NBER WORKING PAPER SERIES

PROPENSITY SCORE MATCHING METHODS FOR NON-EXPERIMENTAL CAUSAL STUDIES

Rajeev H. Dehejia
Sadek Wahba

Working Paper 6829
http://www.nber.org/papers/w6829

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 1998

Portions of this paper circulated under the title "An Oversampling Algorithm for Non-experimental Causal Studies with Incomplete Matching and Missing Outcome Variables" (Dehejia and Wahba 1995). We gratefully acknowledge the support and encouragement of Gary Chamberlain, Guido Imbens, and Donald Rubin. We thank Joshua Angrist and George Cave for detailed comments, as well as Robert Lalonde for helpful discussions and for kindly providing 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. The first author acknowledges support from a Social Sciences and Humanities Research Council of Canada grant, and the second author acknowledges support from a World Bank Fellowship. The views expressed here are those of the author and do not reflect those of the National Bureau of Economic Research.

Propensity Score Matching Methods for
Non-experimental Causal Studies
Rajeev H. Dehejia and Sadek Wahba
NBER Working Paper No. 6829
December 1998
JEL No. C81, C14

ABSTRACT

This paper considers causal inference and sample selection bias in non-experimental settings in which: (i) few units in the non-experimental 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 pre-treatment characteristics. We propose the use of propensity score matching methods and implement them using data from the NSW experiment. Following Lalonde (1986), we pair the experimental treated units with non-experimental 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. 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.

Rajeev H. Dehejia
Department of Economics
Columbia University
420 West 118th Street, 1022 IAB
New York, NY 10027
and NBER
dehejia@columbia.edu

Sadek Wahba
Morgan Stanley and Company
1585 Broadway
New York, NY 10036

1. 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 non-experimental 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.

Matching methods have been widely used in the statistics literature to address this problem (see, *inter alia*, Cave and Bos 1995; Czajka, *et al*. 1992; Cochran and Rubin 1973; Raynor 1983; Rosenbaum 1995; Rosenbaum and Rubin 1985a; Rubin 1973, 1979; and Rubin and Thomas 1992), but are relatively new to the economics literature. Matching involves pairing together 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 (pre-treatment) covariates (i.e., outcomes are independent of assignment to treatment, conditional on pre-treatment covariates), matching methods can yield an unbiased estimate of the treatment impact.

This paper makes three contributions to the literature on matching methods. First, we discuss and extend propensity score matching methods, which are new to the economics literature (the only other application we are aware of is Heckman, Ichimura, and Todd 1997; see Friedlander, Greenberg, and Robins 1997 for a review). Second, we show how these methods expose the key issue of the comparability of the treatment and control

groups in terms of their observable characteristics. Third, we show that our methods can succeed in producing accurate estimates of the treatment impact even when there exist very few comparison units that are comparable to the treatment units.

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 (e.g., 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 a unit. Propensity score matching methods, as we demonstrate below, are especially useful under such circumstances, and succeed in yielding accurate estimates of the treatment impact.

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. This contrasts with approaches that use the full set of controls to estimate the treatment impact (e.g., Heckman, Ichimura, and Todd's [1997] kernel-based matching estimator). There are two motivations for our approach. First, in some settings of interest, data on the outcome variable for the control group are costly to obtain. For example, in economics, some data sets only provide outcome information for one year; if the outcome of interest takes place in a later period, possibly thousands of controls have to be linked across data sets or re-surveyed. In such settings, the ability to obtain the needed data for a subset of relevant controls, discarding the irrelevant potential controls, 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 is very revealing of the extent of overlap between the treatment and comparison groups in terms of pre-treatment characteristics. Since methods that use the full set of controls extrapolate or smooth across the treatment and control groups, it is extremely useful to know how many of the controls 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 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 set 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). We compare estimates obtained using our non-experimental methods to the experimental benchmark. We show that we succeed in replicating the benchmark treatment impact and in selecting from the large set of comparison units those which are most comparable to the treated units.

The paper is organized as follows. In Section 2, we discuss the theory behind our estimation strategy. In Section 3, we analyze the shortcomings of the standard matching approach and propose algorithms to deal with problems of incomplete matching. In Section 4, we describe the NSW data, which we then use in Section 5 to implement our matching procedures. Section 6 concludes the paper.

2. Matching Methods

2.1 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 non-experimental settings is the expected treatment effect over the treated population; hence:

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

where T_i=1 (=0) if the i-th unit was assigned to treatment (control).[1] 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 is potentially a biased estimator of τ. Intuitively, if the treated and control units systematically differ in their characteristics, then in observing only Y_{i0} for the control group we do not

[1] In a non-experimental setting, because the treatment and control groups may differ systematically, we must consider them to be drawn from different populations with potentially different treatment effects. In contrast, in a randomized experiment, the treatment and control groups are drawn from the same population. Thus, in an experiment, the treatment effect for the treated is identical to the treatment effect for the

correctly estimate Y_{i0} for the treated group. Such bias is of paramount concern in non-experimental 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 $Y_i = T_i Y_{i1} + (1-T_i)Y_{i0}$ (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 $\tau\ |_{T=1} = \tau^{e}$.[2]

2.2 Exact Matching on Covariates

To substitute for the absence of experimental control units, we assume that data can be obtained for a (large) set of potential controls, which of course 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 non-experimental settings:

Proposition 1 (Rubin 1977): *If for each unit we observe a vector of covariates X_i, and $Y_{i1}, Y_{i0} \perp\!\!\!\perp T_i | X_i$, $\forall i$, then the population treatment effect for the treated, $\tau\ |_{T=1}$, is identified: it is equal to the treatment effect conditional on covariates and assignment to treatment, $\tau\ |_{T=1,X}$, averaged over the distribution $X|T_i=1$.*

---

untreated, and therefore to the population average treatment effect.

