Causal Effects in Nonexperimental Studies

Rajeev H. DEHEJIA and Sadek WAHBA (1999)

Patrick Altmeyer, Simon Neumeyer

January, 2021

Overview

- Not always possible to set up a randomized control trial (RCT) – costly
- Question: can we reproduce causal effect estimates obtained through a randomized control trial - the gold standard here taken as *ground truth* - using control data from nonexperimental studies? Enter: propensity scores
- Methodology paper:
  - Show that applying propensity scores to LaLonde (1986)'s nonexperimental data can mitigate these concerns
  - Estimates are not sensitive to the specification of the estimated propensity score, but are sensitive to the assumption of selection on observable

Data and methodology

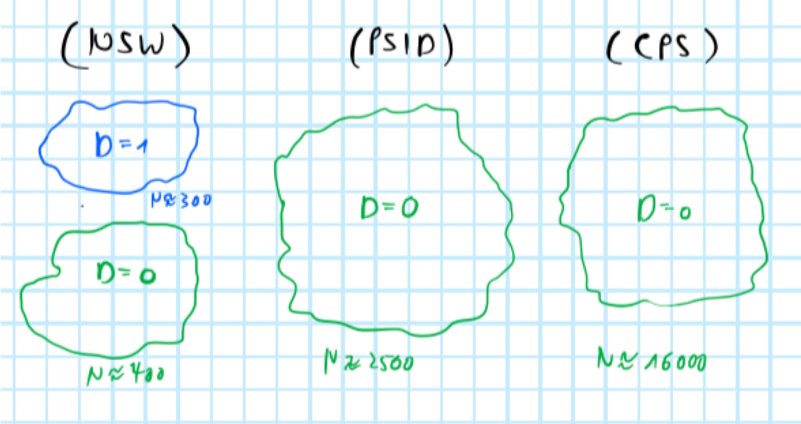
The RCT

NSW – federally funded support program in the 1970s providing opportunities for work experience to unemployed, economically struggling individuals

- Treatment randomly assigned ($N_1 = 297$) among those eligible ($N = 722$)
  - But randomization over a period of two years – *cohort phenomenom*: change in characteristics of individuals within treatment group
- Dehejia and Wahba (1999) further limit themselves to individuals that entered early enough (important to look at several years of preintervention earnings)

Nonexperimental comparison

- Panel Study of Income Dynamics
- Current Population Survey



Summary statistics

- nonexperimental populations differ dramatically from the treatment group in terms of age, marital status, ethnicity, and preintervention earning
- only small adjustments made in LaLonde (1986) (subsets based on one or two pretreatment variables).

