My title*

My subtitle if needed

First author Another author

March 12, 2024

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

1 Introduction

The goal of the Cooperative Election Study (CES) is to investigate how Americans vote and perceive their electoral experiences, how they hold their representatives accountable in elections, and how their behavior and experiences vary depending on the political climate and social milieu. The study used an extremely large sample that allowed it to account for variations across different legislative districts. Actually, the sample size at the state level is sufficient to measure the voter preference distribution within the majority of states with a respectable degree of accuracy. 60 teams participated in 2022CES, which resulted in a 60,000-case generic content sample. Fall of 2022 saw the recruitment of study participants. Every research team invested in a 1,000-person nationwide sample survey from Redwood City, California-based YouGov. Two waves of interviewing were used for the 2022 survey. Pre-election surveys were filled out on the spot between September 29 and November 8. The post-election campaigning frenzy took place between November 10 and December 15. For every 1,000 respondents, half of the questions had common material while the other half were fully created and managed by each study team. Questions that are common to all team modules and have a sample size equal to the total sample size of all team modules make up common content. Every one of the sixty teams bought a 1,000-person survey. Every case was chosen online, and YouGov created a corresponding random sample for this investigation. The first data release happened on March 20, 2023. The Harvard University database contains archived and accessible data from this investigation. Vote verification is included in data release 2 for all respondents. The study on CES 2022 is still under progress. Using a large-scale national survey, the Collaborative Congressional Election project was founded in 2006 to investigate congressional elections and representation. The project built upon the analysis conducted by the Public Opinion Research and Training Lab at the Massachusetts Institute of Technology in 2005.

*Code and data are available at: LINK.

2 Data

2.1 Source and Methodology

The 2022 Cooperative Election Study (CES) (Schaffner, Ansolabehere, and Luks (2021)) involved significant data preprocessing, cleaning, and labeling efforts to ensure the quality and representativeness of the survey data. These efforts are elaborated on through descriptions of sampling methodology, matching procedures, and weighting processes as outlined in the CES Guide 2022.

2.1.1 Sampling and Sampling Matching

The CES 2022 dataset is a sample of adult U.S. citizens, taken from a considerably larger population, rather than an exhaustive list of all possible cases in the target population. The sample was deliberately created to be a true reflection of the population of the United States and to capture the variations among the different legislative districts for in-depth examination in the majority of states. Through the Internet, YouGov conducted the CES 2022 survey. 60,000 adults were interviewed between September and November 2022 (pre-election data) and November and December 2022 (post-election data) for the Common Opinion. The YouGov matching approach for random sampling was used. YouGov employed the matched random sampling method at CES among other sampling techniques. The first step in this process is to count the target population, which in the case of general population research consists of all adults. This is typically done by using an extensive survey such as the American Community Survey. The target sample is then selected at random by CES from this target population. However, because it is frequently impractical to make direct contact with these individuals, a matching sample is chosen from among selectable respondents who share the same characteristics as the target sample. This selection process is carried out based on a multitude of attributes available in the voter and consumer databases.

2.1.2 Weightening

A weighting process is used to account for any residual imbalances between the matched samples, therefore rigorously verifying the representativeness of CES samples. The weighting of the sample was adjusted to align with the framework's distribution along several demographic and political aspects. Entropy balancing and iterative proportional fitting (sequencing) of multiple important variables and their interactions comprise the first stage of the weighting process. The sample is also thought to be representative of every state for the survey's common content and takes into account changes for statewide political races. There are two steps involved in selecting samples when utilizing the matching approach. First, select a sample at random from the intended audience. This sample is known as the target sample. The process of choosing a "representative" sample from a non-random sample of responders is known as

sample matching. Although it works well for Web access panels, it may also be applied to other survey kinds, such phone surveys. The target population is enumerated before sample matching begins. All adults are the target group for general population studies, and they can be counted using either a top-notch survey like the American Community Survey or the decennial Census. This is referred to as a sampling frame in other contexts, although unlike traditional sampling, the sample is not taken out of the frame. Using traditional sampling, participants in the study are chosen at random from a framework for sampling. Because not every person in the framework has access to contact information, particularly email addresses, and because declining to participate would raise the expense of sampling in this manner, it could not be practical or cost-effective.

