Stata Code Sample *

Xinghuan Luo xinghuanluo@uchicago.edu

Aug 28, 2020

Background

This is a code sample extracted from my homework of econometrics class at harris. The questions in this homework are based heavily on the paper Almond et al. 2005. The goal of this assignment is to examine the research question: what is the causal effect of maternal smoking during pregnancy on infant birthweight and other infant health outcomes? The data for the problem set is an extract of all births from the 1993 National Natality Detail Files for Pennsylvania.

This code file has 4 parts:
- Part 0 Initialization and Importing data
- Part 1 Data Checking
- Part 2 Estimating Propensity Score
- Part 3 Analysis with Propensity Score

Basically I first checked the missing pattern of data and cleaned it. Then, I found that the covariates were unbalanced between smoking and non-smoking group. Also because the dataset was not the result of random experiement, to control for selection on observables, I used propensity score matching method. Finally, after estimating p-score, I used it with three different ways: simply including p-score as covariate, weighting both group by p-score and diving whole population by p-score.

0. Initialization and Importing data

```
. clear
. set more off
. local data "D:/OneDrive - The University of Chicago/2021 Fall/job searching/Urban Lab/code_sample/data"
. local output "D:/OneDrive - The University of Chicago/2021 Fall/job searching/Urban Lab/code_sample/out
> put"
. use "`data´/dataset.dta"
.
. if inlist("`c(username)´", "lenovo" ) {
. cd "D:/OneDrive - The University of Chicago/2021 Fall/job searching/Urban Lab/code_sample/output"
D:\OneDrive - The University of Chicago\2021 Fall\job searching\Urban Lab\code_sample\output
. }
```

From the codebook, below variables have missing values. I used Little's test (mcartest) to see if the below variables are missing completely at random.

```
. preserve
.
. local var_mi_99 cigar alcohol wgain
. local var_mi_other tobacco cigar6 alcohol drink herpes
.
. foreach var of varlist `var_mi_99´{
  2. qui replace `var´ =. if `var´ == 99
  3. }
.
. foreach var of varlist `var_mi_other´{
  2. qui replace `var´ =. if inlist(1, tobacco==9, cigar6==6, alcohol==9, drink5==5, herpes==8)
  3. }
.
. mcartest tobacco cigar cigar6 alcohol drink drink5 wgain herpes
Little´s MCAR test
Number of obs       = 97583
```

*I finished this markdown file by `markstat`. The souce code is in my github repository here.

```
Chi-square distance = 86119.9200
Degrees of freedom  = 24
Prob > chi-square   = 0.0000
. restore
.
. // Drop variable with missing values.
.
. drop if 1 == inlist(1, tobacco==9, cigar==99, cigar6==6, alcohol==9, drink==99, drink5==5, wgain==99, h
> erpes==8)
(4,755 observations deleted)
```

1. Data Checking

List a group of predetermiend variables as covariates

```
. local predetermined mrace3 dmeduc dmar dfeduc orfath cntocpop stresfip ormoth nprevist adequacy alcohol
>  drink drink5 preterm pre4000 phyper monpre rectype anemia cardiac lung diabetes herpes chyper disllb i
> sllb10 birmon stresfip pldel3 nlbnl dlivord dtotord totord9 weekday dgestat csex dplural
```

1.1 Balancing Test

To find if there is any selection bias, I did balance check of part of important variables between treatment and control group, because creating a balance table for all variables will make this sample too long. But, practically, I should do it for all variables. If the pregnant women smokes, then tobacco = 1(treatment). Otherwise, tobacco = 0(control).

```
. local predetermined_balance mrace3 dmeduc dmar dfeduc orfath cntocpop stresfip
. label define tobacco_lab 0 "non-smoking" 1 "smoking"
. label val tobacco tobacco_lab
. qui replace tobacco = 0 if tobacco == 2
. iebaltab `predetermined_balance´, grpvar(tobacco) ///
> vce(robust) savetex("`output´/balance_test.tex") replace ///
> rowvarlabels pttest ftest fnoobs pftest
    Balance table saved to: D:/OneDrive - The University of Chicago/2021 Fall/job searching/Urban
        Lab/code_sample/output/balance_test.tex
```

Table 1: Balance Test of Predetermined Variables

	(1) non-smoking		(2) smoking		T-test P-value
Variable	N	Mean/SE	N	Mean/SE	(1)-(2)
race of mother recode	78010	1.260	14818	1.254	0.349
		(0.002)		(0.005)	
detailed educ of mother	78010	13.439	14818	11.997	0.000***
		(0.008)		(0.013)	
marital status of mother	78010	1.208	14818	1.479	0.000***
		(0.001)		(0.004)	
educ of father detail	78010	13.490	14818	12.126	0.000***
		(0.008)		(0.014)	
hispanic origin of father	78010	0.097	14818	0.080	0.000***
		(0.002)		(0.004)	
county of occurence	78010	1.422	14818	1.562	0.000***
		(0.004)		(0.010)	
state of residence	78010	41.722	14818	41.869	0.000***
		(0.008)		(0.013)	
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.

