Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs [*]

Rajeev H. Dehejia and Sadek Wahba

cite as
Rajeev Dehejia and Sadek Wahba, "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs", in Rajeev Dehejia, *Econometric Methods for Program Evaluation*, Ph.D. Dissertation, Harvard University, 1997, Chapter 1.

[*] We gratefully acknowledge the tireless encouragement and support of Gary Chamberlain, Guido Imbens, and Donald Rubin. We thank Robert Lalonde, who kindly provided the data from his 1986 study and provided substantial help in recreating the original data set. We are also grateful to Joshua Angrist, George Cave, David Cutler, Lawrence Katz, Caroline Minter-Hoxby, and participants at the Harvard-MIT labor seminar, the Harvard econometrics and labor lunch seminars, the MIT labor lunch seminar, and a seminar at the Manpower Development Research Corporation (MDRC) for many suggestions and comments. All 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.
Correspondence Address: Department of Economics, University of Toronto. 150 St. George Street, Toronto ON M5S 3G7, Canada. E-mail: rdehejia@chass.utoronto.ca.

Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs

Abstract

The need to use randomized experiments in the context of manpower training programs, and in analyzing causal effects more generally, has been a subject of much debate. Lalonde (1986) considers experimental data from the National Supported Work (NSW) Demonstration and non-experimental comparison groups drawn from the CPS and PSID, and argues that econometric methods fail to replicate the benchmark experimental treatment effect. This paper applies propensity score methods, which have been developed in the statistics literature, to Lalonde's dataset. In contrast with Lalonde's findings, using propensity score methods, we are able closely to replicate the experimental training effect. The methods succeed because they are able flexibly to control for the wide range of observable differences between the (experimental) treatment group and the (non-experimental) comparison group.

1. Introduction

An important question when analyzing causal effects in non-experimental studies is how well techniques of causal inference perform relative to experimental evaluations. For example, how accurately can a researcher hope to estimate the effect of a manpower training program on earnings in an observational study? The question is not a new one.[1] The need to use the classical statistical methodology of randomized experiments in economic applications such as manpower training programs is addressed by Ashenfelter (1978), Ashenfelter and Card (1985), Burtless and Orr (1986) and more recently by Burtless (1995). Lalonde (1986) is the first study to examine the standard econometric procedures used to evaluate the effect of training programs on earnings. He examines a randomized experiment (the National Supported Work Demonstration, NSW) from which he obtains an unbiased estimate of the training effect, and then compares the experimental result to those obtained from a range of parametric selection models (estimated using least squares regressions, instrumental variables, and the Heckman [1979] two-step procedure) applied to the NSW observations that received training and a set of comparison observations constructed from population survey data sets (CPS and PSID).[2] The conclusion in Lalonde (1986), which has been very influential in labor economics and the evaluation of social programs (see Katz [1992]), is that this array of estimators fails robustly to replicate the experimentally determined results.

---

[1] The importance of classical experiments in explaining causal relations in econometrics goes beyond the case of training programs used here as one possible application. See Cox (1992), Leamer (1978), and Pratt and Schlaifer (1988) for various perspectives on the role of randomization in economic analysis.

[2] We use the term control to refer to units that did not receive treatment; this includes experimental units randomized out of treatment and non-experimental comparison units.

In this paper, we apply methods for causal inference developed in the statistics literature that rely on the assumption that conditional on covariates selection is ignorable -- also referred to as selection on observables (Rubin [1974, 1977, 1978], Rosenbaum and Rubin [1983a], and reviewed in Heckman and Robb [1985] and Holland [1986]).[3] We re-estimate the treatment effect in Lalonde's non-experimental dataset and present a range of estimators which employ the propensity score method (Rosenbaum and Rubin [1983a]), successfully and robustly replicating the experimental treatment effect. We show that our methods succeed, where those considered by Lalonde fail, precisely because they control fully for observable differences between the NSW treated units and the CPS and PSID controls. We also demonstrate the importance of using a full set of pre-treatment covariates and allowing for a heterogeneous treatment effect.

There have been many responses to Lalonde's conclusions. Important among these is Heckman and Hotz (1989) who emphasize the importance of using appropriate specification tests to select an estimator. Other related studies include Card and Sullivan [1988] who examine the effect of training on employment; Heckman et al., [1995] who estimate the effect of the JTPA training program on earnings; and Manski et al., [1992] who examine the impact of family structure on school enrollment.

It is clear that selection on observables is a strong assumption, requiring a sufficiently complete set of pre-treatment covariates. However, our view is that before taking recourse to assumptions on functional forms and distributions, there is much merit in exploiting fully the information contained in the variables that are observed. In settings where the selection-on- observables assumption is not adequate, the techniques described in the paper should still be seen as an important and practical complement to

[3] The case of selection on observable characteristics in the econometrics literature was considered first by Goldberger (1972a) and further developed in Barnow, Cain, and Goldberger (1980).

other econometric methods such as instrumental variables, that depend on well specified exclusion restrictions (e.g., Angrist [1990], Angrist, Imbens, and Rubin [1996], and Imbens and Angrist [1994]), and assumptions on the distribution of unobserved characteristics (Heckman [1979]).

The paper is organized as follows. Section 2 reviews Lalonde's results, outlining the econometric framework on selection bias and replicating his results. Section 3 identifies the treatment effect under the potential outcomes causal model, and Section 4 discusses estimation procedures for the treatment effect. In Section 5, we implement the approach of Sections 3 and 4 for Lalonde's dataset, and in Section 6, we discuss the sensitivity of the results to the methodology. Section 7 concludes the paper.

2. Lalonde's Results

2.1 The Econometric Framework

The models considered in Lalonde (1986) fit into the following standard model in the econometrics literature (e.g., Maddala [1983] and Heckman [1990]); there are two outcome equations:

$$Y_{i1} = \alpha_1 + X_i\beta_1 + u_{i1} \quad \text{for participants}$$
$$Y_{i0} = \alpha_0 + X_i\beta_0 + u_{i0} \quad \text{for non-participants,}$$

where Y_{i1} and Y_{i0} are the outcome of interest for participants and non-participants respectively, which are linear functions of a vector of observable characteristics X and some error term. A participation decision rule determines whether individual i participates in the program (T_i=1) or not (T_i=0):

$$T_i = \begin{cases} 1 & \text{if } T_i^* \geq 0 \\ 0 & \text{if } T_i^* < 0 \end{cases}$$
$$T_i^* = \gamma_0 + X_i\gamma + v_i.$$

The error terms are distributed as:

$$\begin{pmatrix} u_{i1} \\ u_{i0} \\ v_i \end{pmatrix} \Big| X \overset{i.i.d.}{\sim} N(0, \Sigma),$$

where a range of specific assumptions can be made about the elements of the covariance matrix, S. A leading case in the literature is the case of a constant additive treatment effect:

$$Y_{i1} = d + Y_{i0},$$

where d is the effect due to treatment. The equation for the observed outcome variable can then be written as:[4]

$$Y_i = b_0 + dT_i + \mathbf{X}_i b_1 + e_i \tag{1}$$

The selection-on-observable-characteristics assumption is expressed as:

$$T_i \perp\!\!\!\perp e_i \,|\, X_i$$

(where $\perp\!\!\!\perp$ refers to independence). Under the above assumptions, equation (1) can be estimated using least squares to obtain an unbiased estimate of the treatment effect, δ:

$$E(Y_i | T_i, X_i) = b_0 + dT_i + X_i b_1$$

It is this model (and variations of it in which the set of conditioning variables is modified) that is applied most widely in Lalonde's paper. He also considers models which exploit the parametric assumption on S (see Heckman [1979]). Our focus is this first class of models.