[2] We are also implicitly making what is sometimes called the stable-unit-treatment-value assumption (see Rubin 1980, 1986). This amounts to the assumption that Y_{i1} (Y_{i0}) does not depend upon which units other

Proof: $Y_{i1}, Y_{i0} \underline{||}\ T_i \big| X_i \Rightarrow\ E(Y_{i0} \big| X_i, T_i = 1) = E(Y_{i0} \big| X_i, T_i = 0) = E(Y_{i0} \big| X_i)$,

which allows us to write:

$$\begin{aligned}
\tau\big|_{T=1} &= E\big(Y_{i1} | T_i = 1\big) - E\big(Y_{i0} | T_i = 1\big) \\
&= E_X\Big[\Big\{E\big(Y_{i1} \big| X_i, T_i = 1\big) - E\big(Y_{i0} \big| X_i, T_i = 1\big)\Big\} \Big| T_i = 1\Big] \\
&= E_X\Big[\Big\{E\big(Y_{i1} \big| X_i, T_i = 1\big) - E\big(Y_{i0} \big| X_i, T_i = 0\big)\Big\} \Big| T_i = 1\Big] \\
&= E_X\Big[\tau\big|_{T=1,X} \Big| T_i = 1\Big].
\end{aligned}$$

Intuitively, this assumes that, conditioning on observable covariates, we can take assignment to treatment as having 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, $\tau\ |_{T=1,X}$, is estimated by an argument identical to the one used above for ignorable assignment, simply by conditioning on X and then averaging over $X|_{T=1}$ to estimate the treatment effect.

One way to estimate this equation would be through matching units on their vector of covariates, X_i. In principle, we could stratify the data into 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 the availability of sufficiently rich data on controls so that no bin containing a treated unit is without a control. For example, if all n variables are dichotomous, the number of possible values for the vector X will be 2^n. Clearly, as the

---

than i were assigned to the treatment group.

number of variables increases, the number of cells will increase exponentially, reducing the probability of finding exact matches for each of the treated units.

2.3 Propensity Score and Dimensionality Reduction

Rosenbaum and Rubin (1983, 1985a,b) suggest the use of the propensity score -- the conditional probability of receiving treatment given a set of covariates -- to reduce the dimensionality of the matching problem alluded to in the previous section:

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

$$\begin{aligned} &(Y_{i1}, Y_{i0}) \perp\!\!\!\perp T_i \mid X_i \\ \Rightarrow &(Y_{i1}, Y_{i0}) \perp\!\!\!\perp T_i \mid p(X_i). \end{aligned}$$

Proof:

$$\begin{aligned} &E(T_i \mid Y_1, Y_0, p(X)) \\ &= E_X\{E(T_i \mid Y_1, Y_0, X) \mid Y_1, Y_0, p(X)\} \\ &= E_X\{E(T_i \mid X) \mid Y_1, Y_0, p(X)\} \\ &= E_X\{p(X) \mid Y_1, Y_0, p(X)\} \\ &= p(X). \end{aligned}$$

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

Proof: Follows immediately from Propositions 1 and 2.

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 $\tau \mid_{T=1,p(X)}$. The point of using the propensity score is that it sub-

stantially reduces the dimensionality of the problem, allowing us to condition on a scalar variable rather than in a general *n*-space.

3. 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.[3]

Matching on the propensity score follows from Proposition 3. An unbiased estimate of the treatment effect arises from conditioning on $p(X_i)$, which entails exact matching on $p(X_i)$. But it is very rare to find two units with exactly the same propensity score, so the objective becomes to match a treated unit to the control units whose propensity scores are sufficiently close to that of the treated unit to consider them as being approximately the same. In particular, we want them to be close enough to consider the conditioning on $p(X_i)$ in Proposition 3 to be approximately valid.

We define a distance metric, which allows us to seek the nearest match. In the context of matching on the propensity score, the simplest metric is:

$$d(i, J) = \left| p(X_i) - \frac{1}{|J|} \sum_{j \in J} p(X_j) \right|,$$

where i is typically a treated unit and J is a set of control units ($|J|$ denotes the cardinality of J). The objective then would be:

[3] Note that the propensity score is used only to select a subset of the control group, and does not enter directly into the estimation of the treatment effect. Nonetheless, standard errors should adjust for the estimation error in the propensity score, which ours currently do not. As far as we know, the best way to do this for our estimator has not yet been addressed in the literature. Heckman, Ichimura, and Todd (1997) do adjust their asymptotic standard errors for the estimation of the propensity score, though for a different estimator.

$$\min_{m(\bullet)} D = \frac{1}{n}\sum_{i=1}^{n} d\big(i, m(i)\big),$$

where *m(i)* denotes the set of control units matched with the treated unit *i*, and where we sum over the *n* treated units since we are estimating the treatment effect for the treated population. If the treated units are exactly matched to controls, then *D*=0.

