# PSet1_Q4_ARE213

October 2, 2023

[1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
#import data packages


import statsmodels.api as sm
#import regression packages
import statsmodels.formula.api as smf
from statsmodels.tools.sm_exceptions import ConvergenceWarning

from sklearn.datasets import load_iris
from sklearn.linear_model import LogisticRegression
```

[2]:
```
d = pd.read_csv('clean_pset1.csv')
```

[18]:
```
#re-define variables\

#variable classification

#outcome
y = ['dbrwt']
#treatment
D = ['tobacco']

#cor with y and D
x1 = ['alcohol', 'mrace3_2', 'mrace3_3', 'ormothhis', 'adeq_2.0', 'adeq_3.0',␣
 ↪'cardiac', 'pre4000', 'phyper',
     'diabetes', 'anemia', 'lung', 'dlivord', 'educ_0.0', 'educ_1.0', 'educ_2.
 ↪0','dmage', 'dmar','tot_2.0',
      'tot_3.0','tot_4.0','tot_5.0','tot_6.0','tot_7.0','tot_8.0','live_1.0',␣
 ↪'live_2.0', 'live_3.0','live_4.0',
      'live_5.0','live_6.0','live_7.0','live_8.0', 'live_9.0']

#cor with y not D
x3 = ['dgestat', 'csex', 'plur_1']
```

[19]:

```python
#running logit to get p-scores
vals = sm.Logit(d[D],sm.add_constant(d[x1+x3]))
out = vals.fit()
print(out.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.376842
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:                tobacco   No. Observations:               114610
Model:                          Logit   Df Residuals:                   114572
Method:                           MLE   Df Model:                           37
Date:                Sun, 01 Oct 2023   Pseudo R-squ.:                  0.1409
Time:                        12:28:40   Log-Likelihood:                -43190.
converged:                       True   LL-Null:                       -50272.
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.0569      1.037     -1.019      0.308      -3.090       0.976
alcohol        1.8939      0.069     27.555      0.000       1.759       2.029
mrace3_2      -1.5274      0.126    -12.167      0.000      -1.773      -1.281
mrace3_3      -1.1389      0.029    -38.747      0.000      -1.197      -1.081
ormothhis     -1.3997      0.055    -25.539      0.000      -1.507      -1.292
adeq_2.0       0.1186      0.021      5.602      0.000       0.077       0.160
adeq_3.0       0.2902      0.038      7.581      0.000       0.215       0.365
cardiac       -0.0906      0.111     -0.819      0.413      -0.307       0.126
pre4000       -0.7352      0.090     -8.212      0.000      -0.911      -0.560
phyper        -0.4186      0.059     -7.098      0.000      -0.534      -0.303
diabetes       0.0699      0.054      1.297      0.195      -0.036       0.175
anemia         0.1082      0.079      1.370      0.171      -0.047       0.263
lung           0.1703      0.093      1.830      0.067      -0.012       0.353
dlivord       -0.0200      0.015     -1.297      0.195      -0.050       0.010
educ_0.0       0.8393      1.027      0.817      0.414      -1.174       2.853
educ_1.0       1.6433      1.026      1.602      0.109      -0.367       3.653
educ_2.0       0.5682      1.026      0.554      0.580      -1.442       2.579
dmage         -0.0305      0.002    -14.464      0.000      -0.035      -0.026
dmar          -1.1809      0.022    -52.722      0.000      -1.225      -1.137
tot_2.0        0.4948      0.033     15.089      0.000       0.430       0.559
tot_3.0        0.7744      0.038     20.497      0.000       0.700       0.848
tot_4.0        0.9328      0.044     21.140      0.000       0.846       1.019
tot_5.0        1.1490      0.054     21.445      0.000       1.044       1.254
tot_6.0        1.2160      0.068     17.941      0.000       1.083       1.349
tot_7.0        1.5290      0.088     17.293      0.000       1.356       1.702
tot_8.0        1.2856      0.103     12.482      0.000       1.084       1.488
live_1.0      -0.2147      0.126     -1.708      0.088      -0.461       0.032
live_2.0       0.1220      0.090      1.355      0.175      -0.054       0.298
```

