WANG Liangjiayi - ECO372 Assignment 2

Student Nb: [1004405789]

a.

```
. reg age treat,robust

Linear regression                               Number of obs     =        722
                                                F(1, 720)         =       0.13
                                                Prob > F          =     0.7216
                                                R-squared         =     0.0002
                                                Root MSE          =       6.63

------------------------------------------------------------------------------
             |               Robust
         age |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   .1792038   .5026692     0.36   0.722    -.8076687    1.166076
       _cons |   24.44706   .3197422    76.46   0.000     23.81932     25.0748
------------------------------------------------------------------------------

. reg education treat,robust

Linear regression                               Number of obs     =        722
                                                F(1, 720)         =       2.14
                                                Prob > F          =     0.1441
                                                R-squared         =     0.0031
                                                Root MSE          =     1.7033

------------------------------------------------------------------------------
             |               Robust
   education |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   .1922361   .1314755     1.46   0.144     -.065885    .4503572
       _cons |   10.18824   .0785342   129.73   0.000     10.03405    10.34242

. reg black treat, robust

Linear regression                               Number of obs     =        722
                                                F(1, 720)         =       0.00
                                                Prob > F          =     0.9645
                                                R-squared         =     0.0000
                                                Root MSE          =     .40014

------------------------------------------------------------------------------
             |               Robust
       black |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
             |
       treat |   .0013468   .0302489     0.04   0.964    -.0580399    .0607335
       _cons |         .8   .0194298    41.17   0.000     .7618542    .8381458
------------------------------------------------------------------------------

. reg married treat,robust

Linear regression                               Number of obs     =        722
                                                F(1, 720)         =       0.15
                                                Prob > F          =     0.7028
                                                R-squared         =     0.0002
                                                Root MSE          =     .36897

------------------------------------------------------------------------------
             |               Robust
     married |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   .0107031   .0280365     0.38   0.703    -.0443399    .0657461
       _cons |   .1576471    .017701     8.91   0.000     .1228953    .1923988
------------------------------------------------------------------------------

. reg hispanic treat, robust

Linear regression                               Number of obs     =        722
                                                F(1, 720)         =       0.66
                                                Prob > F          =     0.4154
                                                R-squared         =     0.0009
                                                Root MSE          =     .30718

------------------------------------------------------------------------------
             |               Robust
    hispanic |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |  -.0186651    .022906    -0.81   0.415    -.0636357    .0263055
       _cons |   .1129412   .0153748     7.35   0.000     .0827563     .143126
------------------------------------------------------------------------------

. reg nodegree treat, robust

Linear regression                               Number of obs     =        722
                                                F(1, 720)         =       6.82
                                                Prob > F          =     0.0092
                                                R-squared         =     0.0098
                                                Root MSE          =     .41293

------------------------------------------------------------------------------
             |               Robust
    nodegree |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |  -.0834779   .0319616    -2.61   0.009     -.146227   -.0207288
       _cons |   .8141176    .018896    43.08   0.000     .7770197    .8512155
```

H_0: *there is no difference between two groups*

H_a: *there is statistically significant difference between two groups*

We set the significance level at 5%. From the table, we conclude that the P-value for the observables characteristics: age, year of school, married, Black, and Hispanic are greater than 0.05. Thus, those variables are not statistically significant at 5%. There is no difference between individuals who were assigned into the training and those who were not assigned into the training in those pre-experiment observables characteristics (age, year of school, married, Black, and Hispanic).

In comparison, the P-value for the observable characteristics: High school dropouts is smaller than 0.05, which represents High school dropouts is statistically significant at 5%. The individuals assigned into the training and those who were not were different in the pre-experiment observables characteristics: High school dropouts.

We expect the means of the characteristics in the experimental groups are the same and we can use this to test the existence of selection bias and whether the randomization is successfully assigned. The selection bias will cause biased estimate. Overall, in this case, most of the pre-experiment observables characteristics are the same in both groups and the randomization is valid.

b.

```
. reg re78 treat,robust

Linear regression                               Number of obs     =        722
                                                F(1, 720)         =       3.30
                                                Prob > F          =     0.0698
                                                R-squared         =     0.0049
                                                Root MSE          =       6242
```

re78	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
treat	886.3037	488.1385	1.82	0.070	-72.04121	1844.649
_cons	5090.048	277.4261	18.35	0.000	4545.388	5634.709

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Y is the earnings of experimental participants in 1978

X is a dummy variable and represent the assigned groups for each participant (1: treatment group; 0: control group)

u is error term.

