Data Issues of Using Matching Methods to Estimate Treatment Effects: An Illustration with NSW Data Set*

Zhong Zhao†

China Center for Economic Research (CCER)

Peking University

July 9, 2003

Keywords: Treatment Effect, Data Issues**,** Matching, Propensity Score

JEL Classification: C14, C21, I38

* I am grateful to Robert Moffitt, Geert Ridder, Carl Christ, Bruce Hamilton, and other seminar participants at Johns Hopkins, Washington University in St. Louis, Manpower Demonstration Research Corporation and Abt, Inc. for their helpful comments. All errors are mine. I would like to thank Rajeev Dehejia for making the data set used in this paper available through his web page.

† Mailing Address: China Center for Economic Research, Peking University, Beijing 100871, China. Email: zzhao@ccer.pku.edu.cn. Tel: 86-10-62758915. Fax: 86-10-62751474

Data Issues of Using Matching Methods to Estimate Treatment Effects: An Illustration with NSW Data Set

Abstract

In this paper, we study data issues of using matching estimators to estimate treatment effect. We first demonstrate that with proper data set, the matching assumptions can be justified for voluntary programs. Next we compare covariate matching and propensity score matching methods, and show that they do not dominate each other in term of data requirement. Finally we use the National Supported Work Demonstration data set to illustrate the issues discussed above.

I. Introduction

How to use observational data (or non-experimental data) to estimate treatment effects is a perpetual research theme in economics, e.g. Heckman (1976, 1979, 1990, 2000, 2001), Barnow et al (1980), Bjorklund and Moffitt (1987) among others. Recently, matching methods attract a lot of attention from economists under the assumption of selection on observables, also known as exogeneity assumption or unconfoundedness assumption (we will formalize this assumption in Section II). Papers in this field by economists include Abadie and Imbens (2002), Angrist (1998), Angrist and Hahn (1999), Dehejia and Wahba (1999, 2003), Hahn (1998), Heckman et al (1997, 1998), Heckman et al (1998), Imbens (2001, 2002), Lechner (2002), Smith and Todd (2001, 2003), and etc.

When the selection bias is only due to observables, matching is a useful tool to estimate treatment effect. The most attractive feature of matching, compared with the regression type estimators, such as that of Barnow et al (1980), is its non-parametric nature. Matching neither imposes functional form restrictions such as linearity nor assumes a homogeneous treatment effect in the population. Both assumptions are usually not justified either by economic theory or by the data. The first issue in this paper is to explore the plausibility of the assumptions for matching estimators. We argue that if proper variables, such as the information on the application indicator and the decision variables used by the program administration, are collected, the assumptions for matching estimators can be justified for voluntary programs.

Using covariate matching to correct the bias due to observables is intuitive, since the source of the bias is the difference of observables in the treated group and comparison group. Matching on covariates by definition will remove this difference and hence the

bias. When there are many covariates, it is impractical to match directly on covariates because of the curse of dimensionality. Taking the study of the Comprehensive Employment and Training Act by Westat (1981) as an example, for controlling only 12 covariates, the covariate matching scheme of Westat led to more than 6 million cells. Since the number of observations is far less than 6 million, most of cells are empty and it is very hard to find a good match on all 12 covariates. It is usually necessary to map the high dimension of covariates into a scalar through some metric, which measures the closeness of two observations. The most often used metric is Mahalanobis metric, e.g. Rubin (1980).

Another way to reduce the dimensionality is through propensity score matching. Rosenbaum and Rubin (1983) show that while covariate is the finest balancing score, propensity score is the coarsest balancing score. A balance score, $b(x)$, is a function of the observed covariates such that the conditional distribution of x given $b(x)$ is the same for treated and comparison groups, as defined in Rosenbaum and Rubin (1983). So matching on covariates and matching on the propensity score will both make the distribution of the covariates in the treated group the same as the distribution of the covariates in the comparison group.

Covariate matching faces the curse of dimensionality, and often encounters empty or small matching cells, while propensity score matching reduces the high dimension of covariates to a scalar, and can also balance the observables between treated group and comparison group. It is natural to ask whether propensity score matching needs less data (requires few observations) than covariate matching? This is the second issues considered in this paper.

The last issue in this paper is using the National Supported Work Demonstration (NSW) data set to illustrate the data issues of matching as well as to show that propensity score matching and covariate matching do not dominate each other. Their performance, like any other econometric evaluation method, crucially depends on the data set.

The remaining paper is organized as follows. Section 2 sets up the model using the potential outcome framework, Section 3 explores the plausibility of the assumptions for matching estimators, Section 4 studies data issues of covariate and propensity score matching methods, Section 5 is an illustration using the NSW data set, and Section 6 concludes the paper.

II. Model Setup

A fruitful framework for estimating treatment effects is the potential outcome framework dated back to Neyman (1923) and is widely used both in economics and statistics, such as, Roy (1951), Quandt (1972), Rubin (1974) and Holland (1986).

In the potential outcome framework, each individual has two potential outcomes (Y_{0i}, Y_{1i}) for a treatment, such as job training, education, or a welfare program. Y_{1i} is the outcome if individual i is treated and Y_{0i} is the outcome if individual i is not treated. Let $T_i = 1$ indicate that individual i is treated and $T_i = 0$ indicate otherwise. With (Y_{0i}, Y_{1i}) we can define different treatment effects, such as those in Heckman and Vytlacil (1999), as follows:

$\Delta_i = Y_{1i} - Y_{0i}$	Treatment Effect for Individual i
$\Delta_{ATE} = E[\Delta_i]$	Average Treatment Effect for the Population (ATE)
$\Delta_S = E[\Delta_i \mid i \in S]$	Average Treatment Effect for the Sub-Population S

When $S = \{i : T_i = 1\}$, Δ_S is the treatment effect on the treated (TT), denoted as Δ_{TT}.

The average treatment effect at population (or sub-population) level can be estimated without bias either by experimental data or by observational data if the selection bias is only due to observables.

That the selection bias is only due to observables is formally characterized by the following two assumptions:

$M-1$: (Y_0, Y_1) $\underline{\perp\!\!\!\perp}$ $T \mid X$ Conditional Independence Assumption

$M-2: 0 < prob(T = 1 \mid X) < 1$ Common Support Assumption

where $\underline{\perp\!\!\!\perp}$ is the notation for statistical independence as in Dawid (1979). $M-1$ is also commonly referred as unconfoundedness assumption or exogeneity assumption.

Under $M-1$ and $M-2$

$$\Delta_{TT} = E_{x|T=1}\{E[Y_1 \mid T = 1, X = x] - E[Y_0 \mid T = 1, X = x]\}$$

$$= E_{x|T=1}\{E[Y_1 \mid T = 1, X = x] - E[Y_0 \mid T = 0, X = x]\} \qquad (1)$$

Unbiased estimates of $E[Y_1 \mid T = 1, X = x]$ and $E[Y_0 \mid T = 0, X = x]$ can be obtained from the data and hence so can Δ_{TT}. This is also true for Δ_{ATE} and for other Δ_S.

Using the so called balancing property:

$$prob(X_i \mid T_i = 1, p(X_i) = p) = prob(X_i \mid T_i = 0, p(X_i) = p) = prob(X_i \mid p)$$

Rosenbaum and Rubin (1983) prove that $M-1$ and $M-2$ imply

$P-1$: (Y_0, Y_1) $\underline{\perp\!\!\!\perp}$ $T \mid p(X)$, and

$P-2: 0 < prob(T = 1 \mid p(X)) < 1$

Following from $P-1$ and $P-2$:

$$\Delta_{TT} = E_{p|T=1}\{E[Y_1 \mid T=1, p(X)=p] - E[Y_0 \mid T=1, p(X)=p]\}$$

$$= E_{p|T=1}\{E[Y_1 \mid T=1, p(X)=p] - E[Y_0 \mid T=0, p(X)=p]\} \qquad (2)$$

Unbiased estimates of $E[Y_1 \mid T=1, p(X)=p]$ and $E[Y_0 \mid T=0, p(X)=p]$ can also be obtained if $p(X)$ is known. The advantage of formula (2) over formula (1) is that instead of controlling for a high-dimensional vector of X, formula (2) only needs to control for a scalar p.

III. The Plausibility of Matching Assumptions

Before going any further, an obvious question is whether the assumptions of $M-1$ and $M-2$ are plausible. Unfortunately in general there is no unambiguous answer to this question. Whether $M-1$ and $M-2$ are plausible or not needs to be argued case by case, and their plausibility depends on many factors, such as the richness of the data set, the nature of the treatment, the treatment effect under estimation, etc. The empirical evidence also suggests that there is no clear-cut answer for this question. Dehejia and Wahba (1999) successfully replicate the experimental benchmark by propensity score matching methods using the NSW data set. But Heckman et al (1998) reject the assumptions of matching estimators and assumptions of their extension, difference-in-difference matching, using the JTPA data set. They also find that selection bias due to observables is much larger than the one due to unobservables. Their work suggests that controlling for the bias due to observables is more important than controlling for the bias due to unobservables. Even if $M-1$ and $M-2$ are not justifiable and there is no prior knowledge on the magnitude of the bias due to unobservable and the bias due to observable, it is still useful to apply matching methods

to eliminate the bias due to observables first and then use different procedures to address the bias due to unobservables.

