Does Having a Black Candidate Running Increase Black Turnout? (with Solutions)

(Based on Bernard Fraga. 2016. "Candidates or Districts? Reevaluating the Role of Race in Voter Turnout." *American Journal of Political Science*, 60: 97-122.)

Some scholars have suggested that having a black candidate running increases black turnout. In this problem set, we explore the causal relationship between black candidates and black turnout using observational data from U.S. elections.

The dataset is in a file called "districts.csv". Table 1 shows the names and descriptions of the variables in this dataset, where the unit of observation is district elections.

variable	description
year	year of the election
state	state where the district is located
district	district number, which is unique within states but not across states
proportion_black	proportion of the district's voting-age population that was black at the time (in percentages)
black_candidate	whether there was a black candidate running at the district-level election in that year: 1=there was a black candidate, 0=there was not
black_turnout	proportion of the district's black voting age population that voted in that year's election (in percentages)

Table 1: Variables in "districts.csv"

In this problem set, we practice fitting a linear model to compute the difference-in-means estimator and fitting a multiple linear regression model to statistically control for confounders.

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

```
## load and look at the data
districts <- read.csv("districts.csv") # reads and stores data
head(districts, n=8) #shows first eight observations
##   year state district proportion_black black_candidate black_turnout
## 1 2006    AK        1         3.178487               0       43.94295
## 2 2006    AL        1        26.122134               0       26.60873
## 3 2006    AL        2        29.478384               0       26.66497
## 4 2006    AL        3        30.843040               0       28.46918
## 5 2006    AL        4         4.996171               0       27.97092
## 6 2006    AL        5        16.597923               0       29.18517
## 7 2006    AL        6        10.682076               0       28.32708
## 8 2006    AL        7        61.952636               1       32.90645
```

In the code above, we asked R to show the first eight observations, instead of the default of six, by using the optional argument `n` in the function `head()`.

1. First, let's make sure we understand the data and identify our X and Y variables.

    a. Substantively interpret the seventh and eight observations in the dataset and do not forget to include the unit of measurements. (2.5 points)

    Answer: The seventh observation in the dataset refers to the election that took place in 2006 in Alabama's district number 6, where about 11% of the voting-age population was black, there was no black candidate running, and a little over 28% of the black voting-age population voted. The eight observation in the dataset refers to the election that took place in 2006 in Alabama's district number 7, where about 62% of the voting-age population was black, there was a black candidate running, and close to 33% of the black voting-age population voted. (Note: The values of the seventh observation are: *year*=2006, *state*="AL", *district*=6, *proportion_black*=10.68%, *black_candidate*=0, and *black_turnout*=28.33%. And, the values of the eight observation are: *year*=2006, *state*="AL", *district*=7, *proportion_black*=61.95%, *black_candidate*=1, and *black_turnout*=32.91%.)

    b. Given that we are interested in estimating the average causal effect of having a black candidate running on black turnout: What should be our Y variable? In other words, which variable is the outcome variable? And, is this variable binary or non-binary? (2.5 points)

    Answer: The Y variable should be *black_turnout*, which is a non-binary variable since it can take more than two values.

    c. Given that we are interested in estimating the average causal effect of having a black candidate running on black turnout: What should be our X variable? In other words, which variable is the treatment variable? And, is this variable binary or non-binary? (2.5 points)

    Answer: The X variable should be *black_candidate*, which is a binary variable since it can only take 1s and 0s. (Recall: All treatment variables in this class are binary. They equal 1 when the observation was treated, 0 when the observation was not treated. In this case, the treatment is the presence of a black candidate running and, thus, the treatment variable is *black_candidate*, which equals 1 when there is a black candidate running, and 0 otherwise.)

2. For now, let's assume that the data we are analyzing came from a randomized experiment, where researchers were able to randomly assign black candidates to district elections. If this were true, then, we could estimate the average causal effect of having a black candidate running on black turnout by computing the difference-in-means estimator.

    a. Let's start by computing the average outcome for the treatment group, that is, the average black turnout in district elections with a black candidate running. Provide an interpretation of the result using a full sentences, and do not forget to include the unit of measurement. (2.5 points)

R code:

```
## compute average outcome for treatment group
mean( districts $black_turnout[ districts $black_candidate==1])
## [1] 45.54975
```

(Recall: The R function to compute a mean is mean(). 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 case above, we extract from *black_turnout* the observations that belong to districts with a black candidate (for which *black_candidate* equals 1) and then calculate the mean of those observations.)

Answer: In district elections with a black candidate running, the average black turnout is of 45.55%.

b. Now, compute the average outcome for the control group, that is, the average black turnout in district elections without a black candidate running. Provide an interpretation of the result using a full sentences, and do not forget to include the unit of measurement. (2.5 points)

R code:

```
## compute average outcome for control group
mean( districts $black_turnout[ districts $black_candidate==0])
## [1] 39.38574
```

Answer: In district elections without a black candidate running, the average black turnout is of 39.39%.

c. Compute the difference-in-means estimator directly and report its value. (2.5 points)

R code:

```
## compute the difference in means estimator
mean( districts $black_turnout[ districts $black_candidate==1]) -
  mean( districts $black_turnout[ districts $black_candidate==0])
## [1] 6.164014
```

Answer: The difference-in-means estimator equals 6.16.

