Evaluating the impact of HISP: Difference-in-Differecnes

In the design of HISP, there are two rounds of data on two groups of households: one group that enrolled in the program, and the another that did not. As in the case of the enrolled and non-enrolled groups, **we realized that we cannot simply compare the average health expentidures of the two groups beacuase of selection bias.** As we have data for two periods for wach household in the sample, we can use those data to solve some of these challenges by comparing the change in health expenditures for the two groups.

Set up

Lauching stata from the jupyter notebook

In [1]:
```
%%capture
import stata_setup
import os
os.chdir('C:\Program Files\Stata17/utilities')
from pystata import config
config.init('mp');
```

Initial set up of log file and load data

In [2]:
```
%%capture
%%stata

clear
set more off, perm

# redirect to workplace
cd "C:\Users\USER\Desktop\Charlene\2022 Charlene at York\Evaluation of Health Policy\practical exercise"

# Load data
use "evaluation.dta", clear
```

Create(rename) variable for treatment effect evaluation

In [3]:
```
%%capture
%%stata

# create generic variable (y)
clonevar y=health_expenditures
label var y "out of pocket health expenditure pc/pa"
clonevar d=enrolled
label var d "Treatment"

# Create global list of regressors
global xs "age_hh age_sp educ_hh educ_sp female_hh indigenous hhsize dirtfloor bathroom land hospital_distar
```

Difference-in-Differecnes

Assumming that the change in the health expenditures of the nonenrolled group reflects what would have happended to the expenditures of the enrolled group in the absence of the program. Here we calculate before-after comparison of **means for nonenrolled households**:

In [4]:
```
%%stata
ttest health_expenditures if enrolled ==0, by(round)
```

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
       0 |   6,949    18.37171    .0678053    5.652299    18.23879    18.50463
       1 |   6,949    20.70746    .1340806    11.17705    20.44462     20.9703
---------+--------------------------------------------------------------------
Combined |  13,898    19.53959    .0757729    8.932852    19.39106    19.68811
---------+--------------------------------------------------------------------
    diff |           -2.335746    .1502504               -2.630257   -2.041235
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t = -15.5457
H0: diff = 0                                     Degrees of freedom =    13896

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

From the table above we get that nonenrolled households have a baseline (before) mean of 18.37 and a follow-up (after) mean of 20.70. Then we calculate before-after comparison of **means for enrolled households**:

In [5]:
```
%%stata
ttest health_expenditures if enrolled ==1, by(round)
```

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
       0 |   2,964    14.48969    .0800166    4.356317     14.3328    14.64659
       1 |   2,965    7.840179    .1468178    7.994495    7.552304    8.128054
---------+--------------------------------------------------------------------
Combined |   5,929    11.16438    .0940975    7.245509    10.97991    11.34884
---------+--------------------------------------------------------------------
    diff |            6.649515    .1672221                6.321699    6.977331
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  39.7646
H0: diff = 0                                     Degrees of freedom =     5927

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

From the table above we het that enrolled households have a baseline (before) mean of 14.49 and a follow-up(after) mean of 7.84. Next we estimate the effect using a **simple linear regression** to compute the simple DiD estimate:

In [6]:
```
%%stata
diff y, t(d) p(round)
```

```
DIFFERENCE-IN-DIFFERENCES ESTIMATION RESULTS
--------------------------------------------
Number of observations in the DIFF-IN-DIFF: 19827
            Before         After
   Control: 6949           6949         13898
   Treated: 2964           2965         5929
            9913           9914

--------------------------------------------------------
 Outcome var.    | y        | S. Err. |    |t|   |  P>|t|
----------------+---------+---------+---------+---------
Before          |         |         |         |
   Control      | 18.372  |         |         |
   Treated      | 14.490  |         |         |
   Diff (T-C)   | -3.882  | 0.180   | -21.56  | 0.000***
After           |         |         |         |
   Control      | 20.707  |         |         |
   Treated      | 7.840   |         |         |
   Diff (T-C)   | -12.867 | 0.180   | 71.46   | 0.000***
                |         |         |         |
Diff-in-Diff    | -8.985  | 0.255   | 35.28   | 0.000***
--------------------------------------------------------
R-square:    0.22
* Means and Standard Errors are estimated by linear regression
**Inference: *** p<0.01; ** p<0.05; * p<0.1
```

