

Reliable Estimation of Average and Quantile Causal Effects in Non-Experimental Settings*

Alexis Diamond

February 19, 2005

Abstract

This paper resolves a longstanding debate about the reliability of matching and regression methods of causal inference in nonexperimental settings. Following Lalonde (1986), data from a randomized experiment is used to establish benchmark estimates of average and quantile effects of job training. Then, to create the kind of observational setting and dataset typically encountered within the social sciences, data from the experimental control group is replaced by data from national surveys. The goal is to determine which statistical methods, if any, are able to use the observational data to recover results obtained from the randomized experiment. Quantile effects are shown to be an important and interesting quantity of interest unable, in many cases, to be reliably calculated via quantile regression. A new matching method called genetic matching is shown to be more robust across datasets than quantile regression and useful for reliably calculating average and quantile effects. Genetic matching results demonstrate that contrary to claims by both Dehejia and Wahba (1999, 2002) and Smith and Todd (2003a, c), it is possible to make reliable causal inferences across a range of models and datasets, even when two years of pretreatment earnings are unavailable.

Introduction

Econometricians intend their empirical studies to reproduce the results of experiments that use random assignment. . . One way, then, to evaluate econometric methods is to compare them against experimentally determined results. Lalonde (1986)

This paper evaluates the ability of quantile regression and matching estimators to recover average and quantile treatment effects in nonexperimental settings. Following in the footsteps of Lalonde (1986) and Dehejia and Wahba (1999, 2002) (hereafter, “DW”), data from a nationwide job training program are used to test the accuracy and precision of these econometric techniques. This job training program was conducted as a field experiment. Participants were randomly selected from the pool of applicants, and therefore the experimentally determined causal effects are easily

*This paper comes out of joint work with Jasjeet Sekhon and I thank him for his guidance and encouragement, for the *Matching* package that he developed for **R**, and for supplying a pre-release copy of his robust glm algorithm. In addition, I thank Alberto Abadie, Gary King, Walter Mebane, Kevin Quinn, Donald Rubin, and Elizabeth Stuart for many fruitful conversations. I am grateful to Roger Koenker for quick, kind, and informative replies to my emailed queries and help with his *quantreg* package. Thanks also to Rajeev Dehejia for making the NSW data available on his website. All errors are of course my own. Comments are welcome at adiamond@fas.harvard.edu. This working draft is not to be circulated without my permission. Before citing, please send me an email to confirm you have the most up-to-date version. This is version 2.1.

estimable. To test the different econometric estimators as they are typically used by practitioners in the field (as tools for causal inference in nonexperimental settings), statistical analyses are run with survey data substituted for the experimental control sample.

Although this dataset has been examined many times (Smith and Todd 2003a, c; Heckman and Hotz 1989; Firpo 2004) and has been widely distributed to serve as a teaching tool for use with matching software (Abadie and Imbens 2003; Ho et al. 2004b; Sekhon 2005c), there are three reasons that this paper is able to extract new information and reach new conclusions. First and most importantly, a new technique called genetic matching is shown to significantly boost the degree of balance achieved via full multivariate matching. Without genetic matching it is difficult (if not impossible) to achieve a high degree of balance with this dataset, even though achieving balance is precisely what the theorems require and conventional matching methods are designed (and purport) to do. Secondly, this paper represents the only detailed study of median and quantile treatment effects across all percentiles, across different model-and-data combinations discussed in the literature, and across full matching, propensity score matching, and quantile regression methods.¹ By performing so many tests, with variation across models, data sets, and estimands, the reliability and accuracy of the matching and regression methods under consideration are able to be rigorously investigated. DW and others have focused on a single estimand, the average treatment effect, which leaves open the possibility that their method works only for this particular causal question in conjunction with this data set.

This paper's third major design innovation is the way it estimates the experimentally determined causal effects that serve as benchmarks against which matching and regression are judged. All prior work has taken the experimentally determined causal effect as a fixed-point target. This paper bootstraps confidence intervals for the experimentally-determined causal effects, estimating the uncertainty associated with the true estimate. The magnitude of this underlying uncertainty puts matching and regression-generated results in proper perspective. For example, a matching estimate of median treatment effect that is \$500 away from the original experimental estimate is not problematic if the average bootstrapped experimental estimate is also \$500 away from the original experimental estimate.²

This research design is able to shed new light on a longstanding and important debate in the econometrics literature and reveal inaccuracies in claims by both sides. Contrary to claims by Smith and Todd (2003a, c), it *is* possible to use matching methodologies to reliably estimate

¹Firpo (2004) very thoroughly explores the statistical properties of propensity-weighted quantile treatment effect estimation methods. His paper includes a cursory examination of a single model and treatment/control group combination for this data.

²The bootstrap, a simple resampling procedure, allows for inference about what would have happened had the original experiment been repeated with new individuals drawn from the same population.

average and quantile treatment effects. Contrary to claims by Dehejia and Wahba (1999, 2002), it *is* possible to produce reliable, robust results without including two years of pre-intervention earnings in one’s model of treatment assignment.³ Moreover, my results show that conventional approaches to quantile regression are conspicuously vulnerable to data contamination and misspecification bias, and are less reliable than intuitive matching estimators.

This paper deals with a single data set, and the generalizability of these results has yet to be determined. Nevertheless, this is *the* canonical data set in the causal inference econometrics literature, and this job training program was chosen by Lalonde and DW and continues to provoke debate because it is representative of a simple, common, and important type of causal inferential problem. A scientist setting out to make causal inference from observational or imperfectly-randomized data faces a daunting array of methodological choices. The goal of this paper is to point the way toward a strategy for making good decisions on the road to reliable causal inference.

The following section presents a brief summary of Lalonde’s research design and explains how his article sparked the present debate over Dehejia and Wahba’s results. Section 2 introduces the Rubin model of causal inference—the theoretical framework upon which this entire enterprise rests—and then goes on to define quantile regression, matching estimators, and associated diagnostic tests. Section 3 discusses the data. Section 4 describes the analysis and section 5 interprets the results. Section 6 concludes and suggests paths for future research.

1 Background

Lalonde (1986) drew upon data from a nationwide randomized job training experiment (the National Supported Work Demonstration, or “NSW”) to illustrate the shortcomings of econometric techniques that were sophisticated at the time and remain widely in use today. He obtained the experimental treatment effect of training on earnings by subtracting the average outcome in the control group from the average outcome among the treated. Then he simply compared this estimate to the results of various statistical analyses that would have been reported by econometricians evaluating the treatment effect without the benefit of a randomized control group. Lalonde’s 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 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.

³Both sets of claims were made in the context of average treatment effects, but the authors’ statements give every indication of expressing broader truth within the wider context of causal inference.

Lalonde’s paper was one of several at the time (Hendry 1980; Leamer 1983) to openly challenge the efficacy of widely-accepted methods and fuel the debate over causal inference in observational settings.

More than a decade later, Dehejia and Wahba (1999) reconstructed Lalonde’s NSW data and adopted his research design, claiming that methods based on matching 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” (p. 1053). To many (though perhaps not to DW themselves)⁴, propensity score matching addressed Lalonde’s challenge; their results were widely interpreted as evidence that there *was* a reliable way to estimate average causal effects in nonexperimental settings under certain testable conditions.

But not everyone was convinced. Jeffrey Smith and Petra Todd replicated DW’s results and published a series 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. This finding is consistent with the fact that the NSW data combined with Lalonde’s nonexperimental comparison groups do not place nonparticipants in the same local labor markets as participants, do not measure the dependent variable (earnings) in the same way across samples, and do not include a particularly rich set of covariates for matching (Smith and Todd 2003b, p. 113).

These criticisms sparked a vigorous intellectual debate that continues to the present day. Dehejia (2005) addresses some of the criticisms, but fails to adequately engage the question of how the DW subsample of treated units was chosen. Even if one accepts DW’s argument for their sample—that including and controlling for two years of pretreatment earnings is essential for one’s model of treatment assignment—this seems inadequate justification for the subsample that DW actually selected: “Why were individuals with zero earnings in months 13 to 24 before random assignment who were randomly assigned after April 1976 included, while individuals with non-zero earnings in months 13 to 24 before random assignment who were randomly assigned after April 1976 excluded?” (Smith and Todd 2003a)

Following ST’s criticisms, Dehejia (2005) became more explicit in his claims about robustness

⁴Contrary 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” (p. 1062).

to model misspecification: “There is no reason to believe that...specifications selected specifically for DW’s samples will balance the covariates in alternative samples” (p. 4). Smith and Todd (2003a), in their rejoinder, “. . .in general agree that different applications will require different propensity score specifications, it seems to us worrisome that small changes, such as restricting the dates of random assignment in the treated sample or including and excluding some observations with zero earnings, should require major changes in the propensity score” (p. 3).

The literature spawned by this debate identifies several different models of treatment assignment for use with propensity score matching estimators, each model customized to work with a particular experimental treatment group and a particular nonexperimental control group.⁵ These specifications define very similar collections of pre-intervention characteristics that differ only by their higher-order and interaction terms; none of these models are definitively focal.⁶ Nevertheless, since these are the models suggested by the literature they are the first models to be assessed (and roundly rejected) in section 4.

2 Methodology

2.1 The Rubin Model of Causal Inference

Rubin’s Causal Model (Holland 1986; Rubin 1974, 1977, 1978), the prevailing framework for causal inference throughout the sciences, prompts three simple questions: (1) What are the units under consideration? (2) What is the treatment and how is it assigned? (3) What are the outcomes of interest? Under the Rubin Model, causal inference is defined in terms of the difference between two sets of potential outcomes: potential outcomes under treatment and potential outcomes under control. The fundamental problem of causal inference is that for each unit (be it an individual person, a country-year, a DNA sequence, etc.), only one of these potential outcomes is actually observed. A unit cannot be in both the treatment group *and* the control group (King and Zeng 2004). In the applied statistics community, causal inference is frequently likened to a missing data problem, and the solution is to make appropriate inferences about missing potential outcomes.