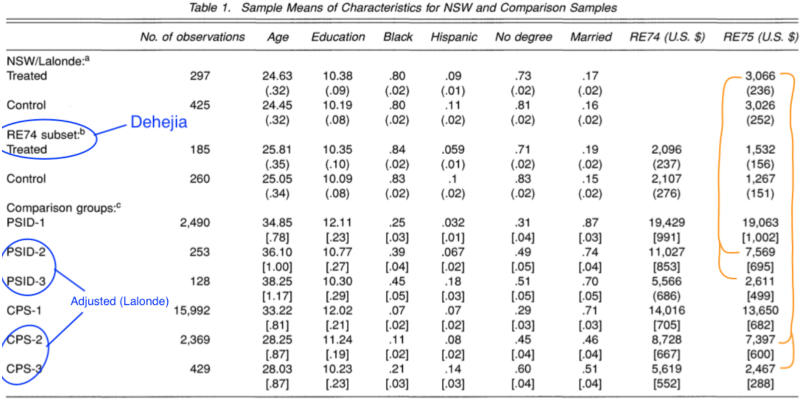


Table 1. Sample Means of Characteristics for NSW and Comparison Samples

	No. of observations	Age	Education	Black	Hispanic	No degree	Married	RE74 (U.S. $)	RE75 (U.S. $)
NSW/Lalonde:[a]									
Treated	297	24.63	10.38	.80	.09	.73	.17		3,066
		(.32)	(.09)	(.02)	(.01)	(.02)	(.02)		(236)
Control	425	24.45	10.19	.80	.11	.81	.16		3,026
		(.32)	(.08)	(.02)	(.02)	(.02)	(.02)		(252)
RE74 subset:[b]									
Treated	185	25.81	10.35	.84	.059	.71	.19	2,096	1,532
		(.35)	(.10)	(.02)	(.01)	(.02)	(.02)	(237)	(156)
Control	260	25.05	10.09	.83	.1	.83	.15	2,107	1,267
		(.34)	(.08)	(.02)	(.02)	(.02)	(.02)	(276)	(151)
Comparison groups:[c]									
PSID-1	2,490	34.85	12.11	.25	.032	.31	.87	19,429	19,063
		[.78]	[.23]	[.03]	[.01]	[.04]	[.03]	[991]	[1,002]
PSID-2	253	36.10	10.77	.39	.067	.49	.74	11,027	7,569
		[1.00]	[.27]	[.04]	[.02]	[.05]	[.04]	[853]	[695]
PSID-3	128	38.25	10.30	.45	.18	.51	.70	5,566	2,611
		[1.17]	[.29]	[.05]	[.03]	[.05]	[.05]	(686)	[499]
CPS-1	15,992	33.22	12.02	.07	.07	.29	.71	14,016	13,650
		[.81]	[.21]	[.02]	[.02]	[.03]	[.03]	[705]	[682]
CPS-2	2,369	28.25	11.24	.11	.08	.45	.46	8,728	7,397
		[.87]	[.19]	[.02]	[.02]	[.04]	[.04]	[667]	[600]
CPS-3	429	28.03	10.23	.21	.14	.60	.51	5,619	2,467
		[.87]	[.23]	[.03]	[.03]	[.04]	[.04]	[552]	[288]

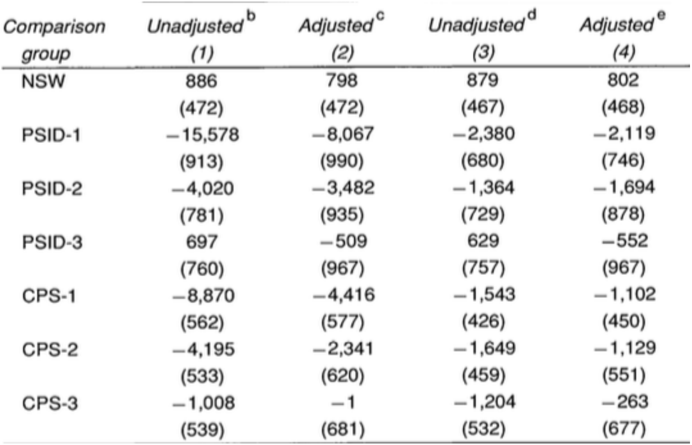
Dehejia

Adjusted (Lalonde)

Lalonde approach

Lalondes's main findings

- LaLonde (1986) finds large differences between RCT and nonexperimental estimates:

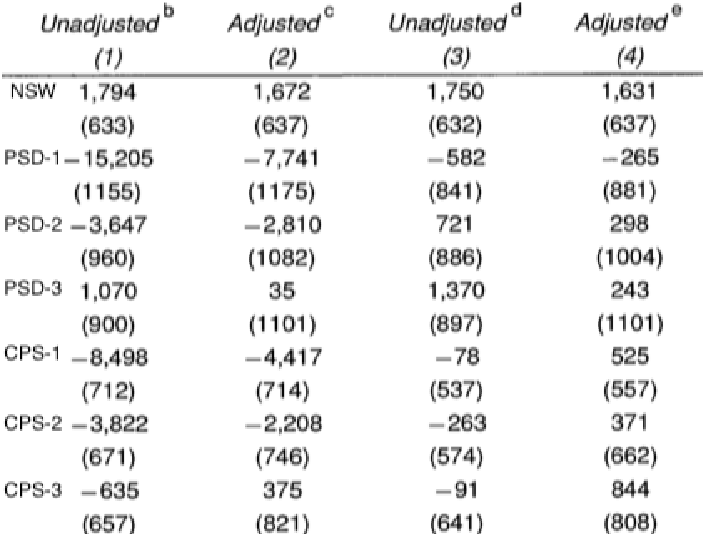


Comparison group	*Unadjusted*[b] *(1)*	*Adjusted*[c] *(2)*	*Unadjusted*[d] *(3)*	*Adjusted*[e] *(4)*
NSW	886	798	879	802
	(472)	(472)	(467)	(468)
PSID-1	−15,578	−8,067	−2,380	−2,119
	(913)	(990)	(680)	(746)
PSID-2	−4,020	−3,482	−1,364	−1,694
	(781)	(935)	(729)	(878)
PSID-3	697	−509	629	−552
	(760)	(967)	(757)	(967)
CPS-1	−8,870	−4,416	−1,543	−1,102
	(562)	(577)	(426)	(450)
CPS-2	−4,195	−2,341	−1,649	−1,129
	(533)	(620)	(459)	(551)
CPS-3	−1,008	−1	−1,204	−263
	(539)	(681)	(532)	(677)

