Stata Code Sample *

Xiling (Celia) 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, arrrests information, and academic performance. We 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

```
. clear all
. set more off
. set varabbrev off
.
. * Set up directories
. global data /Users/celiazhu/Box/projects/ra_code_sample/data
. global output /Users/celiazhu/Box/projects/ra_code_sample/stata/output
. global processed /Users/celiazhu/Box/projects/ra_code_sample/stata/processed
```

1. Data Cleaning

1.1 Clean demographic data

```
. * Import 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)
```

*Generated by `markstat`. For souce code, please see my github repository here.

```
. isid person_id
```

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/stata/processed/demo_clean.dta saved
```

1.2 Clean arrests data (data on arrests is named as "case")

Merge the case and demo datasets together so that each row in the case dataset also contains the demographics of the defendant. It's possible to have one person with multiple cases.

I didn't find other variables (except person_id) in demo and case sharing same variable names.

```
. * Import arrests data
. import delimited "$data/case.csv", clear
(8 vars, 26,000 obs)
.
. * Make sure caseid is uid
. isid caseid
.
. merge m:1 person_id using "$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.

I first change all addresses to lower case because it's possible to have "Chicago" inconsistently capitalized as "CHICAGO", "Chicago", or even "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.

```
. * Glimpse the `arrest_date` and `bdate` to know their formats.
. codebook arrest_date bdate
-------------------------------------------------------------------------------
arrest_date
-------------------------------------------------------------------------------

                  type:  string (str10)
         unique values:  666                      missing "":  0/25,000
              examples:  "2012-05-12"
                         "2012-09-25"
                         "2013-02-06"
                         "2013-06-23"
-------------------------------------------------------------------------------
bdate
-------------------------------------------------------------------------------

                  type:  string (str10)
         unique values:  7,603                    missing "":  0/25,000
              examples:  "1976-05-05"
                         "1981-04-18"
                         "1985-02-25"
                         "1989-04-10"
. 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/stata/processed/case_demo_clean.dta saved
```

1.3 Clean grade data for defendants in their early early adulthood

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 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)
(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 gpa*
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        gpa9 |     11,251    2.661426    .6469981          0          4
       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 cleaned grades data
. save "$processed/grades_clean.dta", replace
file /Users/celiazhu/Box/projects/ra_code_sample/stata/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**.

Because we only have grades data for young adults, I did not use these data to inform your statistical analysis.

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

```
. * The study population should have 25,000 subjects
. assert(_N == 25000)
.
. * 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"
.
. * Label variables for cleaner output.
. label var age "Age"
. label var prior_arrests "Number of prior arrests"
. label var re_arrest "Re-arrested"
. label var treat "Enrolled into program"
.
. label define treat_lab 0 "Unenrolled" 1 "Enrolled"
. label define male_lab 0"Female" 1"Male"
. label define black_lab 0"Non-Black" 1"Black"
. label define white_lab 0"Non-White" 1"White"
.
. foreach var of varlist treat male black white {
  2.   label values `var´ `var´_lab
  3. }
.
. * Save cleaned data with dummies
. save "$processed/analysis_data.dta", replace
file /Users/celiazhu/Box/projects/ra_code_sample/stata/processed/analysis_data.dta saved
.
. * Define a local macro for covariates.
. local balancevar "female male asian black white prior_arrests age"
```

2.1 Summary statistics of study population

The study population are predominantly male, with only 20% cases having female defendants. As to race, 73% of the cases involve Black defendants, only 22% are white, and 5% are Asian. On average, the study population has around 4 prior arrests before the case arrest date, and their average age is approximately 30.

```
. eststo clear
. qui estpost su `balancevar´
. esttab using "$output/summary_statistics.tex", replace ///
> cells("mean(fmt(2)) sd(fmt(0 0 0 0 0 0 1)) min(fmt(0 0 0 0 0 0 1)) max(fmt(0 0 0 0 0 0 1))") ///
> collabel("Mean" "Standard Deviation" "Min" "Max" ) ///
> width(\textwidth) nonumber  label
(output written to /Users/celiazhu/Box/projects/ra_code_sample/stata/output/summary_statistics.tex)
```

Table 1: Summary statistics of study population

	Mean	Standard Deviation	Min	Max
Female	0.20	0	0	1
Male	0.80	0	0	1
Asian	0.05	0	0	1
Black	0.73	0	0	1
White	0.22	0	0	1
Number of prior arrests	3.80	2	0	16
Age	30.34	7.8	9.5	70.1
Observations	25000			

2.2 Balance tests for demographic characteristics

The enrolled and unenrolled groups are not balanced at the baseline.

