Evaluation of the Training Process

20307130030 Shen Jianzhi

2023.6.12

Abstract

This project tests different methods on estimating the ATT of the training program on the future earning, and estimates the effect modification. The methods include regression, Nearest Neighbor Matching, Full Matching, IP Weighting, CBPS Weighting, EBAL Weighting, DID, PSM-DID and DRDID. The project finds that the earning in 1974 is an important confounder, and regression is sensitive to both observations and specifications, while p-score methods are more robust on observations, and DID methods are also applicable in the task, but only on datasets with a filtered treated group.

1 Introduction

In order to study the effect of a job training program on the participants' earning, National Supported Work Demonstration conducted a randomized experiment on economically disadvantaged individuals. Paired with non-experimental control groups from PSID and CPS, the experimental data serves as a benchmark on observational studies. Since the program is targeted at the disadvantaged group, and the non-experimental control groups are representative of the whole population, the estimate of interests is the average treatment effect on the treated. Meanwhile, knowledge about effect modification may help the training program better serve the targeted group, so this project also considers it.

This project will test regression methods, propensity score methods and difference-in-difference methods on different NSW treated-control combinations, in Section 4, 5 and 6 respectively. After that, Section 8 will discuss the results and some prerequisites for further study.

2 Related Work

LaLonde (1986) showed the discrepancy between experimental result and estimations through econometric methods at that time, including simple adjusted regression, difference-in-difference and two-step selection model.

Dehejia and Wahba (1999) highlighted the importance of information about the earning in another pre-training year, selected a subgroup whose earning in 1974 could be inferred, and achieved lower bias using propensity score methods.

Smith and Todd (2003) pointed out the limitation of propensity score matching on this task and evaluated the robustness of a difference in difference matching strategy.

Abadie (2005) proposed semi-parametric DID estimators and Sant'Anna and Zhao (2020) proposed doubly robust DID estimators, which were even more robust on this task.

3 Data

3.1 Source

NSW: The original experimental data introduced by LaLonde (1986), including information about age, race, marriage, years of education, being a high school dropout or not, real earning in 1975, participating in the program or not, and post-training earning in 1978. The included individuals are socially disadvantaged individuals like ex-drug addicts, ex-criminal offenders, and high school dropouts. The data only includes those who completed the follow-up interviews, which affects the population but keeps the integrity of the experimental design. Since the observational estimates concerning female, including AFDC women, are generally close to experiment results, this project focuses on the male individuals. It is worth noticing that the randomized assignment had a 2-year period from 1975 to 1977, which according to Dehejia and Wahba (1999), may lead to "cohort phenomenon".

DW: Dehejia and Wahba (1999) further restricted the population to those whose earning in 1974 could be inferred, which means they either provided their 1974 earning information, or was unemployed that year, so that the real earning in 1974 is included in DW. The ratio of the unemployed rises in the selected sample, which further impairs the generalization capacity of the data.

PSID: From The Panel Study of Income Dynamics, PSID-1 includes all interviewed male household heads under age 55 who did not classify themselves as retired in 1975. In order to get closer to the treatment group, PSID-2 selects from PSID-1 men who were not working when surveyed in the spring of 1976, and PSID-3 selects from PSID-2 men who were not working in 1975 either.

CPS: From Current Population Survey, CPS-1 includes all surveyed male under 55. Similarly, CPS-2 selects from CPS-1 males who were not working when surveyed in March 1976, and CPS-3 selects from CPS-2 all the unemployed males in 1976 whose income in 1975 was below the poverty level.