- Concluded that existing econometric techniques could not produce unbiased estimates

Subsample in Dehejia and Wahba (1999)

- Results for subsample that includes on individuals that entered early enough:



	Unadjusted[b] *(1)*	*Adjusted*[c] *(2)*	*Unadjusted*[d] *(3)*	*Adjusted*[e] *(4)*
NSW	1,794	1,672	1,750	1,631
	(633)	(637)	(632)	(637)
PSD-1	−15,205	−7,741	−582	−265
	(1155)	(1175)	(841)	(881)
PSD-2	−3,647	−2,810	721	298
	(960)	(1082)	(886)	(1004)
PSD-3	1,070	35	1,370	243
	(900)	(1101)	(897)	(1101)
CPS-1	−8,498	−4,417	−78	525
	(712)	(714)	(537)	(557)
CPS-2	−3,822	−2,208	−263	371
	(671)	(746)	(574)	(662)
CPS-3	−635	375	−91	844
	(657)	(821)	(641)	(808)

Subsample – including 1974 earnings

- When including 1974 earnings $\mathbf{x}_{1974}$ as a predictor, findings overall are still similar and essence of LaLonde (1986)'s claim still holds
- Mild improvements:
  - estimates remain largely negative when not controlling for pretreatment covariate, although less so than without controlling for $\mathbf{x}_{1974}$
  - results in column 5 – controlling for all pretreatment covariates – getting us much closer to RCT estimates

Incorporating propensity scores

Recap: treatment effect of the treated

- Goal is to identify the average treatment effect of the treated

$$\alpha_{ATT} = \mathbb{E}\left(Y_{i1}|D_i = 1\right) - \mathbb{E}\left(Y_{i0}|D_i = 1\right) \tag{1}$$

- Under the assumption of conditional unconfoundedness – $(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i|\mathbf{X_i}$ – we can identify α_{ATT} as

$$\mathbb{E}_{\mathbf{X_i}|D_i=1}\left(\mathbb{E}\left(Y_{i1}|D_i = 1\right) - \mathbb{E}\left(Y_{i0}|D_i = 1\right)|D_i = 1\right) \tag{2}$$

Note: Dehejia and Wahba (1999) in fact integrate over the entire NSW population (“group of interest”). Not entirely clear why, but anyway treatment and control group in NSW almost the same in terms of pretreatment variables.

Recap: propensity scores

- In order to estimate α_{ATT} we need to draw subsets from the nonexperimental control groups that conditional on their covariates $\mathbf{X}_i$ have similar propensities to be treated (as the NWS guys)
- Denoting the propensity score as $\pi(\mathbf{X}_i) = p(D_i = 1|\mathbf{X}_i)$ and assuming that $0 < p(\mathbf{X}_i) < 1 \;\; \forall \; i$ (common support) then

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i|\mathbf{X_i} = (Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i|\pi(\mathbf{X_i}) \qquad (3)$$

Intuition: observations with the same propensity score have the same distribution of the full vector of covariates.

Estimation of propensity scores

Step 1

- To estimate propensity scores Dehejia and Wahba (1999) use logistic regression

$$\pi(\mathbf{X}_i) = \frac{\exp(\mathbf{X}\beta)}{1 + \exp(\mathbf{X}\beta)} \tag{4}$$

- They test for interactions and non-linearities but ultimately find that letting all covariates enter linearly does the job

Estimates

- Estimated propensity scores for PSID nonexperimental sample – limited overlap:

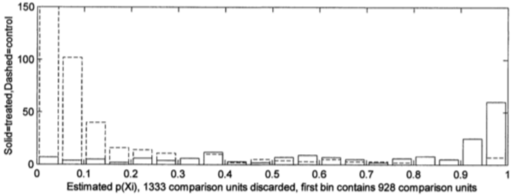


Figure 1. Histogram of the Estimated Propensity Score for NSW Treated Units and PSID Comparison Units. The 1,333 PSID units whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The first bin contains 928 PSID units. There is minimal overlap between the two groups. Three bins (.8–.85, .85–.9, and .9–.95) contain no comparison units. There are 97 treated units with an estimated propensity score greater than .8 and only 7 comparison units.

