Estimating the Bias in Self-Reported Turnout
Part I: Loading and Making Sense of Data

As we will learn in Chapter 3, surveys are frequently used to collect data from a sample of individuals for the purpose of inferring the characteristics of the population from which the sample of individuals was drawn. One of the complications we face when conducting survey research is misreporting, that is, the fact that participants might provide inaccurate or false information. This is particularly likely when one answer is more socially acceptable or desirable than the others. For example, in federal elections in the United States, official turnout rates are systematically lower than self-reported turnout rates. Voting is often perceived as to be a civic duty, so respondents might feel social pressure to lie about their voting behavior.

In a few problem sets, we will estimate the bias in self-reported turnout data in the American National Election Studies (ANES), which is a survey that collects voting data on a representative sample of adults in the United States. The dataset we will use is in a file called "ANES.csv". Table 1 shows the names and descriptions of the variables in this dataset, where the unit of observation is federal elections in the U.S.

variable	description
year	year of the election
presidential	whether it was a presidential election: 1=yes, 0=no
midterm	whether it was a midterm election: 1=yes, 0=no
ANES_turnout	proportion of ANES respondents who reported to have voted in the election (in percentages)
votes	number of ballots officially cast in the election (in thousands)
VEP	voting eligible population at the time (in thousands)
VAP	voting age population at the time (in thousands)
felons	number of felons not eligible to vote (in thousands)
noncitizens	number of non-citizens living in the U.S. (in thousands)

Table 1: Variables in "ANES.csv"

In this problem set, we practice how to load and make sense of data.

1. Use the function read.csv() to read the CSV file "ANES.csv" and use the assignment operator <- to store the data in an object called *anes*. (Do not forget to set the working directory first.) Provide the R code you used (without the output). (10 points)

2. Use the function head() to view the first few observations of the dataset. Provide the R code you used (without the output). (5 points)

3. What does each observation in this dataset represent? (5 points)

4. Use the function View() to open a tab with the entire contents of the dataframe. What is the time period covered in the dataset? In other words, what's the first and last election the dataset contains? (5 points)

5. Please substantively interpret the first observation in the dataset. (5 points)

6. For each variable in the dataset, please identify the type of variable (character vs. numeric binary vs. numeric non-binary) (10 points)

7. How many observations are in the dataset? In other words, how many federal elections are part of this dataset? (Hint: the function dim() might be helpful here.) Provide the R code you used (without the output) and provide the substantive answer. (10 points)