Stata Code Sample

Xiling Zhu xiling@uchicago.edu

Aug 28, 2020

Background

In January 2012, the Cook County State's Attorney's Office established a program intended to reduce re-arrest among people on bail awaiting trial. The program ran through October 2013.

The objective of our analysis is to evaluate the effectiveness of the program. We start by cleaning data sets on demographics, cases, and academic performance. Next we proceed to provide descriptive statistics for the study population and test their baseline equivalence. The final step is to evaluate whether participating in the program reduces the likelihood of re-arrest before disposition.

0. Preamble [1]

```
. clear all
. set more off
. set varabbrev off

. global data /Users/celiazhu/Box/projects/ra_code_sample/data
. global output /Users/celiazhu/Box/projects/ra_code_sample/output
. global processed /Users/celiazhu/Box/projects/ra_code_sample/processed
```

1. Data Cleaning

1.1 Clean demographic Data

```
. import delimited "$data/demo.csv", clear
(4 vars, 20,436 obs)
```

Make sure person_id is uid

```
. duplicates drop
Duplicates in terms of all variables
(4,721 observations deleted)
. isid person_id
```

[1] Generated by `markstat`. Please see here for source code.

The demographic data were extracted from a system that inconsistently coded gender. Recode it so that males are consistently coded as "M" and females are consistently coded as "F".

```
. tab gender, m
```

gender	Freq.	Percent	Cum.
F	2,936	18.68	18.68
M	11,707	74.50	93.18
female	179	1.14	94.32
male	893	5.68	100.00
Total	15,715	100.00	

```
. replace gender = "F" if gender == "female"
(179 real changes made)
. replace gender = "M" if gender == "male"
(893 real changes made)
```

Check if gender is consistently coded

```
. assert gender == "M" | gender == "F"
```

Save cleaned demographic data

```
. save "$processed/demo_clean.dta", replace
file /Users/celiazhu/Box/projects/ra_code_sample/processed/demo_clean.dta saved
```

1.2 Clean case data

```
. import delimited "$data/case.csv", clear
(8 vars, 26,000 obs)
```

Make sure caseid is uid

```
. isid caseid
```

Merge the case and demo datasets together so that each row in the case dataset also contains the demographics of the defendant.

Note: No other variables (except person_id) in demo and case sharing the same variable name

```
. merge m:1 person_id using "$processed/demo_clean.dta", nogen keep(3)
```

Result	# of obs.
not matched	0
matched	26,000

While the program was mostly rolled out to defendants in Chicago, the State's Attorney's Office also ran a pilot serving a small number of individuals arrested in other parts of Cook County.

For the purpose of this analysis, please restrict the data to only individuals who were arrested in Chicago.

```
.   replace address = lower(address)
(26,000 real changes made)
.   keep if strpos(address, ", chicago") > 0
(1,000 observations deleted)
```

Create an age variable equal to the defendant's age at the time of arrest for each case.

When constructing GPA, please use a 4 point scale, where: A=4, B=3, C=2, D=1, and F=0.

```
.   qui: codebook arrest_date bdate

.   foreach var of varlist arrest_date bdate{
  2.     gen `var´_dt = date(`var´, "YMD")
  3.   }
.   gen age = round((arrest_date_dt - bdate_dt)/365.25,0.1)
```

Save cleaned case and demographic data

```
.   save "$processed/case_demo_clean.dta", replace
file /Users/celiazhu/Box/projects/ra_code_sample/processed/case_demo_clean.dta saved
```

1.3 Clean grade data for defendants in their early early adulthood (18-24)

The State's Attorney's Office has requested 9th and 10th grade course grade data from defendants between the ages of 18 and 24. These data are included in grades.csv. Construct measures for 9th and 10th grade GPA for this target population. When constructing GPA, use a 4 point scale, where: A=4, B=3, C=2, D=1, and F=0.

