2024 Election Model Methodology

Alex Bass

2024-01-01

Introduction

Last US Presidential Election, I created a model that was a t-test of the 5 most recent quality polls by state. This time, I wanted to apply a more rigorous approach using a heirarchal bayesian linear regression.

Model Choice Justification

T-Test

While in the end, my last model performed decently well, a simple t-test can't control for variables such as survey audience, survey question type, or survey distribution type. So, if you are only performing t-tests, you have to make one of two concessions:

1. use only a fraction of the data (e.g. I will only use likely-voter polls)
2. use all data and try and implement corrections (e.g. all voter polls over estimated Trump by 4 percentage points, so subtract 4 before t-test). However, this sacrifices some statistical rigor since it may be an oversimplification of the effects. Things likely have changed in the last 4 years and the true effect of total population choice and voter choice will vary from poll to poll.

I tried to do the latter last time, but then the question becomes how granular of changes are you going to correct for? For example, If I want to correct for audience, question type, pollster effects, survey mode, etc. the list starts to add up fast. Can I correct for all of these? Also, what if there are omitted variables interacting with these variables and the dependent variable? For instance, maybe registered voters look different in Florida than they do in Maine etc. These questions present challenges for this methodology

Hierarchal Model

This election, I am opting for a hierarchal regression model for a few reasons:

1. Unlike a T-Test, we can naturally control for important variables in our predictions.
2. We can use the natural structure of the data to our advantage. In the electoral college, elections are won by winning states. Each state is unqiue in makeup and attitudes toward candidates and parties. Allowing our model intercept to vary by states helps us account for their differences while utilizing a larger n-size and use information from control variables across states.
3. Predictions can be easily generated for this type of model.

Model Structure

$$TrumpVotePercent = \beta_{state} + \beta_{controls} \cdot X_{controls} + \epsilon$$

- β_{state} represents the intercept term for each state
- $\beta_{controls}$ and $X_{controls}$ represent the control coefficients and data sampled in our model.
- ϵ represents the error term

$$\beta_{var} \sim \mathcal{N}(\mu_{var}, \sigma^2_{var})$$

In the model, I assume each independent variable in our model is normally distributed with a unique μ_{var} and σ_{var}. I use uninformative priors for these variables.

Data Processing and Feature Engineering

Obtaining Predictions