Matching using Semiparametric Propensity Scores

Gregory Kordas
Department of Economics
University of Pennsylvania
kordas@ssc.upenn.edu

Steven F. Lehrer
Wharton School HCSD
University of Pennsylvania
lehrers@wharton.upenn.edu

June 19, 2003

Abstract

This paper considers microeconometric evaluation by matching methods when selection in to the program under consideration is heterogeneous. Existing studies generally use parametric estimators of binary response models such as the probit and logit to estimate the propensity score, which allows for very limited forms of heterogeteity and imposes strong distributional assumptions on the error term that are often violated with the underlying data. We introduce an easy to implement matching strategy that incorporates semiparametric propensity scores that allow for very general forms of heterogeneity in response across observed covariates along the conditional willingness to participate in the treatment intervention distribution. Data from the NSW experiment, CPS and PSID are used to evaluate the performance of alternative matching estimators. We find significant evidence of heterogeneity and that the proposed algorithm generally exhibits lower bias and accurately captures the experimental treatment impact. A detailed examination of the average absolute bias errors between our procedure and matching algorithms based on parametric propensity scores indicate reductions between 6.2% and 706% of the experimental program impact.

*We are grateful to Mianna Plesca and seminar participants at the 2003 CEA meetings, 2003 IHEA World Conress, Concordia University, Florida State University, Lehigh University, McGill University, Queens University, Simon Fraser University, SUNY Albany, UNC-Greensboro, University of South Florida and the Wharton School, University of Pennsylvania for comments and suggestions which have helped to improve this paper. We are responsible for all errors.

1 Introduction

An increasing body of evidence has found that there is significant diversity and heterogeneity in response to a given policy. Heckman (2001) argues that this has profound consequences for economic theory and for economic practice. In particular, accounting for heterogeneity may improve the performance of non-experimental estimators. In this paper, we introduce and evaluate the performance of an easy to implement propensity score matching estimation strategy that explicitly accounts for heterogeneity in response across observed covariates along the conditional willingness to participate in the treatment intervention distribution.

Matching estimators evaluate the effects of a treatment intervention by comparing outcomes such as wages, employment, fertility or mortality for treated persons to those of similar persons in a comparison group. The use of the propensity score as a basis for matching treated and untreated individuals (and thus for evaluating the magnitude of treatment effects) is becoming increasingly common in clinical medicine, demographic and economic research. The propensity score is defined as the conditional probability of being treated given the individual's covariates and requires the assumption of selection on observables.[1]

Existing studies use parametric estimators of binary response models, such as the probit and logit which imposes strong distributional assumptions on the underlying data. In particular, the dangers of misspecification may be severe if the error terms are not independent and identically distributed from their known parametric distributions.[2] Kordas (2002) outlines the benefits of using Manski's (1975, 1985) binary regression quantiles to provide consistent estimates of the conditional probability at different points of the distribution. This estimator avoids the distributional restrictions embedded in the parametric approach and has the advantage that it is robust and can accommodate heteroskedasticity of unknown form. This property is extremely valuable in our setting as the estimator can accommodate problems of heterogeneity, self-selection and misclassification.

Todd (1999) presents the only other study that we are aware of that considers matching using

[1] The assumption of selection on observables requires that conditioning on the observed variables the assignment to treatment is random. Propensity score matching (Rosenbaum and Rubin (1983)) estimators reduce the dimensionality of having to match participants and non participants on the set of conditioning variables (X) by matching solely on the basis estimated propensity scores ($P(X)$).

[2] Horowitz (1993) demonstrates that misspecification of the conditional distribution of the residual in parametric binary response model is likely to be severe under heteroskedasticity and bimodality.

semiparametrically estimated propensity scores. She considers matching using the index estimated from both the semiparametric least squares estimator of Ichimura (1993) and the quasi maximum likelihood estimator of Klein and Spady (1993).[3] Her Monte Carlo study demonstrates that the gains from using the semiparametric least squares procedure relative to parametric alternatives are greatest when either the systematic component of the model is misspecified or when the error distribution is highly asymmetric.

