Title

Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies

Authors

Diamond, Alexis
Sekhon, Jasjeet S.

Publication Date

2006-09-18

Genetic Matching for Estimating Causal Effects:

A General Multivariate Matching Method for Achieving Balance in Observational Studies*

Alexis Diamond and Jasjeet S. Sekhon†

Corresponding Author

Ph. D. Candidate
Political Economy and Government
Harvard University

Associate Professor
Department of Political Science
UC Berkeley

Version: 1.2 (23:50)

*Matching software developed by Sekhon which implements the technology outlined in this paper can be downloading from `http://sekhon.berkeley.edu/matching/`. Winner of the 2005 Gosnell Prize for Excellence in Political Methodology. We thank Alberto Abadie, Walter Mebane, Jr., Donald Rubin and Jonathan Wand for many helpful discussions. We also thank Henry Brady, Rajeev Dehejia, Joseph Hotz, Kosuke Imai, Guido Imbens, Gary King, Kevin Quinn, Jamie Robins, Phil Schrodt, Jeffrey Smith and Petra Todd for valuable comments. Previous versions have been presented at the Society for Political Methodology Meeting, FSU, July 21–23, 2005; Empirical Evaluation of Labour Market Programmes, Congress Centre of Bundesagentur für Arbeit, Nuremberg, Germany, June 16–17, 2005; Northeast Methods Conference, NYU, April 22, 2005; and the MPSA Annual Meeting, Chicago, Illinois, April 7–10, 2005. All errors are our responsibility.

†`sekhon@berkeley.edu`, `http://sekhon.berkeley.edu/`, Survey Research Center, 2538 Channing Way, UC Berkeley, Berkeley, CA, 94720.

Abstract

Genetic matching is a new method for performing multivariate matching which uses an evolutionary search algorithm to determine the weight each covariate is given. The method utilizes an evolutionary algorithm developed by Mebane and Sekhon (1998; Sekhon and Mebane 1998) that maximizes the balance of observed potential confounders across matched treated and control units. The method is nonparametric and does not depend on knowing or estimating the propensity score, but the method is greatly improved when a known or estimated propensity score is incorporated. Genetic matching reliably reduces both the bias and the mean square error of the estimated causal effect even when the property of equal percent bias reduction (EPBR) does not hold. When this property does not hold, matching methods—such as Mahalanobis distance and propensity score matching—often perform poorly. Even if the EPBR property does hold and the propensity score is correctly specified, in finite samples, estimates based on genetic matching have lower mean square error than those based on the usual matching methods. We present a reanalysis of the LaLonde (1986) job training dataset which demonstrates the benefits of genetic matching and which helps to resolve a longstanding debate between Dehejia and Wahba (1997; 1999; 2002; Dehejia 2005) and Smith and Todd (2001, 2005a,b) over the ability of matching to overcome LaLonde's critique of nonexperimental estimators. Monte Carlos are also presented to demonstrate the properties of our method.

1 Introduction

Matching has become an increasingly popular method of causal inference in many fields including statistics (e.g., Rosenbaum 2002), medicine (e.g., Christakis and Iwashyna 2003; Rubin 1997), economics (e.g., Abadie and Imbens forthcoming; Dehejia and Wahba 1999, 2002; Galiani, Gertler, and Schargrodsky 2005), political science (e.g., Imai 2005; Sekhon 2004), sociology (e.g., Diprete and Engelhardt 2004; Smith 1997; Winship and Morgan 1999) and even law (e.g., Epstein, Ho, King, and Segal 2005; Rubin 2001). There is, however, no consensus on how exactly matching ought to be done, how to measure the success of the matching procedure, and whether or not matching estimators are sufficiently robust to misspecification so as to be useful in practice (Heckman, Ichimura, Smith, and Todd 1998). These issues have been central to an ongoing debate over how matching and other nonexperimental estimators perform when analyzing data from a nationwide job training experiment (LaLonde 1986). The experimental results are used to establish benchmark estimates for causal effects. Then, to create the kind of observational data typically analyzed by social scientists, individuals from the experimental control group are replaced by individuals from national observational surveys. The goal is to determine which methods, if any, are able to use the observational data to recover results obtained from the randomized experiment. We show that the debate between Dehejia and Wahba (1997; 1999; 2002; Dehejia 2005) and Smith and Todd (2001, 2005a,b) over the ability of matching to overcome LaLonde's critique of nonexperimental estimators is largely driven by the fact that the existing matching methods fail to obtain reliable levels of balance in this dataset. We show that our proposed matching method is able to reliably estimate the causal effects when other methods fail because it achieves substantially better balance.

When using matching methods to estimate causal effects, a central problem is deciding how best to perform the matching. Two common approaches are propensity score matching (Rosenbaum and Rubin 1983) and multivariate matching based on Mahalanobis distance (Cochran and Rubin 1973; Rubin 1979, 1980). Matching methods based on the propensity

score (estimated by logistic regression), Mahalanobis distance or a combination of the two have appealing theoretical properties if covariates have ellipsoidal distributions—e.g., distributions such as the normal or t. If the covariates are so distributed, these methods (more generally affinely invariant matching methods[1]) have the property of "equal percent bias reduction" (EPBR) (Rubin 1976a,b; Rubin and Thomas 1992a).[2] This property, which is formally defined below, ensures that matching methods will reduce bias in all linear combinations of the covariates. If a matching method is not EPBR, then that method will, in general, increase the bias for some linear function of the covariates even if all univariate means are closer in the matched data than the unmatched (Rubin 1976a).

Our proposed method, genetic matching (GenMatch), is an affinely invariant method for performing multivariate matching that uses an evolutionary search algorithm to determine the weight given to each baseline covariate. The method utilizes an evolutionary algorithm (EA) developed by Mebane and Sekhon (1998; Sekhon and Mebane 1998) that maximizes the balance of observed baseline covariates across matched treated and control units.

GenMatch is shown to have better properties than the usual alternative matching methods both when the EPBR property holds and when it does not. Even when the EPBR property holds and the mapping from X to Y is linear, GenMatch has better efficiency—i.e., lower mean square error (MSE)—in finite samples. When the EPBR property does not hold as it generally does not, GenMatch retains appealing properties and the differences in performance between GenMatch and the other matching methods can become substantial both in terms of bias and MSE reduction. In short, at the expense of computer time, GenMatch dominates the other matching methods in terms of MSE when assumptions required for EPBR hold and, even more so, when they do not.

GenMatch is able to retain good properties even when EPBR does not hold because a

[1] Affine invariance means that the matching output is invariant to matching on X or an affine transformation of X.

[2] The EPBR results of Rubin and Thomas (1992a) have been extended by Rubin and Stuart (2005) to the case of discriminant mixtures of proportional ellipsoidally symmetric (DMPES) distributions. This extension is important, but it is restricted to a limited set of mixtures. See Section 3.1.

set of constraints is imposed by the loss function optimized by the EA. The loss function depends on a large number of functions of covariate imbalance across matched treatment and control groups. The precise measures of covariate imbalance can be selected by the analyst depending on the application, but given these measures, GenMatch will optimize covariate balance. In the examples in this paper, we apply the algorithm to greedy matching (with replacement). But the algorithm can also be used with optimal full matching (Rosenbaum 1989, 1991).

Both propensity score matching and matching based on Mahalanobis distance can be considered special limiting cases of our method. If the propensity score is known or estimated, genetic matching will use it along with information about baseline covariates uncorrelated with the propensity score; if the propensity score contains all of the relevant information in a given sample, the other variables will be given zero weight.[3] GenMatch will converge to Mahalanobis distance if that proves to be the appropriate distance measure.

The paper is organized as follows. Section 2 outlines the Rubin causal model; Section 3 describes Mahalanobis and propensity score matching; Section 4 describes GenMatch; Section 5 presents our Monte Carlo results which help demonstrate the properties of our estimator; Section 6 presents our reanalysis of the job training data; Section 7 concludes.

2 Rubin Causal Model

The Rubin causal model conceptualizes causal inference in terms of potential outcomes under treatment and control, only one of which is observed for each unit (Holland 1986; Splawa-Neyman 1990 [1923]; Rubin 1974, 1978, 1990). A causal effect is defined as the difference between an observed outcome and its counterfactual.

Let Y_{i1} denote the potential outcome for unit i if the unit receives treatment, and let Y_{i0} denote the potential outcome for unit i in the control regime. The treatment effect for

[3]Technically, the other variables will be given weights just large enough to ensure that the weight matrix is positive definite.

observation i is defined by $\tau_i = Y_{i1} - Y_{i0}$. Causal inference is essentially a missing data problem because Y_{i1} and Y_{i0} are never both observed. Let T_i be a treatment indicator: 1 when i is in the treatment regime and 0 otherwise. The observed outcome for observation i is then $Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$.

In principle, if assignment to treatment is randomized, causal inference is straightforward because the two groups are drawn from the same population by construction, and treatment assignment is independent of all baseline variables. As the sample size grows, observed and unobserved baseline variables are balanced across treatment and control groups with arbitrarily high probability. Because treatment assignment is independent of Y_0 and Y_1—following Dawid's (1979) notation, $\{Y_{i0}, Y_{i1} \perp\!\!\!\perp T_i\}$. Hence, for $j = 0, 1$

$$E(Y_{ij} \mid T_i = 1) = E(Y_{ij} \mid T_i = 0) = E(Y_i \mid T_i = j)$$

Therefore, the average treatment effect (ATE) can be estimated by:

$$\begin{aligned} \tau &= E(Y_{i1} \mid T_i = 1) - E(Y_{i0} \mid T_i = 0) \\ &= E(Y_i \mid T_i = 1) - E(Y_i \mid T_i = 0) \end{aligned} \tag{1}$$

Equation 1 is estimable in an experimental setting because observations in treatment and control groups are exchangeable.[4] In the simplest experimental setup, individuals in both groups are equally likely to receive the treatment, and hence assignment to treatment will not be associated with the outcome. Even in an experimental setup, much can go wrong which requires statistical correction (e.g., Barnard, Frangakis, Hill, and Rubin 2003; Imai 2005).

