NI, Sailing - ECO372 Assignment 3

Student Number: 1004936019

(a) This part consists of four questions:

i. The unit of observation is student/applicant

ii. There are 283 observations in this subsample

iii. The variable "*vouch0*" records whether a student won a voucher

iv. Variable "*math*" records math scores, variable "*reading*" records reading scores, variable "*writing"* records writing scores, and variable "*totalpts*" records total scores.

(b)

```
. summarize math reading writing
```

Variable	Obs	Mean	Std. Dev.	Min	Max
math	282	.0028718	1.002821	-2.024617	3.239387
reading	283	.0006201	1.003512	-3.972432	2.097389
writing	283	.0049475	1.001712	-3.295184	2.20036

The mean figures of all three variables *math, reading,* and *writing*, are close to 0, and sample standard deviations are about 1. This fact suggests that all three test scores are standardized to have mean of 0 and standard deviation of 1.

(c)

```
. reg totalpts vouch0 i.t_site, robust

Linear regression                               Number of obs     =        282
                                                F(3, 278)         =       4.08
                                                Prob > F          =     0.0074
                                                R-squared         =     0.0415
                                                Root MSE          =       .987
```

totalpts	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
vouch0	.2167926	.1208689	1.79	0.074	-.0211419	.4547272
t_site						
2	-.555319	.173769	-3.20	0.002	-.8973892	-.2132488
3	-.3569731	.1650089	-2.16	0.031	-.6817987	-.0321476
_cons	.2352034	.1476332	1.59	0.112	-.0554176	.5258244

```
. estimates store regression1
```

```
. reg math vouch0 i.t_site, robust

Linear regression                               Number of obs     =        282
                                                F(3, 278)         =       4.84
                                                Prob > F          =     0.0027
                                                R-squared         =     0.0480
                                                Root MSE          =     .98374

------------------------------------------------------------------------------
             |               Robust
        math |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      vouch0 |   .1776212   .1184469     1.50   0.135    -.0555454    .4107879
             |
      t_site |
          2  |  -.6321908   .1757982    -3.60   0.000    -.9782555   -.2861262
          3  |  -.3956517   .1682034    -2.35   0.019    -.7267657   -.0645376
             |
       _cons |   .2998769   .1530854     1.96   0.051    -.0014769    .6012306
------------------------------------------------------------------------------

. estimates store regression2
. reg reading vouch0 i.t_site, robust

Linear regression                               Number of obs     =        283
                                                F(3, 279)         =       2.82
                                                Prob > F          =     0.0394
                                                R-squared         =     0.0265
                                                Root MSE          =     .99542

------------------------------------------------------------------------------
             |               Robust
     reading |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      vouch0 |   .2035837   .1227233     1.66   0.098    -.0379975    .4451648
             |
      t_site |
          2  |  -.4035848   .1623344    -2.49   0.013    -.7231405   -.0840291
          3  |  -.3255434   .1480893    -2.20   0.029    -.6170576   -.0340293
             |
       _cons |   .1771546   .1305668     1.36   0.176    -.0798667    .4341758
------------------------------------------------------------------------------

. estimates store regression3

. reg writing vouch0 i.t_site, robust

Linear regression                               Number of obs     =        283
                                                F(3, 279)         =       0.82
                                                Prob > F          =     0.4835
                                                R-squared         =     0.0092
                                                Root MSE          =     1.0025

------------------------------------------------------------------------------
             |               Robust
     writing |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      vouch0 |   .1258825   .1222843     1.03   0.304    -.1148345    .3665994
             |
      t_site |
          2  |  -.2034199   .1819608    -1.12   0.265    -.5616102    .1547705
          3  |  -.0608489   .1665305    -0.37   0.715    -.3886647    .2669669
             |
       _cons |   .0261343   .1525117     0.17   0.864    -.2740855    .3263541
------------------------------------------------------------------------------

. estimates store regression4
```

Footnote 13 from the paper says "the results in columns (1) and (3) are from models that include site dummies only." The variable *t_site* is not a dummy variable but can be any values from 1, 2 or 3. We ran the above four regressions separately on *totalpts, math, reading,* and *writing*, and store them in estimates to prepare for the replication of Table 5 Panel A. Then use commands in *estout* library to obtain this table below:

	(1) totalpts	(2) math	(3) reading	(4) writing
vouch0	0.217	0.178	0.204	0.126
	(0.121)	(0.118)	(0.123)	(0.122)
Observations	282	282	283	283
R-squared	0.0415	0.0480	0.0265	0.00918
F-stat	4.081	4.844	2.819	0.820

Standard errors in parentheses
$^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

Comparing this obtained results and the table in the paper, we find similar results of point estimates across all variables *totalpts, math, reading,* and *writing*. Even the standard errors are a bit different, overall results are not largely different. As quoted in the paper notes, "standard errors in columns (1) and (2) are corrected for within-school-of-application clustering," which might be the cause of this little difference.