3.2 Overview

Dataset	treat	no. obs	age	education	black	hispanic	married	nodegree	re74	re75	re78
NSW	0	425	24.45(0.32)	10.19(0.08)	0.8(0.02)	0.11(0.02)	0.16(0.02)	0.81(0.02)	NA	3026.68(252.3)	5090.05(277.37)
NSW	1	297	24.63(0.39)	10.38(0.11)	0.8(0.02)	0.09(0.02)	0.17(0.02)	0.73(0.03)	NA	3066.1(282.87)	5976.35(401.76)
DW	0	260	25.05(0.44)	10.09(0.1)	0.83(0.02)	0.11(0.02)	0.15(0.02)	0.83(0.02)	2107.03(352.75)	1266.91(192.44)	4554.8(340.09)
DW	1	185	25.82(0.53)	10.35(0.15)	0.84(0.03)	0.06(0.02)	0.19(0.03)	0.71(0.03)	2095.57(359.27)	1532.06(236.68)	6349.14(578.42)
PSID-1	0	2490	34.85(0.21)	12.12(0.06)	0.25(0.01)	0.03(0)	0.87(0.01)	0.31(0.01)	19428.75(268.68)	19063.34(272.48)	21553.92(311.73)
PSID-2	0	253	36.09(0.76)	10.77(0.2)	0.39(0.03)	0.07(0.02)	0.74(0.03)	0.49(0.03)	11027.3(679.91)	7569.22(568.46)	9995.95(703.16)
PSID-3	0	128	38.26(1.14)	10.3(0.28)	0.45(0.04)	0.12(0.03)	0.7(0.04)	0.51(0.04)	5566.87(641.27)	2610.7(492.54)	5279.29(686.14)
CPS-1	0	15992	33.23(0.09)	12.03(0.02)	0.07(0)	0.07(0)	0.71(0)	0.3(0)	14016.8(75.67)	13650.8(73.31)	14846.66(76.29)
CPS-2	0	2369	28.25(0.24)	11.24(0.05)	0.11(0.01)	0.08(0.01)	0.46(0.01)	0.45(0.01)	8727.96(184.25)	7397.23(166.67)	10171.11(181.87)
CPS-3	0	429	28.03(0.52)	10.24(0.14)	0.2(0.02)	0.14(0.02)	0.51(0.02)	0.6(0.02)	5619.24(327.76)	2466.48(158.94)	6984.17(352.17)

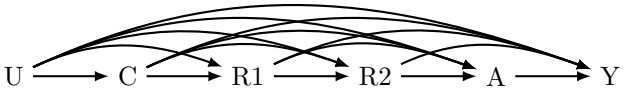
Table 1: Data Overview

As can be shown in Table 1, the treatment and control group within the raw and selected experimental data have relatively balanced covariate. Even though the absolute SMDs of education in NSW and age, education, whether dropout in DW are over 0.1, they are still within an acceptable range (below 0.14).

However, the covariates in the non-experiment control groups are distributed quite differently, since the treated group are selected as socially disadvantaged by design. Therefore, in order to derive a low-biased ATT estimate, we have to treat the paired treated and non-experimental control groups as observational data, and control the covariates according to the Causal DAG, where C denotes the personal background (Age, Education, Race, Marriage, Dropout), R1 and R2 denote the real earning in 1974 and 1975 respectively, A denotes the treatment and Y denotes the real earning in 1978.

It needs to be specified that the relationship between Marriage and real earning in 1974 and 1975 maybe different from what is shown in the DAG, since we do not know exactly when it happened, and thus either cases that not being in a marriage is due to low earning and that marriage may boost the will to earn may happen. However, the treatment was always preceded by all the three covariates, so treating them simply as the DAG shows suffices.

It is also worth noticing that unmeasured confounders U like location exists. Since an individual was only selected where there was a program site, and the effect of training program might vary from site to site.



Causal DAG

3.3 Benchmark

Dataset	ATE(SE)
NSW	886.30(488.20)
DW	1794.34(671.00)

Table 2: Experimental Result

Because the treated and control population are approximately identical in randomized experiment, ATE is equivalent to ATT. From Table 2 the effect modification of being unemployed is significant, since DW dataset overrepresent the unemployed. The effect modifiers will be further explored in Section 4.

4 Regression Methods

4.1 ATT

Assumptions: SUTVA, NUCA, positivity, linear relationship and uncorrelated Gaussian noise (for SE estimate)

Specifications tested are:

Unadjusted: Y on A

C: Y on A and C (Age, Education, Race, Marriage and Dropout)

Sq. Age: Y on A, C and an additional age squared

R2: Y on A and R2

All: Y on A, R2, C

All with Sq. Age: Y on A, R2, C and age squared

R1+R2: Y on A, R1 and R2

R1+All: Y on A, C, R1 and R2

R1+All but Marriage: Y on A, C/Marriage, R1 and R2

R1+All with Sq. Age: Y on A, C, R1, R2 and age squared

According to Causal DAG, R1+All and R1+All with Sq. Age should produce results with the lowest biases.

Since R1 is only available in DW dataset, the last four specifications are only tested on DW related datasets.