Our approach offers several additional benefits for empirical researchers. First, this estimator does not require the researcher to select higher order or interaction terms to ensure balancing of the covariates across the treatment and non treatment groups. Recent work in economics (Dehejia and Wahba, 2002) has proposed the use of balancing tests to determine if additional higher order or interaction terms should be included in the estimates of the propensity scores but does not provide guidance on precisely which of these terms should be included. Second, the results from this estimator can also be used to calculate quantile treatment effects. Researchers can determine the average treatment effect on the treated at different points along the probability of participation distribution. Accounting for heterogeneity in the impact of the program across individuals provides a more complete picture of the effectiveness of the treatment employed.

To demonstrate the performance of our estimation strategy we use experimental data originally employed in LaLonde (1986). This data has been used in a number of studies that have evaluated the performance of different non-experimental estimators including propensity score matching. While early evidence (Dehejia and Wahba (1999)) found that propensity score matching estimators were able to replicate experimental treatment effects, more recent evidence calls these findings in question (Smith and Todd (2002)) and indicate that accounting for permanent unobserved heterogeneity does lower the estimated bias with propensity score matching estimators. If accounting for heterogeneity is indeed a major source of bias then our estimation strategy will account for it.[4]

[3] These methods estimate a conditional mean and overcome the distributional restrictions embedded in the parametric approach but allows for only limited forms of heterogeneity.

[4] Our strategy is unable to account for selection on unobservables that would result in an omitted variable bias problem.

2 Econometric Methods

2.1 Framework

Cross-sectional matching estimators compare outcomes for treatment (Y_1) and comparison group (Y_0) individuals measured at some time period after the program. We define $D_i = 1$ indicate if person i received treatment and $D_i = 0$ if not. The goal of any evaluation study is to estimate the causal effect of the treatment program. One parameter of interest is the effect of the treatment on the treated ($ATT_{D=1}(X)$), which can be defined conditional on some characteristics X as $ATT_{D=1}(X) = E(Y_1 - Y_0|X, D = 1)$. The propensity score reduces the dimension of the conditioning problem in matching by replacing an estimate of $E(Y_{0i}|D = 0, X_i)$ with an estimate of $E(Y_{0i}|D = 0, P(X_i))$; where $P(X) = Pr(D_i = 1|X)$.

. Conditioning on the propensity score yields,

$$ATT_{D=1}(X) = E\{E(Y_{1i}|P(X_i), D_i = 1) - E(Y_{0i}|P(X_i), D_i = 1)\} \tag{1}$$

where Y_{1i} and Y_{0i} are the potential outcomes in two counterfactual situations. To derive equation 1 given the definition of $P(X)$ requires that matching is to be performed over an area of common support ($0 < \Pr(D = 1|X) < 1$) and a balancing hypothesis. $D \perp X|P(X)$. The balancing hypothesis requires observations with the same propensity score to have the same distribution of observable and unobservable characteristics independent of treatment status.

The ease of implementation of these estimators has resulted in a substantial increase in their application in economics and other fields. Implementation involves two steps. In the first step, the conditional probability of participating in the treatment intervention is estimated using either a probit or logit estimator. In the second step, the researcher uses a matching algorithm to construct the matched outcomes for the treated group. Algorithms differ in the distance metric they use to determine which individuals are suitable matches to the treated persons so they can be included in the comparison group of individuals.[5] Our approach differs by estimating this probability using the following semiparametric procedure.

[5] See Smith and Todd (2002) for a comprehensive overview of alternative cross sectional matching algorithms.

2.2 Calculating the Propensity Score Semiparametrically

To avoid the distributional and other restrictions embedded in the parametric specification of $\Pr(D_i = 1|X_i)$ we use Manski's (1975, 1985) binary regression quantiles. Our use of binary regression quantiles is motivated from an estimation viewpoint as with heterogenous populations, a family of quantile estimates can provide a more complete picture of how covariates affect various conditional quantiles of the latent response variable underlying the observed binary indicator. Define the latent variable D_i^* and assume that we may write it's q-th conditional quantile function as linear index

$$Q_{D_i^*}(q|X_i) = X_i'\alpha(q), \qquad q \in (0,1). \tag{2}$$

where $\alpha(q)$ is the coefficient vector for the q-th conditional quantile. Using the equivariance property of quantile functions with respect to monotonic transformations we write the conditional quantile function of $D_i = 1\{D_i^* \geq 0\}$ as

$$Q_{D_i}(q|X_i) \equiv Q_{1\{D_i^* \geq 0\}}(q|X_i) = 1\{Q_{D_i^*}(q|X_i) \geq 0\} = 1\{X_i'\alpha(q) \geq 0\}. \tag{3}$$

This estimator is the binary response analogue to the linear quantile regression estimator introduced by Koenker and Bassett (1978) and offers a robust and efficient semiparametric alternative to commonly used parametric models. From an empirical point of view, their main advantage is their ability to model very general forms of population heterogeneity by allowing the coefficient vector $(\alpha(q))$ to vary across the conditional quantiles of the dependent variable.