When no exact matches are available, matters become more complicated. Examples of simple algorithms applied in the literature (see Rubin 1973, 1979) include ranking the treated observations in descending (or ascending) order of the estimated propensity score, and then matching (without replacement) each treated unit, in turn, to the closest control. An even simpler method involves randomly ordering the treated units, and again matching without replacement.

When there is substantial overlap in the distribution of the propensity score between the control and treatment groups, any matching protocol will usually bring *D* close to its minimum; the randomized protocols are simply easy ways to cut through the problem. But when the control and the treated units are very different, finding a satisfactory match using the standard algorithm can be very problematic. In particular, if there are only a handful of control units comparable to the treated units, then once these controls have been matched, the remaining treated units will have to be matched to controls that are very different. There are two solutions to the problem. Rosenbaum (1995 and references cited therein) considers the use of network flow methods in finding the best matching function, $m(\cdot)$, independent of the order in which units are matched.

In this paper, we explore another approach: matching units *with* replacement. When control units are very different from the treated group, matching with replacement

allows many treated units to be matched to the same control unit. The simplest method is to match each treated unit to the single control unit with the closest propensity score (we call this the nearest-match method). This method selects the smallest possible control group.

But to the extent that our motivation for matching is to condition on $p(X_i)$, we might be willing to admit more than the single best match. In particular, if we consider all units within some tolerance level, δ (chosen by the researcher), to have approximately the same propensity score, then when a treated unit has several controls within a δ -radius, we could use all of these controls. When implementing this method (the radius method), we make one modification, namely that if a treated unit has no control units within a δ -radius, we take the nearest control. Note that like the nearest-match method, a given control may be matched to more than one treated unit.

In switching from the nearest-match to the radius method, we end up using more controls. Adding control units has two effects. The first is to worsen the quality of the match on the propensity score; this follows immediately from the algorithm described earlier. The second is to change the variance of the estimate by using a larger and different sample. The fact that the sample is larger will tend to increase the precision of the estimates, but the larger sample may also embody greater variability; which effect dominates will depend on the application. In essence, then, we face a potential bias-variance tradeoff. In general it is difficult to know which point on the tradeoff is desired, since in applications one does not know the relationship between δ and the resulting bias. In addition to

demonstrating the efficacy of these methods in general, our application also explores this bias-variance tradeoff.

4. The Data

4.1 The National Supported Work Program

The NSW was a U.S. federally funded program which aimed to provide work experience for individuals who had faced economic and social problems prior to enrollment in the program (see Manpower Demonstration Research Corporation 1983).[4] 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. Table 1 provides the characteristics of the sample we use (185 treated and 260 control observations).[5] The table highlights the role of randomization: the distribution of the covariates for the treatment and control groups are not significantly different. We use two non-experimental control groups, drawn from the CPS and PSID (see Lalonde 1986 for further details).

[4] 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).

[5] The data we use are a sub-sample of the data used in Lalonde (1986). The analysis in Lalonde (1986) is based on one year of pre-treatment earnings. But as Ashenfelter (1978) and Ashenfelter and Card (1985) suggest, the use of more than one year of pre-treatment 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 pre-treatment covariates, we do not disturb the balance in observed and unobserved characteristics between the experimental treated and control groups.

4.2 Distribution of the Treatment and Control Samples

Tables 2 and 3 (rows 1 and 2) present the sample characteristics of the two control groups and the treatment group. The differences are striking: the PSID and CPS sample units are 8 to 9 years older than those in the NSW group; their ethnic composition is different; they have on average completed high school degrees, while NSW participants were by and large high school dropouts; and, most dramatically, pre-treatment earnings are much higher for the control 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 (age squared, education squared). 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 controls (11,168 units for the CPS and 1,254 units for the PSID) whose estimated propensity score is less than the minimum estimated propensity score for the treated units. As well, the first bins of both diagrams contain most of the remaining controls (4,398 for the CPS and 1,007 for the PSID). Hence, it is clear that very few of the control 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 controls 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 control units. This last observation will be important in interpreting the results of the next section.