To more formally characterize the Rubin model, I draw upon the notation and discussion in Imbens (2003). Consider N units, indexed by $i = 1, \dots, N$, randomly drawn from a large population. Let $Y_i(1)$ denote the potential outcome for individual i following treatment, and $Y_i(0)$ denote the potential outcome for that individual in the absence of treatment. Let W be a treatment

⁵Experimental treatment groups include: (1) the full Lalonde group selected for treatment in the randomized job training experiment, (2) DW’s group limited to those with two years of pre-treatment income data, and (3) a third group constructed by ST to address their problems with DW’s chosen treatment group. This third group is not utilized in this paper.

⁶A focal model is one that independent analyses would be expected to converge upon.

indicator: 1 when i is in the treatment regime and 0 otherwise. SUTVA (Rubin 1978)—the stable unit treatment value assumption, frequently referred to as noninterference among units—is assumed to hold, assuring that each unit’s potential outcomes are independent of any other unit’s assignment status, and that treatment is identically defined for all units (Imbens N.d). The observed outcome for observation i is then

$$Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0), \quad (1)$$

and the effect of treatment for unit i may be defined as $Y_i(1) - Y_i(0)$. The population average treatment effect (ATE)⁷ is

$$\tau^\mu = \mathbf{E}[Y(1) - Y(0)], \quad (2)$$

and the population average treatment effect for the treated (ATT) is defined as

$$\tau_T^\mu = \mathbf{E}[Y(1) - Y(0)|W = 1]. \quad (3)$$

ATE is the average difference between the means of the two distributions of potential outcomes, $Y(1)$ and $Y(0)$. ATT is this difference for the potential outcome distributions associated with the treated units.⁸ Of the two estimands, ATT is frequently considered the more important because analysts and policymakers tend to care about the effect of treatment on those receiving the treatment.

Quantile treatment effects are defined in an analogous fashion. The population quantile treatment effect (QTE), at a quantile θ , is defined as

$$\tau^\theta = Q_\theta[(Y(1)) - Q_\theta[(Y(0))], \quad (4)$$

and the population quantile treatment effect for the treated (QTET) is defined as

$$\tau_T^\theta = Q_\theta[(Y(1|W = 1)) - Q_\theta[(Y(0|W = 1))]. \quad (5)$$

QTE is the difference between the θ quantile of the $Y(1)$ and $Y(0)$ distributions. QTET is this difference for the treated units, and it is the quantity of interest for all of the analysis in this paper. Quantile treatment effects have conventionally been estimated via quantile regression, and

⁷There are sample analogues for these population estimands. For example, the sample average treatment effect for the treated is defined as $\frac{1}{N_T} \sum_{i:W=1} [Y_i(1) - Y_i(0)]$. See Imbens (2003), pages 4-5, for an excellent discussion.

⁸Note that in a perfectly randomized experiment, $\tau^\mu = \tau_T^\mu$.

this paper is among the first to comprehensively explore matching-based alternatives.

If treatment is assumed to have a homogenous effect, ATE may be interpreted as the effect of the treatment on a single unit. In general, however, ATE is often interpreted as the typical effect on a typical unit, even though this estimand may never actually be experienced by any unit, even when estimated correctly. After all, the average of -2, -3, -4, -1, and 10 is zero, but zero neither a good approximation of the central tendency, nor is it a quantity that is actually observed. In this example the median, -2, is probably more informative than the mean, and is also a quantity that is present within the data itself. The median (the 0.5th quantile) is also far less vulnerable than the mean to influence by outliers at the tail of a distribution. Substitute 10,000 for 10 in the distribution above and the mean skyrockets while the median remains constant.

ATE, ATT, QTE, and QTET are all relatively easy to estimate with data from a randomized experiment because randomization eliminates all problems of selection bias. Given randomization, there is no systematic difference between control and treated units prior to treatment. SATTE and ATE are equal and equivalent to $\mathbf{E}[Y(1)] - \mathbf{E}[Y(0)]$, and the analogous equivalence obtains for QTE and QTET. In nonexperimental settings, selection is nonrandom and these equivalences may no longer be assumed. Let X_i denote a vector of unit-level characteristics (covariates) that are assumed to be unaffected by the treatment.⁹

The two key assumptions that allow for causal inference in nonexperimental settings are:

Assumption 1 (Unconfoundedness) $(Y(1), Y(0)) \perp\!\!\!\perp W|X$

Assumption 2 (Overlap) $0 < Pr(W = 1|X) < 1$.

Unconfoundedness (Rosenbaum and Rubin 1983a) and overlap allow for identification of full marginal distributions of $Y(1)$ and $Y(0)$, and therefore causal effects may be estimated at any moment or quantile. For estimation of ATT and QTET, a weaker unconfoundedness assumption is sufficient:

$$Y(0) \perp\!\!\!\perp W|X \tag{6}$$

This assumption is sufficient because the effects under treatment for the treated units are observed: it is their *unobserved* potential outcomes under control that must be assumed conditionally independent of treatment assignment.

Rosenbaum and Rubin (1983a) showed that directly conditioning on all covariates is not necessary if it is possible to accurately estimate a propensity score, defined as

⁹In practice, the validity of this assumption is often ensured by only including covariates measured prior to treatment. Imbens (2003) notes that these covariates can include lagged outcomes.

$$e(X) \equiv Pr(W = 1|X = x) = \mathbf{E}[W|X = x]. \quad (7)$$

Conditioning on the true propensity score eliminates all biases due to observable covariates. In practice, propensity scores are typically estimated via logit, probit, or nonparametric regression.

2.2 Quantile Regression

Ordinary least squares regression (OLS) is the most commonly used method of causal inference, even though the mantra of every introductory econometrics course is that regression results suggest correlation, not causation. Still, regression may be used for causal inference when treatment assignment is conditionally independent of outcomes, given X . In the absence of this assumption, regression is simply a device for descriptive inference and prediction: it expresses the expected value of the dependent variable as a function of observed covariates. For example, regression is well-suited to answering questions like “What is the expected income for an unmarried 26-year old college-educated black male living in a rented apartment in Cambridge, MA?” Another perfectly reasonable OLS-type question is “What is the expected difference in weight after the first year of life for a baby born after nine months versus a baby born after six months?”

One might be tempted to characterize the latter question in a causal way, as the causal effect of an extra three months of gestation on weight. Rubin’s framework suggests that it would be premature to make the leap to causal inference without first carefully considering the units, the treatment assignment, and the potential outcomes. Were one to define treatment as an additional three months in the uterus, valid causal inference would require achieving the conditional independence of treatment assignment and outcomes—ie., balance across relevant unit-level pretreatment characteristics. In this context the causal thought experiment may fail because the choice of unit is not self-evident—is it a fetus, a mother, or some construct that combines the two? Or, the thought experiment may fail because there is something fundamentally different about six-month fetuses born prematurely and six-month fetuses born after an additional three months of gestation. It may be interesting and important to estimate the difference in weights after one year of life for premature versus non-premature fetuses, but it may be difficult (and intellectually incoherent) to consider this question in the context of causal inference.

In quantile regression (QR) the expected value of conditional quantiles of the dependent variable is expressed as a function of observed covariates. QR is well-suited for questions like, “Among unmarried, college-educated black males in Cambridge, what are the expected earnings for individuals at the 95th conditional percentile of income? And how does this upper-tail of the conditional

distribution change if we look at individuals similar in every way, except that they are married?” Another good QR question is “What is the expected difference in weights between the median premature baby and the median regular baby? What about the difference in weights between the babies in each group who are at the bottom 5% of their group’s weight distribution?”

Just as OLS defines the sample mean to be the solution to the problem of minimizing the sum of the squared residuals, so too can the median be defined as the solution to the problem of minimizing the sum of the absolute value of the residuals. Quantiles other than the median are calculated by minimizing the sum of asymmetrically weighted absolute residuals, giving positive residuals a different weight from negative residuals (Koenker and Hallock 2001). Let ρ_τ denote the weighted absolute value function that yields the τ^{th} sample quantile. This sample quantile is then identified by solving

$$\min_{\xi \in \mathbb{R}} \sum \rho_\tau(y_i - \xi). \quad (8)$$

The formulas for quantile regression are very similar to those of OLS. Consider a dataset (y_1, y_2, \dots, y_n) . Following the notation in Koenker and Hallock (2001), OLS solves

$$\min_{\mu \in \mathbb{R}} \sum_{i=1}^n (y_i - \mu)^2, \quad (9)$$

producing the sample mean, an estimate of the unconditional population mean. Replacing the scalar μ by a parametric function $\mu(x_i, \beta)$ and solving

$$\min_{\beta \in \mathbb{R}} \sum_{i=1}^n \left(y_i - \mu(x_i, \beta) \right)^2 \quad (10)$$

provides an estimate of the conditional expectation function $E(Y|x)$. In quantile regression, an estimate of the conditional quantile function at quantile τ is obtained by replacing ξ in the first equation by the parametric function $\xi(x_i, \beta)$ and solving

$$\min_{\beta \in \mathbb{R}} \sum \rho_\tau \left(y_i - \xi(x_i, \beta) \right). \quad (11)$$

2.3 Matching

Matching, an increasingly popular method of estimating treatment effects, involves constructing an artificial matched data set by matching units that share the same (or at least similar) pretreatment characteristics. Units that do not get matched are typically discarded; the intuition here is that these observations cannot support causal inferences about missing potential outcomes. To the

degree that one believes that treatment assignment is ignorable and the appropriate matches have been made, one may proceed to use the artificially constructed data set as though it were the result of a matched-sample randomized experiment. Note that matching methodologies make no assumption about the data-generating process; instead, all identification restrictions relate to treatment assignment. This section describes the four matching methodologies used in this paper: full matching, propensity score matching, robust propensity score matching (hereafter referred to as robust matching), and full matching in concert with a genetic algorithm (hereafter referred to as genetic matching).

