LABOR ECONOMISTS KNOW THE SCORE

Josh Angrist

MIT 14.381: Fall 2019

Research Flow

- My generation trained on the challenge of identifying training effects
  - Trainees are, by definition, special
- *The case for random assignment*
  - Trainees have a distinctive earnings profile, with low and *declining* earnings before treatment (Ashenfelter, 1978)
  - Diff-in-diff-type strategies generate highly inconsistent estimates that therefore seem unreliable (Ashenfelter and Card, 1985)
  - Lalonde's (1986) landmark study shows that non-experimental methods fail to replicate experimental findings for the NSW (**Tables 1,3,5)**
  - Non-experimental estimates seem highly variable and likely to be misleading, especially for men (Lalonde, 1995)

TABLE 1—THE SAMPLE MEANS AND STANDARD DEVIATIONS OF PRE-TRAINING EARNINGS AND OTHER CHARACTERISTICS FOR THE NSW AFDC AND MALE PARTICIPANTS

	Full National Supported Work Sample			
	AFDC Participants		Male Participants	
Variable	Treatments	Controls	Treatments	Controls
Age	33.37	33.63	24.49	23.99
	(7.43)	(7.18)	(6.58)	(6.54)
Years of School	10.30	10.27	10.17	10.17
	(1.92)	(2.00)	(1.75)	(1.76)
Proportion	.70	.69	.79	.80
High School Dropouts	(.46)	(.46)	(.41)	(.40)
Proportion Married	.02	.04	.14	.13
	(.15)	(.20)	(.35)	(.35)
Proportion Black	.84	.82	.76	.75
	(.37)	(.39)	(.43)	(.43)
Proportion Hispanic	.12	.13	.12	.14
	(.32)	(.33)	(.33)	(.35)
Real Earnings	$393	$395	1472	1558
1 year Before	(1,203)	(1,149)	(2656)	(2961)
Training	[43]	[41]	[58]	[63]
Real Earnings	$854	$894	2860	3030
2 years Before	(2,087)	(2,240)	(4729)	(5293)
Training	[74]	[79]	[104]	[113]
Hours Worked	90	92	278	274
1 year Before	(251)	(253)	(466)	(458)
Training	[9]	[9]	[10]	[10]
Hours Worked	186	188	458	469
2 years Before	(434)	(450)	(654)	(689)
Training	[15]	[16]	[14]	[15]
Month of Assignment	−12.26	−12.30	−16.08	−15.91
(Jan. 78 = 0)	(4.30)	(4.23)	(5.97)	(5.89)
Number of Observations	800	802	2083	2193

Note: The numbers shown in parentheses are the standard deviations and those in the

TABLE 3—ANNUAL EARNINGS OF NSW MALE TREATMENTS, CONTROLS, AND SIX CANDIDATE COMPARISON GROUPS FROM THE *PSID* AND *CPS-SSA*

			Comparison Group[a,b]					
Year	Treatments	Controls	*PSID*-1	*PSID*-2	*PSID*-3	*CPS-SSA*-1	*CPS-SSA*-2	*CPS-SSA*-3
1975	\$3,066	\$3,027	19,056[a]	7,569	2,611	13,650	7,387	2,729
	(283)	(252)	(272)	(568)	(492)	(73)	(206)	(197)
1976	\$4,035	\$2,121	20,267	6,152	3,191	14,579	6,390	3,863
	(215)	(163)	(296)	(601)	(609)	(75)	(187)	(267)
1977	\$6,335	\$3,403	20,898	7,985	3,981	15,046	9,305	6,399
	(376)	(228)	(296)	(621)	(594)	(76)	(225)	(398)
1978	\$5,976	\$5,090	21,542	9,996	5,279	14,846	10,071	7,277
	(402)	(227)	(311)	(703)	(686)	(76)	(241)	(431)
Number of Observations	297	425	2,493	253	128	15,992	1,283	305