5. Matching Results

Figures 3 to 6 provide a snapshot of the matching methods described in Section 3 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-control counterparts (the corresponding figures for the NSW-PSID sample look very similar). Figures 3 to 5 share the common feature that the first 100 or so treated units are well matched to their control counterparts: the solid and the dashed lines virtually overlap. But the 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 control 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 controls with ever-higher propensity scores. Finally, for the remaining units (from approximately the 170th position on), the controls with high propensity scores are exhausted and matches are found among controls 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. 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 controls more carefully, we note that it has flat sections

not seen on the line for treated units. These flats are exactly the regions in which a single control 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 score 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 population. The aim of matching is to choose sub-samples whose characteristics more closely resemble the NSW population. 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 control groups. (Figures 3 to 5 obscure this fact because they compare the order in which units are matched, not the resulting control groups.) The matched samples are much closer to the NSW sample than the full CPS control 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); Nodegree and marital status align; and, perhaps most significantly, the pre-treatment earnings are similar for both 1974 and 1975. None of the differences between the matched groups and the NSW sample are statistically significant. Looking at the nearest-match and radius 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 sufficient number of controls 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. Using the full CPS control 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, though the regression-adjusted estimate is better. The matching estimates are much closer, most dramatically for the difference in means, where the estimates range from $1,559 to $1,605; the regression-adjusted estimates are similar. The fact that the difference in means and regression-adjusted estimates are very similar to the benchmark and to each other demonstrates the success of this method in selecting a suitable control group.

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 pre-treatment income we observe a large (and statistically significant) difference. As well, the estimates of the treatment impact, both by a difference in means and through regression adjustment, are far from the experimental benchmark. In contrast, the matched-with-replacement samples use even fewer (56) controls, but are able to match the pre-treatment 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 controls 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. The treatment impact as estimated by the nearest-match method through a difference in means ($1,890) is very similar to the experimental

benchmark, but differs by $425 when estimated through regression adjustment (though 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 controls from the PSID. The disappointment, then, is not that the regression estimate is poor, but that there are so few controls comparable to the treated units.

In both Tables 2 and 3 the radius method of matching yields broadly similar results to the nearest-match method. As we increase the radius we use more and more controls. For the CPS we expand the number from 119 to 1,731 (for a radius of δ =0.0001), and for the PSID the number expands from 56 to 337. For both samples the estimates of the treatment impact become worse (the bias increases), and the standard errors do not appreciably decline. As more controls are used, the regression-adjusted treatment impact increasingly differs from the difference-in-means treatment impact, because as the composite sample becomes less and less well-balanced in terms of pre-treatment covariates, controlling for these characteristics has a greater impact. Thus, in this application there seems to be little value in using additional controls beyond the nearest matches; of course, this may differ in other applications.

6. Conclusion

This paper has presented a propensity score matching method that is able to yield accurate estimates of the treatment effect in non-experimental settings where the treated group differs substantially from the pool of potential controls. The method is able to pare the large pool of potential controls down to the relevant comparisons without using information on outcomes, thereby, if necessary, allowing outcome data to be collected only for the subset

of relevant controls. 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 demonstrated 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 estimator against the experimental-benchmark estimate, the conclusion of our analysis is that it is extremely valuable to check the comparability of the treatment and control units in terms of pre-treatment characteristics, which the researcher *can* check in most applications.

In particular, the propensity score method dramatically highlights the fact that most of the controls are very different from the treated units. In addition to this, when there are very few control units remaining after having discarded the irrelevant controls, the choice of matching algorithm becomes important. We demonstrated that when there are a sufficient number of comparable controls (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 where there are very few comparable controls (in our application, when using the PSID), matching with replacement fares better than the alternatives. Extensions of matching with replacement (radius matching), though interesting in principal, were of little value in our application.

It is something of an irony that the data which we used were originally employed by Lalonde (1986) to demonstrate the failure of standard non-experimental methods in accurately estimating the treatment effect. Using matching methods on both of his samples, we were able to replicate the experimental benchmark, but beyond this we focused attention on the value of flexibly adjusting for observable differences between the treat-

ment and control groups. The process of trying to find a subset of the PSID controls comparable to the NSW units led us to realize that the PSID is a poor comparison group, especially when compared to the CPS.

Because matching methods are focused on the process of constructing a suitable control group in non-experimental settings, the methods which we discuss are a useful addition and complement to the standard techniques in the researcher's arsenal.

Appendix: Estimating the Propensity Score[6]

The first step in estimating the treatment effect is to estimate the propensity score. Any standard probability model can be used, e.g., logit or probit. It is important to remember that the role of the 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.[7]

In estimating the 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 (see 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 (or block) the difference in score for treated and control observations is insignificant (a t-test on a difference in means between the treated and control groups is a criterion used in this algorithm). Proposition A tells us that within each stratum the distribution of the covariates should be approximately the same across the treated and control groups, once the score is controlled for. Within each stratum, we can test for statistically significant differences between the distribution of covariates for treated and control units; operationally, t-tests on differences in the first moments are often sufficient but a joint F-test for the difference in means for all the variables within each block could also be performed.[8] When the covariates are not balanced within a particular block, the block may be too coarsely defined; recall that Proposition A deals with observations with an identical

---

[6] This discussion is drawn from a related paper, Dehejia and Wahba (1997).

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

[8] More generally, one can also consider higher moments or interactions, but usually there is little difference in the results.

propensity score. The solution adopted is to divide the block into finer blocks and test again for no difference in the distribution of the covariates within the finer blocks. If, however, some covariates remain unbalanced for many blocks, 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 block until covariates are balanced. The algorithm is summarized below.

A Simple Algorithm for Estimating the Propensity Score

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

A key property of this estimation 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.

