My title*

My subtitle if needed

Hannah Yu

April 11, 2024

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

```
ces2020 <-  read_parquet(here::here("data/analysis_data/cleaned_ces2020.parquet"))
```

1 Introduction

You can and should cross-reference sections and sub-sections. We use R Core Team (2023) and Wickham et al. (2019).

The remainder of this paper is structured as follows. Section **??**….

2 Data

Data analysis is performed using statistical programming language R (R Core Team 2023), with packages `tidyverse` (**citeTidyverse?**), `here` (**citeHere?**), `rstanarm` (**citeRstanarm?**), `modelsummary` (**citeModleSummary?**), `ggplot2` (**citeGgplot2?**), `knitr` (**citeKnitr?**), `marginaleffects` (**citeMarginalEffects?**), `plotly` (**citePlotly?**), `tibble` (**citeTibble?**), `margins` (**citeMargins?**), `testthat` (**citetestthat?**) and `kableExtra` (**citeKableExtra?**).

*Code and data are available at: LINK.

2.1 Data Source

Data for this research comes from the 2020 Cooperative Election Study (CES) (**citeCes2020?**), a annual US political survey. The CES contains information about Americans' political view, voting behaviours and experiences across various political geography and social context. 61,000 American adults participated in the survey in 2020.

2.2 Data Cleaning and Variables

Table 1: Preview of the cleaned 2020 CES dataset

voted_for	ABC	CBS	NBC	CNN	Fox_News	MSNBC	PBS	Other	TV_type	Party
Trump	Yes	Yes	Yes	Yes	Yes	No	No	No	Both	Republican
Biden	Yes	Yes	Yes	No	No	Yes	No	No	Both	Independent
Biden	Yes	No	No	No	No	No	No	No	Both	Independent
Trump	No	No	No	No	Yes	Yes	No	No	Both	Republican
Biden	No	No	Yes	Yes	No	Yes	No	No	Both	Democrat

Table 2: Statistics summary of the cleaned 2020 CES dataset

voted_for	ABC	CBS	NBC	CNN	Fox_News	MSNBC	PBS	Other	TV_type	Party
Trump: 7884	Yes: 7402	Yes: 6650	Yes: 7333	Yes: 8311	Yes: 8384	No :14444	No :18073	No :19222	Local Newscast : 0	Democrat :9819
Biden:13328	No :13810	No :14562	No :13879	No :12901	No :12828	Yes: 6768	Yes: 3139	Yes: 1990	National Newscast: 7422	Republican :5642
NA	NA	NA	NA	NA	NA	NA	NA	NA	Both :13790	Independent:5217
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	Other : 534

Table **??** presentts a summary of the cleaned data, showing detailed statistics about the dataset. Iis evident that Biden has more supporters in this election. The seven listed media networks capture the majority networks people watch; while `Other` captures the rest. Due to the small number of respondents that only watch local newscast,

As we can see from the table, there are more people who support Biden. Respondents covered a wide range of races, with white people being the most heavily represented. Also, CES data came from a variety of locations, but the majority came from the southern region of the United

States. The employment status of those who completed the survey varied, with most of them being full-time or retired.

Since this paper focuses on analyzing the influence of media usage on registered voters decision, I perform the following data cleaning process and selected the related variables. The dataset is cleaned by renaming the column names, changing the variable for from categorical to dummy, selecting the variables of interest, and filtering out missing information and information not related to the study. After cleaning, there are 10331 rows of data remain in the cleaned dataset. Table **??** shows a preview of the cleaned dataset.

The dependent variable of my examination is: `presvote20post`, renamed to `voted_for`. This variable represents the presidential candidate the respondent voted for in the 2020 election in the form of a numerical variable. The variable `votereg` represents whether the respondent is registered to vote using numerical numbers. This paper will only analyze respondents who are registered to vote and focus on the outcome of two candidates, Joe Biden representing the Democratic party and Donald Trump representing the Republican party. To analyze the observations of interest, I first limited the observations to the ones that responded "Yes" in `votereg`. And I filtered the responses in `presvote20post` to only "Biden" or "Trump" and converted the variable into a dummy variable where 1 represents "Biden" and 0 represents "Trump". Rest of the variable this paper focuses on are being divided into three categories: Media Use - Networks, Media Use - TV Type, and Party Affiliation.

Media Use - Networks:

- `CC20_300b_1`, renamed to `ABC`,
  - This variable reports if the respondent watches ABC. A value of 1 signifies that the respondent does watch ABC, while 2 indicates otherwise. This variable was converted into a dummy variable, where 1 represents "Yes" and 0 represents "No".
- `CC20_300b_2`, renamed to `CBS`,
  - This variable reports if the respondent watches CBS. A value of 1 signifies that the respondent does watch CBS, while 2 indicates otherwise. This variable was converted into a dummy variable, where 1 represents "Yes" and 0 represents "No".
- `CC20_300b_3`, renamed to `NBC`; `CC20_300b_4`, renamed to `CNN`; `CC20_300b_5`, renamed to `Fox_News`; `CC20_300b_6`, renamed to `MSNBC`; `CC20_300b_7`, renamed to `PBS`; `CC20_300b_8`, renamed to `Other`,
  - The interpretation and cleaning process of these variables are the same as `CBS` and `ABC`.

