

Basic and matching theory:

1. Using properties of the conditional expectation function, show that $E(Y|X)$ is the minimum mean squared error (MMSE) predictor for $Y|X$.
2. Discuss the evaluation problem and the interpretation of selection bias using the potential outcomes model
3. Discuss the identifying assumptions needed to estimate the ATT and ATE using OLS or matching. How does assuming common overlap help?
4. If we think of matching as a 2-step approach, which step is more important in reducing bias and why? Provide the intuition.
5. Angrist (Ecma 1998) shows that the ATT estimated by OLS and nonparametric matching can be expressed as weighted averages of x -cell specific treatment effects, δ_x . Derive and compare these weights. Discuss the differences.
6. Propensity score theorem: Show that, under the Conditional Independence Assumption (CIA), $Y_0, Y_1 \perp D | p(X)$, where $p(X) = p(D = 1|X)$

Matching empirics:

This part of the problem set asks you to work with the datasets used in the LaLonde (LL, 1986) and Dehejia-Wahba (DW, 2002) papers. We will replicate some of the tables/figures in those papers as well as do additional analyses. Use the NSW, NSW-DW, CPS-1 and PSID-1 files for this exercise.

Note: (1) the files only contain data on the male participants and controls from the NSW experiment.

(2) In the past, **students have not been able to replicate the numbers in DW2002**, but they get somewhat close and replicate the qualitative patterns. So please do not worry about getting the numbers exactly right.

1. Prepare a table of descriptive statistics on key X variables and earnings for treated and comparison groups using each of these data files. The different variables should be in the rows, and the different data files span the columns. Briefly comment on the differences and what concerns you have about using CPS/PSID to replicate the experimental estimates.
2. Replicate results in rows 1 (experiment controls), 2 (PSID-1), and 5 (CPS-SSA-1) from Table 5 of LL. Ignore column 10. Use robust standard errors (regardless of what the paper uses). In this step you will not use the NSW-DW file.

Now, we will implement the propensity score matching (PSM) method following DW2002. Use the NSW-DW sample to obtain the treatment group observations, and CPS-1 and PSID-1 files as the non-experimental control groups.

3. Before diving into the empirics, briefly discuss whether PSM can be helpful in this setting. In class we discussed HIT1997 and their suggestions for when matching can perform well. Are those conditions satisfied by the data available here?
4. Estimate the propensity score separately for CPS and PSID samples as in the DW2002 paper. Discuss briefly the steps you took and the final specification you settled on. (It's ok if it is slightly

- different than what they had in DW. They mention their specification in the notes to Table 2). Present the mean values of the Xs for the treated and matched comparison groups.
5. Construct the equivalent versions of figures 1 and 2 from DW. A simple way to think about overlap is the % of treated units that fall in the overlap range of p-scores between the treated and comparison groups. Which dataset (CPS or PSID) appears to have more overlap with the treatment group from the experiment?
 6. Replicate Tables 2 and 3 from DW (you won't get them perfectly). No need to present the columns on means of X variables. Obtain standard errors using bootstrap (100 reps is sufficient).
 7. Comment on the performance of matching with replacement versus matching without replacement in this empirical application. How does it differ between the CPS and PSID samples?
 8. Estimate simple OLS estimates of the training effect using the matched samples obtained with and without replacement. How do they compare with the corresponding PSM estimates?

Next, we will do some additional analysis, beyond what was in the DW paper. Specifically, instead of using the NSW-DW sample, go back to the full NSW sample from the LL paper.

9. Discuss the main differences you observe between the NSW and NSW-DW files. Feel free to review the Smith and Todd (*J Metrics*, 2005) article if you would like more background on the differences.
10. Re-estimate propensity score for the full NSW sample along with the CPS-1 file. Did you have to change the specification to get a valid propensity score?
11. Create an equivalent version of Table 2 from DW, but with the full NSW sample. No need to present mean X values.
12. Does the performance of PSM drop using the full NSW relative to the NSW-DW file?

Basic and Matching Theory

- Using properties of the conditional expectation function, show that $E(Y|X)$ is the minimum mean squared error (MMSE) predictor for $Y|X$.

$$\begin{aligned}
 \text{MSE} &= E[(Y - m(x))^2] \quad \text{for any function of } m(x) \\
 &= E[((Y - E[Y|X]) + (E[Y|X] - m(x)))^2] \\
 &= E[\underbrace{(Y - E[Y|X])^2}_{\textcircled{L}} + \underbrace{2(Y - E[Y|X])(E[Y|X] - m(x))}_{\textcircled{R}} \\
 &\quad + \underbrace{(E[Y|X] - m(x))^2}_{\textcircled{M}}]
 \end{aligned}$$

\textcircled{L} is constant

\textcircled{R} is minimized by $m(x) = E[Y|X]$

$E[\textcircled{M}] = 0$ as I will now show:

$$\begin{aligned}
 E[\textcircled{M}] &= E[2(E[Y|X] - m(x))(Y - E[Y|X])] \\
 &= E[2(E[Y|X] - m(x)) \underbrace{E[Y - E[Y|X] | X]}_{= E[Y|X] - E[Y|X] = 0} \text{ by LIE}] \\
 &= E[0] = 0
 \end{aligned}$$

$\Rightarrow E[Y|X]$ minimizes the MSE

2. Discuss the evaluation problem and the interpretation of selection bias using the potential outcomes model

In the evaluation problem, we are typically interested in estimating the ATT (sometimes the ATE)

$$\tau_{ATT} = E[Y(1) - Y(0) | D=1]$$

yet in practice we can only observe:

$$E[Y|D=1] - E[Y|D=0]$$

$$= \underbrace{E[Y(1)|D=1] - E[Y(0)|D=1]}_{= \hat{\tau}_{ATT}} + \underbrace{E[Y(0)|D=1] - E[Y(0)|D=0]}_{\equiv "Selection\ bias"}$$

Thus with the potential outcomes model, we interpret selection bias as the mean difference in untreated potential outcomes between the treatment and control groups

Random assignment of D solves this problem by making selection bias 0

3. Discuss the identifying assumptions needed to estimate the ATT and ATE using OLS or matching. How does assuming common overlap help?