[a]The Comparison Groups are defined as follows: *PSID*-1: All male household heads continuously from 1975 through 1978, who were less than 55-years-old and did not classify themselves as retired in 1975; *PSID*-2: Selects from the *PSID*-1 group all men who were not working when surveyed in the spring of 1976; *PSID*-3: Selects from the *PSID*-1 group all men who were not working when surveyed in either spring of 1975 or 1976; *CPS-SSA*-1: All males based on Westat's criteria, except those over 55-years-old; *CPS-SSA*-2: Selects from *CPS-SSA*-1 all males who were not working when surveyed in March 1976; *CPS-SSA*-3: Selects from the *CPS-SSA*-1 unemployed males in 1976 whose income in 1975 was below the poverty level.

[b]All earnings are expressed in 1982 dollars. The numbers in parentheses are the standard errors. The number of observations refer only to 1975 and 1978. In the other years there are fewer observations. The sample of treatments is smaller than the sample of controls because treatments still in Supported Work as of January 1978 are excluded from the sample, and in the young high school target group there were by design more controls than treatments.

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE *PSID* AND THE *CPS-SSA*[a,b]

		NSW Treatment Earnings Less Comparison Group Earnings				Difference in Differences: Difference in Earnings Growth 1975–78 Treatments Less Comparisons		Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975–78		
		Pre-Training Year, 1975		Post-Training Year, 1978						
Name of Comparison Group[d]	Comparison Group Earnings Growth 1975–78 (1)	Unad-justed (2)	Ad-justed[c] (3)	Unad-justed (4)	Ad-justed[c] (5)	Without Age (6)	With Age (7)	Unad-justed (8)	Ad-justed[c] (9)	Controlling for All Observed Variables and Pre-Training Earnings (10)
Controls	$2,063	$39	$ − 21	$886	$798	$847	$856	$897	$802	$662
	(325)	(383)	(378)	(476)	(472)	(560)	(558)	(467)	(467)	(506)
PSID-1	$2,043	− $15,997	− $7,624	− $15,578	− $8,067	$425	− $749	− $2,380	− $2,119	− $1,228
	(237)	(795)	(851)	(913)	(990)	(650)	(692)	(680)	(746)	(896)
PSID-2	$6,071	− $4,503	− $3,669	− $4,020	− $3,482	$484	− $650	− $1,364	− $1,694	− $792
	(637)	(608)	(757)	(781)	(935)	(738)	(850)	(729)	(878)	(1024)
PSID-3	($3,322	($455	$455	$697	− $509	$242	− $1,325	$629	− $552	$397
	(780)	(539)	(704)	(760)	(967)	(884)	(1078)	(757)	(967)	(1103)
CPS-SSA-1	$1,196	− $10,585	− $4,654	− $8,870	− $4,416	$1,714	$195	− $1,543	− $1,102	− $805
	(61)	(539)	(509)	(562)	(557)	(452)	(441)	(426)	(450)	(484)
CPS-SSA-2	$2,684	− $4,321	− $1,824	− $4,095	− $1,675	$226	− $488	− $1,850	− $782	− $319
	(229)	(450)	(535)	(537)	(672)	(539)	(530)	(497)	(621)	(761)
CPS-SSA-3	$4,548	$337	$878	− $1,300	$224	− $1,637	− $1,388	− $1,396	$17	$1,466
	(409)	(343)	(447)	(590)	(766)	(631)	(655)	(582)	(761)	(984)

[a]The columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is $886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

[b]Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

[c]The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

[d]See Table 3 for definitions of the comparison groups.