2.1.3 Vote Vaildation

The CES incorporated vote validation to verify the accuracy of respondents' reported voting behaviors. Individual records were matched to the TargetSmart database of registered voters, and only records with a high level of confidence were considered matched. This process helped identify respondents who voted in the 2022 General and Primary Elections, as well as their modes of voting (e.g., absentee, early, mail, etc.). This validation step is crucial for analyses focused on voting behaviors and ensures that the CES data reflects actual voter participation.

These methodologies indicate a comprehensive approach to data preprocessing, cleaning, and labeling within the CES 2022, aiming to enhance the accuracy, reliability, and representativeness of the survey data.

2.2 Variables

```
# A tibble: 60,000 x 4
   votereg presvote20post  race gender4
     <dbl>          <dbl> <dbl>   <dbl>
 1       1              1     1       1
 2       1              1     1       1
 3       1              1     1       2
 4       1              1     1       3
 5       1              6     1       1
 6       1              2     7       1
 7       1              1     2       2
 8       1              1     1       1
 9       1              1     1       2
10       1              1     1       1
# i 59,990 more rows
```

After downloading (**dataverse?**), We using (**get_dataframe_by_name?**) to access the CES. We select and save the data that we are most care about, which are "votereg", "presvote20post", "race", and "gender4". "votereg" represent the voter registration status. To be more specific, whether they are registered to vote or not. "presvote20post" records the president that voter select in the election of 2020. "race" describe the voter's raical or ethnic group, and "gender4" refer to the gender of voters. However, when we access the raw data, there are some variables that we are not interested in. Therefore, we use the codebook to delve into the details. We only care about the voter who are registered to vote, and we are mainly focus on investigating the voter who select Biden or Trump in 2020. We found out that when the "presvote20post" is 1, then this indicates the register vote for Biden, and when it is 2 represents that the voter stand for Trump. We use (**tidyverse?**) to filter to the voters that we care about and label them with meaningful title. In addition CES also provide the information for the gender of the voter. However, we noticed that there are four types of gender, which are man, woman, non-binary, other, and none and we are more interested in the group of man and woman. Therefore, we filter the gender by using (**tidyverse?**). When the variable "gender" is 1, this indicates "man", but we rename all "man" to "male". When the variable "gender" is 2, it refers to "woman". We also rename all "woman" to "female". Finally, the codebook also define different "race" from 1 to 8, which are White, Black, Hispanic, Asian, Native American, Middle Eastern, Two or more races, and other. We keep all those races because we are not only interested in the major racial group, but also curious about the respond from minority group. To better understand the raw data, a summary table had been drawn in order to provide more details for each variables, explaining the variables we select.

2.3 Measurements

The collected dataset from the Cooperative Election Study (CES) offers a comprehensive view of U.S. adult citizens' electoral behaviors, political opinions, and demographic characteristics. The CES 2022 dataset, gathered through an online survey by YouGov, encompasses 60,000 instances, meticulously detailed and categorized by various factors such as election experiences, contextual data, demographic information (e.g., gender, race, education), political affiliations, and opinions on various policy issues. This dataset ensures a reliable representation of the population through its meticulous sampling methodology and validation processes, guaranteeing that it accurately reflects the diverse perspectives within the U.S. electorate. The dataset employs a matched random sampling methodology, with its representativeness further confirmed through a two-stage weighting process involving entropy balancing and iterative proportional fitting. Preprocessing steps, including sample matching, weighting, and vote validation, are conducted to ensure data quality and representativeness, thereby minimizing the risk of representative bias.