1.2 Regression on Whether Smoking

After I controlled the predermined variables, I simply estimated the impact of smoking on birth weight, one minute apgar score and five minute agpar score.

```
. eststo clear
. local i = 1
. foreach var of varlist dbrwt omaps fmaps {
  2. eststo model_`i´: qui reg `var´ tobacco `predetermined´, robust
  3. local i = `i´ + 1
  4. }
. esttab model_* using "`output´/smk_weight.tex", se keep(_cons tobacco ) replace label title(Simple Esti
> mate of Impact of Smoking)
(output written to D:/OneDrive - The University of Chicago/2021 Fall/job searching/Urban Lab/code_sample/
> output/smk_weight.tex)
```

Table 2: Regression on Whether Smoking

Table 3: Simple Estimate of Impact of Smoking

	(1)	(2)	(3)
	birthweight in grams	one minute agpar score	five minute agpar score
tobacco use during pregnancy	-286.8***	0.0183	-0.0208**
	(4.436)	(0.0120)	(0.00652)
Constant	-22.48	3.719***	5.812***
	(127.1)	(0.355)	(0.225)
Observations	92828	92828	92828

Standard errors in parentheses

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

1.3 Categorical Regression on Parent's Age

I simply regressed infant weight on mother and father age controlling those predetermined variable

```
. recode dmage min/26 = 1 27/32 = 2 33/max = 3, gen(dmage_factor)
(92828 differences between dmage and dmage_factor)
. recode dfage min/28 = 1 29/34 = 2 35/max = 3, gen(dfage_factor)
(92828 differences between dfage and dfage_factor)
. label define dmage_lab 1 "below 26" 2 "27-32" 3 "above 33"
. label define dfage_lab 1 "below 28" 2 "29-34" 3 "above 35"
. label val dmage_factor dmage_lab
. label val dfage_factor dfage_lab
. label var dmage_factor "Factorial variable of mother age"
. label var dfage_factor "Factorial variable of father age"
.
. local outcome_weight dbrwt fmaps omaps
. forvalues j = 1/2 {
  2. if `j´ == 1 {
  3.     local idpdt_var dmage_factor
  4. }
  5. else if `j´ == 2 {
  6.     local idpdt_var dfage_factor
  7. }
  8.
. local k = 1
  9. foreach var of varlist `outcome_weight´ {
 10.     qui reg `var ´ i.`idpdt_var´ `predetermined´, robust
 11.     mat b_`k´ = r(table)
 12.     mat b_`k´ = b_`k´[1..2, 2..4]´
 13.     local lbl_var: var label `var´
 14.     mat colnames b_`k´ = "`lbl_var´" "`lbl_var´_se"
 15.     local val_lbls
 16.     forvalues i = 2/3 {
 17.         local val_lbl: label(`idpdt_var´) `i´
 18.         local val_lbls `" `val_lbls´ "`val_lbl´" "´
 19.     }
 20.     local val_lbls `" `val_lbls´ "The Constant" "´
```

```
 21.     mat rownames b_`k´ = `val_lbls´
 22.     local k = `k´ + 1
 23. }
 24. mat final_`j´ = b_1
 25. forvalues k = 2/3 {
 26.     mat final_`j´ = final_`j´, b_`k´
 27. }
 28. }
.
. mat final_matrix = final_1\final_2
. mat list final_matrix
final_matrix[6,6]
              birthweigh_s  birthweigh_e  five minut_e  five minut_e  one minute_e  one minute_e
       27-32      14.92037     4.0710424     .00651721     .00588324    -.00469025     .01059453
    above 33     14.959313     5.1688804     .00548429      .0075422    -.02619256     .01378549
The Constant    -50.855175     2.8304249     -.0284952     .00434988    -.05554367     .00817196
       29-34     8.4427372     3.9343207     .00559474     .00562163     .01426528     .01026885
    above 35      1.893783     4.6807759     .00676338     .00690075      .0057815     .01240228
The Constant    -51.273841     2.8280366    -.02847045     .00433872     -.0541011     .00815766
```

For the above matrix, the first three rows are categorical regression of birth weight and five minute apgar score on mother's age group. "Below 26" group is omitted. The rest of rows are categorical regression of same outcome variables on father's age group, "below 28" is omitted.
The first column is coefficient and the second column is the corresponding standerd error.