Import grade data

```
.   import delimited "$data/grades.csv", clear
(17 vars, 11,251 obs)
```

Construct 9th and 10th GPA for defendants between the age of 18 and 24

```
.   foreach g of varlist gr* {
  2.     gen byte n_`g´ = cond(`g´ == "A", 4, ///
>                 cond(`g´ == "B", 3, ///
>                 cond(`g´ == "C", 2, ///
>                 cond(`g´ == "D", 1, ///
>                 cond(`g´ == "F", 0, .)))))
  3.   }
(2,251 missing values generated)
(2,251 missing values generated)
(2,251 missing values generated)
(2,251 missing values generated)
(2,251 missing values generated)
(2,251 missing values generated)
(2,251 missing values generated)
(2,251 missing values generated)
(2,251 missing values generated)
(2,251 missing values generated)
```

```
(2,251 missing values generated)
(2,251 missing values generated)
(2,251 missing values generated)
(2,251 missing values generated)
(2,251 missing values generated)

.   forvalues i = 9/10{
  2.     local grade`i´ n_gr`i´_*
  3.      egen gpa`i´ = rmean(`grade`i´´)
  4.   }
```

Sanity check on GPAs

```
.   su gpa9
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        gpa9 |     11,251    2.661426    .6469981          0          4

.   su gpa10
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       gpa10 |     11,251    2.661316    .6479062        .25          4
```

Keep person id and gpa for grade 9 and 10

```
.   keep person_id gpa*

.   isid person_id

.   save "$processed/grades_clean.dta", replace
file /Users/celiazhu/Box/projects/ra_code_sample/processed/grades_clean.dta saved
```

2. Statistical Analysis

Determine if the program should be continued/expanded by estimating the program's effect on re-arrests prior to disposition.

Do not use gpa to inform analysis because theses are only for young adults

```
.   use "$processed/case_demo_clean.dta", clear
```

The study population should have 25,000 subjects

```
.   assert(_N == 25000)
```

Label variables

```
.   label var age "Age"

.   label var prior_arrests "Number of prior arrests"

.   label var re_arrest "Re-arrested"

.   label var treat "Enrolled into program"
```

Create dummies for gender and race

```
.   tab gender, m gen(gender_)
     gender |      Freq.     Percent        Cum.
------------+-----------------------------------
          F |      4,936       19.74       19.74
          M |     20,064       80.26      100.00
```

```
------------+-----------------------------------
      Total |      25,000      100.00

. rename gender_1 female

. label var female "Female"

. rename gender_2 male

. label var male "Male"

. tab race, m gen(race_)

       race |      Freq.     Percent        Cum.
------------+-----------------------------------
      ASIAN |      1,239        4.96        4.96
      BLACK |     18,249       73.00       77.95
      WHITE |      5,512       22.05      100.00
------------+-----------------------------------
      Total |     25,000      100.00

. rename race_1 asian

. rename race_2 black

. rename race_3 white

. label var asian "Asian"

. label var black "Black"

. label var white "White"

. local balancevar "female male asian black white prior_arrests age"
```

2.1 Summary statistics of study population

Describe the demographic characteristics of the study population based on the data available. (Hint: the study population has 25,000 subjects).

```
. eststo clear

. qui estpost su `balancevar´

. esttab using "$output/summary_statistics.tex", replace ///
>   cells("mean sd min max") nonumber booktabs
(output written to /Users/celiazhu/Box/projects/ra_code_sample/output/summary_statistics.tex)
```

Table 1: Summary statistics of study population

	mean	sd	min	max
female	.19744	.3980751	0	1
male	.80256	.3980751	0	1
asian	.04956	.2170385	0	1
black	.72996	.4439891	0	1
white	.22048	.4145786	0	1
prior_arrests	3.79848	2.138311	0	16
age	30.34043	7.802865	9.5	70.1
N	25000			

2.2 Balance tests for demographic characteristics

The treatment and control groups are not balanced. The average numbers of arrests prior to the case arrest date are significantly different in two groups. Cases with more prior arrests are more likely to be treated. The age characteristic is also imbalanced. Older cases are more likely to be treated. It signals the existence of selection.