2.3.1 Full Matching

An intuitive and nonparametric way to condition on X_i is to attempt exact matching on the covariates. If the dimensionality of X_i is large, this approach becomes difficult (if not impossible). Abadie and Imbens (2004) explore the properties of these estimators and show that if a fixed number of matches is used, exact matching is inefficient, falling short of the semi-parametric efficiency bound for treatment effects. Another problem with this technique is that the full matching estimator will produce biases that do not disappear asymptotically under the standard \sqrt{N} normalization, and therefore a bias-correction adjustment is typically performed in these cases.¹⁰ These limitations aside, full matching offers the benefits of an approach that makes no functional form assumptions. It is also the only matching methodology for which confidence intervals and standard errors can be reliably calculated at the present time (Abadie and Imbens 2004).

2.3.2 Propensity Score Matching

An alternative way to condition on X_i is to match on the propensity score, the probability of being assigned to treatment (Rosenbaum and Rubin 1983a). Instead of trying to match each treated unit to a control unit along multiple dimensions (age, income, blood pressure, etc.), one matches along a single number between zero and one. Observations with the same propensity score have the same probability of assignment to treatment, and therefore the covariates are independent of treatment assignment conditional on the propensity score. Propensity score matching has long been a popular method of causal inference in the biomedical sciences and program evaluation literature and is now beginning to make inroads into political science (Epstein et al. 2004; Imai 2004; Sekhon 2005a, b; Wand 2004).

In nonexperimental settings, $p(X_i)$ is almost always unknown and must be estimated, typically by a regression-based technique like logit or probit. In this paper, I use logit to perform propensity

¹⁰Bias correction was used for results presented herein.

score matching, and I also use a new robust logit estimator Sekhon (2005d) designed to perform reliably despite data contamination and model misspecification. Sekhon’s estimator downweights y -misclassification outliers using a conditionally-unbiased bounded-influence approach (Kunsch, Stefanski and Carroll 1989).¹¹

Within the robust estimation literature, the influence function is defined as a measure of the effect of an infinitesimal data contamination at x, y , standardized by the mass of the contamination. The influence, therefore, of a particular observation, is an approximation of the effect of including or deleting that observation. The main idea of bounded-influence estimation is to impose a bound on the influence function, and then to find the most efficient estimator subject to the chosen bound (Kunsch, Stefanski and Carroll 1989). Sekhon’s estimator downweights observations according to their leverage and outlyingness, and thus should be less affected than classical estimators by the inclusion or deletion of a few observations. The tradeoff here is that when classical assumptions are met, classical estimators are going to be more efficient. Of course, when classical assumptions are *not* met, classical estimates will be incorrect, and robust estimators are in general much more likely to produce results in the neighborhood of the correct answer.

One of the most appealing aspects of propensity score matching is that even though regression is typically utilized to estimate propensity scores, the values of the coefficients are of no interest. All that matters is that the propensity score is an unbiased estimate of the probability of receiving treatment. Of course, as is usually the case in empirical work, this critical condition is untestable.

2.3.3 Genetic Matching

As noted above, full matching makes no functional form assumptions, allows for reliable standard errors and confidence intervals, and is highly intuitive. The problem with full matching is that it is very difficult to find good matches when there are multiple pretreatment covariates, especially when some of these covariates are continuous. In practice, analysts attempt to match units as closely as possible. But what does it mean for two units to be similar? Consider the task of finding a match for a black male with a college education, age 30. Which of the following is the better match: a white male with a college education, age 30, or a black male with a college education, age 35? The answer depends on the relative weight given to discrepancies in race versus age.

There are several variants on full matching and all require the specification of relative weights, even though this part of the process is typically hidden from and forgotten by the data analyst. Here are the steps associated with the simplest variant of one-to-one nearest-neighbor matching:

1. Selected covariates are normalized to mean zero, unit variance

¹¹Sekhon’s estimator is also capable of downweighting x -outliers. I do not use this feature in the present analysis.

2. Each of these normalized covariates are given a weight of one
3. One by one, each treated unit is matched to its most similar control unit, the control unit with the smallest sum of squared discrepancies across all covariates.

The math behind the multidimensional weighting is very simple. Following Abadie, Drukker and Herr (2001), let $\|x\|_V = \sqrt{x'Vx}$ be the vector norm with positive definite weight matrix V , and $\|z - x\|_V$ be defined as the distance between vectors x and z . The full matching algorithm takes each (normalized) treated unit v_t , a $k \times 1$ vector, and identifies the corresponding (normalized) control unit¹² v_c that minimizes $\|v_t - v_c\|_V$. For each treated unit, the most similar control unit is defined as the one closest in multidimensional covariate space.

The key insight of genetic matching (GM) is that the weighting scheme determines this closeness in space, which in turn determines which units are matched and the consequent degree of balance achieved. Some weighting schemes are strictly better than others, in that they lead to matches that create better balance for all variables. GM gets its name from the genetic algorithm (GA) used to search over the space of weighting schemes and identify those that improve ultimate balance.¹³ GAs are widely considered to be best-in-class for difficult problems of this type, where the search space is large, complex, and poorly understood (Mebane and Sekhon 2004; Sekhon and Mebane 1998).

The GA works by mimicking natural evolutionary problem-solving processes whereby the fittest individuals in a population produce offspring, generation after generation, that manifest their own genetic legacy and some degree of random variation. The population evolves over time and when fitness ceases to improve, the fittest individual represents the solution to the evolutionary problem. In the context of genetic matching the GA evolves weight matrices V , potential solutions to the search problem. Fitness is defined as a precise measure of balance achieved in the matched dataset resulting from a given weighting scheme.

The set of covariates for which balance is required is not necessarily equivalent to the set of covariates for which weights are assigned. For example, consider the NSW dataset, which contains seven basic covariates (age, education, marital status, etc.). A strict test for balance (Sekhon 2005f) would require it for each variable, each variable squared, and all their two-way interactions (35 covariates in all). Should the GA be sent searching for an optimal set of 35 weights? Probably not. All else equal, fewer weights translate into a smaller search space and a simpler, quicker search.

¹²There may be more than one control unit matched to a treated unit, depending on the tolerance specified for ties.

¹³The genetic algorithm used is available in Sekhon's *Rgenoud* package developed for **R** (Sekhon 2005e), and is also now built into his *Matching* package.

The trade-off is that a smaller search space may exclude a solution producing matches that achieve satisfactory balance.

The particular way that one defines balance should inform one's choice of fitness measure; the fitness measure is merely a tool, a means to the end of creating satisfactory balance. The balance measures used in this project are the p -values associated with the paired t -test and the Kolmogorov-Smirnov test. Section 2.5 discusses these tests in greater detail. Given that balance is measured by looking at p -values, and that all else equal, higher p -values are associated with higher balance, this project defined the fitness of a set of weights as the lowest p -value (after full matching) of any covariate for which balance was required.¹⁴ More research is necessary to ascertain the generalizability of this fitness measure and its appropriateness for use with other datasets.

2.4 Using Matching to Estimate Mean and Quantile Effects

Assuming that some form of matching has successfully taken place, one's matched treated and control units now resemble the observable results of a well-executed randomized experiment: the distribution of characteristics in the treated group should not be significantly different from the distribution of characteristics in the control group. To recover the mean treatment effect, one may simply subtract the mean outcome in the control group from the mean outcome among the treated. For the median treatment effect, subtract the median outcome in the control group from the median outcome among the treated. The same basic procedure is performed if a quantile other than the median is desired.

Alternatively, one could match and then perform OLS or quantile regression on the artificial data set. This two-stage regression-adjusted approach has the advantage of allowing the analyst to potentially control for key covariates, reduce estimation uncertainty, and recover standard errors in the usual fashion¹⁵ (Ho et al. 2004a). If one has matched effectively in the first-stage and achieved a high degree of balance, the matching estimate should not change significantly when the regression is applied in the second stage.

2.5 Testing for Balance

Balance across treated and control groups is an essential prerequisite for inference within the Rubin Causal framework, and achieving balance is a primary objective of all matching estimators.

¹⁴I am very grateful to Alberto Abadie for suggesting this fitness criterion.

¹⁵Note that these standard errors do not reflect the estimation uncertainty associated with the first-stage matching procedure.

Unfortunately, perfect balance is almost always an impossible goal. In nonexperimental settings, covariates selected by the analyst are mere approximations to the true forces that determine treatment and outcomes. Even if one's covariates really *do* matter, and *are* correctly chosen, it is impossible to be certain that all necessary covariates have been identified.¹⁶ Precisely which covariates should be tested is an open question. The most common and least rigorous approach is to test a small group of key covariates, or the collection of covariates included in the model of treatment assignment. Sekhon has called for the adoption of much stricter standards, requiring balance across all key covariates, these covariates squared, and two-way interactions of these covariates (Sekhon 2005f).

Given that balance is so important, it is essential to be able to obtain some sense of the degree to which balance has been achieved. There are several ways to proceed. The most common approach is to use paired *t*-tests, covariate-by-covariate, to assess whether the means of variables differ significantly across treatment and control groups. Higher *t*-statistics are associated with lower *p*-values and greater imbalance across treated and control groups.

The Kolmogorov-Smirnov (KS) test is designed to detect differences across the entire distributions of two samples. It is sensitive to differences in the means and differences across all higher moments as well. Applying the KS-test produces a *D*-statistic, and, as with the *t*-statistic, higher *D*-statistics are associated with lower *p*-values and greater imbalance. A special feature of the KS test is that it requires no distributional assumptions for the samples being tested, except that the distributions be continuous. In practice, this continuity requirement has been frequently ignored because alternative tests have been unavailable. Sekhon's *Matching* package includes a new variant of the KS-test that bootstraps to obtain correct coverage even when distributions are not continuous (Abadie 2002). This paper applies Sekhon's bootstrapped KS-test on a covariate-by-covariate basis. Sekhon (2005f) also proposes a new multivariate bootstrap KS-test that assesses balance across the estimated probabilities of treatment across matched treated and control groups. The multivariate test requires an estimation step (a logit regression to estimate the probabilities of receiving treatment), as well as an additional sampling step at the end to deal with problems caused by the logit's nuisance parameters. This paper applies Sekhon's multivariate KS-test as well.