1.4 Bar Chart by Mother Age and Smoking Status

I created a bar chart by mother age and whether smoking.

```
. preserve
.
. gen avg = .
(92,828 missing values generated)
. gen ci_low = .
(92,828 missing values generated)
. gen ci_high =.
(92,828 missing values generated)
```

I calculated means and confidence intervals

```
. qui: mean dbrwt, over(tobacco dmage_factor)
. matrix M = r(table)
.
. forvalues i = 1/6 {
  2. if inrange(`i´, 1, 3) == 1 {
  3.     qui replace avg = M[1, `i´] if tobacco == 0 & dmage_factor == `i´
  4.     qui replace ci_low = M[5, `i´] if tobacco == 0 & dmage_factor == `i´
  5.     qui replace ci_high = M[6, `i´] if tobacco == 0 & dmage_factor == `i´
  6. }
  7. else if inrange(`i´, 4, 6) == 1 {
  8.     qui replace avg = M[1, `i´] if tobacco == 1 & dmage_factor == `i´ - 3
  9.     qui replace ci_low = M[5, `i´] if tobacco == 1 & dmage_factor == `i´ - 3
 10.     qui replace ci_high = M[6, `i´] if tobacco == 1 & dmage_factor == `i´ - 3
 11. }
 12. }
```

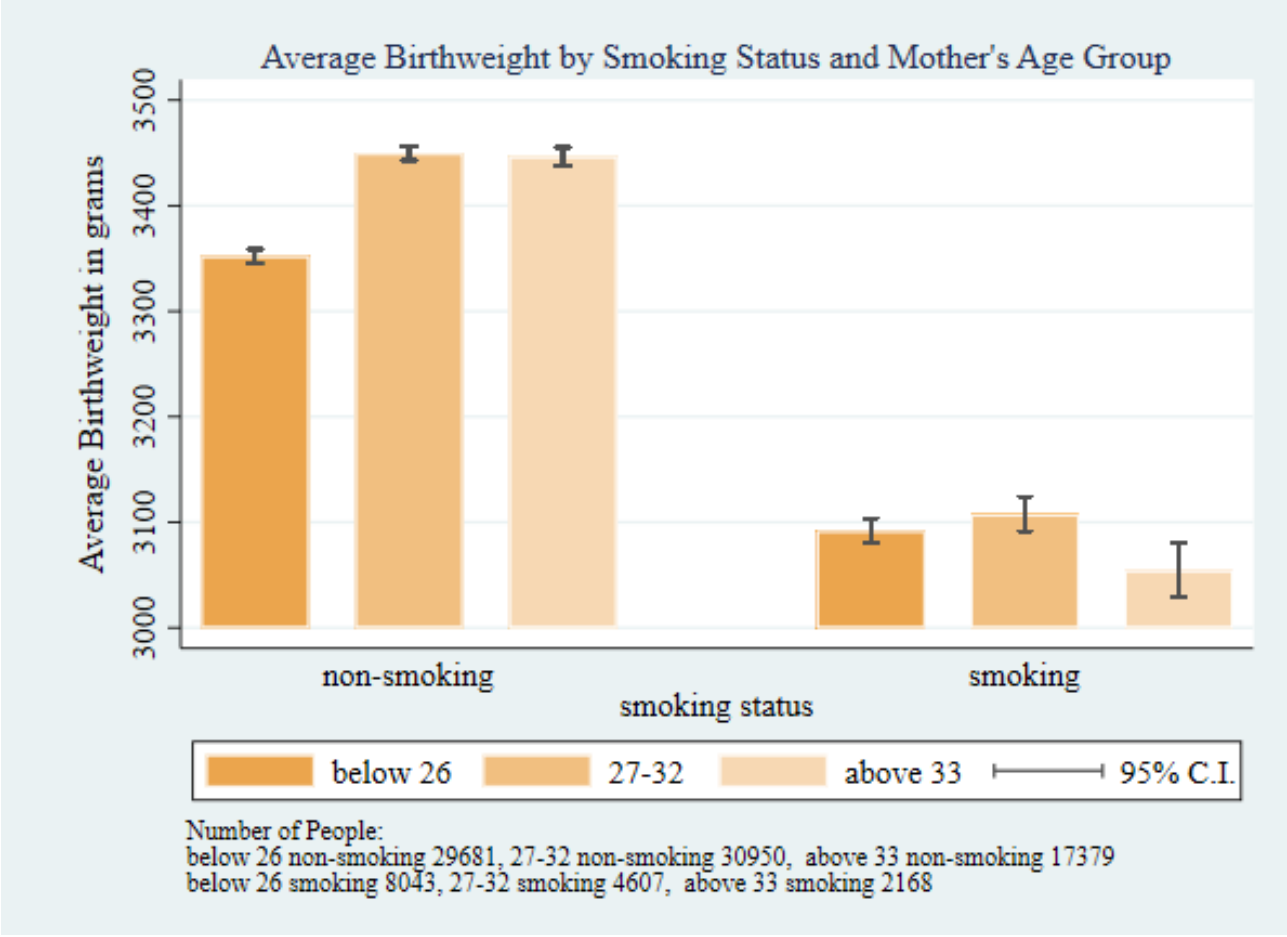
I counted observations

```
. forvalues i = 1/6 {
  2. if inrange(`i´, 1, 3) == 1 {
  3.     count if tobacco == 0 & dmage_factor == `i´
  4. }
  5. else if inrange(`i´, 4, 6) == 1 {
  6.     count if tobacco == 1 & dmage_factor == `i´ - 3
  7. }
  8.
. local `i´N = r(N)
  9.
. }
  29,681
  30,950
  17,379
  8,043
  4,607
  2,168
```