Test the baseline equivalence of demographic characteristics

```
.   balancetable treat `balancevar´ using "$output/balance_test.tex", ///
>   vce(robust) pval ///
>   ctitles("Control group" "Treatment group" "Difference" "P value") ///
>   booktabs varlabels replace
```

Table 2: Balance test

Variable	(1) Control group	(2) Treatment group	(3) Difference
Female	0.200	0.195	-0.005
	(0.400)	(0.396)	(0.338)
Male	0.800	0.805	0.005
	(0.400)	(0.396)	(0.338)
Asian	0.049	0.050	0.002
	(0.215)	(0.219)	(0.508)
Black	0.728	0.732	0.004
	(0.445)	(0.443)	(0.445)
White	0.224	0.218	-0.006
	(0.417)	(0.413)	(0.245)
Number of prior arrests	3.152	4.381	1.229***
	(1.849)	(2.213)	(0.000)
Age	28.727	31.795	3.068***
	(7.170)	(8.061)	(0.000)
Observations	11,851	13,149	25,000

2.3 Visualize number of prior arrests by enrollment status and race

Create a numerical variable for race

```
.   preserve
.   encode race, gen(n_race)

.   qui: codebook n_race

.   gen avg = .
(25,000 missing values generated)
.   gen ci_low = .
(25,000 missing values generated)
.   gen ci_high = .
```

```
(25,000 missing values generated)
```

Calculate means and confidence intervals

```
.   qui: mean prior_arrests, over(treat n_race)
.   matrix M = r(table)
.   matrix list M
M[9,6]
         prior_arr~s:  prior_arr~s:  prior_arr~s:  prior_arr~s:  prior_arr~s:  prior_arr~s:
           _subpop_1     _subpop_2     _subpop_3     _subpop_4     _subpop_5     _subpop_6
     b     3.1284722     3.1516698      3.158808     4.2880845     4.3943896     4.3575673
    se     .07721162     .01987445     .03608994     .07927011      .0227264     .04104244
     t     40.518155     158.57896     87.526007     54.094593     193.36061     106.17222
pvalue             0             0             0             0             0             0
    ll     2.9771329     3.1127147     3.0880696     4.1327104     4.3498445     4.2771217
    ul     3.2798115     3.1906249     3.2295464     4.4434586     4.4389347     4.4380129
    df         24999         24999         24999         24999         24999         24999
  crit     1.9600589     1.9600589     1.9600589     1.9600589     1.9600589     1.9600589
 eform             0             0             0             0             0             0

.   forvalues i = 1/6 {
  2.       if inrange(`i´, 1, 3) == 1{
  3.       replace avg = M[1, `i´] if treat == 0 & n_race == `i´
  4.       replace ci_low = M[5, `i´] if treat == 0 & n_race == `i´
  5.       replace ci_high = M[6, `i´] if treat == 0 & n_race == `i´
  6.       }

.       if inrange(`i´, 4, 6) == 1 {
  8.       replace avg = M[1, `i´] if treat == 1 & n_race == `i´-3
  9.       replace ci_low = M[5, `i´] if treat == 1 & n_race == `i´-3
 10.       replace ci_high = M[6, `i´] if treat == 1 & n_race == `i´-3
 11.       }
 12.   }
(576 real changes made)
(576 real changes made)
(576 real changes made)
(8,624 real changes made)
(8,624 real changes made)
(8,624 real changes made)
(2,651 real changes made)
(2,651 real changes made)
(2,651 real changes made)
(663 real changes made)
(663 real changes made)
(663 real changes made)
(9,625 real changes made)
(9,625 real changes made)
(9,625 real changes made)
(2,861 real changes made)
(2,861 real changes made)
(2,861 real changes made)
```

Count observations

```
.   forvalues i = 1/6 {
  2.       if inrange(`i´, 1, 3) == 1 {
  3.       count if treat == 0 & n_race == `i´ & !missing(prior_arrests)
  4.       }

.       if inrange(`i´, 4, 6) == 1 {
```

```
  6.       count if treat == 1 & n_race == `i´-3 & !missing(prior_arrests)
  7.       }

.       local `i´N = r(N)
  9.   }
  576
  8,624
  2,651
  663
  9,625
  2,861
```