(d) From notes of Table 5: "the estimates in columns (2) and (4) are from models that include controls for applicant's age, gender, parents' schooling, strata of residence, and type of survey and instrument."

```
. reg totalpts vouch0 i.t_site age sex dad_sch mom_sch strata svy hsvisit, robust

Linear regression                               Number of obs     =        189
                                                F(10, 178)        =       3.69
                                                Prob > F          =     0.0002
                                                R-squared         =     0.1793
                                                Root MSE          =     .91241
```

totalpts	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
vouch0	.1007768	.1421145	0.71	0.479	-.1796693	.381223
t_site						
2	-.7119973	.2420668	-2.94	0.004	-1.189687	-.2343073
3	-.5114409	.2419333	-2.11	0.036	-.9888674	-.0340143
age	-.2145499	.0537146	-3.99	0.000	-.3205493	-.1085505
sex	.0812321	.1354219	0.60	0.549	-.1860068	.348471
dad_sch	.020914	.0270925	0.77	0.441	-.0325499	.0743778
mom_sch	.0379183	.0282083	1.34	0.181	-.0177476	.0935841
strata	-.0047916	.126997	-0.04	0.970	-.255405	.2458218
svy	.0012225	.1557238	0.01	0.994	-.3060798	.3085248
hsvisit	-.3169003	.2664674	-1.19	0.236	-.842742	.2089414
_cons	3.362796	.8980749	3.74	0.000	1.590553	5.13504

In the above regression, to control the parents' schooling, I used both parents' years of school, *dad_sch* and *mon_sch*. Form and instrument of survey is controlled by *svy* and *hsvisit*, which represent if the applicant uses the new survey and if the survey is conducted in person respectively. In the paper, the estimate was 0.205 with standard error of 0.108 while in the analysis here, we get 0.101 with standard error of 0.108. This concludes a large difference between obtained result and paper data. One possible reason that can explain this huge difference is the observations used in analysis. In dataset we had 282 observations while here only 189 valid observations has been applied due to missing data. All data with missing information is omitted in my analysis, but in the investigation conducted by the researcher they may assign default values to the missing values. Also, as mentioned before, standard errors in columns (1) and (2) are corrected for within-school-of-application clustering, which can be another source that introduces differences.

	(1) totalpts with covariates
vouch0	0.101
	(0.142)
Observations	189
R-squared	0.179
F-stat	3.690

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(e) We want to test the statistical significance for the point estimate of 0.217, with the robust standard error of 0.116 standard deviation. Use the hypothesis test, the t-stat is calculated to be 0.217/0.116=1.87, which is smaller than 1.96. Therefore, we do not have sufficient evidence to reject the null hypothesis at the 5% significance level. This result means it is not sufficient to conclude statistical significance and infer causal relationship between the total points a student scores and if they won a voucher. In the following interpretation, we just conclude association between the variables:

Among all students that participated in the voucher lottery, with the test sites being controlled, we observe that on average students who won vouchers scores 0.217 standard deviation higher than those who did not receive the voucher, as investigated in 1999, after one year of household survey and three years after children applied for the program.

(f) This part consists of four questions:

i. The unit of observation is student/applicant, and there are 1, 135 observations in this subsample.

ii. The variable "*vouch0*" records whether a student won a voucher.

iii. The variable "*usesch*" records if a student actually used the school voucher.

iv. Variable "*scyfnsh*" records the highest grade of the student, and the variable "*inschl*" records if a student is still in school at the time of the survey.

(g) First, use *tabulate* command to list all possible values of these two categorical variables.

```
. tabulate month
```

Month of survey	Freq.	Percent	Cum.
1	25	2.20	2.20
2	42	3.70	5.90
3	469	41.32	47.22
4	253	22.29	69.52
5	73	6.43	75.95
6	75	6.61	82.56
7	77	6.78	89.34
8	98	8.63	97.97
9	11	0.97	98.94
10	1	0.09	99.03
11	5	0.44	99.47
12	6	0.53	100.00
Total	1,135	100.00	

```
. tabulate strata
```

Strata of residence	Freq.	Percent	Cum.
0	176	15.51	15.51
1	165	14.54	30.04
2	638	56.21	86.26
3	148	13.04	99.30
4	6	0.53	99.82
5	2	0.18	100.00
Total	1,135	100.00	

Then regress highest grade of student and if they are in school at the time of survey on the vouch dummy variable. All procedures are similar with question c and d, the regression results are omitted and the table is as shown below.

	(1) scyfnsh	(2) inschl
vouch0	0.126*	0.00764
	(0.0516)	(0.0203)
Observations	1135	1135
R-squared	0.108	0.165
F-stat	.	.