After running that regression with robust, the estimation of coefficient is the same in table and our regression. The NSW treatment groups' earning are $886 more than control groups' earnings on average. However, the Standard Error are not the same. In the table 5, the standard error equals 476 and in regression the standard error equals 488. The deviation between a sample mean and the actual mean of a population are larger in our regression. Also, from the regression table, we find that the difference of earnings between two groups $(\widehat{\beta_1})$ are statistically significant at 10% significance level (p-value<0.1).

c.

```
. gen age_sqr=age^2

. reg re78 age age_sqr education nodegree black hispanic treat, robust

Linear regression                               Number of obs     =        722
                                                F(7, 714)         =       2.71
                                                Prob > F          =     0.0089
                                                R-squared         =     0.0238
                                                Root MSE          =     6208.4
```

re78	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-3.805475	193.7217	-0.02	0.984	-384.1378	376.5268
age_sqr	.5296508	3.197424	0.17	0.868	-5.747827	6.807128
education	219.7946	165.5418	1.33	0.185	-105.2124	544.8015
nodegree	-494.2816	756.7119	-0.65	0.514	-1979.928	991.3648
black	-1762.833	774.898	-2.27	0.023	-3284.184	-241.4815
hispanic	-117.148	983.6556	-0.12	0.905	-2048.351	1814.055
treat	798.3512	488.168	1.64	0.102	-160.0653	1756.768
_cons	4430.163	3594.725	1.23	0.218	-2627.333	11487.66

$$Y_i = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 Education_i + \beta_4 Nodegree_i + \beta_5 Black_i + \beta_6 Hispanic + \beta_7 Treat + u_i$$

Note:

Age: Age for each participant

Education: Year of schools for each participant

Nodegree: High school dropout status for each participant (1: yes; 0: no)

Black: Whether participant identify as Black (1: yes; 0: no)

Hispanic: Whether participant identify as Hispanic (1: yes; 0: no)

Treat: Assigned groups for each participant (1: participants in treatment group; 0: participants in control group)

Y: the earnings of experimental participants in 1978

u: error term.

After running that regression with robust, the estimation of coefficient is the same in table and our regression. After control the exogenous variable (age, age_sqared, year of schooling, high school dropout status and race) that used in adjusted equations, the NSW treatment groups' earning are $798 more than control groups' earnings on average. However, the Standard Error are not the same. In the paper table 5, the standard error equals 472 and in our regression the standard error equals 488. The deviation between a sample mean and the actual mean of a population are larger in our regression. Also, from the regression table, we find that the difference of earnings between two groups ($\widehat{\beta_7}$) are not statistically significant at 5% significance level (p-value>0.05).

d.

```
. reg re78 age age_sqr education nodegree black hispanic treat

      Source |       SS           df       MS      Number of obs   =       722
-------------+----------------------------------   F(7, 714)       =      2.48
       Model |   670296792         7  95756684.6   Prob > F        =    0.0159
    Residual |  2.7520e+10       714  38543836.8   R-squared       =    0.0238
-------------+----------------------------------   Adj R-squared   =    0.0142
       Total |  2.8191e+10       721  39099301.3   Root MSE        =    6208.4

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -3.805475   211.1663    -0.02   0.986    -418.3866    410.7756
     age_sqr |   .5296508   3.556177     0.15   0.882    -6.452164    7.511466
   education |   219.7946   182.9296     1.20   0.230    -139.3496    578.9387
    nodegree |  -494.2816   749.2561    -0.66   0.510     -1965.29     976.727
       black |  -1762.833     803.88    -2.19   0.029    -3341.084   -184.5814
    hispanic |   -117.148   1054.228    -0.11   0.912    -2186.906     1952.61
       treat |   798.3512   472.1283     1.69   0.091    -128.5747    1725.277
       _cons |   4430.163   3653.224     1.21   0.226    -2742.183     11602.51
------------------------------------------------------------------------------
```

The original paper uses classic robust standard errors. After running that regression without robust, the estimation of coefficient ($\widehat{\beta_{treat}} = 798$) and standard error (472) is the same in table and our regression. Also, from the regression table, we find that the difference of earnings between two groups ($\widehat{\beta_{treat}}$) are statistically significant at 10% significance level (p-value<0.1).

e.

The pre-experiment characteristics should have no difference in both treatment group and control group.

