PSet1_Q2_ARE213

October 2, 2023

[11]:
```
import pandas as pd
import numpy as np
#import data packages
import statsmodels.api as sm
#import regression packages
```

[6]:
```
data = pd.read_csv('clean_pset1.csv')
```

[9]:
```
#calculate mean diff
mean_dif = data[data['tobacco'] == 1]['dbrwt'].mean() - data[data['tobacco'] ==␣
 ↪0]['dbrwt'].mean()
print(mean_dif)

#calculate se for means (not robust)
se = mean_dif = np.sqrt((float(data[data['tobacco'] == 1]['dbrwt'].std())**2)/
 ↪float(data[data['tobacco'] == 1]['dbrwt'].size) +
                        float((data[data['tobacco'] == 0]['dbrwt'].std())**2)/
 ↪float(data[data['tobacco']==0]['dbrwt'].size))
print(se)
```

```
-240.47781100396332
4.6266507330346105
```

[14]:
```
#or... do it via regression!
vals = sm.OLS(data[y], sm.add_constant(data[D]))
out = vals.fit(cov_type = 'HC0')
print(out.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  dbrwt   R-squared:                       0.023
Model:                            OLS   Adj. R-squared:                  0.023
Method:                 Least Squares   F-statistic:                     2702.
Date:                Sun, 01 Oct 2023   Prob (F-statistic):               0.00
Time:                        12:07:42   Log-Likelihood:            -8.9160e+05
No. Observations:              114610   AIC:                         1.783e+06
Df Residuals:                  114608   BIC:                         1.783e+06
Df Model:                           1
Covariance Type:                  HC0
```

```
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       3411.6170      1.868   1826.620      0.000    3407.956    3415.278
tobacco     -240.4778      4.627    -51.978      0.000    -249.546    -231.410
==============================================================================
Omnibus:                    15963.198   Durbin-Watson:                   1.870
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            43808.162
Skew:                          -0.766   Prob(JB):                         0.00
Kurtosis:                       5.613   Cond. No.                         2.81
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
```

[16]:
```
#variable classification


#outcome
y = ['dbrwt']
#treatment
D = ['tobacco']
#cor with y and D

#we believe that drinking alcohol, a mother's race, adequacy of care, other␣
 ↪health problems, education, age,
#marriage, and birth order all affect whether a woman smokes during pregnancy␣
 ↪and the weight of the child
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
#we believe that the sex of the child, length of gestation, and the plurality of␣
 ↪the pregnancy
#only affect the birth weight of the child; they do not affect whether the␣
 ↪mother smokes
x3 = ['dgestat', 'csex', 'plur_1']
```

[ ]:
```

```