The average numbers of prior arrests are significantly different in the two groups. Cases with more prior arrests are more likely to be enrolled into the program.

Their age is also imbalanced. Cases with older defendants are more likely to be enrolled into the program.

The imbalance are not due to random coincidence. The F-test indicates that these covariates didn't pass joint orthogonality, either.

The imbalance at baseline signals the problem of selection.

```
. iebaltab `balancevar´, grpvar(treat) ///
> vce(robust) savetex("$output/balance_test.tex") replace ///
> rowvarlabels pttest ftest fnoobs pftest
    Balance table saved to: /Users/celiazhu/Box/projects/ra_code_sample/stata/output/balance_test.tex
```

Table 2: Balance test

Variable	(1) Unenrolled N	(1) Unenrolled Mean/SE	(2) Enrolled N	(2) Enrolled Mean/SE	T-test P-value (1)-(2)
Female	11851	0.200	13149	0.195	0.338
		(0.004)		(0.003)	
Male	11851	0.800	13149	0.805	0.338
		(0.004)		(0.003)	
Asian	11851	0.049	13149	0.050	0.508
		(0.002)		(0.002)	
Black	11851	0.728	13149	0.732	0.445
		(0.004)		(0.004)	
White	11851	0.224	13149	0.218	0.245
		(0.004)		(0.004)	
Number of prior arrests	11851	3.152	13149	4.381	0.000***
		(0.017)		(0.019)	
Age	11851	28.727	13149	31.795	0.000***
		(0.066)		(0.070)	
F-test of joint significance (p-value)					0.000***

Notes: The value displayed for t-tests are p-values. The value displayed for F-tests are p-values. Standard errors are robust. ***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

2.3 Visualize number of prior arrests by enrollment status and race

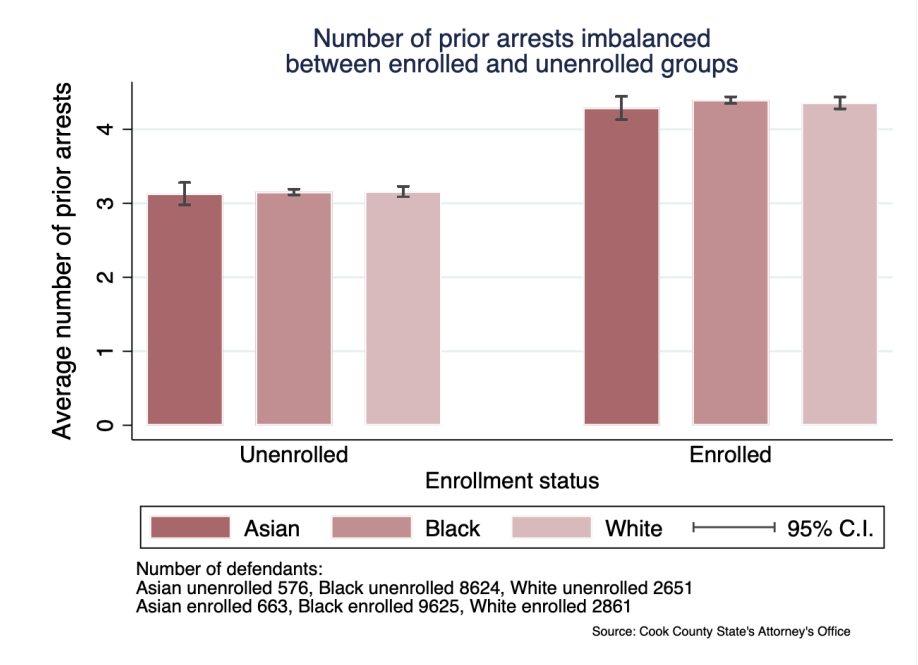
Create a numerical variable for race

```
. 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.   }
  7. 
.   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.   }
  5. 
.   if inrange(`i´, 4, 6) == 1 {
  6.   count if treat == 1 & n_race == `i´-3 & !missing(prior_arrests)
  7.    }
  8. 
.    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/stata/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 in estimating the effect of the program is that I don't have enough information about the program: if program was an randomized controlled trial and if so, how was the compliance, or if it was an observational study.

Specification 1: OLS (or Linear Probaility Model) We start with "naive" OLS model.

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

where i is the individual, c is the case, X_{ic} is the vector for race, gender, age, and number of prior arrests. We could have added individual fixed effects u_i to control for those individual-invariant characteristics. But by examining the unique values of `person_id`, we can conclude that for most defendants, they only have one or two cases. Individual fixed effect is not desirable here.

It can correctly estimate the average treatment effect (ATE) 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.

That said, suppose the program was a randomized controlled trial, the treatment was administered on the case level, not on the individual level. Hence, we don't cluster standard errors here.

Specification 2: Logit Assume the program was not an experimental study, 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.

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

Specification 3: Inverse Probability Weighting (IPW) To estimate the causal effect via propensity score matching (more specifically, inverse probability weighting (IPW)), 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. One improvement from OLS or logit model is that we can relax the assumption on the functional form.

One improvement from OLS or logit model is that we can relax the assumption on the functional form. Still, the estimate based upon IPW is not entirely valid, because we omit important characteristics such as grades, household income, neighborhood, etc.. However, this is less restrictive and therefore more plausible than the other two models.

First, calculate propensity and inverse probability weight (ipw).