Key assumption is Conditional Independence Assumption (CIA)

$$\{Y(0), Y(1)\} \perp D | X \quad] \quad (1)$$

$$\Rightarrow F(Y_0 | X, D=1) = F(Y_0 | X, D=0) = F(Y_0 | X) \\ F(Y_1 | X, D=1) = F(Y_1 | X, D=0) = F(Y_1 | X) \quad] \quad (2)$$

$$\Rightarrow E[Y(0) | X, D=1] = E[Y(0) | X, D=0] = E[Y(0) | X] \\ E[Y(1) | X, D=1] = E[Y(1) | X, D=0] = E[Y(1) | X] \quad] \quad (3)$$

To estimate ATT or ATE we need some form of the CIA
 (1) is often assumed, but (2) or even (3) is sufficient

We also need an overlap assumption $0 < P(D=1 | X) < 1$ (common overlap)
 to estimate the ATE

For the ATT, $P(D=1 | X) < 1$ is sufficient

But common overlap is nice because it also allows ATT estimation

4. If we think of matching as a 2-step approach, which step is more important in reducing bias and why? Provide the intuition.

Step 1: Identify a comparison group which matches the treatment group along a number of pre-treatment observables

Step 2: Estimate treatment effects by comparing treated units and matched comparison units

The first step is more important, because, as we saw in Q2, selection bias results from differences in the composition of treatment and control groups.

Step 1 fixes this problem for observables, and hopefully unobservables as well.

Step 2 just amounts to selecting an estimator to calculate a difference; Step 1 ensures you are taking differences between the correct groups

5. Angrist (Ecma 1998) shows that the ATT estimated by OLS and nonparametric matching can be expressed as weighted averages of x -cell specific treatment effects, δ_x . Derive and compare these weights. Discuss the differences.

Matching weights:

$$\begin{aligned}
 \text{ATT} &= E[Y(1) - Y(0) | D=1] = E[Y(1) | D=1] - E[Y(0) | D=0] \quad \text{by CIA} \\
 &= E\left[E[Y(1) | X, D=1] - E[Y(0) | X, D=0] \mid D=1 \right] \quad \text{by LIE} \\
 &= E\left[\underbrace{E[Y | X, D=1]}_{\equiv \delta_x} - E[Y | X, D=0] \mid D=1 \right] \\
 &= \sum_x \delta_x P(X=x | D=1) = \frac{\sum_x \delta_x P(D=1 | X=x) P(X=x)}{\sum_x P(D=1 | X=x) P(X=x)} \quad \text{by Bayes Rule}
 \end{aligned}$$

OLS weights:

$$Y_i = \alpha + \delta_R D_i + \beta X_i + \varepsilon_i \quad (\text{I will drop } i \text{ subscripts as before for convenience})$$

$$\begin{aligned}
 \text{Define } \tilde{D}_i &= D_i - \gamma X_i \\
 &= D_i - E[D_i | X_i]
 \end{aligned}$$

$$\Rightarrow \delta_R = \frac{\text{cov}(Y, \tilde{D})}{\text{var}(\tilde{D})} = \frac{E[(D - E[D|X])Y]}{E[(D - E[D|X])^2]} = \frac{\textcircled{1}}{\textcircled{2}}$$

$$\begin{aligned}
 \textcircled{1} &= E[(D - E[D|X])(E[Y | D=0, X] + \delta_x D)] \\
 &= E[(D - E[D|X])(E[Y | D=0, X])] + E[(D - E[D|X])\delta_x D] \\
 &= 0 \quad \text{since } \text{cov}(\tilde{D}, E[Y | D=0, X]) = 0 \quad = E[(D - E[D|X])^2 \delta_x]
 \end{aligned}$$

$$\begin{aligned}
 \Rightarrow \delta_R &= \frac{E[(D - E[D|X])^2 \delta_x]}{E[(D - E[D|X])^2]} \\
 &= \frac{\sum_x \delta_x [P(D=1 | X=x) (1 - P(D=1 | X=x))] P(X=x)}{\sum_x [P(D=1 | X=x) (1 - P(D=1 | X=x))] P(X=x)}
 \end{aligned}$$

So cells with probability of treatment closer to 1/2 are weighted more heavily for OLS, but in matching, cells with higher probability of treatment are weighted more heavily.

6. Propensity score theorem: Show that, under the Conditional Independence Assumption (CIA),
 $Y_0, Y_1 \perp D \mid p(X)$, where $p(X) = p(D=1 \mid X)$

WTS: $Y(0), Y(1) \perp D \mid p(X) = p(D=1 \mid X)$

Holds if $p(D=1 \mid Y(j), p(X))$ doesn't depend on $Y(j)$ $j \in \{0, 1\}$

$$\begin{aligned} p(D=1 \mid Y(j), p(X)) &= E[D \mid Y(j), p(X)] \\ &= E[E[D \mid Y(j), p(X), X] \mid Y(j), p(X)] \quad \text{by LIE} \\ &= E[E[D \mid X] \mid Y(j), p(X)] \quad \text{by CIA} \\ &= E[p(X) \mid Y(j), p(X)] = p(X) \end{aligned}$$

which doesn't depend on $Y(j)$

HCMG 901 Problem Set 1

Matching Empirics Section

Andrés Rovira

February 15, 2023

Part 1

Prepare a table of descriptive statistics on key X variables and earnings for treated and comparison groups using each of these data files. The different variables should be in the rows, and the different data files span the columns. Briefly comment on the differences and what concerns you have about using CPS/PSID to replicate the experimental estimates.

See Table 1. The NSW treated and control groups are pretty similar to one another, as expected, and the Dehejia and Wahba sample has the advantage of data on real earnings in 1974. However, the CPS and PSID differ dramatically from the experimental samples on most variables. They feature higher age and years of education, a much lower share of black participants, a far higher fraction married and lower fraction with no high school degree. Perhaps most critically, they exhibit far higher earnings in all years, and their earnings do not display a downward trend from 1974 to 1975, unlike the experimental samples. I am concerned about using these samples to replicate the experimental results because on these observables they appear far from representative of the same type of people.

Table 1: Descriptive Statistics

	NSW Treated	NSW Control	NSW-DW Treated	NSW-DW Control	CPS	PSID
age	24.63 (6.69)	24.45 (6.59)	25.82 (7.16)	25.05 (7.06)	33.23 (11.05)	34.85 (10.44)
education	10.38 (1.82)	10.19 (1.62)	10.35 (2.01)	10.09 (1.61)	12.03 (2.87)	12.12 (3.08)
black	0.80 (0.40)	0.80 (0.40)	0.84 (0.36)	0.83 (0.38)	0.07 (0.26)	0.25 (0.43)
hispanic	0.09 (0.29)	0.11 (0.32)	0.06 (0.24)	0.11 (0.31)	0.07 (0.26)	0.03 (0.18)
married	0.17 (0.37)	0.16 (0.36)	0.19 (0.39)	0.15 (0.36)	0.71 (0.45)	0.87 (0.34)
nodegree	0.73 (0.44)	0.81 (0.39)	0.71 (0.46)	0.83 (0.37)	0.30 (0.46)	0.31 (0.46)
re74			2,095.57 (4,886.62)	2,107.03 (5,687.91)	14,016.80 (9,569.80)	19,428.75 (13,406.88)
re75	3,066.10 (4,874.89)	3,026.68 (5,201.25)	1,532.06 (3,219.25)	1,266.91 (3,102.98)	13,650.80 (9,270.40)	19,063.34 (13,596.96)
re78	5,976.35 (6,923.80)	5,090.05 (5,718.09)	6,349.14 (7,867.40)	4,554.80 (5,483.84)	14,846.66 (9,647.39)	21,553.92 (15,555.35)

Part 2

Replicate results in rows 1 (experiment controls), 2 (PSID-1), and 5 (CPS-SSA-1) from Table 5 of LL. Ignore column 10. Use robust standard errors (regardless of what the paper uses). In this step you will not use the NSW-DW file.

