Data 102 Final Project Checkpoint 2

December 2, 2021

1 Checkpoint 2

This notebook includes our work on checkpoint 2 of the Data 102 Final Project.

Collaborators and Student IDs Alan Jian 3033730509

Shreya Chowdhury 3033623454

Matilda Ju 3033728143

Yannie Li 3034042574

1.1 Final Study Design

In our final study design we decided upon the following:

Outcome: Score based on the number of scenarios in which a respondent supported police brutality (more details in later section)

Confounders: INCOME, AGE, SEX, EDUC, MARITAL, WRKSTAT

Treatment Variable: Whether or not an individual identifies as African American.

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

2 Data Cleaning

Here, we take a bunch of steps to clean up our data and make it suitable for future work. Here are a series of steps that we took:

2.1 Importing Data

We started by importing the data. The columns we are interested in include YEAR, RACE, INCOME, POLABUSEY, POLMURDR, POLESCAP, POLATTAKY, POLHITOKY.

If we separate out the variables based on what they represent in our causal inference, they are as follows:

These will form the basis of our outcome variable: POLABUSE, POLMURDR, POLESCAP, POLATTAK, POLHITOK

These are our confounders: INCOME, AGE, SEX, EDUC, MARITAL, WRKSTAT,

This will form the foundation of our treatment variable: RACE

[59]:
```
y = ['POLABUSE', 'POLMURDR', 'POLESCAP', 'POLATTAK', 'POLHITOK']
x = ['INCOME', 'AGE', 'SEX', 'EDUC', 'MARITAL', 'WRKSTAT']
z = ['RACE']

labels = ['YEAR']
labels.extend(y)
labels.extend(x)
labels.extend(z)
```

[109]:
```
origlocation = "./GSS/GSS7218_R3"
alanLoc = "~/Downloads/GSS_spss/GSS7218_R3.sav"

causal = pd.read_spss(alanLoc, usecols=labels,
                      convert_categoricals=True)
```

2.2 Initial Transformations

Additionally, we make some further modifications, chopping off any data prior to 2000s (because we're interested in the period between 2000-2018).

[110]:
```
causal = causal[causal['YEAR'] >= 2000]
causal.head()
```

[110]:
```
         YEAR          WRKSTAT        MARITAL AGE  EDUC     SEX   RACE  \
38116  2000.0  WORKING FULLTIME  NEVER MARRIED  26  16.0    MALE  WHITE
38117  2000.0  WORKING FULLTIME       DIVORCED  48  15.0  FEMALE  WHITE
38118  2000.0     KEEPING HOUSE        WIDOWED  67  13.0  FEMALE  WHITE
38119  2000.0  WORKING FULLTIME  NEVER MARRIED  39  14.0  FEMALE  WHITE
38120  2000.0  WORKING FULLTIME       DIVORCED  25  14.0  FEMALE  WHITE

                INCOME POLHITOK POLABUSE POLMURDR POLESCAP POLATTAK
38116              NaN      NaN      NaN      NaN      NaN      NaN
38117    $8000 TO 9999       NO       NO       NO       NO      YES
38118  $15000 - 19999      YES       NO       NO      YES      YES
38119  $25000 OR MORE      YES       NO       NO      YES      YES
38120  $25000 OR MORE      YES       NO      NaN      NaN      YES
```

[111]:
```
# sanity check to make sure we're including all the right years.
sorted(causal['YEAR'].value_counts().index)
```

[111]:
```
[2000.0,
 2002.0,
```

```
 2004.0,
 2006.0,
 2008.0,
 2010.0,
 2012.0,
 2014.0,
 2016.0,
 2018.0]
```

We were interested in police brutality, and we wanted to compile a score based off of their responses to 5 different questions. In order to create a valid score, we had to limit our study to individuals who answered all 5 questions.

After removing all the individuals who did not respond to all of the questions, we are left with 12433 responses.

