PROPENSITY SCORE-MATCHING METHODS FOR NONEXPERIMENTAL CAUSAL STUDIES

Rajeev H. Dehejia and Sadek Wahba*

Abstract—This paper considers causal inference and sample selection bias in nonexperimental settings in which (i) few units in the nonexperimental comparison group are comparable to the treatment units, and (ii) selecting a subset of comparison units similar to the treatment units is difficult because units must be compared across a high-dimensional set of pretreatment characteristics. We discuss the use of propensity score-matching methods, and implement them using data from the National Supported Work experiment. Following LaLonde (1986), we pair the experimental treated units with nonexperimental comparison units from the CPS and PSID, and compare the estimates of the treatment effect obtained using our methods to the benchmark results from the experiment. For both comparison groups, we show that the methods succeed in focusing attention on the small subset of the comparison units comparable to the treated units and, hence, in alleviating the bias due to systematic differences between the treated and comparison units.

I. Introduction

An important problem of causal inference is how to estimate treatment effects in observational studies, situations (like an experiment) in which a group of units is exposed to a well-defined treatment, but (unlike an experiment) no systematic methods of experimental design are used to maintain a control group. It is well recognized that the estimate of a causal effect obtained by comparing a treatment group with a nonexperimental comparison group could be biased because of problems such as self-selection or some systematic judgment by the researcher in selecting units to be assigned to the treatment. This paper discusses the use of propensity score-matching methods to correct for sample selection bias due to observable differences between the treatment and comparison groups.

Matching involves pairing treatment and comparison units that are similar in terms of their observable characteristics. When the relevant differences between any two units are captured in the observable (pretreatment) covariates, which occurs when outcomes are independent of assignment to treatment conditional on pretreatment covariates, matching methods can yield an unbiased estimate of the treatment impact.[1] The first generation of matching methods paired observations based on either a single variable or weighting several variables. (See, inter alia, Bassi (1984), Cave and Bos (1995), Czajka et al. (1992), Cochran and Rubin (1973), Raynor (1983), Rosenbaum (1995), Rubin (1973, 1979), Westat (1981), and studies cited by Barnow (1987).)

The motivation for focusing on propensity score-matching methods is that, in many applications of interest, the dimensionality of the observable characteristics is high. With a small number of characteristics (for example, two binary variables), matching is straightforward (one would group units in four cells). However, when there are many variables, it is difficult to determine along which dimensions to match units or which weighting scheme to adopt. Propensity score-matching methods, as we demonstrate, are especially useful under such circumstances because they provide a natural weighting scheme that yields unbiased estimates of the treatment impact.

The key contribution of this paper is to discuss and apply propensity score-matching methods, which are new to the economics literature. (Previous papers include Dehejia and Wahba (1999), Heckman et al. (1996, 1998), Heckman, Ichimura, and Todd (1997, 1998). See Friedlander, Greenberg, and Robins (1997) for a review.) This paper differs from Dehejia and Wahba (1999) by focusing on matching methods in detail, and it complements the Heckman et al. papers by discussing a different array of matching estimators in the context of a different data set.

An important feature of our method is that, after units are matched, the unmatched comparison units are discarded and are not directly used in estimating the treatment impact. Our approach has two motivations. First, in some settings of interest, data on the outcome variable for the comparison group are costly to obtain. For example, in economics, some data sets provide outcome information for only one year; if the outcome of interest takes place in a later period, possibly thousands of comparison units have to be linked across data sets or resurveyed. In such settings, the ability to obtain the needed data for a subset of relevant comparison units, discarding the irrelevant potential comparison units, is extremely valuable. Second, even if information on the outcome is available for all comparison units (as it is in our data), the process of searching for the best subset from the comparison group reveals the extent of overlap between the treatment and comparison groups in terms of pretreatment characteristics. Because methods that use the full set of

Received for publication February 12, 1998. Revision accepted for publication January 24, 2001.

* Columbia University and Morgan Stanley, respectively.

Previous versions of this paper were circulated under the title "An Oversampling Algorithm for Nonexperimental Causal Studies with Incomplete Matching and Missing Outcome Variables" (1995) and as National Bureau of Economic Research working paper no. 6829. We thank Robert Moffitt and two referees for detailed comments and suggestions that have improved the paper. We are grateful to Gary Chamberlain, Guido Imbens, and Donald Rubin for their support and encouragement, and greatly appreciate comments from Joshua Angrist, George Cave, and Jeff Smith. Special thanks are due to Robert LaLonde for providing, and helping to reconstruct, the data from his 1986 study. Valuable comments were received from seminar participants at Harvard, MIT, and the Manpower Demonstration Research Corporation. Any remaining errors are the authors' responsibility.

[1] More precisely, to estimate the treatment impact on the treated, the outcome in the untreated state must be independent of the treatment assignment.

© 2002 by the President and Fellows of Harvard College and the Massachusetts Institute of Technology

Downloaded from http://direct.mit.edu/rest/article-pdf/84/1/151/1613304/0034653023173331982.pdf by AMERICAN UNIV LIBRARY user on 21 December 2024

comparison units extrapolate or smooth across the treatment and comparison groups, it is extremely useful to know how many of the comparison units are in fact comparable and hence how much smoothing one's estimator is expected to perform.

The data we use, obtained from LaLonde (1986), are from the National Supported Work (NSW) Demonstration, a labor market experiment in which participants were randomized between treatment (on-the-job training lasting between nine months and a year) and control groups. Following LaLonde, we use the experimental controls to obtain a benchmark estimate for the treatment impact and then set them aside, wedding the treated units from the experiment to comparison units from the Population Survey of Income Dynamics (PSID) and the Current Population Survey (CPS).[2] We compare estimates obtained using our nonexperimental methods to the experimental benchmark. We show that most of the nonexperimental comparison units are not good matches for the treated group. We succeed in selecting the comparison units that are most comparable to the treated units and in replicating the benchmark treatment impact.

The paper is organized as follows. In section II, we discuss the theory behind our estimation strategy. In section III, we discuss propensity score-matching methods. In section IV, we describe the NSW data, which we then use in section V to implement our matching procedures. Section VI tests the matching assumption and examines the sensitivity of our estimates to the specification of the propensity score. Section VII concludes the paper.

II. Matching Methods

A. The Role of Randomization

A cause is viewed as a manipulation or treatment that brings about a change in the variable of interest, compared to some baseline, called the control (Cox, 1992; Holland, 1986). The basic problem in identifying a causal effect is that the variable of interest is observed under either the treatment or control regimes, but never both.

Formally, let i index the population under consideration. Y_{i1} is the value of the variable of interest when unit i is subject to treatment (1), and Y_{i0} is the value of the same variable when the unit is exposed to the control (0). The treatment effect for a single unit, τ_i, is defined as $\tau_i = Y_{i1} - Y_{i0}$. The primary treatment effect of interest in nonexperimental settings is the expected treatment effect for the treated population; hence

$$\begin{aligned}\tau|_{T=1} &= E(\tau_i|T_i = 1)\\ &= E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 1),\end{aligned}$$

where $T_i = 1$ $(= 0)$ if the ith unit was assigned to treatment (control).[3] The problem of unobservability is summarized by the fact that we can estimate $E(Y_{i1}|T_i = 1)$, but not $E(Y_{i0}|T_i = 1)$.

The difference, $\tau^e = E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 0)$, can be estimated, but it is potentially a biased estimator of τ. Intuitively, if Y_{i0} for the treated and comparison units systematically differs, then in observing only Y_{i0} for the comparison group we do not correctly estimate Y_{i0} for the treated group. Such bias is of paramount concern in nonexperimental studies. The role of randomization is to prevent this:

$$Y_{i1}, Y_{i0} \perp\!\!\!\perp T_i \Rightarrow E(Y_{i0}|T_i = 0) = E(Y_{i0}|T_i = 1) = E(Y_i|T_i = 0),$$

where $Y_i = T_iY_{i1} + (1 - T_i)Y_{i0}$ (the observed value of the outcome) and $\perp\!\!\!\perp$ is the symbol for independence. The treated and control groups do not systematically differ from each other, making the conditioning on T_i in the expectation unnecessary (ignorable treatment assignment, in the terminology of Rubin (1977)), and yielding $\tau|_{T=1} = \tau^e$.[4]

B. Exact Matching on Covariates

To substitute for the absence of experimental control units, we assume that data can be obtained for a set of potential comparison units, which are not necessarily drawn from the same population as the treated units but for whom we observe the same set of pretreatment covariates, X_i. The following proposition extends the framework of the previous section to nonexperimental settings:

Proposition 1 (Rubin, 1977). If for each unit we observe a vector of covariates X_i and $Y_{i0} \perp\!\!\!\perp T_i|X_i$, $\forall i$, then the population treatment effect for the treated, $\tau|_{T=1}$, is identified: it is equal to the treatment effect conditional on covariates and on assignment to treatment, $\tau|_{T=1,X}$, averaged over the distribution $X|T_i = 1$[5].