Lalonde Reactionaries

- Initial Reactions
  - Heckman and Hotz (1985): *Specification testing gets it right*
  - Heckman, Ichimura, and Todd (1997): *Better data helps* (same survey or admin data sources; same labor market for treated, controls)
- Dehejia and Wahba (1999): *The propensity score answers the Lalonde challenge*
  - Smith and Todd (2001, 2005): No, it doesn't; the DW sample is special
  - Dehejia (2005): Yes, it does
  - *MHE*: You don't need to keep (the) score, but you must get the controls right
  - Kline (2011): a fully-interacted regression model does well
- Cook, Shadish and Wong (2008): *Well-designed* observational studies (e.g., RD) come close to an RCT benchmark
  - Card, Kluve, and Weber (2010; 2015) find that more recent (mostly European) non-experimental training evaluations seem to generate average effects about the same as randomized trials

CIA Recap

- Training evals focus on effect of treatment on the treated (TOT),

$$E[Y_{1i} - Y_{0i}|D_i = 1] = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]$$

- The earnings differential by training status is a biased measure of TOT unless D_i and Y_{0i} are independent:

$$\begin{aligned} & E\left[Y_i|D_i = 1\right] - E\left[Y_i|D_i = 0\right] \\ = \quad & E\left[Y_{1i} - Y_{0i}|D_i = 1\right] + \left\{E\left[Y_{0i}|D_i = 1\right] - E\left[Y_{0i}|D_i = 0\right]\right\} \end{aligned}$$

- Assume that, conditional on X_i, training is as good as randomly assigned (CIA):

$$\{Y_{0i}, Y_{1i}\} \amalg D_i|X_i$$

- Use the CIA to reveal counterfactual $E[Y_{0i}|X_i, D_i = 1]$, we have,

$$\begin{aligned} \delta_{TOT} &= E\left\{E\left[Y_{1i}|X_i, D_i = 1\right] - E\left[Y_{0i}|X_i, D_i = 0\right]|D_i = 1\right\} \quad (1) \\ &= E[\delta_X|D_i = 1] \end{aligned}$$

where δ_X is cell-specific treatment effect

The P-Score Steps In

- What if X_i is many/multi/continuous? Matching now requires grouping or parametric assumptions, a fact that induces bias (Abadie and Imbens 2006; 2011)
  - How to pick controls when full non-parametric control is impractical?
- The OVB formula is regression's golden rule: coefficients on D_i are unaffected by omission of vars uncorrelated with D_i
- The *propensity score theorem* (Rosenbaum and Rubin, 1983) extends the regression OVB idea to nonparametric conditioning
- If potential outcomes are independent of treatment status conditional on vector X_i, then the CIA holds conditional on the (scalar) propensity score, $p(\mathrm{X}_i) \equiv E[\mathrm{D}_i|\mathrm{X}_i]$
- *Condition on the score and you kill OVB*: controls matter for treatment effects only if they predict treatment

Rosenbaum and Rubin's Insight

Theorem

The Propensity-Score Property
Suppose the conditional independence assumption (CIA) holds for Y_{ji}*;* $j = 0, 1$*. Then,* $\mathrm{Y}_{ji} \amalg \mathrm{D}_i | p(\mathrm{X}_i)$.

Proof.

The claim is true if $P[\mathrm{D}_i = 1 | \mathrm{Y}_{ji}, p(\mathrm{X}_i)]$ does not depend on Y_{ji}.

$$
\begin{aligned}
P[\mathrm{D}_i = 1 | \mathrm{Y}_{ji}, p(\mathrm{X}_i)] &= E[\mathrm{D}_i | \mathrm{Y}_{ji}, p(\mathrm{X}_i)] \\
&= E\{E[\mathrm{D}_i | \mathrm{Y}_{ji}, p(\mathrm{X}_i), \mathrm{X}_i] | \mathrm{Y}_{ji}, p(\mathrm{X}_i)\} \\
&= E\{E[\mathrm{D}_i | \mathrm{Y}_{ji}, \mathrm{X}_i] | \mathrm{Y}_{ji}, p(\mathrm{X}_i)\} \\
&= E\{E[\mathrm{D}_i | \mathrm{X}_i] | \mathrm{Y}_{ji}, p(\mathrm{X}_i)\}, \text{by the CIA.}
\end{aligned}
$$