In non-experimental settings where data on the control group is either a self-selected sample, or in some cases (as in Lalonde's paper) is drawn from an altogether different population, the distribution

[4] Where $Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$, $b_0 = a_0$, $X_i b_1 = X_i m_0 = X_i m_1$, $d = a_1 - a_0$, and $e_i = u_{i1} = u_{i0}$.

of the observable characteristics between treated and control units need not overlap very much. In this case, estimating treatment effects through models such as equation (1) amounts to extrapolating between two very different groups. If the groups are sufficiently different, such an extrapolation can be extremely misleading, as will be demonstrated. Our objective in this paper is to show that by relaxing the linearity assumptions in (1), the assumption of selection on observables can still be maintained, and that the resulting estimators perform well under the evaluation considered by Lalonde.

2.2 Replicating Lalonde's Results

The characteristics of the NSW sample are presented in Table 1. The analysis in Lalonde (1986) uses only one year of pre-treatment earnings. But as Ashenfelter (1978), Ashenfelter and Card (1985), Card and Sullivan (1988) and others indicate, the use of several years of earnings is key in estimating the variability in the training effect. Since the methodology we explore relies on fully exploiting the selection on observables assumption, we obtain from the NSW survey additional information on the earnings profile of the participants. This additional information allows us to test the sensitivity of the methodology to selection on observable characteristics such as pre-treatment earnings. Table 1 also provides the characteristics of the reduced sample used throughout this paper (185 treated observations and 260 controls).[5]

[5] From the sample of 297 treated and 425 control used by Lalonde, we exclude those observations for which earnings in calendar 1974 could not be obtained, thus arriving at a reduced sample of 185 treated observations and 260 control observations. The two samples do, however, differ from each other. For example, earnings in 1975 are substantially lower for our sample than for the entire sample ($ 1,532 against $ 3,066), as are the earnings two years prior to assignment (which is equivalent to earnings in 1974 for the second sample). As expected the average month of assignment increases for the second sample, 18.5 compared to 16.5 for the first sample. These differences simply reflect the "cohort phenomenon" noticed by the designers of the NSW program and do not compromise the validity of a simple comparison of sample means as an unbiased estimate of the treatment impact (see MDRC [1983]).

Non-experimental estimates of the treatment effect are based on the two distinct comparison groups used by Lalonde (1986), the Panel Study of Income Dynamics (PSID) and Westat's Matched Current Population Survey-Social Security Administration File (CPS-SSA). From these two control groups, several sub-groups are created following criteria outlined in Table 2. Following Lalonde, the training effect is estimated in two ways: first as a difference in means of earnings between the treated and control groups (the unadjusted treatment effect), and second, through an estimate of equation (1) by regressing earnings in 1978 on a dummy variable for treatment and a set of covariates (hereafter the adjusted treatment effect). These two estimators are reported throughout the analysis.

Table 3 presents the complete set of estimators used by Lalonde (1986), with the first row reproducing the experimental treatment effect using the NSW control group. The simple difference in means, reported in column (4), yields highly negative treatment effects for the CPS and PSID controls. The adjusted treatment effect which controls for pre-treatment earnings and covariates is reported in column (10). Applying one of the specification tests suggested by Lalonde, of regressing pre-treatment earnings in 1975 and in 1974 over the same functional form used to estimate the adjusted treatment effect, the researcher would have to reject all the estimators in column (10), since the difference in pre-treatment earnings (1974 and 1975) of the two groups is statistically significant.[6] Likewise, the estimators in the other columns fail to produce a stable estimate replicating the experimental benchmark (see Lalonde [1986] for additional details).

The essential insight of Lalonde's study is that adjustment through linear regression on the composite sample of NSW treated and CPS or PSID controls yields estimated treatment effects which

[6] The usefulness of one additional year of pre-treatment earnings becomes apparent when applying the specification test; both for PSID-1,2, and CPS-1,2 the test fails for earnings in 1974 and 1975. However, for PSID-3, and CPS-3

fail robustly to replicate the experimental treatment effect. In the next sections we demonstrate that maintaining the assumption of ignorable assignment conditional on covariates, but relaxing the linearity assumptions, one can successfully replicate the experimental treatment effect.[7]

3. Identifying the Average Treatment Effect

3.1 Causality and the Role of Randomization

We begin with a brief review of the notion of causality that we use. A cause is viewed as a manipulation or treatment which brings about a change in the variable of interest as compared to a baseline, called the control. If, as in the previous section, Y_{i1} (Y_{i0}) is the value of the outcome when unit i is subject to treatment 1 (treatment 0, called control), the treatment effect for a single unit, t_i, is defined by: $t_i \equiv Y_{i1} - Y_{i0}$. As compared to $Y_i = T_i Y_{i1} + (1-T_i)Y_{i0}$, the observed value of the outcome variable, only one of Y_{i0} or Y_{i1} is observed for any i (referred to as the fundamental problem of causal inference in Holland [1986]). Likewise the average treatment effect,

$$\begin{aligned} t \equiv E(t_i) &= E(Y_{i1}) - E(Y_{i0}) \\ &= E(Y_{i1}|T_i = 1)\cdot p(T_i = 1) + E(Y_{i1}|T_i = 0)\cdot p(T_i = 0) \\ &\quad - \left[E(Y_{i0}|T_i = 0)\cdot p(T_i = 0) + E(Y_{i0}|T_i = 1)\cdot p(T_i = 1)\right], \end{aligned}$$

---

adjusted pre-treatment earnings in 1975 are statistically insignificantly different for treated and control units, but are significantly different for earnings in 1974.

[7] Another way to relax the linearity assumption of the model in Section 2 is suggest by Goldberger (1972b). Equation (1) is re-written as:

$$E(Y_i|T_i, X_i) = \beta_0 + \delta T_i + X_i\beta_1 + T_i X_i \beta_2. \qquad (2)$$

As well, one could allow for higher order and interaction terms of the covariates X_i. Although, in principle it is a flexible approach for estimating the treatment effect, estimating such a model when X_i is multi-dimensional (and includes many continuous variables) is an econometric (non-parametric) problem of a high order of difficulty. Simply saturating a regression with higher order and interaction terms would quickly exhaust the number of observations available, and which interaction terms to include (or exclude) is an issue that increases in complexity as the number of possible terms increases exponentially (see Härdle [1990], as well as Angrist [1995]).

cannot be estimated directly because using observed data we can only estimate $E(Y_{i1}|T_i=1)$ and $E(Y_{i0}|T_i=0)$.[8] Intuitively, if the treated and control units systematically differ in their characteristics, then in observing only the treated group we cannot in general correctly estimate Y_{i1} for the whole population ($E(Y_{i1}) \neq E(Y_{i1}|T_i = 1)$), and likewise for Y_{i0} and the control group.