[112]:
```python
causal[y].dropna()
```

[112]:
```
      POLABUSE POLMURDR POLESCAP POLATTAK POLHITOK
38117       NO       NO       NO      YES       NO
38118       NO       NO      YES      YES      YES
38119       NO       NO      YES      YES      YES
38121       NO       NO      YES      YES      YES
38123       NO       NO      YES      YES      YES
…         …      …      …      …      …
64804       NO       NO      YES      YES      YES
64806      YES       NO      YES      YES      YES
64808      YES      YES      YES      YES      YES
64811       NO       NO      YES      YES      YES
64812       NO       NO       NO      YES      YES

[12433 rows x 5 columns]
```

[113]:
```python
causal = causal.loc[causal[y].dropna().index]
```

2.3 Visualizing and Understanding NaNs

Additionally, we tried to understand whether any of the variables we wanted to control for, or any of the treatment or response variables had any NaNs that we had to worry about.

Any columns containing NaNs were then removed, since we were interested in potentially going for exact matching.

[114]:
```python
causal.isna()['WRKSTAT'].value_counts()
```

[114]:
```
False    12427
True         6
Name: WRKSTAT, dtype: int64
```

[115]:
```
causal.isna()['MARITAL'].value_counts()
```

[115]:
```
False    12428
True         5
Name: MARITAL, dtype: int64
```

[116]:
```
causal.isna()['AGE'].value_counts()
```

[116]:
```
False    12400
True        33
Name: AGE, dtype: int64
```

[117]:
```
causal.isna()['EDUC'].value_counts()
```

[117]:
```
False    12415
True        18
Name: EDUC, dtype: int64
```

[118]:
```
causal.isna()['SEX'].value_counts()
```

[118]:
```
False    12433
Name: SEX, dtype: int64
```

[119]:
```
causal.isna()['YEAR'].value_counts()
```

[119]:
```
False    12433
Name: YEAR, dtype: int64
```

[120]:
```
causal.isna()['RACE'].value_counts()
```

[120]:
```
False    12433
Name: RACE, dtype: int64
```

We noticed that the `INCOME` column had a bunch of NaNs, so we wanted to see if nonresponse was correlated with anything like race or year.

[121]:
```
causal.isna()['INCOME'].value_counts()
```

[121]:
```
False    11022
True      1411
Name: INCOME, dtype: int64
```

[122]:
```
eda = causal.copy()
eda['INC_NAN'] = causal['INCOME'].isna().astype(int)

eda[['RACE', 'INC_NAN']].groupby('RACE').mean()
```

[122]:
```
        INC_NAN
RACE
BLACK  0.118994
OTHER  0.114414
WHITE  0.112347
```

[123]:
```
eda[['YEAR', 'INC_NAN']].groupby('YEAR').mean()
```

[123]:
```
         INC_NAN
YEAR
2000.0  0.120782
2002.0  0.073394
2004.0  0.099455
2006.0  0.118765
2008.0  0.092437
2010.0  0.102415
2012.0  0.101019
2014.0  0.081867
2016.0  0.149033
2018.0  0.146341
```

Looks like NaN frequency is roughly uncorrelated with year and race!

Now, I'm going to remove a bunch of the NaNs. Since we want to do exact matching, we have no tolerance for non-response.

[128]:
```
causal = causal.dropna()
causal.head()
```

[128]:
```
         YEAR           WRKSTAT        MARITAL AGE  EDUC     SEX   RACE  \
38117  2000.0  WORKING FULLTIME       DIVORCED  48  15.0  FEMALE  WHITE
38118  2000.0     KEEPING HOUSE        WIDOWED  67  13.0  FEMALE  WHITE
38119  2000.0  WORKING FULLTIME  NEVER MARRIED  39  14.0  FEMALE  WHITE
38123  2000.0  WORKING FULLTIME       DIVORCED  44  14.0  FEMALE  WHITE
38124  2000.0  WORKING FULLTIME        MARRIED  44  18.0    MALE  WHITE

                INCOME POLHITOK POLABUSE POLMURDR POLESCAP POLATTAK
38117   $8000 TO 9999       NO       NO       NO       NO      YES
38118  $15000 - 19999      YES       NO       NO      YES      YES
38119  $25000 OR MORE      YES       NO       NO      YES      YES
38123  $20000 - 24999      YES       NO       NO      YES      YES
38124  $25000 OR MORE      YES       NO       NO      YES      YES
```

