5304 Econometrics Autumn 2023[1]

Problem set 1
Data management and exploratory analyses

For this assignment you will be using a data excerpt from the American National Election Studies survey (ANES 2020). This data set contains survey responses from a sample of the U.S. population – the number of available variables is huge, but we will be focusing only on a small subset of them. The data is available in *PS1_ANES2020.dta*.

1. Load the data set. How many observations and variables are there? Are there any duplicates in terms of the unique observation identifier? If there are, drop them. *Hint: use "duplicates"*

2. Clean and recode the variables in the data set in the following way:

   (a) Drop observations that have missing values in any of the variables. *Hint: missing variables are recoded to nonsensical values, use "codebook" to find them*

   (b) Drop all respondents that did not vote for either Trump or Biden. Create a dummy variable for voting Trump.

   (c) Generate a dummy for having at least a Bachelor's degree.

   (d) Generate a dummy for being white.

   (e) Generate a dummy for *not* working for pay last week.

   (f) Label your variables.

3. Create a table with descriptive statistics for the groups of Trump and Biden voters, as well as the whole pooled sample, in separate columns. Include the dummy variables that you created above, as well as age and the "feeling toward conservatives" variable. Make sure that your table shows both the means of the variables as well as their standard deviations. Describe the typical Trump and Biden voter and their differences. *Hint: use "estpost summarize" by groups and compile the table with "esttab". Read the help files – learning to work with the esttab package is time well invested!*

4. A friend wants you to investigate the relationship between voting for Trump and age:

   (a) Begin by plotting the distribution of the age variable. Will you be able to say anything about the relationship of interest for people over 80? If not, exclude this group from the remainder of this question. *Hint: when working with histograms, you need to think clearly about the size of your bins – having bins too wide could hide something important in the data. Work with the width option for Stata's "histogram" function!*

   (b) Your friend likes to see data visualized. You first try plotting the share voting for Trump versus age in a scatter plot (do this), which looks terrible since the dependent variable is binary. Instead of a proper scatter plot, your friend asks you to compute means of Trump voting in 20 equally sized bins[2] of age, and then plotting the relationship. Do this and include a linear and quadratic fit in separate figures. Which seems to fit the data best? *Hint: use "binscatter" which does exactly this*

---

[1] TA: Petter Berg, petter.berg@phdstudent.hhs.se

[2] The exact number is not so important here, you can choose what you like as long as the number is not too small.

(c) Let's say you were to run a (linear probability) regression of Trump voting (binary, 0 or 1) on age (integer, 18-80) and race (categorical, values 1-6). Your friend wants to allow for age to enter as a quadratic polynomial, and control for race indicators appropriately. How would you write down the econometric specification for this regression (e.g. $Trump_i = ...$)?

5. Your friend also wants to know whether there is a difference in Trump voting for those employed vs. unemployed, and wants to see this visually. What is the problem in using the "not working" variable that you created? Given that you only have the data you have, do your best to answer your friend's question. *Hint: use "cibar"*

References

ANES (2020): "ANES 2020 Time Series Study," *American National Election Studies. University of Michigan, Stanford University.*