Randomizing assignment of individuals into treatment and control allows us to estimate the average treatment effect over the population of interest, because it implies that

$$Y_{i1}, Y_{i0} \perp\!\!\!\perp T_i .$$

This in turn implies that $E(Y_{i1}|T_i = 1) = E(Y_{i1}|T_i = 0)$ and $E(Y_{i0}|T_i = 0) = E(Y_{i0}|T_i = 1)$, so that

$$\begin{aligned} t &= E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 0) \\ &= E(Y_i|T_i = 1) - E(Y_i|T_i = 0). \end{aligned}$$

Because the treated and control groups are random subsamples of the participants, Y_{i1} for the treated group is representative of Y_{i1} for the population (likewise for Y_{i0} and the control group), so that the treatment effect is identified.

3.2 Non-Experimental Settings and the Role of the Propensity Score

The extension of the classical randomized framework to a non-experimental setting when assignment to treatment occurs on observable characteristics is due to Rubin (1974, 1977, 1978). In non-experimental studies, data is typically available only on a treated group made up of a systematic sub-sample of the population (e.g., volunteers). The control group is either a systematic sub-sample of the

[8] An important assumption is that the conditional expectation of the outcome for unit *i* does not depend on the treatment status of other units. Otherwise we would have to condition throughout on the entire vector of treatment assignments. This is referred to as the stable unit treatment value assumption (SUTVA) (Holland [1986] and Rubin [1978]).

population (e.g., those who did not volunteer or were not chosen for treatment), or it may not have been collected alongside the treatment group and may have to be created by turning to other data sets (for example, potential controls are available through periodic population surveys). In these cases, it is (usually) the treated group which is drawn from the population of interest.[9] The treatment effect is then defined as:

$$t|_{T=1} = E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 1). \tag{3}$$

Again, as in the previous section, equation (3) is not identified since Y_{i0} is never observed for units with T_i=1.

In a non-experimental setting, identification is possible under the assumption of ignorable assignment conditional on covariates, i.e., assignment to treatment or control is a (stochastic) function of a vector of (observable) covariates. In this case, conditional on the vector X, the assignment mechanism is like a randomized experiment (Rubin [1977]):

Proposition 1: *If for each unit we observe a vector of covariates* X_i *and*

$$Y_{i1},\ Y_{i0} \perp T_i | X_i,\ \forall i,$$

then:

$$\begin{aligned} t|_{T=1} &\equiv E(Y_{i1}|T_i = 1) - E(Y_{i0}|T_i = 1) \\ &= E_X\left\{E\left(Y_i|X_i, T_i = 1\right) - E\left(Y_i|X_i, T_i = 0\right)\middle|T_i = 1\right\}, \end{aligned} \tag{4}$$

where $Y_i = T_i Y_{i1} + (1 - T_i)Y_{i0}$.

[9] A less natural but consistent case would be to use a treated group to estimate the causal effect for a given control population of interest. Note that in the setting of a randomized experiment, the treatment effect for the treated population is identical to the treatment effect for the untreated population:

$$t_{|T=1} = t_{|T=0} = E(Y_i|T_i = 1) - E(Y_i|T_i = 0).$$

Proof: See Appendix A.

Taken literally, the notion of conditioning corresponds to matching or grouping the observations on the covariate X. But implementing this approach requires a sufficiently simple set of (discrete) covariates to keep the task of conditioning on the exact value of X a tractable exercise. In many instances, this is extremely difficult; for example if there are k dichotomous covariates, the number of matching cells is 2^k. The Propensity Score Theorem (Rosenbaum and Rubin [1983a]) offers a potential solution to this problem:

Proposition 2: *Let $p(X_i)$ be the probability of unit* i *having been assigned to treatment, defined as $p(X_i)=\Pr(T_i=1|X_i)=E(T_i|X_i)$, where $0<p(X_i)<1$, $\forall i$. Then:*

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

implies

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

Proof: See Appendix A.

Corollary 2.1:

$$t|_{T=1} = E_{p(X)}\Big[E(Y_i|T_i=1, p(X_i)) - E(Y_i|T_i=0, p(X_i))\,\Big|\,T_i=1\Big]. \tag{5}$$

Proof: See Appendix A.

Thus, independence conditional on covariates extends to the propensity score, as does by immediate implication our result on the computation of the treatment effect. The achievement of the theorem is that equation (5) only requires matches on a univariate scale, rather than on X.

Proposition 2 essentially reduces the exercise of estimating the treatment effect to estimating the following two non-parametric functions:

$$E\left(Y_{i1} | p(X_i)\right) = E\left(Y_i | T = 1, p(X_i)\right)$$

$$E\left(Y_{i0} | p(X_i)\right) = E\left(Y_i | T_i = 0, p(X_i)\right),$$

which would be univariate non-parametric regressions if the propensity score were known.[10] In this paper we deliberately focus on relatively intuitive methods for obtaining a flexible functional form, but in principle one could use any one of the standard array of non-parametric techniques (see for example Härdle [1990]).

A complication in implementing this procedure is that the propensity score is unknown. We estimate it using a logit model (see Appendix B for details). At some level, this merely transfers the burden of estimating a high-order non-parametric regression from equation (4) to this step. There are a number of reasons to prefer our approach. First, as indicated earlier, tackling equation (4) directly with a non-parametric regression would encounter the curse of dimensionality as a problem in many datasets. This is also true about estimating the propensity score using standard non-parametric techniques. Hence, we use a parametric model for the propensity score. But this is preferable to applying a parametric model to the outcome equation, such as equation (2), because there is a well-defined criterion to

[10] Since we are estimating the average treatment effect for the treated population, and Y_{i1} for the treated population is known, in effect we do not have to estimate the second of these.

determine how many interaction terms to include in the specification, embodied in the following proposition (Rosenbaum and Rubin 1983a):

Proposition 3:

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

Proof: See Appendix B.

Though elementary, Proposition 3 is fundamental in providing a framework to validate estimates of the propensity score and hence in choosing which higher order and interaction terms to use. For equal values of the propensity score, the theorem tells us that the covariates are also balanced (in distribution). This provides an easy diagnostic for how well the score has been estimated (discussed in greater detail in Appendix B). Finally, as we will see in the next section, depending on the estimation strategy one adopts, an extremely precise estimate of the propensity score is not even needed.

4. Estimating Treatment Effect

This section suggests three straightforward methods of using the propensity score to obtain an estimate of the estimate treatment effect.

4.1 Stratifying on the Score

A commonly used method to control for a single covariate is stratification. Stratifying on the propensity score entails dividing the unit interval into blocks or strata that are sufficiently fine to allow us to consider the treated and control units within each stratum as having approximately the same propensity score,

because when units have the same propensity score, it follows that the distribution of the entire vector of covariates will be the same.[11] It is in this sense that stratifying on the score ensures an overlap in the distribution of characteristics across the treated and control groups.

Within each such stratum, the treatment effect is the difference of two expectations that are a function of observables, $E(Y_{i1}|T_i=1,p(X_i))-E(Y_{i0}|T_i=0,p(X_i))$. Estimating $t|_{T=1,p(X)}$ requires only point estimates of each term, and like a randomized experiment, the difference of means is an (approximately) unbiased estimator.[12] Within each block very little modeling is required, and the choice of functional forms is no longer a major issue. A weighted average of the treatment effect within each block (where the weights are the number of treated units in each block) estimates the average treatment effect for the treated.