References

Ashenfelter, O. (1978), "Estimating the Effects of Training Programs on Earnings," *Review of Economics and Statistics*, 60, 47-57.

Ashenfelter, O., and D. Card (1985), "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics*, 67, 648-660.

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

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

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

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

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

--------- and --------- (1997), "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," mimeo., University of Toronto.

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

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

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

Lalonde, Robert (1986), "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review*, 76, 604-620.

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

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

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

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

--------- and --------- (1985a), "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity," *American Statistician*, 39, 33-38.

--------- and --------- (1985b), "The Bias Due to Incomplete Matching," *Biometrics*, 41, 103-116.

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

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

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

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

--------- (1986), Discussion of Holland (1986), *Journal of the American Statistical Association*, 81, 961-964.

Rubin, Donald B., and Neal Thomas (1992), "Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Distributions," *Biometrika*, 79, 797-809.

Table 1: Sample Means and Standard Errors of Covariates For Male NSW Participants

National Supported Work Sample
(Treatment and Control)

Variable	Dehejia-Wahba Sample	
	Treatment	Control
Age	25.81	25.05
	(0.52)	(0.45)
Years of schooling	10.35	10.09
	(0.15)	(0.1)
Proportion of school dropouts	0.71	0.83
	(0.03)	(0.02)
Proportion of blacks	0.84	0.83
	(0.03)	(0.02)
Proportion of hispanic	0.059	0.1
	(0.017)	(0.019)
Proportion married	0.19	0.15
	(0.03)	(0.02)
Number of children	0.41	0.37
	(0.07)	(0.06)
No-show variable	0	n/a
	(0)	
Month of assignment (Jan. 1978=0)	18.49	17.86
	(0.36)	(0.35)
Real earnings 12 months before training	1,689	1,425
	(235)	(182)
Real earnings 24 months before training	2,096	2,107
	(359)	(353)
Hours worked 1 year before training	294	243
	(36)	(27)
Hours worked 2 years before training	306	267
	(46)	(37)
Sample size	185	260

Data Legend: Age, age of participant; **Educ**, number of school years; **Black**, 1 if black, 0 otherwise; **Hisp**, 1 if Hispanic, 0 otherwise; **Nodegree**, 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

Table 2: Sample Characteristics of the NSW and CPS 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[D]
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
													(633)	*(638)*
Full CPS (s.e.)[C]	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)	*(0.53)*	*(0.15)*	*(0.03)*	*(0.02)*	*(0.03)*	*(0.03)*	*(367)*	*(248)*	*(0.03)*	*(0.04)*	*(583)*	*(554)*
Without replacement:														
Random	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)*	*(719)*
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)*	*(719)*
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)*	*(730)*	*(719)*
With replacement:														
Minimum Bias	119	0.37	25.36	10.31	0.84	0.06	0.69	0.17	2407	1516	0.35	0.49	1360	1374
		(0.03)	*(1.04)*	*(0.31)*	*(0.06)*	*(0.04)*	*(0.07)*	*(0.06)*	*(727)*	*(506)*	*(0.07)*	*(0.07)*	*(974)*	*(743)*
Radius:														
δ=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)*	*(1077)*	*(780)*
δ=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)*	*(0.06)*	*(1086)*	*(774)*
δ=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)*	*(1080)*	*(774)*

Notes:.(A) The propensity score is estimated using a logit of treatment status on: Age, Age^2, Age^3, Educ, Educ^2, Married, Nodegree, Black, Hisp, RE74, RE75, U74,. U75, Educ·RE74. (B) The treatment effect for the NSW sample is estimated using the experimental control group. (C) 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. (D) 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.

Table 3: Sample Characteristics of the NSW and PSID Samples

Control Sample	No. of Observations[A]	Mean Propensity Score	Age	School	Black	Hispanic	No Degree	Married	RE74 US$	RE75 US$	U74	U75	Treatment effect (diff. in means)	Regression treatment effect[D]
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
													(633)	*(638)*
Full PSID (s.e.)[C]	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)	*(0.57)*	*(0.16)*	*(0.03)*	*(0.02)*	*(0.03)*	*(0.03)*	*(449)*	*(361)*	*(0.04)*	*(0.03)*	*(657)*	*(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)*	*(1002)*	*(1152)*
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)*	*(1002)*	*(1152)*
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)*	*(1002)*	*(1152)*
With replacement:														
Minimum Bias	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)*	*(1791)*	*(809)*
Radius:														
δ=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)*	*(2410)*	*(853)*
δ=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)*	*(2523)*	*(955)*
δ=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)*	*(2691)*	*(1094)*
δ=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)*	*(3459)*	*(1404)*

Notes:.(A) The propensity score is estimated using a logit of treatment status on: Age, Age^2, Educ, Educ^2, Married, Nodegree, 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 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. (D) 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.