```
live_3.0       0.0320      0.049      0.657      0.511      -0.064       0.128
live_4.0      -0.1302      0.046     -2.802      0.005      -0.221      -0.039
live_5.0      -0.1620      0.042     -3.900      0.000      -0.243      -0.081
live_6.0      -0.0638      0.043     -1.469      0.142      -0.149       0.021
live_7.0      -0.0055      0.049     -0.112      0.911      -0.101       0.090
live_8.0       0.1200      0.055      2.197      0.028       0.013       0.227
live_9.0       0.4235      0.043      9.965      0.000       0.340       0.507
dgestat       -0.0193      0.003     -5.603      0.000      -0.026      -0.013
csex           0.0183      0.017      1.055      0.292      -0.016       0.052
plur_1         0.1531      0.072      2.136      0.033       0.013       0.294
==============================================================================
```

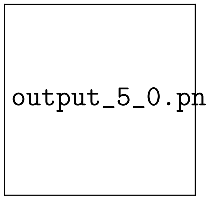
[9]:

```
#obtaining p-scores by prediction

post = out.predict(sm.add_constant(d[x1+x3]))
d['post'] = post
```

```
count    114610.000000
mean          0.159375
std           0.135541
min           0.002354
25%           0.058771
50%           0.116412
75%           0.214498
max           0.950695
Name: post, dtype: float64
```

[10]:

```
#plotting overlap of p-scores

plt.hist(d[d['tobacco']==1]['post'], bins =30, alpha = 0.5)
plt.hist(d[d['tobacco']==0]['post'], bins = 30, alpha = 0.5)
plt.show()
```



[12]:

```
#creating bins of unequal size, but equal distances in propensity score
bin_sizes = [(0,.1)]
#creating bins
for i in range(1, 10, 1):
    bin_sizes.append((i/10, (i+1)/10))
```

```
#cutting dataset into bins
bins2 = pd.IntervalIndex.from_tuples(bin_sizes)
d['cat'] = pd.cut(d['post'], bins2)


#indexing into bin dummies
df2 = pd.get_dummies(d['cat'], prefix = 'cat')
```

[13]:

```
#testing balance in phat
for item in x1+x3:

    vals = sm.OLS(d[item], sm.add_constant(pd.concat([d['tobacco'],df2[['cat_(0.
 ↪1, 0.2]', 'cat_(0.2, 0.3]', 'cat_(0.3, 0.4]',
       'cat_(0.4, 0.5]', 'cat_(0.5, 0.6]', 'cat_(0.6, 0.7]', 'cat_(0.7, 0.8]',
       'cat_(0.8, 0.9]', 'cat_(0.9, 1.0]']]], axis = 1)))
    out = vals.fit()
    print(item)
    print('coefficients')
    print(out.params[1:2])
    print('standard errors')
    print(out.bse[1:2])
    print('test stat')
    print(out.params[1:2]/out.bse[1:2])
```

```
alcohol
coefficients
tobacco   -0.000562
dtype: float64
standard errors
tobacco    0.000676
dtype: float64
test stat
tobacco   -0.832203
dtype: float64
mrace3_2
coefficients
tobacco   -0.004851
dtype: float64
standard errors
tobacco    0.001196
dtype: float64
test stat
tobacco   -4.056366
dtype: float64
mrace3_3
coefficients
tobacco   -0.002303
```

```
dtype: float64
standard errors
tobacco    0.002788
dtype: float64
test stat
tobacco   -0.825951
dtype: float64
ormothhis
coefficients
tobacco   -0.002634
dtype: float64
standard errors
tobacco    0.001603
dtype: float64
test stat
tobacco   -1.643124
dtype: float64
adeq_2.0
coefficients
tobacco    0.004483
dtype: float64
standard errors
tobacco    0.003481
dtype: float64
test stat
tobacco    1.287901
dtype: float64
adeq_3.0
coefficients
tobacco    0.000687
dtype: float64
standard errors
tobacco    0.001759
dtype: float64
test stat
tobacco    0.390506
dtype: float64
cardiac
coefficients
tobacco   -0.000102
dtype: float64
standard errors
tobacco    0.000717
dtype: float64
test stat
tobacco   -0.142569
dtype: float64
pre4000
```