A simple and immediate test of whether an estimate of the propensity score is sufficiently accurate is that one can find a partition structure such that, within each stratum, observable characteristics are balanced across treated and control units. In this case, the estimate of the propensity score is used only in grouping units, not directly in the estimator.

---

[11] The method suggested in Appendix B for estimating the propensity score also involves stratifying on the estimated propensity score. A natural choice here would be to use the same strata. Cochran (1968) and Rosenbaum and Rubin (1985a) show that under certain restrictions, including normality of covariate distribution, five equal size blocks reduce 95 percent of the bias. Although these results can be taken as a benchmark for the number of blocks that would reduce most of the bias from differences in the distribution of the covariates, ultimately the blocking is a function of the overlap between the distribution of the score for treated and control samples as well as sample size. It is also evident that the simple stratification procedure adopted in the appendix is not the only way of determining the number of blocks that balance the covariates; other non-parametric techniques could be used such as kernel or nearest neighbor (see Härdle [1990]).

[12] Using a difference in means within each stratum amounts to specifying a step-function functional form. More generally one could use more complicated forms within each stratum (linear, quadratic, etc.), and impose continuity or differentiability requirements as well. These strategies are simpler versions of standard non-parametric regression techniques.

4.2 Matching on the Propensity Score

A second estimation strategy that follows from equation (5) is pair-wise matching on the score. The conditioning on the propensity score is implemented by matching techniques that pair each treated unit to the single control unit with the closest propensity score.[13] The matched sample will have the property that the distribution of observed covariates for the treated and control groups is approximately the same. Given the assumption of conditional ignorability (Proposition 1), the treatment effect is estimated by taking a difference in means or using least squares adjustment appropriately weighted, to correct for any remaining imbalance. An algorithm for determining how to match units within some bands of tolerance on the inexactitude of the match needs to be specified, and there are delicate issues regarding the order in which to match the treated units (see Rosenbaum and Rubin [1985b], and Rubin and Thomas [1992]). Results reported in this paper for matching follow a procedure that accounts for the minimal overlap between the treated and control distributions by allowing a given control unit to be matched with more than one treated unit (see Dehejia and Wahba [1995]).[14]

4.3 Using the Propensity Score as Weights

The score can be used directly as a weight in estimating the average treatment effect, as stated in the following proposition:

Proposition 4:

[13] In some sense matching is an extreme form of stratification where each treated observation is in a separate stratum, which is sufficiently narrow to include only one control. Note that control units would be thrown out in a matching exercise even if they were previously included in blocks determined by the score.

[14] Early work on matching revolved around matching on a covariate or a set of covariates. For a theoretical examination of matching on a set of covariates see Rubin (1973, 1979).

$$\frac{1}{N^T}\sum_{i=1}^{N^T+N^C}\left(T_iY_i-(1-T_i)\frac{p(X_i)}{1-p(X_i)}Y_i\right)$$

is a consistent estimator of $t|_{T=1}$*, where* Y_i *is the observed value of outcome for unit* i*,* T_i *is an index variable (*=1 *if treated and* =0 *if control),* $0<p(X_i)<1$*, and* N^T *and* N^C *are the total number of treated and control observations respectively.*

Proof: See Appendix C

This estimator differs from the first two to the extent that the objective of the method adopted in this paper is to be agnostic about which functional form needs to be assumed. So even if in estimating the propensity score the functional form is not exactly correct, but within each stratum the observable covariates of treated and control units balance, then given the assumption of selection on observables this is sufficient for an unbiased estimate of the treatment effect -- independent of the estimate of the propensity score. There is no such comfort in using weighting. A very different estimate of the treatment obtained when using weighting, compared to using stratification or matching, would suggest that either the score is mis-specified (an issue that can be corrected through additional interaction terms and higher order terms) or that the treatment effect is not ignorable conditional on the score. In this sense, Proposition 4 provides an additional self-diagnostic test.

5. Results Using the Propensity Score

5.1 Estimating the Propensity Score

Following the argument in Section 3, the propensity score is estimated using a logit with the treatment status as the dependent variable and the pre-treatment covariates as independent variables. The final

choice of interaction and higher order terms included in the logistic function was determined solely by the need to balance the covariates within blocks as defined in the algorithm (see Appendix B). This procedure was repeated for each of the six control groups of Table 2 separately, and the resulting logistic functions are presented in the footnote of Table 4. Note that in this procedure the outcome variable (earnings in 1978) plays no role.[15] Note also, that the procedure embodies specification tests of the type suggested by Lalonde. Within each block unadjusted as well as regression-adjusted differences in all pre-treatment covariates between the two groups, including pre-training earnings, were estimated. Only if the difference was not significant were the blocks maintained.

5.2 Treatment Effect Stratifying on the Propensity Score

The first of the estimators discussed in Section 4 uses stratification on the estimated propensity score. The treatment effect is estimated by summing over the blocks the difference of the within-block means of the outcome variable for the treated and control observations, where the sum is weighted by the number of treated observations within each block, obtaining the unadjusted treatment effect.[16] Alternatively, a treatment effect could be obtained using the same regression specification as column (10) of Table 4 within each block, and again taking a weighted sum over the blocks to obtain the adjusted treatment effect.

Tables 5 and 6 present the disaggregated treatment effect using the stratification resulting from the estimation of the propensity score for PSID-1 and CPS-1 respectively. Note that all control observations with an estimated score lower than the minimum estimated score for treated observations

[15] Earnings do enter however in lagged form as pre-treatment earnings in 1975 and 1974.

are excluded. For Table 5, this is indicated in the first block where the lowest estimated score is 0.0004. The number of control observations used is determined only by the degree of overlap between the distribution of the score for the treated and control groups, resulting in 1070 control observations from PSID-1. The estimated training effect is $1,509 and $1,647 for the unadjusted and adjusted estimates respectively. In Table 6 the minimum estimated score is 0.001, which is the lowest estimated value for the treated observations. The total number of CPS-1 control actually used is 3,992 implying that 12,060 control observations (a full 75 percent of the total number of control observations) have an estimated score less than the minimum estimated score for the treated observations. This illustrates well the weakness of the standard model; in linear models such as those tested by Lalonde (1986), one is extrapolating from a group made up mostly of irrelevant controls. With CPS as a control group, after controlling for observable characteristics through the propensity score, the unadjusted training effect is $1,713, and the adjusted training effect $1,774.

Several other characteristics of the tables should be mentioned. Blocks vary in their score range because though a greater number of observations within a block is desirable, ultimately the block size will depend on balancing the covariates. The treatment effect varies within each block since it depends on the particular sample characteristics represented in the block. For example, treated observations in block 1 of Table 6 had an average age of 27, 20 percent of them were black participants, and the average earnings was $6,620 in 1975. In contrast, the last block was made up of treated participants with an average age of 26, all of whom were black, with an average income of $194 in 1975. The heterogeneity of the treatment effect is taken up further in Section 5.5. Finally, as in a randomized

[16] In the non-experimental set-up this corresponds to the average treatment effect conditional on having been in the NSW treated group. Because the NSW is a randomized experiment its treated and control groups are drawn from the

experiment, the unadjusted and adjusted treatment effects within each block are similar, which demonstrates that by conditioning on the propensity score we are balancing the other covariates as well.