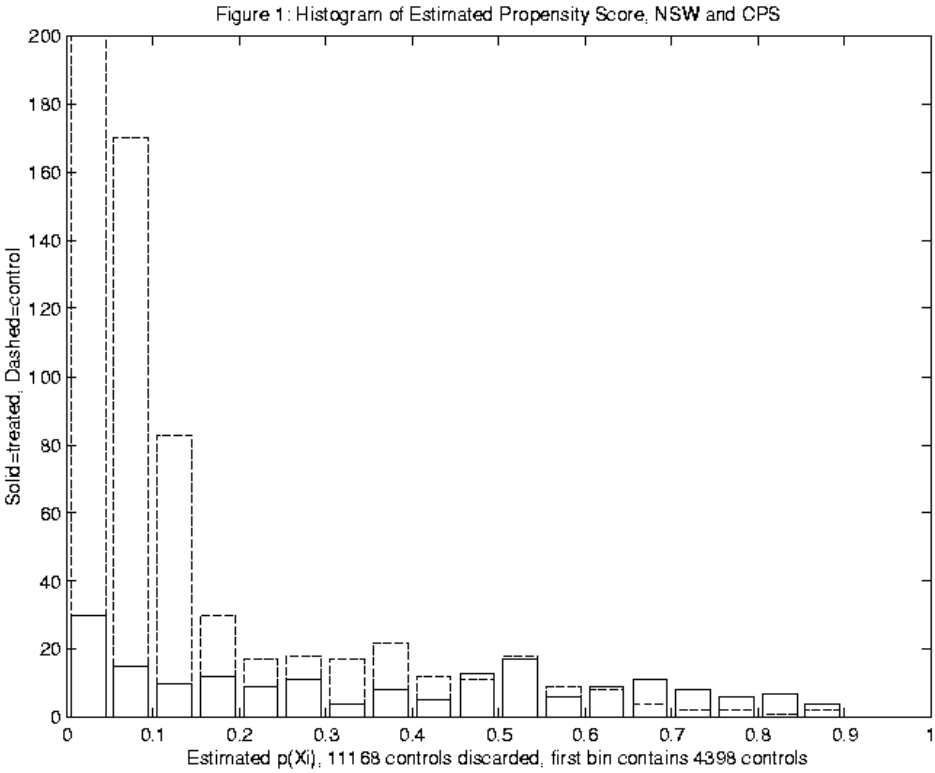


Figure 1: Histogram of Estimated Propensity Score, NSW and CPS
Solid=treated, Dashed=control
Estimated p(Xi), 11168 controls discarded, first bin contains 4398 controls
0
20
40
60
80
100
120
140
160
180
200
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

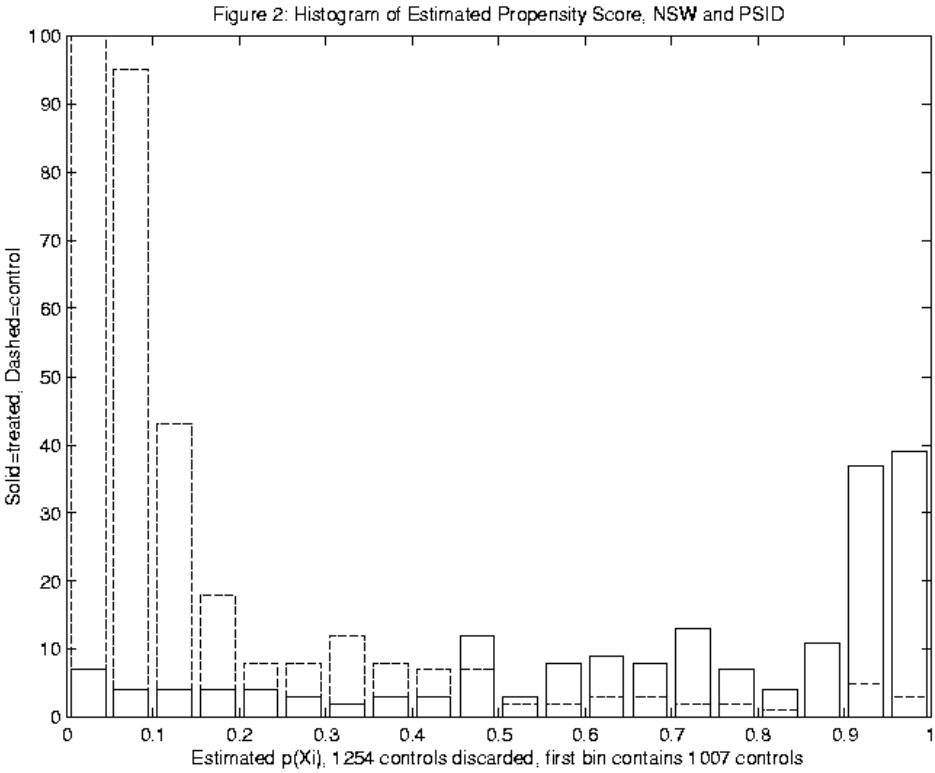


Figure 2: Histogram of Estimated Propensity Score, NSW and PSID
Solid=treated, Dashed=control
Estimated p(Xi), 1254 controls discarded, first bin contains 1007 controls
0
10
20
30
40
50
60
70
80
90
100
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