Nevertheless $M-1$ and $M-2$ can be justified for a wide variety of applications if a proper data set is available. In the United States as well as in many other countries, like countries in the European Union, participation of the social programs is often voluntary. For these programs, the treatment status is the result of two decisions: the application decision made by each individual and the admission decision made by the program administration. Let A_i and B_i be the indicators of these two decisions, so $T_i = A_i B_i$. Under this scenario different treatment effects can be defined for the people who apply for the program (i.e. for i with $A_i = 1$):

$\Delta_{ATT-A} = [\Delta_i \mid A_i = 1]$ Average Treatment Effect for the Applicants

$\Delta_{TT-A} = E[\Delta_i \mid A_i = 1, T_i = 1]$ Treatment Effect on the Treated for Applicants

$\Delta_{UTT-A} = E[\Delta_i \mid A_i = 1, T_i = 0]$ Treatment Effect on the Untreated for Applicants

Essentially these are the treatment effects we are interested in,[1] except in some special cases such as if we try to evaluate what will happen if a program is expanded to cover the whole population and is changed from a voluntary program into a mandatory program. In that case we also need to know the treatment effect for the non-applicants. These treatment effects for the applicants can be used to answer many interesting questions, for instance, can the benefit of a program cover its cost? What will happen if the coverage of a program is expanded?

[1] For the non-applicants, since they never participate in the program, their responses to the program have little policy interest.

$M-1$ and $M-2$ for these treatment effects for the applicants can be justified if A_i and B_i can be observed separately. Taking Δ_{TT-A}, treatment effect on the treated for applicants, as an example, it can be written as:[2]

$$\begin{aligned}\Delta_{TT-A} &= E[\Delta_i \mid A_i = 1, T_i = 1] \\ &= E[Y_{1i} - Y_{0i} \mid A_i = 1, B_i = 1] \\ &= \{E[Y_{1i} \mid A_i = 1, B_i = 1] - E[Y_{0i} \mid A_i = 1, B_i = 0]\} \\ &\quad + \{E[Y_{0i} \mid A_i = 1, B_i = 0] - E[Y_{0i} \mid A_i = 1, B_i = 1]\}\end{aligned}$$

First, $\Delta_{TT-A} = \Delta_{TT}$, i.e. the treatment effect on the treated for the applicants is the same as the treatment effect on the treated for the whole population, since both of them are the average treatment effect for the same group of people.[3] Second, the first term $\{E[Y_{1i} \mid A_i = 1, B_i = 1] - E[Y_{0i} \mid A_i = 1, B_i = 0]\}$ is directly estimable from the data. The second term $\{E[Y_{0i} \mid A_i = 1, B_i = 0] - E[Y_{0i} \mid A_i = 1, B_i = 1]\}$ is the bias term and is due to observables if the program administration decision variables are collected, i.e. the program decision indicator, B_i, is independent of Y_{0i} conditioning on the decision variables of B_i.

When applying propensity score matching methods, we need to decide which *X* should be included in the propensity score so *P*-1 will be satisfied and what functional form the propensity score should have. The balancing test in Dehejia and Wahba (1999) is useful to find the functional form of the propensity score but cannot tell which *X* should be used, see Smith and Todd (2003). If we have sufficient knowledge of the

---

[2] Δ_{UTT-A} can be justified in the same manner and Δ_{ATT-A} can be written as a weighted average of Δ_{TT-A} and Δ_{UTT-A}.

[3] Generally $\Delta_{ATT-A} \neq \Delta_{ATE}$ and $\Delta_{UTT-A} \neq \Delta_{UTT}$, where $UTT = \{i : T_i = 0\}$.

program and restrict our attention on the applicants, we need only include the X used by the program administrator to select the applicants into the treatment.

The above argument relies on the assumption that we have the information on A_i and B_i. Instead of focusing on devising different kinds of estimators based on ultimately untestable assumption(s), a feasible and more reliable alternative is to collect good data.[4] The application status indicator A_i should be (but unfortunately has not been) included in many data sets, such as the Survey of Income and Program Participation (SIPP).

IV. Data Issues of Covariate Matching and Propensity Matching Estimators

The common approaches to control for the bias due to observable variables in the matching literature include matching on covariates or on the propensity score, sub-classification by covariates or by the propensity score, and weighting by the propensity score. Imbens (2003) provides an excellent survey. We will focus on one-to-one matching estimators since one-to-one matching estimators are widely used in the empirical studies and it is important to understand their properties. One-to-one matching involves selecting a single observation from the comparison sample to match each observation in the treated sample by some metric. Though matching on covariates or on the propensity score can both remove the bias due to observables, if there are many covariates, especially continuous ones, matching on covariates runs into the curse of dimensionality. Since the work of Rosenbaum and Rubin (1983), propensity score matching has dominated the literature on matching. In most cases, it is easier to find observations with similar values of propensity score than with similar values of covariates, as argued in Rosenbaum (1995). Does this mean that propensity score matching requires fewer observations than

[4] See Moffitt (1991) and Heckman et al (1998) on the importance of good (rich) data from other perspectives.

covariate matching? In order to answer this question, we need to examine more closely how covariate matching and propensity score matching work.

Define the two potential outcome equations and the selection equation as:

$Y_{1i} = f_1(X_i) + \varepsilon_{1i}$, ε_{1i} is iid with $E[\varepsilon_{1i} \mid X_i] = 0$;

$Y_{0i} = f_0(X_i) + \varepsilon_{0i}$, ε_{0i} is iid with $E[\varepsilon_{0i} \mid X_i] = 0$;

$T_i = I(T_i^* > 0)$, $I(\cdot)$ is the indicator function; and

$T^* = h(X_i) + \mu_i$, μ_i is iid with $E[\mu_i \mid X_i] = 0$ and CDF $G(\cdot)$.

The basic ideas of covariate matching are:

(1) $X_i = X_j \Rightarrow f_t(X_i) = f_t(X_j)$, $t = 0, 1$; and

(2) $d(X_i, X_j) < \varepsilon \Rightarrow d'(f_t(X_i), f_t(X_j)) < \delta$, $t = 0, 1$, where d and d' are some metrics in mathematical sense.

Assumption (1) justifies exact matching. Assumption (2) means that f_t is continuous at X and it justifies neighborhood matching.

For simplicity, we assume that exact matching is possible.[5] Through covariate matching, the observation i in the treated sample is matched with the observation j in the comparison sample if $X_i = X_j = x$. Define:

$$\begin{aligned}\hat{\Delta}_i^C &= Y_{1i} - Y_{0j} \\ &= f_1(X_i) + \varepsilon_{1i} - f_0(X_j) - \varepsilon_{0j} \\ &= \{f_1(x) + \varepsilon_{1i} - [f_0(x) + \varepsilon_{0i}]\} + \{\varepsilon_{0i} - \varepsilon_{0j}\} \\ &= \Delta_i + \{\varepsilon_{0i} - \varepsilon_{0j}\}\end{aligned}$$

[5] When exact matching is impossible, the discussions are still approximately true if we can match on some sufficiently small neighborhood of X.

where Δ_i is the true treatment effect for individual i. Denote $m(x)$ as the number of matching pairs in an $x-cell$ which have the same covariate x; then the average treatment effect at x can be estimated by:

$$\hat{\Delta}^C(x) = \frac{1}{m(x)}\sum_{i=1}^{m(x)}\hat{\Delta}_i^C$$

$$= \frac{1}{m(x)}\sum_{i=1}^{m(x)}\Delta_i + \frac{1}{m(x)}\sum_{i=1}^{m(x)}\varepsilon_{0i} - \frac{1}{m(x)}\sum_{j=1}^{m(x)}\varepsilon_{0j}$$

$$= \Delta(x) + \frac{1}{m(x)}\sum_{i=1}^{m(x)}\varepsilon_{0i} - \frac{1}{m(x)}\sum_{j=1}^{m(x)}\varepsilon_{0j}$$

where $\Delta(x)$ is the average treatment effect at x. It is clear that $\hat{\Delta}^C(x)$ is an unbiased and consistent estimator of the average treatment effect at x. Let r^C be the number of covariate matching cells, and let $n^C = \sum_{r^C} m(x)$ be the total number of matched pairs in the whole sample (n^C also equals the number of observations in the treated sample). The treatment effect on the treated can be estimated by the following estimator:

$$\hat{\Delta}_{TT}^C = \frac{1}{n^C}\sum_{i=1}^{n^C}\hat{\Delta}_i^C$$

$$= \frac{1}{n^C}\sum_{i=1}^{n^C}\Delta(x) + \frac{1}{n^C}\sum_{i=1}^{n^C}\varepsilon_{0i} - \frac{1}{n^C}\sum_{j=1}^{n^C}\varepsilon_{0j}$$

$$= \Delta_{TT} + \frac{1}{n^C}\sum_{i=1}^{n^C}\varepsilon_{0i} - \frac{1}{n^C}\sum_{j=1}^{n^C}\varepsilon_{0j}$$

The closeness of the covariates of each matching pair plays a crucial role in the covariate matching and itself is enough to guarantee the reliability of the estimator under the assumptions of $M-1$ and $M-2$ and the continuity of f_t.

