Does Social Pressure Affect Turnout?
Part II: Estimating an Average Causal Effect (with Solutions)

Let's continue working with the data from the randomized experiment conducted in Michigan and estimate the average causal effect of receiving the social pressure message on the probability of voting.

The dataset we will use is in a file called "voting.csv". Table 1 shows the names and descriptions of the variables in this dataset, where the unit of observation is registered voters.

variable	description
birth	year of birth of registered voter
message	whether registered voter was assigned to receive the social pressure message: "yes", "no"
voted	whether registered voter voted in the August 2006 election: 1=voted, 0=didn't vote

Table 1: Variables in "voting.csv"

In this problem set, we practice how to create new variables, how to estimate an average treatment effect using data from a randomized experiment, and how to write a conclusion statement.

As always, we start by loading and looking at the data:

```
## load and look at the data
voting <- read.csv("voting.csv") # reads and stores data
head(voting) # shows first observations
##   birth message voted
## 1  1981      no     0
## 2  1959      no     1
## 3  1956      no     1
## 4  1939     yes     1
## 5  1968      no     0
## 6  1967      no     0
```

1. To estimate the average causal effect of receiving the social pressure message on the probability of voting:

   a. What should be our outcome variable (Y) and what type of variable is it? Please just provide the name of the variable and indicate whether the variable is character, numeric binary, or numeric non-binary. (2.5 points)

      Answer: The outcome variable should be *voted*, which is a numeric binary variable since it can only take 1s and 0s.

   b. Do we already have in the dataframe our treatment variable (X), that is, a numeric binary variable that identifies the observations that received the treatment? A yes/no answer will suffice. (2.5 points)

Answer: No, we do not have the treatment variable already in the dataframe. We have a character variable, *message*, that identifies the observations that received the treatment with a "yes", but it is not numeric.

c. Let's create our treatment variable by using the function ifelse() and the existing variable *message*. Name this new variable *pressure* and make sure to store it in the dataframe *voting*. (R code only.) (2.5 points)

R code:

```
## create variable pressure inside dataframe voting
voting$pressure <- # stores return values in new variable
    ifelse (voting$message=="yes", # logical test
      1, # return value if logical test is true
      0) # return value if logical test is false
```

(Recall: The function ifelse() requires three arguments and in this order: (1) the logical test (specified using == in this class), (2) the return value if test is false, (3) the return value if test is true. The existing variable we use to create this new binary variable is *message* since that is the variable that identifies receipt of the treatment. When the value of *message* equals "yes", we want the new variable *pressure* to equal 1; and to equal 0 otherwise. Hence, we can use ifelse(voting$message=="yes", 1, 0) to produce the values of this new variable. Note, however, that since there are only two possible values of *message*, we would create the same values if we ran ifelse(voting$message=="no", 0, 1) instead. Now, to store these values as a variable named *pressure* inside the dataframe *voting*, we use the assignment operator <- and specify voting$pressure to the left of the assignment operator. Without voting$ in front of pressure, you would be creating a variable outside the dataframe, as a new object, which is not what we want.)

d. Look at the first few observations of *voting* to confirm that the new variable was created correctly. (R code only.) (2.5 points)

R code:

```
head(voting)
##   birth message voted pressure
## 1  1981      no     0        0
## 2  1959      no     1        0
## 3  1956      no     1        0
## 4  1939     yes     1        1
## 5  1968      no     0        0
## 6  1967      no     0        0
```

(Note: We can confirm here that whenever *message* equals "yes", *pressure* equals 1, and whenever *message* equals "no", *pressure* equals 0, which is exactly what we want.)

e. Considering that the dataset comes from a randomized experiment, what can we compute to estimate the average causal effect of receiving the social pressure message on the probability of voting? Please provide the name of the estimator and its definition. (2.5 points)

Answer: We can compute the difference-in-means estimator, which is defined as the average outcome for the treatment group minus the average outcome for the control group.

f. What are the individuals or entities in the treatment group and what are the individual or entities in the control group? Please be precise. (2.5 points)

Answer: The treatment group are the registered voters who were randomly assigned to receive the social pressure message and the control group are the registered voters who were randomly assigned not to receive the social pressure message.