In an observational setting, unless something special is done, treatment and non-treatment

[4]It is standard practice to assume the Stable Unit Treatment Value assumption, also known as SUTVA (Holland 1986; Rubin 1978). SUTVA requires that the treatment status of any unit be independent of potential outcomes for all other units, and that treatment is defined identically for all units. Throughout the rest of the paper, we take SUTVA as given. If SUTVA does not hold, even in an experimental setting the estimation of the causal effect become difficult.

groups are almost never balanced because the two groups are not ordinarily drawn from the same population. Thus, a common quantity of interest is the average treatment effect for the treated (ATT):

$$\tau \mid (T = 1) \quad = \quad E(Y_{i1} \mid T_i = 1) - E(Y_{i0} \mid T_i = 1). \tag{2}$$

Equation 2 cannot be directly estimated because Y_{i0} is not observed for the treated. Progress can be made by assuming that selection for treatment depends on observable covariates X. Following Rosenbaum and Rubin (1983), one can assume that conditional on X, treatment assignment is unconfounded ($\{Y_0, Y_1 \perp\!\!\!\perp T\} \mid X$) and that there is overlap: $0 < Pr(T = 1 \mid X) < 1$. Together, unconfoundedness and overlap constitute a property known as strong ignorability of assignment which is necessary for identifying ATE. Heckman et al. (1998) shows that for ATT, the unconfoundedness assumption can be weakened to mean independence: $E\left(Y_{ij} \mid T_i, X_i\right) = E\left(Y_{ij} \mid X_i\right)$.[5] The overlap assumption for ATT only requires that the support of X for the treated be a subset of the support of X for control observations.

Then, following Rubin (1974, 1977) we obtain

$$E(Y_{ij} \mid X_i, T_i = 1) = E(Y_{ij} \mid X_i, T_i = 0) = E(Y_i \mid X_i, T_i = j). \tag{3}$$

By conditioning on observed covariates, X_i, treatment and control groups are balanced. The average treatment effect for the treated is estimated as

$$\tau \mid (T = 1) \quad = \quad E\left\{E(Y_i \mid X_i, T_i = 1) - E(Y_i \mid X_i, T_i = 0) \mid T_i = 1\right\}, \tag{4}$$

where the outer expectation is taken over the distribution of $X_i \mid (T_i = 1)$ which is the distribution of baseline variables in the treated group.

The most straightforward and nonparametric way to condition on X is to exactly match on the covariates. This is an old approach going back to at least Fechner (1966 [1860]), the

[5] Also see Abadie and Imbens (forthcoming).

father of psychophysics. This approach fails in finite samples if the dimensionality of X is large or if X contains continuous covariates. Thus, in general, alternative methods must be used.

3 Mahalanobis and Propensity Score Matching

The most common method of multivariate matching is based on Mahalanobis distance (Cochran and Rubin 1973; Rubin 1979, 1980). The Mahalanobis distance between any two column vectors is:

$$md(X_i, X_j) \quad = \{(X_i - X_j)' S^{-1} (X_i - X_j)\}^{\frac{1}{2}}$$

where S is the sample covariance matrix of X. To estimate ATT by matching with replacement, one matches each treated unit with the M closest control units, as defined by this distance measure, $md()$.[6] If X consists of more than one continuous variable, multivariate matching estimates contain a bias term which does not asymptotically go to zero at $\sqrt{n}$ (Abadie and Imbens forthcoming).

An alternative way to condition on X is to match on the probability of assignment to treatment, known as the propensity score.[7] As one's sample size grows large, matching on the propensity score produces balance on the vector of covariates X (Rosenbaum and Rubin 1983).

Let $e(X_i) \equiv Pr(T_i = 1 \mid X_i) = E(T_i \mid X_i)$, defining $e(X_i)$ to be the propensity score. Given $0 < Pr(T_i \mid X_i) < 1$ and that $Pr(T_1, T_2, \cdots T_N \mid X_1, X_2, \cdots X_N) = \Pi_{i=1}^{N} e(X_i)^{T_i} (1 -$

[6] Alternatively one can do optimal full matching (Rosenbaum 1989, 1991) instead of the greedy matching (with replacement) which we focus on in this paper. But this decision is a separate one from the choice of a distance metric.

[7] The first estimator of treatment effects to be based on a weighted function of the probability of treatment was the Horvitz-Thompson statistic (Horvitz and Thompson 1952).

$e(X_i))^{(1-T_i)}$, then as Rosenbaum and Rubin (1983) prove,

$$\tau \mid (T = 1) \;\; = \;\; E\left\{E(Y_i \mid e(X_i), T_i = 1) - E(Y_i \mid e(X_i), T_i = 0) \mid T_i = 1\right\},$$

where the outer expectation is taken over the distribution of $e(X_i) \mid (T_i = 1)$. Since the propensity score is generally unknown, it must be estimated.

Propensity score matching involves matching each treated unit to the nearest control unit on the unidimensional metric of the propensity score vector. If the propensity score is estimated by logistic regression, as is typically the case, much is to be gained by matching not on the predicted probabilities (bounded between zero and one) but on the linear predictor: $\hat{\mu} = X\hat{\beta}$. Matching on the linear predictor avoids compression of propensity scores near zero and one. Moreover, the linear predictor is often more nearly normally distributed which is of some importance given the EPBR results if the propensity score is matched along with other covariates.

Mahalanobis distance and propensity score matching can be combined in various ways (Rubin 2001; Rosenbaum and Rubin 1985). It is useful to combine the propensity score with Mahalanobis distance matching because propensity score matching is particularly good at minimizing the discrepancy along the propensity score and Mahalanobis distance is particularly good at minimizing the distance between individual coordinates of X (orthogonal to the propensity score) (Rosenbaum and Rubin 1985).

3.1 Equal Percent Bias Reduction (EPBR)

Affinely invariant matching methods, such as Mahalanobis metric matching and propensity score matching (if the propensity score is estimated by logistic regression), are equal percent bias reducing if all of the covariates used have ellipsoidal distributions (Rubin and Thomas 1992a)—e.g., distributions such as the normal or t—or if the covariates are mixtures

of proportional ellipsoidally symmetric (DMPES) distributions Rubin and Stuart (2005).[8]

To formally define EPBR, let Z be the expected value of X in the matched control group. Then, as outlined in Rubin (1976a), a matching procedure is EPBR if

$$E(X \mid T = 1) - Z \ = \gamma \{E(X \mid T = 1) - E(X \mid T = 0)\}$$

for a scalar $0 \leq \gamma \leq 1$. In other words, we say that a matching method is EPBR for X because the percent reduction in the biases of each of the matching variables is the same. One obtains the same percent reduction in bias for any linear function of X if and only if the matching method is EPBR for X. Moreover, if a matching method is not EPBR for X, the bias for some linear function of X is increased even if all univariate covariate means are closer in the matched data than the unmatched (Rubin 1976a).

A significant shortcoming of common matching methods such as Mahalanobis distance and propensity score matching is that they may (and in practice, frequently do) make balance worse across measured potential confounders. These methods may make balance worse, in practice, even if covariates *are* distributed ellipsoidally symmetric, because EPBR is a property that obtains in expectation, not necessarily for any particular set of data. Moreover, if covariates are neither ellipsoidally symmetric nor are mixtures of DMPES distributions, propensity score matching has good theoretical properties if and only if the true propensity score model is known with certainty and the sample size is large.

Even if the covariates have elliptic distributions, in finite samples they may not. Then Mahalanobis distance may not be optimal because the matrix used to scale the distances, the covariance matrix of X, can be improved upon.[9]

The EPBR property itself is limited and in a given substantive problem it may not be

[8]Note that DMPES defines a limited set of mixtures. In particular, countably infinite mixtures of ellipsoidal distributions where: (1) all inner products are proportional and (2) where the centers of each constituent ellipsoidal distribution are such that all best linear discriminants between any two components are also proportional.

[9]For justifications of Mahalanobis distance based on distributional considerations see Mitchell and Krzanowski (1985, 1989).

desirable. This can arise if it is known based on theory that one covariate has a large nonlinear relationship with the outcome while another does not—e.g., $Y = X_1^4 + X_2$, where $X > 1$. In such a case, reducing bias in X_1 will be more important than X_2.

4 Genetic Matching

The idea underlying the GenMatch algorithm is that if Mahalanobis distance is not optimal for achieving balance in a given dataset, one should be able to search over the space of distance metrics and find something better. One way of generalizing the Mahalanobis metric is including an additional weight matrix:

$$d(X_i, X_j) \quad = \left\{ (X_i - X_j)' \left(S^{-1/2}\right)' W S^{-1/2} (X_i - X_j) \right\}^{\frac{1}{2}}$$

where W is a $k \times k$ positive definite weight matrix and $S^{1/2}$ is the Cholesky decomposition of S which is the variance-covariance matrix of X.[10]

GenMatch is an affinely invariant matching algorithm that uses the distance measure $d()$, in which all elements of W are zero except down the main diagonal. The main diagonal consists of k parameters which must be chosen. Note that if each of these k parameters are set equal to 1, $d()$ is the same as Mahalanobis distance. Like Mahalanobis distance, this distance metric can be used to conduct either greedy or optimal full matching.