I plotted the bar chart

```
. gen tobacco_age = dmage_factor if tobacco == 0
(14,818 missing values generated)
. qui replace tobacco_age = dmage_factor + 4 if tobacco == 1
. twoway (bar avg tobacco_age if dmage_factor == 1, fcolor(dkorange) ///
>     fintensity(inten70) lcolor(white) barw(0.7)) ///
>     (bar avg tobacco_age if dmage_factor == 2, fcolor(dkorange) ///
>     fintensity(inten50) lcolor(white) barw(0.7)) ///
>     (bar avg tobacco_age if dmage_factor == 3, fcolor(dkorange) ///
>     fintensity(inten30) lcolor(white) barw(0.7)) ///
>     (rcap ci_low ci_high tobacco_age, lcolor(gs5)), ///
>     legend(row(1) order(1 "below 26" 2 "27-32" 3 "above 33" 4 "95% C.I.")) ///
>     xlabel(2 "non-smoking" 6 "smoking", noticks) xtitle("smoking status") ///
>     ytitle("Average Birthweight in grams", ///
>     margin(medium) size(medium)) ///
>     title("Average Birthweight by Smoking Status and Mother´s Age Group", size(medium)) ///
>     note("Number of People: " ///
>         "below 26 non-smoking `1N´, 27-32 non-smoking `2N´,  above 33 non-smoking `3N´" ///
>         "below 26 smoking `4N´, 27-32 smoking `5N´,  above 33 smoking `6N´" )
.
. graph export "`output´/birthweight_by_age.png", replace
(file D:/OneDrive - The University of Chicago/2021 Fall/job searching/Urban Lab/code_sample/output/birthw
> eight_by_age.png written in PNG format)
. restore
```

Figure 1: This is a caption



2. Estimating Propensity Score

2.1 Creating Propensity Score

To better control selection on observables, I used propensity score matching to find the treatment effect. I used logit specification to generate propensity score.

```
. qui logit tobacco `predetermined´,vce(r)
. qui predict pscore_1
```

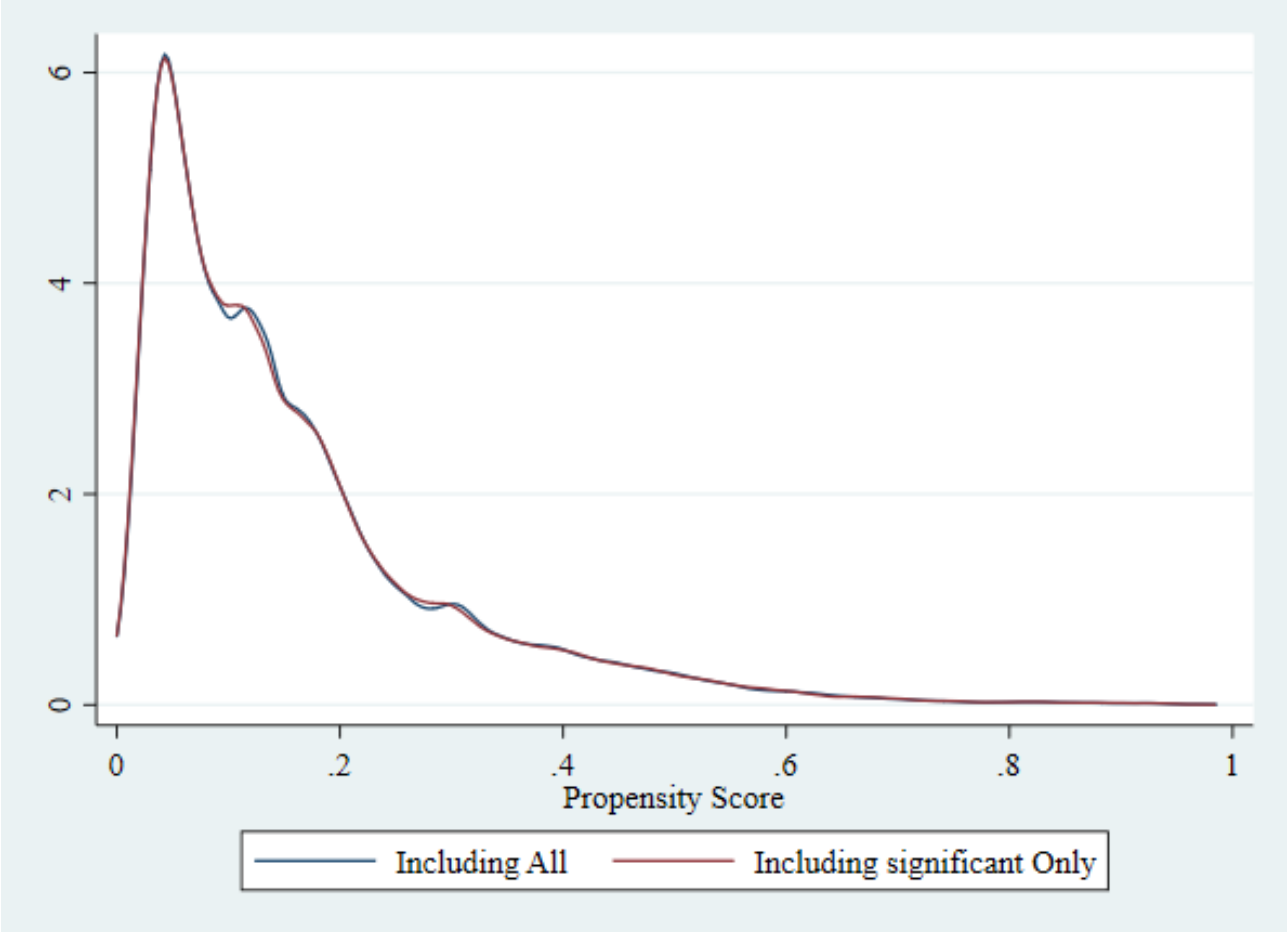
I only included predetermined covariates who were significant in the last logit regression

```
. local predetermined_sig stresfip rectype adequacy dmage mrace3 dmeduc dmar dfage dfeduc orfath ormoth n
> previst alcohol drink5 preterm pre4000 phyper isllb10 pldel3 dlivord dtotord totord9
. qui logit tobacco `predetermined_sig´,vce(r)
. qui predict pscore_2
```