2. Let's start by computing the average outcome for the treatment group and the average outcome for the control group separately. Please interpret the output of both with a full sentence and make sure to include the unit of measurement. (10 points)

R code:

```
## compute mean of voted
mean(voting$voted[voting$pressure==1]) # for registered voters who received message
## [1] 0.3779482
mean(voting$voted[voting$pressure==0]) # for registered voters who did not receive it
## [1] 0.2966383
```

(Recall: we use [] to subset a variable; inside the square brackets, we specify the criterion of selection. For example, we can use the relational operator == to set a logical test; only the observations for which the logical test are true will be extracted. In the first case above, for example, we extract from *voted* the observations that belong to the registered voters who were assigned to receive the social pressure message (for which *pressure* equals 1) and then calculate the mean of those observations.)

Answer: We can interpret the first mean above as: 38% of the registered voters who received the social pressure message voted. We can interpret the second mean above as: 30% of the registered voters who did not receive the social pressure message voted. (Note: The word "average" does not appear in these interpretations because the mean of a binary variable should be interpreted as the proportion of the observations that have the characteristic identified by the variable. Here, *voted* is binary and identifies the registered voters that voted. The unit of measurement is %, after multiplying the output by 100.)

3. Now, let's answer the question: What is the estimated average causal effect of receiving the social pressure message on the probability of voting? Please provide a full substantive answer (make sure to include the assumption, why the assumption is reasonable, the treatment, the outcome, as well as the direction, size, and unit of measurement of the average treatment effect) (25 points)

R code:

```
## compute the difference in means estimator
mean(voting$voted[voting$pressure==1]) - #average outcome for treatment group MINUS
  mean(voting$voted[voting$pressure==0]) #average outcome for control group
##  [1] 0.08130991
```

Answer: Let's start by figuring out each key element separately.

- *What's the assumption?* We assume that the registered voters who were randomly assigned to receive the social pressure message (the treatment group) are comparable to the registered voters who were randomly assigned NOT to receive the social pressure message (the control group). (Note: If this assumption were not true the difference-in-means estimator would NOT produce a valid estimate of the average treatment effect.)

- *Why is the assumption reasonable?* Because the social pressure message was assigned at random OR because the data come from a randomized experiment. (Recall: Random treatment assignment makes the treatment and control groups on average identical to each other in all observed and unobserved pre-treatment characteristics.)

- *What's the treatment?* Receiving the social pressure message.

- *What's the outcome?* Probability of voting (Note: When the outcome variable is binary, we speak of "probability of 1," where 1 is whatever the 1 in the variable identifies.)

- *What's the direction, size, and unit of measurement of the average causal effect?* An increase of 8 percentage points, on average. (Note: It is an increase because we are measuring change—the change in the outcome variable caused by the treatment—and the difference-in-means estimator is positive. The difference-in-means estimator = proportion of registered voters who received the social pressure message who voted - proportion of registered voters who did NOT receive the social pressure message who voted= 38% - 30% = 8 percentage points. Recall that percentage point is the unit of measurement of the arithmetic difference between two percentages.)

Full answer: Assuming that the registered voters who were randomly assigned to receive the social pressure message are comparable to the registered voters who were randomly assigned NOT to receive the social pressure message (a reasonable assumption since the social pressure message was assigned at random), we estimate that receiving the social pressure message increases the probability of voting by 8 percentage points, on average.

(Note: the following statement is not the correct answer because it doesn't make a causal claim; it only makes an observation: Assuming that the registered voters who were randomly assigned to receive the social pressure message are comparable to the registered voters who were randomly assigned NOT to receive the social pressure message (a reasonable assumption since the social pressure message was assigned at random), we estimate that registered voters who received the message were 8 percentage points more likely to vote than the registered voters who did not receive the message, on average.)