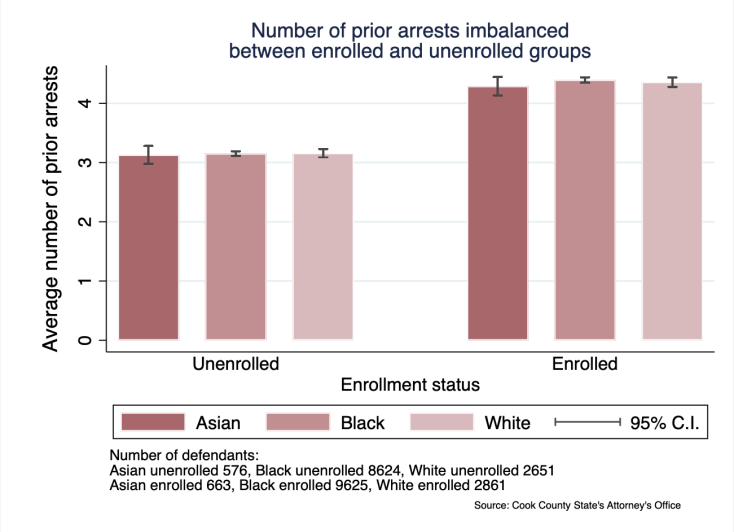
Plot the bar chart

```
.   gen treat_race = n_race if treat == 0
(13,149 missing values generated)
.   replace treat_race = n_race + 4 if treat == 1
(13,149 real changes made)

.   twoway (bar avg treat_race if n_race == 1, fcolor(maroon) ///
>           fintensity(inten70) lcolor(white) barw(0.7)) ///
>          (bar avg treat_race if n_race == 2, fcolor(maroon) ///
>          fintensity(inten50) lcolor(white) barw(0.7)) ///
>          (bar avg treat_race if n_race == 3, fcolor(maroon) ///
>          fintensity(inten30) lcolor(white) barw(0.7)) ///
>              (rcap ci_low ci_high treat_race, lcolor(gs5)), ///
>          legend(row(1) order(1 "Asian" 2 "Black" 3 "White" 4 "95% C.I.")) ///
>          xlabel(2"Unenrolled" 6"Enrolled", noticks) xtitle("Enrollment status") ///
>          ylabel(0(1)4) ytitle("Average number of prior arrests", ///
>          margin(medium) size(medium)) ///
>          title("Number of prior arrests imbalanced" ///
>          "between enrolled and unenrolled groups", size(medium)) ///
>          note("Number of defendants:" ///
>          "Asian unenrolled `1N´, Black unenrolled `2N´, White unenrolled `3N´" ///
>            "Asian enrolled `4N´, Black enrolled `5N´, White enrolled `6N´") ///
>          caption("Source: Cook County State´s Attorney´s Office", ///
>          justification(left) size(vsmall) linegap(0.8) position(5) span) ///
>          graphregion( color(white) ) plotregion(fcolor(white))

.   graph export "$output/prior_arrests.png", replace
(file /Users/celiazhu/Box/projects/ra_code_sample/output/prior_arrests.png written in PNG format)
.   restore
```

Figure 1: Number of prior arrests imbalanced bewteen enrolled and unerolled groups



2.4 Estimate the effect of the program on reducing the likelihood of re-arrest before disposition

One difficulty is that we don't have enough information about the implementation of the program. We are not sure if the program was randomized and how was the compliance.

Specification 1: OLS

$$rearrest_{ic} = \tau treat_{ic} + \beta X_{ic} + \epsilon_{ic} + \epsilon_i$$

where i is the individual, c is the case, X is the vector for prior arrests, gender, race, and age.

We start with "naive" OLS. It can correctly estimate the average treatment effect if this program was a **randomized experiment** with perfect compliance; or it was an **observational study** satisfying the two conditions: 1) there is no selection on unobservables and we've controlled all observables that could be selected upon; and 2) how people are self-selected based on those variables can be approximated by a linear function. But these conditions are unlikely to be true.

If the program was a randomized trial, the treatment was administered on the case level, not on the individual level. Hence, we don't cluter standard errors here.