[2] Fraker and Maynard (1987) also conduct an evaluation of nonexperimental methods using the NSW data. Their findings were similar to LaLonde's.

[3] In a nonexperimental setting, the treatment and comparison samples are either drawn from distinct groups or are nonrandom samples from a common population. In the former case, typically the interest is the treatment impact for the group from which the treatment sample is drawn. In the latter case, the interest could be in knowing the treatment effect for the subpopulation from which the treatment sample is drawn or the treatment effect for the full population from which both treatment and comparison samples were drawn. In contrast, in a randomized experiment, the treatment and control samples are randomly drawn from the same population, and thus the treatment effect for the treated group is identical to the treatment effect for the untreated group.

[4] We are also implicitly making what is sometimes called the stable-unit-treatment-value assumption (Rubin, 1980, 1986). This amounts to the assumption that $Y_{i1}(Y_{i0})$ does not depend upon which units other than i were assigned to the treatment group; that is, there are no within-group spillovers or general equilibrium effects.

[5] Randomization implies $Y_{i1}, Y_{i0} \perp\!\!\!\perp T_i$, but $Y_{i0} \perp\!\!\!\perp T_i|X_i$ is all that is required to estimate the treatment effect on the treated. The stronger assumption, $Y_{i1}, Y_{i0} \perp\!\!\!\perp T_i|X_i$, would be needed to identify the treatment effect on the comparison group or the overall average. Note that we are estimating the treatment effect for the treatment group as it exists at the time of analysis. We are not estimating any program entry or exit effects that might arise if

Downloaded from http://direct.mit.edu/rest/article-pdf/84/1/151/1613304/0034653023173319 82.pdf by AMERICAN UNIV LIBRARY user on 21 December 2024

Intuitively, this assumes that, conditioning on observable covariates, we can take assignment to treatment to have been random and that, in particular, unobservables play no role in the treatment assignment; comparing two individuals with the same observable characteristics, one of whom was treated and one of whom was not, is by proposition 1 like comparing those two individuals in a randomized experiment. Under this assumption, the conditional treatment effect, $\tau|_{T=1}$, is estimated by first estimating $\tau|_{T=1,X}$ and then averaging over the distribution of X conditional on $T = 1$.

One way to estimate this equation would be by matching units on their vector of covariates, X_i. In principle, we could stratify the data into subgroups (or bins), each defined by a particular value of X; within each bin, this amounts to conditioning on X. The limitation of this method is that it relies on a sufficiently rich comparison group so that no bin containing a treated unit is without a comparison unit. For example, if all n variables are dichotomous, the number of possible values for the vector X will be 2^n. Clearly, as the number of variables increases, the number of cells increases exponentially, increasing the difficulty of finding exact matches for each of the treated units.

C. *Propensity Score and Dimensionality Reduction*

Rosenbaum and Rubin (1983, 1985a, b) suggest the use of the propensity score—the probability of receiving treatment conditional on covariates—to reduce the dimensionality of the matching problem discussed in the previous section.

Proposition 2 (Rosenbaum and Rubin, 1983). Let $p(X_i)$ be the probability of a unit i having been assigned to treatment, defined as $p(X_i) \equiv \Pr(T_i = 1|X_i) = E(T_i|X_i)$. Then,

$$(Y_{i1}, Y_{i0}) \perp\!\!\!\perp T_i | X_i \Rightarrow (Y_{i1}, Y_{i0}) \perp\!\!\!\perp T_i | p(X_i).$$

Proposition 3. $\tau|_{T=1} = E_{p(X)}[(\tau|_{T=1,p(X)})|T_i = 1]$.

Thus, the conditional independence result extends to the use of the propensity score, as does by immediate implication our result on the computation of the conditional treatment effect, now $\tau|_{T=1,p(X)}$. The point of using the propensity score is that it substantially reduces the dimensionality of the problem, allowing us to condition on a scalar variable rather than in a general n-space.

III. Propensity Score-Matching Algorithms

In the discussion that follows, we assume that the propensity score is known, which of course it is not. The appendix discusses a straightforward method for estimating it.[6]

Matching on the propensity score is essentially a weighting scheme, which determines what weights are placed on comparison units when computing the estimated treatment effect:

$$\hat{\tau}|_{T=1} = \frac{1}{|N|} \sum_{i \in N} \left(Y_i - \frac{1}{|J_i|} \sum_{j \in J_i} Y_j \right),$$

where N is the treatment group, $|N|$ the number of units in the treatment group, J_i is the set of comparison units matched to treatment unit i (see Heckman, Ichimura, and Todd (1998), who discuss more general weighting schemes), and $|J_i|$ is the number of comparison units in J_i.

This estimator follows from proposition 3. Expectations are replaced by sample means, and we condition on $p(X_i)$ by matching each treatment unit i to a set of comparison units, J_i, with a similar propensity score. Taken literally, conditioning on $p(X_i)$ implies exact matching on $p(X_i)$. This is difficult in practice, so the objective becomes to match treated units to comparison units whose propensity scores are sufficiently close to consider the conditioning on $p(X_i)$ in proposition 3 to be approximately valid.

Three issues arise in implementing matching: whether or not to match with replacement, how many comparison units to match to each treated unit, and finally which matching method to choose. We consider each in turn.

Matching with replacement minimizes the propensity-score distance between the matched comparison units and the treatment unit: each treatment unit can be matched to the nearest comparison unit, even if a comparison unit is matched more than once. This is beneficial in terms of bias reduction. In contrast, by matching without replacement, when there are few comparison units similar to the treated units, we may be forced to match treated units to comparison units that are quite different in terms of the estimated propensity score. This increases bias, but it could improve the precision of the estimates. An additional complication of matching without replacement is that the results are potentially sensitive to the order in which the treatment units are matched (Rosenbaum, 1995).

The question of how many comparison units to match with each treatment unit is closely related. By using a single comparison unit for each treatment unit, we ensure the smallest propensity-score distance between the treatment and comparison units. By using more comparison units, one increases the precision of the estimates, but at the cost of increased bias. One method of selecting a set of comparison units is the nearest-neighbor method, which selects the m comparison units whose propensity scores are closest to the treated unit in question. Another method is caliper matching, which uses all of the comparison units within a predefined propensity score radius (or "caliper"). A benefit of

the treatment were made more widely available. Estimation of such effects would require additional data as described by Moffitt (1992).

[6] Standard errors should adjust for the estimation error in the propensity score and the variation that it induces in the matching process. In the application, we use bootstrap standard errors. Heckman, Ichimura, and Todd (1998) provide asymptotic standard errors for propensity score estimators, but in their application paper, Heckman, Ichimura, and Todd (1997) also use bootstrap standard errors.

Downloaded from http://direct.mit.edu/rest/article-pdf/84/1/151/1613304/0034653023173319 82.pdf by AMERICAN UNIV LIBRARY user on 21 December 2024

caliper matching is that it uses only as many comparison units as are available within the calipers, allowing for the use of extra (fewer) units when good matches are (not) available.

In the application that follows, we consider a range of simple estimators. For matching without replacement, we consider low-to-high, high-to-low, and random matching. In these methods, the treated units are ranked (from lowest to highest or highest to lowest propensity score, or randomly). The highest-ranked unit is matched first, and the matched comparison unit is removed from further matching. For matching with replacement, we consider single-nearest-neighbor matching and caliper matching for a range of calipers. In addition to using a weighted difference in means to estimate the treatment effect, we also consider a weighted regression using the treatment and matched comparison units, with the comparison units weighted by the number of times that they are matched to a treated unit. A regression can potentially improve the precision of the estimates.