The choice of setting the nondiagnal elements of W to zero is made for reasons of computational power alone. The optimization problem grows exponentially with the number of free parameters so it is important that the problem be parameterized so as to limit the number of parameters which must be estimated. Moreover, as we shall see, our proposed generalization works well in this example and in a set of Monte Carlos, but there is no claim that it is generally the best parameterization.

The weight matrix W has an infinity of equivalent solutions because the matches produced

[10]The Cholesky decomposition is parameterized such that $S = LL'$, $S^{1/2} = L$. In other words, L is a lower triangular matrix with positive diagonal elements.

are invariant to a constant scale change to the distance measure. In particular, the matches produced are the same for every $W = cW$ for any positive scalar c. The matrix can be uniquely identified in many ways.

As noted in Section 3, it is beneficial to combine both propensity score matching and Mahalanobis distance. There are a variety of ways of doing this. For example, Rosenbaum and Rubin (1985) suggest doing nearest neighbor matching within calipers defined by the propensity score. We use an alternative approach (Rubin 2001). The linear predictor of the estimated propensity score, $\hat{\mu}$, is matched upon along with the covariates X once they have been adjusted so as to be uncorrelated with the linear predictor. Adjustment is accomplished accomplished by regressing each covariate on the estimated linear predictor:

$$X_k \quad = \hat{\alpha} + \hat{\mu} + \hat{\epsilon}_k$$

where k indexes the covariate number. By construction, $cor(\hat{\epsilon}_k, \hat{\mu}) = 0$. The covariates for GenMatch are defined by the following column vectors: $\hat{\mu}, \hat{\epsilon}_1, \cdots, \hat{\epsilon}_k$. We identify the weight matrix, W, by setting the weight for the propensity score to a constant such as 100.

This leaves the problem of how to choose the $k - 1$ free elements of W. Many loss criteria recommend themselves. The one we have chosen attempts to minimize a measure of the maximum observed discrepancy between the matched treated and control covariates at every iteration of optimization. For a given set of matches resulting from a given W, the loss is defined as the minimum p-value observed across a series of balance tests performed on distributions of matched baseline covariates. By default, tests are conducted for all univariate baseline covariates, as well as their first-order interactions and quadratic terms. In practice, the analyst may add tests of any function of X desired, including additional nonlinear functions and higher order interactions. The tests conducted are t-tests for the difference of means and nonparametric (bootstrap) Kolmogorov-Smirnov distributional tests. Further details of these tests are provided in Section 4.2.

The algorithm attempts to maximize this loss function by minimizing the largest discrep-

ancy at every step. Because GenMatch is minimizing the maximum discrepancy observed at each step, it is minimizing the infinity norm. This property holds even when, because of the distribution of X, the EPBR property does not hold. Therefore, if an analyst is concerned that matching may increase the bias in some linear combination of X even if the means are reduced, GenMatch allows the analyst to put in the loss function all of the linear combinations of X which may be of concern. Indeed, any nonlinear function of X can also be included in the loss function, which would ensure that bias in some nonlinear functions of X is not made inordinately large by matching.

The GenMatch loss function does allow for imbalance in functions of X to worsen as long as the maximum discrepancy is reduced. Hence, it is important the maximum discrepancy be small—i.e., that the smallest p-value be large. As the data example makes clear (Section 6), p-values conventionally understood to signal balance (e.g., 0.10), may be too low to produce reliable estimates. After GenMatch optimization, the p-values from these balance tests cannot be interpreted as true probabilities because of standard pre-test problems, but they remain useful measures of balance.

The optimization problem described above is difficult and irregular, and we utilize an evolutionary algorithm developed by Mebane and Sekhon (1998) called GENOUD, discussed in detail by Sekhon and Mebane (1998). We offer a brief overview of the algorithm in the following subsection.

4.1 Genetic Optimization

An evolutionary algorithm uses a collection of heuristic rules to modify a population of trial solutions in such a way that each generation of trial values tends to be, on average, better than its predecessor. GENOUD works for cases in which a solution is a vector of numbers that serve as the parameters of a function to be optimized. The search for a solution proceeds via a set of heuristic rules, or *operators*, each of which acts on one or more trial solutions from the current population to produce one or more trial solutions to be included in the new

population. EAs do not require derivatives to exist or the function to be continuous in order find the global optimum.

The EA in GENOUD is fundamentally a genetic algorithm (GA) in which the code-strings are vectors of floating point numbers rather than bit strings, and the GA operators take special forms tuned for the floating-point vector representation. A GA uses a set of randomized genetic operators to evolve a finite population of finite code-strings over a series of generations (Holland 1975; Goldberg 1989; Grefenstette and Baker 1989). The operators used in GA implementations vary (Davis 1991; Filho and Alippi 1994), but in an analytical sense the basic set of operators can be defined as reproduction, mutation, crossover and inversion. The variations across these operators reflect the variety of codes best suited for different applications. Reproduction entails selecting a code-string with a probability that increases with the code-string's fitness value. Crossover and inversion use pairs or larger sets of the selected code-strings to create new code-strings. Mutation randomly changes the values of elements of a single selected code-string.

Used in suitable combinations, the genetic operators tend to improve average fitness of each successive generation, though there is no guarantee that average fitness will improve between every pair of successive generations. Average fitness may well decline. But theorems exist to prove that code-substrings that have above average fitness values for the current population are sampled at an exponential rate for inclusion in the subsequent population (Holland 1975, 139–140). Each generation's population contains a biased sample of code-strings, so that a substring's performance in that population is a biased estimate of its average performance over all possible populations (De Jong 1993; Grefenstette 1993).

The long-run properties of a GA are best understood by thinking of the GA as a Markov chain. A state of the chain is a code-string population of the size used in the GA. For code-strings of finite length and GA populations of finite size, the state space is finite. If such a GA uses random reproduction and random mutation, all states always have a positive probability of occurring. A finite GA with random reproduction and mutation is therefore

a finite and irreducible Markov chain.[11] An irreducible, finite Markov chain converges at an exponential rate to a unique stationary distribution Billingsley (1986, 128). This means that the probability that each population occurs rapidly converges to a constant, positive value. Nix and Vose (1992; Vose 1993) use a Markov chain model to show that in a GA where the probability that each code-string is selected to reproduce is proportional to its observed fitness, the stationary distribution strongly emphasizes populations that contain code-strings that have high fitness values. They show that asymptotic in the population size—i.e., in the limit for a series of GAs with successively larger populations—populations that have suboptimal average fitness have probabilities approaching zero in the stationary distribution, while the probability for the population that has optimal average fitness approaches one. If $k > 1$ populations have optimal average fitness, then in the limiting stationary distribution the probability for each approaches $1/k$.

The crucial practical implication from the theoretical results of Nix and Vose is that a GA's success as an optimizer depends on having a sufficiently large population of code-strings. If the GA population is not sufficiently large, then the Markov chain that the GA implements is converging to a stationary distribution in which the probabilities of optimal and suboptimal states are not sharply distinguished. Suboptimal populations can be as likely or even more likely to occur than optimal ones. Because the Markov chain is irreducible, the GA will necessarily generate an optimal code-string if it is allowed to run for an unlimited number of generations. But if the stationary distribution is not favorable, the run time in terms of generations needed to produce an optimal code-string will be excessive. If π_j is the probability of an optimal code-string s_j, then the expected number of generations until s_j occurs (from an arbitrary starting population) is reasonably approximated by the recurrence time, $\mu_j = 1/\pi_j$ (Feller 1970, 393). For all but trivially small state spaces, an unfavorable stationary distribution can easily imply an expected running time in the millions of generations. But if the stationary distribution strongly emphasizes optimal populations,

[11] Feller (1970, 372–419) and Billingsley (1986, 107–142) review the relevant properties of Markov chains.

relatively few generations may be needed to find an optimal code-string. In general, the probability of producing an optimum in a fixed number of generations increases with the GA population size.

GAs have much in common with simulating annealing (SA) algorithms. SA is modeled on annealing, a physical process common in metallurgy in which a solid is slowly cooled so that when it eventually has its structure frozen, it is frozen at its minimum energy state (Cerny 1985; Kirkpatrick, Gelatt, and Vecchi 1983). Extensions to simulating annealing, called simulating tempering techniques, have been made where the temperature at which the process is cooled is taken to be an additional random variable during optimization (Geyer and Thompson 1995; Marinari and Parisi 1992). All are probabilistic methods for finding the global minimum of a loss function that may be irregular and contain several local minima. The theoretical properties of GAs and simulating annealing algorithms are very similar. Both are best understood by considering the properties of Markov chains discussed above (Bertsimas and Tsitsiklis 1993). Details of the precise GA we are using, GENOUD, are offered in Sekhon and Mebane (1998).

4.2 Balance Tests and their Properties

In order to maximize balance across treatment and control groups, it is necessary to be able to measure and test for balance. There are many issues involved with choosing appropriate tests, and we have little to add to this vast literature. The best choice of what tests to use is obviously dependent on the precise application. For example, if the application at hand is an experiment where there was randomization but the randomization failed to produce covariate balance, balance tests based on randomization inference may be both powerful and require only weak assumptions (e.g., Bowers and Hansen 2005). Since we are using observational data, we cannot use such tests.