One of the first things we're doing is combining the race. We're interested only in black vs. non-black, so the WHITE and OTHER denominations are being combined, so that the category is binary.

[129]:
```
causal['BLACK'] = causal['RACE'] == 'BLACK'
causal.head()
```

[129]:
```
         YEAR            WRKSTAT        MARITAL AGE  EDUC     SEX   RACE  \
38117  2000.0  WORKING FULLTIME       DIVORCED  48  15.0  FEMALE  WHITE
38118  2000.0     KEEPING HOUSE        WIDOWED  67  13.0  FEMALE  WHITE
38119  2000.0  WORKING FULLTIME  NEVER MARRIED  39  14.0  FEMALE  WHITE
38123  2000.0  WORKING FULLTIME       DIVORCED  44  14.0  FEMALE  WHITE
38124  2000.0  WORKING FULLTIME        MARRIED  44  18.0    MALE  WHITE

                INCOME POLHITOK POLABUSE POLMURDR POLESCAP POLATTAK  BLACK
38117   $8000 TO 9999       NO       NO       NO       NO      YES  False
38118  $15000 - 19999      YES       NO       NO      YES      YES  False
38119  $25000 OR MORE      YES       NO       NO      YES      YES  False
38123  $20000 - 24999      YES       NO       NO      YES      YES  False
38124  $25000 OR MORE      YES       NO       NO      YES      YES  False
```

When we binarize the variable, we actually note a change in the demographic make-up of our population. Only ~17% of our sample population is black, which could affect results down the road.

[130]:
```
causal['BLACK'].value_counts()
```

[130]:
```
False    9424
True     1566
Name: BLACK, dtype: int64
```

3 Formulating our Outcome Variable

So one of the design considerations that we thought a lot about it is how to capture feelings about police brutality from these 5 questions. We wanted to binarize the results of the 5 questions, so we thought of 3 thresholds:

We decided upon a simple score based on how many scenarios in which an individual supports police brutality.

The scenarios that respondents were querried about were as follows: 1. (Would you approve of a policeman striking a citizen who…) Had said vulgar and obscene things to the policeman? 2. (Would you approve of a policeman striking a citizen who…) Was being questioned as a suspect in a murder case? 3. (Would you approve of a policeman striking a citizen who…) Was attacking the policeman with his fists? 4. (Would you approve of a policeman striking a citizen who…) Was attempting to escape from custody? 5. Are there any situations you can imagine in which you would approve of a policeman striking an adult male citizen?

Respondents answered Yes or No to each of these questions. We simply summed the number of yes's to formulate our score out of 5.

[132]:
```
for label in y:
    causal[label + '_bin'] = (causal[label] == 'YES').astype(int)

causal['brutal'] = np.sum(causal[['POLABUSE_bin', 'POLMURDR_bin',␣
↪'POLESCAP_bin', 'POLATTAK_bin', 'POLHITOK_bin']], axis=1)
causal['brutal_bin'] = causal['brutal'] >= 3
causal.head()
```