See Table 2.

Table 2: Replication of Lalonde (1986) Table 5

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Comparison Group	Earnings Growth	Unadjusted	Adjusted	Unadjusted	Adjusted	Without Age	With Age	Unadjusted	Adjusted
Controls	2,063 (375)	39 (379)	-21 (381)	886 (488)	798 (488)	847 (582)	852 (582)	879 (486)	802 (486)
PSID-1	2,491 (414)	-15,997 (393)	-7,624 (560)	-15,578 (508)	-8,067 (668)	420 (529)	-896 (570)	-2,380 (640)	-2,119 (680)
CPS-SSA-1	1,196 (106)	-10,585 (292)	-4,654 (412)	-8,870 (408)	-4,416 (464)	1,714 (486)	361 (496)	-1,543 (451)	-1,102 (471)

Part 3

Before diving into the empirics, briefly discuss whether PSM can be helpful in this setting. In class we discussed HIT1997 and their suggestions for when matching can perform well. Are those conditions satisfied by the data available here?

PSM can be helpful here because our issue is that the non-experimental control groups are unbalanced. PSM will adjust them to make them balanced on observables, hopefully improving balance on unobservables as well.

However, the criteria offered by HIT1997 temper our expectations somewhat, because they are not satisfactorily met by our data here. After matching, treated and matched units do not have the same distributions of observables. As seen in Table 3, some observables like age become close, but others like real earnings still differ considerably. Additionally, outcomes and characteristics were not measured in the same survey for the treated and matched units, so they may have been measured differently. Finally, the treated and matched groups were not located in the same economic environment, since the NSW program was isolated to 10 cities, while the CPS and PSID are national samples.

Part 4

Estimate the propensity score separately for CPS and PSID samples as in the DW2002 paper. Discuss briefly the steps you took and the final specification you settled on. (It's ok if it is slightly different than what they had in DW. They mention their specification in the notes to Table 2). Present the mean values of the Xs for the treated and matched comparison groups.

For my propensity score model I estimate a logit of treatment status on age, age squared, age cubed, years of education, years of education squared, marital status, high school degree status, a black indicator, a hispanic indicator, real earnings in 1974 and 1975, and a an interaction term between years of education and real earnings in 1974

After trying several different specifications, I opted for this one since it was closest to replicating the specification in DW2002 and, from examining the R^2 's, it explained a satisfactory share of the variation in treatment.

See Table 3 below for the mean values of Xs for treated and matched comparison groups.

Table 3: Group Means of X Variables

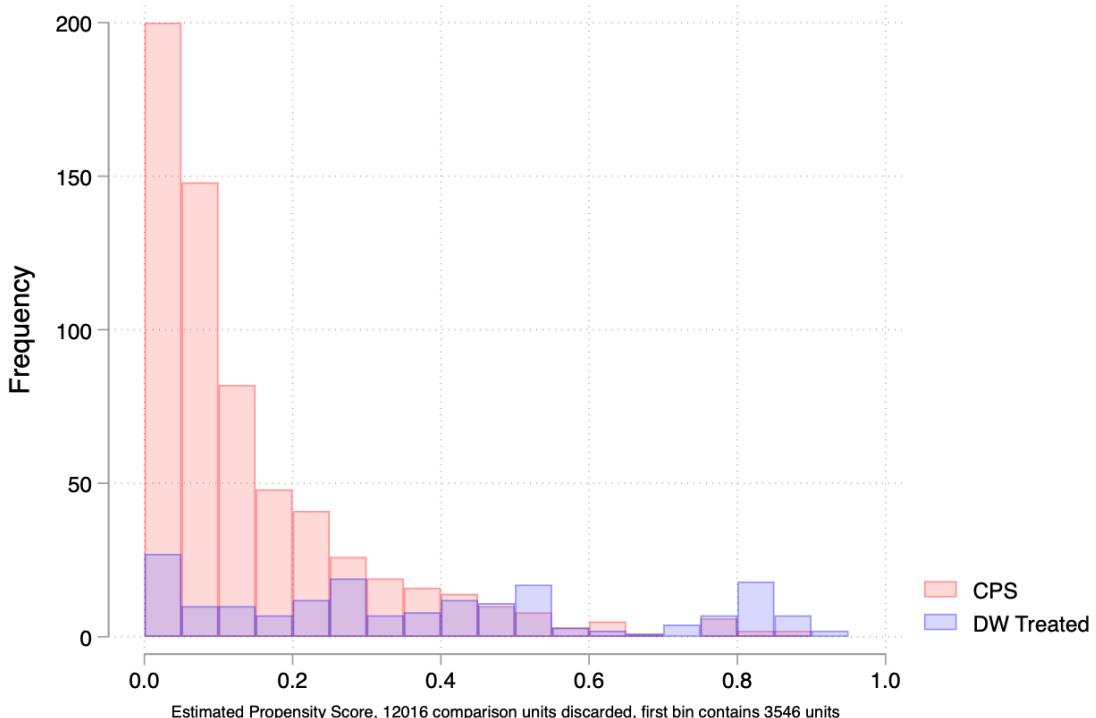
Variable	Treated NSW	Matched CPS	Matched PSID
age	25.82	25.80	27.81
education	10.35	10.51	10.58
nodegree	0.71	0.63	0.61
black	0.84	0.77	0.74
hispanic	0.06	0.08	0.06
married	0.19	0.24	0.37
re74	2095.57	2867.35	4294.90
re75	1532.06	2030.94	3005.81
re78	6349.14	5502.25	6636.40

Part 5

Construct the equivalent versions of figures 1 and 2 from DW. A simple way to think about overlap is the % of treated units that fall in the overlap range of p-scores between the treated and comparison groups. Which dataset (CPS or PSID) appears to have more overlap with the treatment group from the experiment?