The question that remains is which method to select in practice. In general, this depends on the data in question, and in particular on the degree of overlap between the treatment and comparison groups in terms of the propensity score. When there is substantial overlap in the distribution of the propensity score between the comparison and treatment groups, most of the matching algorithms will yield similar results. When the treatment and comparison units are very different, finding a satisfactory match by matching without replacement can be very problematic. In particular, if there are only a handful of comparison units comparable to the treated units, then once these comparison units have been matched, the remaining treated units will have to be matched to comparison units that are very different. In such settings, matching with replacement is the natural choice. If there are no comparison units for a range of propensity scores, then for that range the treatment effect could not be estimated. The application that follows will further clarify the choices that the researcher faces in practice.

IV. The Data

A. The National Supported Work Program

The NSW was a U.S. federally and privately funded program that aimed to provide work experience for individuals who had faced economic and social problems prior to enrollment in the program (Hollister, Kemper, and Maynard, 1984; Manpower Demonstration Research Corporation, 1983).[7] Candidates for the experiment were selected on the basis of eligibility criteria, and then were either randomly assigned to, or excluded from, the training program.

[7] Four groups were targeted: Women on Aid to Families with Dependent Children (AFDC), former addicts, former offenders, and young school dropouts. Several reports extensively document the NSW program. For a general summary of the findings, see Manpower Demonstration Research Corporation (1983).

TABLE 1.—SAMPLE MEANS AND STANDARD ERRORS OF COVARIATES FOR MALE NSW PARTICIPANTS

National Supported Work Sample (Treatment and Control)		
	Dehejia-Wahba Sample	
Variable	Treatment	Control
Age	25.81 (0.52)	25.05 (0.45)
Years of schooling	10.35 (0.15)	10.09 (0.1)
Proportion of school dropouts	0.71 (0.03)	0.83 (0.02)
Proportion of blacks	0.84 (0.03)	0.83 (0.02)
Proportion of Hispanic	0.06 (0.017)	0.10 (0.019)
Proportion married	0.19 (0.03)	0.15 (0.02)
Number of children	0.41 (0.07)	0.37 (0.06)
No-show variable	0 (0)	n/a
Month of assignment (Jan. 1978 = 0)	18.49 (0.36)	17.86 (0.35)
Real earnings 12 months before training	1,689 (235)	1,425 (182)
Real earnings 24 months before training	2,096 (359)	2,107 (353)
Hours worked 1 year before training	294 (36)	243 (27)
Hours worked 2 years before training	306 (46)	267 (37)
Sample size	185	260

Table 1 provides the characteristics of the sample we use, LaLonde's male sample (185 treated and 260 control observations).[8] The table highlights the role of randomization: the distribution of the covariates for the treatment and control groups are not significantly different. We use the two nonexperimental comparison groups constructed by LaLonde (1986), drawn from the CPS and PSID.[9]

B. Distribution of the Treatment and Comparison Samples

Tables 2 and 3 (rows 1 and 2) present the sample characteristics of the two comparison groups and the treatment group. The differences are striking: the PSID and CPS sample units are eight to nine years older than those in the NSW group, their ethnic composition is different, and they have on average completed high school degrees, whereas NSW participants were by and large high school dropouts, and, most dramatically, pretreatment earnings are much higher for the comparison units than for the treated units, by more than $10,000. A more synoptic way to view these differences is to use the estimated propensity score as a summary statistic. Using the method outlined in the appendix, we estimate the propensity score for the two composite samples (NSW-CPS and NSW-PSID), incorporating the covariates linearly and with some higher-order terms.

[8] The data we use are a subsample of the data used in LaLonde (1986). The analysis in LaLonde is based on one year of pretreatment earnings. But, as Ashenfelter (1978) and Ashenfelter and Card (1985) suggest, the use of more than one year of pretreatment earnings is key in accurately estimating the treatment effect, because many people who volunteer for training programs experience a drop in their earnings just prior to entering the training program. Using the LaLonde sample of 297 treated and 425 control units, we exclude the observations for which earnings in 1974 could not be obtained, thus arriving at a reduced sample of 185 treated observations and 260 control observations. Because we obtain this subset by looking at pretreatment covariates, we do not disturb the balance in observed and unobserved characteristics between the experimental treated and control groups. See Dehejia and Wahba (1999) for a comparison of the two samples.

[9] These are the CPS-1 and PSID-1 comparison groups from LaLonde's paper.

Downloaded from http://direct.mit.edu/rest/article-pdf/84/1/151/1613304/0034653023173331982.pdf by AMERICAN UNIV LIBRARY user on 21 December 2024

TABLE 2.—SAMPLE CHARACTERISTICS AND ESTIMATED IMPACTS FROM THE NSW AND CPS SAMPLES

Control Sample	No. of Observations	Mean Propensity Score[A]	Age	School	Black	Hispanic	No Degree	Married	RE74	RE75	U74	U75	Treatment Effect (Diff. in Means)	Regression Treatment Effect
NSW	185	0.37	25.82	10.35	0.84	0.06	0.71	0.19	2095	1532	0.29	0.40	1794[B]	1672[C]
													(633)	(638)
Full CPS	15992	0.01	33.23	12.03	0.07	0.07	0.30	0.71	14017	13651	0.88	0.89	−8498	1066
		(0.02)[D]	(0.53)	(0.15)	(0.03)	(0.02)	(0.03)	(0.03)	(367)	(248)	(0.03)	(0.04)	(583)[E]	(554)
Without replacement:														
Random	185	0.32	25.26	10.30	0.84	0.06	0.65	0.22	2305	1687	0.37	0.51	1559	1651
		(0.03)	(0.79)	(0.23)	(0.04)	(0.03)	(0.05)	(0.04)	(495)	(341)	(0.05)	(0.05)	(733)	(709)
Low to high	185	0.32	25.23	10.28	0.84	0.06	0.66	0.22	2286	1687	0.37	0.51	1605	1681
		(0.03)	(0.79)	(0.23)	(0.04)	(0.03)	(0.05)	(0.04)	(495)	(341)	(0.05)	(0.05)	(730)	(704)
High to low	185	0.32	25.26	10.30	0.84	0.06	0.65	0.22	2305	1687	0.37	0.51	1559	1651
		(0.03)	(0.79)	(0.23)	(0.04)	(0.03)	(0.05)	(0.04)	(495)	(341)	(0.05)	(0.05)	(733)	(709)
With replacement:														
Nearest neighbor	119	0.37	25.36	10.31	0.84	0.06	0.69	0.17	2407	1516	0.35	0.49	1360	1375
		(0.03)	(1.04)	(0.31)	(0.06)	(0.04)	(0.07)	(0.06)	(727)	(506)	(0.07)	(0.07)	(913)	(907)
Caliper, $\delta = 0.00001$	325	0.37	25.26	10.31	0.84	0.07	0.69	0.17	2424	1509	0.36	0.50	1119	1142
		(0.03)	(1.03)	(0.30)	(0.06)	(0.04)	(0.07)	(0.06)	(845)	(647)	(0.06)	(0.06)	(875)	(874)
Caliper, $\delta = 0.00005$	1043	0.37	25.29	10.28	0.84	0.07	0.69	0.17	2305	1523	0.35	0.49	1158	1139
		(0.02)	(1.03)	(0.32)	(0.05)	(0.04)	(0.06)	(0.06)	(877)	(675)	(0.06)	(0.60)	(852)	(851)
Caliper, $\delta = 0.0001$	1731	0.37	25.19	10.36	0.84	0.07	0.69	0.17	2213	1545	0.34	0.50	1122	1119
		(0.02)	(1.03)	(0.31)	(0.05)	(0.04)	(0.06)	(0.06)	(890)	(701)	(0.06)	(0.06)	(850)	(843)

Variables: Age, age of participant; School, number of school years; Black, 1 if black, 0 otherwise; Hisp, 1 if Hispanic, 0 otherwise; No degree, 1 if participant had no school degrees, 0 otherwise; Married, 1 if married, 0 otherwise; RE74, real earnings (1982US$) in 1974; RE75, real earnings (1982US$) in 1975; U74, 1 if unemployed in 1974, 0 otherwise; U75, 1 if unemployed in 1975, 0 otherwise; and RE78, real earnings (1982US$) in 1978.