[132]:
```
         YEAR          WRKSTAT        MARITAL AGE  EDUC     SEX   RACE  \
38117  2000.0  WORKING FULLTIME       DIVORCED  48  15.0  FEMALE  WHITE
38118  2000.0     KEEPING HOUSE        WIDOWED  67  13.0  FEMALE  WHITE
38119  2000.0  WORKING FULLTIME  NEVER MARRIED  39  14.0  FEMALE  WHITE
38123  2000.0  WORKING FULLTIME       DIVORCED  44  14.0  FEMALE  WHITE
38124  2000.0  WORKING FULLTIME        MARRIED  44  18.0    MALE  WHITE

                INCOME POLHITOK POLABUSE  … POLESCAP POLATTAK  BLACK  \
38117   $8000 TO 9999       NO       NO  …       NO      YES  False
38118  $15000 - 19999      YES       NO  …      YES      YES  False
38119  $25000 OR MORE      YES       NO  …      YES      YES  False
38123  $20000 - 24999      YES       NO  …      YES      YES  False
38124  $25000 OR MORE      YES       NO  …      YES      YES  False

       POLABUSE_bin  POLMURDR_bin  POLESCAP_bin  POLATTAK_bin  POLHITOK_bin  \
38117             0             0             0             1             0
38118             0             0             1             1             1
38119             0             0             1             1             1
38123             0             0             1             1             1
38124             0             0             1             1             1

       brutal  brutal_bin
38117       1       False
38118       3        True
38119       3        True
38123       3        True
38124       3        True

[5 rows x 21 columns]
```

4 Causal Inference via Matching and Conditional SDO

Here, I'm just putting some possible matching ideas that I scraped from this doc: https://www.researchgate.net/publication/46428171_Matching_Methods_for_Causal_Inference_A_Review_a

We can try: 1. 1-to-1 matching 2. k-to-1 matching

Instead, we're using the conditional SDO to estimate the conditional ATE, which we will integrate using tower property to get ATE.

Note: Matching based on bins is actually called *sub-classification*, and in certains settings, is an unbiased estimator of the ATE.

Unconfoundedness Assumption: By controlling for income, age, marital status, work status, education, and sex, a respondent's race is conditionally independent from his/her the number of scenarios in which he/she supports police brutality.

Why does the unconfoundedness assumption hold: See sociology study.

[167]:
```
test_df = causal.copy()
test_df = test_df[x + ['BLACK', 'brutal']]
test_df = test_df.groupby(x + ['BLACK']).mean()
```

[168]:
```
test_df['brutal'].value_counts()
```

[168]:
```
3.000000    3231
2.000000    1625
1.000000    1000
4.000000     503
0.000000     403
            …
1.818182       1
2.700000       1
2.615385       1
2.571429       1
2.444444       1
Name: brutal, Length: 70, dtype: int64
```

[194]:
```
cond_exp_unpaired = test_df.dropna()
cond_exp_unpaired
```

[194]:
```
                                                                       brutal
INCOME        AGE  SEX    EDUC MARITAL       WRKSTAT          BLACK
$1000 TO 2999 18.0 FEMALE 11.0 NEVER MARRIED SCHOOL           False       3.0
              19.0 FEMALE 12.0 NEVER MARRIED WORKING PARTTIME False       1.0
                          13.0 NEVER MARRIED SCHOOL           False       3.0
                   MALE   12.0 NEVER MARRIED SCHOOL           True        1.0
              20.0 FEMALE 11.0 MARRIED       KEEPING HOUSE    True        1.0
…                                                                      …
LT $1000      81.0 FEMALE 6.0  SEPARATED     RETIRED          False       2.0
              82.0 FEMALE 12.0 DIVORCED      RETIRED          True        2.0
              84.0 MALE   4.0  MARRIED       RETIRED          True        2.0
                          11.0 WIDOWED       RETIRED          False       4.0
              87.0 MALE   12.0 MARRIED       RETIRED          False       2.0

[7743 rows x 1 columns]
```

[205]:

```
# all of the matches have pairs of values
cond_exp_counts = cond_exp_unpaired.reset_index().groupby(x).size().
 ↪sort_values(ascending=False)

# now we extract the indices
indices_of_matches = cond_exp_counts[cond_exp_counts == 2].index.to_list()

# and filter our original set
cond_exp_filtered = cond_exp_unpaired.reset_index('BLACK').
 ↪loc[indices_of_matches]
```