For our application we employ paired t-tests to test for differences in means (because matching produces matched pairs). And because it is important to not simply test for

differences of means, bootstrapped Kolmogorov-Smirnov tests are used for nonparametric tests of the equality of distributions. The p-value of the usual Kolmogorov-Smirnov test is not consistent when the distributions being compared are not continuous. However, as Abadie (2002) proves, bootstrapped Kolmogorov-Smirnov tests are consistent even when there are point masses in the distributions. A second issue arises when trying to conduct tests on estimated propensity scores. Estimated propensity scores contain nuisance parameters associated with the estimation method used, usually logistic regression. Also, estimated propensity scores often have point masses when the baseline covariates are not continuous. We resolve both issues by computer simulation—see Appendix A for details.

5 Monte Carlo Experiments

Two different Monte Carlo experiments are presented. In the first, the experimental conditions satisfy assumptions outlined in Rubin and Thomas (1992a). In this experiment, the true propensity score is known, all baseline covariates are distributed following a normal distribution, and the mapping between X and Y is linear. Although the true propensity score is known, the estimated propensity score (with correct specification) is used in each Monte Carlo sample because it will perform better for efficiency reasons (Rosenbaum and Rubin 1983).

In the second Monte Carlo experiment, the assumptions required for EPBR are not satisfied. This experiment is a difficult case for matching. Some of the baseline variables are discrete and others contain point masses and skewed distributions. The propensity score is not correctly specified, and the mapping between X and Y is nonlinear. One thousand Monte Carlo samples are performed for both experiments.

For each Monte Carlo sample in Experiment 1, there are 50 treated observations and 100 control observations. There are three baseline covariates all of which are normally distributed with variance 1 and zero covariances. The baseline covariates for the treated observations

all have means equal to 1 and the covariates for the control group all have means equal to 0.2. The effect of treatment is zero and the outcome, Y, is generated as follows:

$$Y = X\beta + \epsilon$$

where $\epsilon \sim N(0, .5)$ and all of the β parameters are equal to 1.

For the second Monte Carlo sampling experiment, the distribution of covariates was chosen to make the setting as realistic as possible, with variables taken from the Dehejia and Wahba (1999) experimental sample of the LaLonde (1986) data. There are eight baseline variables, none of which have ellipsoidal distributions. They are age, years of education, real earnings in 1974, real earnings in 1975 and a series of indicator variables. The indicator variables are Black, Hispanic, married and high school diploma. The two earnings variables have large point masses at zero, have fat tails and are heavily skewed distributions (for more details, see Section 6). Given this, the EPBR property is unlikely to hold. In this simulation we assume a homogeneous treatment effect of \$1000. The equation that determines outcomes Y (fictional earnings) is:

$$Y = 1000\, T + .1 \exp\left[.7 \log(\text{re74} + .01) + .7 \log(\text{re75} + 0.01)\right] + \epsilon$$

where $\epsilon \sim N(0, 10)$, re74 is real earnings in 1974, re75 is real earnings in 1975 and T is the treatment indicator. The mapping from baseline covariates to Y is obviously nonlinear and only two of the baseline variables are directly related to Y.

The true propensity score for each observation, π_i, is defined by:

$$\pi_i = \text{logit}^{-1}\left[1 + .5\hat{\mu} + .01\ \text{age}^2 - .3\ \text{educ}^2 - .01 \log(\text{re74} + .01)^2 + .01 \log(\text{re75} + .01)^2\right]$$

where $\hat{\mu}$ equals the linear predictor obtained by estimating a logistic regression model, where the dependent variable is the actually observed treatment indicator in the Dehejia Wahba (1999) experimental sample of the LaLonde (1986) data. The true propensity score in the

Monte Carlo experiment is a mix of the estimated propensity score in the Dehejia Wahba sample plus extra variables in Equation 5, because we want to ensure that the propensity model estimated in the Monte Carlos samples would be badly misspecified. The linear predictor is:

$$\begin{aligned}\hat{\mu} &= 1 + 1.428 \times 10^{-4}\text{age}^2 - 2.918 \times 10^{-3}\text{educ}^2 - .2275 \text{ black} + -.8276 \text{ Hisp} \\ &+ .2071 \text{ married} - .8232 \text{ nodegree} - 1.236 \times 10^{-9}\text{re74}^2 + 5.865 \times 10^{-10}\text{re75}^2 \\ &- .04328 \text{ u74} - .3804 \text{ u75}\end{aligned}$$

where u74 is an indicator variable for if real earnings in 1974 are zero and u75 is an indicator variable for if real earnings in 1975 are zero.

In each Monte Carlo sample of this experiment, the propensity score is estimated using logistic regression and the following incorrect functional form:

$$\begin{aligned}\hat{\mu}^* &= \alpha + \alpha_1 \text{ age} + \alpha_2 \text{ educ} + \alpha_3 \text{ black} + \alpha_4 \text{ Hisp} \\ &+ \alpha_5 \text{ married} + \alpha_6 \text{ nodegree} + \alpha_7 \text{ re74} + \alpha_8 \text{ re75} \\ &+ \alpha_9 \text{ u74} + \alpha_{10} \text{ u75}\end{aligned}$$

Table 1 presents results of the first Monte Carlo experiment. The first column of the table presents the mean estimate of a given estimator and the second column the root mean square error over 1000 Monte Carlo samples. Recall that the true estimate is 0. In this experiment, where the propensity score is correctly specified and EPBR holds, GenMatch has, as expected, the lowest mean square error and the second–lowest bias. The "Raw" estimate refers the naive unadjusted ATE which is simply, in a given sample, the mean treatment outcome minus the mean control outcome. The raw bias is 604. The bias of Mahalanobis distance matching is -8.63, for the joint propensity score Mahalanobis distance estimator the bias is -5.96, and for propensity score matching the bias is -2.45. GenMatch has a bias of -2.47, which is the second lowest result, almost indistinguishable from the bias

of propensity score matching. Consistent with the results of Rosenbaum and Rubin (1983), the estimator with the second lowest root means square error is the joint propensity score Mahalanobis distance estimator.

The last two columns of Table 1 present the ratios of bias and root mean square error of a given estimator relative to GenMatch. The two other multivariate matching methods, Mahalanobis distance and the joint estimator, both have significantly more bias than GenMatch: Mahalanobis distance has 3.5 times the bias and the joint estimator 2.4 times the bias. Propensity score matching, however, only has .993 times the bias of GenMatch. This is to be expected because as Abadie and Imbens (forthcoming) prove, if one is matching on more than one continuous variable the bias is not $\sqrt{n}$ consistent. What **is** surprising is how close the GenMatch bias is to that of propensity score matching.

The Mahalanobis distance root mean square error is 1.33 times as large as that of GenMatch, for the propensity score estimator it is 1.61 times as large, and for the joint estimator it is 1.21 times as large. Although the propensity score estimator has a slightly lower bias than GenMatch (its bias is 0.993 times that of GenMatch), its root mean square error is 1.61 times as large. GenMatch dominates both other multivariate matching methods both in terms of bias and MSE, and dominates propensity score matching in terms of MSE.

Table 2 presents the results for the second Monte Carlo experiment. In this experiment the EPBR conditions do not hold and the propensity score is misspecified. GenMatch now clearly dominates all other estimators both in terms of means square error and bias. The other matching estimators have a bias which range from 16.8 times to 28 times as large as that of GenMatch. The root mean square error of the other matching estimators ranges from 1.6 times to 2.8 times as large as that of GenMatch. Once again, the joint matching estimator has the second lowest mean square error (1.63 times that of GenMatch). But this time it also has the second lowest bias (at 16.8 times that of GenMatch). This is even though the propensity score is not correctly specified. It is worth noting that both Mahalanobis distance and propensity score matching have a higher bias than the "Raw"

difference between treatment and control groups across the Monte Carlo samples.

GenMatch is the only estimator which, across samples, produces a reliable estimate of the true effect. The true causal estimate is \$1000 and the average GenMatch bias is only 25.6 (0.0256%), with root mean square error of 378.

These Monte Carlos show that in a given finite sample genetic matching will produce better balance in covariates than Mahalanobis distance even if the variables are multivariate normal. And in a given sample, it will improve the balance obtained even if the propensity score is correct. In cases where the propensity score is not correctly estimated or where the variables are discrete or otherwise such that the EPBR property is unlikely to hold, genetic matching performs well and the other matching methods perform poorly.

6 Example: Job Training Experiment

Following a research design pioneered by LaLonde (1986) and later duplicated by Dehejia and Wahba (1997; 1999; 2002; Dehejia 2005) (DW) and Smith and Todd (2001, 2005a,b) (ST), we use data from a randomized job training experiment, the National Supported Work Demonstration Program (NSW), to illustrate the reliability of GenMatch versus conventional matching methods and help resolve a longstanding debate in the literature. First we use the NSW experimental data to establish benchmark estimates of average treatment effects. Then, to create the type of observational setting and dataset typically encountered within the social sciences, data from the experimental control group are replaced by data from the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID). The goal is to determine which statistical methods, if any, are able to use observational survey data to recover the results obtained from the randomized experiment. This dataset and research design is canonical in the causal inference literature and has continued to provoke debate because it is representative of a common and important type of inferential problem in non-experimental settings.

Although the NSW data has been examined many times by DW, ST, and others (Heckman and Hotz 1989; Firpo 2004) and has been widely distributed as a teaching tool for use with matching software (Abadie and Imbens 2003; Ho, Imai, King, and Stuart 2004b; Sekhon 2005), this paper extracts new information and reaches new conclusions because genetic matching significantly boosts the degree of balance achieved. Without genetic matching it is difficult (if not impossible) to achieve a high degree of balance with the NSW data, and this is precisely what is required for this inferential problem.