Dataset	Unadjusted	C	Sq. Age	R2	All	All with Sq. Age
NSW	886.3(472.09)	793.61(471.9)	791.44(472.28)	878.78(466.71)	806.51(467.89)	800.54(468.1)
NSW-PSID-1	-15577.57(913.33)	-6410.36(1070.22)	-5947.6(1072.06)	-2380.08(680.27)	-1457.91(801.63)	-1347.8(804.48)
NSW-PSID-2	-4019.6(781.4)	-1895.69(1012.88)	-2275.53(995.67)	-1363.82(729.04)	-605.71(937.13)	-951.73(930.43)
NSW-PSID-3	697.06(759.8)	623.64(1009.72)	-7.07(1010.71)	628.91(757.04)	518.22(1010.7)	-79.39(1011.63)
NSW-CPS-1	-8870.31(562.48)	-3236.27(579.62)	-3498.06(574.64)	-1543.49(425.69)	-992.92(451.55)	-999.61(451.88)
NSW-CPS-2	-4194.76(533)	-1080.04(621.62)	-1855.26(623.77)	-1648.94(458.63)	-683.45(549.76)	-1100.75(556.04)
NSW-CPS-3	-1007.82(539.35)	633.34(679.77)	427.02(695.27)	-1204.45(531.91)	211.21(682.48)	89.98(695.45)
DW	1794.34(632.85)	1671.13(637.97)	1669.97(638.55)	1750.15(632.09)	1636.28(637.66)	1636.11(638.3)
DW-PSID-1	-15204.78(1154.61)	-5928.11(1250.25)	-5613.38(1247.1)	-581.83(841.26)	395.78(931.25)	455.57(931.65)
DW-PSID-2	-3646.81(959.7)	-1087.58(1180.37)	-1614.28(1151.72)	720.5(886.35)	1385.12(1066.74)	873.69(1057.27)
DW-PSID-3	1069.85(899.62)	1212.07(1172.35)	475.08(1158.63)	1369.83(896.97)	1320.05(1166.85)	594.79(1156.01)
DW-CPS-1	-8497.52(712.02)	-2973.49(716.27)	-3436.79(710.24)	-77.71(536.6)	632.38(557.24)	622.55(558.01)
DW-CPS-2	-3821.97(670.6)	-697.72(747.38)	-1697.14(749.88)	-262.97(573.65)	911.39(657.5)	362.3(666.81)
DW-CPS-3	-635.03(657.14)	1163.92(811.62)	771.41(836.73)	-90.8(641.4)	1221.22(794.47)	1009.96(822.04)

Table 3: Different Specifications without R1

As can be seen in Table 3 and Table 4, without controlling R1, no specification produces an experimental-like result robustly. In general, they work better on PSID-3 and CPS-3, where population has already been trimmed to resemble the treated group.

After R1 is controlled, the estimates' biases are lower. This is due to the fact that R1 is also an important confounder, since the assignment in early 1975 might have referred to earning in 1974. Furthermore, two time stamps in the regression covariate provide the model with some information that can control some time-invariant unmeasured confounders, which will be further discussed in Section 6. However, the estimate is still sensitive to the observation, for CIs derived on some observational pair cover 0. Therefore, the E-value for those OLS results are 1, which means the estimated treatment effect may be solely brought by unmeasured confounders.

Dataset	R1+R2	R1+All	R1+All but Marriage	R1+All with Sq. Age
DW	1772.6(632.61)	1676.34(638.68)	1673.48(637.71)	1675.86(639.34)
DW-PSID-1	219.5(829.01)	751.95(915.26)	105.07(863.26)	795.02(915.79)
DW-PSID-2	1725.35(905.3)	1873.77(1060.56)	1301.65(1016.59)	1360.22(1052.39)
DW-PSID-3	2228(907.13)	1833.13(1159.78)	1491.57(1110.3)	1107.13(1152.55)
DW-CPS-1	169.05(530.17)	699.13(547.64)	684.29(546.86)	793.59(548.25)
DW-CPS-2	50.14(566.99)	1172.7(645.86)	1173.76(645.02)	813.32(657.77)
DW-CPS-3	798.13(635.69)	1548.24(781.28)	1500.52(776.58)	1368.98(808.95)

Table 4: Different Specifications with R1

An age squared term, which is commonly used in econometrics, does not reduce the bias of the estimate in this task.

Therefore, the regression method is sensitive to the specification, and observation as well. To provide a low-biased regression model needs to correctly specify the model and use less unbalanced datasets.

4.2 Effect Modification

This project uses NSW dataset and regression with interaction terms for low-biased effect modification estimate. U75 is the indicator whether an individual is unemployed in 1975, i.e. R2=0.