[206]:

```
# split the columns based on treatment
black_df = cond_exp_filtered[cond_exp_filtered['BLACK']]
nblack_df = cond_exp_filtered[~cond_exp_filtered['BLACK']]

# and pair them back together via merge on confounders
cond_exp_paired = black_df.merge(nblack_df, left_index=True, right_index=True)
cond_exp_paired.head()
```

[206]:

```
                                                                BLACK_x  \
INCOME          AGE  SEX    EDUC MARITAL       WRKSTAT
$25000 OR MORE 40.0 MALE   16.0 MARRIED       WORKING FULLTIME     True
               48.0 MALE   19.0 MARRIED       WORKING FULLTIME     True
               49.0 FEMALE 12.0 MARRIED       WORKING FULLTIME     True
               31.0 FEMALE 16.0 NEVER MARRIED WORKING FULLTIME     True
$10000 - 14999 77.0 FEMALE 9.0  WIDOWED       RETIRED              True

                                                                brutal_x  \
INCOME          AGE  SEX    EDUC MARITAL       WRKSTAT
$25000 OR MORE 40.0 MALE   16.0 MARRIED       WORKING FULLTIME       2.0
               48.0 MALE   19.0 MARRIED       WORKING FULLTIME       2.0
               49.0 FEMALE 12.0 MARRIED       WORKING FULLTIME       3.0
               31.0 FEMALE 16.0 NEVER MARRIED WORKING FULLTIME       3.0
$10000 - 14999 77.0 FEMALE 9.0  WIDOWED       RETIRED                0.0

                                                                BLACK_y  \
INCOME          AGE  SEX    EDUC MARITAL       WRKSTAT
$25000 OR MORE 40.0 MALE   16.0 MARRIED       WORKING FULLTIME    False
               48.0 MALE   19.0 MARRIED       WORKING FULLTIME    False
               49.0 FEMALE 12.0 MARRIED       WORKING FULLTIME    False
               31.0 FEMALE 16.0 NEVER MARRIED WORKING FULLTIME    False
$10000 - 14999 77.0 FEMALE 9.0  WIDOWED       RETIRED             False

                                                                brutal_y
INCOME          AGE  SEX    EDUC MARITAL       WRKSTAT
$25000 OR MORE 40.0 MALE   16.0 MARRIED       WORKING FULLTIME  2.800000
               48.0 MALE   19.0 MARRIED       WORKING FULLTIME  3.000000
```

```
                 49.0 FEMALE 12.0 MARRIED       WORKING FULLTIME  1.250000
                 31.0 FEMALE 16.0 NEVER MARRIED WORKING FULLTIME  2.666667
$10000 - 14999 77.0 FEMALE 9.0  WIDOWED       RETIRED           3.000000
```

[208]:
```
# Take the ATE
cond_exp_paired['conditional ATE'] = cond_exp_paired['brutal_x'] -␣
↪cond_exp_paired['brutal_y']
cond_exp_paired
```