6.1 Background

The NSW was a federally and privately funded program implemented in the mid-1970s to provide work experience for 6–18 months to individuals facing economic and social disadvantages. Those randomly selected to join the program participated in various types of work. Information on pre-intervention variables (pre-intervention earnings, as well as education, age, ethnicity, and marital status) was obtained from initial surveys and Social Security Administration records. By limiting himself to those assigned to treatment after December 1975, LaLonde (1986) ensured that retrospective earnings information from the experiment included calendar 1975 earnings, a covariate which he included in his models. Limiting the dataset to those who were no longer participating in the program by January 1978 ensured that the post-intervention data included calendar 1978 earnings, the outcome of interest.[12]

LaLonde (1986) obtained the experimental treatment effect of training on earnings and then compared this estimate to the results of various statistical analyzes that would have been reported by econometricians evaluating the treatment effect without the benefit of a randomized control group. His work revealed that standard econometric procedures (ordinary least squares and instrumental variable regression) were unable to replicate the experimental results, and that conventional statistical diagnostics and specification tests were of little

[12]In this paper, as in all papers following LaLonde (1986), only male participants are included in the analysis.

value. An investigator with no knowledge of the true experimental outcomes would have no way of knowing which models, techniques, and non-experimental control groups were able to produce accurate estimates of average treatment effects. LaLonde's paper was one of several of that period (Hendry 1980; Leamer 1983) to openly challenge widely-accepted methods and spark the debate over causal inference in observational settings.

More than a decade later, Dehejia and Wahba (1997, 1999) moved the debate forward by reconstructing the NSW data and adopting the LaLonde (1986) research design—with one important difference. DW cited cited theoretical and empirical labor economics literature to support the claim that it was necessary to control for more than one year of pre-intervention earnings.[13] Thus, DW limited themselves to the subset of LaLonde's NSW data for which 1974 earnings could be obtained: those individuals who joined the program early enough for the retrospective earnings information to include 1974, as well as those individuals who joined later but were known to have been unemployed prior to randomization.[14] LaLonde's original sample was composed of 297 treated observations and 425 control observations; the DW subset contains 185 treated and 260 control observations.

DW reported that nearest-neighbor propensity score matching methods were able to successfully recover the average treatment effect for the treatment group "when the range of estimated propensity scores of the treatment and comparison groups overlap, and when the variables determining assignment to treatment are observed" (1053). Their results were widely interpreted as evidence that there *was* a reliable way to estimate average causal effects in non-experimental settings under certain testable conditions.[15] These claims were questioned by Smith and Todd (2001), who replicated DW's results and published a series

[13]See Ashenfelter (1978); Ashenfelter and Card (1985); Card and Sullivan (1988).

[14]The selection of this subset provoked considerable criticism from ST.

[15]Contrary to the way some have interpreted their paper, Dehejia and Wahba (1999) did not claim that matching estimators provide a magic bullet method for evaluating social experiments: "The methods we suggest are not relevant in all situations. There may be important unobservable covariates... However, rather than giving up, or relying on assumptions about the unobserved variables, there is substantial reward in exploring first the information contained in the variables that *are* observed. In this regard, propensity score methods can offer both a diagnostic on the quality of the comparison group and a means to estimate the treatment impact" (1062).

of papers arguing that "...except in the special case of DW's sample and their propensity score specification, the matching estimators applied to the NSW data often exhibit substantial biases" (113). The debate continues, most recently with Dehejia (2005), which suggests researchers should confirm the quality of matched comparison groups by checking the sensitivity of estimates to small changes in the propensity score specification.[16]

6.2 Data

LaLonde's non-experimental estimates were based on two sets of comparison groups: the Panel Study of Income Dynamics (PSID-1) and Westat's Matched Current Population Survey-Social Security Administration file (CPS-1). Both PSID-1 and CPS-1 differ substantially from the NSW experimental treatment group in terms of age, marital status, ethnicity, and pre-intervention earnings. All mean differences across treated and control groups are significantly different from zero at any conventional significance level, except the indicator for Hispanic ethnic background.

To bridge the gap between treatment and comparison group pre-intervention characteristics, LaLonde extracted subsets from PSID-1 and CPS-1 (denoted PSID-2 and -3, and CPS-2 and -3) that he deemed similar to the treatment group in terms of particular covariates. PSID-2 selects from PSID-1 all men not working when surveyed in 1976; PSID-3 selects from PSID-1 all men not working when surveyed in either 1975 or 1976; CPS-2 selects from CPS-1 all males who not working in 1976; CPS-3 selects from CPS-1 all males unemployed in 1976 with 1975 income below the poverty level. CPS-1 has 15,992 observations, CPS-2 has 2,369 observations, and CPS-3 has 429 observations; PSID-1 has 2,490 observations, PSID-2

[16] A feature of this sensitivity test is that it requires the analyst to consider outcomes in the process of judging the quality of the matched samples. As noted by Rubin (2001), one of the great benefits of experiments is the discipline that comes of deciding how data is collected prior to observing outcomes: "If we could try hundreds of designs and for each see the resultant answer, we could...choose the design that generated the answer we wanted! The lack of availability of outcome data when designing experiments is a tremendous stimulus for 'honesty' in experiments, and can be in well–designed observational studies as well" (169). The GenMatch algorithm adheres to this principle, performing balance tests entirely blind to the outcomes of interest.

has 253 observations, and PSID-3 has 128 observations.

According to LaLonde, these smaller comparison groups were composed of individuals whose characteristics were similar to the eligibility criteria used to admit applicants into the NSW program. Even so, the subsets remain substantially different from the control group and from each other.[17] Because LaLonde's study and the NSW experiment took place so many decades ago, it is impossible to know precisely how and why the CPS- and PSID-2, and -3 subsets were constructed.

6.3 Analysis

The NSW data and LaLonde (1986) research design presents an exceptionally difficult evaluation problem. As observed by Smith and Todd (2005a), the data does not include a rich set of baseline covariates, the non-experimental comparison groups are not drawn from the same local labor market as participants, and the dependent variable is not measured identically for participants and non-participants. It is also important to note that there is *not* one uniquely well-defined experimental target result, but rather several candidate target estimates, all of which have wide confidence intervals.[18] Much of the prior literature has focused on the experimental target as the simple difference in the means of outcomes across treatment and control groups, and we shall do the same. By this metric, the estimated average treatment effect for the treated is \$886 in the LaLonde sample and \$1794 in the DW subsample, and in both cases the range of the 95% bootstrapped confidence interval exceeds \$900 (see Table 3).

[17] LaLonde's paper reports that he experimented with matching the comparison groups even more closely to the pre-training characteristics of the experimental sample, but found these closely matched comparison groups were extremely small.

[18] One might propose other experimental target estimates produced via matching, regression adjustment, or difference-in-difference estimation: all produce similar, though different, answers.

6.3.1 GenMatch

Given that most of the NSW covariates are discrete and given that the income variables are highly skewed with fat tails and contain point masses, the EPBR property is unlikely to hold. Moreover, the correct specification of the propensity score is unknown. Given these difficulties, it is not surprising that Mahalanobis distance and propensity score matching methods are unreliable in the NSW setting (see Sections 6.3.3 and 6.3.2 below) because they fail to achieve an adequate degree of balance in this very difficult evaluation problem. GenMatch reliably recovers accurate estimates of the experimental results in both DW and LaLonde samples because it incorporates sufficiently strict tests for balance and it harnesses a search algorithm capable of achieving the balance required to satisfy these tests.

Figure 1 shows the relationship between the fitness value (the lowest p-value obtained, after genetic matching, from covariate-by-covariate paired t- and KS-tests across all covariates' interaction and quadratic terms) and the square error associated with treatment effect estimates in LaLonde and DW samples. Each point represents an attempt at matching resulting in a balance test and an estimate of causal effect. Square error is calculated using the simple experimental difference in means. In both cases, the square error declines as the balance improves. Note that square error does not decline sharply until we achieve far better balance than that required by commonly administered conventional tests. In the DW sample, square error declines sharply only after fitness values exceed approximately 0.15. The LaLonde sample is even more demanding, with square error declining sharply only after fitness exceeds approximately 0.20.

Figure 2 shows how the distribution of GenMatch estimates vary with fitness. The upper panel shows the DW sample, with estimates distributed above and below the target experimental result. The 64 best-balancing estimates at the maximum fitness value are all extremely accurate, within \$52 of the experimental difference in means. The average of the best-balancing estimates is \$1734 (see Table 3). Note that in the DW sample, it is possible to get lucky and produce a reliable result even when balance has not been attained, which

helps explain why it is possible for DW to obtain accurate results with propensity scores models that do not achieve a high degree of balance. Reliable results are obtained only at the highest fitness values.

The lower panel of Figure 2 shows that in the LaLonde sample, all the GenMatch estimates are negatively biased, which is to be expected given the omission of earnings in 1974, the second year prior to training. As noted by Dehejia and Wahba (1999), there is a wealth of economics literature to show that failing to control for more than one year of pre-training income should induce bias. The sign of the bias is due to the fact that the omitted variable compromises the quality of the matches; matched control individuals are better off than the treated individuals would be in the absence of treatment. This is not surprising, given that many in the treatment group experienced two years of zero pretreatment earnings and are being matched with control individuals who did not.[19] Even so, in this difficult case, GenMatch obtains results within a small neighborhood of the experimental result.

In the LaLonde sample we obtained 451 estimates with equally best-balancing fitness values of 0.23. Of these estimates, the mean was \$281.64, the lowest was \$234 and the highest was \$345; the mode, \$285.41, was shared by 383 of these estimates. Standard errors ranged from \$702 to \$715. Recall that the experimental difference in means for the LaLonde sample was \$886, with a bootstrapped 95% confidence interval bounded at \$-55 and \$1864 (Table 3). While the GenMatch point estimates do appear biased, all 64 best-balancing estimates are well within the confidence interval of the experimental result. Furthermore, the experimental result is less than one standard error away from all the best-balancing GenMatch point estimates.[20]

[19] We have results confirming that balancing on all covariates *except* 1974 earnings does not automatically induce balance on 1974 earnings.