To better compare the above two propensity scores, I calculated the correlation between those 2 propensity scores and simulated their density function by kdensity command

```
. corr(pscore_1 pscore_2)
(obs=92,828)
             | pscore_1 pscore_2
-------------+------------------
    pscore_1 |   1.0000
    pscore_2 |   0.9929   1.0000
. twoway (kdensity pscore_1) || (kdensity pscore_2),  xtitle("Propensity Score") legend(label(1 "Includin
> g All") label(2 "Including significant Only"))
. graph export "`output´/corr.png", replace
(file D:/OneDrive - The University of Chicago/2021 Fall/job searching/Urban Lab/code_sample/output/corr.p
> ng written in PNG format)
```

Figure 2: Correlation of Two Different Propensity Scores



2.2 Checking Common Support

Generate common support

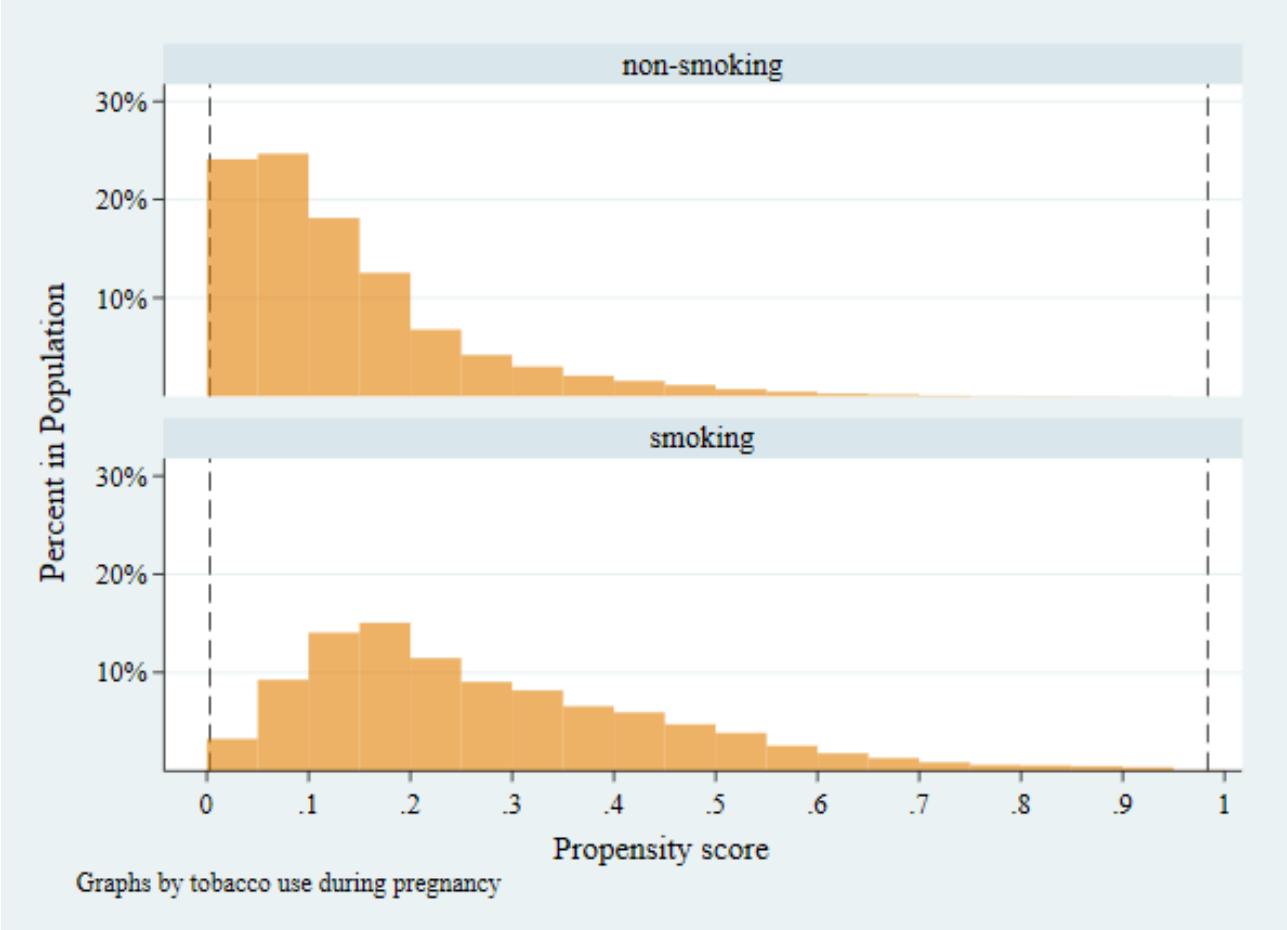
```
. cap drop common_support
. gen common_support = 1
. forvalues i = 1/2 {
  2. su pscore_2 if tobacco == `i´ - 1
  3. qui replace common_support = 0 if inrange(pscore_2, r(min), r(max)) == 0
  4. }
    Variable |        Obs        Mean    Std. Dev.       Min        Max
```

```
    pscore_2 |     78,010    .1367458    .1185099    .0000374    .9842785