As mentioned above, Sekhon's test requires balance across the key variables as well as their higher-order terms and interactions. Balance is assessed by all the tests above. All *p*-values are required to be insignificant at conventional levels: the higher, the better. As a rule-of-thumb, balance is considered to have been confidently achieved when all *p*-values are greater than 0.15

¹⁶There are methods of checking the sensitivity of matching results to threats from confoundedness (omitted variables bias (Rosenbaum and Rubin 1983b)).

(Sekhon 2005f).

3 Data

The NSW Demonstration 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. In this paper, as in all papers following the Dehejia and Wahba research design, only male participants are included in the analysis.

Candidates eligible for the NSW were randomized into the program between March 1975 and July 1977. According to Dehejia and Wahba (1999) one consequence of randomization over a 2-year period was that individuals who joined early in the program had different characteristics than those who entered later. DW also note another consequence: that data from the NSW are delineated in terms of experimental time. Lalonde (1986) annualized the experimental earnings data because his nonexperimental comparison groups were delineated in calendar time. By limiting himself to those assigned to treatment after December 1975, Lalonde ensured that retrospective earnings information from the experiment included calendar 1975 earnings, a covariate which he very sensibly included in his models. By likewise limiting himself to those who were no longer participating in the program by January 1978, he ensured that the post-intervention data included calendar 1978 earnings, his outcome of interest. The 1975 pre-intervention earnings and 1978 post-interventions earnings data were available for both nonexperimental comparison groups, and the size of the resulting sample was 297 treated observations and 425 control observations.

When they reviewed Lalonde’s work, Dehejia and Wahba (1999) cited theoretical and empirical labor economics literature to support the claim that more than one year of pre-intervention earnings were a necessary prerequisite for causal inference. 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. It is the selection of this subset that provoked considerable criticism from ST¹⁷ ST’s objections aside, DW’s subset of the randomized treated and control units retains a key property of the full experimental data: treatment and control groups have essentially identical distributions of pre-

¹⁷See page 4 above.

intervention variables, though this distribution differs from the distribution of covariates for the larger sample. A difference in means for the relatively small DW sample remains an unbiased estimator of the average treatment effect for this sample. This DW subset of the original Lalonde experimental sample contains 185 treated and 260 control observations.

Table 1 replicates table 1 in Dehejia and Wahba (1999). The smaller DW subset differs from Lalonde’s original sample, especially in terms of 1975 earnings. Nevertheless, the distributions of pre-intervention variables within each sample are very similar across the treatment and control groups; none of the differences is significantly different from zero at a 5% level of significance, with the exception of the indicator for whether the individual graduated high school (“no degree”).

Lalonde’s nonexperimental estimates of the treatment effect were based on two different comparison groups: the Panel Study of Income Dynamics (PSID-1) and Westat’s Matched Current Population Survey-Social Security Administration file (CPS-1). Table 1 shows these groups’ pre-intervention characteristics as well. Both PSID-1 and CPS-1 differ nontrivially 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 0 at any conventional significance level, except the indicator for “Hispanic”. To bridge the gap between treatment and comparison groups in terms of pre-intervention characteristics, Lalonde extracted subsets from PSID-1 and CPS-1 (denoted PSID-2 and -3, and CPS-2 and -3) that were similar to the treatment group in terms of single pre-intervention characteristics (such as age or employment status. PSID-2 selects from the PSID-1 group all men who were not working when surveyed in spring of 1976; PSID-3 selects from the PSID-1 group all men who were not working when surveyed in either spring of 1975 or 1976; CPS-2 selects from CPS-1 all males who were not working when surveyed in March 1976; CPS-3 selects from the CPS-1 unemployment males in 1976 whose income in 1975 was 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 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, although this table clearly shows that the subsets remain substantially different from the control group, and indeed from each other.¹⁸ 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. Given the 2 different treatment groups and the six

¹⁸Lalonde’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.

different control groups, there are a total of 12 different data combinations in all.

4 Analysis

There is wide agreement in the literature that regression-based methods are unable to reliably estimate average treatment effect of the NSW, but the reliability of quantile regression remains untested. Prior work with this dataset has taken the experimentally-determined causal effect to be a fixed point—i.e., the true ATT for Lalonde’s sample is \$886 and the true ATT for DW’s sample is \$1794. This paper bootstraps confidence intervals for the experimentally-determined causal effects, estimating the uncertainty associated with the true estimate. The magnitude of this underlying uncertainty puts matching and regression-generated results in proper perspective. Bootstrapped confidence intervals for the original experiment results (the ATT and the 10th, 50th, and 90th quantile effects) are provided in table 2.

Quantile regressions were performed using each of the six models identified by (Dehejia 2005), on each of 12 different treated/control group combinations, for a total of 72 separate analyses.¹⁹ The quantity and variety of regressions run is a reflection of the fact that none of these models are definitively focal. There is no particularly compelling reason for an analyst to choose one model over another when the only difference between them is the inclusion of a single interaction or higher-order term. Evaluating so many different models and datasets allows for a rigorous examination of the results’ robustness to model misspecification. Estimates of treatment effects at the 10th, 50th, and 90th quantiles are shown in table 3. Figures 1 and 2 show quantile regression estimates across all quantiles for CPS-2 and CPS-3.²⁰

The theorems that support matching methods require balance across treatment and control groups; it makes no sense to use matching to estimate causal effects if balance has not been achieved. Yet Dehejia’s models do not achieve adequate balance, even when judged by the most lenient standards.²¹ Consider the data combination that should be most favorable to Dehejia’s

¹⁹Four of Dehejia’s six models (models 2, 3, 5, and 6) involve 1974 earnings (RE74), a covariate that is missing for more than half of Lalonde’s experimental treated sample and available for all nonexperimental comparison groups. Matching on missingness here would not work because *none* of the control groups have missing data, so balance could never be achieved. In the four models that incorporate RE74, I simply leave it out when using Lalonde’s experimental sample (eg., in model 5, the interaction term “No Degree*RE74=0” becomes simply “No Degree”). Results in the following section show that these models performed very reliably even when used on the Lalonde sample for which they were never intended. A list of all six models follow: (1) RE75, married, black, hispanic, age, education, married*unemployed1975, nodegree*RE75, age²; (2) RE74, RE75, married, black, hispanic, education, age, black*age; (3) constant, RE74, married, black, hispanic, education, age, nodegree*(RE75=0), RE74²; (4) RE75, married, black, hispanic, age, education, black*education, hispanic*RE75, nodegree*education; (5) RE74, RE75, married, black, hispanic, education, age, married*(RE75=0), nodegree*(RE74=0); (6) RE74, RE75, married, black, hispanic, education, age, hispanic*education, RE74². All models include a constant term, and the components of interaction terms are not included on their own unless explicitly specified as covariates in their own right.

²⁰Results for CPS-1 were not shown because quantile regression’s poor performance in this case is clear enough from table 3.

²¹For this observation I am grateful to Jasjeet Sekhon.

position, the DW treated subset combined with the largest control group, CPS-1. He purports to have a particular propensity score model that achieves balance for these data. Yet, after using his model to create a matched dataset, a simple paired t -test performed covariate-by-covariate across treated and control groups shows a p -value for the variable 'no degree' of 0.01, signaling significant imbalance. The p -value for 'age' is 0.09, significant at the 10% level. Many interaction terms (eg., age*nodegree) that show a significant difference at the 1% level. Moreover, when bootstrap KS-tests are performed covariate-by-covariate, nearly all variables show very significant differences. These results, and corresponding results for the PSID data and the original Lalonde treated group are provided in table 4.

In fact, Dehejia's customized models do not achieve a high degree of balance for *any* of the dataset combinations, even those combinations for which they were specifically designed. After an exhaustive search for balance performing full, propensity score, and robust propensity score matching methods across the six models and 12 data combinations, none of the resulting 216 cases satisfied Sekhon's strict balance test (Sekhon 2005f). When the standard for balance was lowered—requiring t -test p -values greater than 0.05 for only the variables included in the model of treatment assignment—eight cases passed muster. Five were cases of propensity score matching, two were robust matching, and one was full matching. Seven involved the DW subsample; only one utilized the original Lalonde treatment group. Estimates of these models' mean and median treatment effects are provided in table 5, and are clearly unreliable.

Genetic matching was also performed on the Lalonde and DW treated groups in combination with CPS-1 and PSID-1 control groups. CPS-2, CPS-3, PSID-2, and PSID-3 were excluded because in a real analysis, they would never be constructed or used by anyone doing genetic matching; the larger and more diverse the control group, the easier it is to make good matches. GM requires specification of a weight vector and a fitness measure. To test the sensitivity to particular specifications, analyses were rerun several times with different sets of weight vectors. The fitness measure was defined as the smallest p -value obtained from paired t -tests and KS-tests across all the covariates, their quadratic terms, and all two-way interactions. Only cases that passed Sekhon's strict test were included in the analysis, which eliminated all attempts with PSID-1. Estimates of mean and median treatment effects are provided in table 6, showing results from the first ten cases that met the strict test. Figure 3 shows how the GM-estimated quantile treatment effects compare to the experimental results at every quantile. These results are reliable for nearly all quantiles, across all models, even when using the Lalonde sample that includes only a single year of pretreatment earnings.

5 Discussion

Experimental results in table 2 suggest that quantile effects are increasing across the quantiles for both the Lalonde and DW samples, and the effects in DW’s sample are considerably larger. The black circles and dotted lines in figures 1, 2, and 3 tell the same story in greater detail, showing no effects whatsoever until about the 25th quantile, followed by a slow, steady rise to the 80th quantile. Above this level, quantile effects peak very sharply but also reflect much greater estimation uncertainty.

These results demonstrate that the NSW had no effect on the lower tail of the earnings distribution. Note that this is *not* the same as saying that it had no effect on the individuals at the lower tail of the distribution—inference is at the level of the distribution, not at the level of individual trainees. The typical person in a world with training would earn somewhat more than the typical person in the counterfactual setting. Compare this to the effect at the 90th quantile, where the treatment effect is more than \$3000 and the upper bound on its 95% confidence interval is more than \$6000. These large estimates at the upper tail are the reason that the average effects are so much larger than the median effects: average effects are nothing more than an average of the quantile effects across the quantiles. Only by looking across the quantiles can we see that the average treatment effect is, in a sense, misleading. Most of the effects—for most individuals across the distribution—are much smaller than the average effect that is so commonly reported.