Estimates of the scaled coefficients $\alpha(q)$ such that $||\alpha(q) = 1||$, are obtained by solving the quantile regression problem

$$\alpha(q) = \text{argmin}_{a:||a||=1} \left\{ S_N(a) = N^{-1} \sum_{i=1}^{N} \rho_q(D_i - 1\{X_i'a \geq 0\}) \right\}, \tag{4}$$

where $\rho_q(u) = (q - 1\{u < 0\}) \cdot u$, and $S_N(\cdot)$ is the *score function*.

Since S_N is a multimodal step function of a, binary quantile regression estimators are solutions to difficult optimization problems.[6] The discontinuities of the objective function also affect the asymptotic behavior of the estimators that have been shown to converge at the slow $N^{1/3}$ rate to a non-gaussian random variable (Kim and Pollard, 1990). To overcome these problems Horowitz

[6] Optimization is performed using the simulated annealing algorithm. See Goffe et al. (1994) for details.

(1992) smoothed the median score function and derived a smoothed median estimator that is asymptotically normally distributed [7]. Kordas (2002) extended these results to show joint asymptotic normality of families of smoothed binary quantile estimates and showed how these smoothed estimates may be optimally combined for efficient estimation.

Our main objective is to use quantile estimates to derive semiparametric estimates of the propensity score, or equivalently, semiparametric estimates of the probability that a given individual receives treatment. To this effect the un-smoothed binary quantile estimates will suffice. Thus we only consider un-smoothed estimation. With these estimates we can compute the counterfactual outcome $E(Y_{0i}|P(X_i), D_i = 1)$.

Turning to the issue of computing probabilities from quantile estimates, note that the quantile regression model in (3) implies that if an individual's q-th conditional quantile $X_i'\alpha(q)$ is (approximately) equal to zero, his conditional probability of receiving treatment is (approximately) equal to $1 - q$, i.e.,

$$\Pr(D_i = 1|X'\alpha(q) = 0) = 1 - q. \tag{5}$$

Given estimates of $\alpha(q)$ over a grid $\theta = \{q_1, q_2, \cdots, q_M | q_1 < q_2 <, \cdots < q_M\}$ of quantiles, this equation may be used to derive semiparametric *interval* probability estimates as follows. Let

$$\hat{q}_i = \text{argmin}_{q\in\theta}\{q : X_i'\alpha(q) \geq 0\} \tag{6}$$

be the smallest quantile in the grid for which i's index function is positive. Then an interval estimate of the conditional probability of $P_{i,1|X_i} \equiv Pr(D_i = 1|X_i)$ is given by

$$\hat{P}_{i,1|X_i} = [1 - \hat{q}_i, 1 - \hat{q}_{i-1}), \tag{7}$$

[7] It easy to see that the score function may be rewritten as $S_N(a)$ is the score function and it

$$S_N(a) \propto^a N^{-1} \sum_{i=1}^{N} (y_i - (1 - \tau)) \cdot 1\{X_i'a \geq 0\}$$

where the notation $\propto^a$ means "proportional in a". Horowitz (1992) proposed replacing the indicator function by a smooth function $J : R \to [0, 1]$ (e.g. a distribution function) and defined the smoothed maximum score estimator as

$$\alpha^*(q) = \left\{ S_N^*(a) \propto^a N^{-1} \sum_{i=1}^{N} (y_i - (1 - \tau)) \cdot J\left(\frac{X_i'a}{h_N}\right) \right\}$$

where h_N is a smoothing parameter that tends to zero as N becomes large.

where $\hat{q}_{i-1}$ denotes the quantile immediately preceding $\hat{q}_i$ in θ. In our application, $\theta = \{0.05, 0.10, \cdots, 0.95\}$, so, for example, if i's quantile indices are negative for quantiles below 0.70 and are positive for quantiles 0.70 and above $\hat{q}_i = 0.70$ and $P_i|x_i(\theta) = [0.30, 0.35)$.