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    pscore_2 |     14,818    .2800961    .1792051    .0032281    .9833701
```

histogram of propensity frequency

```
. su pscore_2 if common_support == 1
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    pscore_2 |     92,517    .1601475     .140217    .0032281    .9833701
. local rhs = r(max)
. local lhs = r(min)
.
. histogram pscore_2, xline(`rhs´, lcolor(black) lwidth(thin) lpattern(dash)) ///
>       xline(`lhs´, lcolor(black) lwidth(thin) lpattern(dash)) ///
>       xlabel(0(0.1)1) ///
>       ylabel(10 "10%" 20 "20%" 30 "30%", angle(0)) ///
>       percent width(0.05) ///
>       by (tobacco, row(2)) ///
>       fcolor(dkorange%60) lcolor(white%0) ///
>       xtitle("Propensity score", size(medsmall)) ///
>       ytitle("Percent in Population", size(medsamll))
(note:  named style medsamll not found in class gsize, default attributes used)
(note:  named style medsamll not found in class gsize, default attributes used)
.
. graph export "`output´/histogram.png", replace
(file D:/OneDrive - The University of Chicago/2021 Fall/job searching/Urban Lab/code_sample/output/histog
> ram.png written in PNG format)
.
. drop if common_support == 0
(311 observations deleted)
```