The estimated treatment effects from stratification on the score for all six groups are summarized in Table 4 in columns (4) and (5). Columns (1) and (2) repeat the benchmark estimates discussed earlier for convenience. The main feature of these results is that the use of the propensity score has eliminated those observations in the control groups that are not comparable to the treated observations, without resorting to any *ad hoc* assumptions on the characteristics of the control observations used to derive PSID-2 and PSID-3 and CPS-1 and CPS-3. This is not to say that in going from CPS-1 to CPS-3 and PSID-1 to PSID-3 one may not improve the estimate; instead, the basic point is that ensuring overlap through the score is a more systematic way to eliminate irrelevant controls.

Furthermore, comparing these estimates with the estimated training effect over the entire sample (columns 1 and 2) demonstrates the problem associated with the extrapolation implicit in the least squares training effect when there is minimal overlap in the two distributions. Unlike the estimates in columns (1) and (2), the treatment effect estimated in column (3) is estimated using the same specification as (2), adding the score as a variable and regressing over the overlap sample specified in column (6). The estimated treatment effect for PSID-1 and CPS-1 is $542 and $893 respectively. A constant additive treatment effect estimated in the overlap sample does not result in substantially higher estimates. As results in columns (4) and (5) indicate, ensuring that the distributions overlap *and* relaxing the constant effect treatment assumption through a flexible functional form yield estimates that are considerably closer to the benchmark estimate.

---

same population, so the correct benchmark for comparison remains the treatment effect of $1,794.

Moving down the various control groups does not significantly alter the estimated treatment effect in columns (4) and (5). The estimated treatment effects range from a low of $1,335 to a high of $1,713 for the unadjusted estimate with CPS controls and from a low of $1,509 to a high of $1,829 for PSID controls. In the case of adjusted estimates the training effect varies from $1,023 to $1,774 and $1,647 to $2,538 for the CPS and PSID controls respectively. Note however that under stratification on the propensity score there is no need to construct further control groups, since non-comparable controls are already discarded through stratification. Thus, under this approach a researcher would no longer need to construct somewhat arbitrary control groups such as PSID-2, PSID-3 and CPS-2, CPS-3 and would report only the adjusted training effects that vary between $1,509 and $1,774.

5.3 Matching on the Propensity Score

As suggested in Section 4 an alternative to stratifying on the score is pair-wise matching. By matching each treated unit to the control with the nearest propensity score (with replacement), we focus attention on a much smaller subset of the overall control group. For PSID-1 to 3, 52, 31, and 43 controls are used respectively and for CPS-1 to 3, the number of controls matched to the treated observations are 106, 87, and 63 respectively. The characteristics of the matched control samples are reported in Table 7. Comparing the sample characteristics of the matched sample with unmatched samples in Table 2 shows precisely the result of matching on the propensity score. Columns (7) and (8) of Table 4 present the unadjusted and adjusted treatment effects.[17] The treatment effect varies from $870 to $2,190 (unadjusted) and $826 and $1,740 (adjusted) with PSID controls. With CPS controls, the treatment

[17] Note however that weights need to be used in matched samples to take into account the matching of more than one treated observation to the same control observation. For more details see Dehejia and Wahba (1995).

effect varies from $-466 to $1,445 (unadjusted) and $-372 to $1,589 (adjusted). Again, a researcher following our approach would not need to construct control groups other than the original control group, so that the adjusted estimated treatment effect under matching methods would vary between $1,174 and $1,690. Although the researcher would miss estimates somewhat closer to the experimental benchmark by not using control groups such as PSID-3 and CPS-2, she would also eschew particularly poor results by not using ad hoc sub-groups such as PSID-2 or CPS-3.

5.4 Using the Propensity Score as Weights

The estimates using the score directly in a weighting scheme are presented in column (9) of Table 4. The treatment effects for the PSID-1 and CPS-1 samples are $1,129 and $1,485 respectively. As we vary the control sample (and accordingly re-run the logistic regression), there is noticeable variation in the reported treatment effects, though they do remain positive. It is difficult to give a proper interpretation to the estimates under the reduced samples; the observations dropped from the control group could be those which are least likely to be treated (low score), or if the chosen criterion for reducing the sample is an inappropriate one, the observations dropped could be those that are most likely to be treated (high score). Either way, by removing them from the sample the information they contribute to estimating the score accurately is lost.

The critical issue concerning sub-groups such as those created by Lalonde (1986) to reduce the bias is that forming subsets of the control group based on single characteristics such as employment status between PSID-1 and -2, for example, imposes a lexicographic preference in terms of suitability of matches on that characteristic. Instead, by allowing the score to choose from the full data set, one incorporates all observable characteristics weighted by the probability of selection.

5.5 Estimating the Treatment Effect by Sample Characteristics

A notable feature of the results presented in the previous sections is the high standard errors on the treatment effects. One possible explanation is heterogeneity in the treatment effect, an issue which is of independent interest.[18]

As Tables 5 and 6 indicate, the treatment effect for the blocks vary from a highly positive to a highly negative effect. These estimated treatment effects are for observations with similar propensity scores and socio-economic characteristics within blocks, but different across the blocks. For example, block 6 of Table 6 with a score range of 0.6 to 0.85 is made up of 26 treated observations and 12 control observations all of whom are blacks, with an average age of 26, 10 years of high school, less than a third of them married, with no earnings in 1974, and very low earnings ($250) in 1975. For this group the average training effect was $2,364 (unadjusted) and $3,683 (adjusted). But the blocking cannot provide a sharp characterization of the treatment effect in terms of specific pre-treatment characteristics. This is offered in Tables 8 and 9.

Table 8 (rows 1 and 2) presents the training effect using the two randomized samples of Table 1 and the training effect for sub-groups selected by sample characteristics from the second sample. The treatment effect reported in column 3 is estimated by taking a difference-in-means (the unadjusted treatment effect) of 1978 earnings between the treated and control groups. The relatively high standard error for the treatment effect of $886 (s.e.= 476) suggests the possibility of heterogeneity of the treatment effect among units. The higher treatment effect for the second sample ($1794) is a reflection

of the cohort phenomenon as explained previously. Within the second sample, the variation in training effect is indicated in column (4). Those participants, for example, that completed high school or that have more than 11 years of schooling have a treatment effect which is much higher than the average ($3,085), and significantly different from the treatment effect of their complement (no degree or less than eleven years of schooling). Unemployment in 1974 is an important covariate that distinguishes participants (the treatment effect is $3,376 for those unemployed in 1974 and $-685 for those employed in 1974), whereas whether an individual was employed or not in 1975 makes little difference in terms of treatment effect. The importance of the earnings profile for 1974 in determining the probability of training participation is discussed in Section 6.2 in the context of the sensitivity to the selection on observables assumption. Table 8 exposes a significant degree of heterogeneity in the benchmark training effect, suggesting that a model with a constant treatment effect such as equation (1) can be substantially misleading. In Tables 9a and 9b we see that the potential heterogeneity of treatment effects is readily explored through the estimation strategy followed in this paper, even in the absence of a randomized experiment.