(A) The propensity score is estimated using a logit of treatment status on: Age, Age^2, Age^3, School, School^2, Married, No degree, Black, Hisp, RE74, RE75, U74, U75, School · RE74.

(B) The treatment effect for the NSW sample is estimated using the experimental control group.

(C) The regression treatment effect controls for all covariates linearly. For matching with replacement, weighted least squares is used, where treatment units are weighted at 1 and the weight for a control is the number of times it is matched to a treatment unit.

(D) The standard error applies to the difference in means between the matched and the NSW sample, except in the last two columns, where the standard error applies to the treatment effect.

(E) Standard errors for the treatment effect and regression treatment effect are computed using a bootstrap with 500 replications.

Figures 1 and 2 provide a simple diagnostic on the data examined, plotting the histograms of the estimated propensity scores for the NSW-CPS and NSW-PSID samples. Note that the histograms do not include the comparison units (11,168 units for the CPS and 1,254 units for the PSID) whose estimated propensity score is less

TABLE 3.—SAMPLE CHARACTERISTICS AND ESTIMATED IMPACTS FROM THE NSW AND PSID SAMPLES

Control Sample	No. of Observations	Mean Propensity Score[A]	Age	School	Black	Hispanic	No Degree	Married	RE74 US$	RE75 US$	U74	U75	Treatment Effect (Diff. in Means)	Regression Treatment Effect
NSW	185	0.37	25.82	10.35	0.84	0.06	0.71	0.19	2095	1532	0.29	0.40	1794[B]	1672[C]
													(633)	(638)
Full PSID	2490	0.02	34.85	12.12	0.25	0.03	0.31	0.87	19429	19063	0.10	0.09	−15205	4
		(0.02)[D]	(0.57)	(0.16)	(0.03)	(0.02)	(0.03)	(0.03)	(449)	(361)	(0.04)	(0.03)	(657)[E]	(1014)
Without replacement:														
Random	185	0.25	29.17	10.30	0.68	0.07	0.60	0.52	4659	3263	0.40	0.40	−916	77
		(0.03)	(0.90)	(0.25)	(0.04)	(0.03)	(0.05)	(0.05)	(554)	(361)	(0.05)	(0.05)	(1035)	(983)
Low to high	185	0.25	29.17	10.30	0.68	0.07	0.60	0.52	4659	3263	0.40	0.40	−916	77
		(0.03)	(0.90)	(0.25)	(0.04)	(0.03)	(0.05)	(0.05)	(554)	(361)	(0.05)	(0.05)	(1135)	(983)
High to low	185	0.25	29.17	10.30	0.68	0.07	0.60	0.52	4659	3263	0.40	0.40	−916	77
		(0.03)	(0.90)	(0.25)	(0.04)	(0.03)	(0.05)	(0.05)	(554)	(361)	(0.05)	(0.05)	(1135)	(983)
With replacement:														
Nearest Neighbor	56	0.70	24.81	10.72	0.78	0.09	0.53	0.14	2206	1801	0.54	0.69	1890	2315
		(0.07)	(1.78)	(0.54)	(0.11)	(0.05)	(0.12)	(0.11)	(1248)	(963)	(0.11)	(0.11)	(1202)	(1131)
Caliper, $\delta = 0.00001$	85	0.70	24.85	10.72	0.78	0.09	0.53	0.13	2216	1819	0.54	0.69	1893	2327
		(0.08)	(1.80)	(0.56)	(0.12)	(0.05)	(0.12)	(0.12)	(1859)	(1896)	(0.10)	(0.11)	(1198)	(1129)
Caliper, $\delta = 0.00005$	193	0.70	24.83	10.72	0.78	0.09	0.53	0.14	2247	1778	0.54	0.69	1928	2349
		(0.06)	(2.17)	(0.60)	(0.11)	(0.04)	(0.11)	(0.10)	(1983)	(1869)	(0.09)	(0.09)	(1196)	(1121)
Caliper, $\delta = 0.0001$	337	0.70	24.92	10.73	0.78	0.09	0.53	0.14	2228	1763	0.54	0.70	1973	2411
		(0.05)	(2.30)	(0.67)	(0.11)	(0.04)	(0.11)	(0.09)	(1965)	(1777)	(0.07)	(0.08)	(1191)	(1122)
Caliper, $\delta = 0.001$	2021	0.70	24.98	10.74	0.79	0.09	0.53	0.13	2398	1882	0.53	0.69	1824	2333
		(0.03)	(2.37)	(0.70)	(0.09)	(0.04)	(0.10)	(0.07)	(2950)	(2943)	(0.06)	(0.06)	(1187)	(1101)

(A) The propensity score is estimated using a logit of treatment status on: Age, Age^2, School, School^2, Married, No degree, Black, Hisp, RE74, RE74^2, RE75, RE75^2, U74, U75, U74 · Hisp.

(B) The treatment effect for the NSW sample is estimated using the experimental control group.

(C) The regression treatment effect controls for all covariates linearly. For matching with replacement, weighted least squares is used, where treatment units are weighted at 1 and the weight for a control is the number of times it is matched to a treatment unit.

(D) The standard error applies to the difference in means between the matched and the NSW sample, except in the last two columns, where the standard error applies to the treatment effect.

(E) Standard errors for the treatment effect and regression treatment effect are computed using a bootstrap with 500 replications.

Downloaded from http://direct.mit.edu/rest/article-pdf/84/1/151/1613304/0034653023173319 82.pdf by AMERICAN UNIV LIBRARY user on 21 December 2024

FIGURE 1.—HISTOGRAM OF ESTIMATED PROPENSITY SCORE, NSW AND CPS

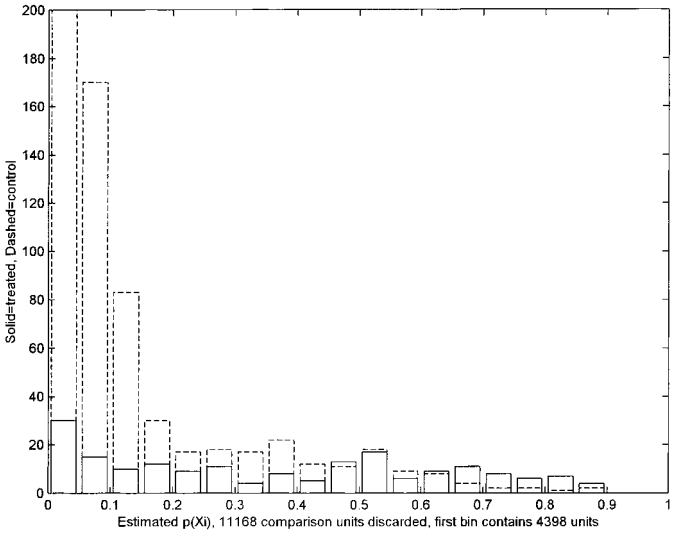
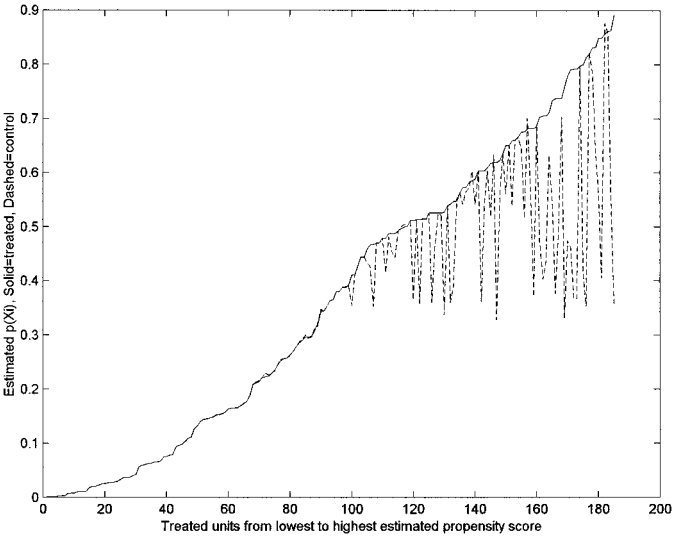


FIGURE 3.—PROPENSITY SCORE FOR TREATED AND MATCHED COMPARISON UNITS, RANDOM WITHOUT REPLACEMENT



than the minimum estimated propensity score for the treated units. As well, the first bins of both diagrams contain most of the remaining comparison units (4,398 for the CPS and 1,007 for the PSID). Hence, it is clear that very few of the comparison units are comparable to the treated units. In fact, one of the strengths of the propensity score method is that it dramatically highlights this fact. In comparing the other bins, we note that the number of comparison units in each bin is approximately equal to the number of treated units in the NSW-CPS sample, but, in the NSW-PSID sample, many of the upper bins have far more treated units than comparison units. This last observation will be important in interpreting the results of the next section.