If CIA holds, the earning difference can be identified as causal effect of the training program. In our case, the CIA holds since the participants are randomly assigned to treatment and control group which eliminate the selection bias. And from Question a, we noticed that most of characteristics have same mean in both groups

f.

```
. ttest re78, by(treat) unequal

Two-sample t test with unequal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
0	425	5090.048	277.368	5718.089	4544.861	5635.236
1	297	5976.352	401.7594	6923.796	5185.685	6767.019
combined	722	5454.636	232.7105	6252.943	4997.765	5911.507
diff		-886.3037	488.2045		-1845.251	72.64306

```
    diff = mean(0) - mean(1)                                      t =  -1.8154
Ho: diff = 0                     Satterthwaite's degrees of freedom =  557.062

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0350         Pr(|T| > |t|) = 0.0700          Pr(T > t) = 0.9650
```

$H_0: Earning\ 1978_{control} - Earning\ 1978_{treatment} = 0$

$H_a: Earning\ 1978_{control} - Earning\ 1978_{treatment} \neq 0$

$p - value = 0.0700 > 0.05$

Since p-value > 0.05, we fail to reject null hypothesis H0 that Earnings in 1978 for participants in treatment group is the same for participant in control group. At 5% significance level, we have no evidence that there is a statistically significant difference between 1978 earnings for participants in treatment group ($Earning\ 1978_{treatment}$) and 1978 earnings for participants in control group ($Earning\ 1978_{control}$). Thus, we conclude that the earnings for participant who join the training (treatment group) is as same as the earnings for participant who are not join the training (control group).

The effectiveness of training program is not remarkable in our case and it may due to several reasons. In the question a, we noticed that the means of high school dropout is not the same in both groups which will influence the accuracy of regression result and lead to small bias. Also, most of our independent variables in regression are not significant at 5% level. If we eliminate those potential problems, the effectiveness of training program may be significant.

g.

PSID-3 is all male household heads continuously from 197-1978, who were less than 55-years-old and did not classify themselves as retried in 1975 and were not working when surveyed in either spring of 1975 or 1976.

h.

```
. reg age treat,robust

Linear regression                               Number of obs     =        425
                                                F(1, 423)         =     128.56
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3261
                                                Root MSE          =     9.0109

------------------------------------------------------------------------------
             |               Robust
         age |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |  -13.63155   1.202261   -11.34   0.000     -15.9947    -11.2684
       _cons |   38.25781   1.137848    33.62   0.000     36.02127    40.49435
------------------------------------------------------------------------------
```

```
. reg education treat,robust

Linear regression                               Number of obs     =        425
                                                F(1, 423)         =       0.06
                                                Prob > F          =     0.8004
                                                R-squared         =     0.0002
                                                Root MSE          =      2.311

------------------------------------------------------------------------------
             |               Robust
   education |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   .0757839   .2995041     0.25   0.800    -.5129178    .6644855
       _cons |   10.30469   .2802907    36.76   0.000     9.753751    10.85562
------------------------------------------------------------------------------
```

```
. reg black treat, robust

Linear regression                               Number of obs     =        425
                                                F(1, 423)         =      48.82
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1207
                                                Root MSE          =     .43215

------------------------------------------------------------------------------
             |               Robust
       black |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   .3482218   .0498361     6.99   0.000     .2502645    .4461791
       _cons |    .453125   .0441034    10.27   0.000     .3664358    .5398142
------------------------------------------------------------------------------
```

```
. reg nodegree treat, robust

Linear regression                               Number of obs     =        425
                                                F(1, 423)         =      18.90
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0468
                                                Root MSE          =      .4624

------------------------------------------------------------------------------
             |               Robust
    nodegree |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |   .2228272   .0512607     4.35   0.000     .1220698    .3235846
       _cons |   .5078125   .0442931    11.46   0.000     .4207505    .5948745
------------------------------------------------------------------------------
```

```
. reg married treat,robust

Linear regression                               Number of obs     =        425
                                                F(1, 423)         =     129.97
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2655
                                                Root MSE          =       .403

------------------------------------------------------------------------------
             |               Robust
     married |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |  -.5269623    .046223   -11.40   0.000    -.6178177    -.436107
       _cons |   .6953125    .040779    17.05   0.000     .6151578    .7754672
------------------------------------------------------------------------------
```

```
. reg hispanic treat, robust

Linear regression                               Number of obs     =        425
                                                F(1, 423)         =       0.48
                                                Prob > F          =     0.4902
                                                R-squared         =     0.0012
                                                Root MSE          =     .30209

------------------------------------------------------------------------------
             |               Robust
    hispanic |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       treat |  -.0229114   .0331801    -0.69   0.490    -.0881299    .0423071
       _cons |   .1171875   .0284967     4.11   0.000     .0611748    .1732002
------------------------------------------------------------------------------
```

H_0: *there is no difference between two groups*

H_a: *there is statistically significant difference between two groups*

We set the significance level at 5%. From the table, we conclude that the P-value for the observables characteristics: age, married, black, and nodegree are smaller than 0.05. Thus, those variables are statistically significant at 5% and we reject H0. There is a statistically significant difference between individuals who were assigned into the training and those who were not in those pre-experiment observables characteristics (age, married, black and nodegree). In this case, the participants are not randomly assigned. Thus, the selection bias occurs and leads to the estimation bias in the regression model which affect the labor market outcome.

In comparison, the P-value for the observable characteristics: Education and Hispanic is greater than 0.05, which represents year of education is not statistically significant at 5%. The individuals assigned into the training and those who were not has no different in the pre-experiment observables characteristics: Education, Hispanic.

i.