5.1 Quantile Regression

Table 3 shows quantile regression estimates varying widely across the different datasets and model combinations, but the unreliability of quantile regression is not surprising. Regression is notoriously vulnerable to model misspecification and data contamination, and the data from participants is intrinsically different from that of non-participants (Smith and Todd 2003a).

In the case of the Lalonde data matched with CPS-1 and PSID-1, *none* of the 36 different estimates quantile treatment effects are within the 95% intervals of the experimental results. Estimates do improve for CPS-2 and PSID-2, though many are still outside the intervals. Even the best results, with CPS-3 and PSID-3, show serious heterogeneity and inaccuracy across different models for estimates of effects at the 90th quantile, and two of the PSID-3 median-effect estimates are outside the intervals. When considering these results, one should recall that the Lalonde data does not include two years of pre-treatment earnings, information that some have claimed to be a necessary prerequisite for reliable causal inference in the context of job-training program evaluation.

Results from the control datasets combined with Dehejia-Wahba treated units show more accu-

racy but results are mixed, even though this dataset controls for an additional year of pre-treatment earnings. CPS-1 and PSID-1 perform miserably, but there are a few pockets of good results. PSID-2 does well with the median and 90th quantile; CPS-2, CPS-3, and PSID-3 do well with the 10th quantile and the median.

For both the Lalonde and DW datasets, estimates at the 90th quantile seem to be particularly unreliable, a problem largely due to the fact that the very wealthiest in the non-experimental control groups are very different from (and far wealthier than) the wealthiest NSW trainees. This problem tends to drag down the estimates of upper-tail quantile effects,²² and in turn contributes to a second source of difficulty: there are not very many observations nearby to inform the regression estimate at the 90th quantile, and those that are nearby show great heterogeneity.²³ Note that both factors would be visible to a data analyst exploring the NSW data in an observational setting, and an astute analyst would be wary of making inferences about the effects on the upper tails.

Though the Dehejia-Wahba results do vary widely, the CPS-3 and PSID-3 estimates of 10th quantile and median treatment effects are relatively stable and accurate. This is probably due to the fact that the CPS-3 and PSID-3 subsets are more similar to the trainees than the other subsets, and therefore the regression is relying less upon extrapolating outside the support of the data. A quick look at figure 2 confirms the accuracy of inferences conducted with the DW sample and CPS-2, and with both samples combined with CPS-3. Of course, if one’s goal is to make treated and control units more similar, it is difficult to understand why one would form ad-hoc subsets like CPS-2 and CPS-3 in the first place. Selecting controls that look like the treated is the *raison d’être* of matching, which affords a coherent, systematic strategy for achieving this objective.

Theory dictates that the median treatment effect is, in general, a more robust estimand than the average treatment effect. Table 3 shows that quantile regression is able to provide reliable and accurate median estimates in only certain cases, with certain datasets. It should be noted, however, that this performance is an improvement from that of OLS and its estimand, which Lalonde showed to be completely unreliable.

5.2 Matching

One look at table 4 is enough to ascertain that the propensity score models of Dehejia (2005) and Dehejia and Wahba (2002) do not obtain balance. The models in table 5 do obtain a higher degree of balance and even pass the most commonly utilized (weak) balance test (see page 18, above), but

²²This problem does not plague estimates at the lower tail, because dependent variables for both treated and control units are truncated from below at zero.

²³Note that the neighborhood of the 90th quantile is truncated only ten centiles away. Though the 10th quantile treatment effect is also truncated 10 centiles away, the homogeneity and similarity of treated and control lower tails allow for far more reliable estimates.

their estimates are also completely unreliable for estimation of both average and quantile effects.

Unlike these conventional matching methods, genetic matching produces a very high degree of balance. Results in table 6 come from matched datasets for which the lowest p -values of covariate-by-covariate paired t -tests and KS-tests (across all variables, interactions, and quadratic terms) exceed 0.15. All the estimates for average, 10th quantile and median estimates are within the 95% interval of the experimental results. Figure 3 tells a similar story, showing how the quantile effects at all quantiles track the experimental results for the vast majority of the quantiles. Note that the genetic matching estimator performs well in both the DW *and* the Lalonde datasets, across a wide variety of models and specifications, countering claims that two years of pre-treatment income are essential for reliable causal inference in this context.

The only problem encountered by the genetic matching estimator, in some cases, relates to effects at the upper tail of the distribution. Here, the estimator faces the same obstacles faced by quantile regression. Unlike the quantile regression results, however, genetic matching estimates do not uniformly underreport the quantile effects across all quantiles.

6 Conclusion

Prior empirical analysis of the NSW data has stopped at the identification of the average effects of training on earnings, omitting many potentially useful and interesting inferences about the effect of training on the entire distribution of earnings. For example, it might be useful to know if training improved outcomes for all those who received it, or how training affected the percentage of participants living in poverty. One might want to know the effect of training on income inequality. Or, one might wonder if the typical individual (or the wealthiest individuals) in a world with training would differ significantly from their counterparts in an alternative world where training did not exist. These questions require estimation of quantile treatment effects.

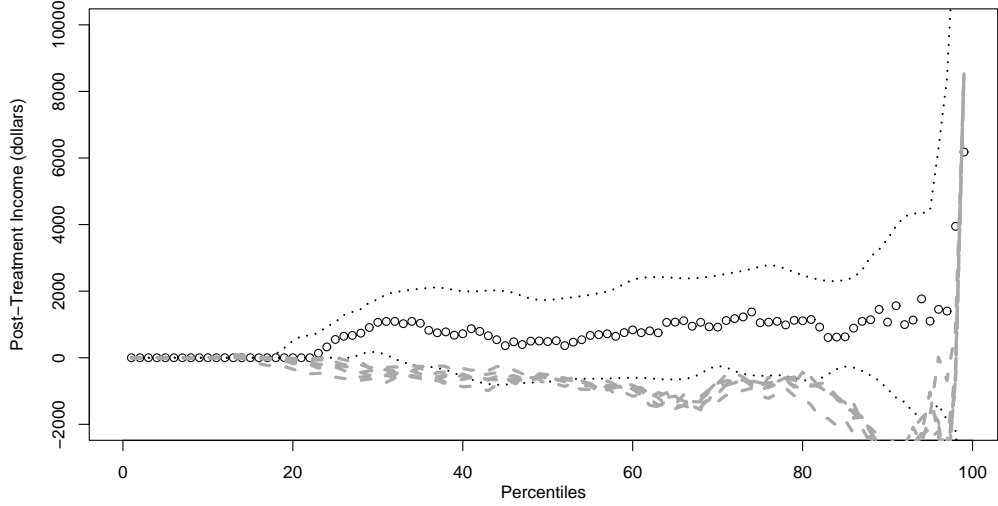
Nearly all prior work with this estimand has been performed by econometricians using quantile regression techniques, but to my knowledge, this is the first paper to evaluate this method's reliability with real data from a social experiment. Lalonde (1986) showed the poor performance of regression-based methods of estimating average treatment effects in observational settings. This paper shows that quantile regression can be an accurate and reliable approach in certain limited cases (such as the DW sample combined with CPS-3) when control groups are culled to ensure that they are similar to program participants. But even this modest performance in culled subsamples is not necessarily a victory for quantile regression: the culling was really nothing more than a rudimentary matching method, because Lalonde selected his subsamples on the basis of

a true model of treatment eligibility. Indeed, Lalonde was able to cull quite successfully without a formal matching procedure because he knew the precise eligibility requirements of the NSW training program. Most analysts estimating causal effects in observational studies do not have the luxury of published selection and eligibility criteria. In the absence of such information, culling would become ad-hoc matching, and there is no evidence to suggest that quantile regression would produce reliable results.

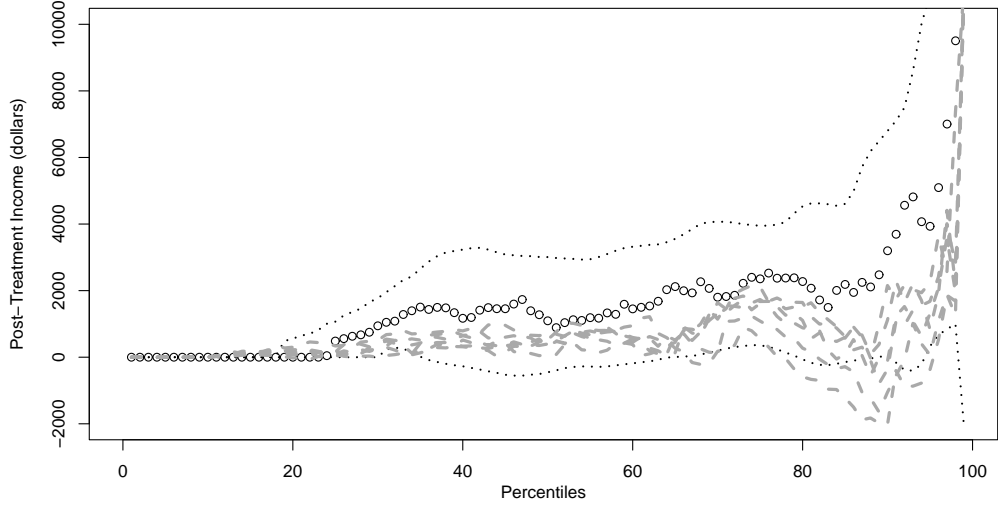
This paper is the first to estimate quantile effects using full matching estimators, and it is also the first to demonstrate that these methods are reliable when strict balance has been achieved. In the Dehejia-Wahba versus Smith-Todd debate, each side was partially right. Smith and Todd were right to argue that the matching estimators, as they were applied, were completely unreliable. But Dehejia and Wahba were on the right track—the matching theorems are useful in practice, as long as the appropriate identification assumptions are satisfied. The problem for Dehejia and Wahba was that in the NSW dataset, and in many other applications more generally, it is prohibitively difficult (if not impossible) to achieve conditional independence of treatment and outcomes by propensity score or conventional full matching. Genetic matching offers the promise of being able to balance observable characteristics much better than conventional methods, and with this degree of balance comes much greater reliability, robustness, and the possibility of easily and intuitively estimating quantile as well as mean treatment effects.