The theory behind propensity score matching is quite different from covariate matching.

The basic ideas of propensity score matching are:

(1) $prob(X_i \mid T_i = 1, p(X_i) = p) = prob(X_i \mid T_i = 0, p(X_i) = p) = prob(X_i \mid p)$,

the balancing property; and

(2) $d(p_k, p_l) < \varepsilon \Rightarrow d'(prob(X_i \mid p_k), prob(X_j \mid p_l)) < \delta$

These two ideas are parallel to the two ideas of covariate matching. Assumption (1) says that when the matching is exact at the propensity score p, then the distribution of X will be the same for the treated sample and the comparison sample at p. Assumption (2) says if exact matching is impossible and instead matching is on some neighborhood of p, the distribution of X is still approximately the same for the treated sample and the comparison sample within the neighborhood of p.

In the propensity score matching methods, the observation i in the treated sample is matched with the observation j in the comparison sample if $p(X_i) = p(X_j) = p$. Define:

$$\begin{aligned}
\hat{\Delta}_i^P &= Y_{1i} - Y_{0j} \\
&= f_1(X_i) + \varepsilon_{1i} - \{f_0(X_j) + \varepsilon_{0j}\} \\
&= \{f_1(X_i) + \varepsilon_{1i} - [f_0(X_i) + \varepsilon_{0i}]\} + \{f_0(X_i) - f_0(X_j) + \varepsilon_{0i} - \varepsilon_{0j}\} \\
&= \Delta_i + \{f_0(X_i) - f_0(X_j) + \varepsilon_{0i} - \varepsilon_{0j}\}
\end{aligned}$$

Using $\hat{\Delta}_i^P$ as the building block, and denoting $m(p)$ as the number of matching pairs in a $p-cell$ which has the same propensity score p, we can estimate the average treatment effect at p by:

$$\hat{\Delta}^p(p) = \frac{1}{m(p)}\sum_{i=1}^{m(p)}\hat{\Delta}_i^p$$

$$= \frac{1}{m(p)}\sum_{i=1}^{m(p)}\Delta_i + \frac{1}{m(p)}\sum_{i=1}^{m(p)}\varepsilon_{0i} - \frac{1}{m(p)}\sum_{j=1}^{m(p)}\varepsilon_{0j} + \frac{1}{m(p)}\sum_{i=1}^{m(p)} f_0(x_i) - \frac{1}{m(p)}\sum_{j=1}^{m(p)} f_0(x_j)$$

$$= \Delta(p) + \frac{1}{m(p)}\sum_{i=1}^{m(p)}\varepsilon_{0i} - \frac{1}{m(p)}\sum_{j=1}^{m(p)}\varepsilon_{0j} + \frac{1}{m(p)}\sum_{i=1}^{m(p)} f_0(x_i) - \frac{1}{m(p)}\sum_{j=1}^{m(p)} f_0(x_j)$$

where $\Delta(p)$ is the average treatment effect at p. $\hat{\Delta}^p(p)$ is an unbiased and consistent estimator of the average treatment effect at p.

Let r^p be the number of propensity score matching cells, and $n^p = \sum_{r^p} m(p)$ be the number of matched pairs in the whole sample (n^P also equals the number of observations in the treated sample). We can estimate the treatment effect on the treated by the following estimator, which is widely used in the matching literature:

$$\hat{\Delta}_{TT}^P = \frac{1}{n^p}\sum_{i=1}^{n^p}\hat{\Delta}_i^P$$

$$= \frac{1}{n^p}\sum_{i=1}^{n^p}\Delta_i + \frac{1}{n^p}\sum_{i=1}^{n^p}\varepsilon_{0i} - \frac{1}{n^p}\sum_{j=1}^{n^p}\varepsilon_{0j} + \frac{1}{n^p}\sum_{i=1}^{n^p} f_0(x_i) - \frac{1}{n^p}\sum_{j=1}^{n^p} f_0(x_j)$$

$$= \Delta_{TT} + \{\frac{1}{n^p}\sum_{i=1}^{n^p}\varepsilon_{0i} - \frac{1}{n^p}\sum_{j=1}^{n^p}\varepsilon_{0j}\} + \{\frac{1}{n^p}\sum_{i=1}^{n^p} f_0(x_i) - \frac{1}{n^p}\sum_{j=1}^{n^p} f_0(x_j)\}$$

It is clear that $\hat{\Delta}_{TT}^{P}$ is also an unbiased estimator for Δ_{TT}. The second term, $\{1/n^p \sum_{i=1}^{n^p} \varepsilon_{0i} - 1/n^p \sum_{j=1}^{n^p} \varepsilon_{0j}\}$, will go to zero as sample size goes to infinity. The third term, $\{1/n^p \sum_{i=1}^{n^p} f_0(x_i) - 1/n^p \sum_{j=1}^{n^p} f_0(x_j)\}$, needs to be balanced out.

It is very possible that individuals with the same propensity score will have very different treatment outcomes, i.e. p approximately the same does not imply X hence doesn't imply treatment outcome, $f(X)$, approximately the same. Because of the balancing property this will not be a problem if the number of observations at each propensity score is large. This can be easily seen if we compare propensity score matching methods to a randomized experiment. The foundation of a randomized experiment is $prob(X, \nu \mid treated) = prob(X, \nu \mid control)$ where X is observable and ν is unobservable. The balancing property plays a similar role in propensity score matching, but propensity score matching methods differ from randomization in two important ways. First, a randomized experiment balances the distributions of both observables and unobservables between treated and comparison samples, but propensity score matching only balances the observables. This is why the independence assumption $M-1$ is needed. Second, a randomized experiment balances the distributions for the whole sample, but propensity score matching balances the distributions at each individual propensity score value. In other words, under $M-1$ and $M-2$, the matched sample at each propensity score value p is equivalent to a randomized sample. The estimate of propensity score matching can be thought as a weighted average of the estimates from many mini randomized experiments (at different p's). The overall quality of the estimation depends on the quality of each of these mini randomized experiments. A

substantial sample size is needed to obtain a meaningful estimate from a randomized experiment and this is translated into a sufficiently large sample size at each p for a meaningful propensity score matching estimate.

When comparing covariate matching with propensity score matching, we note that the advantage of propensity score matching over covariate matching is often characterized by dimensionality reduction, which composes two aspects. One aspect of dimensionality reduction is that instead of controlling high-dimension X, controlling the propensity score p, a scalar, is enough. Nonetheless the data requirement we discussed is related to the other aspect of dimensionality reduction, namely in general the number of *p-cell*s, r^p is less than the number of *X-cell*s, r^C (also see the discussion in Angrist and Hahn 1999). Let us consider two polar cases. One polar case is a randomized experiment. This is the strongest case for the propensity score matching. Since $p(X_i)$ is the same for every individual in the randomization, r^p is 1. The dramatic reduction of the data requirement for the randomized experiment is the result of the drastic reduction of r^p compared with r^C. The other polar case is that in which the correspondence between p and X is one-to-one. In this case if exact matching is possible, then matching on the propensity score or on covariates is equivalent and both require same amount of data since in this case people with same *X* must have same *p*, and vice versa. If exact matching is impossible and instead we match on some neighborhood of the propensity score, the story is different. We note that it is a fact there does *not* exist a one-to-one and bi-continuous (i.e. both the function and its inverse function are continuous) correspondence between R^n space and R^1 space, i.e. R^n space and R^1 space are not a homoeomorphism. It is natural to assume that $p(X)$ is a continuous function of X and this implies that

$p^{-1}(X)$ is *not* a continuous function of p. The implication of this mathematical fact is shown in Figure 1. On the one hand, if $X's$, like X_1 and X_2, lie in the set *A*, then their $p(X)$'s, i.e. p_1 and p_2, must lie in the set *B* (this follows from the continuity of $p(X)$). On the other hand, there must be always some X's, e.g. X_3 and X_4, that lie outside the set *A*, but whose $p(X)$'s, i.e. p_3 and p_4, are in the set *B* (this follows from the discontinuity of $p^{-1}(X)$). Their corresponding treatment outcomes can be quite different from the ones in the set *A*. Matching by the propensity score on some neighborhood of the propensity score has the risk of matching p_1 with p_4, whose $f(X_1)$ and $f(X_4)$ are quite different even though their propensity scores are similar and the correspondence between X and p is one to one. To average this kind of mismatching out, propensity score matching must rely on the balancing property and needs the neighborhood of p to contain a relatively large number of observations. In this case, the advantage of matching on covariates is obvious.

