), reflecting greater uncertainty in its baseline effect. Additionally, the standard deviation σ, which governs the noise in the data, is given an exponential prior, favoring smaller values to encourage a tighter fit. This hierarchical structure allows the model to borrow strength across predictors, promoting more stable estimates and helping to regularize the model by introducing prior beliefs about the distribution of parameters, thereby preventing overfitting.

Steps of prediction

Load Models and Data

The linear models for both Trump and Harris were loaded using readRDS(). We use ths featured value mentioned before as input values in our model.

Data Preprocessing

The hypothetical column was converted to logical data type where necessary, ensuring data compatibility with prediction models.

Prediction and Visualization