Tables 9a and 9b estimate the treatment effect by sample characteristics using the PSID-1 and CPS-1 controls respectively and the propensity score stratification scheme discussed above. We note that for many of the characteristics the estimates reasonably match the experimental results. This provides added confidence in the accuracy of the non-experimental results, since not only do they track the average treatment effect for the NSW group, but they also track the average treatment effect for sub-sets of the original group. Note however that standard errors are still relatively high for many of the

[18] Another explanation for the high standard errors is that there is minimal overlap between the distributions of the treated and control observations, which implies small sample sizes for a number of the blocks, as indicated in Tables 5

treatment effects controlling for individual characteristics, especially when using PSID-1 as control group.

Thus in summary, using both the PSID and CPS, we estimate treatment effects which come reasonably close to the experimental benchmark. Lalonde's message from his analysis was that the researcher is presented with an array of estimates which differ dramatically (from $-15,205 to $1,326) and with no clear way to choose between them. In contrast, from our array of estimators, the answer which emerges is much more focused. Furthermore, the estimates are based on a simple method for comparing observations as summarized by their propensity score. The flexibility of the approach is also demonstrated in the way it is able to replicate to a large extent non-constant treatment effect embodied in the original experimental data set.

6. Sensitivity Analysis

6.1 Sensitivity to Specification of the Propensity Score

Under the algorithm defined in Appendix B, the choice of interaction terms in the logistic function is entirely determined by the need to balance the covariates within blocks. Table 10 presents various point estimates of the treatment effect with CPS-1 and PSID-1 as control groups, starting with the logit function reported in Table 4 and then excluding higher order terms (squared and cubic) followed by excluding interaction terms from the logit. Although none of the resulting logistic functions completely balance the covariates for equal values of the score (as did the logit function reported in Table 4), the results indicate that the point estimates stratifying on the score are not highly sensitive to logit specifications. Estimates in column (3) where the score enters linearly in the regression are also not very

---

and 6, resulting in higher standard errors than treatment effects estimated directly over the full sample.

sensitive to propensity score specification. This points to the crucial characteristic of our approach, namely that the choice of terms in the logit specification is driven only by the need to balance covariates of observations with similar propensity scores. In contrast, estimating the treatment effect through generalizations of the regression model in equation (1) (e.g., equation (2)) requires prior information on which terms to include. Note also that with CPS controls, standard errors are significantly lower.[19] Sensitivity analysis to starting parameters in the logit for score estimation (see the first step in Figure 1, Appendix B) was also conducted and generally produced the same logit specification. Also, results were not sensitive to changes in the initial blocking rule (see the second step in Figure 1).

6.2 Sensitivity to Selection on Observables

The key assumption driving the above analysis is that all the variables generating assignment to treatment (and correlated with potential outcomes Y_{i1} and Y_{i0}) are observed. It is clear that rarely are all the relevant variables observed by the researcher. Thus, it is of interest to examine how far we can go in removing the bias from the results through conditioning on observables. In this section we examine this issue by excluding pre-treatment earnings in 1974 and re-estimating the treatment effect using the estimators described in Section 4. The results of Table 4 are re-computed and presented in Table 11.

The first apparent difference between Tables 4 and 11 is the sensitivity of the PSID sample to pre-treatment earnings in 1974. When 1974 earnings are dropped, estimates of the training effect are negative with very high standard errors. As expected, the use of Lalonde's PSID-1 and PSID-2 samples does not change the results very much. Estimates of the treatment effect using matching or the

[19]The lower standard errors that come with coarser specification of the logistic function suggest a tradeoff between efficiency and unbiasedness. The properties of the algorithm proposed in Section 3 and any other algorithm need to

score as weights also perform poorly. In contrast, using CPS as a control group results in estimates that are more robust. Stratification on the score with CPS-1 produces an adjusted training effect of $1,207 (s.e.=880). Pair-wise matching on the score also produces a significant effect of $1,969 (s.e.=808), only $175 higher than the benchmark case. The reason for this important difference between the two control groups is found by examining the distribution of earnings in 1974 and 1975 across the propensity score blocks. Whereas earnings in 1974 are not balanced for most of the blocks in the case of PSID-1, the opposite is true for most (but not all) of the blocks with CPS-1. The difference in the two samples comes from relatively different pre-treatment earnings profiles. In PSID-1, earnings in 1975 do not follow closely earnings in 1974, controlling for the propensity score; for higher propensity score levels, earnings in 1975 do not fall as sharply as earnings in 1974, resulting in a negative correlation between the two years. In contrast, earnings in CPS-1 for 1974 and 1975 follow each other closely, both dropping substantially with higher score levels, resulting in a positive correlation across all blocks. With the dip in earnings captured earlier in the CPS sample, dropping earnings in 1974 affects the estimates of the training effect for PSID-1 but not CPS-1 (see Ashenfelter [1974, 1978] and Ashenfelter and Card [1985] on what has been referred to as the "Ashenfelter dip" in earnings prior to enrollment in training programs).

6.3 The Use of More than One Control Group

By comparing the overall results in Table 11 to those in Table 4 the value of using several control groups becomes evident. Whereas a coherent estimate of the treatment effect emerges in Table 4, Table 11 shows that if the researcher did not know that an important covariate was missing, she would report a

---

be evaluated fully, a direction for future research.

treatment effect that varies substantially depending on the control group used. How does one compare the estimates of the treatment effect for two (or more) control groups? In the above analysis we tested sensitivity to the available set of covariates by using our knowledge of the experimental benchmark to see how far we strayed from the true estimate when a key covariate was set aside. In applications, such randomized data sets are not typically available, but though it is more difficult to assess sensitivity to unobservable characteristics, it is not impossible (see Rosenbaum and Rubin [1983b]).

The use of more than one control group provides additional information regarding the sensitivity of the results to unobservable covariates. The practice of using multiple control groups in economics and more specifically in the manpower training literature is not uncommon, but studies generally report only the final control group used in the evaluation.[20] There is however a fundamental difference between sensitivity to the choice of control group *within* a specific data set (and sub-groups obtained from it), as was addressed in the previous section, and *between* two distinct control groups. The former is an issue already addressed by making use of the propensity score. Comparing results between two distinct control groups is a more delicate exercise. Some studies (Rosenbaum [1984, 1987]) suggest that the use of a second control group in non-experimental settings can sometimes help detect the presence of important variables not observed in the data. The intuition is simple, and was illustrated by Tables 4 and 11. When a variable that determines assignment to treatment is not observed there are two possibilities. If the estimated treatment effect across the two samples is quite similar (as in Table 4), this suggests either that all important variables are observed or that the unobserved variable affects the observed covariates of both samples in a similar way. If instead the estimates differ substantially (as in Table 11),

[20] Fraker and Maynard (1987) provide a detailed analysis of the treatment effect for the NSW program using a series of control groups. They conclude that the results are generally sensitive to the choice of control groups.

this suggests quite strongly the presence of some unobserved variable which affects each sample differently. Without an experimental data set, the use of multiple control groups can provide a partial test for the presence of unobserved variables.