Whether propensity score matching needs less data hinges on how large the difference between r^p and r^C is. Briefly, the reduction of the data requirement of propensity score matching relies on the reduction of the cell number. The reduction of the cell number creates the risk of mismatching. To average out the risk of mismatching requires a large cell size $m(p)$. The combination of r^p, r^C and $m(p)$ ultimately determines the relative data requirements between propensity score matching and covariate matching. Propensity score matching and covariate matching do not dominate each other regarding the data requirement.

V. An Illustration with the National Supported Work Demonstration Data Set

The NSW Demonstration is a randomized experiment conducted from 1975 to 1980 to estimate the effects of a "supported" work experience on the disadvantaged population, such AFDC recipients and ex-offenders. This experiment has 3,214 observations in the treated sample and 3,402 in the control sample. The NSW data set has played an important role in the treatment effect literature. Lalonde (1986) uses a subset of the NSW data set, combined with the PSID and the CPS data, to evaluate different non-experimental estimators. He uses the estimate from the NSW data set as the benchmark, then drops the control group in the NSW data and constructs other comparison groups from the CPS and the PSID. Different estimators have been applied to the constructed data sets. Different estimators have produced very different estimates and often have failed to replicate the benchmark. Fraker and Maynard (1987) use a similar approach but emphasize the sensitivity of estimates to the selection of the comparison groups from the NSW and the CPS data, and they reach a similar conclusion as LaLonde. As a response, Heckman and Hotz (1989) devise tests to aid the choice among estimators. Using propensity score methods, Dehejia and Wahba (1999) successfully replicate the benchmark result, but Smith and Todd (2003) suggest that the success of propensity score matching methods in Dehejia and Wahba (1999) has something to do with the data selected by Dehejia and Wahba instead of the propensity score matching method, per se.[6]

The data sets used in LaLonde (1986), Dehejia and Wahba (1999) and Smith and Todd (2003) are different. The Dehejia and Wahba data set is a subset of LaLonde data set and it excludes the observations with missing earnings variable in 1974. Smith and

[6] The data set in Heckman and Hotz (1989) is different from the one in LaLonde (1986), though both are subsets of the NSW data set.

Todd data set is a subset of Dehejia and Wahba data set and it excludes the observations that were randomized after April of 1976 and Smith and Todd argues that including these observations is problematic.

Similar to their work, we apply propensity score matching methods, Mahalanobis metric, and other two matching metrics proposed in Zhao (2003) and discussed in Imbens (2003) to both the LaLonde data set and the Dehejia and Wahba data set.

The first metric considered in Zhao (2003) and Imbens (2003) is as follows: let the propensity score $p(X) = G(X\beta')$, and consider the following metric:

$$d_1 = \sum_{k=1}^{K} | X_{ki} - X_{kj} | \cdot | \beta_k |$$

This metric incorporates information on both X and p, and weights each coordinate of X by its marginal effect on the propensity score.[7]

The second metric is incorporated outcome information. Assume (Y_{0i}, Y_{1i}) and X have linear relationships, such that

$$Y_{ti} = f_t(X_i) + \varepsilon_{ti} = X_i \alpha_t' + \varepsilon_{ti}, \ t = 0, 1$$

Define the metric as:

$$d_2 = \sum_{k=1}^{K} | X_{ki} - X_{kj} | \cdot | \alpha_{kt} |$$

This metric weights the coordinates of X by their marginal effects to the potential outcomes. It is a natural measurement for closeness of two observations in term of their potential outcomes.

[7] Strictly speaking, β can be interpreted as the marginal effect only if it is estimated from the LPM. For other models, like probit and logit, though β is not the marginal effect but it is still proportioned to the marginal effect.

We refer matching by metrics d_1 and d_2 as covariate & propensity score matching and covariate & outcome matching, respectively. All treatment effects are estimated by one to one matching. Since the CPS is more representative than the PSID and since we want to examine the effectiveness of different matching methods and do not want other sample selection procedures to contaminate the matching process, our estimation is focused on the whole CPS sample, i.e. CPS-SSA-1 in LaLonde (1986) and CPS-1 in Dehejia and Wahba (1999).

Table 1 shows results from different matching metrics using the Dehejia and Wahba data set. The propensity score specification in Table 1 is the same as the specification in Dehejia and Wahba (1999). Measured by the closeness to the benchmark, results from different metrics are very similar and there is no evidence that one estimator dominates the other estimator. Matching without replacement performs more poorly than matching with replacement, which is consistent with Dehejia and Wahba (1999). Imposing the common support condition has little effect on the results. For some estimators it increases the bias, and for others it reduces the bias, but these changes are small. This is not surprising. Unlike sub-classification estimators, one-to-one matching estimators automatically solve the common support problem.

As discussed earlier, the estimate from propensity score matching is the weighted average of the estimates at different propensity score values. The overall quality of the estimation relies on the quality of estimation at each propensity score value. Besides the final estimate, it is also interesting to examine more closely the intermediate estimates. Figure 2 shows the propensity scores of the matched pairs. They are very well matched in terms of the propensity score. Figure 3 is the treatment effect estimated at a pair level. It

highlights that people with similar propensity score can have very different treatment effects. Using age as example, Figure 4 shows that people with a very similar propensity score value can have very different covariates. As discussed in Section 4, the difference of covariates at pair level does not matter as long as the averages of covariates are similar at cell level. We stratify the matched pairs into 18 cells by the propensity score of the treated observation in each pair. The width of each cell is 0.05 (since there is no treated observation with propensity score value larger than 0.9, there are 18 cells). Figure 5 shows the treatment effects for each cell estimated from the NSW experiment and propensity score matching. There is less volatility than at the pair level, but they are still very noisy. Contrary to the common intuition that the people have higher propensity score values also have larger treatment effects, it seems that the treatment effect is independent of the propensity score. The independence of the treatment effect and the propensity score partially explains why matching methods are successful in the Dehejia and Wahba (1999). Table 2 shows that the means of the covariates of the treated sample and the comparison sample in each cell. The balance of the covariates in each cell is a necessary condition for the propensity score matching methods to work. Table 2 shows some cells and some covariates in certain cells are indeed balanced, but the majority of them are not. Dehejia and Wahba (1999) devised a test to make the covariates balanced in each cell before carrying out matching, which is different from the issue we discuss here. Even if the covariates are balanced before matching, they could become unbalanced after matching, but the unbalance will become unlikely as the sample size increases.

The estimates from the LaLonde data set are shown in Table 3. Contrary to the estimates from the Dehejia and Wahba data set, all methods fail to replicate the

benchmark from the NSW experiment and most estimates even do not have right sign. The estimation from OLS is as good as the estimations from propensity score matching and the majority of the estimations from covariate matching, if not better.

The only difference of the LaLonde data set from the Dehejia and Wahba data set is that the Dehejia and Wahba data set has information on the earnings in 1974 (two years prior the treatment) but the LaLonde data set does not. It is hard to imagine that this difference is responsible for the failure of the matching estimators in the LaLonde data set, though the importance of pre-program earning history in the program evaluation is well known since the discovery of the famous Ashenfelter dip in Ashenfelter (1978). In order to explore this issue further, we pretend that we do not have the 1974 earning information in Dehejia and Wahba data set and estimate the treatment effects without using the earning variable of 1974. The results are reported in Table 4. Compared Table 1 with Table 4, it can be seen that the contribution of the 1974 earning history in improving the estimation of the treatment effects is marginal at most. With or without the 1974 earning history, the estimates from Dehejia and Wahba data set are closed to the NSW experimental benchmark.

VI. Conclusions

Selection bias due only to observables is a strong assumption, but for voluntary programs if we have the data on the application indicator and on the variables used by the program administrator to make the selection decision, we can justify this assumption.

With a proper data set and if the selection on observables assumption is justifiable, matching methods are useful tools to estimate treatment effects. There is no clear winner among matching estimators considered here. Propensity score matching

methods rely on the balancing property and require a large number of observations at each propensity score value, which is not required by covariate matching methods, while covariate matching methods face the curse of dimensionality and often encounter small cell or empty cell problem, so they do not dominate each other in term of data requirement in finite samples.