V. Matching Results

Figures 3 to 6 provide a snapshot of the matching methods described in section III and applied to the NSW-CPS sample, where the horizontal axis displays treated units (indexed from lowest to highest estimated propensity score) and the vertical axis depicts the propensity scores of the treated units and their matched comparison counterparts. (The corresponding figures for the NSW-PSID sample look very similar.) Figures 3 to 5, which consider matching without replacement, share the common feature that the first 100 or so treated units are well matched to their comparison group counterparts: the solid and the dashed lines virtually overlap. But the

FIGURE 2.—HISTOGRAM OF ESTIMATED PROPENSITY SCORE, NSW AND PSID

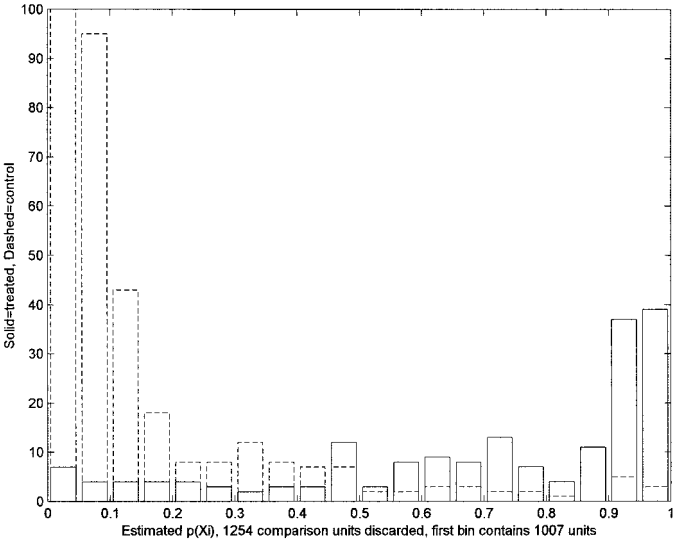
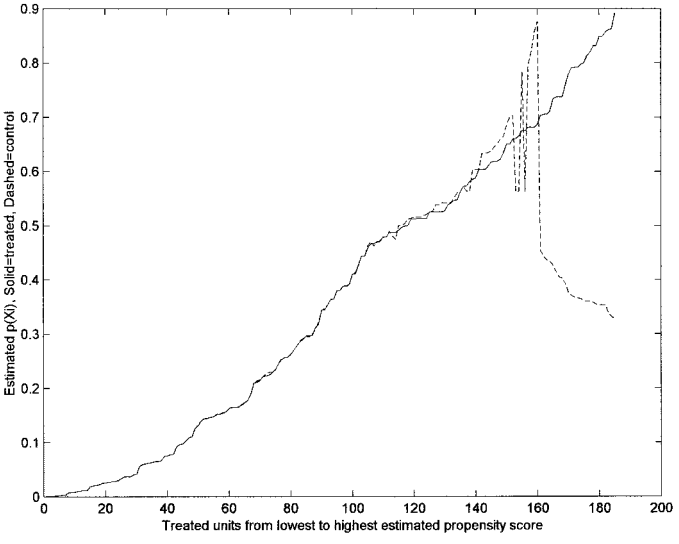


FIGURE 4.—PROPENSITY SCORE FOR TREATED AND MATCHED COMPARISON UNITS, LOWEST TO HIGHEST



Downloaded from http://direct.mit.edu/rest/article-pdf/84/1/151/1613304/0034653023173319 82.pdf by AMERICAN UNIV LIBRARY user on 21 December 2024

FIGURE 5.—PROPENSITY SCORE FOR TREATED AND MATCHED COMPARISON UNITS, HIGHEST TO LOWEST

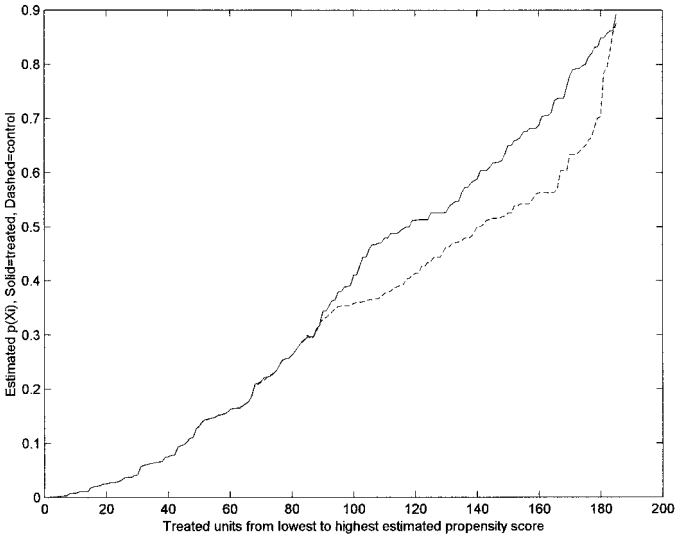
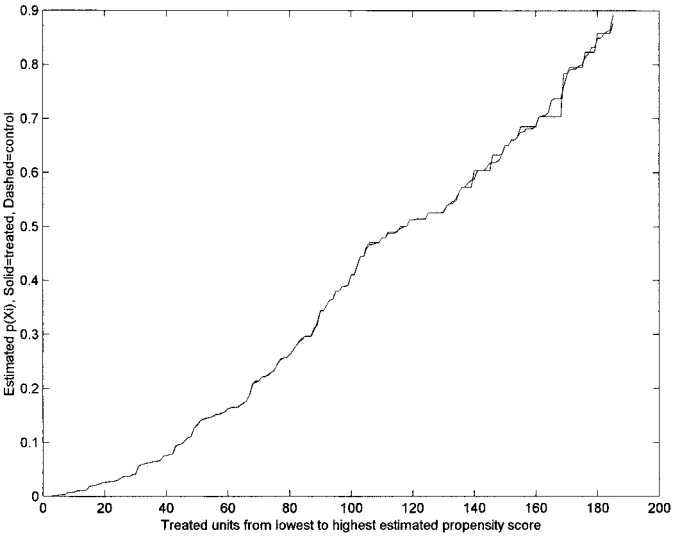


FIGURE 6.—PROPENSITY SCORE FOR TREATED AND MATCHED COMPARISON UNITS, NEAREST MATCH



treated units with estimated propensity scores of 0.4 or higher are not well matched.

In figure 3, units that are randomly selected to be matched earlier find better matches, but those matched later are poorly matched because the few comparison units comparable to the treated units have already been used. Likewise, in figure 4, where units are matched from lowest to highest, treated units in the 140th to 170th positions are forced to use comparison units with ever-higher propensity scores. Finally, for the remaining units (from approximately the 170th position on), the comparison units with high propensity scores are exhausted and matches are found among comparison units with much lower estimated propensity scores. Similarly, when we match from highest to lowest, the quality of matches begins to decline after the first few treated units, until we reach treated units whose propensity score is (approximately) 0.4.

Figure 6 depicts the matching achieved by the nearest-match method.[10] We note immediately that by matching with replacement we are able to avoid the deterioration in the quality of matches noted in figures 3 to 5; the solid and the dashed lines largely coincide. Looking at the line depicting comparison units more carefully, we note that it has flat sections that correspond to ranges in which a single comparison unit is being matched to more than one treated unit. Thus, even though there is a smaller sample size, we are better able to match the distribution of the propensity scores of the treated units.