Covariate	EM
U75	1425.79(965.34)
Black	1360.11(1176.15)
Hispanic	1714.80(955.77)
Dropout	-352.24(1126.02)
Marriage	1738.49(1274.89)
Age(Cont.)	64.15(71.17)
Education(Cont.)	396.09(272.94)

Table 5: Estimate of Effect Modification

As shown in Table 5, although not significant, U75, Black, Hispanic, Marriage all have positive effect modification. However, it can be infer that academically disadvantaged individuals may benefit less from the program, since education is also a positive effect modifier. So it is advisable to figure out how to help those academically disadvantaged better.

5 Propensity Score Methods

Assumptions: Since the estimand is ATT, the assumptions are SUTVA and

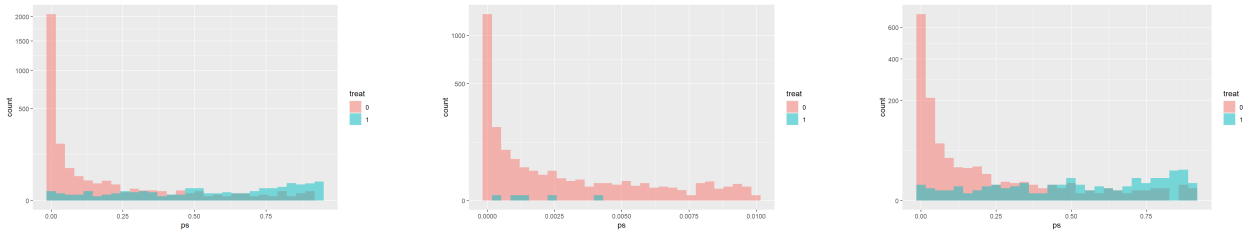
$$E[Y_0|C, A = 1] = E[Y_0|C, A = 0] = E[Y_0|C] \tag{1}$$

$$P(A = 1|C) < 1 \tag{2}$$

The propensity score is estimated through logistic regression, and a suggested specification is regression of Treat on C, R1 and R2.

5.1 Checking Overlap

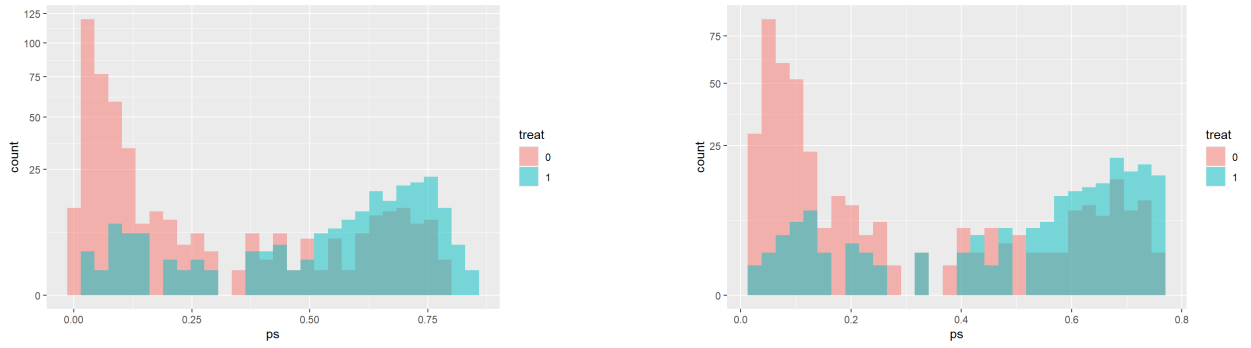
Figure 1 is an instance done on DW paired with PSID-1. For better vision effect, the y-axis of the histogram has been square-root transformed. At first glance the overlap seems acceptable. However, when the histogram is zoomed over interval [0, 0.01], it can be seen that over 1000 control units have no treated counterparts. Therefore, this project screens the observations by dropping the control units whose estimated propensity scores are below the treated minimum and the treated units whose p-scores are above the control maximum. As can be seen in Figure 1, the overlap is acceptable after screening.



(a) Before Screening (b) Zoomed (c) After Screening

Figure 1: Estimated p-score plot on DW-PSID-1

The effect of screening is easier to see on DW paired with CPS-3, as Figure 2 shows.



