Simplified Model for U.S. Election Polling Analysis*

STA302 Tutorial 10

Talia Fabregas

March 19, 2024

Pollsters often use logistic regression to predict vote choice in upcoming U.S. elections. It is extremely important to start with a sample that is representative of the general electorate. However, polling error is inevitable and different qualified pollsters can produce slightly different results.

1 Introduction

In this paper, I studied the models used by pollsters for the 2016 U.S. presidential election. Inevitably, polling results are dependent equally dependent on pollster judgement and survey methodology (Cohn 2016). In fact, it is possible for two excellent pollsters who use the exact same data set to produce two contrasting predictions (Cohn 2016). There is no clear-cut way to design a survey, choose variables, and build a model to predict vote choice. However, race, sex, and age are widely accepted as the base variables to include when designing a model to predict vote choice (Cohn 2016). This is somewhat expected, because the exploratory data analysis (EDA) process, which is used to gain a sense of a data set, has no clear-cut or one-size-fits-all process (Alexander 2023). There are multiple ways to reasonably analyze a survey data set, however, as I learned when doing my U.S. election forecast project, a model used to predict vote choice can only ever be as good as the survey data set used to fit it (Alexander 2023).

I use R programming language (R Core Team 2023) and the `rstanarm` package to build a simplified version of the model mentioned in the article titled "We Gave Four Good Pollsters the same Raw Data. They Had Four Different Results" (Cohn 2016).

*Code and data are available at: https://github.com/taliafabs/tutorial10.git

2 Data

I downloaded the America's Political Pulse Week 5 2024 data set. This data set is provided by the Polarization Research lab. The America's Political Pulse survey is a weekly survey of 1000 U.S. adults that aims to learn about political polarization, acceptance of democratic norms, and support for political violence in America (Iyengar, Lelkes, and Westwood 2024). This is the survey data that I used for my U.S. election forecast paper, however I have chosen a different week. This survey does not explicitly ask respondents who they plan to vote for in the 2024 U.S. presidential election, but it does ask about political affiliation, ideology, 2016 presidential vote choice, and 2020 presidential vote choice. Using these existing variables, I created a binary indicator variable `vote_biden` that is equal to 1 if the respondent's preferred presidential candidate is Joe Biden, and 0 if their preferred presidential candidate is Trump. I only considered the binary case because I'm using a very simplified version of a vote choice predicting model, or in other words, logistic regression. Logistic regression is only capable of doing binary classification, and America has a two-party system where only Democrats and Republicans win electoral votes and hold seats in Congress and the U.S. Senate.

I was not able to access the data set that was given to Charles Franklin, Patrick Ruffini, Maggie Omero, and Sam Corbett-Davies in the study. So, I used a different version of the survey data set that I used in term paper 3 as well as the same data cleaning steps that I previously used to construct the `vote_biden` variable, to illustrate an extremely simplified version of the models used in the Cohn (2016) article.

3 Model

I built a simple logistic regression model to predict vote choice. This model is fit and estimated using my polling data. The generalized linear model that I chose to use is logistic regression.

3.1 Model set-up

The very simple logistic regression model that I estimated is as follows:

$$Pr(vote\_biden_i = 1) = logit^{-1}(\beta_0 + \beta_1 age_i + \beta_2 gender_i + \beta_3 race_i)$$

3.1.1 Model justification

I opted for a logistic regression model because it is very suitable for analyzing outcomes for various events, including elections and horse racing (Alexander 2023). Logistic regression is appropriate because vote choice in the United States can be considered as a binary outcome, 0 or 1, for the Republican presidential nominee and the Democratic presidential nominee. In

this case, I have used logistic regression to model the probability that each survey respondent will vote for Joe Biden in the 2024 election, using age, gender, and race as predictors. Based on my understanding of the Cohn (2016) article, I would assume that the models built by Charles Franklin, Patrick Ruffini, Maggie Omero, and Sam Corbett-Davies all use regression to predict vote choice, started with age, gender, and race, and ended up producing different results because the additional predictors that they chose vary from pollster to pollster. Poisson and negative binomial regression would not be appropriate for modeling elections or vote choice. Poisson regression focuses on modeling how frequently an event occurs, while negative binomial is a generalized version of poison regression (Alexander 2023).

4 Discussion

Cohn (2016) mentions region, party registration, education, and past turnout as possible predictors that can also be included in the model. Other socio-demographic variables such as whether the respondent lives in an urban or rural area, marital status, and ideology are also reasonable predictors to use. Choice of variables is one reason why the different pollsters' results varied. In general, varying decisions on how to determine who a likely voter is, whether to weight traditionally or using the model, and whether to post-stratify using census or voter-file data can also cause pollsters' results to vary (Cohn 2016). Logistic regression is limited to binary classification. This is generally okay because of America's two-party political system, but it might not be suitable for predicting Canadian election outcomes because Canada has a multi-party system. It is important to note that, while less common than voting for the Democratic or Republican nominee, voting third-party or abstaining are possible voting outcomes for American adults. A more complex, real-life implementation of this might consider using softmax regression, which is a generalized version of logistic regression that can do multi-class classification. If the survey data set is large enough, then softmax regression might be an option to fit a model to classify respondents as Biden, Trump, third-party, or abstaining voters.

References

Alexander, Rohan. 2023. *Telling Stories with Data*. "University of Toronto". https://www.tellingstorieswithdata.com.

Cohn, Nate. 2016. *We Gave Four Good Pollsters the Same Raw Data. They Had Four Different Results.* New York Times. https://www.nytimes.com/interactive/2016/09/20/upshot/the-error-the-polling-world-rarely-talks-about.html.

Iyengar, Shanto, Yphtach Lelkes, and Sean Westwood. 2024. *America's Political Pulse*. https://polarizationresearchlab.org/americas-political-pulse/.

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