- CPS looks similar, tough less individual with high propensities to be treated

Estimation

Step 2

- Regression:

$$Y_i = (\mathbf{1}, D_i, \mathbf{X}_i)\beta \tag{5}$$

$$Y_i = (D_i, \pi(\mathbf{X}_i))\,\beta \tag{6}$$

- Difference in sample means
  - Sum the within-stratum differences weighted by the number of treated observations within stratum
  - Propensity score matching (nearest neighbour)

Results

Results

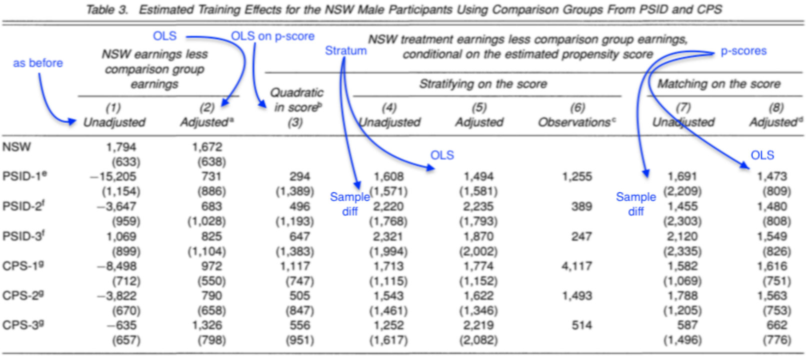

as before
OLS
OLS on p-score
Stratum
p-scores
OLS
Sample diff
OLS
Sample diff

Table 3. Estimated Training Effects for the NSW Male Participants Using Comparison Groups From PSID and CPS

	NSW earnings less comparison group earnings		NSW treatment earnings less comparison group earnings, conditional on the estimated propensity score					
			Quadratic in score[b]	Stratifying on the score			Matching on the score	
	(1) Unadjusted	(2) Adjusted[a]	(3)	(4) Unadjusted	(5) Adjusted	(6) Observations[c]	(7) Unadjusted	(8) Adjusted[d]
NSW	1,794	1,672						
	(633)	(638)						
PSID-1[e]	−15,205	731	294	1,608	1,494	1,255	1,691	1,473
	(1,154)	(886)	(1,389)	(1,571)	(1,581)		(2,209)	(809)
PSID-2[f]	−3,647	683	496	2,220	2,235	389	1,455	1,480
	(959)	(1,028)	(1,193)	(1,768)	(1,793)		(2,303)	(808)
PSID-3[f]	1,069	825	647	2,321	1,870	247	2,120	1,549
	(899)	(1,104)	(1,383)	(1,994)	(2,002)		(2,335)	(826)
CPS-1[g]	−8,498	972	1,117	1,713	1,774	4,117	1,582	1,616
	(712)	(550)	(747)	(1,115)	(1,152)		(1,069)	(751)
CPS-2[g]	−3,822	790	505	1,543	1,622	1,493	1,788	1,563
	(670)	(658)	(847)	(1,461)	(1,346)		(1,205)	(753)
CPS-3[g]	−635	1,326	556	1,252	2,219	514	587	662
	(657)	(798)	(951)	(1,617)	(2,082)		(1,496)	(776)

Sensitivity analysis

Estimates for treatment impact are:

- not sensitive to specification of propensity score
- quite sensitive to selection on observables: When dropping 1974 earnings, results change quite a lot
    - shows us importance of including pre-intervention variables (lengthy earnings history)
    - demonstrates value of using multiple comparison groups

