Causal Inference

Yingying Ji,15220162202134

2019/06/08

Introduction

In this report, I will do some simple empirical analyses about learning causal effect from randomized experiments, using Stata language to implement. The goal is to estimate the causal effect of small class on student's test score. The data are from CEPS database originally and have been already processed for simplicity.

Prepare the data

```
*HW7 for Microeconometrics
*By Yingying Ji, Stu.ID:15220162202134
*2019/06/08

clear //clear memory and remove data
set more off
cd "C:\Users\15068\Desktop\Microeconometrics\HW7"
//Change the working directory to a specific folder


use hw7data.dta //load the data
des //read the data
sum //read the data
```

-
```
. des //read the data

Contains data from hw7data.dta
  obs:       100,966
 vars:             7                          29 Mar 2018 11:41
 size:     2,827,048
```

variable name	storage type	display format	value label	variable label
stdid	float	%9.0g		student id
classid	float	%9.0g		class id
income	float	%9.0g		family annual income (1000$)
girl	float	%9.0g		indicator of girl
small	float	%9.0g	vetlb	indicator of small classes: true attendence
testscore	float	%9.0g		test score
classsize	float	%9.0g		

```
Sorted by: classid

. sum //read the data
```

Variable	Obs	Mean	Std. Dev.	Min	Max
stdid	100,966	50797.21	29335.15	1	101698
classid	100,966	2852.895	1645.026	1	5714
income	100,966	50.01178	10.02449	6.611536	98.59592
girl	100,966	.4818058	.4996713	0	1
small	100,966	.4040568	.490711	0	1
testscore	100,966	79.99507	5.16517	55.67772	98.48424
classsize	100,966	21.11501	5.822513	10	32

-

Check random assignment

- In this step, I check that if students are randomly assigned to different types of classes.

```
ttest income, by(small) unequal
//Use t-test to compare the mean of income between two groups(small and regular class)
ttest girl, by(small) unequal
//Use t-test to compare the mean of the share of girls between two groups.
regress girl small
// or regress girl on the indicator of small classes
regress income small
//or regress income on the indicator of small classes
```

-

```
. ttest income, by(small) unequal

Two-sample t test with unequal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
0	60,170	50.00338	.04075	9.995795	49.92351	50.08325
1	40,796	50.02417	.0498404	10.06677	49.92648	50.12186
combined	100,966	50.01178	.0315482	10.02449	49.94995	50.07362
diff		-.0207878	.0643788		-.1469696	.105394

```
    diff = mean(0) - mean(1)                                      t =  -0.3229
Ho: diff = 0                     Satterthwaite's degrees of freedom =  87159.2

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.3734         Pr(|T| > |t|) = 0.7468          Pr(T > t) = 0.6266
```

-

```
. ttest girl, by(small) unequal

Two-sample t test with unequal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
0	60,170	.4795247	.0020367	.4995847	.4755328	.4835165
1	40,796	.4851701	.0024744	.4997862	.4803202	.4900201
combined	100,966	.4818058	.0015725	.4996713	.4787236	.4848879
diff		-.0056454	.0032048		-.0119268	.000636

```
    diff = mean(0) - mean(1)                                      t =  -1.7616
Ho: diff = 0                     Satterthwaite's degrees of freedom =  87549.3

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0391         Pr(|T| > |t|) = 0.0781          Pr(T > t) = 0.9609
```

-
- The students are randomly assinged to different types of classes. As the output shows, for income, $P = 0.7468 > 0.1$, we should not reject H_0, the difference is not statistically significant at 10% significance level. For the share of girl, $P = 0.0781 > 0.05$, it's not statistically significant at 5% significance level. We should not reject H_0, the difference is also not statistically significant. I also do a formal test by regressing "girl" and "income" on the indicator of small classes to check again. And the results are the same.

Necessary assumptions

- To use a regression to estimate the causal effect of small class on student's test score, we need these assumptions:
  - $y_{0i}, y_{1i} \perp D_i$. That is, with random assignment, the test score should be independent with the treatment (small class).So Selection Bias goes to 0. A regression of Y on D (and a constant term) gives the treatment effect (ATE and ATET).
  - The error term ϵ_i has conditional mean zero given the independent variable X_i: $E(\epsilon_i|X_i) = 0$;
  - (X_i, Y_i) are i.i.d(identical independent distributed) for i=1,2,...,n.
  - Large outliers are unlikely: the independent variable X_i and the dependent variable Y_i have nonzero finite fourth moments.
- The model I use in this setting:

$$testscore_i = \alpha + \rho small_i + \mathbf{x}_i\prime\gamma + \epsilon_i$$

- Meaning of each variable:
  - $testscore_i$:the test score of student i;
  - ρ:the average treatment effect of class type on students scores;
  - $small_i$:a dummy variable that indicate small class when equals to 1, regular class when equal to 0;
  - $\mathbf{x}_i$:control variables, in this case, income and gender;
  - γ:the coefficient on control variables;
  - ϵ_i:the error term.
- I include variable "girl" and "income" in my model as control variables to make the results much preciser and therefore reduces the residual variance, which in turn lowers the standard error of the regression estimates.

Get causal effects

```
reg testscore small girl income, cluster(classid)
//regress testscore over small, girl, and income
//with standard errors allow to correlated within the same classid
```

-
- The type of standard error I use is clustered, which means regression with standard errors allow to correlated within the same class id.

```
. reg testscore small girl income, cluster(classid)

Linear regression                               Number of obs     =    100,966
                                                F(3, 5227)        =   16523.95
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2986
                                                Root MSE          =     4.3259

                            (Std. Err. adjusted for 5,228 clusters in classid)
```

testscore	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
small	4.980887	.0252245	197.46	0.000	4.931437	5.030338
girl	1.985354	.028376	69.97	0.000	1.929725	2.040983
income	.0984024	.0013883	70.88	0.000	.0956807	.101124
_cons	72.10468	.0733729	982.71	0.000	71.96084	72.24852

-

- The coefficient ρ captrues the difference in students' test scores when they're assigned to different class size types. ρ equals to 4.98 and it's statistically significant at 1% significant level, which means that the students who are assigned to small class tend to get 4.98 points higher than those assigned to regular class, holding others constant. In addition, γ equals to 1.99 and it's statistically significant at 1% significant level, which means that girls tend to get 1.99 points higher than boys.