[20] It has not escaped our attention that the experimental median estimate in the LaLonde sample is \$485, suggesting that the experimental difference in means may be overstating the the central tendency of the true effect. Our best-balancing matching estimate with experimental data, \$528 (fitness value of 0.22), lends support to this hypothesis.

6.3.2 Propensity Score Matching

ST were correct to criticize the unreliability of simple nearest-neighbor propensity score methods in this dataset, but the GenMatch results clearly repudiate the claim that matching could not reliably solve the evaluation problem. The root of the problem is not with matching, but with overly-lenient conventional tests for balance, and the difficulty of achieving a high degree of balance using nearest-neighbor propensity score methods. There are numerous propensity score models that achieve weak but conventionally-accepted degrees of balance and produce inaccurate estimates of the effect of training. DW follow a conventional approach to balance-testing, checking balance across variables within blocks of a given propensity-score range. The DW papers do not provide detailed information on the degree of balance achieved on each variable. Instead, the authors plot the distributions of treated and control propensity scores and claim overlap. Upon our replication of Dehejia and Wahba (2002) and Dehejia (2005), it is clear that while their figures indicate overlap and their results satisfy conventional notions of balance, performing paired t-tests and Kolmogorov-Smirnov tests across matched treated and control covariates yields significant p-values.

For example, consider the case that should be most favorable to the DW argument: propensity score matching with the largest of the non-experimental control groups, CPS-1, and the most recent propensity score specification from Dehejia (2005).[21] In this case, the dummy variable for *high school degree* has a significant t-test p-value, as does its interaction with *age*, *education*, and *Black*.[22] We obtain Kolmogorov-Smirnov p-values less than 0.01 for all non-dichotomous covariates: *age*, *education*, and two years of pre-treatment income. Moreover, the ratios of covariate variances across control and treatment groups exceed 2 in several cases.

Conventional balance tests, like those used by DW and ST, typically involve t-tests

[21] We have replicated the earlier DW results across their models and datasets and results are much the same. Their propensity score matching methods do not achieve a very high degree of balance across all the confounders, their interactions, and the quadratic terms.

[22] These p-values are significant regardless of whether the paired or unpaired t-test is used.

performed across the variables included in one's model of treatment assignment. After an extensive search across propensity score models and non-experimental comparison groups, we identified five cases capable of producing paired t-test p-values greater than 0.05 for all covariates included in the propensity score model. Four involved the DW subsample, and one utilized the original LaLonde treatment group. Estimates of these models' mean treatment effects are provided in Table 3, and are clearly unreliable. One particular propensity score model utilized in association with the DW dataset produced matches such that the lowest paired t-test p-value was 0.56. The lowest KS-test p-value was zero, and, therefore, this model would have been rejected by our strict test. Thus, unsurprisingly, the resulting point estimate was outside the 95% confidence interval of the experimental result.[23]

6.3.3 Mahalanobis Matching

In the Dehejia and Wahba subsample, both simple Mahalanobis matching and Mahalanobis matching incorporating the propensity score fail to balance all baseline covariates and their interaction and quadratic terms. For example, when using the propensity score, the paired t-test p-value associated with age and several covariates interacted with age show significant differences at conventional levels. Yet even though the balance tests, the lack of distributional symmetry, and the uncertainty surrounding the propensity score give no cause for confidence in the Mahalanobis estimates, the results are, in this case, very near the experimental benchmark of \$1794. Simple Mahalanobis matching produces an estimate of \$1807. When the joint propensity score–Mahalanobis method is used, the estimate is \$1950. These estimates underscore the same fact illustrated in the upper panel of Figure 2: for any given dataset and given causal question, it is possible to get lucky and obtain good results even when the appropriate identification assumptions do not obtain.

In the LaLonde dataset, which does not include two years of pre-treatment earnings, sim-

[23]We achieve substantively similar results when we disallow matching on control units outside the support of treated units' propensity scores. After culling these control units and ensuring all t-test p-values are greater than 0.05, we still produce estimates outside the 95% confidence interval of the experimental estimate.

ple Mahalanobis matching fails a strict balance test, but multivariate Mahalanobis matching with the propensity score produces balance such that the lowest p-value for univariate KS and paired t-tests across all interaction and quadratic terms is 0.053. This is a higher degree of balance than is typically required of conventional matching estimators and much higher than the balance found in the Dehejia and Wahba subsample, but the resulting estimate of -$484 is highly biased and far below the lower bound of the experimental estimate's 95% confidence interval. As is evident from the lower panel of Figures 1 and 2, fitness values considerably greater than 0.05 are required for reliable causal inference in this setting.

7 Conclusion

The debate arising from the NSW dataset makes clear the need to find algorithms which produce matched datasets with high levels of covariate balance. Small and arbitrary changes in the propensity score model resulted in radically different causal estimates. But this is unsurprising given that although these propensity score models reduced covariate bias, significant bias remained. Some specifications simply got lucky and landed near the experimental benchmark while others, with an equally poor degree of balance, did not. The fact that so many talented researchers over several years failed to produce a propensity score model which had a high degree of covariate balance is a cautionary tale. In situations like these, machine learning can come to the rescue. There is little reason for a human to try all of the multitude of models possible to achieve balance when a computer can do this more systematically and much faster.

Historically, the matching literature, like much of statistics, has been limited by computational power. The rise of optimal full matching is a good example of what is now possible with fast computers. We think that genetic matching is another example. It is an algorithm that allows the researcher to include her substantive knowledge of the data when choosing the covariates to match on, the balance tests to conduct and the propensity score model

to include. It is also possible to start the algorithm with suggested weights and indeed it is possible for the researcher to bound the weights at will. From this substantive base, the algorithm will search and improve balance to a degree far greater than what is realistically achievable by human researchers. The debate regarding matching and the LaLonde dataset has gone on for over **eight years**, but in this time, smart and diligent researchers were unable to produce a propensity score model with anywhere near the balance achieved by our algorithm in a few hours.

There are many outstanding questions and issues. There are other ways to generalize Mahalanobis distance and these should be examined. Our proposed generalization works well in this example and in a set of Monte Carlos, but there is no claim that it is generally the best. It is unclear how to best measure the degree of covariate balance. It is surprising how little attention this issue has received in the matching literature. Finally, it is possible to use alternative optimization methods to search the space of possible solutions. New theoretical work may be especially useful in helping to design a more efficient optimization algorithm.

A Bootstrap Kolmogorov-Smirnov Test

The bootstrap is used to account for the sampling distribution of nuisance parameters, and Monte Carlo simulation is used to construct the correct test level when data contain point masses. We first outline the algorithm for calculating Kolmogorov-Smirnov p-values that are consistent even when there are point masses. In the next subsection we discuss the complete algorithm that corrects for both point masses and nuisance parameters.

A.1 Point Mass

Let Y be the $n \times 1$ data vector of interest. Let Y_1 be the first sample of Y and Y_2 the second each of which has, respectively, length n_1 and n_2. Let ks_s denote the usual Kolmogorov-Smirnov (KS) test statistic, and let ks_p denote the probability of observing ks_s

as determined by the usual formulas (Conover 1971, 295–301, 309–314).

The point mass algorithm is:

Step 1 Calculate the KS statistic in the original (full) sample using Y_1 and Y_2; denote this $\hat{ks}_s^f$.

Step 2 Resample n observations from Y with replacement B times. Denote a given resample by Y^b. Divide Y^b into two samples equal in size to n_1 and n_2, denoted Y_1^b and Y_2^b. Compute ks_s using Y_1^b and Y_2^b, denote this statistic as $\hat{ks}_s^b$.

Step 3 Calculate ks_{mc} which is the Monte Carlo p-value as: $\hat{ks}_{mc} = \sum_{b=1}^{B} 1\left\{\hat{ks}_s^b >= \hat{ks}_s^f\right\}/B$

A.2 Nuisance Parameters

The bootstrap is used to integrate over the distribution of the parameters when estimating nuisance parameters (Hall 1992). In this case, we want to obtain the p-value for the KS test applied to $\hat{Y}_1$ and $\hat{Y}_2$. The algorithm is:

Step 1 Calculate the KS statistic in the original (full) sample using $\hat{Y}_1$ and $\hat{Y}_2$; denote this $\hat{ks}_s^f$.

Step 2 Calculate the Monte Carlos KS test in the full sample; denote this as $\hat{ks}_{mc}^f$.

Step 3 Take B samples from the multivariate distribution of the nuisance parameters. For logistic regression, this would be draws from the multivariate normal distribution of the coefficient vector, where each draw is denoted by $\hat{\beta}^b$. For each bootstrap draw calculate $\hat{Y}^b$. In the case of logistic regression this would be $\hat{\mu} = X\hat{\beta}^b$. And $\hat{Y}^b = exp(\hat{\mu})/(1 + exp(\hat{\mu}))$.

Step 4 For each $\hat{Y}^b$ calculate ks_{mc} denoted as $\hat{ks}_{mc}^b$.

Step 5 The bootstrap p-value is calculated as: $\hat{ks}_{bs} = \sum_{b=1}^{B} 1\left\{\hat{ks}_{mc}^b >= \hat{ks}_{mc}^f\right\}/B$

$\hat{ks}_{bs}$ has taken into account both the sampling distribution of the nuisance parameters and calculated the empirical p-value in each Monte Carlos step, so the data may contain point masses.