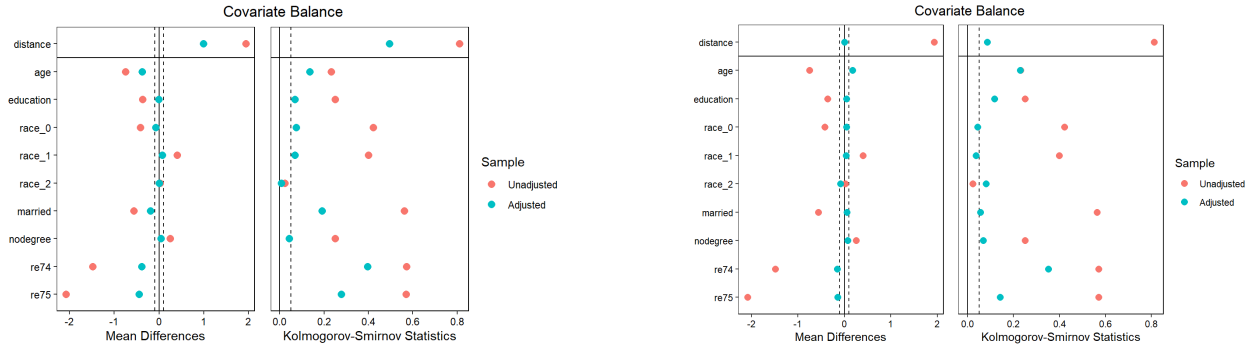
(a) Before Screening (b) After Screening

Figure 2: Estimated p-score plot on DW-CPS-3

5.2 Checking Covariate Balance

5.2.1 Matching

This project displays the balance result with nearest matching. As can be seen in Figure 3, matching without replacement fails to balance the covariate very well, which is due to the fact that the control group does not have enough economically disadvantaged observations to pair with treated units. The balance derived with replacement is acceptable, so this project resorts to matching with replacement, and the standard error is derived with bootstrap.



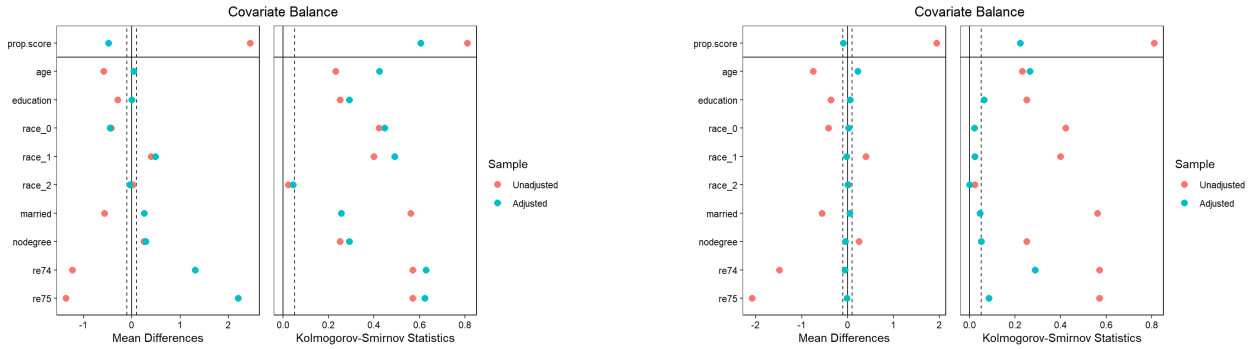
(a) Without Replacement (b) With Replacement

Figure 3: Love plot for matching on DW-PSID-1

5.2.2 Weighting

As Figure 4 shows, traditional weighting for ATE has problem balancing the covariates, because ATE requires the treated group to present the control group, which is not realistic in this situation where the treated group is from a specific population. But weighting for ATT works well, because it multiplies the ATE weights by the p-score to make the control group resemble the treated group. Therefore, this project uses weighting for ATT.

However, according to KS statistics, the distribution of Age and R1 in two groups is still varied.



(a) For ATE (b) For ATT

Figure 4: Love plot for weighting on DW-PSID-1

5.3 Result

The tested methods are:

NM (Nearest neighbor matching),

FM (Full matching): Introduced by Hansen and Klopfer (2006), a one-to-N subclassification strategy minimizing a weighted average of the estimated distance measure between each treated and each control unit within each subclass.

IPW,

CBPS weighting: Proposed by Imai and Ratkovic (2013), a method modeling assignment mechanism while balancing the covariates under the framework of general method-of-moments or empirical likelihood.

EBAL weighting: Proposed by Hainmueller (2012), an entropy balancing method to construct a weight for each control observation such that the sample moments of observed covariates are identical between the treatment and weighted control groups

The standard errors of matching estimates are derived with bootstrap.