In table 2, we explore the matched samples and the estimated treatment impacts for the CPS. From rows 1 and 2, we already noted that the CPS sample is very different from the NSW. The aim of matching is to choose subsamples whose characteristics more closely resemble the NSW. Rows 3 to 5 of table 2 depict the matched samples that emerge from matching without replacement. Note that the characteristics of these samples are essentially identical, suggesting that these three methods yield the same comparison groups. (Figures 3 to 5 obscure this fact because they compare the order in which units are matched, not the resulting comparison groups.) The matched samples are much closer to the NSW sample than the full CPS comparison group. The matched CPS group has an age of 25.3 (compared with 25.8 and 33.2 for the NSW and full CPS samples), its ethnic composition is the same as the NSW sample (note especially the difference in the full CPS in terms of the variable Black), no degree and marital status align, and, perhaps most significantly, the pretreatment earnings are similar for both 1974 and 1975.[11] None of the differences between the matched groups and the NSW sample are statistically significant.[12] Looking at the nearest-match and caliper methods, little significant improvement can be discerned, although most of the variables are marginally better matched. This suggests that the observation made regarding figure 1 (that the CPS, in fact, has a

[10] Note that, in implementing this method, if the set of comparison units within a given caliper is empty for a treated unit, we match it to the nearest comparison unit. The alternative is to drop unmatched treated units, but then one would no longer be estimating the treatment effect for the entire treated group.

[11] The matched earnings, like the NSW sample, exhibit the Ashenfelter (1978) "dip" in earnings in the year prior to program participation.

[12] Note that both LaLonde (1986) and Fraker and Maynard (1987) attempt to use "first-generation" matching methods to reduce differences between the treatment and comparison groups. LaLonde creates subsets of CPS-1 and PSID-1 by matching single characteristics (employment status and income). Dehejia and Wahba (1999) demonstrates that significant differences remain between the reduced comparison groups and the treatment group. Fraker and Maynard match on predicted earnings. Their matching method also fails to balance pretreatment characteristics (especially earnings) between the treatment and comparison group. (See Fraker and Maynard (1987, p. 205).)

Downloaded from http://direct.mit.edu/rest/article-pdf/84/1/151/1613304/0034653023173331982.pdf by AMERICAN UNIV LIBRARY user on 21 December 2024

sufficient number of comparison units overlapping with the NSW) is borne out in terms of the matched sample.

Turning to the estimates of the treatment impact, in row 1 we see that the benchmark estimate of the treatment impact from the randomized experiment is $1,794. For the full CPS comparison group, the estimate is −$8,498 using a difference in means and $1,066 using regression adjustment. The raw estimate is very misleading when compared with the benchmark, although the regression-adjusted estimate is better. The matching estimates are closer. For the without-replacement estimators, the estimate ranges from $1,559 to $1,605 for the difference in means and from $1,651 to $1,681 for the regression-adjusted estimator. The nearest-neighbor with-replacement estimates are $1,360 and $1,375. Essentially, these methods succeed by picking out the subset of the CPS that is the best comparison for the NSW. Based on these estimates, one might conclude that matching without replacement is the best strategy. The reason why all the methods perform well is that there is reasonable overlap between the treatment and CPS comparison samples. As we will see, for the PSID comparison group the estimates are very different.

When using caliper matching, a larger comparison group is selected: 325 for a caliper of 0.00001, 1,043 for a caliper of 0.0001, and 1,731 for a caliper of 0.0001. In terms of the characteristics of the sample, few significant differences are observed, although we know that the quality of the matches in terms of the propensity score is poorer. This is reflected in the estimated treatment impact which ranges from $1,122 to $1,149.

Using the PSID sample (table 3), somewhat different conclusions are reached. Like the CPS, the PSID sample is very different from the NSW sample. Unlike the CPS, the matched-without-replacement samples are not fully comparable to the NSW. They are reasonably comparable in terms of age, schooling, and ethnicity, but, in terms of pretreatment earnings, we observe a large (and statistically significant) difference. As a result, it is not surprising that the estimates of the treatment impact, both by a difference in means and through regression adjustment, are far from the experimental benchmark (ranging from $−916 to $77). In contrast, the matched-with-replacement samples use even fewer (56) comparison units, but they are able to match the pretreatment earnings of the NSW sample and the other variables as well. This corresponds to our observation regarding figure 2, namely that there are very few comparison units in the PSID that are similar to units in the NSW; when this is the case, we expect more sensitivity to the method used to match observations, and we expect matching with replacement to perform better. The treatment impact as estimated by the nearest-neighbor method through a difference in means ($1,890) is very similar to the experimental benchmark, but differs by $425 when estimated through regression adjustment (although it is still closer than the estimates in rows 1 to 4). The difference in the two estimates is less surprising when we consider the sample size involved: we are using only 56 of the 2,490 potential comparison units from the PSID. For the PSID, caliper matching also performs well. The estimates range from $1,824 to $2,411. Slightly lower standard errors are achieved than nearest-neighbor matching.

In conclusion, propensity score-matching methods are able to yield reasonably accurate estimates of the treatment impact, especially when contrasted with the range of estimates that emerged in LaLonde's paper. By selecting an appropriate subset from the comparison group, a simple difference in means yields an estimate of the treatment effect close to the experimental benchmark. The choice among matching methods becomes important when there is minimal overlap between the treatment and comparison groups. When there is minimal overlap, matching with replacement emerges as a better choice. In principle, caliper matching can also improve standard errors relative to nearest-neighbor matching, although at the cost of greater bias. At least in our application, the benefits of caliper matching were limited. When there is greater overlap, the without-replacement estimators perform as well as the nearest-neighbor method, and their standard errors are somewhat lower than the nearest-neighbor method, so, when many comparison units overlap with the treatment group, matching without replacement is probably a better choice.

VI. Testing

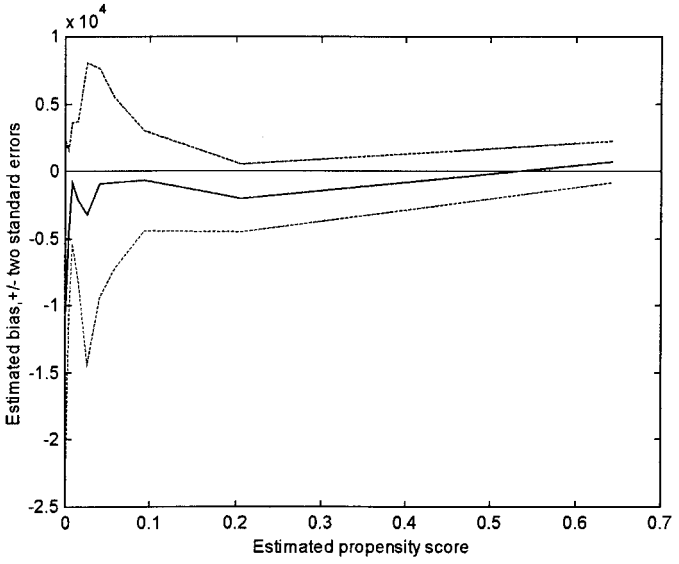
A. Testing the Matching Assumption

The special structure of the data we use allows us to test the assumption that underlies propensity score matching. Because we have both an experimental control group (which we use to estimate the experimental benchmark estimate in row 1 of tables 2 and 3) and two nonexperimental comparison groups, we can test the assumption that, conditional on the propensity score, earnings in the nontreated state are independent of assignment to treatment (Heckman et al., 1998; Heckman, Ichimura, and Todd, 1997). In practice, this amounts to comparing earnings for the experimental control group with earnings for the two comparison groups using the propensity score. We apply the propensity score specifications from section V to the composite sample of NSW control units and CPS (or PSID) comparison units. Following Heckman et al. (1998), we compute the bias within strata defined on the propensity score.

The bias estimates—earnings for the experimental control group less earnings for the nonexperimental comparison group conditional on the estimated propensity score—are presented graphically in figures 7 and 8. For both the CPS and PSID, we see a range of bias estimates that are particularly large for low values of the estimated propensity score. This group represents those who are least likely to have been in the treatment group, and, based on tables 2 and 3, this group has much higher earnings than those in the NSW. But none of the bias estimates are statistically significant.