Two few caveats are in order. First, the generalizability of these results remains to be tested. There have been several large-scale social experiments in addition to the NSW, and it would be fruitful to attempt a similar project with a similar—but different—dataset. Second, one of the features distinctive to genetic matching is that there are likely to be many different solutions (weighting schemes) that produce excellent balance. Only time and future replication will tell if there are weighting schemes that produce a high degree of balance but do not give good results. Such a discovery would be a serious blow to the claims in this paper, and to matching methods more generally.

Quantile Treatment Effects For the Treated: Lalonde Sample (Not Controlling for 1974 Income)



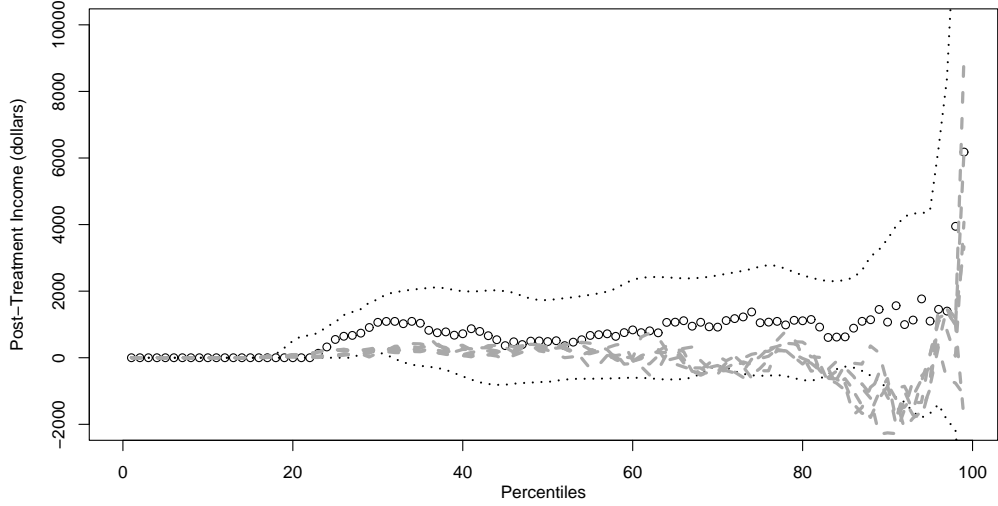
Quantile Treatment Effects For the Treated: DW Sample (Controlling for 1974 Income)



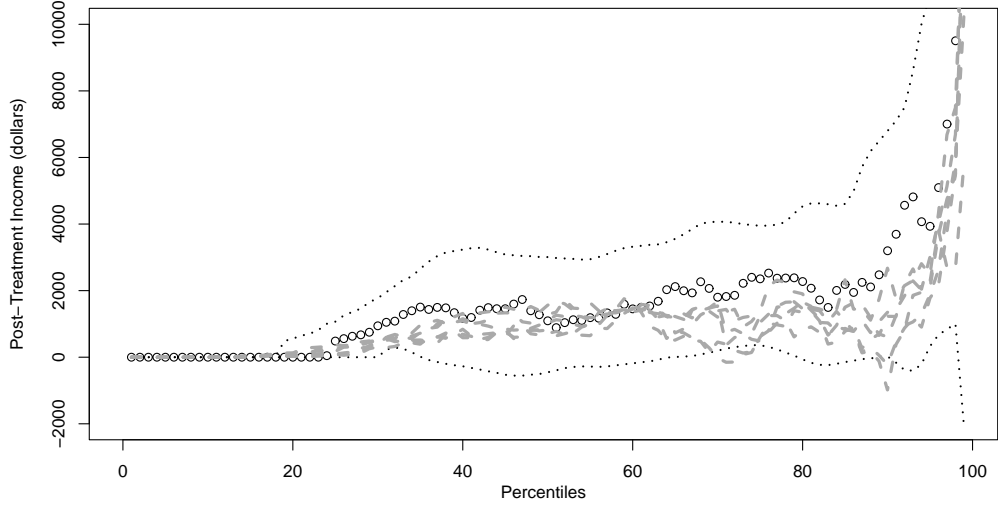
Results from Randomized Experiment in Black (95% confidence intervals) --- Quantile Regression Estimates in Gray

Figure 1: *Quantile Regression, CPS-2*: The black circles show the point estimates of quantile treatment effects at each quantile; the black dotted lines above and below are 95% bootstrap confidence intervals of this experimental result. The gray dashed lines show quantile regression-estimated quantile treatment effects (using CPS-2 controls). See table 3 for particular quantities of interest.

Quantile Treatment Effects For the Treated: Lalonde Sample (Not Controlling for 1974 Income)



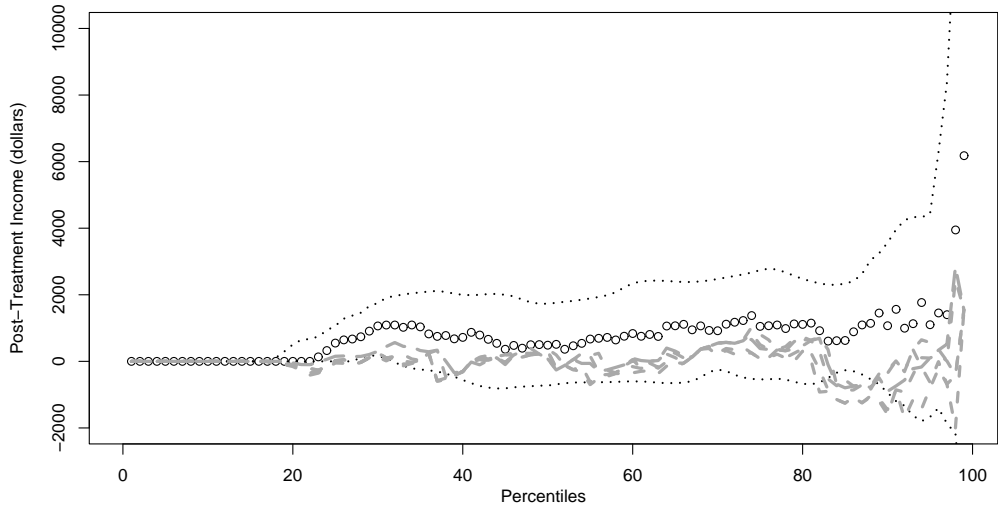
Quantile Treatment Effects For the Treated: DW Sample (Controlling for 1974 Income)



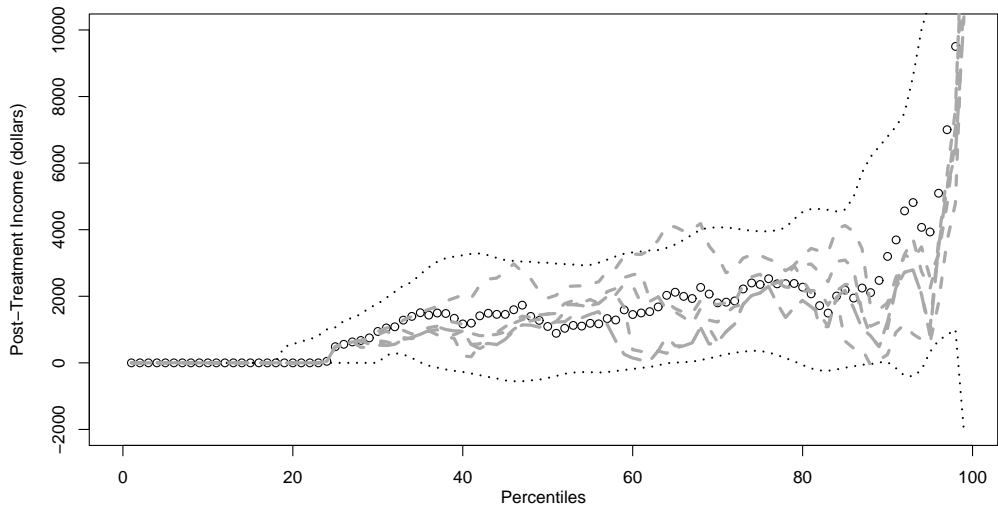
Results from Randomized Experiment in Black (95% confidence intervals) --- Quantile Regression Estimates in Gray

Figure 2: *Quantile Regression, CPS-3*: The black circles show the point estimates of quantile treatment effects at each quantile; the black dotted lines above and below are 95% bootstrap confidence intervals of this experimental result. The gray dashed lines show quantile regression-estimated quantile treatment effects (using CPS-3 controls). See table 3 for particular quantities of interest.

Quantile Treatment Effects For the Treated: Lalonde Sample (Not Controlling for 1974 Income)



Quantile Treatment Effects For the Treated: DW Sample (Controlling for 1974 Income)



Results from Randomized Experiment in Black (95% confidence intervals) --- Genetic Matching Estimates in Gray

Figure 3: *Genetic Matching, CPS-1*: The black circles show the point estimates of quantile treatment effects at each quantile; the black dotted lines above and below are 95% bootstrap confidence intervals of this experimental result. The gray dashed lines show genetic matching-estimated quantile treatment effects (using CPS-1 controls). See table 6 for particular quantities of interest.