```
. // Get propensity score
. * Let female and asian be base groups
. local covlist "male black white prior_arrests age"
. qui logit treat `covlist´
. predict pscore, pr
.
. // Inverse probability weighting (IPW)
. * Weight for average treatment effect (ATE)
. gen double ate_weight = 1.treat/pscore + 0.treat/(1-pscore)
. * Weight for average treatment effect on the treated (ATET)
. gen double atet_weight = 1.treat/1 + 0.treat * pscore/(1-pscore)
.
. // Common support
. gen support = 1
. forvalues i = 1/2 {
  2.     su pscore if treat == `i´ -1
  3.     replace support = 0 if inrange(pscore, r(min), r(max)) == 0
  4.   }
    Variable |        Obs        Mean    Std. Dev.       Min        Max
```

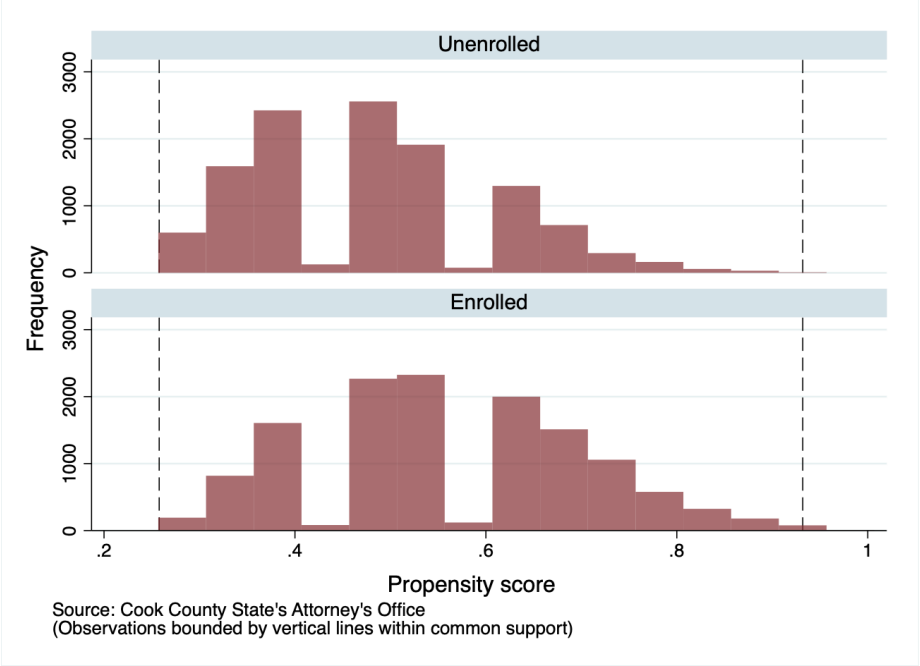
```
      pscore |     11,851    .4821746    .1287615    .2568446    .9319222
(16 real changes made)
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      pscore |     13,149    .5654231    .1457166    .2577653    .9778187
(4 real changes made)
```

Before we estimate average treatment effect (ATE) and average treatment effect on the treated (ATET) of this program, we first assess the validity of propensity score by 1) the distribution of the propensity score, 2) summary statistics of observations within the common support, and 3) the balance of the sample within common support, weighted by the IPW for ATE.

1) Assess the distribution of the propensity score in enrolled and unerolled groups