See Figure 1 and Figure 2 below. CPS appears to have more overlap with the treatment group from the experiment than PSID.

Figure 1: Replication of Dehejia and Wahba (2002) Figure 1



Part 6

Replicate Tables 2 and 3 from DW (you won't get them perfectly). No need to present the columns on means of X variables. Obtain standard errors using bootstrap (100 reps is sufficient).

See Table 4 and Table 5 below.

Note that with caliper matching, DW say that they match all controls within the caliper radius of a treated observation, but this then does not fit with how, when estimating treatment effects, they say to weight observations equally by the number of times they are matched. To do so would overrepresent treated observations with very low propensity scores, since most control observations have very low propensity scores. To avoid this, I only match one control within the caliper radius, rather than all available controls.

Figure 2: Replication of Dehejia and Wahba (2002) Figure 2

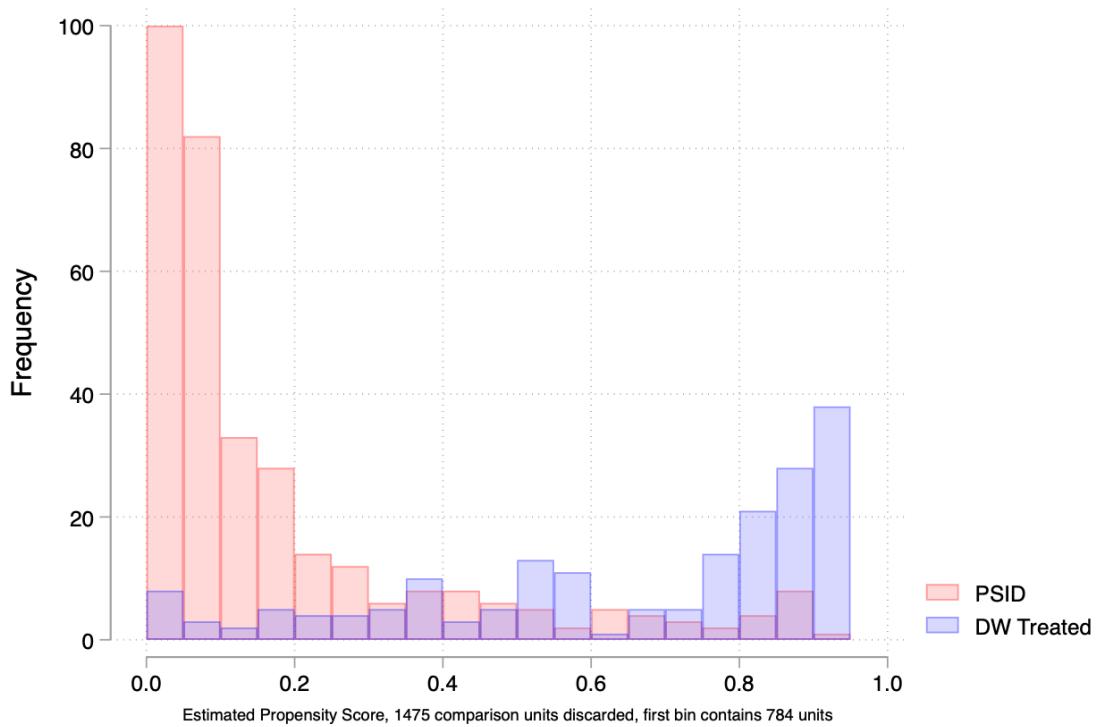


Table 4: Replication of Dehejia and Wahba Table 2

Control Sample	Obs	Mean Propen-sity Score	Treatment Effect (Diff. in Means)	Regression Treatment Effect	Q8: Simple OLS
NSW	185	0.38 (0.02)	1,794 (613)	1,676 (722)	
Full CPS	15,992	0.01 (0.02)	-8,498 (592)	699 (679)	
W/o Replacement: Random	185	0.30 (0.03)	996 (673)	1,119 (671)	1,119 (719)
W/o Replacement: Low to High	185	0.30 (0.02)	1,223 (741)	1,333 (682)	1,333 (719)
W/o Replacement: High to Low	185	0.30 (0.03)	1,179 (760)	1,281 (670)	1,281 (718)
With Replacement: Nearest Neighbor	117	0.38 (0.03)	1,625 (678)	1,546 (699)	1,358 (818)
With Replacement: Caliper = 0.00001	39	0.20 (0.05)	1,970 (1,519)	1,736 (1,604)	1,852 (1,520)
With Replacement: Caliper = 0.00005	47	0.19 (0.04)	1,537 (1,365)	1,473 (1,081)	1,538 (1,319)
With Replacement: Caliper = 0.0001	57	0.20 (0.03)	1,949 (1,201)	1,888 (1,237)	1,824 (1,343)

Table 5: Replication of Dehejia and Wahba Table 3

Control Sample	Obs	Mean Propen- sity Score	Treatment Effect (Diff. in Means)	Regression Treat- ment Effect	Q8: Simple OLS
NSW	185	0.65	1,794 (641)	1,676 (643)	
Full PSID	2,490	0.03 (0.02)	-15,205 (669)	752 (738)	
W/o Replacement: Random	185	0.29 (0.03)	-53 (750)	1,045 (895)	1,045 (856)
W/o Replacement: Low to High	185	0.29 (0.03)	-53 (789)	1,045 (889)	1,045 (856)
W/o Replacement: High to Low	185	0.29 (0.03)	-53 (822)	1,045 (737)	1,045 (856)
With Replacement: Nearest Neighbor	63	0.65 (0.03)	879 (790)	1,173 (717)	-201 (1,137)
With Replacement: Caliper = 0.00001	9	0.59 (0.13)	751 (3,168)	400 (2,157)	222 (2,655)
With Replacement: Caliper = 0.00005	18	0.48 (0.11)	-524 (2,831)	-1,643 (1,938)	-2,760 (2,472)
With Replacement: Caliper = 0.0001	20	0.46 (0.09)	174 (2,222)	-793 (1,857)	-1,684 (2,355)
With Replacement: Caliper = 0.001	41	0.51 (0.06)	200 (1,405)	33 (1,147)	-728 (1,444)

Part 7

Comment on the performance of matching with replacement versus matching without replacement in this empirical application. How does it differ between the CPS and PSID samples?

Matching with replacement performs significantly better than matching without replacement with both the CPS and PSID. The resulting treatment effect estimates are much closer to the experimental estimates. However with the PSID sample, matching without replacement performs quite poorly, with negative point estimates, and with replacement performs better, but not even as good as any of the estimates with the CPS sample. The results with the PSID are generally quite poor.