To better understand the American election and its voting system that occurred in 2022, we need to follow these steps:

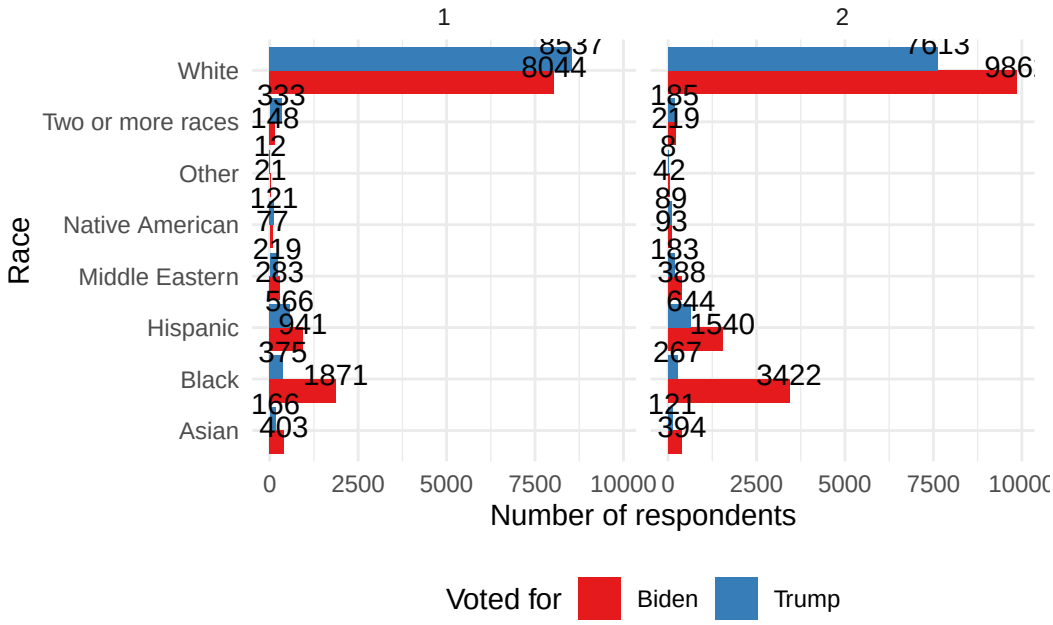

Figure 1: The distribution of presidential preferences, by gender, and race

Define the population: The survey involves 60,000 cases, based on the survey of adult participants in the fall of 2022. Individual respondents participate in the survey, which is designed to record various information related to electoral behavior, political perspectives, and demographic characteristics.

Gather Data: Acquire information regarding the total number of confirmed survey cases. This dataset should include various factors such as, election experiences, contextual data, demographic details (e.g., gender, race, education), political affiliations, and viewpoints on diverse policy matters.

Determine the Time Period: Specify the time frame to calculate when the survey object was collected. The timeframe is relevant to survey participants about their opinions, voting intentions in the 2022 election. The first period is the pre-election data collection that occurred from September 29 to November 8, 2022, and the second period is post- election data gathered from November 10 to December 15, 2022.

Calculate the Case: Calculate each individual's distribution to the survey object. And consider the outcome that will be given with the distributions.

3 Model

3.1 Model set-up

Define y_i is the political preference of the respondent and equal to 1 if Biden and 0 if Trump. Then gender_i is the gender of the respondent and race_i is the race of the respondent.
We could estimate the parameters using `stan_glm()`. Note that the model is a generally accepted short-hand. In practice rstanarm converts categorical variables into a series of indicator variables and there are multiple coefficients estimated. In the interest of run-time we will randomly sample 500 observations and fit the model on that, rather than the full dataset.

$$y_i|\pi_i \sim \text{Bern}(\pi_i) \tag{1}$$
$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \times \text{gender}_i + \beta_2 \times \text{education}_i \tag{2}$$
$$\beta_0 \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta_1 \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\beta_2 \sim \text{Normal}(0.2.5) \tag{5}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022), `here` package of Müller (2020) and `model_summary` package of Arel-Bundock (2022), we use the default priors from `rstanarm`

Through feature selection, the influence of redundant features and noise features is reduced, and the generalization ability of the model is improved. By means of L1 regularization and L2 regularization, the complexity of the model is limited and overfitting is prevented. By integrating the results of multiple classifiers, the accuracy and robustness of the model are improved. By changing the structure of the model, such as increasing the depth of the network, increasing the hidden layer and changing the activation function, the expressibility of the model is improved. By expanding, rotating and scaling the data, the diversity of the data is increased and the generalization ability of the model is improved.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. Logistic regression models learn and predict very quickly Fast, because logistic regression is a linear model, its decision function is a very simple linear function, so its computational complexity is low. Logistic regression model is easy to interpret, and the coefficient of the model can give the degree of influence of each feature on the model. This can be used in real business applications to help decision makers better understand the model's predictions. In addition, logistic regression models can handle large scale data well, because logistic regression models

only need to dimension a small number of parameters to adapt well to a large number of parameters.