[208]:
```
                                                                 BLACK_x  \
INCOME         AGE  SEX    EDUC MARITAL       WRKSTAT
$25000 OR MORE 40.0 MALE   16.0 MARRIED       WORKING FULLTIME     True
               48.0 MALE   19.0 MARRIED       WORKING FULLTIME     True
               49.0 FEMALE 12.0 MARRIED       WORKING FULLTIME     True
               31.0 FEMALE 16.0 NEVER MARRIED WORKING FULLTIME     True
$10000 - 14999 77.0 FEMALE 9.0  WIDOWED       RETIRED              True
…                                                                   …
$25000 OR MORE 41.0 FEMALE 16.0 MARRIED       WORKING PARTTIME     True
                                              WORKING FULLTIME     True
               37.0 MALE   12.0 NEVER MARRIED WORKING FULLTIME     True
$8000 TO 9999  26.0 FEMALE 12.0 NEVER MARRIED WORKING FULLTIME     True
$25000 OR MORE 43.0 MALE   12.0 NEVER MARRIED WORKING FULLTIME     True

                                                                 brutal_x  \
INCOME         AGE  SEX    EDUC MARITAL       WRKSTAT
$25000 OR MORE 40.0 MALE   16.0 MARRIED       WORKING FULLTIME       2.0
               48.0 MALE   19.0 MARRIED       WORKING FULLTIME       2.0
               49.0 FEMALE 12.0 MARRIED       WORKING FULLTIME       3.0
               31.0 FEMALE 16.0 NEVER MARRIED WORKING FULLTIME       3.0
$10000 - 14999 77.0 FEMALE 9.0  WIDOWED       RETIRED                0.0
…                                                                    …
$25000 OR MORE 41.0 FEMALE 16.0 MARRIED       WORKING PARTTIME       3.0
                                              WORKING FULLTIME       3.0
               37.0 MALE   12.0 NEVER MARRIED WORKING FULLTIME       1.5
$8000 TO 9999  26.0 FEMALE 12.0 NEVER MARRIED WORKING FULLTIME       4.0
$25000 OR MORE 43.0 MALE   12.0 NEVER MARRIED WORKING FULLTIME       3.0

                                                                 BLACK_y  \
INCOME         AGE  SEX    EDUC MARITAL       WRKSTAT
$25000 OR MORE 40.0 MALE   16.0 MARRIED       WORKING FULLTIME    False
               48.0 MALE   19.0 MARRIED       WORKING FULLTIME    False
               49.0 FEMALE 12.0 MARRIED       WORKING FULLTIME    False
               31.0 FEMALE 16.0 NEVER MARRIED WORKING FULLTIME    False
$10000 - 14999 77.0 FEMALE 9.0  WIDOWED       RETIRED             False
…                                                                 …
$25000 OR MORE 41.0 FEMALE 16.0 MARRIED       WORKING PARTTIME    False
                                              WORKING FULLTIME    False
```

```
                37.0 MALE   12.0 NEVER MARRIED WORKING FULLTIME    False
$8000 TO 9999   26.0 FEMALE 12.0 NEVER MARRIED WORKING FULLTIME    False
$25000 OR MORE 43.0 MALE   12.0 NEVER MARRIED WORKING FULLTIME    False

                                                                  brutal_y  \
INCOME         AGE  SEX    EDUC MARITAL       WRKSTAT
$25000 OR MORE 40.0 MALE   16.0 MARRIED       WORKING FULLTIME  2.800000
               48.0 MALE   19.0 MARRIED       WORKING FULLTIME  3.000000
               49.0 FEMALE 12.0 MARRIED       WORKING FULLTIME  1.250000
               31.0 FEMALE 16.0 NEVER MARRIED WORKING FULLTIME  2.666667
$10000 - 14999 77.0 FEMALE 9.0  WIDOWED       RETIRED           3.000000
…                                                                    …
$25000 OR MORE 41.0 FEMALE 16.0 MARRIED       WORKING PARTTIME  3.000000
                                              WORKING FULLTIME  2.500000
               37.0 MALE   12.0 NEVER MARRIED WORKING FULLTIME  2.750000
$8000 TO 9999  26.0 FEMALE 12.0 NEVER MARRIED WORKING FULLTIME  0.000000
$25000 OR MORE 43.0 MALE   12.0 NEVER MARRIED WORKING FULLTIME  2.500000

                                                                  conditional ATE
INCOME         AGE  SEX    EDUC MARITAL       WRKSTAT
$25000 OR MORE 40.0 MALE   16.0 MARRIED       WORKING FULLTIME        -0.800000
               48.0 MALE   19.0 MARRIED       WORKING FULLTIME        -1.000000
               49.0 FEMALE 12.0 MARRIED       WORKING FULLTIME         1.750000
               31.0 FEMALE 16.0 NEVER MARRIED WORKING FULLTIME         0.333333
$10000 - 14999 77.0 FEMALE 9.0  WIDOWED       RETIRED                 -3.000000
…                                                                           …
$25000 OR MORE 41.0 FEMALE 16.0 MARRIED       WORKING PARTTIME         0.000000
                                              WORKING FULLTIME         0.500000
               37.0 MALE   12.0 NEVER MARRIED WORKING FULLTIME        -1.250000
$8000 TO 9999  26.0 FEMALE 12.0 NEVER MARRIED WORKING FULLTIME         4.000000
$25000 OR MORE 43.0 MALE   12.0 NEVER MARRIED WORKING FULLTIME         0.500000

[463 rows x 5 columns]
```