Sensitivity analysis

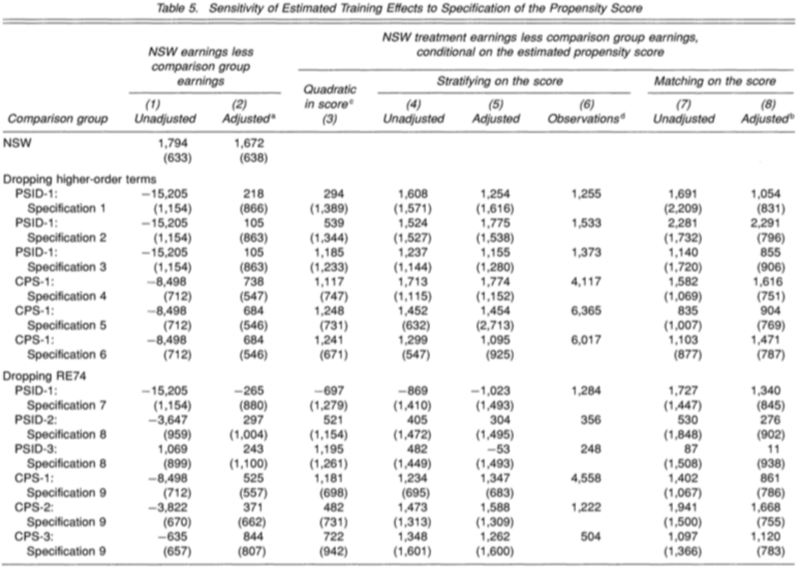


Table 5. Sensitivity of Estimated Training Effects to Specification of the Propensity Score

	NSW earnings less comparison group earnings		NSW treatment earnings less comparison group earnings, conditional on the estimated propensity score					
			Quadratic in score[c]	Stratifying on the score			Matching on the score	
Comparison group	(1) Unadjusted	(2) Adjusted[a]	(3)	(4) Unadjusted	(5) Adjusted	(6) Observations[d]	(7) Unadjusted	(8) Adjusted[b]
NSW	1,794	1,672						
	(633)	(638)						
Dropping higher-order terms								
PSID-1:	−15,205	218	294	1,608	1,254	1,255	1,691	1,054
Specification 1	(1,154)	(866)	(1,389)	(1,571)	(1,616)		(2,209)	(831)
PSID-1:	−15,205	105	539	1,524	1,775	1,533	2,281	2,291
Specification 2	(1,154)	(863)	(1,344)	(1,527)	(1,538)		(1,732)	(796)
PSID-1:	−15,205	105	1,185	1,237	1,155	1,373	1,140	855
Specification 3	(1,154)	(863)	(1,233)	(1,144)	(1,280)		(1,720)	(906)
CPS-1:	−8,498	738	1,117	1,713	1,774	4,117	1,582	1,616
Specification 4	(712)	(547)	(747)	(1,115)	(1,152)		(1,069)	(751)
CPS-1:	−8,498	684	1,248	1,452	1,454	6,365	835	904
Specification 5	(712)	(546)	(731)	(632)	(2,713)		(1,007)	(769)
CPS-1:	−8,498	684	1,241	1,299	1,095	6,017	1,103	1,471
Specification 6	(712)	(546)	(671)	(547)	(925)		(877)	(787)
Dropping RE74								
PSID-1:	−15,205	−265	−697	−869	−1,023	1,284	1,727	1,340
Specification 7	(1,154)	(880)	(1,279)	(1,410)	(1,493)		(1,447)	(845)
PSID-2:	−3,647	297	521	405	304	356	530	276
Specification 8	(959)	(1,004)	(1,154)	(1,472)	(1,495)		(1,848)	(902)
PSID-3:	1,069	243	1,195	482	−53	248	87	11
Specification 8	(899)	(1,100)	(1,261)	(1,449)	(1,493)		(1,508)	(938)
CPS-1:	−8,498	525	1,181	1,234	1,347	4,558	1,402	861
Specification 9	(712)	(557)	(698)	(695)	(683)		(1,067)	(786)
CPS-2:	−3,822	371	482	1,473	1,588	1,222	1,941	1,668
Specification 9	(670)	(662)	(731)	(1,313)	(1,309)		(1,500)	(755)
CPS-3:	−635	844	722	1,348	1,262	504	1,097	1,120
Specification 9	(657)	(807)	(942)	(1,601)	(1,600)		(1,366)	(783)

Conclusion

- Propensity methods can mitigate selection bias and thereby help facilitate the use of nonexperimental data as a comparison population for an RCT
- This offers a way to test previous RCT results for robustness without the need to set up an expensive follow-up study

References

Dehejia, Rajeev H, and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94 (448): 1053–62.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *The American Economic Review*, 604–20.