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Then we compare the obtained results with Table 3. Since the above analysis did not control for "*19 barrio*", our results must correspond to column 3. Also, because our outcomes are 1/ the highest grade achieved by the student, and 2/ whether they are still in school at the time of the survey, we only care about the rows say "highest grade completed" and "currently in school". In column (3) and row "highest grade completed", the estimate is 0.130 with a standard error of 0.051 (in obtained data, the estimate is 0.126 with standard error of 0.0516); in column (3) and row "currently in school", the estimate is 0.007 with a standard error of 0.020 (in obtained data, the estimate is 0.00764 with standard error of 0.0203). Overall, our obtained results comply to the data in table 3.

(h) No, regressing the required outcomes on the dummy variable *"UsedVoucher"* does not suggest any causal relationship of using the voucher. The assignment of voucher is random but the choice of using voucher is not: those who wins a voucher can decide if use it or not, but those who did not win the voucher had no choice on use of voucher. This suggests the data we collect about "*UsedVoucher"* would be subject to endogeneity bias, and there are confounding variables that affect both the use of voucher and the outcomes. For example, family support will influence both these two variables. Parents who do care about children may not let them use the voucher and might not encourage these students to study for higher grade or even go to school.

(i) The causal chain:

Win_Voucher (Z_i) $\Rightarrow$ Use_Voucher (D_i) $\Rightarrow$ Outcome: length of schooling (Y_i)

Since both the highest grade of student and if they are currently in school indicate the length of schooling, the analysis is interchangeable.

(a) Requirement 1: First stage/strong instrument: Z_i must have a causal effect on D_i

CIA has to be satisfied to establish a causal relationship. The investigation has suggested the voucher results are randomly generated, and it fulfills CIA. Also, the paper mentioned that those who did not get a voucher cannot use voucher by other means, meaning the outcome of Z_i influence Y_i which is the schooling length.

(b) Requirement 2: Exogeneity/independence: Z_i must be as good as random

The fact that voucher results are randomly generated means wining a voucher is not correlated with any unobserved variables that might affect the length of schooling Y_i.

(c) Requirement 3: Exclusion restriction: Z_i has a causal effect on Y_i only through D_i

In this research, voucher winners will be granted the access to private school. Getting the voucher does not affect any other factors including how hard you work, and how supportive your family, which might affect the length of schooling. In other words, voucher changes a student's highest achieved grade or if they are currently in school, only because receiving a voucher give then the right to choose go to private school or not.

(j) We regress the highest grade of student in OLS model and 2SLS models, and the results are shown below. Under the Bogota 1995 sample panel on the left, column (1) of the obtained result corresponds to the OLS column in the paper; while column (2) corresponds to the 2SLS column of the paper.

The obtained column (1) gives an estimate of 0.171 with standard error of 0.0544 (in the paper research on OLS model gives 0.167 with standard error of 0.053. The obtained column (2) gives an estimate of 0.191 with standard error of 0.0770 (in the paper research on 2SLS gives 0.196 with standard error of 0.078. We can see that there is no big difference between the two sets of results, and both estimates and the standard errors are quite close to the paper.

	(1) scyfnsh-ols	(2) scyfnsh-2sls
usesch	0.171**	0.191*
	(0.0544)	(0.0770)
N	1135	1135

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE 7—OLS AND 2SLS ESTIMATES OF THE EFFECT OF EVER USING A PRIVATE SCHOOL SCHOLARSHIP

	Coefficient on "Ever used a private-school scholarship"				
		Bogotá 1995		Combined sample	
Dependent variable	Loser means	OLS	2SLS	OLS	2SLS
Highest grade completed	7.5 (0.965)	0.167 (0.053)	0.196 (0.078)	0.141 (0.042)	0.134 (0.065)

(k) Since the 2SLS model is of higher accuracy as it controls for confounding and measurement error and allow us to infer causality conclusions on observational data. Running the results for 2SLS model as in question (j), we can test the statistical significance of our data. We calculate the t-stat by 0.191/0.077 to 2.48, which is larger than 1.96 at the confidence level of 5%. Hence, the null is rejected, and we have sufficient evidence to conclude statistical significance. Therefore, the causality conclusion is that use of school vouchers results in higher school grades. Specifically, using a school voucher causes highest grade of a student to increase 0.191 standard deviations on average.

(l) This question is quite similar with what we do in question (j) and (k), with the outcome variable replaced with in-school variable *inschl*. The results is shown below.

(m)	(1) inschl-2sls
usesch	0.0115
	(0.0304)
N	1135

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

The interpretation is similar with previous questions too. The results are similar with Table 7 figures in the paper. Here, we obtained an estimate of 0.0115 with standard error of 0.304, which would give us a relevant t-stat of 0.38, significantly smaller than 1.96 at the confidence level of 5%. Hence, we do not have sufficient evidence to reject the null, and then we cannot conclude any causality since the result might be subject to sampling error. Therefore, there is no statistically significant effect of school vouchers on the likelihood of being in school by the time of survey on the population level.