[203]:
```
cond_exp_paired = black_df.merge(nblack_df, left_index=True, right_index=True)
cond_exp_paired['conditional ATE'] = cond_exp_paired['brutal_x'] -␣
 ↪cond_exp_paired['brutal_y']
cond_exp_paired.head()
```

[203]:
```
                                                                  BLACK_x  \
INCOME         AGE  SEX    EDUC MARITAL       WRKSTAT
$25000 OR MORE 40.0 MALE   16.0 MARRIED       WORKING FULLTIME     True
               48.0 MALE   19.0 MARRIED       WORKING FULLTIME     True
               49.0 FEMALE 12.0 MARRIED       WORKING FULLTIME     True
               31.0 FEMALE 16.0 NEVER MARRIED WORKING FULLTIME     True
$10000 - 14999 77.0 FEMALE 9.0  WIDOWED       RETIRED              True
```

```
                                                              brutal_x  \
INCOME          AGE  SEX    EDUC MARITAL       WRKSTAT
$25000 OR MORE 40.0 MALE    16.0 MARRIED       WORKING FULLTIME       2.0
               48.0 MALE    19.0 MARRIED       WORKING FULLTIME       2.0
               49.0 FEMALE 12.0 MARRIED       WORKING FULLTIME       3.0
               31.0 FEMALE 16.0 NEVER MARRIED WORKING FULLTIME       3.0
$10000 - 14999 77.0 FEMALE 9.0  WIDOWED       RETIRED                0.0

                                                              BLACK_y  \
INCOME          AGE  SEX    EDUC MARITAL       WRKSTAT
$25000 OR MORE 40.0 MALE    16.0 MARRIED       WORKING FULLTIME    False
               48.0 MALE    19.0 MARRIED       WORKING FULLTIME    False
               49.0 FEMALE 12.0 MARRIED       WORKING FULLTIME    False
               31.0 FEMALE 16.0 NEVER MARRIED WORKING FULLTIME    False
$10000 - 14999 77.0 FEMALE 9.0  WIDOWED       RETIRED             False

                                                              brutal_y  \
INCOME          AGE  SEX    EDUC MARITAL       WRKSTAT
$25000 OR MORE 40.0 MALE    16.0 MARRIED       WORKING FULLTIME  2.800000
               48.0 MALE    19.0 MARRIED       WORKING FULLTIME  3.000000
               49.0 FEMALE 12.0 MARRIED       WORKING FULLTIME  1.250000
               31.0 FEMALE 16.0 NEVER MARRIED WORKING FULLTIME  2.666667
$10000 - 14999 77.0 FEMALE 9.0  WIDOWED       RETIRED           3.000000

                                                              conditional ATE
INCOME          AGE  SEX    EDUC MARITAL       WRKSTAT
$25000 OR MORE 40.0 MALE    16.0 MARRIED       WORKING FULLTIME        -0.800000
               48.0 MALE    19.0 MARRIED       WORKING FULLTIME        -1.000000
               49.0 FEMALE 12.0 MARRIED       WORKING FULLTIME         1.750000
               31.0 FEMALE 16.0 NEVER MARRIED WORKING FULLTIME         0.333333
$10000 - 14999 77.0 FEMALE 9.0  WIDOWED       RETIRED                 -3.000000
```

Weighting by true prevalence needed to get ATE!