2.3 Matching using Semiparametric Propensity Scores

Since the estimated choice probabilities are discrete (interval probabilities) the average treatment effect on the treated ($ATT_{D=1}(X)$) is calculated using stratification matching. At each probability interval, we compute the difference in average outcomes of treated and controls, providing an estimates of a quantile treatment effect ($ATT_{D=1}(X)^q$),

$$ATT_{D=1}(X)^q = \frac{\sum_{i \in L_q} Y_{1i}}{N_q^1} - \frac{\sum_{j \in L_q} Y_{0j}}{N_q^0} \tag{8}$$

where N_q^1 and N_q^0 number of treated and untreated individuals at quantile q respectively. The average treatment effect on the treated is computed using a weighted (by the number of treated) average of these quantile treatment effects as

$$ATT_{D=1}(X) = \sum_{q=1}^{Q} ATT_{D=1}(X)^q * \frac{\sum_{i \in N_q^1} D_i}{\sum_{i \in N_1} D_i} \tag{9}$$

where Q is the total number of quantiles estimated and N_1 is the total number of treated individuals that are matched.

Assuming independence of outcomes across units, the variance of $ATT_{D=1}(X)$ is given by

$$Var(ATT_{D=1}(X)) = \frac{1}{N_1} \left\{ Var(Y_{1i}) + \sum_{q=1}^{Q} \frac{N_q^1}{N_1} * \frac{N_q^1}{N_q^0} Var(Y_{0j}) \right\} \tag{10}$$

Bootstrapped standard errors could be calculated as well.[8]

[8]Notice that if a quantile contains numerous treated units and few controls it will increase the variance of the estimated mean effect of treatment on the treated. Quantiles with few treated and many controls work in an opposite manner but receive little weight in the calculation of the average treatment effect on the treated. In our empirical application we present bootstrapped standard errors since we are matching on the estimated and not the actual propensity score.

3 Returns to the NSW Job Training Program

3.1 Data

To evaluate the performance of our procedure we employ the same data used by LaLonde (1986), Heckman and Hotz (1989), Dehejia and Wahba (1999,2002), Abadie and Imbens (2002) and Smith and Todd (2002) in our study to assist in any comparisons. This literature examines whether econometric (non-experimental) estimators recover impacts on post-intervention earnings that are similar to those produced from a randomized experiment. The experimental data is drawn from a labor training program known as the National Supported Work Demonstration program.

Conducted during the 1970s, the National Supported Work Demonstration, looked at the effects of supported work on individuals with identified employment problems. Eligible applicants were assigned randomly to an experimental (participant) group, which could enroll in supported work, or to a control group, which was precluded from enrolling. Through close supervision, peer-group support, and graduated performance standards, supported work programs prepared participants to make the transition to unsubsidized employment after 12 to 18 months of program experience.

Since control and treated units were randomly assigned the experimental benchmark estimate of the treatment effect is simple to calculate. To evaluate the performance of nonexperimental estimators, treated and control units from the NSW experiment are combined with nonexperimental comparison units drawn from two national survey datasets; CPS and PSID.[9] Following Smith and Todd (2002), we consider three experimental samples (LaLonde's full sample, the Dehejia and Wahba extract, an extract containing only subjects assigned in the first four months of the program) in addition to the survey data. Summary statistics for each sample employed in the study are presented in Appendix Table 1. While there are no significant differences between the treated and control groups for each experimental sample therearesubstantial differences between these samples and the non experimental samples. The experimental sample contains more minorities particularly blacks, is younger, poorer educated, less likely to be married than the non experimental samples. Further, the earnings in all three years are substantially lower and the PSID subjects have the highest incomes. These substantial differences present challenges for any non experimental estimator.

[9] See Smith and Todd (2002) for further information on the construction of the CPS and PSID comparison group samples as well as a detailed discussion on the construction of the Dehejia and Wahba (1999) sub-sample.