But $E\{E[\mathrm{D}_i | \mathrm{X}_i] | \mathrm{Y}_{ji}, p(\mathrm{X}_i)\} = E\{p(\mathrm{X}_i) | \mathrm{Y}_{ji}, p(\mathrm{X}_i)\}$, which is just $p(\mathrm{X}_i)$. $\square$

Using the Score

- Direct propensity-score matching works like covariate matching, except that we match on the score-as-covariate. By the propensity score theorem and the CIA, $E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i} | \mathrm{D}_i = 1]$ is

$$E\left\{E[\mathrm{Y}_i | p(\mathrm{X}_i), \mathrm{D}_i = 1] - E[\mathrm{Y}_i | p(\mathrm{X}_i), \mathrm{D}_i = 0] | \mathrm{D}_i = 1\right\}$$

- Stratify on an estimate of $p(\mathrm{X}_i)$ and substitute conditional sample averages for expectations, or match each treated observation to controls with similar values of the propensity score (both of these approaches appear in DW99)
- Alternately, a model-based or non-parametric estimate of $e_{Di} \equiv E[\mathrm{Y}_i | p(\mathrm{X}_i), \mathrm{D}_i]$ can be substituted for these conditional mean functions, replacing the outer expectation with a sum (as in Heckman, Ichimura, and Todd, 1998)
- **DW Tables 1, 3, and 4**: pretty impressive! (tho the DW subsample isn't quite the Lalonde sample)

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]

NOTE: Standard errors are in parentheses. Standard error on difference in means with RE74 subset/treated is given in brackets. Age = age in years; Education = number of years of schooling; Black = 1 if black, 0 otherwise; Hispanic = 1 if Hispanic, 0 otherwise; No degree = 1 if no high school degree, 0 otherwise; Married = 1 if married, 0 otherwise; REx = earnings in calendar year 19x.

[a] NSW sample as constructed by Lalonde (1986).

[b] The subset of the Lalonde sample for which RE74 is available.

[c] Definition of comparison groups (Lalonde 1986):

PSID-1: All male household heads under age 55 who did not classify themselves as retired in 1975.

PSID-2: Selects from PSID-1 all men who were not working when surveyed in the spring of 1976.

PSID-3: Selects from PSID-2 all men who were not working in 1975.

CPS-1: All CPS males under age 55.

CPS-2: Selects from CPS-1 all males who were not working when surveyed in March 1976.

CPS-3: Selects from CPS-2 all the unemployed males in 1976 whose income in 1975 was below the poverty level.

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)

[a] Least squares regression: RE78 on a constant, a treatment indicator, age, age^2, education, no degree, black, Hispanic, RE74, RE75.

[b] Least squares regression of RE78 on a quadratic on the estimated propensity score and a treatment indicator, for observations used under stratification; see note (g).

[c] Number of observations refers to the actual number of comparison and treatment units used for (3)–(5); namely, all treatment units and those comparison units whose estimated propensity score is greater than the minimum, and less than the maximum, estimated propensity score for the treatment group.

[d] Weighted least squares: treatment observations weighted as 1, and control observations weighted by the number of times they are matched to a treatment observation [same covariates as (a)].

Propensity scores are estimated using the logistic model, with specifications as follows:

[e] PSID-1: Prob (T_i = 1) = F(age, age^2, education, education^2, married, no degree, black, Hispanic, RE74, RE75, RE74^2, RE75^2, u74*black).

[f] PSID-2 and PSID-3: Prob (T_i = 1) = F(age, age^2, education, education^2, no degree, married, black, Hispanic, RE74, RE74^2, RE75, RE75^2, u74, u75).

[g] CPS-1, CPS-2, and CPS-3: Prob (T_i = 1) = F(age, age^2, education, education^2, no degree, married, black, Hispanic, RE74, RE75, u74, u75, education*RE74, age^3).