Figure 3: Distribution of Propensity Score in Population

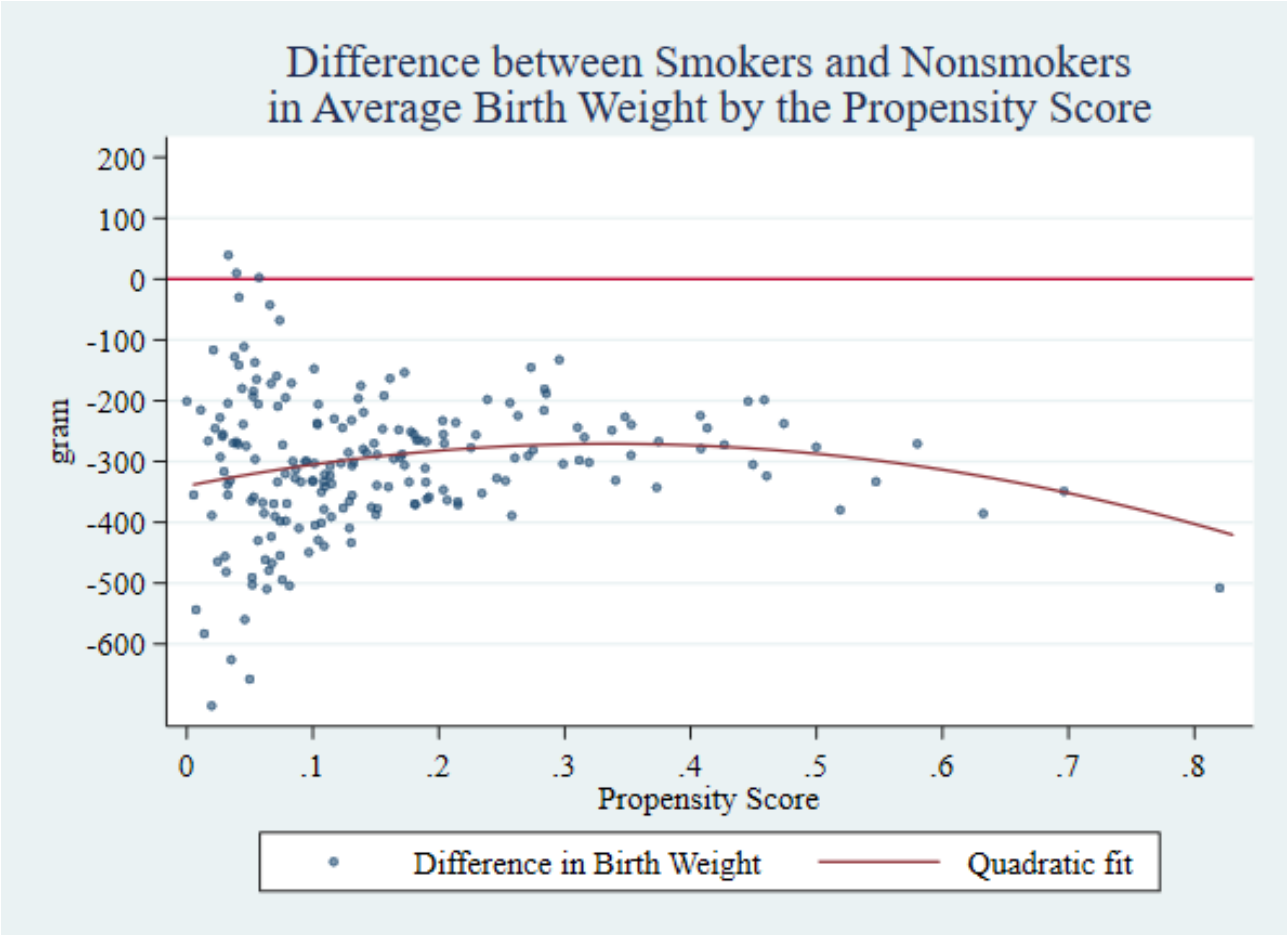


3. Analysis with Propensity Score

3.1 Plotting the Difference of Birth Weight by Propensity Score

Generate infant birth weight difference graph by propensity score

```
. egen rank = rank(pscore_2), unique
. egen group = cut(rank), group(200)
. gen smk_weight =.
(92,517 missing values generated)
. gen non_smk_weight =.
(92,517 missing values generated)
.
. forvalues i = 1/200 {
  2. local j = `i´ - 1
  3. qui count if group == `j´ & tobacco == 0
  4. qui sum dbrwt if group == `j´ & tobacco == 0
  5. qui replace smk_weight = r(sum) / r(N) if group == `j´
  6.
. qui count if group == `j´ & tobacco == 1
  7. qui sum dbrwt if group == `j´ & tobacco == 1
  8. qui replace non_smk_weight = r(sum) / r(N) if group == `j´
  9. }
.
. sort group
. gen weighted_diff = non_smk_weight - smk_weight
.
. preserve
. collapse(mean) pscore_2 weighted_diff, by(group)
. twoway scatter weighted_diff pscore_2 , msize(vsmall) mcolor(%50) connect(i) jitter(5) xlabel(0(0.1)0.8
> ) xtitle("Propensity Score") ytitle("gram") ///
>             ylabel(-600(100)200, angle(0)) yline(0) legend(order(1 "Difference in Birth Weight" 2 " Qua
> dratic fit ") rows(1)) ///
>             title("Difference between Smokers and Nonsmokers" "in Average Birth Weight by the Propensit
> y Score") || qfit weighted_diff pscore_2
.
. graph export "`output´/weight_diff.png", replace
(file D:/OneDrive - The University of Chicago/2021 Fall/job searching/Urban Lab/code_sample/output/weight
> _diff.png written in PNG format)
. restore
```



3.2 Three Different Analysis with Propensity Score

First, I simply included ps_score as covariates.

```
. reg dbrwt tobacco pscore_2
```