Different specification for sensitivity analysis are:

Suggested: A on C, R1 and R2

Sq. Age: A on C, R1, R2 with an additional age squared

No R1: A on C and R2 NSW can be tested on this specification

As shown in Table 6, with the suggested specification, all methods achieve robust performance over different datasets, except nearest neighbor matching on DW-CPS-2 and DW-CPS-3 (with E-value 2.66 and 3.16). It is worth noticing that before screening, the matching result is 1991.62 (not listed in the table). One possible explanation is that screening DW to more specific population CPS-2 and CPS-3 again changes the distribution of the treated group.

Across different specifications, Sq. Age still sees robust performance, but No R1 leads to large biases. These results indicates the great observational confounding brought by R1, and the sensitivity of p-score methods to specification of assignment mechanism.

Based on the results, on this task full matching is more robust than nearest neighbor matching, and weighting, on average, is more robust than matching. Each weighting method performs slightly better or worse depending on the dataset.

Nevertheless, since most estimated CIs still covers 0, the sensitivity for unmeasured confounding is still high in most results. The results derived by IPW, CBPS and EBAL on DW-PSID-1 is relatively significant, and the E-values are respectively 1.82, 1.81 and 1.84, which means the unmeasured confounder needs to approximately double the treating probability as well as double the probability to have high 1978 earnings to fully explain the ATT. Therefore, the sensitivity of these results is relatively low.

6 Difference in Difference Methods

Assumptions: SUTVA, positivity, Parallel Trend Assumption, linear relationship and uncorrelated Gaussian noise

The basic DID model treated R2 also as dependent variables:

Specification	Dataset	NM	FM	IPW	CBPS	EBAL
Suggested	DW	1973.39(856.65)	1877.92(690.57)	1781.28(697.47)	1780.49(697.8)	1780.7(697.79)
	DW-PSID-1	1740.57(1242.93)	1293.71(776.64)	1962.52(954.78)	1927.21(908.92)	2013.92(914.54)
	DW-PSID-2	1162.91(1089.03)	1438.13(975.63)	2017.6(1065.33)	2529.07(999.63)	2519.41(1003.71)
	DW-PSID-3	717.09(1197.75)	1002.41(1067.98)	1929.61(1275.99)	1864.92(1165.34)	2313.55(1154.44)
	DW-CPS-1	1681.97(872.71)	1800.54(607.02)	1173.11(635.15)	1268.83(635.29)	1268.6(635.32)
	DW-CPS-2	530.55(980.55)	1236.81(604.69)	1085.85(685.28)	1141.85(688.32)	1143.17(688.33)
	DW-CPS-3	586.18(847.46)	976.1(640.52)	975.02(772.65)	1043.93(765.42)	1041.1(765.68)
Sq. Age	DW	1956.97(836.96)	1952.3(681.25)	1777.46(697.55)	1774.19(697.71)	1774.58(697.74)
	DW-PSID-1	808.77(1189.55)	557.13(701.76)	1308.45(1064.56)	1860.1(943.44)	1862.43(949.17)
	DW-PSID-2	1154.04(1199.12)	1261.33(958.74)	1125.93(1222.28)	1601.45(1028.59)	2053.82(1058.86)
	DW-PSID-3	2399.1(987.39)	2830.09(1120.79)	847.56(1359.58)	960.44(1116.72)	1768.77(1181.92)
	DW-CPS-1	2201.13(975.22)	1805.65(532.2)	1384.66(697.79)	1321.87(696.07)	1324.96(696.48)
	DW-CPS-2	1053.1(1000.23)	1059.83(720.54)	1795.21(787.51)	1370.33(779.7)	1394.39(779.62)
	DW-CPS-3	2249.1(870.87)	2117.94(609.77)	1278.1(822.89)	1131.06(804.19)	1101.77(816.39)
No R1	NSW	604.17(667.39)	627.99(484.57)	822.73(498.51)	838.34(497.99)	837.91(498.01)
	NSW-PSID-1	-1947.19(1276.17)	-1572.98(501.99)	-672.18(881.97)	-753.47(872.07)	-742.49(872.91)
	NSW-PSID-2	-198.5(959.5)	-580.61(850.68)	89.24(1044.82)	-215.28(959.36)	33.17(976.21)
	NSW-PSID-3	499.49(1197.96)	796.7(863.88)	520.07(1413.82)	233.29(1313.51)	-49.24(1380.1)
	NSW-CPS-1	-679.3(757.59)	-622.29(355.45)	-578.22(478.31)	-583.44(476.12)	-580.66(476.13)
	NSW-CPS-2	-517.62(707.26)	-700.97(469.65)	-581.43(551.82)	-646.16(552.69)	-644.24(552.82)
	NSW-CPS-3	-651.33(913.53)	-755.35(582.31)	-182.08(835.51)	-74.58(799.66)	-77.04(800.58)
	DW	1141.15(814.48)	1365.77(638.13)	1647.12(694.32)	1654.23(693.49)	1654.05(693.52)
	DW-PSID-1	-74.18(1721.17)	146.38(607.75)	707.93(1163.59)	448.02(1147.57)	469.55(1147.44)
	DW-PSID-2	604.33(1156.04)	356.59(913.39)	1198.89(1172.98)	1555.56(1149.19)	1556.21(1149.35)
	DW-PSID-3	254.41(1062.62)	68.23(914.58)	1398.11(1413.21)	1254.98(1325.03)	1361.21(1310.15)
	DW-CPS-1	784.58(814.32)	1416.37(581.14)	1105.43(635.93)	1166.32(636.19)	1164.85(636.18)
	DW-CPS-2	820.19(926.57)	495.83(599)	902.57(688.48)	916.87(690.25)	917.18(690.17)
	DW-CPS-3	761.63(1005.36)	964.35(654.85)	987.27(846.85)	1005.84(844.84)	1004.77(844.82)