Downloaded from http://direct.mit.edu/rest/article-pdf/84/1/151/1613304/0034653023173319 82.pdf by AMERICAN UNIV LIBRARY user on 21 December 2024

FIGURE 7.—BIAS ESTIMATES, CPS



Of course, in practice a researcher will not be able to perform such tests, but it is a useful exercise when possible. It confirms that matching succeeds because the nontreated earnings of the comparison and control groups are not statistically significantly different, conditional on the estimated propensity score.

B. Testing Sensitivity to the Specification of the Propensity Score

One potential limitation of propensity score methods is the need to estimate the propensity score. In LaLonde's (1986) paper, one of the cautionary findings was the sensitivity of the nonexperimental estimators to the specification adopted. The appendix suggests a simple method to choose a specification for the propensity score. In table 4, we consider sensitivity of the estimates to the choice of specification.

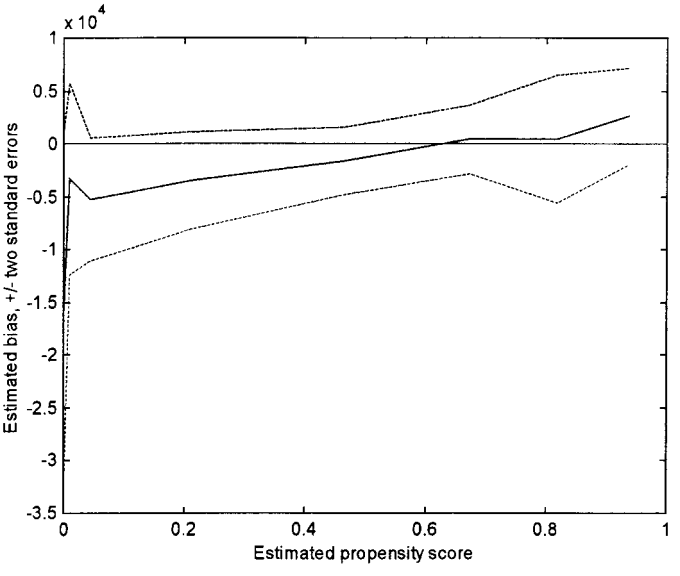
In table 4, we consider dropping in succession the interactions and cubes, the indicators for unemployment, and finally squares of covariates in the specification. The final specification for both samples contains the covariates linearly. For the CPS, the estimate bounces from \$1,037 to \$1,874, and for the PSID from \$1,004 to \$1,845. The estimates are not particularly sensitive, especially compared to the variability of estimators in LaLonde's original paper. Furthermore, a researcher who did not have the benefit of the experimental benchmark estimate would choose the full-specification estimates because (as explained in the appendix) these specifications succeed in balancing all the observed covariates, conditional on the estimated propensity score.

VII. Conclusion

This paper has presented a propensity score-matching method that is able to yield accurate estimates of the treatment effect in nonexperimental settings in which the treated group differs substantially from the pool of potential comparison units. The method is able to pare the large comparison group down to the relevant comparisons without using information on outcomes, thereby, if necessary, allowing outcome data to be collected only for the relevant subset of comparison units. Of course, the quality of the estimate that emerges from the resulting comparison is limited by the overall quality of the comparison group that is used. Using LaLonde's (1986) data set, we demonstrate the ability of this technique to work in practice. Even though in a typical application the researcher would not have the benefit of checking his or her estimate against the experimental-benchmark estimate, the conclusion of our analysis is that it is extremely valuable to check the comparability of the treatment and comparison units in terms of pretreatment characteristics, which the researcher can check in most applications.

In particular, the propensity score method dramatically highlights the fact that most of the comparison units are very different from the treated units. In addition to this, when there are very few comparison units remaining after having discarded the irrelevant comparison units, the choice of matching algorithm becomes important. We demonstrate that, when there are a sufficient number of relevant comparison units (in our application, when using the CPS), the nearest-match method does no worse than the matching-without-replacement methods that would typically be applied, and, in situations in which there are very few relevant comparison units (in our application, when using the PSID), matching with replacement fares better than the alternatives. Extensions of matching with replacement (caliper matching), although interesting in principal, were of little value in our application.

FIGURE 8.—ESTIMATED BIAS, PSID



Downloaded from http://direct.mit.edu/rest/article-pdf/84/1/151/1613304/0034653023173331982.pdf by AMERICAN UNIV LIBRARY user on 21 December 2024

TABLE 4.—SENSITIVITY OF MATCHING WITH REPLACEMENT TO THE SPECIFICATION OF THE ESTIMATED PROPENSITY SCORE

Specification	Number of Observations	Difference-in-Means Treatment Effect (Standard Error)[B]	Regression Treatment Effect[A] (Standard Error)[B]
CPS			
Full specification	119	1360 (633)	1375 (638)
Dropping interactions and cubes	124	1037 (1005)	1109 (966)
Dropping indicators:	142	1874 (911)	1529 (928)
Dropping squares	134	1637 (944)	1705 (965)
PSID			
Full specification	56	1890 (1202)	2315 (1131)
Dropping interactions and cubes	61	1004 (2412)	1729 (3621)
Dropping indicators:	65	1845 (1720)	1592 (1624)
Dropping squares	69	1428 (1126)	1400 (1157)

For all specifications other than the full specifications, some covariates are not balanced across the treatment and comparison groups.

(A) The regression treatment effect controls for all covariates linearly. Weighted least squares is used where treatment units are weighted at 1 and the weight for a control is the number of times it is matched to a treatment unit.

(B) Standard errors for the treatment effect and regression treatment effect are computed using a bootstrap with 500 replications.

It is something of an irony that the data that we use were originally employed by LaLonde (1986) to demonstrate the failure of standard nonexperimental methods in accurately estimating the treatment effect. Using matching methods on both of his samples, we are able to replicate the experimental benchmark, but beyond this we focus attention on the value of flexibly adjusting for observable differences between the treatment and comparison groups. The process of trying to find a subset of the PSID group comparable to the NSW units demonstrated that the PSID is a poor comparison group, especially when compared to the CPS.

Given the success of propensity score methods in this application, how might a researcher choose which method to use in other settings? An important issue is whether the assumption of selection on observable covariates is valid, or whether the selection process depends on variables that are unobserved (Heckman and Robb, 1985). Only when the researcher is comfortable with the former assumption do propensity score methods come into play. Even then, the researcher still can use standard regression techniques with suitably flexible functional forms (Cain, 1975; Barnow, Cain, and Goldberger, 1980). The methods that we discuss in this paper should be viewed as a complement to the standard techniques in the researcher's arsenal. By starting with a propensity score analysis, the researcher will have a better sense of the extent to which the treatment and comparison groups overlap and consequently of how sensitive estimates will be to the choice of functional form.

REFERENCES

Ashenfelter, Orley, "Estimating the Effects of Training Programs on Earnings," this REVIEW 60:1 (February 1978), 47–57.

Ashenfelter, Orley, and D. Card, "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," this REVIEW 67:4 (November 1985), 648–660.

Barnow, Burt, "The Impact of CETA Programs on Earnings: A Review of the Literature," *Journal of Human Resources* 22:2 (Spring 1987), 157–193.

Barnow, Burt, Glen Cain, and Arthur Goldberger, "Issues in the Analysis of Selectivity Bias," (pp. 42–59), in Ernst W. Stromsdorfer and George Farkas (Eds.), *Evaluation Studies Review Annual,* 5 (Beverly Hills: Sage Publications, 1980).

Bassi, Laurie, "Estimating the Effects of Training Programs with Nonrandom Selection," this REVIEW 66:1 (February 1984), 36–43.

Cain, Glen, "Regression and Selection Models to Improve Nonexperimental Comparisons" (pp. 297–317), in C. A. Bennett and A. A. Lumsdaine (Eds.), *Evaluation and Experiments: Some Critical Issues in Assessing Social Programs* (New York: Academic Press, 1975).

Cave, George, and Hans Bos, "The Value of a GED in a Choice-Based Experimental Sample," (New York: Manpower Demonstration Research Corporation, 1995).

Cochran, W. G., and D. B. Rubin, "Controlling Bias in Observational Studies: A Review," *Sankhya,* ser. A, 35:4 (December 1973), 417–446.

Cox, D. R., "Causality: Some Statistical Aspects," *Journal of the Royal Statistical Society,* series A, 155, part 2 (1992), 291–301.