Part 8

Estimate simple OLS estimates of the training effect using the matched samples obtained with and without replacement. How do they compare with the corresponding PSM estimates?

See the last columns in Tables 4 and 5.

By “simple OLS estimates” I assume you want the regression treatment effect with controls, without weighting observations by the number of times they are used as matched controls, and using robust SEs instead of bootstrapping. Mechanically, these estimates are identical when matching without replacement, since each control observation can only be used once. For matching with replacement, the OLS results are similar with the CPS sample, but generally much worse (lower) for the PSID sample.

Part 9

Discuss the main differences you observe between the NSW and NSW-DW files. Feel free to review the Smith and Todd (J Metrics, 2005) article if you would like more background on the differences.

The most critical difference I observe between the files is that NSW-DW has a variable for real earnings in 1974, whereas NSW does not. As a consequence, however, it also has much fewer observations. From Table 1, we can see how they differ on other covariates: age, education, black, hispanic, married, and nodegree are all fairly similar. But the real earnings variables have very different means. NSW-DW has much lower average real earnings in 1975 and somewhat different real earnings in 1978. It thus likely captures a substantively different sample of people.

Part 10

Re-estimate propensity score for the full NSW sample along with the CPS-1 file. Did you have to change the specification to get a valid propensity score?

Yes, due to the lack of a variable on real earnings in 1974, I had to drop all terms using that variable from my propensity score model.

Part 11

Create an equivalent version of Table 2 from DW, but with the full NSW sample. No need to present mean X values.

See Table 6 below.

Table 6: Replication of Dehejia and Wahba Table 2 with Full NSW Sample

Control Sample	Obs	Mean Propen- sity Score	Treatment Effect (Diff. in Means)	Regression Treat- ment Effect	Q8: Simple OLS
NSW	297	0.38	886 (492)	807 (482)	
Full CPS	15,992	0.01 (0.02)	-8,870 (373)	-993 (477)	
W/o Replacement: Random	297	0.30 (0.02)	-1,033 (614)	-700 (462)	-700 (544)
W/o Replacement: Low to High	297	0.30 (0.02)	-990 (573)	-657 (504)	-657 (544)
W/o Replacement: High to Low	297	0.30 (0.02)	-1,019 (564)	-695 (524)	-695 (543)
With Replacement: Nearest Neighbor	187	0.38 (0.02)	-78 (528)	-0 (524)	-476 (609)
With Replacement: Caliper = 0.00001	60	0.23 (0.05)	1,664 (1,393)	1,481 (1,360)	1,298 (1,343)
With Replacement: Caliper = 0.00005	82	0.21 (0.03)	595 (1,234)	647 (1,083)	626 (1,087)
With Replacement: Caliper = 0.0001	94	0.20 (0.03)	27 (1,169)	-19 (923)	16 (1,008)

Part 12

Does the performance of PSM drop using the full NSW relative to the NSW-DW file?

Yes, PSM performance drops substantially. Point estimates on treatment effects are often negative, which is the wrong direction entirely. The only variant that performs decently well is matching with replacement with the smallest caliper, indicating that the concern with the sample and PSM model broadly may be an insufficient propensity score overlap.

Code

```
*****
* AUTHORS: Andres Rovira (with Stephanie Grove, Xianya Zhou, Fabio Schanaider)
* CREATED: 2023-02-08
* PURPOSE: Solve coding portions of HCMG 901 Problem Set 1
*****  
  
global home "˜"  
if "$S_OS" == "Windows" global home `:env USERPROFILE'  
cd "$home/Dropbox (Penn)/Classes/2_health_applied_metrics/ps1/code"  
global ol "$home/Dropbox (Penn)/Apps/Overleaf/hcmg901_ps1"  
  
// ssc install texsave  
// ssc install blindschemes  
// ssc install psmatch2  
  
set scheme plotplainblind  
  
*****  
***# Append datasets together  
*****  
  
clear all  
gen dataset = ""  
foreach f in cps1 nsw nsw_dw psid1 {  
append using "../input/'f'.dta"  
replace dataset = "'f'" if missing(dataset)  
}  
save "../intermediate/combined.dta", replace  
  
*****  
***# Part 1  
*****  
  
use "../intermediate/combined.dta", clear  
  
 tostring treat, replace  
gen group = dataset + "_" + treat  
  
levelsof group, local(groups)  
local glist ""  
foreach g of local groups {  
local glist "'glist' 'g'"  
}  
local xvars age education black hispanic married nodegree re74 re75 re78  
  
tempname memhold  
tempfile results  
postfile `memhold' str20 varname sd_ind `glist' using `results'
```

```

foreach var of varlist `xvars' {
    local mean_postline "(0)"
    local sd_postline   "(1)"
    foreach g of local groups {
        quietly summarize `var' if group == "'g'"
        local mean_postline "'mean_postline' (`r(mean))"
        local sd_postline   "'sd_postline' (`r(sd))"
    }
    local mean_postline = subinstr("`mean_postline'", "()", "(.)", .)
    local sd_postline   = subinstr("`sd_postline'", "()", "(.)", .)
    post `memhold' ("`var'") `mean_postline'
    post `memhold' ("`var'") `sd_postline'
}
postclose `memhold'
use `results', clear

// Format latex table
foreach var of local glist {
    format `var' %13.2fc
    tostring `var', replace force usedisplayformat
    replace `var' = "" if `var' == "."
    replace `var' = "(" + `var' + ")" if sd_ind == 1 & !missing(`var')
}
replace varname = "" if sd_ind == 1
drop sd_ind
order varname nsw_1 nsw_0 nsw_dw_1 nsw_dw_0 cps1_0 psid1_0
label var varname ""
label var nsw_1      "NSW Treated"
label var nsw_0      "NSW Control"
label var nsw_dw_1   "NSW-DW Treated"
label var nsw_dw_0   "NSW-DW Control"
label var cps1_0     "CPS"
label var psid1_0    "PSID"
texsave * using "$ol/tabl1.tex", varlabels frag replace title("Descriptive Statistics") location("hbt")

*****
**# Part 2
*****


global adj_vars age age2 education nodegree black hispanic

tempname memhold
 tempfile results
postfile `memhold' str20 comp_grp sd_ind c1 c2 c3 c4 c5 c6 c7 c8 c9 using `results'
foreach d in nsw psid1 cps1 {

use "../intermediate/combined.dta", clear
keep if (dataset == "nsw" & treat == 1) | (dataset == "'d'" & treat == 0)
gen age2 = age^2
gen re_growth = re78 - re75

