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

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)

   R code:

   ```
   anes <- read.csv("ANES.csv") # reads and stores data
   ```

(Recall: to the left of the assignment operator <-, we specify the name of the object, anes in this case; to the right of the assignment operator <-, we specify the contents, which, in this case, are produced by reading the CSV file "ANES.csv". Also, we do not use quotes around the name of an object such as anes or around the name of a function such as read.csv(), but we do use quotes around the name of a file: "ANES.csv".)

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)

   R code:

```
head(anes) # shows first observations
##   year presidential midterm ANES_turnout votes    VEP    VAP felons noncitizens
## 1 1980            1       0           71 86515 159635 164445    802        5756
## 2 1982            0       1           60 67616 160467 166028    960        6641
## 3 1984            1       0           74 92653 167702 173995   1165        7482
## 4 1986            0       1           53 64991 170396 177922   1367        8362
## 5 1988            1       0           70 91595 173579 181955   1594        9280
## 6 1990            0       1           47 67859 176629 186159   1901       10239
```

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

   Answer: Each observation represents a U.S. federal election. (Note: We know this because, as stated above, the unit of observation in this dataset is federal elections in the U.S.)

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)

   R code:

```
View(anes) # opens a new tab with contents of dataset
```

   Answer: The dataset covers the time period from 1980 to 2004. In other words, the first election is the U.S. federal election that took place in 1980 and the last election is the U.S. federal election that took place in 2004. (Note: We can see this by scrolling up and down on the tab opened by View() and looking at the minimum and maximum value of *year*.)

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

   Answer: The first observation in the dataset represents the federal election that took place in 1980, which was a presidential election, not a midterm election, 71% of the respondents of the ANES survey reported to have voted in this election, 86,515 thousand ballots were officially cast in this election, the number of voting eligible population at the time was 159,635 thousand, the number of voting age population at the time was 164,445 thousand, the number of felons not eligible to vote was 802 thousand, and the number of non-citizens living in the U.S. was 5,756 thousand. (Note: the first observation consists of the following values: *year*=1980, *presidential*=1, *midterm*=0, *ANES_turnout*=71, *votes*=86515, *VEP*=159635, *VAP*=164445, *felons*=802, *noncitizens*=5756; we can interpret each value by using the description of the variables in Table 1.)

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

   Answer: *job_id* is a numeric non-binary variable, *criminal* is a numeric binary variable, *call* is a numeric binary variable, and *race* is a character variable. (Recall: binary variables can only take two values, 0s and 1s, and non-binary variables can take more than two values.)

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)

   R code:

   ```
   dim(anes) # provides dimensions of dataframe: rows, columns
   ## [1] 13 9
   ```

   Answer: There were 13 federal elections that are part of this dataset. (Recall: the first number provided by dim() corresponds to the number of observations in the dataframe, the second number corresponds to the number of variables. Based on the output of dim(), there are 13 observations in the dataframe *anes*. Since the unit of observation is federal elections, the dataframe *anes* has data on 13 federal elections.)