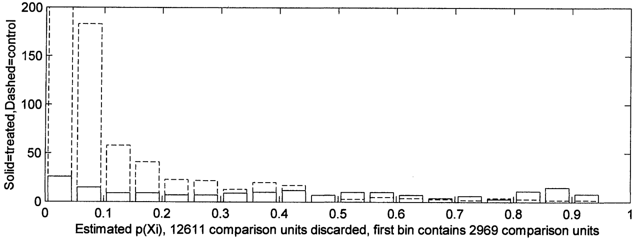

Figure 2. Histogram of the Estimated Propensity Score for NSW Treated Units and CPS Comparison Units. The 12,611 CPS units estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The first bin c 2,969 CPS units. There is minimal overlap between the two groups, but the overlap is greater than in Figure 1; only one bin (.45–.5) cont comparison units, and there are 35 treated and 7 comparison units with an estimated propensity score greater than .8.

Table 4. Sample Means of Characteristics for Matched Control Samples

Matched samples	*No. of observations*	*Age*	*Education*	*Black*	*Hispanic*	*No degree*	*Married*	*RE74 (U.S. $)*	*RE75 (U.S. $)*
NSW	185	25.81	10.35	.84	.06	.71	.19	2,096	1,532
MPSID-1	56	26.39	10.62	.86	.02	.55	.15	1,794	1,126
		[2.56]	[.63]	[.13]	[.06]	[.13]	[.12]	[1,406]	[1,146]
MPSID-2	49	25.32	11.10	.89	.02	.57	.19	1,599	2,225
		[2.63]	[.83]	[.14]	[.08]	[.16]	[.16]	[1,905]	[1,228]
MPSID-3	30	26.86	10.96	.91	.01	.52	.25	1,386	1,863
		[2.97]	[.84]	[.13]	[.08]	[.16]	[.16]	[1,680]	[1,494]
MCPS-1	119	26.91	10.52	.86	.04	.64	.19	2,110	1,396
		[1.25]	[.32]	[.06]	[.04]	[.07]	[.06]	[841]	[563]
MCPS-2	87	26.21	10.21	.85	.04	.68	.20	1,758	1,204
		[1.43]	[.37]	[.08]	[.05]	[.09]	.08	[896]	[661]
MCPS-3	63	25.94	10.69	.87	.06	.53	.13	2,709	1,587
		[1.68]	[.48]	[.09]	[.06]	[.10]	[.09]	[1,285]	[760]

NOTE: Standard error on the difference in means with NSW sample is given in brackets.
MPSID1–3 and MCPS1–3 are the subsamples of PSID1–3 and CPS1–3 that are matched to the treatment group.

Do We *Need* the Score?

- *MHE* argues that CIA-based identification turns more on what you control for than how. See OLS estimates in **MHE Table 3.3.3**
- Kline (2011) goes one better, using OLS methods to construct TOT. Recall that:

$$E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i} | \mathrm{D}_i = 1] = E[\mathrm{Y}_{1i} | \mathrm{D}_i = 1] - E[\mathrm{Y}_{0i} | \mathrm{D}_i = 1]$$

  The trick, of course, is how to produce a credible $E[\mathrm{Y}_{0i} | \mathrm{D}_i = 1]$
- Under selection on observables

$$E[Y_{0i} | X_i, D_i = 1] = E[Y_{0i} | X_i, D_i = 0] = E[Y_i | X_i, D_i = 0]$$

- Suppose we model $E[Y_i | X_i, D_i = 0] = X_i' \beta_0$. Iterating expectations:

$$E[Y_{0i} | D_i = 1] = E[X_i | D_i = 1]' \beta_0$$

- The linear reweighting estimator is the sample analog of

$$E[\mathrm{Y}_{1i} - \mathrm{Y}_{0i} | \mathrm{D}_i = 1] = E[Y_i | D_i = 1] - E[X_i | D_i = 1]' \beta_0,$$

  generating impressive **results in the NSW data**

Table 3.3.3
Regression estimates of NSW training effects using alternative controls