```

```

preserve
keep treat re75 re78
gen id = _n
reshape long re, i(id treat) j(year)
regress re i.year if treat == 0, robust
local c1b = _b[78.year]
local c1se = _se[78.year]
restore

regress re75 treat, robust
local c2b = _b[treat]
local c2se = _se[treat]
regress re75 treat $adj_vars, robust
local c3b = _b[treat]
local c3se = _se[treat]
regress re78 treat, robust
local c4b = _b[treat]
local c4se = _se[treat]
regress re78 treat $adj_vars, robust
local c5b = _b[treat]
local c5se = _se[treat]
regress re_growth treat, robust
local c6b = _b[treat]
local c6se = _se[treat]
regress re_growth treat age, robust
local c7b = _b[treat]
local c7se = _se[treat]
regress re_growth treat re75, robust
local c8b = _b[treat]
local c8se = _se[treat]
regress re_growth treat re75 $adj_vars, robust
local c9b = _b[treat]
local c9se = _se[treat]

post `memhold' ("`d'") (0) (`c1b') (`c2b') (`c3b') (`c4b') (`c5b') (`c6b') (`c7b') (`c8b') (`c9b')
post `memhold' ("`d'") (1) (`c1se') (`c2se') (`c3se') (`c4se') (`c5se') (`c6se') (`c7se') (`c8se') (`c9se')
}
postclose `memhold'
use `results', clear

// Format latex table
format %13.0fc c?
 tostring c?, replace force usedisplayformat
foreach var of varlist c? {
replace `var' = "(" + `var' + ")" if sd_ind == 1 & !missing(`var')
}
replace comp_grp = "" if sd_ind == 1
drop sd_ind
replace comp_grp = "PSID-1" if comp_grp == "psid1"
replace comp_grp = "CPS-SSA-1" if comp_grp == "cps1"

```

```

replace comp_grp = "Controls" if comp_grp == "nsw"
label var comp_grp "Comparison Group"
label var c1 "Earnings Growth"
label var c2 "Unadjusted"
label var c3 "Adjusted"
label var c4 "Unadjusted"
label var c5 "Adjusted"
label var c6 "Without Age"
label var c7 "With Age"
label var c8 "Unadjusted"
label var c9 "Adjusted"
texsave * using "$ol/tab2.tex", varlabels frag replace autonumber ///
title("Replication of Lalonde (1986) Table 5")

*****
***# Part 4
*****
use "../intermediate/combined.dta", clear
drop if dataset == "nsw"

gen age2 = age^2
gen age3 = age^3
gen education2 = education^2

gen ps_sample_cps = (dataset == "nsw_dw" & treat == 1) | (dataset == "cps1" & treat == 0)
gen ps_sample_psid = (dataset == "nsw_dw" & treat == 1) | (dataset == "psid1" & treat == 0)

global dw_psvars age age2 age3 education education2 married nodegree black hispanic re74 re75 c.education

logit treat $dw_psvars if ps_sample_cps == 1
predict ps_cps if ps_sample_cps == 1

logit treat $dw_psvars if ps_sample_psid == 1
predict ps_psid if ps_sample_psid == 1

save "../intermediate/combined_w_pscore.dta", replace

local covars age education nodegree black hispanic married re74 re75 re78
foreach var of varlist `covars' {
    summarize `var' if dataset == "cps1" [iweight=ps_cps]
    local mean_`var'_cps = string(r(mean), "%4.2f")
    dis "'mewn_`var'_cps'"
    summarize `var' if dataset == "psid1" [iweight=ps_psid]
    local mean_`var'_psid = string(r(mean), "%4.2f")
    summarize `var' if dataset == "nsw_dw" & treat == 1
    local mean_`var'_treat = string(r(mean), "%4.2f")
}

```

```

cap erase "$ol/tab4.tex"
file open fh using "$ol/tab4.tex", write replace
#delimit ;
file write fh
"\begin{table}![htbp]" _n
"\centering" _n
"\caption{\textbf{Group Means of X Variables}}" _n
"\resizebox{\textwidth}{!}{" _n
"\begin{tabular}{l|ccc}" _n
"\toprule" _n
"Variable & Treated NSW & Matched CPS & Matched PSID \\" _n
"\midrule" _n
"age & 'mean_age_treat' & 'mean_age_cps' & 'mean_age_psid' \\" _n
"education & 'mean_education_treat' & 'mean_education_cps' & 'mean_education_psid' \\" _n
"nodegree & 'mean_nodegree_treat' & 'mean_nodegree_cps' & 'mean_nodegree_psid' \\" _n
"black & 'mean_black_treat' & 'mean_black_cps' & 'mean_black_psid' \\" _n
"hispanic & 'mean_hispanic_treat' & 'mean_hispanic_cps' & 'mean_hispanic_psid' \\" _n
"married & 'mean_married_treat' & 'mean_married_cps' & 'mean_married_psid' \\" _n
"re74 & 'mean_re74_treat' & 'mean_re74_cps' & 'mean_re74_psid' \\" _n
"re75 & 'mean_re75_treat' & 'mean_re75_cps' & 'mean_re75_psid' \\" _n
"re78 & 'mean_re78_treat' & 'mean_re78_cps' & 'mean_re78_psid' \\" _n
"\bottomrule" _n
"\end{tabular}" _n
"}" _n
"\label{tab:addlabel}" _n
"\end{table}" _n
;
#delimit cr
file close fh

*****
*** Part 5
*****



// CPS version
use "../intermediate/combined_w_pscore.dta", clear
keep if ps_sample_cps == 1
summarize ps_cps if treat == 1
drop if ps_cps < `r(min)'
local n_discarded = `r(N_drop)'

sort treat ps_cps
gen first_ctrl_bin = (treat == 0) & (ps_cps <= 0.05)
gsort first_ctrl_bin -ps_cps
by first_ctrl_bin: gen num = _n
count if first_ctrl_bin == 1
local n_first_bin = `r(N)'
drop if first_ctrl_bin == 1 & num > 200

twoway ///