A major data requirement for propensity score matching is that at each propensity score value the number of observations is large. If this condition fails, it could affect the final results. It is important to check whether the covariates are balanced *after* matching at each propensity score value.

The failure of matching methods in the LaLonde data set highlights that, like any non-experimental estimator, the behavior of matching estimators largely depends on the data structure at hand. Matching is a useful estimator under suitable conditions, but it is definitely not *the* estimator for every evaluation. There is no easy way out in social program evaluation. A successful evaluation study requires detailed knowledge of the program, a good data set, and a careful consideration and choice of the estimation strategy.

Reference

Abadie, Albert and Guido W. Imbens (2002), "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," unpublished manuscript, Department of Economics, UC Berkeley (August 2002).

Angrist, Joshua. D. (1998), "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," Econometrica, Vol. 66 (March 1998), 249-288.

Angrist, Joshua D. and Jinyong Hahn (1999), "When to Control For Covariates? Panel-Asymptotic Results for Estimates of Treatment Effects," NBER Technical Working Paper 241 (1999)

Ashenfelter, Orley (1978), "Estimating the Effect of Training Programs on Earnings," The Review of Economic and Statistics 60 (February 1978), 47-57

Barnow, Burt S., Glen G. Cain, and Arthur S. Goldberger (1980), "Issues in the Analysis of Selection Bias," Evaluation Studies Review Annual 5 (1980), edited by Stromsdorfer, E. and Farkas, G.

Bjorklund, Anders, and Robert Moffitt (1987), "The Estimation of Wage Gains and Welfare Gains in Self-Selection Models," The Review of Economics and Statistics 69 (February 1987), 42-49.

Dawid, A. Philip (1980), "Conditional Independence for Statistical Operations", Annals of Statistics 8 (May 1980), 598-617.

_____ (2000), "Causal Inference without Counterfactuals", Journal of the American Statistical Association 95 (June 2000), 407-424.

Dehejia, Rajeev H. and Sadek Wahba (1999), "Causal Effects in Nonexperimental

Studies: Reevaluating the Evaluation of Training Programs," Journal of the American Statistical Association 94 (December 1999), 1053-1062.

_____ (2002), "Propensity Score Matching Methods for Non-Experimental Causal Studies," Review of Economics and Statistics 84, (February 2002), 151-175.

Fraker, Thomas and Rebecca Maynard (1987), "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs," The Journal of Human Resources 22 (Spring 1987), 194-227.

Hahn, Jinyong (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," Econometrica 66 (March 1998), 315-331.

Heckman, James J. (1974), "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," Annals of Economic and Social Measurement 5 (Fall 1976), 475-492.

_____ (1979), "Sample Selection Bias as a Specification Error," Econometrica 47 (January 1979), 153-162.

_____ (1990) , "Varieties of selection bias," American Economic Review 80 (May 1990), 313-318.

_____ (2000), "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective," Quarterly Journal of Economics 115 (February 2000), 45-97.

_____ (2001), "Microdata, Heterogeneity and the Evaluation of Public Policy: Nobel Lecture," Journal of Political Economy 109 (August 2001), 673-748.

Heckman, James J. and Joseph V. 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 (September 1989), 862-874.

Heckman, James J., Hidehiko Ichimura, Jeffrey A. Smith, and Petra E. Todd (1998), "Characterizing Selection Bias Using Experimental Dada," Econometrica 66 (September 1998), 1017-1098.

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

_____ (1998), "Matching as an Econometric Evaluation Estimator," Review of Economic Studies 65 (April 1998), 261-294.

Heckman, James J. and Edward Vytlacil (1999), "Local Instrumental Variables and Latent Variables Models for Identifying and Bounding Treatment Effects," Proceedings of the National Academy of Sciences of USA, 96 (February 1999), 4730-4734.

Holland, Paul W. (1986), "Statistics and Causal Inference," Journal of the American Statistical Association 81 (December 1986), 945-970.

Imbens, Guido W. (2000), "The Role of Propensity Score in Estimating Dose-Response Functions," Biometrika 87 (September 2000), 706-710.

Imbens, Guido W. (2003), "Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review," unpublished manuscript, Department of Economics, UC Berkeley (May 2003).

LaLonde, Robert J. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," The American Economic Review 76 (September 1986), 604-620.

Lechner, Michael (2002), "Program Heterogeneity and Propensity Score Matching: An Application to the Evaluation of Active Labor Market Policies," Review of Economics and Statistics 84 (May 2002), 205-220.

Moffitt, Robert A. (1991), "Program Evaluation With Nonexperimental Data," Evaluation Review 15 (June 1991), 291-315.

Neyman, Jerzy S. (1923), "On the Application of Probability Theory to Agriculture Experiments. Essay on Principles. Section 9.," Statistical Science 5 (1990), 465-485 (Translated from the Polish origin in Roczniki Nauk Rolniczych Tom X, 1923, 1-51).

Quandt, Richard E. (1972), "A New Approach to Estimating Switching Regressions," Journal of American Statistical Association 67 (June 1972), 306-310.

Rosenbaum, Paul R. (1995), "Observational Studies," New York: Springer-Verlag, (1995).

Rosenbaum, Paul R. and Donald B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," Biometrika 70 (April 1983), 41-55.

Roy, Andrew D. (1951), "Some Thoughts on The Distribution of Earnings," Oxford Economics Paper 3 (1951), 135-146.

Rubin, Donald B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," Journal of Educational Psychology 66 (1974), 688-

701.

Smith, Jeffrey A. and Petra E. Todd (2001), "Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods," The American Economic Review 91 (May 2001), 112-118.

Smith, Jeffrey A. and Petra E. Todd (2003), "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" Journal of Econometrics, forthcoming.

Westat (1981), "Continuous Longitudinal Manpower Survey Net Impact Report No. 1: Impact on 1977 Earnings of New FY 1976 CETA Enrollees in Selected Program Activities," Report prepared for US Department of Labor under Contract No. 23-23-74 (1981).

Zhao, Zhong (2003), "Using Matching to Estimate Treatment Effect: Data Requirement, Matching Metrics and Monte Carlo Evidence," unpublished manuscript, China Center for Economic Research, Peking University (June 2003).

Table 1 Estimates from Various Matching Estimators Using Dehejia and Wahba Data

Panel A: Matching with Replacement						
	With Common Support Condition			Without Common Support Condition		
Methods	Treatment Effect	Bias	Boothtrap Std. Error	Treatment Effect	Bias	Boothtrap Std. Error
NSW Experiment (Benchmark)	1794.3424	0	619.1345	1794.3424	0	577.865
Simple Mean Difference	-1855.9763	-3650.3187	562.7947	-8497.5161	-10291.8585	523.7342
OLS	1418.4384	-375.904	669.5362	1213.4147	-580.9277	519.514
Propensity Score Matching (Probit)	1677.5946	-116.7478	1017.8633	1702.3805	-91.9619	1005.5016
Propensity Score Matching (Logit)	1352.5425	-441.7999	1050.6117	1223.0937	-571.2487	905.453
Propensity Score Matching (LPM)	1600.0339	-194.3085	1414.7686	1600.0339	-194.3085	864.7236
Propensity Score Matching (Weighted LPM)	1084.8266	-709.5158	1106.3448	1084.8266	-709.5158	1055.1431
Covariate Matching (Mahalanobis)	1991.0404	196.698	1044.4974	2088.6502	294.3078	800.7334
Covariate & Propensity Score (Probit)	1941.6922	147.3498	946.8705	1761.7929	-32.5495	823.4981
Covariate & Propensity Score (Logit)	2108.5395	314.1971	966.3276	1928.6402	134.2978	838.8226
Covariate & Propensity Score (LPM)	1524.5499	-269.7925	977.4047	1482.5745	-311.7679	839.7936
Covariate & Propensity Score (Weighted LPM)	1434.5363	-359.8061	894.3857	1430.3668	-363.9756	920.8582
Covariate & Outcome	1927.8255	133.4831	1102.4222	2062.5942	268.2518	718.3613
Panel B: Matching without Replacement						
	With Common Support Condition			Without Common Support Condition		
Methods	Treatment Effect	Bias	Boothtrap Std. Error	Treatment Effect	Bias	Boothtrap Std. Error
Propensity Score Matching (Probit)	1341.4319	-452.9105	802.7994	1366.2177	-428.1247	717.4967
Propensity Score Matching (Logit)	1560.2072	-234.1352	814.4174	1430.7584	-363.584	711.5508
Propensity Score Matching (LPM)	1312.7879	-481.5545	918.0792	1312.7879	-481.5545	705.9952
Propensity Score Matching (Weighted LPM)	792.5403	-1001.8021	717.9579	792.5403	-1001.8021	752.4914
Covariate Matching (Mahalanobis)	986.3221	-808.0203	709.0015	854.1095	-940.2329	640.6439
Covariate & Propensity Score (Probit)	1351.1288	-443.2136	843.9903	1202.2937	-592.0487	742.9999
Covariate & Propensity Score (Logit)	1365.7223	-428.6201	836.9152	1111.705	-682.6374	740.1312
Covariate & Propensity Score (LPM)	625.4851	-1168.8573	761.8081	583.5097	-1210.8327	703.7281
Covariate & Propensity Score (Weighted LPM)	1085.7131	-708.6293	741.2391	1081.5435	-712.7989	744.7629
Covariate & Outcome	939.1385	-855.2039	718.8523	1073.9072	-720.4352	548.8117

Note: 1. The specification of the propensity score is the same as in Dehejia and Wahba (1999), including age, age squared, education, education squared, no degree, married, black, hispanic, re74, re75, u74, u75, education*re74 and age cubed.