7. Conclusion

This paper presents a framework for estimating treatment effects in non-experimental settings when assignment to treatment is assumed to be ignorable conditional on observable characteristics. Drawing from the statistics literature on causal inference analysis, the paper defines the role of the propensity score in identifying the treatment effect with conditionally ignorable assignment. The paper then proposes an algorithm for estimating the propensity score, and three types of estimators of the treatment effect based on the score.

The estimators are evaluated using Lalonde's seminal re-creation of a non-experimental setting. Results show that the estimates of the training effect are close to the benchmark randomized case, and are robust to specification of the control groups defined by Lalonde. By stratifying observations on the score, a researcher need only use the original control groups to estimate the training effect and would report an effect that varies between $1,509 and $1,774 compared to the randomized treatment effect of $1,794. Using estimators based on matching or weighting by the propensity score lead to similar estimates. The paper also evaluates sensitivity to the specification of the propensity score as well as sensitivity to the selection on observables assumption. Results indicate the robustness of the estimated training effect to changes in the benchmark logit specification and to blocking methods. Excluding earnings in 1974 from the analysis affects the estimated training effect when using PSID as control but

less so with CPS, a result that underscores the importance of using more than one control group in non-experimental studies.

In most of the estimates the standard errors are high, and although the heterogeneity of the treatment effect as well as the minimal overlap in the distribution of covariates between treated and control go far in explaining the high standard errors, further research is needed to examine the optimality properties of the rule specifying the score, such as the tradeoff between bias and efficiency. While the results obtained in the paper are specific to the data set, further studies based on non-experimental evaluation of randomized studies should provide additional evidence on the merits of this approach and on how general (or specific) are these methods to the data at hand. This does not deny the importance of randomized experiments; indeed, it is thanks to a randomized data set that such an evaluation was made possible.

The Lalonde paper and the ensuing debate may have cast a negative light on standard econometric methods of evaluating social programs. This paper attempts to rehabilitate the assumption of selection on observables with the use of the propensity score to exploit more carefully the information contained in observable covariates. There are however many settings in which the assumption of selection on observables is not sufficient to identify the treatment. The conclusion to draw from this paper is that even when the researcher suspects that important characteristics are unobserved and that exclusion restrictions that identify the treatment may be available, the self-diagnostic nature of the approach reveals valuable information to the researcher by examining the comparability of the distributions of the treated and control units. The techniques exposed in this paper are powerful enough to sort out which observations from a large pool of potential controls are relevant comparisons to treated units under consideration and to help guide the researcher in other possible directions. Our

argument would be: before recourse to modeling through assumptions on functional forms and distribution -- assumptions on unobservables, which by their very nature are difficult to test in the data -- there is substantial reward in exploring first the information contained in the variables that *are* observed.

References

Angrist, J. (1990). "Lifetime Earnings and the Vietnam Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, 80, 313-335.

------ (1995). "Using Social Security Data on Military Applicants to Estimate the Effect of Voluntary Military Service on Earnings." Massachusetts Institute of Technology, unpublished.

Angrist, J., G.W. Imbens, and D. Rubin (1996). "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444-454.

Ashenfelter, O. (1974). "The Effect of Manpower Training on Earnings: Preliminary Results," in J. Stern and B. Dennis (eds.), *Proceedings of the Twenty-Seventh Annual Winter Meetings of the Industrial Relations Research Association*. Madison: Industrial Relations Research Association.

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

------ 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.

Barnow, B., G. Cain, and Arthur Goldberger (1980). "Issues in the Analysis of Selectivity Bias," *Evaluation Studies*, 5, 42-59.

Burtless, Gary (1995). "The Case for Randomized Field Trials in Economic and Policy Research," *Journal of Economic Perspectives*, 9, 61-84.

------ and Orr, L. (1986). "Are Classical Experiments Needed for Manpower Policy?" *Journal of Human Resources*, 21, 606-639.

Card, David and Daniel Sullivan (1988). "Measuring the Effect of Subsidized Training Programs on Movements in and out of Employment," *Econometrica*, 56, 497-530.

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

Cochran, W. G. (1968). "The Effectiveness of Adjustment by Sub-Classification in Removing Bias in Observational Studies" *Biometrics*, 24, 295-313.

Dehejia, Rajeev H. and Sadek Wahba (1994). "Re-evaluating the Evaluation of Training Programs: On the Methodology of Causal Inference," Harvard University, unpublished.

------ (1995). "An Oversampling Algorithm for Causal Inference in Non-Experimental Studies with Incomplete Matching and Missing Outcome Variables," Harvard University, unpublished.

Fisher, R. (1935). *The Design of Experiments.* London: Oliver and Boyd.

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

Goldberger, Arthur (1972a). "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations," University of Wisconsin, Institute for Research on Poverty, Discussion paper, 123-72.

------ (1972b). "Selection Bias in Evaluating Treatment Effects: The case of Interaction," University of Wisconsin, Institute for Research on Poverty, Discussion paper, 129-72.

Härdle, Wolfgang (1990). *Applied Nonparametric Regression*. Econometric Society Monographs, Cambridge: Cambridge University Press.

Heckman, J. (1979). "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 931-961.

------ (1990). "Varieties of Selection Bias," *American Economic Review*, 80, 313-318.

------ and J. Hotz (1989). "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, 84, 862-880.

------ and Richard Robb (1985). "Alternative Methods for Evaluating the Impact of Interventions", in James Heckman and Burton Singer (eds.), *Longitudinal Analysis of Labor Market Data.* Econometric Society Monograph No. 10, Cambridge: Cambridge University Press.

------ (1986). "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatment on Outcomes," in Howard Rainer (ed.), *Drawing Inferences from Self-Selected Samples*. New York: Springer-Verlag.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd (1995). "Non-Parametric Characterization of Selection Bias Using Experimental Data: A Study of Adult Males in JTPA," University of Chicago, unpublished.

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

Imbens, Guido W. and J. Angrist (1994). "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467-475.

Katz, Lawrence (1992). "Recent Developments in Labor Economics," American Economic Association Meetings, January 1992.

Lalonde, Robert (1984). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," Princeton University, Industrial Relations Section, Working Paper No 183.

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

Leamer, Edward (1983). "Let's Take the Con Out of Econometrics," *American Economic Review*, 73, 31-43.

Maddala, G.S. (1983). *Qualitative and Limited Dependent Variable Models in Econometrics*. Econometric Monograph No. 3, Cambridge: Cambridge University Press.

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

Manski, Charles F. (1995). *Identification Problems in the Social Sciences*. Cambridge: Harvard University Press.

------, G. Sandefur, S. McLanahan, and D. Powers (1992). "Alternative Estimates of the Effect of Family Structure During Adolescence on High School Graduation," *Journal of the American Statistical Association*, 87, 25-37.

Neyman, J. (1935). "Statistical Problems in Agricultural Experimentation," *Journal of the Royal Statistical Society*, supplement, II, 107-180.

Pratt, John and Robert Schlaifer (1988). "On the Interpretation and Observations of Laws," *Journal of Econometrics*, 39, 23-52.

Reinisch, June, Stephanie Sanders, E. Mortensen, and Donald Rubin (1993). "Prenatal Exposure to Phenobarbital and Intelligence Deficits in Adult Human Males," The Kinsey Institute for Research in Sex, Gender and Reproduction, unpublished.