References

Abadie, Alberto. 2002. "Bootstrap Tests for Distributional Treatment Effect in Instrumental Variable Models." *Journal of the American Statistical Association* 97 (457): 284–292.

Abadie, Alberto and Guido Imbens. 2003. "Matching Software for STATA and MATLAB." `http://emlab.berkeley.edu/users/imbens/estimators.shtml`.

Abadie, Alberto and Guido Imbens. forthcoming. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica*.

Ashenfelter, Orley. 1978. "Estimating the Effects of Training Programs on Earnings." *Review of Economics and Statistics* (February): 47–57.

Ashenfelter, Orley and David Card. 1985. "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs." *Review of Economics and Statistics* (November): 648–660.

Barnard, John, Constantine E. Frangakis, Jennifer L. Hill, and Donald B. Rubin. 2003. "Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City." *Journal of the American Statistical Association* 98 (462): 299–323.

Bertsimas, Dimitris and John Tsitsiklis. 1993. "Simulated Annealing." *Statistical Science* 8 (1): 10–15.

Billingsley, Patrick. 1986. *Probability and Measure*. New York: Wiley.

Bowers, Jake and Ben Hansen. 2005. "Attributing Effects to a Get-Out-The-Vote Campaign Using Full Matching and Randomization Inference." Working Paper.

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

Cerny, V. 1985. "Thermodynamical Approach to the Traveling Salesman Problem: An Efficient Simulation Algorithm." *Journal of Optimization Theory and Applications* 45 (1): 41–51.

Christakis, Nicholas A. and Theodore I. Iwashyna. 2003. "The Health Impact of Health Care on Families: A matched cohort study of hospice use by decedents and mortality outcomes in surviving, widowed spouses." *Social Science & Medicine* 57 (3): 465–475.

Cochran, William G. and Donald B. Rubin. 1973. "Controlling Bias in Observational Studies: A Review." *Sankhya,* Ser. A 35: 417–446.

Conover, William J. 1971. *Practical Nonparametric Statistics*. New York: John Wiley & Sons.

Davis, ed, Lawrence. 1991. *Handbook of Genetic Algorithms*. New York: Van Nostrand Reinhold.

Dawid, A. Phillip. 1979. "Conditional Independence in Statistical Theory." *Journal of the Royal Statistical Society, Series B* 41 (1): 1–31.

De Jong, Kenneth A. 1993. "Genetic Algorithms Are Not Function Optimizers." In L. Darrell Whitley, editor, *Foundations of Genetic Algorithms 2* San Mateo, CA: Morgan Kaufmann.

Dehejia, Rajeev. 2005. "Practical Propensity Score Matching: A Reply to Smith and Todd." *Journal of Econometrics* 125 (1–2): 355–364.

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

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

Dehejia, Rajeev H. and Sadek Wahba. 2002. "Propensity Score Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics* 84 (1): 151–161.

Diprete, Thomas A. and Henriette Engelhardt. 2004. "Estimating Causal Effects With Matching Methods in the Presence and Absence of Bias Cancellation." *Sociological Methods & Research* 32 (4): 501–528.

Epstein, Lee, Daniel E. Ho, Gary King, and Jeffrey A. Segal. 2005. "The Supreme Court During Crisis: How War Affects only Non-War Cases." *New York University Law Review* 80 (1): 1–116.

Fechner, Gustav Theodor. 1966 [1860]. *Elements of psychophysics, Vol 1.*. New York: Rinehart & Winston. Translated by Helmut E. Adler and edited by D.H. Howes and E.G. Boring.

Feller, William. 1970. *An Introduction to Probability Theory and Its Applications*. New York: Wiley. Vol. 1, 3d ed., revised printing.

Filho, Philip C. Treleaven, Jose L. Ribeiro and Cesare Alippi. 1994. "Genetic Algorithm Programming Environments." *Computer* 27: 28–43.

Firpo, Sergio. 2004. "Efficient Semiparametric Estimation of Quantile Treatment Effects." `http://www.econ.ubc.ca/sfirpo/research/qte/qtefirpo_AUG2004.pdf`.

Galiani, Sebastian, Paul Gertler, and Ernesto Schargrodsky. 2005. "Water for Life: The Impact of the Privatization of Water Services on Child Mortality." *Journal of Political Economy* 113 (1): 83–120.

Geyer, Charles J. and Elizabeth A. Thompson. 1995. "Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference." *Journal of the American Statistical Association* 90 (431): 909–920.

Goldberg, David E. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison-Wesley.

Grefenstette, John J. 1993. "Deception Considered Harmful." In L. Darrell Whitley, editor, *Foundations of Genetic Algorithms 2* San Mateo, CA: Morgan Kaufmann.

Grefenstette, John J. and James E. Baker. 1989. "How Genetic Algorithms Work: A Critical Look at Implicit Parallelism." In *Proceedings of the Third International Conference on Genetic Algorithms* San Mateo, CA: Morgan Kaufmann. pages 20–27.

Hall, Peter. 1992. *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.

Heckman, James and Joseph Hotz. 1989. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical* 84 (408): 862–74.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66 (5): 1017–1098.

Hendry, David. 1980. "Econometrices: Alchemy or Science?" *Economica* 47 (November): 387–406.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2004b. "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference." `http://gking.harvard.edu/matchit/docs/The_Lalonde_Data.html`.

Holland, John H. 1975. *Adaptation in Natural and Artificial Systems*. Ann Arbor: University of Michigan Press.

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

Horvitz, D. G. and D. J. Thompson. 1952. "A Generalization of Sampling without Replacement from a Finite Universe." *Journal of the American Statistical Association* 47: 663–685.

Imai, Kosuke. 2005. "Do Get-Out-The-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments." *American Political Science Review* 99 (2): 283–300.

Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. "Optimization by Simulated Annealing." *Science* 220 (4598): 671–680.

LaLonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76 (September): 604–20.

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

Marinari, E. and G. Parisi. 1992. "Simulated tempering: A New Monte Carlo Scheme." *Europhysics Letters* 6 (19): 451–455.

Mebane, Walter R. Jr. and Jasjeet S. Sekhon. 1998. "GENetic Optimization Using Derivatives (GENOUD)." Software Package. `http://sekhon.polisci.berkeley.edu/rgenoud/`.

Mitchell, Ann F. S. and Wojtek J. Krzanowski. 1985. "The Mahalanobis Distance and Elliptic Distributions." *Biometrika* 72 (2): 464–467.

Mitchell, Ann F. S. and Wojtek J. Krzanowski. 1989. "Amendments and Corrections: The Mahalanobis Distance and Elliptic Distributions." *Biometrika* 76 (2): 407.

Nix, Allen E. and Michael D. Vose. 1992. "Modeling Genetic Algorithms with Markov Chains." *Annals of Mathematics and Artificial Intelligence* 5: 79–88.

Rosenbaum, Paul R. 1989. "Optimal Matching for Observational Studies." *Journal of the American Statistical Association* 84 (408): 1024–1032.

Rosenbaum, Paul R. 1991. "A Characterization of Optimal Designs for Observational Studies." *Journal of the Royal Statistical Society, Series B* 53 (3): 597–610.

Rosenbaum, Paul R. 2002. *Observational Studies*. New York: Springer-Verlag 2nd edition.

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

Rosenbaum, Paul R. and Donald B. Rubin. 1985. "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score." *The American Statistician* 39 (1): 33–38.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66 (5): 688–701.

Rubin, Donald B. 1976a. "Multivariate Matching Methods That are Equal Percent Bias Reducing, I: Some Examples." *Biometrics* 32 (1): 109–120.

Rubin, Donald B. 1976b. "Multivariate Matching Methods That are Equal Percent Bias Reducing, II: Maximums on Bias Reduction for Fixed Sample Sizes." *Biometrics* 32 (1): 121–132.

Rubin, Donald B. 1977. "Assignment to a Treatment Group on the Basis of a Covariate." *Journal of Educational Statistics* 2: 1–26.

Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *Annals of Statistics* 6 (1): 34–58.

Rubin, Donald B. 1979. "Using Multivariate Sampling and Regression Adjustment to Control Bias in Observational Studies." *Journal of the American Statistical Association* 74: 318–328.

Rubin, Donald B. 1980. "Bias Reduction Using Mahalanobis-Metric Matching." *Biometrics* 36 (2): 293–298.

Rubin, Donald B. 1990. "Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies." *Statistical Science* 5 (4): 472–480.

Rubin, Donald B. 1997. "Estimating Causal Effects from Large Data Sets Using Propensity Scores." *Annals of Internal Medicine* 127 (8S): 757–763.

Rubin, Donald B. 2001. "Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation." *Health Services & Outcomes Research Methodology* 2 (1): 169–188.

Rubin, Donald B. and Elizabeth A. Stuart. 2005. "Affinely Invariant Matching Methods with Discriminant Mixtures of Proportional Ellipsoidally Symmetric Distributions." Working Paper.

Rubin, Donald B. and Neal Thomas. 1992a. "Affinely Invariant Matching Methods with Ellipsoidal Distributions." *Annals of Statistics* 20 (2): 1079–1093.

Rubin, Donald B. and Neal Thomas. 1992b. "Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Distributions." *Biometrika* 79 (4): 797–809.

Sekhon, Jasjeet S. 2004. "The Varying Role of Voter Information Across Democratic Societies." Working Paper.
URL http://sekhon.polisci.berkeley.edu/papers/SekhonInformation.pdf

Sekhon, Jasjeet S. 2005. "Matching: Multivariate and Propensity Score Matching with Automated Balance Search." Computer program available at http://sekhon.polisci.berkeley.edu/matching/.