Using a simple linear regression to compute the simple DiD sestimatem, I find that the program reduced household expenditures by US$8.985. I then refine my analysis by adding additional control variables. In other words, I use a **mutivariate linear regression** that takes into accoutn a host of other factors:

In [7]:
```
%%stata
diff y, t(d) p(round) cov($xs)
```

```
DIFFERENCE-IN-DIFFERENCES WITH COVARIATES
-----------------------------------------

DIFFERENCE-IN-DIFFERENCES ESTIMATION RESULTS
--------------------------------------------
Number of observations in the DIFF-IN-DIFF: 19827
            Before         After
   Control: 6949           6949        13898
   Treated: 2964           2965        5929
            9913           9914
--------------------------------------------------------
 Outcome var.   | y        | S. Err. |   |t|   |  P>|t|
----------------+---------+---------+---------+---------
Before          |          |         |         |
   Control      | 26.154   |         |         |
   Treated      | 25.325   |         |         |
   Diff (T-C)   | -0.829   | 0.147   | -5.65   | 0.000***
After           |          |         |         |
   Control      | 28.418   |         |         |
   Treated      | 18.604   |         |         |
   Diff (T-C)   | -9.814   | 0.147   | 66.91   | 0.000***
                |          |         |         |
Diff-in-Diff    | -8.985   | 0.202   | 44.48   | 0.000***
--------------------------------------------------------
R-square:    0.51
* Means and Standard Errors are estimated by linear regression
**Inference: *** p<0.01; ** p<0.05; * p<0.1
```

From the multivariate linear regression result, I find the same reduction in household health expenditure.

Questions

What are the basic assumptions required to accept this result from difference-in-differences?

To accept this result, **we assume that there are no differential time varying factors between the two groups other than the program.** We assume that the treatment and comparison groups would have equal trends or changes in outcomes in the absence of treatment. While this assumption can't be tested in the postintervention period, we can compare trends before the intervention starts.

Based on the result from difference-in-differences, should HISP be scaled up nationally?

No, based on this result, the HISP should not be scaled up nationally because it has decreased health expenditures by less than the $10 threshold level. Taking the estimated impact under random assignment as the "true" impact of the program suggests that the difference in difference estimate may be biased. In fact, in this case, using the nonenrolled households as a comparison group does not accurately represent the counterfactual trend in health expenditures.

Additional Commend

1. Estimating a fixed effects regression with `xtest`

In [8]:
```
%%stata
qui xtset household_identifier round
qui gen treated = d*round
xtreg y treated round, fe
```

```
. qui xtset household_identifier round

. qui gen treated = d*round

. xtreg y treated round, fe

Fixed-effects (within) regression               Number of obs     =     19,827
Group variable: household_~r                    Number of groups  =      9,914

R-squared:                                      Obs per group:
     Within  = 0.1698                                         min =          1
     Between = 0.2401                                         avg =        2.0
     Overall = 0.2013                                         max =          2

                                                F(2,9911)         =    1013.79
corr(u_i, Xb) = 0.1779                          Prob > F          =     0.0000

------------------------------------------------------------------------------
           y | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     treated |  -8.985667   .2002792   -44.87   0.000    -9.378255   -8.593079
       round |   2.335746   .1095147    21.33   0.000     2.121075    2.550417
       _cons |   17.21091   .0648368   265.45   0.000     17.08382      17.338
-------------+----------------------------------------------------------------
     sigma_u |   7.049476
     sigma_e |  6.4553311
         rho |  .54391007   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(9913, 9911) = 2.31                  Prob > F = 0.0000

.
```

2. Estimating DiD with `xtdidregress`

In [10]:
```
%%stata
xtdidregress (y) (treated), group(d) time(round)
```

```
Number of groups and treatment time

Time variable: round
Control:       treated = 0
Treatment:     treated = 1
-----------------------------------
             |   Control  Treatment
-------------+---------------------
Group        |
           d |         1          1
-------------+---------------------
Time         |
     Minimum |         0          1
     Maximum |         0          1
-----------------------------------

Difference-in-differences regression                    Number of obs = 19,827
Data type: Longitudinal

                                      (Std. err. adjusted for 2 clusters in d)
------------------------------------------------------------------------------
             |               Robust
           y | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
ATET         |
     treated |
   (1 vs 0)  |  -8.985667   1.29e-15 -7.0e+15   0.000    -8.985667   -8.985667
------------------------------------------------------------------------------
Note: ATET estimate adjusted for panel effects and time effects.
```