Media Use - TV Type:

- `CC20_300a`, renamed to `TV_type`,

  - This variable reports on what kind of TV news the respondent watches. A value of 1 signifies that the respondent watches "Local Newcast", 2 signifies "National Newscast"; 3 "Both".

Party Affiliation:

- `CC20_433a`, renamed to `Party`,
  - This variable reports if the respondent political party affiliation. A value of 1 signifies that the respondent identifies as "Democrat", 2 signifies "Republican"; 3 "Indepdent", and 4 "Other".

Table **??** presentts a summary of the cleaned data, showing detailed statistics about the dataset. Iis evident that Biden has more supporters in this election.

As we can see from the table, there are more people who support Biden. Respondents covered a wide range of races, with white people being the most heavily represented. Also, CES data came from a variety of locations, but the majority came from the southern region of the United States. The employment status of those who completed the survey varied, with most of them being full-time or retired.

2.3 Data Measurement

Because the data was collected from surveys, there might be some inconsistencies and misinterpretations of the questions in people's responses. Therefore this section disects on what the variables are measuring.

each of our variable is only representable for the following scenarios.

Firstly, for the variable "votereg" indicating whether a respondent is registered to vote, the variable relies on self-reported information from survey respondents. Due to the inefficiencies of the US voter registration system, people who misunderstood their voter registration status might falsely reported their condition. For example, some individuals might believe they reached the status simply because they are of age but did not actually register at their local office. According to (**inaccurate?**), there are millions of voter registrations that are no longer valid or inaccurate.

Representing the presidential candidate the respondent voted for, "CC20_410" takes into account of recall bias and social desirability bias along with people's reported presidential candidate preference. Trump had made many controversial speeches throughout his presidency that contributed to his polarizing image and unpopularity in mainstream media. Therefore, many of his voters would conceal their support due to social pressure, potentially leading to under-reporting of votes for Trump in the survey data.

People's media use networks is being represented by if "ABC", "CBS", "NBC", "CNN", "Fox_News", "MSNBC", "PBS", and "Other". also rely on people's self-reported preferences.

For the variable "race," which denotes the racial or ethnic group of the respondent, categorization relies on self-identification. However, racial identity is complex and can be influenced by cultural, social, and historical factors. Mixed-race individuals, for instance, may choose to identify with one racial group over another based on social preferences or personal experiences. Moreover, the race variable includes a category for "two or more races" to classify individuals identifying with multiple racial backgrounds. While this acknowledges the diversity within the population, it may not capture the nuances of each individual's identity. Additionally, the absence of a distinct category for "Indians or South Asians" presents a limitation. While these individuals may technically fall under the category "Asian," many Indian and South Asian Americans may prefer to identify separately due to differences in appearance and cultural background. The term "Asians" are more generally associated with East Asians such as Chinese, Japanese, and Koreans. As a result, there's a potential for misclassification, with some individuals opting to report themselves as belonging to the "other" race category instead.

In the case of "region," which represents the census region where the respondent lives, data collection relies on categorization based on geographical location. However, it's important to note that regional boundaries can sometimes be arbitrary and may not fully capture the diverse cultural, economic, and social affiliations of individuals. Additionally, people living in bordering states or areas near regional boundaries may have affiliations with multiple regions or may identify more strongly with a neighboring region. For example, states like Missouri and Kentucky are often considered part of the Midwest, but they also share cultural and economic ties with the South. This complexity can lead to challenges in accurately categorizing individuals based on their region of residence, potentially resulting in oversimplification or misrepresentation of regional identities and characteristics.

The "region" variable categorizes respondents based on their geographical location into four census regions: Northeast, Midwest, South, and West. However, regional boundaries can be arbitrary, and individuals living in bordering areas may have affiliations with multiple regions or identify more strongly with a neighboring region. For instance, states like Missouri and Kentucky are often considered part of the Midwest but also share cultural and economic ties with the South. This complexity can make individuals accurately reporting their region challenging.

Finally, for "employ", indicating the respondent's self-reporting employment status, can be influenced by various factors such as job changes, seasonal work, and personal circumstances. In addition, individuals may intentionally misreport their employment status due to societal pressures or to access benefits. For instance, some may falsely claim employment to avoid stigma, while others may report unemployment to qualify for unemployment benefits. Additionally, students working part-time jobs may selectively report their status based on their perceived benefit. These inconsistencies can introduce bias and inaccuracies into the data, impacting the reliability of our analyses.