Sekhon, Jasjeet Singh and Walter R. Mebane, Jr. 1998. "Genetic Optimization Using Derivatives: Theory and Application to Nonlinear Models." *Political Analysis* 7: 189–203.

Smith, Herbert L. 1997. "Matching with Multiple Controls to Estimate Treatment Effects in Observational Studies." *Sociological Methodology* 27: 305–353.

Smith, Jeffrey and Petra Todd. 2005a. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125 (1–2): 305–353.

Smith, Jeffrey and Petra Todd. 2005b. "Rejoinder." *Journal of Econometrics* 125 (1–2): 365–375.

Smith, Jeffrey A. and Petra E. Todd. 2001. "Reconciling Conflicting Evidence on the Performance of Propensity Score Matching Methods." *AEA Papers and Proceedings* 91 (2): 112–118.

Splawa-Neyman, Jerzy. 1990 [1923]. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." *Statistical Science* 5 (4): 465–472. Trans. Dorota M. Dabrowska and Terence P. Speed.

Vose, Michael D. 1993. "Modeling Simple Genetic Algorithms." In L. Darrell Whitley, editor, *Foundations of Genetic Algorithms 2* San Mateo, CA: Morgan Kaufmann.

Winship, Christopher and Stephen Morgan. 1999. "The estimation of causal effects from observational data." *Annual Review of Sociology* 25: 659–707.

Table 1: Experimental Condition 1: Multivariate Normal Distribution of Covariates

Estimator	Bias	RMSE	$\frac{\text{Bias}}{\text{Bias Genmatch}}$	$\frac{\text{RMSE}}{\text{RMSE Genmatch}}$
GenMatch	−2.47	13.1		
Pscore	−2.45	21.0	.993	1.61
Mahalanobis (MH)	−8.63	17.3	3.50	1.33
Pscore + MH	−5.96	16.0	2.41	1.21
Raw	−60.4	68.6	24.6	5.31

The true treatment effect is \$0. The last three columns present the ratios of bias, rmse and mse of a given estimator relative to GenMatch. The experiment is done under assumptions which satisfy the conditions outlined in Rubin and Thomas (1992a,b). In particular, the baseline covariates are multivariate normal. And the mapping between the baseline covariates and the outcome is linear.

Table 2: Experimental Condition 2: Distribution of Lalonde Covariates

Estimator	Bias	RMSE	$\frac{\text{Bias}}{\text{Bias Genmatch}}$	$\frac{\text{RMSE}}{\text{RMSE Genmatch}}$
GenMatch	25.6	455		
Pscore	512	1294	20.0	2.84
Mahalanobis (MH)	−717	959	28.0	2.11
Pscore + MH	428	743	16.8	1.63
Raw	485	1611	19.0	3.54

The true treatment effect is $1000. The last three columns present the ratios of bias, rmse and mse of a given estimator relative to GenMatch. The experiment is done under assumptions which do **not** satisfy the conditions outlined in Rubin and Thomas (1992a,b). In particular, the baseline covariates do not have symmetric ellipsoidal distributions. The covariates include discrete variables as well as semi-continuous variables with significant skew. And the mapping between the baseline covariates and the outcome is not linear.

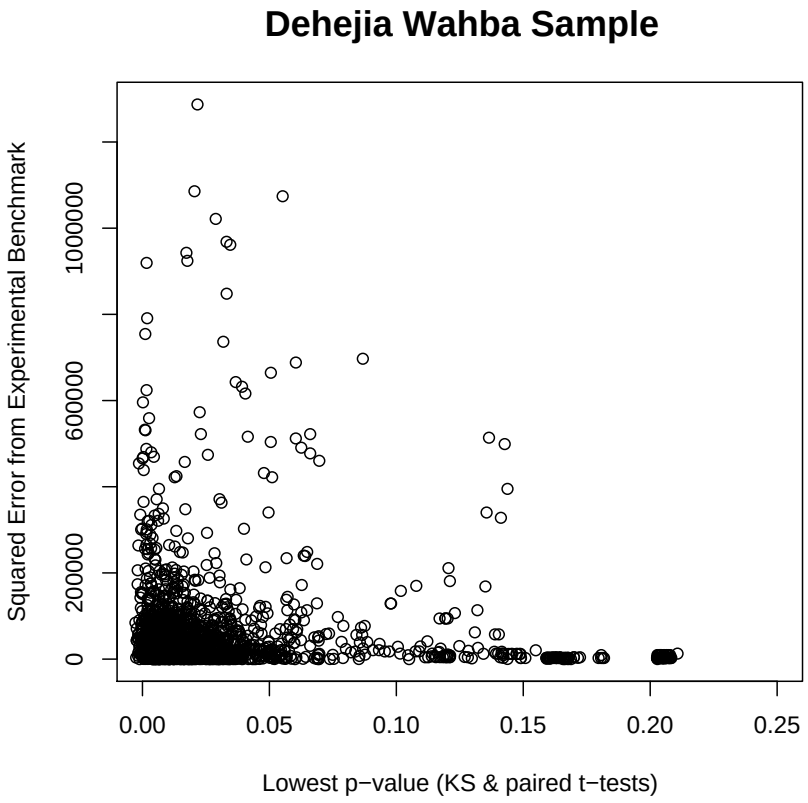
Table 3: Recovering the Result of the NSW Randomized Experiment using Nonexperimental Survey Data

			Est. Effects of Treatment on Treated		
				95% Confidence Interval	
Data	*Method*	*Balance Measure*	*Point Estimate*	Lower Bound	*Upper Bound*
DW Subsample (Benchmark)	*Experiment*	*NA*	\$1794	\$512	\$3146
DW Sample & CPS-1	*GenMatch*	fitness value = 0.21	\$1734	-\$298	\$3766
DW Subsample & PSID-2	P Score Matching	t-test p-val > 0.05	-\$487	-\$3469	\$2493
DW Subsample & PSID-3	P Score Matching	t-test p-val > 0.05	-\$1044	-\$4688	\$2600
DW Subsample & CPS-3	P Score Matching	t-test p-val > 0.05	\$705	-\$1553	\$2962
DW Subsample & CPS-3	P Score Matching	t-test p-val > 0.05	-\$295	-\$2745	\$2155

				95% Confidence Interval	
Data	*Method*	*Balance Measure*	*Point Estimate*	Lower Bound	*Upper Bound*
Lalonde (Benchmark)	*Experiment*	*NA*	\$886	-\$54	\$1864
Lalonde & CPS-1	*GenMatch*	fitness value = 0.23	\$281	-\$1122	\$1686
Lalonde & CPS-3	P Score Matching	t-test p-val > 0.05	-\$1512	-\$3748	\$724

Experimental estimates were calculated by taking the difference in the means of outcomes across participants and nonparticipants. GenMatch results average over the best-balancing estimates (64 for the DW subsample and 451 for the Lalonde sample); balance was evaluated via the GenMatch fitness value, the lowest p-value obtained via paired t- and Kolmogorov-Smirnov tests performed across the raw variables and their quadratic and interaction terms. Propensity score results show different models that achieve balance by conventional standards, covariate-by-covariate t-tests across those variables included in the model of treatment assignment—all these models fail our stricter test, with KS-test p-values ¡ 0.01 for some of these covariates. Matching-based estimates were obtained via a difference in means of matched sample outcomes. Regression adjustment does not change any these estimates by more than \$30.

Figure 1: Squared Error Declines with Balance Achieved



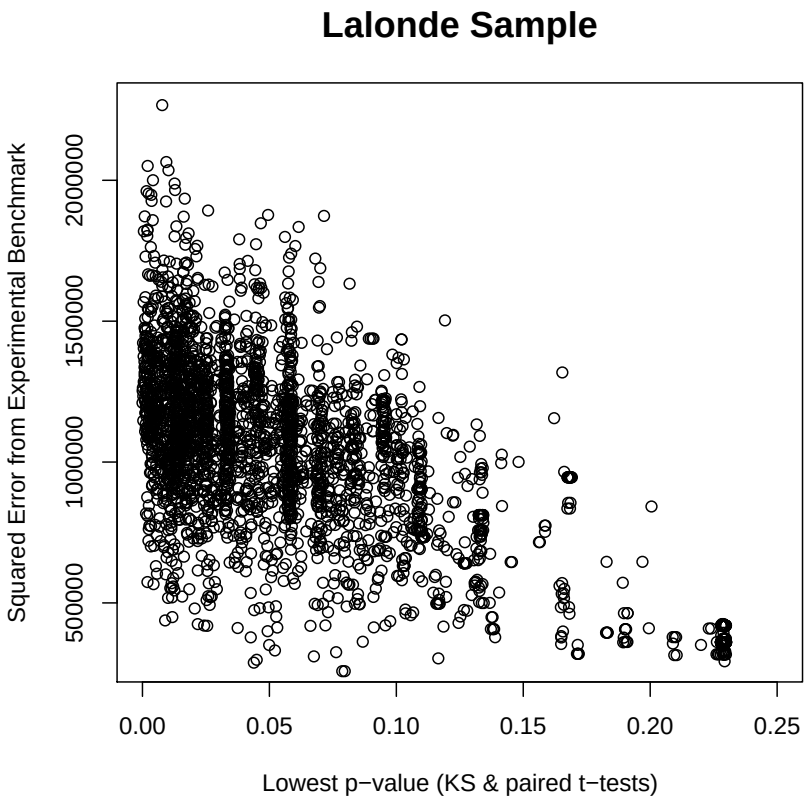


Figure 2: Reliable Estimates Require High Degree of Balance