```

```

(histogram ps_cps if treat == 0, start(0) width(0.05) color(red%30) frequency) ///
(histogram ps_cps if treat == 1, start(0) width(0.05) color(blue%30) frequency), ///
legend(order(1 "CPS" 2 "DW Treated")) xlabel(, format(%2.1fc)) yscale(range(0 200)) ///
xtitle("Estimated Propensity Score, 'n_discarded' comparison units discarded, first bin contains 'n_f")
graph export "$ol/fig5_cps.png", replace

// PSID version
use "../intermediate/combined_w_pscore.dta", clear
keep if ps_sample_psid == 1
summarize ps_psid if treat == 1
drop if ps_psid < `r(min)'
local n_discarded = `r(N_drop)'

sort treat ps_psid
gen first_ctrl_bin = (treat == 0) & (ps_psid <= 0.05)
gsort first_ctrl_bin -ps_psid
by first_ctrl_bin: gen num = _n
count if first_ctrl_bin == 1
local n_first_bin = `r(N)'
drop if first_ctrl_bin == 1 & num > 100

twoway ///
(histogram ps_psid if treat == 0, start(0) width(0.05) color(red%30) frequency) ///
(histogram ps_psid if treat == 1, start(0) width(0.05) color(blue%30) frequency), ///
legend(order(1 "PSID" 2 "DW Treated")) xlabel(, format(%2.1fc)) yscale(range(0 100)) ///
xtitle("Estimated Propensity Score, 'n_discarded' comparison units discarded, first bin contains 'n_f")
graph export "$ol/fig5_psid.png", replace

*****
*** Parts 6 and 8
*****


global dw_regvars age education married nodegree black hispanic re74 re75

local cps_full_row "Full CPS"
local psid_full_row "Full PSID"
local cps_title "Replication of Dehejia and Wahba Table 2"
local psid_title "Replication of Dehejia and Wahba Table 3"

foreach d in cps psid {

tempname memhold
tempfile results
postfile `memhold' str50 ctrl_samp sd_ind obs mean_ps te_dim reg_te ols_te using `results'

// Row 1
use "../intermediate/combined_w_pscore.dta", clear
count if dataset == "nsw_dw" & treat == 1
local obs = `r(N)'

```

```

summarize ps_`d' if dataset == "nsw_dw" & treat == 1
local mean_ps = `r(mean)'
bootstrap, reps(100): regress re78 treat if dataset == "nsw_dw"
local te_dim = _b[treat]
local te_dim_se = _se[treat]
bootstrap, reps(100): regress re78 treat $dw_regvars if dataset == "nsw_dw"
local reg_te = _b[treat]
local reg_te_se = _se[treat]

post 'memhold' ("NSW") (0) ('obs') ('mean_ps') ('te_dim') ('reg_te') (.)
post 'memhold' ("NSW") (1) (.) (.) ('te_dim_se') ('reg_te_se') (.)

// Row 2
use "../intermediate/combined_w_pscore.dta", clear
count if dataset == "'d'1"
local obs = `r(N)'
summarize ps_`d' if dataset == "'d'1"
local mean_ps = `r(mean)'
bootstrap, reps(100): regress ps_`d' treat if dataset == "'d'1" | (dataset == "nsw_dw" & treat == 1)
local mean_ps_se = _se[treat]
bootstrap, reps(100): regress re78 treat if dataset == "'d'1" | (dataset == "nsw_dw" & treat == 1)
local te_dim = _b[treat]
local te_dim_se = _se[treat]
bootstrap, reps(100): regress re78 treat $dw_regvars if dataset == "'d'1" | (dataset == "nsw_dw" & tr)
local reg_te = _b[treat]
local reg_te_se = _se[treat]

post 'memhold' ("'d'_full_row") (0) ('obs') ('mean_ps') ('te_dim') ('reg_te') (.)
post 'memhold' ("'d'_full_row") (1) (.) ('mean_ps_se') ('te_dim_se') ('reg_te_se') (.)

// Remaining rows
forvalues row = 3/10 {

use "../intermediate/combined_w_pscore.dta", clear
keep if ps_sample_`d' == 1

if `row' == 3 {
local samp_name "W/o Replacement: Random"
tempvar sortorder
gen 'sortorder' = runiform()
sort 'sortorder'
psmatch2 treat, outcome(re78) pscore(ps_`d') neighbor(1) noreplacement
}
else if `row' == 4 {
local samp_name "W/o Replacement: Low to High"
psmatch2 treat, outcome(re78) pscore(ps_`d') neighbor(1) noreplacement
}
else if `row' == 5 {
local samp_name "W/o Replacement: High to Low"
psmatch2 treat, outcome(re78) pscore(ps_`d') neighbor(1) noreplacement descending
}
}

```

```

else if `row' == 6 {
    local samp_name "With Replacement: Nearest Neighbor"
    psmatch2 treat, outcome(re78) pscore(ps_`d') neighbor(1)
}
else if `row' == 7 {
    local samp_name "With Replacement: Caliper = 0.00001"
    psmatch2 treat, outcome(re78) pscore(ps_`d') caliper(0.00001)
}
else if `row' == 8 {
    local samp_name "With Replacement: Caliper = 0.00005"
    psmatch2 treat, outcome(re78) pscore(ps_`d') caliper(0.00005)
}
else if `row' == 9 {
    local samp_name "With Replacement: Caliper = 0.0001"
    psmatch2 treat, outcome(re78) pscore(ps_`d') caliper(0.0001)
}
else if `row' == 10 {
    local samp_name "With Replacement: Caliper = 0.001"
    if "'d'" == "cps" continue
    psmatch2 treat, outcome(re78) pscore(ps_`d') caliper(0.001)
}

keep if !missing(_weight)

count if treat == 0
local obs = `r(N)'
regress re78 treat $dw_regvars, robust
local ols_te = _b[treat]
local ols_te_se = _se[treat]

expand _weight, generate(num) // applies weighting to subsequent estimations

summarize ps_`d' if treat == 0
local mean_ps = `r(mean)'
bootstrap, reps(100): regress ps_`d' treat
local mean_ps_se = _se[treat]
bootstrap, reps(100): regress re78 treat
local te_dim = _b[treat]
local te_dim_se = _se[treat]
bootstrap, reps(100): regress re78 treat $dw_regvars
local reg_te = _b[treat]
local reg_te_se = _se[treat]

post `memhold' ("`samp_name'") (0) (`obs') (`mean_ps') (`te_dim') (`reg_te') (`ols_te')
post `memhold' ("`samp_name'") (1) (.) (`mean_ps_se') (`te_dim_se') (`reg_te_se') (`ols_te_se')

}

postclose `memhold'
use `results', clear