3.2 Results

3.2.1 Propensity Score Estimates

We present matching estimates based on two alternative specifications of the propensity score, $\Pr(D = 1|X)$. As in Smith and Todd (2002) both the experimental treatment and control groups are included in estimating the propensity score for efficiency reasons. The first specification (henceforth referred to as specification one) is based on Dehejia and Wahba (1999,2002) includes higher order and interaction terms to satisfy balancing tests.[10] The second specification we consider omits these higher order and interaction terms from the estimating equation since in theory the inclusion of higher order and interaction terms should not affect estimates from binary regression quantiles as they are robust to heteroskedasticity of unknown form.[11]

While parametric binary response models do not allow for heterogeneity a concern exists to whether they are misspecified. We conducted simple likelihood ratio tests between the heteroskedastic logit and logit for specification one and two respectively and the null hypothesis of a homoskedastic residual is strongly rejected.[12]

The importance of heterogeneity in response to covariates is illustrated in figure 1. The figure demonstrates how the normalized quantile coefficient estimates vary across quantiles when using the early random assignment sample and PSID samples for specification 1.[13] Notice that black

[10] Balancing tests determine whether a covariate adds information on the selection process conditional on the propensity score. A slightly different set of higher order and interaction terms are used for the specifications with the CPS and PSID samples. These specification were also used in Smith and Todd (2002). See table 3 of their paper for the estimated coefficients and standard errors for a logistic regression. Note that the selection of variables to include in the estimation of the propensity score is very important since even small changes in the estimated probabilities can dramatically affect the magnitude of treatment effects in the matching stage and cause a substantial difference in the amount of bias present in the matching estimator. See Heckman, Ichimura, Smith and Todd (1998) for a discussion.

[11] While, Dehejia and Wahba (1999, 2002) did not find evidence that the treatment effect estimated was sensitive to the inclusion of these terms, they stress the importance of variable selection to ensure that the balancing hypothesis is satisfied.

[12] This assumption is rejected below the 5% level for all columns and specification with the exception of column 5 in specification 1 which is rehjected at the 15% level.

[13] Note that to improve the performance of our estimation algorithm we rescaled the covariates so that the ratio of each covariates logit coefficient relative to the logit coefficient on the education parameter ranged between 1 and 10. This sample corresponds to column 6 in Smith and Todd (2002). Our estimation algorithm and software used in

and hispanic individuals receive an increasing weight as we move from low towards higher quantiles, indicating that individuals higher in the willingness to participate distribution assign less importance to race (recall the probability interval is 1-q). Similarly and consistent with the summary statistics the coefficients on marital and dropout status become less important as individuals move higher in the willingness to participate distribution. The logit estimates seem to capture behavior fairly accurately at all but the extreme quantiles for many of the covariates.[14]

3.2.2 Treatment Effects

Table 1 presents estimates of the causal effect of the NSW Work Demonstration on earnings based on stratification matching with semiparametric propensity scores for specifications 1 and 2 in the top and bottom panel respectively. The outcome variable throughout the paper is earnings in calender year 1978. The rows differ solely in the number of bins that are employed and the lowest probability bin is excluded from the analysis.[15] For each specification, we find that the treatment impact is captured within a 95% bootstrapped confidence interval. The estimates are extremely accurate for each (even numbered column) experimental treatment sample matched with the PSID non experimental sample.

The results with twenty bins are practically identical between specification 1 and 2. Further, the results do not appear to be very sensitive to the number of bins that are used to stratify the sample match. For certain subsamples the results improve with fewer bins while for other samples the results are not as positive. Yet, as the number of bins are reduced to five, the estimates in column 2 and 6 of Table 1 decrease by approximately 67%.

The bottom panel of table 1 demonstrates how the estimated treatment effect changes when the

this study is available at http://acadfs01.whacad.wharton.upenn.edu/lehrers/software.htm.

[14] A graphical examination of the average propensity score computed by logit for each individual assigned to a given 5% quantile was also conducted. Disagreements between parametric and semiparametric propensity scores become larger at higher quantiles as the parametric models under predict the probability of participation. The general pattern of over and under prediction in these figures provides further evidence of the restrictiveness of the parametric model which tend to extrapolate the behavior of individuals near the mean to individuals that belong in the tails of the willingness to participate distribution..

[15] We present results where this bin is included in the lower half of Table 1. Table 2 presents evidence for why this bin should be excluded when conducting analysis using the CPS non experimental sample. Stratification matching estimates based on parametric propensity scores that correspond to Table 1 are presented in Appendix Table 2.

lowest interval probability bin is included in the analysis. While it is a concern that after discarding individuals, the matched sample is no longer representative recall the evidence from table 1 that demonstrated how different these samples were. The addition of this quintile dramatically reduces the magnitude of the treatment effect for the columns using the CPS sample. Even if one were to calculate semiparametric propensity score between 1% and 5% as well as 95% to 99% and include all observations the results move the estimates closer to table1 but remain slightly smaller since a few treated individuals remain in the bin with the smallest probability.