```
coefficients
tobacco   -0.001094
dtype: float64
standard errors
tobacco    0.001034
dtype: float64
test stat
tobacco   -1.057491
dtype: float64
phyper
coefficients
tobacco   -0.001306
dtype: float64
standard errors
tobacco    0.001492
dtype: float64
test stat
tobacco   -0.875351
dtype: float64
diabetes
coefficients
tobacco   -0.00003
dtype: float64
standard errors
tobacco    0.001402
dtype: float64
test stat
tobacco   -0.021381
dtype: float64
anemia
coefficients
tobacco    0.0002
dtype: float64
standard errors
tobacco    0.000862
dtype: float64
test stat
tobacco    0.231509
dtype: float64
lung
coefficients
tobacco    0.000109
dtype: float64
standard errors
tobacco    0.000734
dtype: float64
test stat
tobacco    0.148679
```

```
dtype: float64
dlivord
coefficients
tobacco    0.031139
dtype: float64
standard errors
tobacco    0.009849
dtype: float64
test stat
tobacco    3.161608
dtype: float64
educ_0.0
coefficients
tobacco   -0.000635
dtype: float64
standard errors
tobacco    0.001305
dtype: float64
test stat
tobacco   -0.486554
dtype: float64
educ_1.0
coefficients
tobacco    0.003936
dtype: float64
standard errors
tobacco    0.002682
dtype: float64
test stat
tobacco    1.467148
dtype: float64
educ_2.0
coefficients
tobacco   -0.003244
dtype: float64
standard errors
tobacco    0.002665
dtype: float64
test stat
tobacco   -1.216897
dtype: float64
dgestat
coefficients
tobacco   -0.014529
dtype: float64
standard errors
tobacco    0.021169
dtype: float64
```

```
test stat
tobacco   -0.686324
dtype: float64
dmage
coefficients
tobacco   -0.055068
dtype: float64
standard errors
tobacco    0.045574
dtype: float64
test stat
tobacco   -1.208326
dtype: float64
dmar
coefficients
tobacco   -0.010261
dtype: float64
standard errors
tobacco    0.002869
dtype: float64
test stat
tobacco   -3.577171
dtype: float64
csex
coefficients
tobacco    0.000324
dtype: float64
standard errors
tobacco    0.004334
dtype: float64
test stat
tobacco    0.074801
dtype: float64
tot_2.0
coefficients
tobacco    0.009262
dtype: float64
standard errors
tobacco    0.003924
dtype: float64
test stat
tobacco    2.36021
dtype: float64
tot_3.0
coefficients
tobacco    0.010606
dtype: float64
standard errors
```

```
tobacco    0.003357
dtype: float64
test stat
tobacco    3.159059
dtype: float64
tot_4.0
coefficients
tobacco    0.003781
dtype: float64
standard errors
tobacco    0.002532
dtype: float64
test stat
tobacco    1.49312
dtype: float64
tot_5.0
coefficients
tobacco    0.000844
dtype: float64
standard errors
tobacco    0.001795
dtype: float64
test stat
tobacco    0.470239
dtype: float64
tot_6.0
coefficients
tobacco    0.000169
dtype: float64
standard errors
tobacco    0.001273
dtype: float64
test stat
tobacco    0.132755
dtype: float64
tot_7.0
coefficients
tobacco    0.000098
dtype: float64
standard errors
tobacco    0.000852
dtype: float64
test stat
tobacco    0.11497
dtype: float64
tot_8.0
coefficients
tobacco    0.000002
```

```
dtype: float64
standard errors
tobacco    0.000879
dtype: float64
test stat
tobacco    0.002688
dtype: float64
live_1.0
coefficients
tobacco    0.000168
dtype: float64
standard errors
tobacco    0.000808
dtype: float64
test stat
tobacco    0.207916
dtype: float64
live_2.0
coefficients
tobacco    0.000332
dtype: float64
standard errors
tobacco    0.000756
dtype: float64
test stat
tobacco    0.438571
dtype: float64
live_3.0
coefficients
tobacco    0.002331
dtype: float64
standard errors
tobacco    0.001984
dtype: float64
test stat
tobacco    1.174588
dtype: float64
live_4.0
coefficients
tobacco    0.003139
dtype: float64
standard errors
tobacco    0.002374
dtype: float64
test stat
tobacco    1.32241
dtype: float64
live_5.0
```