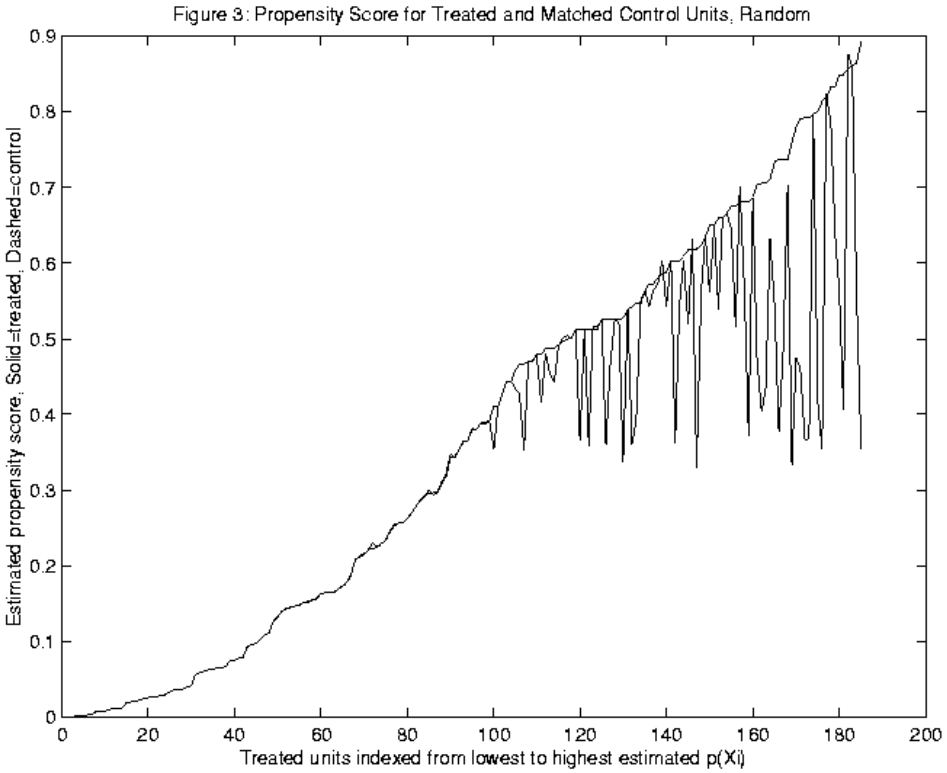
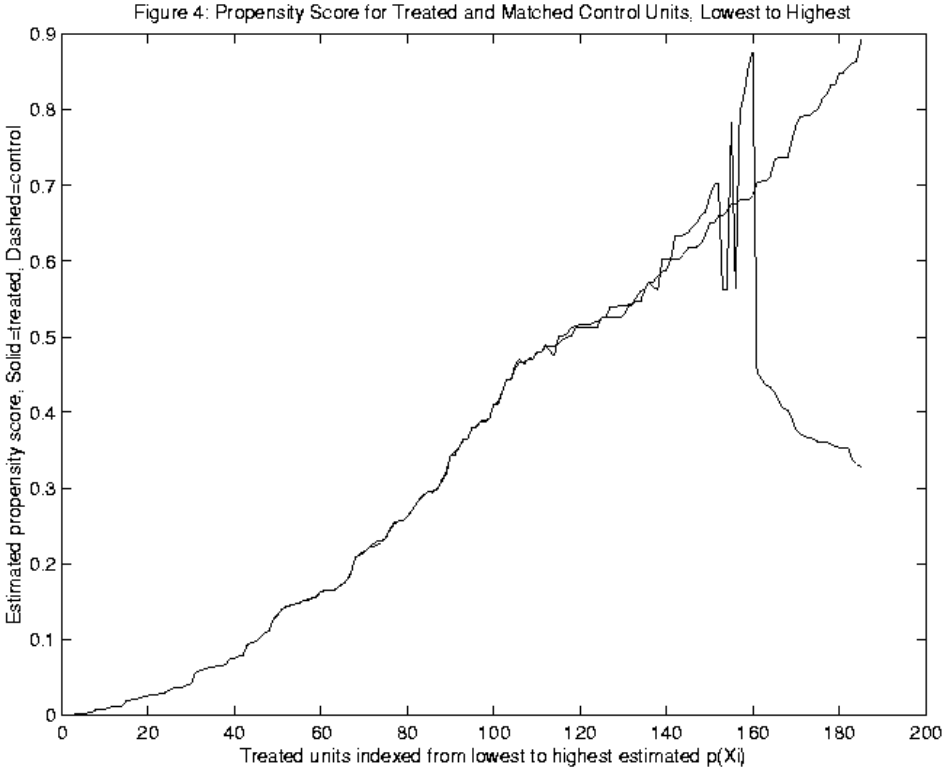


Figure 3: Propensity Score for Treated and Matched Control Units, Random
Estimated propensity score, Solid=treated, Dashed=control
Treated units indexed from lowest to highest estimated p(Xi)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
20
40
60
80
100
120
140
160
180
200

Figure 4: Propensity Score for Treated and Matched Control Units, Lowest to Highest