In table 2 we present Hotelling T^2 tests for differences in means (i.e. balancing tests) for each covariate used to estimate the semiparametric propensity scores within each quintile probability interval. Each entry lists the number of covariates which failed the test at the 5% level.[16] Notice that there are significant failures at the lowest probability interval capturing the dissimilarities between the experimental and CPS non experimental samples. These differences help explain the large swing in the estimated treatment effect between the top and bottom panel of table 1.[17]

If the covariates are balanced, interpretation of the quantile treatment effects are clear. In figure 2, we graph quantile treatment effects for the sample that corresponds to specification 2 and column 6 of table 1. Notice that the largest gains in the training program are received by those who had the highest and lowest probability of participation. Further, analysis indicates that the training program's success was due in part to those individuals characteristic of the experimental sample as well as those individuals who were observationally similar to the PSID sample that suffered low earnings in 1975. The training program had a negative impact for those subjects in the middle quantiles who tend to be either blacks or hispanics that had low earnings in 1974 but high earnings in 1975. This indicates that the supported work program had the largest benefits for individuals who had a permanent history of employment problems if the control group was drawn from the PSID.

[16] The results do not change significantly if we report the 10% or 20% level. Note that the result do improve significantly in all but the lowest probability interval if we report significance 1% level.

[17] Note that the majority of these individuals are not included in the parametric matching procedures due to trimming conditions. For example, Dehejia and Wahba (1999) trim the sample by deleting all observations in the control group whose estimated propensity score is less than the minimum estimated propensity score of the treatment group. Similarly, they delete all treatment observations whose estimated propensity score is above the maximum estimated propensity score of the control group. Appendix Table 2 presents stratification matching estimates with parametric propensity scores using the Dehejia and Wahba (1999) sampling criteria.

3.2.3 Bias Estimates

Evaluation bias estimates are obtained by applying our matching algorithm to the randomized out control group and nonexperimental group. As neither group has received the intervention the difference in earnings between matched individuals from each experimental control group and non experimental sample should be zero. Table 3 presents direct estimates of the bias using stratification matching with semiparametric propensity scores. Notice that with the exception of column 4 of speciication one, the bias is of the order of a few hundred dollars and is less than 15% of the experimental treatment impact in columns 2, 3, 5 and 6 respectively. The inclusion of individuals in the lowest probability quantile has little effect on the bias unlike the treatment effects. This occurs since the majority of individuals from the experimental sample who are assigned to this probability interval were randomly assigned treatment.

The evaluation bias increases by approximately $200 in column 1- 3 when the higher order and interaction terms are omitted from the estimating equation. Column 6 continues to exhibit low bias whereas column 5's bias is also reduced in absolute value. Once again and surprisingly we find a high degree of bias in the Dehejia and Wahba samples. This is striking and contrasts the findings of Smith and Todd (2002) who found that matching algorithms using parametric propensity scores provided low bias only for this subsample. In Appendix table 3 we present interval matching estimates for the bias using propensity scores estimated by a logit for each specification where the lowest probability bin is excluded and included respectively. The bias is significantly lower for this subsample (both columns 3 and 4, Dehejia and Wahba) as compared to the estimates presented in tables 3.

To uncover an explanation as to why the evaluation bias calculated using semiparametric propensity scores exceeded the estimate obtained using parametric propensity scores in column 4 of table 4 we conducted a more detailed examination of how the estimated bias differs across quantiles. Figure 3 presents a graph of the quantile bias effects at each interval for both parametric and semiparametric propensity scores for specification 1. Notice that in almost all quantiles the semiparametric procedure exhibits lower bias.[18] The results in table 3 (and Appendix Table 3) present a number of treated individuals weighted average of these quintile biases and suggest that the lower bias for the Dehejia and Wahba subsample is based in part on having the larger biases across quantiles cancel out.

[18]The pattern is similar for specification 2.

