**GOVT 5001**
**Analysis of Political Data**
**Fall 2022**

**Final Project: A Data Memo**

**Due: Tuesday, December 6th at 11:59 p.m. ET**

# Assignment Description

In this project, you will use R to analyze data from the 2020 Cooperative Election Study (CES)[1]. The CES is a very large survey of 61,000 respondents which collects information on how Americans vote, participate in politics, and view major political issues. The study originated in 2006 under the name "Cooperative Congressional Election Study". Subjects for the 2020 survey were recruited during Fall 2020, and the survey was conducted over the Internet by YouGov in two waves: pre-election (September 29 - November 2) and post-election (November 8 - December 14). A sample of US adults was drawn using YouGov's matched random sample methodology, which selects members from a pool of opt-in respondents to create a sample of respondents who have the same measured characteristics as a true random sample. The CES is designed to be representative of all national adults.

Overall, you will:

(1) get experience using a large publicly available data set;

(2) gain additional experience using R;

(3) select appropriate methods to answer questions;

(4) apply the tools and concepts that you've learned in this course to gain insight into substantive questions;

(5) convey your analysis in a concise, well-organized manner – in the form of a data analysis memo.

A modified version of the 2020 CES is provided with this assignment (`ces.csv`) and a codebook is provided below.

# Project Description

## Theory and Research Motivation

Brady, Verba, and Schlozman (1995) develop a resource model of political participation.[2] The resources they consider are time, money, and civic skills, where civic skills are communication and organizational capacities that are essential to political activity. They argue that education is an important source of civic skills, but that such skills can also be developed in the workplace, voluntary associations, and religious faiths.

You will create a measure of political activity using the CCES to test hypotheses derived from Brady, Verba, and Schlozman's resource model of political participation. Finally, you will consider the implications of unequal rates of political activity on representation. Specifically, if politically active individuals hold different policy views than politically inactive individuals *and* elected officials are more responsive to politically active individuals, then we may worry that policy outcomes will be incongruent with the policy views of inactive individuals. Therefore, we will use the CCES to determine whether politically active and inactive individuals hold different policy views.

[1] Ansolabehere, Stephen, Brian F. Schaffner, and Sam Luks, COOPERATIVE ELECTION STUDY, 2020: COMMON CONTENT. [Computer File] Release 2: August 4th, 2021. Cambridge, MA: Harvard University [producer] http://cces.gov.harvard.edu

[2] Brady, Henry E., Sidney Verba, and Kay Lehman Schlozman. 1995. "Beyond SES: A Resource Model of Political Participation." *American Political Science Review* 89(2): 271-294.

## General Instructions

Answer each question or complete each task below. Use a figure or table to describe a relationship when appropriate. You should describe the magnitude of effects and conduct appropriate statistical tests to characterize the uncertainty in your estimates. Readers should not be left to interpret the results themselves.

## Part 1: Measuring Political Activity

Using the binary variables for political acts - voting, attending meetings, putting up a sign, working for a campaign, protesting, contacting a public official, and donating money - create a scale from 0-7 that represents the number of political acts a respondent engaged in. In other words, create a variable that counts the number of these activities that a respondent reports doing.

Visualize the distribution of this variable and interpret the figure. What do we learn about people's political activity from this measure?

**Note:** For this assignment, we treat the number of political acts as a continuous variable.

## Part 2: Explaining Political Activity

1. Does political activity increase with education? Here we use education as a measure of civic skill.
2. Does political activity increase with family income? You may want to collapse some of the categories of income. If you choose to do so, be clear about how you recoded it.
3. Someone says they worry that the effect of education on political activity is confounded by family income. Explain their concern and execute analysis to address it.
4. Use the CCES data to estimate a model that you believe best explains political activity. Justify your choices. If you added new variables to your model, does including these variables change your conclusions from item 3 above? What are some possible omitted variables not available in this dataset? Can you claim to have estimated causal effects?

## Part 3: Political Activity and Policy Preferences

Select at least two policy variables to use as your dependent variable for this question. Examples of policy variables include `medicare`, `EPA`, and `police_incr`.

For each policy variable you chose, do respondents who are more active in politics have different policy preferences than respondents who are less active in politics?

What, if any, implications do your findings have for the potential represenativeness of public policy?

## Part 4: Measurement

Discuss any concerns you have about the count of political activities as a measure of the concept political participation and the implications of your concerns for your results. This discussion should be included in the analysis section of your memo.

**Extra Credit: 5 points**

If possible execute analysis that addresses your measurement concern above and discuss the results. Or explain why no such analysis is possible with these data. Extra credit will be graded very stringently.