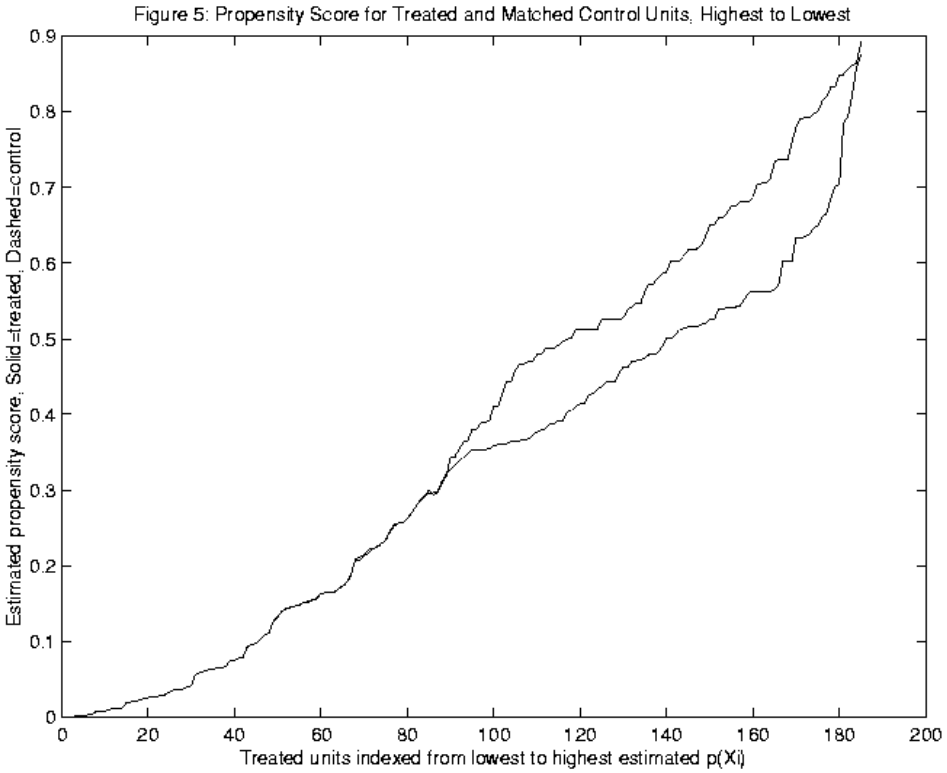


Figure 5: Propensity Score for Treated and Matched Control Units, Highest to Lowest
Estimated propensity score, Solid=treated, Dashed=control
Treated units indexed from lowest to highest estimated p(Xi)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
20
40
60
80
100
120
140
160
180
200

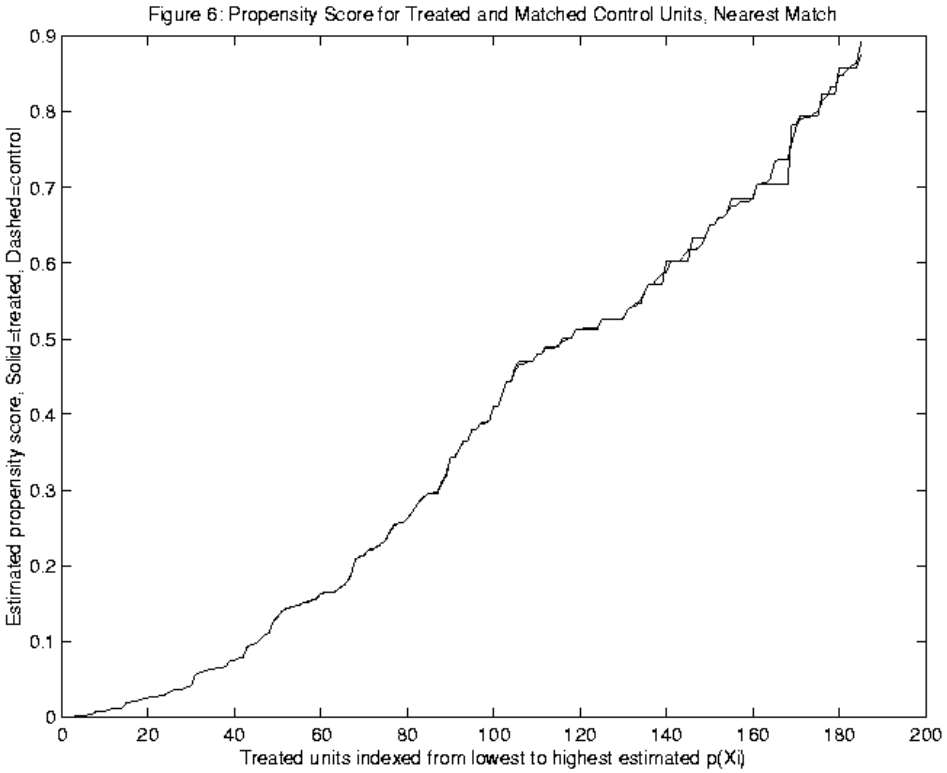


Figure 6: Propensity Score for Treated and Matched Control Units, Nearest Match
Estimated propensity score, Solid=treated, Dashed=control
Treated units indexed from lowest to highest estimated p(Xi)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0
20
40
60
80
100
120
140
160
180
200