```

```

// Format latex table
format %13.0fc obs te_dim reg_te ols_te
format %13.2fc mean_ps
foreach var of varlist obs mean_ps te_dim reg_te ols_te {
    tostring `var', replace force usedisplayformat
    replace `var' = "" if `var' == "."
    replace `var' = "(" + `var' + ")" if sd_ind == 1 & !missing(`var')
}
replace ctrl_samp = "" if sd_ind == 1
drop sd_ind
label var ctrl_samp "Control Sample"
label var obs "Obs"
label var mean_ps "Mean Propensity Score"
label var te_dim "Treatment Effect (Diff. in Means)"
label var reg_te "Regression Treatment Effect"
label var ols_te "Q8: Simple OLS"
texsave * using "$ol/tab6_`d'.tex", varlabels frag replace title("`d'_title")
}

*****
*** Part 10
*****

use "../intermediate/combined.dta", clear
keep if inlist(dataset, "nsw", "cps1")

gen age2 = age^2
gen age3 = age^3
gen education2 = education^2

gen ps_sample_cps = (dataset == "nsw" & treat == 1) | (dataset == "cps1" & treat == 0)

global nondw_psvars age age2 age3 education education2 married nodegree black hispanic re75

logit treat $nondw_psvars if ps_sample_cps == 1
predict ps_cps if ps_sample_cps == 1

save "../intermediate/combined_w_pscore_nondw.dta", replace

*****
*** Part 11
*****


global nondw_regvars age education married nodegree black hispanic re75

local cps_full_row "Full CPS"
local psid_full_row "Full PSID"
local cps_title "Replication of Dehejia and Wahba Table 2 with Full NSW Sample"

```

```

local psid_title "Replication of Dehejia and Wahba Table 3 with Full NSW Sample"

foreach d in cps {

tempname memhold
tempfile results
postfile `memhold' str50 ctrl_samp sd_ind obs mean_ps te_dim reg_te ols_te using `results'

// Row 1
use "../intermediate/combined_w_pscore_nondw.dta", clear
count if dataset == "nsw" & treat == 1
local obs = `r(N)'
summarize ps_`d' if dataset == "nsw" & treat == 1
local mean_ps = `r(mean)'
bootstrap, reps(100): regress re78 treat if dataset == "nsw"
local te_dim = _b[treat]
local te_dim_se = _se[treat]
bootstrap, reps(100): regress re78 treat $nondw_regvars if dataset == "nsw"
local reg_te = _b[treat]
local reg_te_se = _se[treat]

post `memhold' ("NSW") (0) ('obs') ('mean_ps') ('te_dim') ('reg_te') (.)
post `memhold' ("NSW") (1) (.) (.) ('te_dim_se') ('reg_te_se') (.)

// Row 2
use "../intermediate/combined_w_pscore_nondw.dta", clear
count if dataset == "d'1"
local obs = `r(N)'
summarize ps_`d' if dataset == "d'1"
local mean_ps = `r(mean)'
bootstrap, reps(100): regress ps_`d' treat if dataset == "d'1" | (dataset == "nsw" & treat == 1)
local mean_ps_se = _se[treat]
bootstrap, reps(100): regress re78 treat if dataset == "d'1" | (dataset == "nsw" & treat == 1)
local te_dim = _b[treat]
local te_dim_se = _se[treat]
bootstrap, reps(100): regress re78 treat $nondw_regvars if dataset == "d'1" | (dataset == "nsw" & tr
local reg_te = _b[treat]
local reg_te_se = _se[treat]

post `memhold' ("`d'_full_row") (0) ('obs') ('mean_ps') ('te_dim') ('reg_te') (.)
post `memhold' ("`d'_full_row") (1) (.) ('mean_ps_se') ('te_dim_se') ('reg_te_se') (.)

// Remaining rows
forvalues row = 3/10 {

use "../intermediate/combined_w_pscore_nondw.dta", clear
keep if ps_sample_`d' == 1

if `row' == 3 {
local samp_name "W/o Replacement: Random"
tempvar sortorder

```

```

gen `sortorder' = runiform()
sort `sortorder'
psmatch2 treat, outcome(re78) pscore(ps_`d') neighbor(1) noreplacement
}
else if `row' == 4 {
local samp_name "W/o Replacement: Low to High"
psmatch2 treat, outcome(re78) pscore(ps_`d') neighbor(1) noreplacement
}
else if `row' == 5 {
local samp_name "W/o Replacement: High to Low"
psmatch2 treat, outcome(re78) pscore(ps_`d') neighbor(1) noreplacement descending
}
else if `row' == 6 {
local samp_name "With Replacement: Nearest Neighbor"
psmatch2 treat, outcome(re78) pscore(ps_`d') neighbor(1)
}
else if `row' == 7 {
local samp_name "With Replacement: Caliper = 0.00001"
psmatch2 treat, outcome(re78) pscore(ps_`d') caliper(0.00001)
}
else if `row' == 8 {
local samp_name "With Replacement: Caliper = 0.00005"
psmatch2 treat, outcome(re78) pscore(ps_`d') caliper(0.00005)
}
else if `row' == 9 {
local samp_name "With Replacement: Caliper = 0.0001"
psmatch2 treat, outcome(re78) pscore(ps_`d') caliper(0.0001)
}
else if `row' == 10 {
local samp_name "With Replacement: Caliper = 0.001"
if "'d'" == "cps" continue
psmatch2 treat, outcome(re78) pscore(ps_`d') caliper(0.001)
}

keep if !missing(_weight)

count if treat == 0
local obs = `r(N)'
regress re78 treat $nondw_regvars, robust
local ols_te = _b[treat]
local ols_te_se = _se[treat]

expand _weight, generate(num) // applies weighting to subsequent estimations

summarize ps_`d' if treat == 0
local mean_ps = `r(mean)'
bootstrap, reps(100): regress ps_`d' treat
local mean_ps_se = _se[treat]
bootstrap, reps(100): regress re78 treat
local te_dim = _b[treat]
local te_dim_se = _se[treat]

```

```

bootstrap, reps(100): regress re78 treat $nondw_regvars
local reg_te = _b[treat]
local reg_te_se = _se[treat]

post `memhold' ("`samp_name'" (0) (`obs') (`mean_ps') (`te_dim') (`reg_te') (`ols_te')
post `memhold' ("`samp_name'" (1) (.) (`mean_ps_se') (`te_dim_se') (`reg_te_se') (`ols_te_se')

}

postclose `memhold'
use `results', clear

// Format latex table
format %13.0fc obs te_dim reg_te ols_te
format %13.2fc mean_ps
foreach var of varlist obs mean_ps te_dim reg_te ols_te {
    tostring `var', replace force usedisplayformat
    replace `var' = "" if `var' == "."
    replace `var' = "(" + `var' + ")" if sd_ind == 1 & !missing(`var')
}
replace ctrl_samp = "" if sd_ind == 1
drop sd_ind
label var ctrl_samp "Control Sample"
label var obs "Obs"
label var mean_ps "Mean Propensity Score"
label var te_dim "Treatment Effect (Diff. in Means)"
label var reg_te "Regression Treatment Effect"
label var ols_te "Q8: Simple OLS"
texsave * using "$ol/tab11_`d'.tex", varlabels frag replace title("`d'_title")
}

```