## Due Date and Expectations

The final project is due on Canvas no later than Tuesday, December 6th at 11:59 p.m. ET. It should be formatted as a data memo (see Guidelines on Writing Memos below). In addition to turning in your memo (which should include all tables and figures), please also turn in your R script file. Your project will not be considered complete and submitted on time if you do not include an R script file with the code necessary to reproduce the analysis in your memo. Only one member of your group needs to submit the assignment.

All work on this assignment must be discussed and performed only by you and your team members. This is a group project, but not an all-class project. Each group should independently analyze and write up their results.

## Choosing Groups

You should work in groups of 3-4. Please email Ben and me to propose a group. Only one member of your group needs to email the us. We try to accept all proposed groups, but occasionally will need to move a student to a different group. If you don't have a group and would like us to assign you to one, we are happy to do that, just please let us know.

## Guidelines on Writing Memos

Your assignment should be in memo form with:

- A brief introductory/overview section that describes the overall research questions and has a brief review of your findings.
- A section summarizing the data set (e.g., sample size, time period, target population).
- A separate section for each of the questions/analyses requested (referring, when applicable, to supporting tables and figures).
- A concluding section which summarizes your findings, whether you have found a causal relationships (i.e., what is the interval validity of your study), and how broadly your findings generalize (i.e., what is the external validity of your study).
- Tables and figures should be appropriately labeled and formatted (in other words, do not simply paste R output) and refer to it at specific points in the body of the memo.

There is no strict page limit or word limit. However, as a guideline, I think you can complete your analysis in the equivalent of about 3-5 single-spaced pages of text, plus tables and figures as needed. Your memo may be longer or shorter, depending on your use of bullet points, spacing, size of tables and figures, etc.

Do not turn in R output with your memo. It will be your responsibility to convey your findings and analyses through your memo, and in supporting tables or graphs that you create. Your report should be publication-ready. For examples, consider the presentation of tables and figures and the associated discussion in published academic articles that you have read.

Your memo should be accessible to *two* types of readers:

1) Readers who are not familiar with the language of statistics (e.g., an educated reader of a newspaper who has not had a class in statistics), and
2) Readers who will only believe statements you make if you show that you arrived at your answer using rigorous statistical methods (e.g., your instructor and TA for this course).

How do you balance these two audiences?

For audience #1, it is your responsibility to appropriately interpret and translate your statistical analyses into language that is accessible to those who are not trained in statistics. This means being able to convey the gist of the "null hypothesis," "statistical significance," "generalizability", etc. without actually using the terms (when you do use them, you must clearly explain them).

For audience #2, you will want to include information about the type of tests you conducted, the test statistics you obtained, and p-values (perhaps in footnotes, parentheses, or as supporting information to the statements you make for audience #1).

## Coding and Data Wranlging Tips

Lastly, R will import the variables in `ces.csv` as numeric. You may want to turn some into categorical or ordinal variables with the integers labeled per the codebook. One way to do this is to turn the variable into a factor variable using `factor()`. See Section 2.2.5 of *QSS*. For example, the following code chunk would convert `region` into a factor variable with levels corresponding to regions rather than numbers. Additionally, you'll see that `factor()` takes an `ordered` argument that will indicate the variable is ordinal. If you give `lm()` an ordered factor, it will treat as it an n-degree polynomial where n is the number of levels of the factor. If you want `lm()` to turn each level of the factor into an binary variable (like the example from class), **do not** set `ordered = TRUE` in `factor()`.

```
ces$region <- factor(ces$region, levels = 1:4,
                     labels = c("Northeast", "Midwest", "South", "West"))
```

Suppose I want to create a new variable in my data set that is the sum of variables 1 through 3 for each observation. The following code would yield the desired new variable, but be sure to examine what happens if one of variables 1 through 3 is missing. See Section 4.2 of *QSS* for discussion of missing data.

```
data$sum <- data$var_1 + data$var_2 + data$var_3
```