```
. gen age_sqr=age^2

. reg re78 age age_sqr education nodegree black hispanic treat
```

Source	SS	df	MS
Model	1.7849e+09	7	254984265
Residual	2.0102e+10	417	48205230.7
Total	2.1886e+10	424	51619035.5

Number of obs	=	425
F(7, 417)	=	5.29
Prob > F	=	0.0000
R-squared	=	0.0816
Adj R-squared	=	0.0661
Root MSE	=	6943

re78	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	769.8936	232.7513	3.31	0.001	312.3815	1227.406
age_sqr	-12.63406	3.395785	-3.72	0.000	-19.30905	-5.95907
education	192.5652	226.1696	0.85	0.395	-252.0094	637.1397
nodegree	-581.9737	1044.792	-0.56	0.578	-2635.689	1471.742
black	-1516.321	971.6635	-1.56	0.119	-3426.29	393.6479
hispanic	293.7419	1369.466	0.21	0.830	-2398.175	2985.659
treat	-509.2156	967.4217	-0.53	0.599	-2410.847	1392.415
_cons	-4635.431	4526.469	-1.02	0.306	-13532.97	4262.11

From question d, we find that the original paper uses the classic standard error. Then we run a new regression based on new dataset.

After running that regression without robust, the estimation of coefficient and standard error are the same in table and our regression. After control the exogenous variable (age, age_sqared, year of schooling, high school dropout status and race) that used in adjusted equations, the NSW treatment groups' earning are $509 less than PSID-3 groups' earnings on average.

Compared the result in question d, the estimated effect for PSID-3 is -$509 and the estimated effect for control group is $798. Two results are completely different. The non-experimental estimate method does not randomly assign the participants. The observable/unobservable characteristics of trainees and the comparison group members differ. Also, we cannot ensure that the unobservable in the earnings and participation equations are uncorrelated. Thus, the $-509 as a non-experimental estimate cannot replicate the experimental results ($798).

j.

From the result in question h, we find that the treatment group is 13-years old younger than the control group. And treat has greater effect on age compared to other characteristics. And we expect younger participants have higher wage.

```
. reg age treat,robust

Linear regression                          Number of obs     =        425
                                           F(1, 423)         =     128.56
                                           Prob > F          =     0.0000
                                           R-squared         =     0.3261
                                           Root MSE          =     9.0109
```

age	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
treat	-13.63155	1.202261	-11.34	0.000	-15.9947	-11.2684
_cons	38.25781	1.137848	33.62	0.000	36.02127	40.49435

k.

```
. gen diff=re78-re75

. reg diff age age_sqr treat
```

Source	SS	df	MS
Model	700550249	3	233516750
Residual	2.8840e+10	421	68503430.5
Total	2.9540e+10	424	69670977.6

```
Number of obs     =        425
F(3, 421)         =       3.41
Prob > F          =     0.0176
R-squared         =     0.0237
Adj R-squared     =     0.0168
Root MSE          =     8276.7
```

diff	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	594.3688	268.2832	2.22	0.027	67.02736	1121.71
age_sqr	-9.890007	3.854744	-2.57	0.011	-17.46695	-2.313065
treat	-1324.562	1078.325	-1.23	0.220	-3444.134	795.0104
_cons	-3963.773	4268.744	-0.93	0.354	-12354.48	4426.933

$$Y_i = \beta_0 + \beta_1 Age_i + \beta_2 Age_i^2 + \beta_3 Treat + u_i$$

Note:

Age: Age for each participant

Treat: Assigned groups for each participant (1: participants in treatment group; 0: participants in control group)

Y: the earning growth 1975-78 of experimental participants

U: the error term.

From this regression, the estimate ($-1325) is identical to column 7.

l.

In column 7, the only observable characteristics we control for is age. In control group, the estimate of earning growth is $856 and in PSID-3 group the estimate of earning growth is $-1325. In question h and j, we also noticed that there are differences in the characteristics like race, married and high school dropout between treatment and control group, that could influence the outcomes. Without control for those variables and eliminate the correlation between the unobservables, the non-experimental estimate is biased.