	Full Comparison Samples			P-Score Screened Comparison Samples	
Specification	NSW (1)	CPS-1 (2)	CPS-3 (3)	CPS-1 (4)	CPS-3 (5)
Raw difference	1,794	−8,498	−635		
	(633)	(712)	(657)		
Demographic controls	1,670	−3,437	771	−3,361	890
	(639)	(710)	(837)	(811)	(884)
				[139/497]	[154/154]
1975 earnings	1,750	−78	−91	No	166
	(632)	(537)	(641)	obs.	(644)
				[0/0]	[183/427]
Demographics, 1975 earnings	1,636	623	1,010	1,201	1,050
	(638)	(558)	(822)	(722)	(861)
				[149/357]	[157/162]
Demographics, 1974 and 1975 earnings	1,676	794	1,369	1,362	649
	(639)	(548)	(809)	(708)	(853)
				[151/352]	[147/157]

Notes: The table reports regression estimates of training effects using the Dehejia-Wahba (1999) data with alternative sets of controls. The demographic controls are age, years of schooling, and dummies for black, Hispanic, high school dropout, and married. Standard errors are reported in parentheses. Observation counts are reported in brackets [treated/control]. There are

Table 1 assesses this question empirically by comparing treatment effect estimates generated by each estimator using the observational CPS-3 controls and the experimental NSW controls.[13]

Table 1 - Estimated Impact of NSW on Men's 1978 Earnings		
Estimator/Control Group	CPS-3	NSW
Raw Difference	−$635	$1794
	(677)	(671)
OLS	$1369	$1676
	(739)	(677)
Logistic Reweighting*	$1440	$1808
	(863)	(705)
Blinder-Oaxaca	$1701	$1785
	(841)	(677)
Sample Size	614	445
Note: Heteroscedasticity robust standard errors in parentheses.		
*Reweighting standard errors computed from 1,000 bootstrap replications.		

form estimates of the odds of treatment. The latter approach can be thought of as indirectly approximating the unknown odds via a different set of basis functions, albeit a set that imposes the side constraint that the odds are nonnegative. Whether, in the presence of misspecification, the imposition of this side constraint yields a better approximation to the counterfactual of interest is an empirical question and will depend on the data generating process.

Despite its allowance of negative weights, the Blinder-Oaxaca estimator has several features to commend it. It is easily implemented in unbalanced designs with few treated units and many controls and allows for straightforward computation of standard errors and regression diagnostics. It is consistent if either the linear model for the potential outcomes or the implicit log-logistic model for the propensity score are correct. And unlike standard reweighting estimators, the B-O weights yield exact covariate balance and are finite sample unbiased for the counterfactual under proper specification of the out-

'Metrics Mysteries

- Hahn (1998): Full covariate matching attains the semiparametric efficiency bound for estimation of ATE and TOT under the CIA (analogous to the MLE Cramer-Rao bound)
  - Matching on the score doesn't hit the bound; it's therefore inefficient
  - The score plays no role in the likelihood for ATE, though knowledge of the score can help (via weighting) with the estimation of TOT
- Angrist and Hahn (2004): These asymptotic results use a conventional sequence that fixes the number of covariates, letting the sample size grow to infinity in each covariate cell
  - An alternative (Bekker-type) sequence that fixes obs-per-cell shows that score conditioning generates finite-sample efficiency gains
- Cattaneo, Jansson, and Newey (JASA 2017) use a similar sequence to highlight finite-sample problems with robust standard errors in models with many covs

Wait! What About Score-Weighting?

- By the CIA and Horvitz-Thompson (1952), we can construct unconditional potential outcome means using

$$E\left[\frac{\mathrm{Y}_i\mathrm{D}_i}{p(\mathrm{X}_i)}\right] = E[\mathrm{Y}_{1i}]$$
$$E\left[\frac{\mathrm{Y}_i(1-\mathrm{D}_i)}{(1-p(\mathrm{X}_i))}\right] = E[\mathrm{Y}_{0i}]$$

generating HIR's (2003) efficient score-based estimand for ATE (with continuous covs):