```
. su pscore if support == 1
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      pscore |     24,980    .5257297     .143676    .2577653    .9319222
. local lhs = r(min)
. local rhs = r(max)
. di `lhs´
.25776526
. di `rhs´
.9319222
. histogram pscore, xline(`lhs´, lcolor(black) lwidth(thin) lpattern(dash)) ///
> xline(`rhs´, lcolor(black) lwidth(thin) lpattern(dash)) ///
> freq width(0.05) ///
> by(treat, row(2) graphregion(color(white)) note("Source: Cook County State´s Attorney´s Office" ///
> "(Observations bounded by vertical lines within common support)", size(small))) ///
> fcolor(maroon%70) lcolor(white%0) ///
> xtitle("Propensity score", size(medsmall)) ///
> ytitle("Frequency", size(medsmall))
.
. graph export "$output/ps_hist.png", replace
(file /Users/celiazhu/Box/projects/ra_code_sample/stata/output/ps_hist.png written in PNG format)
```

Figure 2: histogram of propensity score by enrollment status



2) Summary statistics of observations within common support, weighted by inverse probability for ATE

```
. eststo clear
. qui estpost su `balancevar´ if support == 1 [aw = ate_weight]
. esttab using "$output/summary_statistics_cs.tex", replace ///
> cells("count(fmt(0)) mean(fmt(2)) sd(fmt(2)) min(fmt(0 0 0 0 0 0 1)) max(fmt(0 0 0 0 0 0 1))") ///
> collabel("N" "Mean" "Standard Deviation" "Min" "Max" ) ///
> width(\textwidth) nonumber  label
(output written to /Users/celiazhu/Box/projects/ra_code_sample/stata/output/summary_statistics_cs.tex)
```

Table 3: Summary statistics for observations within common support, weighted by inverse probability for ATE

	N	Mean	Standard Deviation	Min	Max
Female	24980	0.20	0.40	0	1
Male	24980	0.80	0.40	0	1
Asian	24980	0.05	0.22	0	1
Black	24980	0.73	0.44	0	1
White	24980	0.22	0.41	0	1
Number of prior arrests	24980	3.80	2.13	0	12
Age	24980	30.36	7.83	9.5	70.1
Observations	24980				

3) Balance test for observations within common support, weighted by inverse probability for ATE

```
. iebaltab `balancevar´  if support == 1 [pw = ate_weight], grpvar(treat)  ///
> vce(robust) savetex("$output/balance_test_ipw.tex") replace ///
> rowvarlabels pttest ftest fnoobs pftest
    Balance table saved to: /Users/celiazhu/Box/projects/ra_code_sample/stata/output/balance_test_ipw.tex
```

Table 4: Balance test for observation within common support, weighted by inverse probability for ATE

	(1) Unenrolled		(2) Enrolled		T-test P-value
Variable	N	Mean/SE	N	Mean/SE	(1)-(2)
Female	11847	0.198	13133	0.197	0.913
		(0.004)		(0.004)	
Male	11847	0.802	13133	0.803	0.913
		(0.004)		(0.004)	
Asian	11847	0.050	13133	0.049	0.905
		(0.002)		(0.002)	
Black	11847	0.730	13133	0.730	0.980
		(0.004)		(0.004)	
White	11847	0.221	13133	0.221	0.929
		(0.004)		(0.004)	
Number of prior arrests	11847	3.803	13133	3.793	0.780
		(0.027)		(0.019)	
Age	11847	30.390	13133	30.334	0.635
		(0.094)		(0.069)	
F-test of joint significance (p-value)					0.998

Notes: The value displayed for t-tests are p-values. The value displayed for F-tests are p-values. Standard errors are robust. Observations are weighted using variable ate_weight as pweight weights.***, **, and * indicate significance at the 1, 5, and 10 percent critical level.

OLS regression