To provide additional guidance for empirical researchers on the performance of propensity score matching algorithms we compare the average absolute bias error of our matching algorithm with a variety of different matching algorithms based on parametric propensity scores; described in Smith and Todd (2002).[19] For each matched outcome we first calculate the absolute bias error

$$Bias\ error = |Y_{1i} - \hat{E}(Y_{0i}|P(X_i), D_i = 0)|$$

where $\hat{E}(Y_{0i}|P(X_i), D_i = 0)$ is calculated by the algorithm under investigation. The average absolute bias error is calculate by dividing the sum of these bias errors by the number of individuals in the treatment group who were successfully matched. We report the average absolute bias error and its standard error in table 4. For interval matching estimators this estimate is simply a weighted average of the absolute value of each quantile bias effect. For the parametric propensity score we match on the estimated propensity score for most of these estimator with the exception of kernel and local linear matching estimators where we match on the odds ratio due to the nature of the sample.[20]

Notice that with one exception, the smallest average absolute bias error is attained using stratification matching with propensity scores calculated by binary regression quantiles. In general, bias error estimates obtained by stratification matching procedures are smaller than the nonparametric and distance metric algorithms. In general when using parametric propensity scores algorithms that use a larger distance produce smooth results; whereas narrow intervals produce larger bias errors on average. In part, this occurs since fewer individuals have matches as the distance shrinks. The results from specification 1 find that Kernel and local linear matching estimator exhibit significantly less bias error than nearest neighbor or caliper matching algorithms. Overall, it appears that using 20 bins produces estimates with the smallest mean squared error. The results suggest

[19] We also compared our procedure to the Abadie and Imbens (2002) matching procedure which determines matched outcomes based on a weighting of distance between covariates and not the propensity score. Due to space constraints we do not report the results but we considered both homoskedastic and heteroskeedastic weighting matrices with one and four individuals matched and the results indicated larger absolute bias error than local linear matching with a bandwidth of 0.01.

[20] Since the data are choice based with unknown sampling weights consistent estimates for the probability of program participation are generally not obtained. Heckman and Todd (1995) demonstrate that matching methods can be applied with the odds ratio to gain consistent estimates when the sample is choice based. Note failure to account for choice based samples should not affect nearest neighbor or stratification point estimates.

that adding the lowest probability quantile to the stratification matching algorithm increases bias up to an average of $500 and $670 per treated participant for specification 1 and 2 respectively.

The increased average size of the bias error from parametric procedures ranges from slightly more than $55.00 to approximately $5200 for specification 1. As a percentage of the estimated treatment impact this range is equivalent 6.2% to 586.9%. For specification 2, stratification matching using parametric propensity scores does exhibit smaller bias error for column 3.[21] Of the remaining columns, the size of the average bias error ranges from $91 to $6250 or 10.3% to 706% of the experimental treatment impact per matched treated individual. While the semiparametric procedure yielded the smallest average absolute bias error in 11 of the 12 columns in Table 4, the number is still large relative to the experimental impact. This casts doubt as to whether all observables were included in the estimation of the propensity score and is a potential cause for concern for empirical researchers interested in using these methods.[22]

Stratification matching with parametric and semiparametric propensity scores yield similar average absolute bias error but wildly different treatment effects (Table 1 versus Appendix Table 2). In general if one excludes the lowest probability bin (0.0-0.05%) the procedures rarely placed individuals within the same interval. This is demonstrated by examining the scarcity of individuals lying on the prime diagonal of table 5 and the large number of individuals residing in the off diagonal elements. This table presents information on the horizontal rows of which bin the semiparametric procedure assigns and the columns provide the bins that the parametric procedure assigns. Notice that ignoring the lowest probability quantile, approximately 30% of all the observations fall in the same probability bin for the two methods. For all 12 subsamples the similarities range between 22%-43%.

4 Conclusions

In situations with nonexperimental data matching methods provide a means to estimate program impacts when the variables determining assignment to treatment are observed and the support of treatment and comparison groups overlap. In this paper, we demonstrate that potential gains

[21] Since this column exhibits a significant number of quantiles with failures in the balancing tests (presented in table 5) further investigation is required to see whether these intervals present significantly larger estimated biases.

[22] Further note that the average bias error for one nearest neighbor matching is extremely large which suggests that there are substantial differences even in the case where matched individuals should.be most alike.

can be achieved with stratification matching using propensity scores estimated by binary regression quantiles, a semiparametric estimation technique that does not make any distributional assumptions on unobservables and allows for general forms of heterogeneity in response to observed covariates. Since binary regression quantiles are robust to general forms of heteroskedasticity, the researcher only has to specify which covariates affect program participation in the estimating equation.