[233]:
```
# indices required of us
indices_for_tower = cond_exp_paired.index.to_list()

# find the prevalence from the GSS survey
bin_counts = causal.groupby(x).size()
prevalence = pd.DataFrame(bin_counts / np.sum(bin_counts))
prevalence.columns = ['Prevalence']

# look up prevalence values of interest
prevalence = prevalence.loc[indices_for_tower]

# Normalize in anticipation of calculating expectation
norm_prevalence = prevalence / np.sum(prevalence)
```

[236]:
```python
ATE_df = cond_exp_paired.merge(norm_prevalence, left_index=True,␣
 ↪right_index=True)
np.sum(ATE_df['conditional ATE'] * ATE_df['Prevalence'])
```

[236]: -0.5211832566916026

5 Results

Summarize and interpret your results, providing a clear statement about causality (or a lack thereof) including any assumptions necessary.

Assuming unconfoundedness given income, age, sex, education level, marital status, and work status, we found there to be a negative causal relationship between race and opinion of police brutality.

Given our ATE of -0.52, this indicates that Black Americans support police brutality in slightly fewer scenarios compared to those of non-Black Americans.

Where possible, discuss the uncertainty in your estimate and/or the evidence against the hypotheses you are investigating.

See our discussion below for some of the issues that we had in formulating the study. These represent the largest sources of uncertainty in our estimate of the ATE.

6 Discussion

Elaborate on the limitations of your methods.

- We're conditioning on a shit income thing
  - Bins are horribly designed, really only care about gradations of poverty
  - Level of non-response was much higher here compared to the rest of the questions
- We're using exact matching on sub-classifications bins; it's not exact
- Level of non-response

Our study has one primary flaw: the confounding variable of income. Although the literature says definitively that income is highly correlated with race and has a causal relationship with opinions on police brutality, the confounding variable that we controlled for here is poorly organized for this use case. The income variable is badly binned, with a majority of individuals self-identifying a salary above 25000 for obvious reasons. The question asked in the survey was clearly more interested in those living in poverty, since there were several bins that divided up income levels between $0-25000 in annual salary. As a result, the unconfoundedness assumption likely does not completely hold in our study, since the income bins only limited matches in respondents who earned an income far below the poverty line.

Which additional data would be useful?

- Additional confounders:
  - Income that actually is representative (or exact)
  - Personal experience with law enforcement is a confounder!
  - Neighborhood

- Outcome Variable:
  - More general question about broad support for police brutality
    - We had to create one for the study, and answers are highly correlated

It would be nice if we could get a more representative income variable. Unconfoundedness does not hold in our causal inference question because the quality of this variable is so poor, and does not constrain our matching enough to be useful. If we had a better income variable, controlling for income's confounding effects would be much more effective.

Additionally, there are some other confounders that we did not consider. Since interactions with police can be extremely correlated with race and opinions of police brutality (if you're constantly getting arrested or having run-ins with police there might be a differential amount of empathy you feel in this case), it would have been nice to have access to this information for purposes of matching.

The questions we were able to explore about police brutality were very specific. For example, they generally follow the format "Would you approve of a policeman striking a citizen who…" We would have been more interested in exploring DEFUND, which asks if people favor or oppose reducing funding for police departments (which didn't have enough responses for us to use), or more general questions about police brutality.

How confident are you that there's a causal relationship between your chosen treatment and outcome? Why?

Since our finding is supported in sociology, we are highly confident that there is a causal relationship between race and opinions on police brutality. If there would be an uncertainty that we have about our result, it would be the fact that the ATE is calculated only on the sample data that we have; and given its closeness to 0, it wouldn't be a surprise to see that the 95% confidence interval on the ATE would include 0, indicating a chance of no causal effect.

[ ]:
```
```