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in **?@tbl-modelresults.Logistic** regression is a common classification algorithm used to predict binary classification problems. In logistic regression, the probability that a sample belongs to a certain class is determined by mapping the input data to a probability range between 0 and 1 using a logical function, such as the sigmoid function. When interpreting the results of logistic regression models. Logistic regression models can provide coefficients (or weights) for each feature variable that represent the relative importance of the feature. The positive and negative symbols of the coefficient can tell us whether the feature has positive or negative influence on the target category, and the absolute value of the coefficient can map the importance of the feature to the target category.

5 Discussion

5.1 application of Logistic regression

Logistic regression is also known as generalized linear regression model, and its form is basically the same as that of linear regression model. The biggest difference lies in their different dependent variables. If it is continuous, it is multiple linear regression. If it is a binomial distribution, it is Logistic regression. Although Logistic regression has the name "regression", it is actually a classification method, mainly used for binary classification problems (that is, there are only two outputs, each representing two categories), but also can handle multiple classification problems. Linear regression is used to predict continuous variables, whose value range is (- ", +"), while logistic regression model is used to predict categories. For example, using logistic regression model to predict whether an item belongs to class A or Class B essentially predicts the probability that the item belongs to class A or class B, and the value range of probability is 0~1. Therefore, it is not possible to predict the probability directly with the linear regression equation, which involves the Sigmoid function, which converts the values of the range (-∞, +∞) to the range (0,1).

5.2 linear regression model

In summary, the essence of logistic regression model is to transform the linear regression model Q through a nonlinear Sigmoid function to obtain a probability value between 0 and 1. For binary classification problem (classification 0 and 1), the probability of predicting the

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

	Support Biden
(Intercept)	−0.462
	(0.075)
gender4	−0.388
	(0.019)
raceBlack	−1.030
	(0.082)
raceHispanic	0.354
	(0.076)
raceMiddle Eastern	0.535
	(0.093)
raceNative American	1.245
	(0.118)
raceOther	−0.092
	(0.268)
raceTwo or more races	1.372
	(0.097)
raceWhite	0.940
	(0.070)
Num.Obs.	47 466
R2	0.075
Log.Lik.	−30 135.229
ELPD	−30 144.2
ELPD s.e.	65.9
LOOIC	60 288.5
LOOIC s.e.	131.9
WAIC	60 288.4
RMSE	0.47

classification as 1(or the classification with a larger value in the binary classification) can be calculated using the formula shown below. df has a total of about 7000 groups of historical data, of which about 2000 groups are lost customers, about 5000 groups are not lost customers will "whether to lose" as the target variable, other fields as characteristic variables, through some basic information and transaction records of a customer to predict whether he will lose, or judge the probability of loss.

5.3 model building

The training set is used to train the data and build the model, and the test set is used to check the effect of the model built after training. The purpose of dividing the training set and the test set is to evaluate the model, and to optimize the model through the test set. Dividing the training set and the test set is also, in part, to check for overfitting of the model.

5.4 Weaknesses and next steps

Logistic regression can only deal with linearly separable problems, and for nonlinear separable data, logistic regression will be poor. Logistic regression is sensitive to outliers, and if there are outliers in the data, the effect of logistic regression may be affected. Finally, logistic regression can usually only handle binary classification problems, and for multi-classification problems, some additional processing is required. If the sample size is insufficient or the features are too complex, logistic regression is prone to overfitting.

References

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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

Schaffner, Brian, Stephen Ansolabehere, and Sam Luks. 2021. "Cooperative Election Study Common Content, 2020." Harvard Dataverse. https://doi.org/10.7910/DVN/E9N6PH.