Source	SS	df	MS	Number of obs	=	92,517
				F(2, 92514)	=	2009.02
Model	1.3342e+09	2	667077577	Prob > F	=	0.0000
Residual	3.0718e+10	92,514	332040.53	R-squared	=	0.0416
				Adj R-squared	=	0.0416
Total	3.2053e+10	92,516	346454.157	Root MSE	=	576.23

dbrwt	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tobacco	-292.5458	5.568575	-52.54	0.000	-303.4602	-281.6315
pscore_2	-193.3672	14.56555	-13.28	0.000	-221.9155	-164.8189
_cons	3437.746	2.875968	1195.34	0.000	3432.11	3443.383

Second, I used them to reweight the outcomes and estimated the average treatment effect.

```
. gen ate_weight     = (1/pscore_2)   if tobacco==1
(77,699 missing values generated)
. qui replace ate_weight = 1/(1-pscore_2) if tobacco==0
. reg dbrwt tobacco [pweight=ate_weight]
(sum of wgt is 185,455.493234515)
```

Linear regression

				Number of obs	=	92,517
				F(1, 92515)	=	917.77
				Prob > F	=	0.0000
				R-squared	=	0.0614
				Root MSE	=	578.55

dbrwt	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
tobacco	-295.9967	9.770555	-30.29	0.000	-315.1469	-276.8465
_cons	3406.471	2.198073	1549.75	0.000	3402.163	3410.779

I used them to reweight the outcomes and estimated the average treatment effect on treated.

```
. gen att_weight     = 1                          if tobacco==1
(77,699 missing values generated)
. qui replace att_weight = pscore_2/(1-pscore_2) if tobacco==0
. reg dbrwt tobacco [pweight=att_weight]
(sum of wgt is 30,053.5116870475)
Linear regression                               Number of obs     =     92,517
                                                F(1, 92515)       =    2042.49
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0597
                                                Root MSE          =     577.83
------------------------------------------------------------------------------
             |               Robust
       dbrwt |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     tobacco |  -291.3006   6.445571   -45.19   0.000    -303.9339   -278.6674
       _cons |    3382.34   4.519554   748.38   0.000     3373.482    3391.198
------------------------------------------------------------------------------
```

Third, I made blockings on propensity score. I divided the data into 100 approximately equally spaced bins based on the estimated propensity score (pscore_2). Then, I calculated the average treatment effect(b_tobacco) of each bin and calculate the total treatment effect by weighting the b_tobacco of each bin based on the size of each bin.

```
. sum pscore_2
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    pscore_2 |     92,517    .1601475     .140217   .0032281   .9833701
. local bandwidth = (r(max)-r(min))/100
. gen bin_num =.
(92,517 missing values generated)
. forvalues i = 1/100{
  2. qui replace bin_num =`i´ if pscore_2 <=`i´*`bandwidth´ & bin_num ==.
  3. }
. qui replace bin_num = 100 if bin_num ==.
.
. sort bin_num
. gen weighted_tt =.
(92,517 missing values generated)
.
. local b_tobacco = 0
. foreach i of num 1/100 {
  2. qui reg dbrwt tobacco if bin_num == `i´
  3. qui replace weighted_tt = _b[tobacco]*`e(N)´/_N if bin_num == `i´
  4. local b_tobacco = `b_tobacco´+ _b[tobacco]*`e(N)´/_N
  5. }
.
. di `b_tobacco´
-302.29976
```

As you can see from the above, the effect of smoking on birth weight is -302.29976 gram.

3.3 Analysis of Low Birth Weight

I redo the last part using an indicator for low weight birth (less than 2500 grams) as the outcome. I use different methods here. I calculate the size of each bin first.

```
. gen low_bw = dbrwt < 2500
. sum pscore_2
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    pscore_2 |     92,517    .1601475     .140217   .0032281   .9833701
. gen psc_bins = autocode(pscore_2, 100, r(min), r(max))
. egen bin_size = sum(1), by(psc_bins)
```

Then, for each bin, I calculated the mean value of low weight birth for smoking and non-smoking group respectively.

```
. egen mean_low_1 = mean(low) if tobacco == 1, by(psc_bins)
```

```
(77699 missing values generated)
. egen mean_low_0 = mean(low) if tobacco == 0, by(psc_bins)
(14818 missing values generated)
```

After that, for each bin, to calculate the average outcome of smoking and non-smoking group, I divided mean_low_1 and mean_low_0 by the size of bin. Finally, I got the average ate_low by weighting the smoking effect of each bin based on their size.

```
. preserve
. collapse(mean) mean_low_0 mean_low_1 bin_size, by(psc_bins)
. egen N = sum(bin_size)
. egen ate_low = sum((mean_low_1 - mean_low_0)*(bin_size/N))
. di ate_low
.05866295
. restore
```

As you can see from the above, the effect of smoking on low birth weight is 0.058.