2. The specification of the OLS is the same as the specification of the propensity score.

3. The outcome equation in the covariate & outcome matching is estimated using only the treated data by OLS and without any higher order and interaction term.

Table 2 The Covariate Means in Each Cell after Propensity Score Matching Using Dehejia and Wahba Data

Propensity Score	Cell	Age		Education		Married		Nodgree		Earning74		Earning75	
Cell	Size	Treated	Control	Treated	Control	Treated	Control	Treated	Control	Treated	Control	Treated	Control
0.00-0.05	24	26.42	27.17	11.04	10.79	0.50	0.38	0.46	0.42	5144.77	5716.24	4969.86	3858.59
0.05-0.10	8	27.75	28.25	11.38	10.38	0.25	0.25	0.50	0.50	5173.82	6174.38	2621.67	2340.40
0.10-0.15	14	26.43	27.57	10.14	11.00	0.21	0.43	0.71	0.50	2913.09	5086.12	1773.58	5200.50
0.15-0.20	11	30.09	29.55	8.91	9.55	0.36	0.27	0.73	0.91	1394.18	1696.37	1058.00	1923.78
0.20-0.25	8	25.50	21.00	11.38	9.63	0.00	0.25	0.38	0.88	7055.38	878.98	2632.63	669.36
0.25-0.30	5	23.60	22.40	10.60	10.20	0.20	0.20	0.60	0.80	6346.94	0.00	3288.77	852.19
0.30-0.35	11	26.09	23.36	11.36	11.18	0.09	0.27	0.45	0.55	4653.54	248.29	1822.78	587.55
0.35-0.40	16	21.81	20.25	9.69	9.19	0.06	0.00	0.88	0.88	669.81	407.28	592.81	172.54
0.40-0.45	9	20.44	22.56	10.00	12.33	0.00	0.00	0.78	0.56	1489.46	3.92	612.17	0.00
0.45-0.50	9	22.00	25.44	9.67	10.56	0.11	0.22	0.89	0.78	356.45	697.06	1292.01	2184.99
0.50-0.55	9	26.00	25.11	10.44	11.33	0.22	0.33	0.78	0.44	0.00	0.00	301.62	247.26
0.55-0.60	14	27.29	28.43	11.21	9.57	0.36	0.29	0.71	0.86	0.00	1599.74	619.06	932.89
0.60-0.65	10	27.00	30.20	11.40	11.80	0.20	0.20	0.40	0.20	0.00	0.00	93.44	0.00
0.65-0.70	1	27.00	29.00	10.00	12.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
0.70-0.75	3	26.67	35.00	9.67	11.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00
0.75-0.80	5	27.60	22.40	8.80	10.20	0.00	0.00	1.00	1.00	0.00	0.00	879.79	0.00
0.80-0.85	19	26.32	28.53	10.05	9.95	0.00	0.00	1.00	1.00	0.00	0.00	306.27	279.76
0.85-0.90	9	27.56	26.33	8.67	9.78	0.00	0.00	1.00	1.00	0.00	0.00	0.00	21.09

Table 3 Estimates from Various Matching Estimators Using Lalonde Data

Panel A: Matching with Replacement						
	With Common Support Condition			Without Common Support Condition		
Methods	Treatment Effect	Bias	Boothtrap Std. Error	Treatment Effect	Bias	Boothtrap Std. Error
NSW Expriment (Benchmark)	886.3037	0	552.1865	886.3037	0	520.9488
Simple Mean Difference	-5091.4435	-5977.7472	864.9801	-8870.3076	-9756.6113	428.1363
OLS	-653.4911	-1539.7948	546.1872	-900.2483	-1786.552	474.0258
Propensity Score Matching (Probit)	-830.6882	-1716.9919	884.9409	-894.8972	-1781.2009	794.2928
Propensity Score Matching (Logit)	-834.1216	-1720.4253	835.5473	-834.1216	-1720.4253	784.0141
Propensity Score Matching (LPM)	-988.9594	-1875.2631	927.6027	-988.9594	-1875.2631	942.5992
Propensity Score Matching (Weighted LPM)	-1172.6353	-2058.939	835.4494	-1193.1245	-2079.4282	850.1418
Covariate Matching (Mahalanobis)	38.0357	-848.268	729.5191	60.6543	-825.6494	674.9392
Covariate & Propensity Score (Probit)	-906.7191	-1793.0228	745.0735	-867.5519	-1753.8556	819.766
Covariate & Propensity Score (Logit)	-781.2119	-1667.5156	723.089	-742.0447	-1628.3484	834.9877
Covariate & Propensity Score (LPM)	-1663.6733	-2549.977	786.3256	-1663.6733	-2549.977	785.6147
Covariate & Propensity Score (Weighted LPM)	-1371.5015	-2257.8052	744.3925	-1371.5015	-2257.8052	792.4616
Covariate & Outcome	211.2083	-675.0954	829.2703	250.3755	-635.9282	777.0201
Panel B: Matching without Replacement						
	With Common Support Condition			Without Common Support Condition		
Methods	Treatment Effect	Bias	Boothtrap Std. Error	Treatment Effect	Bias	Boothtrap Std. Error
Propensity Score Matching (Probit)	-1119.5763	-2005.88	748.332	-1176.9589	-2063.2626	607.0538
Propensity Score Matching (Logit)	-763.8225	-1650.1262	651.467	-763.8225	-1650.1262	669.4701
Propensity Score Matching (LPM)	-1537.579	-2423.8827	657.2353	-1537.579	-2423.8827	709.2184
Propensity Score Matching (Weighted LPM)	-1145.6929	-2031.9966	633.6048	-1166.1821	-2052.4858	617.681
Covariate Matching (Mahalanobis)	-761.2005	-1647.5042	621.2306	-626.2149	-1512.5186	602.2818
Covariate & Propensity Score (Probit)	-1824.3924	-2710.6961	672.6644	-1785.2252	-2671.5289	656.9399
Covariate & Propensity Score (Logit)	-1800.1467	-2686.4504	676.5479	-1760.9795	-2647.2832	675.2881
Covariate & Propensity Score (LPM)	-2429.0461	-3315.3498	704.0201	-2429.0461	-3315.3498	700.9407
Covariate & Propensity Score (Weighted LPM)	-2224.2692	-3110.5729	668.49	-2224.2692	-3110.5729	649.8167
Covariate & Outcome	-1073.6276	-1959.9313	817.4817	-1034.4604	-1920.7641	712.3288