```
coefficients
tobacco    0.004676
dtype: float64
standard errors
tobacco    0.003076
dtype: float64
test stat
tobacco    1.520385
dtype: float64
live_6.0
coefficients
tobacco    0.004464
dtype: float64
standard errors
tobacco    0.002597
dtype: float64
test stat
tobacco    1.718783
dtype: float64
live_7.0
coefficients
tobacco    0.002656
dtype: float64
standard errors
tobacco    0.002016
dtype: float64
test stat
tobacco    1.317487
dtype: float64
live_8.0
coefficients
tobacco    0.001338
dtype: float64
standard errors
tobacco    0.001602
dtype: float64
test stat
tobacco    0.835308
dtype: float64
live_9.0
coefficients
tobacco    0.000582
dtype: float64
standard errors
tobacco    0.002359
dtype: float64
test stat
tobacco    0.246871
```

```
dtype: float64
plur_1
coefficients
tobacco    0.000088
dtype: float64
standard errors
tobacco    0.001396
dtype: float64
test stat
tobacco    0.062732
dtype: float64
```

[14]:

```
#ATE estimation
bins = ['cat_(0.0, 0.1]', 'cat_(0.1, 0.2]', 'cat_(0.2, 0.3]', 'cat_(0.3, 0.4]',
       'cat_(0.4, 0.5]', 'cat_(0.5, 0.6]', 'cat_(0.6, 0.7]', 'cat_(0.7, 0.8]',
       'cat_(0.8, 0.9]', 'cat_(0.9, 1.0]']
#initialize helper code
meandifs = []
totaldif = 0.0

#loop over bins
for item in bins:
    #interact bins with tobacco
    d[item] = df2[item]
    d[item+'tobacco']= df2[item]*d['tobacco']
    print('block')
    print(item[5:-1])
    print('mean difference in birthweight between smokers and nonsmokers within␣
 ↪bin')
    #difference means within bins
    print(d[d[item+'tobacco']==1]['dbrwt'].mean() -␣
 ↪d[d[item+'tobacco']+d[item]==1]['dbrwt'].mean())
    print('number of observations in bin')
    #get weights of bins
    print(d[item].sum())
    #append and add to helper code
    meandifs.append(d[d[item+'tobacco']==1]['dbrwt'].mean() -␣
 ↪d[d[item+'tobacco']+d[item]==1]['dbrwt'].mean())
    #weight difference by number of items in bin
    totaldif = totaldif + (d[d[item+'tobacco']==1]['dbrwt'].mean() -␣
 ↪d[d[item+'tobacco']+d[item]==1]['dbrwt'].mean())*d[item].sum()

#report all mean differences
print(meandifs)
print('ATE')
#report avg of differences
print(totaldif/d.shape[:1])
```

```
block
0.0, 0.1
mean difference in birthweight between smokers and nonsmokers within block
-214.03892322082083
number of observations in block
51636
block
0.1, 0.2
mean difference in birthweight between smokers and nonsmokers within block
-217.51571334050413
number of observations in block
30831
block
0.2, 0.3
mean difference in birthweight between smokers and nonsmokers within block
-213.88813757182743
number of observations in block
15990
block
0.3, 0.4
mean difference in birthweight between smokers and nonsmokers within block
-187.36128847081727
number of observations in block
8112
block
0.4, 0.5
mean difference in birthweight between smokers and nonsmokers within block
-157.58378092833573
number of observations in block
4272
block
0.5, 0.6
mean difference in birthweight between smokers and nonsmokers within block
-197.5916129653915
number of observations in block
2614
block
0.6, 0.7
mean difference in birthweight between smokers and nonsmokers within block
-284.9227038183694
number of observations in block
711
block
0.7, 0.8
mean difference in birthweight between smokers and nonsmokers within block
-124.78041237113393
number of observations in block
244
```

```
block
0.8, 0.9
mean difference in birthweight between smokers and nonsmokers within block
-107.8960983884649
number of observations in block
149
block
0.9, 1.0
mean difference in birthweight between smokers and nonsmokers within block
-373.75
number of observations in block
51
[-214.03892322082083, -217.51571334050413, -213.88813757182743,
-187.36128847081727, -157.58378092833573, -197.5916129653915,
-284.9227038183694, -124.78041237113393, -107.8960983884649, -373.75]
ATE
[-210.76828819]
```

[15]:

```
#ATT
bins = ['cat_(0.0, 0.1]', 'cat_(0.1, 0.2]', 'cat_(0.2, 0.3]', 'cat_(0.3, 0.4]',
       'cat_(0.4, 0.5]', 'cat_(0.5, 0.6]', 'cat_(0.6, 0.7]', 'cat_(0.7, 0.8]',
       'cat_(0.8, 0.9]', 'cat_(0.9, 1.0]']

#initialize helpers
meandifs = []
totaldif = 0.0

#loop over bins
for item in bins:
    #interact bin with tobacco
    d[item] = df2[item]
    d[item+'tobacco']= df2[item]*d['tobacco']

    print('block')
    print(item[5:-1])

    #difference in means within block
    print('mean difference in birthweight between smokers and nonsmokers within␣
 ↪block')
    print(d[d[item+'tobacco']==1]['dbrwt'].mean() -␣
 ↪d[d[item+'tobacco']+d[item]==1]['dbrwt'].mean())
    print('number of smokers in block')
    #find number of smokers in block
    print(d[item+'tobacco'].sum())
    #append to helper code
    meandifs.append(d[d[item+'tobacco']==1]['dbrwt'].mean() -␣
 ↪d[d[item+'tobacco']+d[item]==1]['dbrwt'].mean())
```

```
    #weight by number of smokers
    totaldif = totaldif + (d[d[item+'tobacco']==1]['dbrwt'].mean() -␣
 ↪d[d[item+'tobacco']+d[item]==1]['dbrwt'].mean())*d[item+'tobacco'].sum()

print(meandifs)
print('ATT')
#report final average
print(totaldif/d['tobacco'].sum())
```

```
block
0.0, 0.1
mean difference in birthweight between smokers and nonsmokers within block
-214.03892322082083
number of smokers in block
2754.0
block
0.1, 0.2
mean difference in birthweight between smokers and nonsmokers within block
-217.51571334050413
number of smokers in block
4650.0
block
0.2, 0.3
mean difference in birthweight between smokers and nonsmokers within block
-213.88813757182743
number of smokers in block
3987.0
block
0.3, 0.4
mean difference in birthweight between smokers and nonsmokers within block
-187.36128847081727
number of smokers in block
2763.0
block
0.4, 0.5
mean difference in birthweight between smokers and nonsmokers within block
-157.58378092833573
number of smokers in block
1897.0
block
0.5, 0.6
mean difference in birthweight between smokers and nonsmokers within block
-197.5916129653915
number of smokers in block
1386.0
block
0.6, 0.7
mean difference in birthweight between smokers and nonsmokers within block
```

```
-284.9227038183694
number of smokers in block
456.0
block
0.7, 0.8
mean difference in birthweight between smokers and nonsmokers within block
-124.78041237113393
number of smokers in block
194.0
block
0.8, 0.9
mean difference in birthweight between smokers and nonsmokers within block
-107.8960983884649
number of smokers in block
131.0
block
0.9, 1.0
mean difference in birthweight between smokers and nonsmokers within block
-373.75
number of smokers in block
48.0
[-214.03892322082083, -217.51571334050413, -213.88813757182743,
-187.36128847081727, -157.58378092833573, -197.5916129653915,
-284.9227038183694, -124.78041237113393, -107.8960983884649, -373.75]
ATT
-204.22464994790207
```

[16]:

```python
#4e

#Di/p(Xi)
d['wt'] = d['tobacco'] / d['post']

#(1-Di)/(1-p(Xi))
d['wt2'] = (1 - d['tobacco']) /(1 -  d['post'])

#YiDi/p(Xi)
d['plugin'] = d['dbrwt']*d['wt']

#Yi(1-Di)/(1-p(Xi))
d['plugin2'] = d['dbrwt']*d['wt2']

#adding, dividing, and subtracting to get "improved performance" ATE
print('reweighted ATE')
ATE = (d['plugin'].sum()/d['wt'].sum()) - (d['plugin2'].sum()/d['wt2'].sum())
print(ATE)
```

```
reweighted ATE
-213.16820976094368
```

[17]:

```
d['D'] = d['tobacco']

d['X'] = ((d['tobacco'] - d['post'])*d['dbrwt'])/(1 -  d['post'])

ATT = d['X'].sum()/d['D'].sum()
print('ATT- Wooldridge')
print(ATT)
```

```
ATT- Wooldridge
-199.8735150613276
```

[ ]:

```
```