	<i>Number of Observations</i>	<i>Age (years)</i>	<i>Education (years)</i>	<i>Black</i>	<i>Hispanic</i>	<i>High School Degree</i>	<i>Married</i>	<i>1974 Earnings (US \$)</i>	<i>1975 Earnings (US \$)</i>
NSW/Lalonde^a									
<i>Treated</i>	297	24.63	10.38	0.80	0.09	0.73	0.17		3,066
<i>Control</i>	425	24.45	10.19	0.80	0.11	0.81	0.16		3,026
DW subset^b									
<i>Treated</i>	185	25.81	10.35	0.84	0.59	0.71	0.19	2,096	1,532
<i>Control</i>	260	25.05	10.09	0.83	0.1	0.83	0.15	2,107	1,276
Comparison Groups^c									
<i>PSID-1</i>	2,490	34.85	12.11	0.25	0.032	0.31	0.87	19,429	19,063
<i>PSID-2</i>	253	36.10	10.77	0.39	0.67	0.49	0.74	11,027	7,569
<i>PSID-3</i>	128	38.25	10.30	0.45	0.18	0.51	0.70	5,566	2,611
<i>CPS-1</i>	15,992	33.22	12.02	0.07	0.07	0.29	0.71	14,016	13,650
<i>CPS-2</i>	2,3569	28.25	11.24	0.11	0.08	0.45	0.46	8,728	7,397
<i>CPS-3</i>	429	28.03	10.23	0.21	0.14	0.60	0.51	5,619	2,467

Table 1: Sample Means of Pre-Treatment Characteristics for NSW and comparison samples as originally reported in DW (1999). ^aNSW sample as constructed by Lalonde (1986). ^bSubset of the Lalonde sample for which earnings in 1974 are available. ^cDefinition of comparison groups (Lalonde 1986): PSID-1: All male household heads under age 55 not classified as retired in 1975; PSID-2: Selects from PSID-1 all men not working when surveyed in spring of 1976; PSID-3: Selects from PSID-2 all men not working in 1975; CPS-1: All CPS males under age 55; CPS-2: Selects from CPS-2 all the males who were not working when surveyed in March 1976; CPS-2 all unemployed males in poverty in 1976 with income below poverty level.

Lalonde Sample

Estimand	Estimate from Randomized Experiment	Bootstrapped 95% Intervals	
		<i>lower bound</i>	<i>upper bound</i>
<i>Mean</i>	\$886.30	-\$54.61	\$1863.61
<i>Median</i>	\$485.61	-\$727.96	\$1731.05
<i>10th quantile</i>	\$0.00	\$0.00	\$0.00
<i>90th quantile</i>	\$1070.51	-\$937.34	\$3560.87

Dehejia-Wahba Subsample

Estimand	Estimate from Randomized Experiment	Bootstrapped 95% Intervals	
		<i>lower bound</i>	<i>upper bound</i>
<i>Mean</i>	\$1794.32	\$511.51	\$3145.88
<i>Median</i>	\$1093.51	-\$453.13	\$3008.3
<i>10th quantile</i>	\$0.00	\$0.00	\$0.00
<i>90th quantile</i>	\$3197.80	\$25.21	\$6793.98

Table 2: Point estimates of mean and quantile treatment effects from the randomized experiment were calculated by simply differencing mean and quantile outcomes across treatment and control groups. Intervals were calculated for each estimand by performing a bootstrap procedure—resampling with replacement from the treated group and from the control group 1 million times, and differencing outcomes as described above.

Quantile Effects Estimated via Quantile Regression

Lalonde Sample	CPS-1			CPS-2			CPS-3		
	10 th	Median	90 th	10 th	Median	90 th	10 th	Median	90 th
Model 1	-\$229.77	-\$1600.97	-\$6028.56	-\$0.04	-\$726.91	-\$3109.23	\$0.00	\$331.72	-\$814.35
Model 2	-\$319.14	-\$902.88	-\$6446.30	\$0.00	-\$508.55	-\$2706.55	\$0.00	\$237.64	-\$2256.89
Model 3	-\$108.48	-\$916.33	-\$5187.05	\$0.00	-\$653.66	-\$3144.38	\$0.00	\$269.24	-\$864.41
Model 4	-\$217.74	-\$1142.14	-\$6812.01	\$0.00	-\$804.77	-\$2645.25	\$0.00	\$164.20	-\$784.95
Model 5	-\$333.91	-\$1712.95	-\$6677.90	\$0.00	-\$537.52	-\$2965.82	\$0.00	\$329.21	-\$1182.15
Model 6	-\$334.44	-\$1017.54	-\$6873.21	\$0.00	-\$598.34	-\$2974.76	\$0.00	\$226.24	-\$624.72
	PSID-1			PSID-2			PSID-3		
	10 th	Median	90 th	10 th	Median	90 th	10 th	Median	90 th
Model 1	-\$1058.30	-\$1491.69	-\$1038.38	\$0.00	-\$1350.25	-\$1724.91	\$0.00	\$12.54	-\$1672.51
Model 2	-\$1024.53	-\$974.63	-\$12.69	\$7.61	-\$1109.83	-\$607.62	\$0.00	\$240.56	-\$2147.88
Model 3	-\$836.15	-\$1322.91	-\$1187.10	\$0.00	-\$1401.90	-\$534.92	\$0.00	\$152.45	-\$1304.67
Model 4	-\$974.29	-\$1183.25	-\$533.83	\$0.00	-\$757.04	-\$74.09	\$0.00	\$828.32	-\$1521.09
Model 5	-\$1037.90	-\$1254.51	-\$1031.70	\$0.00	-\$1261.37	-\$88.85	\$0.00	-\$538.95	-\$1358.41
Model 6	-\$914.47	-\$1165.49	-\$1102.68	\$0.00	-\$1291.84	-\$105.39	\$0.00	-\$303.51	-\$1373.98
Dehejia-Wahba Sample	CPS-1			CPS-2			CPS-3		
	10 th	Median	90 th	10 th	Median	90 th	10 th	Median	90 th
Model 1	\$120.26	-\$288.33	-\$2699.47	\$1.76	\$73.36	-\$2029.51	\$0.00	\$817.78	\$610.06
Model 2	\$288.01	\$347.91	-\$4051.06	\$0.00	\$528.11	-\$119.28	\$0.00	\$1284.59	\$2675.18
Model 3	-\$56.71	\$360.39	-\$703.34	\$4.28	\$397.52	\$2154.05	\$0.00	\$1032.56	-\$221.67
Model 4	\$341.08	\$18.92	-\$5181.84	\$0.00	\$333.23	-\$1053.60	\$0.00	\$710.80	\$1348.93
Model 5	\$158.59	-\$451.22	-\$2097.24	\$0.53	\$690.40	-\$364.38	\$0.00	\$1562.75	\$1228.84
Model 6	\$39.52	\$454.13	-\$2477.5	\$5.04	\$548.55	\$1112.85	\$0.00	\$1108.64	-\$982.00
	PSID-1			PSID-2			PSID-3		
	10 th	Median	90 th	10 th	Median	90 th	10 th	Median	90 th
Model 1	-\$257.73	-\$76.67	\$1288.36	\$0.00	\$925.39	\$2102.52	\$0.00	\$1734.01	-\$602.54
Model 2	\$323.92	\$294.11	\$2930.61	\$289.12	\$1381.48	\$5271.48	\$0.00	\$2486.66	\$1274.70
Model 3	\$362.14	\$293.79	\$1746.09	\$185.89	\$1033.51	\$5892.32	\$0.00	\$1574.21	\$2007.40
Model 4	\$738.50	\$393.03	\$2109.64	\$135.72	\$954.35	\$2109.23	\$0.00	\$1248.27	-\$309.79
Model 5	-\$76.60	-\$705.16	\$2578.08	\$63.17	\$142.03	\$2304.58	\$0.00	\$1859.06	\$1156.17
Model 6	\$474.43	\$317.21	\$1461.78	\$255.57	\$1379.46	\$6187.62	\$0.00	\$1445.02	-\$40.84

Table 3: Quantile regression estimates of treatment effects at three key quantiles. Models (Dehejia 2003) listed on page 17; CPS/PSID details on page 16.

p-values after propensity score matching per Dehejia (2005) and Dehejia and Wahba (2002)

	Age (years)		Education (years)		Black		Hispanic		Married		No High School Degree		1975 Earnings (US \$)		1974 Earnings (US \$)			
	t-test	KS-test	t-test	KS-test	t-test	KS-test	t-test	KS-test	t-test	KS-test	t-test	KS-test	t-test	KS-test	t-test	KS-test		
Lalonde Sample																		
with CPS-1 ^a	0.00	0.00	0.29	0.04	0.92	0.00	0.76	0.00	0.30	0.00	0.00	0.00	0.88	0.04	NA	NA	NA	NA
with PSID-1 ^a	0.35	0.00	0.00	0.00	0.07	0.00	0.05	0.13	0.00	0.03	0.90	0.06	0.00	0.00	NA	NA	NA	NA
Dehejia-Wahba ^a	t-test	KS-test	t-test	KS-test	t-test	KS-test	t-test	KS-test	t-test	KS-test	t-test	KS-test	t-test	KS-test	t-test	KS-test	t-test	KS-test
with CPS-1 ^a	0.09	0.00	0.50	0.04	0.84	0.00	0.23	0.08	0.87	0.03	0.01	0.00	0.69	0.00	0.96	0.00	0.00	0.00
with CPS-1 ^b	0.00	0.00	0.07	0.05	0.00	0.02	0.01	0.03	0.08	0.01	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00
with PSID-1 ^a	0.00	0.00	0.74	0.01	0.30	0.06	0.17	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
with PSID-1 ^b	0.08	0.00	0.03	0.00	0.00	0.00	0.43	0.15	0.80	0.12	0.00	0.00	0.00	0.00	0.10	0.14	0.00	0.00

Table 4: Dehejia's models do not achieve balance across all variables in the matched samples, as revealed by the plethora of very small p -values following covariate-by-covariate paired t -tests and KS-tests. ^aPropensity score models per corresponding datasets, as defined in Dehejia (2005). ^bPropensity score models per corresponding datasets, as defined in Dehejia and Wahba (2002).