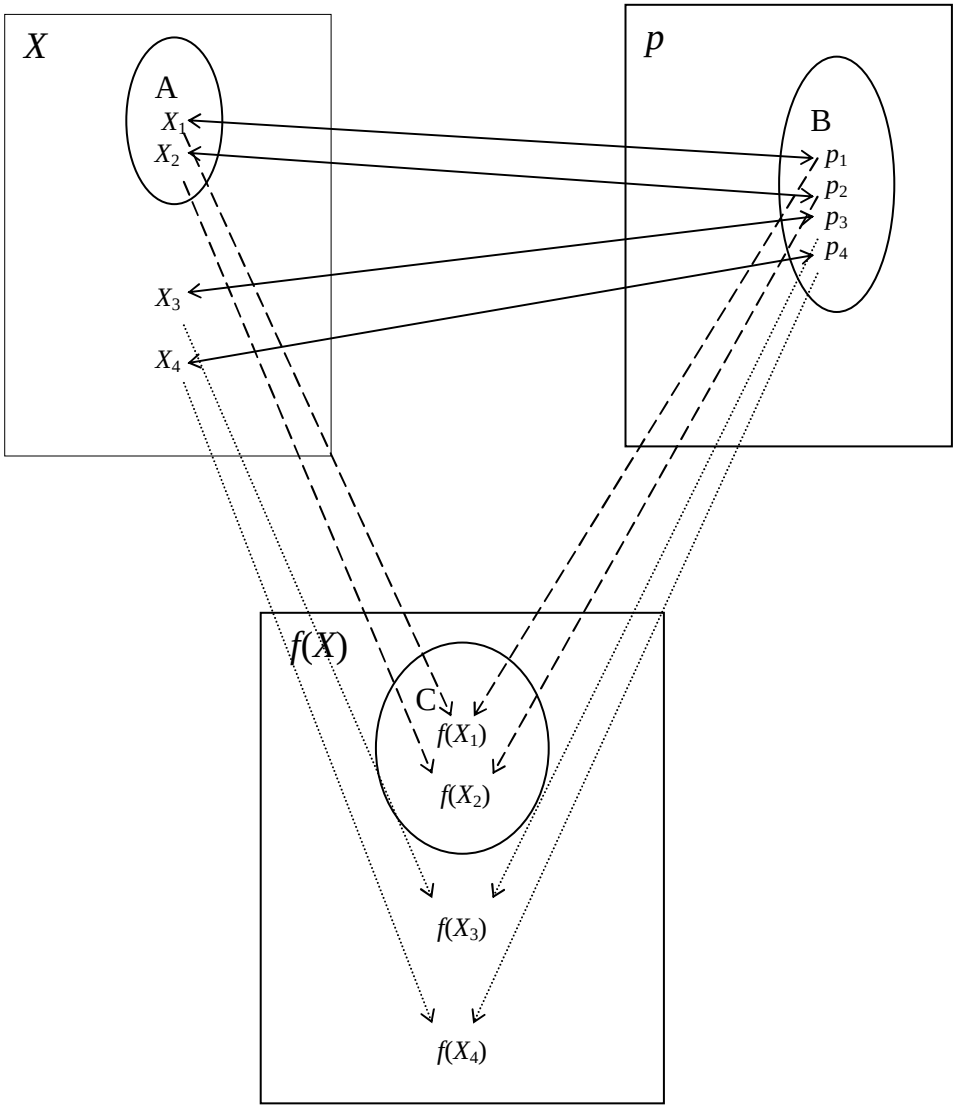
Note: 1. The specification of the propensity score is the same as in Dehejia and Wahba (1999), including age, age squared, education, education squared, no degree, married, black, hispanic, re75, u75 and age cubed.
2. The specification of the OLS is the same as the specification of the propensity score.
3. The outcome equation in the covariate & outcome matching is estimated using only the treated data by OLS and without any higher order and interaction term.

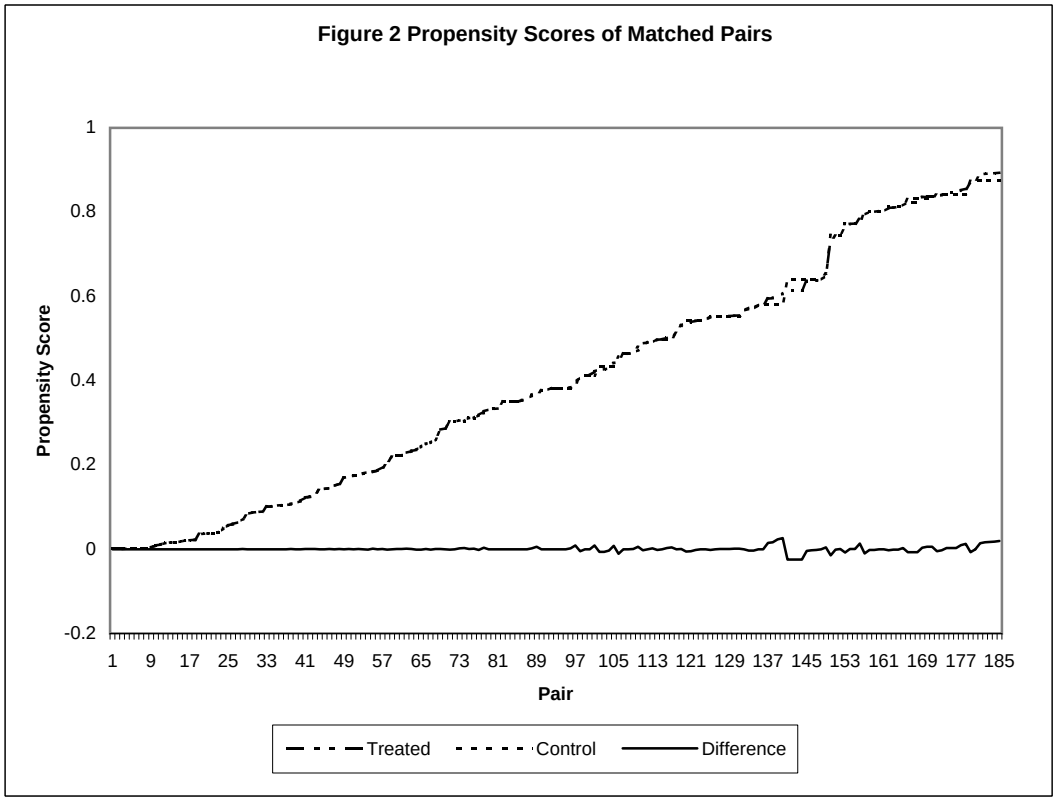
Table 4 Estimates from Various Matching Estimators Using Dehejia and Wahba Data without 1974 Earning History

Panel A: Matching with Replacement						
	With Common Support Condition			Without Common Support Condition		
Methods	Treatment Effect	Bias	Boothtrap Std. Error	Treatment Effect	Bias	Boothtrap Std. Error
NSW Experiment (Benchmark)	1794.3424	0	627.0964	1794.3424	0	697.8393
Simple Mean Difference	-1542.2862	-3336.6286	673.7822	-8497.5161	-10291.8585	623.0656
OLS	947.8409	-846.5015	647.9457	703.5223	-1090.8201	673.6418
Propensity Score Matching (Probit)	1742.4868	-51.8556	929.9885	1742.4868	-51.8556	984.5567
Propensity Score Matching (Logit)	1062.9881	-731.3543	926.9224	1062.9881	-731.3543	1036.6405
Propensity Score Matching (LPM)	-18.6278	-1812.9702	936.6556	-18.6278	-1812.9702	1126.1358
Propensity Score Matching (Weighted LPM)	889.7427	-904.5997	867.7195	1029.7993	-764.5431	952.8389
Covariate Matching (Mahalanobis)	1379.3894	-414.953	969.6251	1803.7764	9.434	944.4436
Covariate & Propensity Score (Probit)	1228.9009	-565.4415	944.9318	1228.9009	-565.4415	1039.7549
Covariate & Propensity Score (Logit)	1837.8201	43.4777	955.4343	1837.8201	43.4777	988.6396
Covariate & Propensity Score (LPM)	1353.0539	-441.2885	955.5663	1408.241	-386.1014	934.5179
Covariate & Propensity Score (Weighted LPM)	1473.6523	-320.6901	977.3127	1548.6969	-245.6455	950.8789
Covariate & Outcome	1848.228	53.8856	1042.6264	1870.2421	75.8997	1044.1177
Panel B: Matching without Replacement						
	With Common Support Condition			Without Common Support Condition		
Methods	Treatment Effect	Bias	Boothtrap Std. Error	Treatment Effect	Bias	Boothtrap Std. Error
Propensity Score Matching (Probit)	1875.3385	80.9961	810.1426	1875.3385	80.9961	819.9448
Propensity Score Matching (Logit)	1296.0695	-498.2729	755.6609	1261.2032	-533.1392	788.5103
Propensity Score Matching (LPM)	-785.504	-2579.8464	726.63	-785.504	-2579.8464	880.0777
Propensity Score Matching (Weighted LPM)	848.9815	-945.3609	753.7384	928.9465	-865.3959	818.2015
Covariate Matching (Mahalanobis)	892.9058	-901.4366	773.56	1060.1125	-734.2299	796.1598
Covariate & Propensity Score (Probit)	1523.9989	-270.3435	789.1705	1517.4889	-276.8535	793.7948
Covariate & Propensity Score (Logit)	1576.4222	-217.9202	781.3752	1534.942	-259.4004	798.2094
Covariate & Propensity Score (LPM)	1251.4742	-542.8682	723.2978	1283.3531	-510.9893	805.018
Covariate & Propensity Score (Weighted LPM)	1397.4815	-396.8609	719.3703	1400.8284	-393.514	809.137
Covariate & Outcome	883.0169	-911.3255	964.801	922.8197	-871.5227	1007.0669