Table 6: Results of different methods on different datasets and specifications

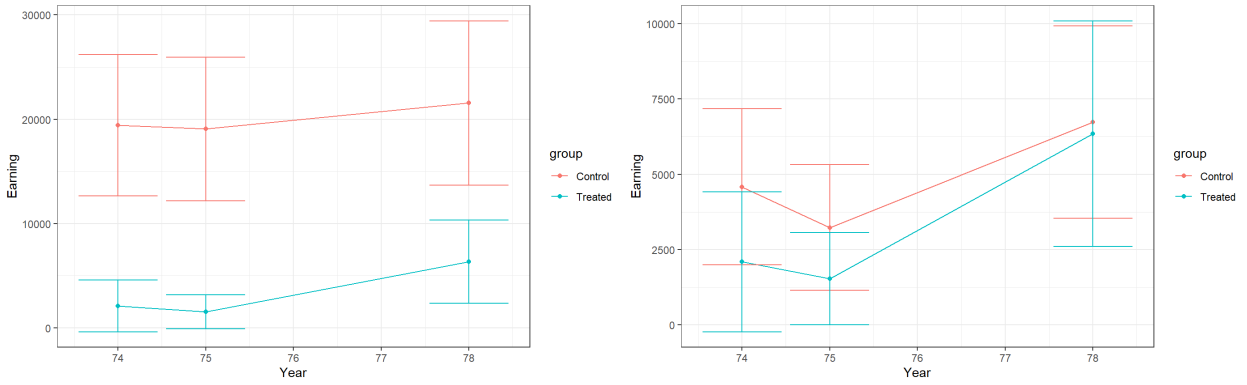
$$Y_t = \alpha + \beta(A \times D_t) + \gamma A + \delta D_t + \theta^{\mathrm{T}} C + \epsilon \tag{3}$$

where Y_1, Y_2 denotes R2, Y respectively, D_t denotes whether the training program has taken place at time t. In theory, DID methods may help control the time-invariant unmeasured confounding like location.

As can be seen in Figure 5, the two trends before intervention is approximately parallel, C controlled or not, and the treatment effect is indicated by the change in trend.

The methods tested are simple DID, PSM-DID, and Doubly Robust DID. Doubly robust estimate combines p-score balancing and regression, and yields a consistent estimate with either correct assignment specification or correct outcome specification.

Table 7 shows that, whether R1 appears in the assignment specification does not affect DRDID on DW