Estimated Treatment Effects
(Effects of the Treatment on the Treated)

<i>Data</i>	<i>Method</i>	<i>Model #^a</i>	<i>Mean</i>	<i>Median</i>	<i>10th Quantile</i>	<i>90th Quantile</i>
<i>Lalonde Sample (Benchmark)</i>						
	<i>Random Exper</i>	NA	\$886.30	\$485.61	\$0.00	\$1070.51
Lalonde Sample and CPS-3 ^a	Propensity	Model 4	-\$1512.78	-\$3307.05	\$0.00	-\$4139.19
<i>DW Subsample (Benchmark)</i>						
	<i>Random Exper</i>	NA	\$1794.32	\$1093.51	\$0.00	\$3197.80
DW Subsample and PSID-2	Propensity	Model 1	-\$487.75	-\$1678.60	\$0.00	-\$1478.07
DW Subsample and PSID-3 ^b	Propensity	Model 1	-\$1044.39	-\$1678.60	\$0.00	-\$1478.07
DW Subsample and CPS-1	Full	Model 2	\$2042.50	\$1842.86	\$0.00	\$1622.20
DW Subsample and CPS-3	Robust	Model 4	\$437.72	\$1061.50	-\$238.41	\$3752.48
DW Subsample and CPS-3	Propensity	Model 5	\$704.80	-\$1068.30	\$0.00	\$160.73
DW Subsample and CPS-3	Robust	Model 5	\$787.92	-\$681.83	-\$38.90	\$4324.06
DW Subsample and CPS-3	Propensity	Model 6	-\$295.00	-\$25.27	\$0.00	-\$3359.22

Table 5: Models and data combinations that satisfy only a weak balance test may not provide reliable estimates. In practice, balance is often considered adequate when covariate-by-covariate *t*-tests and (for continuous variables) Kolmogorov-Smirnov tests produce p-values > 0.05 for all variables in the model of treatment assignment. All cases above satisfy this criterion, but the estimates vary widely and are quite far from the benchmarks. ^aModels as defined by Dehejia (2003). ^bHere, quantile estimates are the same as for the case directly above because nothing much has changed; PSID-2 and PSID-3 are identical except for a few additional observations that do not make a noticeable difference once the units are matched.

Quantile Treatment Effects (\$), Estimated via Genetic Matching — Effects of Treatment on the Treated

<i>Data</i>	<i>Mean^a</i>	<i>Median</i>	<i>10th Quantile</i>	<i>90th Quantile</i>	<i>Fitness^b</i>	<i>Additional Variables Included in Matching Model^c</i>
<i>Lalonde Sample (Benchmark)</i>	886	485.61	0.00	1070.51	NA	NA
(1) Lalonde Sample & CPS-1	79 (730)	367.56	0.00	-1558.51	0.48	re75 ² , age*black, educ*hispan, black*married, hispan*married
(2) Lalonde Sample & CPS-1	296 (713)	78.42	0.00	-901.67	0.32	age*educ, educ*hispan, black*married, hispan*married
(3) Lalonde Sample & CPS-1	5 (710)	78.42	0.00	-1450.50	0.25	age*educ, black*married
(5) Lalonde Sample & CPS-1	225 (708)	311.90	0.00	-272.45	0.24	re75 ² , age*married, age*re75, educ*black
(4) Lalonde Sample & CPS-1	152 (694)	311.90	0.00	-359.64	0.18	re75 ² , educ*re75, hispan*married, married*nodegree
<i>Data</i>	<i>Mean^a</i>	<i>Median</i>	<i>10th Quantile</i>	<i>90th Quantile</i>	<i>Fitness^b</i>	<i>Additional Variables Included in Matching Model^c</i>
<i>DW Subsample (Benchmark)</i>	1794	1093.51	0.00	3197.80	NA	NA
(1) DW Subset & CPS-1	1326 (964)	1151.25	0.00	1235.81	0.21	I(re74 = 0), I(re75 = 0)
(2) DW Subset & CPS-1	1456 (919)	1466.00	0.00	250.47	0.21	re74 ² , age*re74, educ*re75, re74*re75
(3) DW Subset & CPS-1	1527 (967)	1185.23	0.00	1031.00	0.20	age*educ, educ*black, married*nodegree, re74*re75
(4) DW Subset & CPS-1	2136 (919)	1947.74	0.00	1689.78	0.16	re74*re75, hispan*married, re74 ² , nodegree*re74, educ*married... ^d
(5) DW Subset & CPS-1	1876 (929)	1466.00	0.00	1802.67	0.16	black*nodegree, married*re75, re74*re75

Table 6: Five sets of genetic matching results for each of the Lalonde and DW data are shown in the table above; additional sets results could have been shown—these estimates are provided because they were the first to be produced with fitness values greater than 0.15, passing Sekhon’s strict balance test across all quadratic and higher-order terms. ^aThe small numbers in parenthesis are standard errors. ^bThe fitness value is the lowest p -value produced by running covariate-by-covariate t -tests and KS-tests on the matched data. ^cModels (covariates selected for matching) were chosen via trial-and error, and include the basic variables plus additional interaction and higher-order terms. The variables re74 and re75 refer to real earnings in 1974 and 1975. ^d More variables were included in this specification but were omitted in the table above due to space constraints.

References

- Abadie, Alberto. 2002. "Bootstrap Tests for Distributional Treatment Effect in Instrumental Variable Models." *Journal of the American Statistical Association* 97:284–202.
- Abadie, Alberto, D. Drukker and J. Herr. 2001. Implementing Matching Estimators for Average Treatment Effects in STATA. Technical Report.
- 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. 2004. "Large Sample Properties of Matching Estimators for Average Treatment Effects." Working Paper.
- Dehejia, Rajeev. 2005. "Practical Propensity Score Matching: A Reply to Smith and Todd." *Journal of Econometrics*.
- 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.
- Dehejia, Rajeev H. and Sadek Wahba. 2002. "Propensity Score Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics* 84:151–161.
- Epstein, Lee, Daniel E. Ho, Gary King and Jeffrey Segal. 2004. "The Effect of War on the Supreme Court." <http://gking.harvard.edu/files/crisis.pdf>.
- Firpo, Sergio. 2004. "Efficient Semiparametric Estimation of Quantile Treatment Effects." http://www.econ.ubc.ca/sfirpo/research/qte/qtefirpo_AUG2004.pdf.
- 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 Association* 84:862–74.
- Hendry, David. 1980. "Econometrics: Alchemy or Science?" *Economica* 47:387–406.
- Ho, Daniel E., Kosuke Imai, Gary King and Elizabeth Stewart. 2004a. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." <http://gking.harvard.edu/files/matchp.pdf>.
- Ho, Daniel E., Kosuke Imai, Gary King and Elizabeth Stewart. 2004b. "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference." http://gking.harvard.edu/matchit/docs/The_Lalonde_Data.html.
- Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–960.
- Imai, Kosuke. 2004. "Do Get-Out-The-Vote Calls Reduce Turnout?" *American Political Science Review*. (forthcoming).
- Imbens, Guido. 2003. "Semiparametric Estimation of Average Treatment Effects under Exogeneity: A Review." http://elsa.berkeley.edu/imbens/unconf_03may25.pdf.
- Imbens, Guido. N.d. "Draft manuscript on Causal Inference and Program Evaluation."
- King, Gary and Langche Zeng. 2004. "When Can History Be Our Guide? The Pitfalls of Counterfactual Inference." <http://gking.harvard.edu/files/counterf.pdf>.
- Koenker, Roger and Kevin F. Hallock. 2001. "Quantile Regression." *Journal of Economic Perspectives* 15:143–156.

- Kunsch, H.R., A. Stefanski and R.J. Carroll. 1989. "Conditionally Unbiased Bounded-Influence Estimation in General Regression Models." *Journal of the American Statistical Association* 84:460–66.
- Lalonde, Robert. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76:604–20.
- Leamer, Edward. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73:31–43.
- Mebane, Walter R. and Jasjeet S. Sekhon. 2004. "Robust Estimation and Outlier Detection for Overdispersed Multinomial Models of Count Data." *American Journal of Political Science* 48:391–410.
- Rosenbaum, Paul and Donald B. Rubin. 1983a. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*.
- Rosenbaum, Paul and Donald Rubin. 1983b. "Assessing Sensitivity to an Unobserved Binary Covariate in an Observaitonal Study with Binary Outcomes." *Journal of the Royal Statistical Society, Series B*.
- Rubin, Donald. 1977. "Assignment to a Treatment Group on the Basis of a Covariate." *Journal of Educational Statistics* 2:1–26.
- Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66.
- Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics* 6:34–58.
- Sekhon, Jasjeet S. 2005a. "The Varying Role of Voter Information Across Democratic Societies." <http://jsekhon.fas.harvard.edu/papers/SekhonInformation.pdf>.
- Sekhon, Jasjeet S. 2005b. "The 2004 Florida Optical Voting Machine Controversy: A Causal Analysis Using Matching." <http://jsekhon.fas.harvard.edu/papers/SekhonOpticalMatch.pdf>.
- Sekhon, Jasjeet S. 2005c. "Multivariate and Propensity Score Matching Software." <http://jsekhon.fas.harvard.edu/matching/>.
- Sekhon, Jasjeet S. 2005d. "Robust Estimation Software." <http://jsekhon.fas.harvard.edu/robust/>.
- Sekhon, Jasjeet S. 2005e. "R Version of GENetic Optimization Using Derivatives (GENOUD)." <http://jsekhon.fas.harvard.edu/rgenoud/>.
- Sekhon, Jasjeet S. 2005f. "Balance Tests for Matching Estimators." Working Paper.
- Sekhon, Jasjeet S. and Walter R. Mebane. 1998. "Genetic Optimization Using Derivatives: Theory and Application to Nonlinear Models." *Political Analysis* 7:187–210.
- Smith, Jeffrey A. and Petra E. Todd. 2003a. "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*. (forthcoming).
- Smith, Jeffrey A. and Petra E. Todd. 2003b. "Reconciling Conflicting Evidence on the Performance of Propensity Score Matching Methods." *AEA Papers and Proceedings*.
- Smith, Jeffrey A. and Petra E. Todd. 2003c. "Rejoinder." *Journal of Econometrics*. (forthcoming).
- Wand, Jonathan. 2004. "<http://www.stanford.edu/class/polisci353/2004winter/reading.html>." Website dedicated to matching literature.