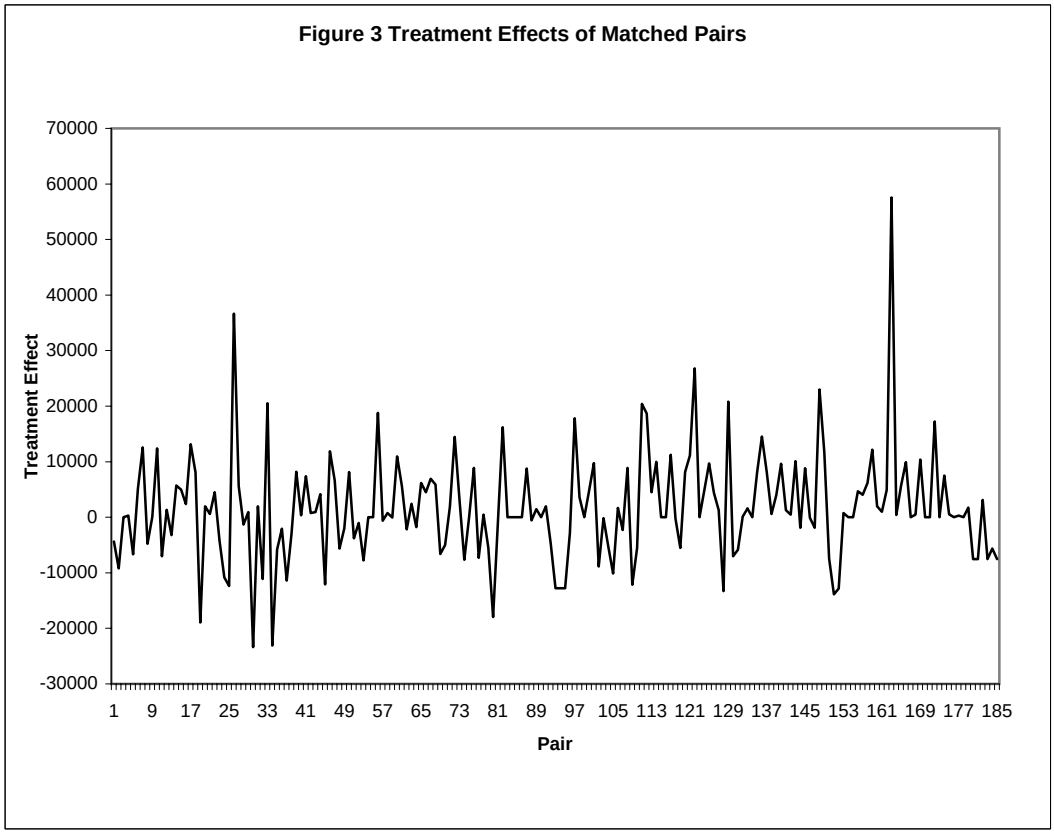
Note: 1. The specification of the propensity score is linear without any higher order and interaction term.
2. The specification of the OLS is the same as the specification of the propensity score.
3. The outcome equation in the covariate & outcome matching is estimated using only the treated data by OLS and without any higher order and interaction term.

Figure 1 Neighborhood Matching
With One to One Correspondence between *X* and *p*

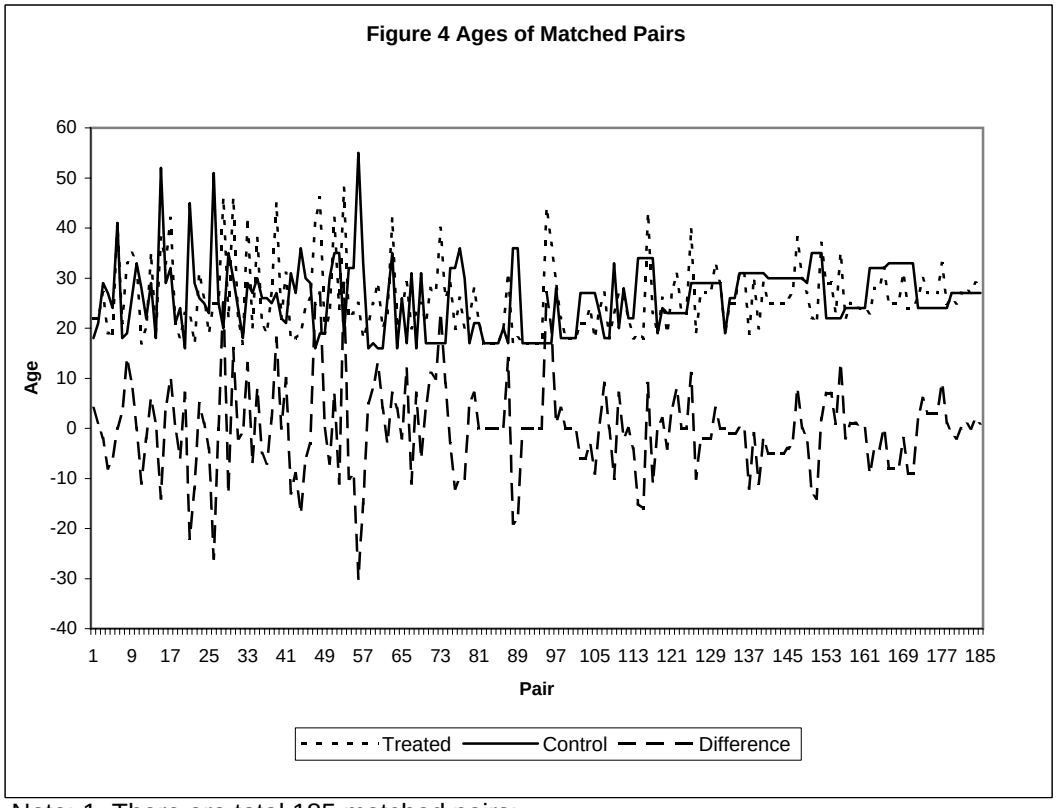




Note: 1. There are total 185 matched pairs;
2. The matched pairs are sorted by the propensity score of the treated;

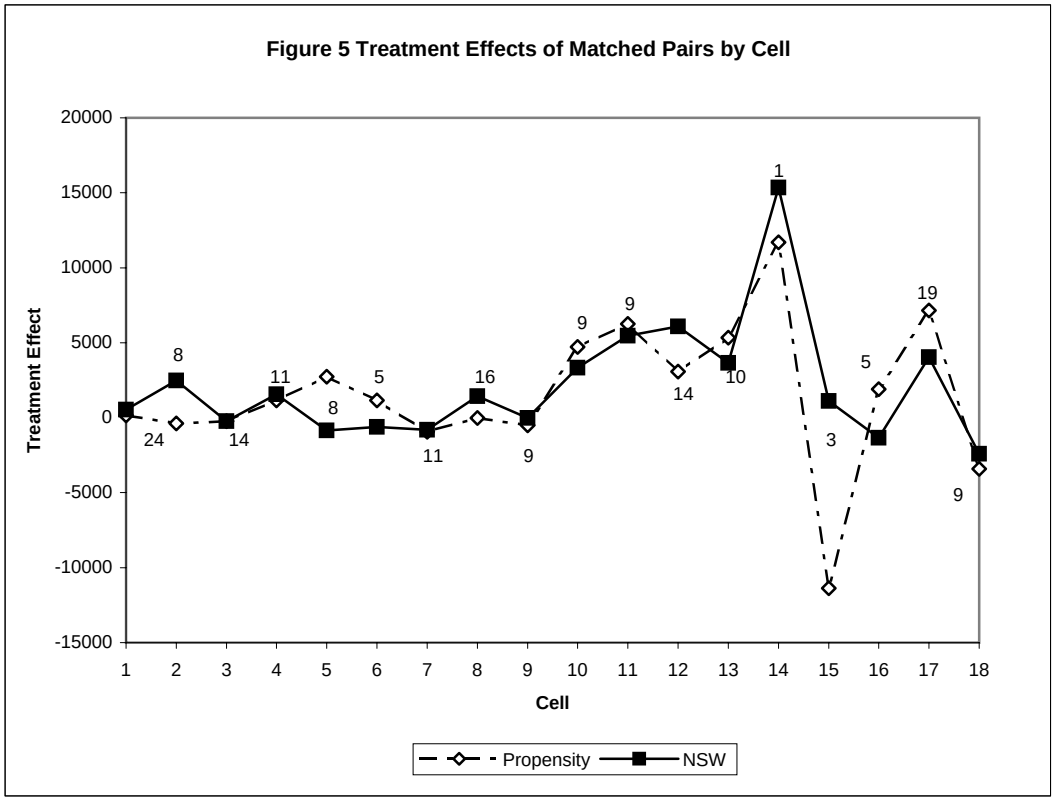


Note: 1. There are total 185 matched pairs;
2. The matched pairs are sorted by the propensity score of the treated;



Note: 1. There are total 185 matched pairs;

2. The matched pairs are sorted by the propensity score of the treated;



Note: 1. There are total 185 matched pairs;

2. The matched pairs are sorted by the propensity score of the treated;
3. The width of the cell in Figure 5 is 0.05. Since there is no observation with propensity score larger than 0.9, there are only 18 cells. The numbers in the plot are the numbers of matched pairs in each cell.