(a) Without Controlling c

(b) C Controlled by matching

Figure 5: DID Plots

Dataset	DID	PSM-DID	PSM-DID without R1	DRDID	DRDID without R1
DW	1529.19(737.08)	1457.12(820.72)	936.62(837.11)	1494.69(707.79)	1524.68(722.72)
DW-PSID-1	2326.50(749.67)	2749.16(1111.84)	-911.04(1921.98)	3374.07(816.66)	3269.86(763.02)
DW-PSID-2	2390.36(1099.17)	1352.80(1356.29)	2126.11(1286.55)	3243.93(987.21)	2713.62(1049.23)
DW-PSID-3	2148.49(1050.70)	653.84(1388.54)	1366.15(1342.57)	2379.76(1000.30)	1547.27(1179.38)
DW-CPS-1	3621.23(633.86)	1193.56(866.49)	551.92(924.35)	1869.52(644.95)	2455.87(653.12)
DW-CPS-2	2043.21(671.89)	301.68(997.32)	983.74(902.3)	1445.51(690.81)	1834.25(710.6)
DW-CPS-3	299.40(734.76)	1405.57(965.37)	851.76(1046.3)	1153.75(805.72)	1036.62(830.19)
NSW	846.89(618.07)	NA	568.29(759.49)	NA	948.87(600.63)
NSW-PSID-1	419.67(642.53)	NA	-1635.5(1489.59)	NA	1271.06(680.98)
NSW-PSID-2	483.53(1029.08)	NA	738.19(1145.3)	NA	756.94(981.77)
NSW-PSID-3	241.66(977.14)	NA	-306.34(1473.71)	NA	-469.95(1086.05)
NSW-CPS-1	1714.4(502.61)	NA	-718.18(813.67)	NA	313.69(539.84)
NSW-CPS-2	136.38(549.8)	NA	-854.1(841.59)	NA	-353.94(603.3)
NSW-CPS-3	-1607.43(625.07)	NA	-109.48(901.39)	NA	-1005.57(739.2)

Table 7: Results of DID methods

datasets too much, while PSM-DID is sensitive to it. On NSW datasets, the simple DID method works relatively well, but still fails on NSW-CPS-3, and other methods all have problems producing an unbiased estimate. The reason may be that NSW data does not meet the parallel assumption. The earning trend of the economically disadvantaged population in NSW may be upward instead, and DW changes this by selecting mostly unemployed units. However, that is not testable.

7 Conclusion

Generally, R1 is an important confounder in this task, since all methods work well only after R1 is controlled. There is positive effect modification in unemployed group, black group, hispanic group, and

married group. Education may brings a positive effect modification as well.

Over different methods, regression is sensitive to both observations and outcome specifications, some p-score methods like FM, CBPS, and EBAL are relatively more robust over different observations, but are still sensitive to assignment specifications. DID methods works robustly on DW, and DRDID is even robust without controlling R1, but still produces large biases on NSW.

8 Discussion

Most methods provide highly-biased estimates without further controlling R1, which may be accounted that only when R1 and R2 are controlled, an individual's will to earn more, which may also be an important unmeasured confounder, is partially controlled. More importantly, this factor may be time-variant, so DID also fails to provide a robust estimate.

The failure on NSW may also be explained by the already insignificant result derived from the experimental data (CI of NSW experimental ATE covers 0) and the effect modification.

Furthermore, there may be unmeasured covariates like location, since the experimental data may have more individual in the poverty-stricken areas while observational data does not, and the site in those areas may be subpar, thus bringing underestimate problem. Besides DID, we might resort to IV analysis if the treated were not all compliers, so future experimental design shall allow never takers.

The indirect effect through social connection is a possible mediator analysis target. Some individuals might earn more in the future simply because they established a good relationship with the site manager, who they would not have met if they were not assigned into the program. Therefore, we can add another dimension by surveying whether he has a good connection with the site manager. Therefore, the direct effect of training program through capability development can emerge.

References

Robert J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620, 1986. ISSN 00028282. URL http://www.jstor.org/stable/1806062.

Rajeev H. Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062, 1999. doi: 10.1080/01621459.1999.10473858. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1999.10473858.

Jeffrey Smith and Petra Todd. Does matching overcome lalonde's critique of nonexperimental estimators? University of Western Ontario, Centre for Human Capital and Productivity (CHCP) Working Papers 20035, University of Western Ontario, Centre for Human Capital and Productivity (CHCP), 2003. URL https://EconPapers.repec.org/RePEc:uwo:hcuwoc:20035.

Alberto Abadie. Semiparametric Difference-in-Differences Estimators. *The Review of Economic Studies*, 72

(1):1–19, 01 2005. ISSN 0034-6527. doi: 10.1111/0034-6527.00321. URL https://doi.org/10.1111/0034-6527.00321.

Pedro H.C. Sant'Anna and Jun Zhao. Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1):101–122, 2020. ISSN 0304-4076. doi: https://doi.org/10.1016/j.jeconom.2020.06.003. URL https://www.sciencedirect.com/science/article/pii/S0304407620301901.

Ben B Hansen and Stephanie Olsen Klopfer. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15(3):609–627, 2006. doi: 10.1198/106186006X137047. URL https://doi.org/10.1198/106186006X137047.

Kosuke Imai and Marc Ratkovic. Covariate Balancing Propensity Score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):243–263, 07 2013. ISSN 1369-7412. doi: 10.1111/rssb.12027. URL https://doi.org/10.1111/rssb.12027.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012. doi: 10.1093/pan/mpr025.