Also, by examining the unique values of `person_id`, we can conclude that for most defendants, they only have one or two cases. Indiviudal fixed effects is not desirable here.

```
.   unique person_id
Number of unique values of person_id is  14353
Number of records is  25000

.   eststo clear

.   eststo: qui reg re_arrest treat `balancevar´, rob
(est1 stored)
```

Specification 2: Logit

$$Pr(rearrest = 1|X_c) = \frac{exp^{X_c'\beta}}{1 + e^{X_c'\beta}}$$

Assume the program was not an experiment, we can improve our **prediction** on the likelihood by using a logit model instead of a linear probability model, which was implemented in specification 1. But the results given by logit specification is for prediction, not for causal inference.

```
.   eststo: qui logit re_arrest treat `balancevar´, rob
(est2 stored)
```

Specification 3: Weighted propensity score matching To estimate the causal effect of this treatment given that the program was observational, not experimental, we still need to assume that there is no selection on unobservables and we've controlled all observables that could be selected upon. But we can relax the assumption on the functional form.

Still, the estimate based upon propensity score matching is not entirely valid, because we omit variables such as grades, household income, etc..

However, this is less restrictive and therefore more plausible than the OLS estimate, and better serves the purpose of causal inference compared to logit estimate.

```
.   qui logit treat `balancevar´
.   predict pscore
(option pr assumed; Pr(treat))
```

Common support

```
.   forvalues i = 1/2 {
  2.     su pscore if treat == `i´ -1
  3.     drop if inrange(pscore, r(min), r(max)) == 0
  4.   }
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      pscore |     11,851    .4821746    .1287615   .2568446   .9319222
(16 observations deleted)
```

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      pscore |     13,133    .5649514    .1451764   .2577653   .9314315
(5 observations deleted)
```

Weight for average treatment effect (ATE)

```
.   gen ate_weight = (1/pscore) if treat == 1
(11,846 missing values generated)
.   replace ate_weight = 1/(1-pscore) if treat == 0
(11,846 real changes made)
.   eststo: qui reg re_arrest treat [pw = ate_weight]
(est3 stored)
```

Weight for average treatment effect on the treated (ATET)

```
.   gen atet_weight = 1 if treat == 1
(11,846 missing values generated)
.   replace atet_weight = pscore/(1-pscore) if treat == 0
(11,846 real changes made)
.   eststo: qui reg re_arrest treat [pw = atet_weight]
(est4 stored)
```

Export results of three specifications

```
. esttab using "$output/estimation.tex", se booktabs label ///
> addnotes("Model 1: OLS; Model 2: Logit;" ///
> "Model 3: Propensity score matching (ATT); Model 4: Propensity Score Matching (ATET)" ///
> "Souce: Cook County State´s Attorney´s Office") ///
> replace numbers
(output written to /Users/celiazhu/Box/projects/ra_code_sample/output/estimation.tex)
```

Table 3: Estimation

	(1) Re-arrested	(2) Re-arrested	(3) Re-arrested	(4) Re-arrested
main				
Enrolled into program	-0.0154**	-0.0936**	-0.0171**	-0.0170**
	(0.00532)	(0.0330)	(0.00549)	(0.00621)
Female	-0.00520	-0.0319		
	(0.00635)	(0.0397)		
Male	0	0		
	(.)	(.)		
Asian	-0.00337	-0.0178		
	(0.0126)	(0.0790)		
Black	0.00237	0.0143		
	(0.00619)	(0.0382)		
White	0	0		
	(.)	(.)		
Number of prior arrests	0.0160***	0.0922***		
	(0.00162)	(0.00942)		
Age	0.00376***	0.0220***		
	(0.000456)	(0.00267)		
Constant	0.0431***	-2.328***	0.219***	0.235***
	(0.0120)	(0.0741)	(0.00421)	(0.00506)
Observations	25000	25000	24979	24979

Standard errors in parentheses

Model 1: OLS; Model 2: Logit;

Model 3: Propensity score matching (ATT); Model 4: Propensity Score Matching (ATET)

Souce: Cook County State's Attorney's Office

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3. Conclusion

Overall, the treatment significantly reduces the likelihood of re-arrest before disposition. With the information we have, we can conclude that the program is effective and should be expanded or continued, or should be furthered examined with an experiment.

However, please note that the causal inference has much room for improvement. It would have better performance if we can obtain more relevant variables, such as grades, household income, neighboorhoods.