$$\begin{aligned} E[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}] &= E\left[\frac{\mathrm{Y}_i\mathrm{D}_i}{p(\mathrm{X}_i)} - \frac{\mathrm{Y}_i(1-\mathrm{D}_i)}{1-p(\mathrm{X}_i)}\right] \\ &= E\left[\frac{(\mathrm{D}_i-p(\mathrm{X}_i))\mathrm{Y}_i}{p(\mathrm{X}_i)(1-p(\mathrm{X}_i))}\right] \end{aligned} \tag{2}$$

while score-weighting for TOT is:

$$E[\mathrm{Y}_{1i}-\mathrm{Y}_{0i}|\mathrm{D}_i=1] = E\left[\frac{(\mathrm{D}_i-p(\mathrm{X}_i))\mathrm{Y}_i}{(1-p(\mathrm{X}_i))E(\mathrm{D}_i)}\right] \tag{3}$$

Regression vs the Score

- The Angrist (1998) regression estimand, δ_R, can be written:

$$\delta_R = \frac{C(\tilde{\mathrm{D}}_i, \tilde{\mathrm{Y}}_i)}{V(\tilde{\mathrm{D}}_i)} = \frac{E[(\mathrm{D}_i - p(\mathrm{X}_i))\mathrm{Y}_i]}{E[p(\mathrm{X}_i)(1 - p(\mathrm{X}_i))]} \tag{4}$$

- Compare this with HIR's (2003) set of efficient score-based estimands,

$$E\left\{g_i\left[\frac{\mathrm{Y}_i\mathrm{D}_i}{p(\mathrm{X}_i)} - \frac{\mathrm{Y}_i(1-\mathrm{D}_i)}{(1-p(\mathrm{X}_i))}\right]\right\} = E\left\{g_i\left[\frac{(\mathrm{D}_i - p(\mathrm{X}_i))\mathrm{Y}_i}{p(\mathrm{X}_i)(1-p(\mathrm{X}_i))}\right]\right\},$$

where $g_i \equiv g(\mathrm{X}_i)$ is a known weighting function

1. For ATE, set $g(\mathrm{X}_i) = 1$
2. For TOT, set $g(\mathrm{X}_i) = \frac{p(\mathrm{X}_i)}{E[\mathrm{D}_i]}$
3. For regression set $g(\mathrm{X}_i) = \frac{p(\mathrm{X}_i)(1-p(\mathrm{X}_i))}{E[p(\mathrm{X}_i)(1-p(\mathrm{X}_i))]}$

- Knowing the score tells us *what* to control for; regression vs. matching implementation ... may not be weighty

To Know the Score Isn't to Love It

- Abadie and Imbens (2015) note that, in practice, when matching on the score, we must use an estimate
  - Does this affect inference? If so, how?
  - Heckman, Ichimura, and Todd (1998), HIR (2003), and others looked at this for semiparametric estimators; the properties of simple score matching schemes have remained, until recently, unexplored
- Answers
  - Matching on the estimated score affects limiting distributions
  - Estimates of ATE are made more precise by matching on an estimated score instead of a known score; *treating the score as known is therefore conservative for ATE*
  - Use of an estimated score for TOT can go move SEs either way
- Other score frontiers
  - Research design meets market design (Abdulkadiroglu, Angrist, Narita, and Pathak, 2017; TBD, after IV)
  - Residual balancing (Athey, Imbens, and Wager, 2018)

FEATURE STORY

In Memory of Robert J. Lalonde (1958-2018)

Robert J. Lalonde, professor at the University of Chicago Harris School of Public Policy, was a passionate scholar who cared about the effects that his work could have on people's lives. He was an important contributor to research focusing on workplace issues, education, and the economic effects of immigration, as well as many other workforce-related subjects. His work will continue to have transformative impact.

He will be deeply missed by those who connected with him, either personally or professionally, throughout his life.

In Memory of Robert J. Lalonde (1958-2018)

Robert Lalonde's friends and colleagues reflect on his many contributions to economics and the broader community in a BFI statement found here.

Posted: 1/19/18