d. Now, let's use the lm() function to fit a line to the data in such a way that the $\widehat{\beta}$ coefficient will be equivalent to the difference-in-means estimator. What is the fitted line? In other words, provide the formula $\widehat{Y} = \widehat{\alpha} + \widehat{\beta}X$ where you specify each term (i.e., substitute Y for the name of the outcome variable, substitute $\widehat{\alpha}$ for the estimated value of the intercept coefficient, substitute $\widehat{\beta}$ for the estimated value of the slope coefficient, and substitute X for the name of the treatment variable.) (Hint: The lm() function requires an argument of the form Y~X) (5 points)

R code:

```
## fit the linear model
lm( districts $black_turnout ~ districts $black_candidate) # or
```

```
lm(black_turnout ~ black_candidate, data = districts )
##
## Call :
## lm(formula = black_turnout ~ black_candidate, data = districts )
##
## Coefficients :
##       ( Intercept )   black_candidate
##              39.386             6.164
```

(Recall: To fit a linear model in R we use the lm() function. This function requires an argument of the type Y ~ X. Here, *black_turnout* is the outcome variable, Y, and *black_candidate* is the treatment variable, X. To specify the dataframe where the variables are stored, we can use either the $ operator or the optional argument data.)

Answer: $\widehat{\text{black\_turnout}} = 39 + 6$ *black_candidate*. (Note: The Y variable is *black_turnout*, $\widehat{\alpha}$=39, $\widehat{\beta}$=6, and the X variable is *black_candidate*.)

e. Is the estimated slope coefficient ($\widehat{\beta}$) equivalent to the value of the difference-in-means estimator in this case? A yes/no answer will suffice. (5 points)

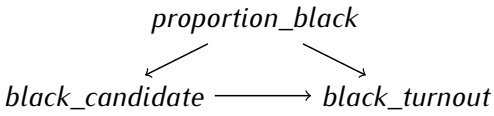
Answer: Yes, the estimated slope coefficient is equivalent to the difference-in-means estimator. (Note: They are both equal to 6.16.)

f. If the data came from a randomized experiment, what would you conclude is the direction, size, and unit of measurement of the estimated average causal effect of having a black candidate running on black turnout? (5 points)

Answer: If the data came from a randomized experiment, I would estimate that having a black candidate running increases black turnout by 6 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 unit of measurement is the same as the unit of measurement of $\triangle\overline{Y}$. Since Y is non-binary and measured in percentages, $\overline{Y}$ is measured in percentages and $\triangle\overline{Y}$, the difference-in-means estimator, and $\widehat{\beta}$ are measured in percentage points. Recall: Percentage points is the unit of measurement of the difference between two percentages. In this case, 45.55% - 39.39% = 6.16 percentage points.

3. Since the data is observational, which means it did not come from a randomized experiment, the district elections with black candidates running might be different in some relevant ways from the district elections without black candidates running. In other words, there might be potential confounding variables present obscuring the causal relationship between *black_candidates* and *black_turnout*. A potential confounder is, for example, the proportion of the district's voting-age population that was black at the time of the election: *proportion_black*. Here is the reasoning: It is likely the case that districts with a higher proportion of black voting-age population are

more likely to have black candidates running in the election ($Z \rightarrow X$). It is also likely the case that districts with a higher proportion of black voting-age population are more likely to have a higher black turnout in the elections ($Z \rightarrow Y$).



a. To further explore the possibility that *proportion_black* is a confounder, compute the correlation between *proportion_black* and *black_candidate*. Are these two variables moderately to highly correlated with each other? A yes/no answer will suffice. (5 points)

   R code:

```
## compute correlations
cor( districts $proportion_black,  districts $black_candidate)
## [1] 0.7281678
```

   (Recall: The function in R to compute a correlation coefficient is cor(). The only two required arguments are the code identifying the two variables. The order of the variables does not matter since the correlation between X and Y is the same as the correlation between Y and X.)

   Answer: Yes, these two variables are moderately to highly correlated with each other so it is possible that *proportion_black* is a confounder. (Note: The correlation coefficient is much closer to 1 than to 0.)

b. Now, statistically control for *proportion_black* by running a multiple linear regression model and estimate the average causal effect of having a black candidate running on black turnout while keeping the proportion of black voting-age population in the district constant. Report the new direction, size, and unit of measurement of the estimated average causal effect of having a black candidate running on black turnout. (10 points)

   R code:

```
## fit the linear model
lm( districts $black_turnout ~  districts $black_candidate + proportion_black) # or
```

```
lm(black_turnout ~ black_candidate + proportion_black, data = districts )
##
## Call :
## lm(formula = black_turnout ~ black_candidate + proportion_black,
##     data = districts )
##
## Coefficients :
##       ( Intercept )    black_candidate    proportion_black
##             37.5275            -0.7364             0.2074
```

   (Recall: To fit a multiple linear regression model in R we also use the lm() function. The required argument is a formula of the type $Y \sim X_1 + X_2 + ... + X_k$.)

Answer: When we hold the proportion of black voting-age population in the district constant, we estimate that having a black candidate running *decreases* black turnout by 0.74 percentage points, on average. (Note: It is a decrease because we are measuring change—the change in the outcome variable caused by the treatment—and the $\widehat{\beta}_1$ is negative.) (Important note: The causal interpretation of this coefficient would be wrong if there are confounding variables other than *proportion_black*.)

c. Given this last analysis, would you conclude that having a black candidate running increases black turnout? A yes/no answer will suffice. (5 points)

Answer: No, I would not conclude that having a black candidate running increases black turnout. Once we control for the confounder *proportion_black*, the effect disappears. In fact, the effect reverses directions and becomes very small.