Czajka, John, Sharon M. Hirabayashi, Roderick J. A. Little, and Donald B. Rubin, "Projecting from Advance Data Using Propensity Modeling: An Application to Income and Tax Statistics," *Journal of Business and Economic Statistics* 10:2 (April 1992), 117–131.

Dehejia, Rajeev, and Sadek Wahba, "An Oversampling Algorithm for Non-experimental Causal Studies with Incomplete Matching and Missing Outcome Variables," Harvard University mimeograph (1995).

———, "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," *Journal of the American Statistical Association* 94:448 (December 1999), 1053–1062.

Fraker, T., and R. Maynard, "Evaluating Comparison Group Designs with Employment-Related Programs," *Journal of Human Resources* 22 (1987), 194–227.

Friedlander, Daniel, David Greenberg, and Philip Robins, "Evaluating Government Training Programs for the Economically Disadvantaged," *Journal of Economic Literature* 35:4 (December 1997), 1809–1855.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd, "Sources of Selection Bias in Evaluating Social Programs: An Interpretation of Conventional Measures and Evidence on the Effectiveness of Matching as a Program Evaluation Method," *Proceedings of the National Academy of Sciences* 93:23 (November 1996), 13416–13420.

———, "Characterizing Selection Bias Using Experimental Data," *Econometrica* 66:5 (September 1998), 1017–1098.

Heckman, James, Hidehiko Ichimura, and Petra Todd, "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies* 64:4 (October 1997), 605–654.

———, "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65:2 (April 1998), 261–294.

Heckman, James, and Richard Robb, "Alternative Methods for Evaluating the Impact of Interventions" (pp. 63–113), in J. Heckman and B. Singer (Eds.), *Longitudinal Analysis of Labor Market Data,* Econometric Society Monograph, No. 10 (Cambridge, UK: Cambridge University Press, 1985).

Downloaded from http://direct.mit.edu/rest/article-pdf/84/1/151/1613304/0034653023173319 82.pdf by AMERICAN UNIV LIBRARY user on 21 December 2024

Holland, Paul W., "Statistics and Causal Inference," *Journal of the American Statistical Association* 81:396 (December 1986), 945–960.

Hollister, Robinson, Peter Kemper, and Rebecca Maynard, *The National Supported Work Demonstration* (Madison, WI: University of Wisconsin Press, 1984).

LaLonde, Robert, "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review* 76:4 (September 1986), 604–620.

Manpower Demonstration Research Corporation, *Summary and Findings of the National Supported Work Demonstration* (Cambridge, MA: Ballinger, 1983).

Moffitt, Robert, "Evaluation Methods for Program Entry Effects" (pp. 231–252), in Charles Manski and Irwin Garfinkel (Eds.), *Evaluating Welfare and Training Programs* (Cambridge, MA: Harvard University Press, 1992).

Raynor, W. J., "Caliper Pair-Matching on a Continuous Variable in Case Control Studies," *Communications in Statistics: Theory and Methods* 12:13 (June 1983), 1499–1509.

Rosenbaum, Paul, *Observational Studies* (New York: Springer Verlag, 1995).

Rosenbaum, P., and D. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70:1 (April 1983), 41–55.

———, "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity," *American Statistician* 39:1 (February 1985a), 33–38.

———, "The Bias Due to Incomplete Matching," *Biometrics* 41 (March 1985b), 103–116.

Rubin, D., "Matching to Remove Bias in Observational Studies," *Biometrics* 29 (March 1973), 159–183.

———, "Assignment to a Treatment Group on the Basis of a Covariate," *Journal of Educational Statistics* 2:1 (Spring 1977), 1–26.

———, "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observation Studies," *Journal of the American Statistical Association* 74:366 (June 1979), 318–328.

———, "Discussion of Randomization Analysis of Experimental Data: The Fisher Randomization Test, by D. Basu," *Journal of the American Statistical Association* 75:371 (September 1980), 591–593.

———, "Discussion of Holland (1986)," *Journal of the American Statistical Association* 81:396 (December 1986), 961–964.

Westat, "Continuous Longitudinal Manpower Survey Net Impact Report No. 1: Impact on 1977 Earnings of New FY 1976 CETA Enrollees in Selected Program Activities," report prepared for U.S. DOL under contract 23-24-75-07 (1981).

APPENDIX: ESTIMATING THE PROPENSITY SCORE

The first step in estimating the treatment effect is to estimate the propensity score. Any standard probability model can be used (for example, logit or probit). It is important to remember that the role of the propensity score is only to reduce the dimensions of the conditioning; as such, it has no behavioral assumptions attached to it. For ease of estimation, most applications in the statistics literature have concentrated on the logit model:

$$\Pr(T_i = 1|X_i) = \frac{e^{\lambda h(X_i)}}{1 + e^{\lambda h(X_i)}},$$

where T_i is the treatment status and $h(X_i)$ is made up of linear and higher-order terms of the covariates on which we condition to obtain an ignorable treatment assignment.[13]

In estimating the propensity score through a probability model, the choice of which interaction or higher-order term to include is determined solely by the need to condition fully on the observable characteristics that make up the assignment mechanism. The following proposition forms the basis of the algorithm we use to estimate the propensity score (Rosenbaum and Rubin, 1983):

Proposition A:

$$X \perp\!\!\!\perp T|p(X).$$

Proof: From the definition of $p(X)$ in proposition 2: $E(T_i|X_i, p(X_i)) = E(T_i|X_i) = p(X_i)$.

The algorithm works as follows. Starting with a parsimonious logistic function with linear covariates to estimate the score, rank all observations by the estimated propensity score (from lowest to highest). Divide the observations into strata such that within each stratum the difference in propensity score for treated and comparison observations is insignificant. Proposition A tells us that within each stratum the distribution of the covariates should be approximately the same across the treated and comparison groups, once the propensity score is controlled for. Within each stratum, we can test for statistically significant differences between the distribution of covariates for treated and comparison units; operationally, *t*-tests on differences in the first moments are often sufficient, but a joint test for the difference in means for all the variables within each stratum could also be performed.[14] When the covariates are not balanced within a particular stratum, the stratum may be too coarsely defined; recall that proposition A deals with observations with an identical propensity score. The solution adopted is to divide the stratum into finer strata and test again for no difference in the distribution of the covariates within the finer strata. If, however, some covariates remain unbalanced for many strata, the score may be poorly estimated, which suggests that additional terms (interaction or higher-order terms) of the unbalanced covariates should be added to the logistic specification to control better for these characteristics. This procedure is repeated for each given stratum until the covariates are balanced. The algorithm is summarized next.

A Simple Algorithm for Estimating the Propensity Score

1. Start with a parsimonious logit specification to estimate the score.
2. Sort data according to estimated propensity score (ranking from lowest to highest).
3. Stratify all observations such that estimated propensity scores within a stratum for treated and comparison units are close (no significant difference); for example, start by dividing observations into strata of equal score range (0–0.2, . . . , 0.8–1).
4. Statistical test: for all covariates, differences in means across treated and comparison units within each stratum are not significantly different from zero.
   a. If covariates are balanced between treated and comparison observations for all strata, stop.
   b. If covariates are not balanced for some stratum, divide the stratum into finer strata and reevaluate.
   c. If a covariate is not balanced for many strata, modify the logit by adding interaction terms and/or higher-order terms of the covariate and reevaluate.

A key property of this procedure is that it uses a well-defined criterion to determine which interaction terms to use in the estimation, namely those terms that balance the covariates. It also makes no use of the outcome variable, and embodies one of the specification tests proposed by LaLonde (1986) and others in the context of evaluating the impact of training on earnings, namely to test for the regression-adjusted difference in the earnings prior to treatment.

[13] Because we allow for higher-order terms in X, this choice is not very restrictive. By rearranging and taking logs, we obtain $\ln(\Pr(T_i = 1|X_i)/1 - \Pr(T_i = 1|X_i)) = \lambda h(X_i)$. A Taylor-series expansion allows us an arbitrarily precise approximation. See also Rosenbaum and Rubin (1983).

[14] More generally, one can also consider higher moments or interactions.

Downloaded from http://direct.mit.edu/rest/article-pdf/84/1/151/1613304/0034653023173319 82.pdf by AMERICAN UNIV LIBRARY user on 21 December 2024