Rosenbaum, P. (1987). "The Role of a Second Control Group in an Observational Study," *Statistical Science*, 2(3).

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

------ (1983b). "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome," *Journal of the Royal Statistical Society*, Series B, vol. 45.

------ (1985a). "Reducing Bias in Observational Studies Using the Subclassification on the Propensity Score," *Journal American Statistical Association*, 79, 516-524.

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

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

------ (1974). "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies," *Journal of Educational Psychology*, 66, 688-701.

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

------ (1978). "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34-58.

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

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

Appendix A. Selection on Observables and the Role of the Propensity Score

Proposition A.1: *If for each unit we observe a vector of covariates* X_i *and*

$$Y_{i1},\ Y_{i0} \perp T_i | X_i,\ \forall i,$$

then:

$$t|_{T=1} \equiv E(Y_{i1}|T_i=1) - E(Y_{i0}|T_i=1)$$

$$= E_X\left\{E\left(Y_i|X_i, T_i=1\right) - E\left(Y_i|X_i, T_i=0\right)\middle|T_i=1\right\},$$

where $Y_i = T_i Y_{i1} + (1-T_i)Y_{i0}$.

Proof:

$$Y_{i1}, Y_{i0} \perp\!\!\!\perp T_i \middle| X_i$$

$$\Rightarrow E\left(Y_{i1}\middle|X_i, T_i=1\right) = E\left(Y_{i1}\middle|X_i, T_i=0\right) = E\left(Y_{i1}\middle|X_i\right),$$

and similarly for Y_{i0}, which allows us to write:

$$t|_{T=1} = E\left(Y_{i1}\middle|T_i=1\right) - E\left(Y_{i0}\middle|T_i=1\right)$$

$$= E_X\left\{E\left(Y_{i1}\middle|X_i, T_i=1\right) - E\left(Y_{i0}\middle|X_i, T_i=1\right)\middle|T_i=1\right\}$$

$$= E_X\left\{E\left(Y_{i1}\middle|X_i, T_i=1\right) - E\left(Y_{i0}\middle|X_i, T_i=0\right)\middle|T_i=1\right\}$$

$$= E_X\left\{E\left(Y_i\middle|X_i, T_i=1\right) - E\left(Y_i\middle|X_i, T_i=0\right)\middle|T_i=1\right\}$$

$$= E_X\left\{\left(t_{T=1,X}\right)\middle|T_i=1\right\},$$

where $t_{T=1,X} \equiv E(Y_i|X_i, T_i=1) - E(Y_i|X_i, T_i=0)$.

•

Proposition A.2: *Let* $p(X_i)$ *be the probability of unit* i *having been assigned to treatment, defined as* $p(X_i)=Pr(T_i=1|X_i)=E(T_i|X_i)$, *where* $0<p(X_i)<1$, $\forall i$. *Then:*

$$(Y_{i1}, Y_{i0}) \perp\!\!\!\perp T_i \middle| X_i$$

implies

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

Proof:

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

Hence,

$$\Rightarrow T \perp\!\!\!\perp Y_1, Y_0 | p(X).$$

•

Corollary A.2.1: $t\,|_{T=1} = E_{p(X)}\left\{E\left(Y_i|T_i=1, p(X_i)\right) - E\left(Y_i|T_i=0, p(X_i)\right)\middle|T_i=1\right\}$.

Proof:

$$\begin{aligned}
t\Big|_{T=1} &= E\left(Y_{i1}|T_i=1\right) - E\left(Y_{i0}|T_i=1\right) \\
&= E_{p(x)}\left\{E\left(Y_{i1}\middle|T_i=1, p(X_i)\right) - E\left(Y_{i0}\middle|T_i=1, p(X_i)\right)\middle|T_i=1\right\} \\
&= E_{p(x)}\left\{E\left(Y_{i1}\middle|T_i=1, p(X_i)\right) - E\left(Y_{i0}\middle|T_i=0, p(X_i)\right)\middle|T_i=1\right\} \\
&= E_{p(x)}\left\{E\left(Y_i\middle|T_i=1, p(X_i)\right) - E\left(Y_i\middle|T_i=0, p(X_i)\right)\middle|T_i=1\right\} \\
&\equiv E_{p(x)}\left\{t\,|_{T=1, p(X)}\,\middle|T_i=1\right\}.
\end{aligned}$$

•

Appendix B. 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, e.g., logit or probit. It is important to remember that the role of the score is only in reducing the dimensions of the conditioning, and, 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:

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

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.[21]

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 propose to estimate the propensity score (see Rosenbaum and Rubin 1983a):

Proposition A.3:

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

Proof: From the definition of $p(X)$ in Proposition A.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.3 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.[22] When the covariates are not balanced within a particular block, the block may be too coarsely defined; recall that Proposition A.3 in fact deals with observations with an identical 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

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

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

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 for these characteristics better. This procedure is repeated for each given block until covariates are balanced. [23] The algorithm is summarized in Figure 1.

Figure 1 - 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 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: difference-in-means for all covariates of treated and control in all blocks are not significant from zero at relevant confidence level.

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 logit by adding interaction terms and/or higher order terms of 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 which 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. Once the propensity score is estimated the treatment effect can be obtained in a number of ways.

Appendix C

Proposition A.4:

$$\frac{1}{N^T}\sum_{i=1}^{N^T+N^C}\left(T_iY_i-(1-T_i)\frac{p(X_i)}{1-p(X_i)}Y_i\right)$$

is a consistent estimator of $t|_{T=1}$*, where* Y_i *is the observed value of outcome for unit* i*,* T_i *is an index variable (=1 if treated and =0 if control),* $0<p(X_i)<1$*, and* N^T *and* N^C *are the total number of treated and control observations respectively.*

Proof: In the population consider q, where:

$$\begin{aligned}
q &\equiv E\left(T_iY_i-\frac{p(X_i)}{1-p(X_i)}(1-T_i)Y_i\right)\\
&= E_X\left(E\left(T_iY_i\middle|X_i\right)-\frac{p(X_i)}{1-p(X_i)}E\left((1-T_i)Y_i\middle|X_i\right)\right)\\
&= E_X\left(p(X_i)E\left(Y_{i1}\middle|X_i\right)-p(X_i)E\left(Y_{i0}\middle|X_i\right)\right)\\
&= \int p(X_i)\left(E\left(Y_{i1}\middle|X_i\right)-E\left(Y_{i0}\middle|X_i\right)\right)f(X_i)d(X_i)\\
&= \int \Pr(T_i=1)\left(E\left(Y_{i1}\middle|X_i\right)-E\left(Y_{i0}\middle|X_i\right)\right)f(X_i|T_i=1)d(X_i)\\
&= \Pr(T_i=1)E_{X|T_i=1}\left(E(Y_{i1}-Y_{i0}|X_i)\right).
\end{aligned}$$

Thus, $t\big|_{T=1}=q/\Pr(T_i=1)$, and the sample analogue of t is:

$$\frac{1}{N^T}\sum_{i=1}^{N^T+N^C}\left(T_iY_i-\frac{p(X_i)}{1-p(X_i)}(1-T_i)Y_i\right).$$

•