# Codebook

| Variable | Level | Description |
|---|---|---|
| gender | | **Gender** |
| | 1 | Male |
| | 2 | Female |
| educ | | **Highest Level of Education Completed** |
| | 1 | No HS |
| | 2 | High school graduate |
| | 3 | Some college |
| | 4 | 2-year degree |
| | 5 | 4-year degree |
| | 6 | Post-grad |
| race | | **Racial/Ethnic Group** |
| | 1 | White |
| | 2 | Black |
| | 3 | Hispanic |
| | 4 | Asian |
| | 5 | Native American |
| | 6 | Middle Eastern |
| | 7 | Two or more races |
| | 8 | Other |
| hispanic | | **Spanish, Latino, or Hispanic origin or descent** |
| | 1 | Yes |
| | 2 | No |
| region | | **Census region of residence** |
| | 1 | Northeast |
| | 2 | Midwest |
| | 3 | South |
| | 4 | West |
| ownhome | | **Homeowner or Renter** |
| | 1 | Own |
| | 2 | Rent |
| | 3 | Other |
| urbancity | | **Neighborhood Type** |
| | 1 | City |
| | 2 | Suburb |
| | 3 | Town |
| | 4 | Rural area |
| | 5 | Other |
| unionhh | | **Are you or any member of your household a union member?** |
| | 1 | Yes |
| | 0 | No |
| religious | | **How important is religion in your life?** |
| | 1 | Very important |
| | 2 | Somewhat important |
| | 3 | Not too important |
| | 4 | Not at all important |

| Variable | Level | Description |
|---|---|---|
| faminc | | **Family Income** |
| | 1 | <10,000 |
| | 2 | 10,000 - 19,999 |
| | 3 | 20,000 - 29,999 |
| | 4 | 30,000 - 39,999 |
| | 5 | 40,000 - 49,999 |
| | 6 | 50,000 - 59,999 |
| | 7 | 60,000 - 69,999 |
| | 8 | 70,000 - 79,999 |
| | 9 | 80,000 - 99,999 |
| | 10 | 100,000 - 119,999 |
| | 11 | 120,000 - 149,999 |
| | 12 | 150,000 - 199,999 |
| | 13 | 200,000 - 249,999 |
| | 14 | 250,000 - 349,999 |
| | 15 | 350,000 - 499,999 |
| | 16 | 500,000 or more |
| voted | | **Voted in 2020 election (Validated)** |
| | 1 | Yes |
| | 0 | No |
| meeting | | **Attend local political meetings** |
| | 1 | Yes |
| | 0 | No |
| sign | | **Put up a political sign** |
| | 1 | Yes |
| | 0 | No |
| campaign | | **Work for a candidate or campaign** |
| | 1 | Yes |
| | 0 | No |
| protest | | **Attend a political protest, march or demonstration** |
| | 1 | Yes |
| | 0 | No |
| contact | | **Contact a public official** |
| | 1 | Yes |
| | 0 | No |
| donate | | **Donate money to a candidate, campaign, or political organization** |
| | 1 | Yes |
| | 0 | No |

| Variable | Level | Description |
|---|---|---|
| ideo5 | | **Ideology** |
| | 1 | Very liberal |
| | 2 | Liberal |
| | 3 | Moderate |
| | 4 | Conservative |
| | 5 | Very conservative |
| pid3 | | **Party Identification** |
| | 1 | Democrat |
| | 2 | Republican |
| | 3 | Independent |
| | 4 | Other |
| | 5 | Not Sure |
| pid7 | | **7 Point Party ID** |
| | 1 | Strong Democrat |
| | 2 | Not very strong Democrat |
| | 3 | Lean Democrat |
| | 4 | Independent |
| | 5 | Lean Republican |
| | 6 | Not very strong Republican |
| | 7 | Strong Republican |
| medicare | | **Expand Medicare to a single comprehensive public health care coverage program that would cover all Americans** |
| | 1 | Support |
| | 0 | Oppose |
| ACA | | **Repeal the entire Affordable Care Act** |
| | 1 | Support |
| | 0 | Oppose |
| abortion | | **Always allow a woman to obtain an abortion as a matter of choice** |
| | 1 | Support |
| | 0 | Oppose |
| EPA | | **Strengthen the Environmental Protection Agency enforcement of the Clean Air Act and Clean Water Act even if it costs U.S. jobs** |
| | 1 | Support |
| | 0 | Oppose |
| police_incr | | **Increase the number of police on the street by 10 percent, even if it means fewer funds for other public services** |
| | 1 | Support |
| | 0 | Oppose |
| police_decr | | **Decrease the number of police on the street by 10 percent, and increase funding for other public services** |
| | 1 | Support |
| | 0 | Oppose |
| trade | | **25 percent tariffs on imported steel and 10 percent on imported aluminum, EXCEPT from Canada and Mexico** |
| | 1 | Support |
| | 0 | Oppose |
| minwage | | **Raise the minimum wage to 15 dollars an hour** |
| | 1 | Support |
| | 0 | Oppose |
| work_req | | **Require able-bodied adults 18 to 49 years of age who do not have dependents to have a job in order to receive food stamps** |
| | 1 | Support |
| | 0 | Oppose |