```
. 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 `covlist´, rob
(est1 stored)
```

Logit regression

```
. eststo: logit re_arrest treat `covlist´, rob cluster(person_id)

Iteration 0:   log pseudolikelihood =  -12852.89
Iteration 1:   log pseudolikelihood = -12621.071
Iteration 2:   log pseudolikelihood = -12618.548
Iteration 3:   log pseudolikelihood = -12618.548

Logistic regression                             Number of obs     =     25,000
                                                Wald chi2(6)      =     680.50
                                                Prob > chi2       =     0.0000
```

```
Log pseudolikelihood = -12618.548               Pseudo R2         =     0.0182
                          (Std. Err. adjusted for 14,353 clusters in person_id)
```

re_arrest	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
treat	-.0935816	.0328717	-2.85	0.004	-.1580089	-.0291543
male	.0318649	.0379127	0.84	0.401	-.0424426	.1061724
black	.0321375	.0730498	0.44	0.660	-.1110376	.1753125
white	.0178003	.0772279	0.23	0.818	-.1335637	.1691642
prior_arrests	.0922383	.008603	10.72	0.000	.0753768	.1090999
age	.0220245	.0025201	8.74	0.000	.0170852	.0269637
_cons	-2.377173	.0996872	-23.85	0.000	-2.572556	-2.181789

```
(est2 stored)
```

Estimate ATE and ATET with inverse probability weighting.

```
. eststo: qui reg re_arrest treat [pw = ate_weight] if support == 1
(est3 stored)
. eststo: qui reg re_arrest treat [pw = atet_weight] if support == 1
(est4 stored)
```

Export results from three specifications

```
. esttab using "$output/estimation.tex", se  label nobaselevels noomitted ///
> addnotes("Model 1: OLS; Model 2: Logit; Model 3: IPW (ATE); Model 4: IPW (ATET)" ///
> "Odds ratio reported in logit model" ///
> "Robust s.e. reported in OLS model; clustered robust s.e. reported in logit model") ///
> eform(0 1 0 0 ) replace eqlabels(none)
(output written to /Users/celiazhu/Box/projects/ra_code_sample/stata/output/estimation.tex)
```

Table 5: Estimation

	(1)	(2)	(3)	(4)
	Re-arrested	Re-arrested	Re-arrested	Re-arrested
Enrolled into program	-0.0154**	0.911**	-0.0169**	-0.0168**
	(0.00532)	(0.0299)	(0.00549)	(0.00621)
Male	0.00520	1.032		
	(0.00635)	(0.0391)		
Black	0.00574	1.033		
	(0.0118)	(0.0754)		
White	0.00337	1.018		
	(0.0126)	(0.0786)		
Number of prior arrests	0.0160***	1.097***		
	(0.00162)	(0.00943)		
Age	0.00376***	1.022***		
	(0.000456)	(0.00258)		
Constant	0.0346*	0.0928***	0.219***	0.234***
	(0.0163)	(0.00925)	(0.00421)	(0.00506)
Observations	25000	25000	24980	24980

Standard errors in parentheses

Model 1: OLS; Model 2: Logit; Model 3: IPW (ATE); Model 4: IPW (ATET)

Odds ratio reported in logit model

Robust s.e. reported in OLS model; clustered robust s.e. reported in logit model

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

3. Conclusion

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

However, please note that the causal inference has much room for improvement. Though the sample on the common support and weighted by the inverse probability for ATE is balanced, the distribution of propensity score is not ideal. Some cases not selected into the program still have relatively high propensity score. And the distribution didn't improve much when I added higher-ordered terms, like age squared.

The causal inference would have better performance if we can obtain more characteristics. To name a few, household income, and the neighborhoods

defendants live in; and for younger defendants, we can also incorporate their academic performance and disciplinary incidents in school, and the school districts they live in.