To examine the performance of stratification matching using semiparametric propensity scores calculated via binary regression quantiles we employ data used in several influential studies evaluating the performance of nonexperimental estimators. We find that our technique accurately captures the experimental treatment impact and generally exhibits lower bias than strategies employing parametric propensity scores. A detailed examination of the average absolute bias errors between our procedure and matching algorithms based on parametric propensity scores indicate reductions between 6.2% and 706% of the experimental program impact. These differences are due in part to misspecifaction and as a result fewer than 50% of the same individuals are assigned to the same probability bin with semiparametric and parametric propensity scores.

While previous work using semiparametric estimators did not find large gains relative to parametric procedures they only allowed for very restrictive forms of heterogeneity relative to binary regression quantiles. In conclusion, since the use of the propensity score as a basis for estimating treatment effects is becoming increasingly common in research in a variety of disciplines researchers should test for possible misspecification and if present, should consider the methods described in this paper to improve inference.

References

[1] Abadie, Alberto and Guido Imbens (2002), "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," *mimeo*, University of California at Berkeley.

[2] 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*, Vol. 94, pp. 1053-1062.

[3] Dehejia, Rajeev and Sadek Wahba (2002), "Propensity Score Matching Methods for Non-Experimental Causal Studies," *Review of Economics and Statistics*, Vol. 84, pp. 151-161.

[4] Goffe, William, Gary D. Ferrier and John Rogers (1994), "Global Optimization of Statistical Functions with Simulated Annealing," *Journal of Econometrics*, Vol. 60, pp. 65–101.

[5] Heckman, James J. (2001), "Heterogeneity, Microdata and the Evaluation of Public Policy: Nobel Lecture," J*ournal of Political Economy*, Vol. 109, pp. 673-748.

[6] Heckman, James J., Hidehiko Ichimura, Jeffrey Smith and Petra Todd (1998), "Characterizing Selection Bias using Experimental Data," *Econometrica*, Vol. 66, pp. 1017-1098.

[7] Heckman, James J., Hidehiko Ichimura and Petra Todd (1998),"Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, Vol. 65, pp. 261-294.

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

[9] Heckman, James J. and Petra Todd, (1995), "Adapting Propensity Score Matching and Selection Models to Choice-based Samples," *mimeo*, University of Chicago.

[10] Heckman, James J., and Joseph V. Hotz (1989), "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," J*ournal of the American Statistical Association*, Vol. 84, pp. 862-880..

[11] Horowitz, Joel (1993), "Semiparametric and Nonparametric Estimation of Quantal Response Models," in *Handbook of Statistics*, Vol 11, Maddala, GS and Vinod, HD (eds.), North-Holland.

[12] Horowitz, Joel (1992), "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, Vol. 60, pp. 505-531.

[13] Ichimura, Hidehiko (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics*, Vol. 58, pp. 71-120.

[14] Kim, Jeankyung and David Pollard (1990), "Cube Root Asymptotics," *Annals of Statistics*, Vol. 18, pp. 191–219.

[15] Klein, Roger and Richard Spady (1993), "An Efficient Semiparametric Estimator for Discrete Choice Models," *Econometrica*, Vol. 61, pp. 387-422.

[16] Koenker, Roger and Gilbert Bassett, Jr. (1978), "Regression Quantiles," *Econometrica*, Vol. 46, pp. 33-50.

[17] Kordas, Gregory (2002), "Smoothed Binary Regression Quantiles," *mimeo*, University of Pennsylvania

[18] LaLonde, Robert (1986), "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review*, Vol. 76, pp. 604-620.

[19] Manski, Charles (1975), "Maximum Score Estimation of the Stochastic Utility Model of Choice ", *Journal of Econometrics*, Vol. 3 pp. 205 - 228.

[20] Rosenbaum, Paul. and Donald B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, Vol. 70, pp. 41-55.

[21] Smith, Jeffrey, and Petra Todd (2002), "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators", forthcoming in *Journal of Econometrics*.

[22] Todd, Petra (1999), "Local Linear Approaches to Program Evaluation using a Semiparametric Propensity Score", *mimeo*, University of Pennsylvania.

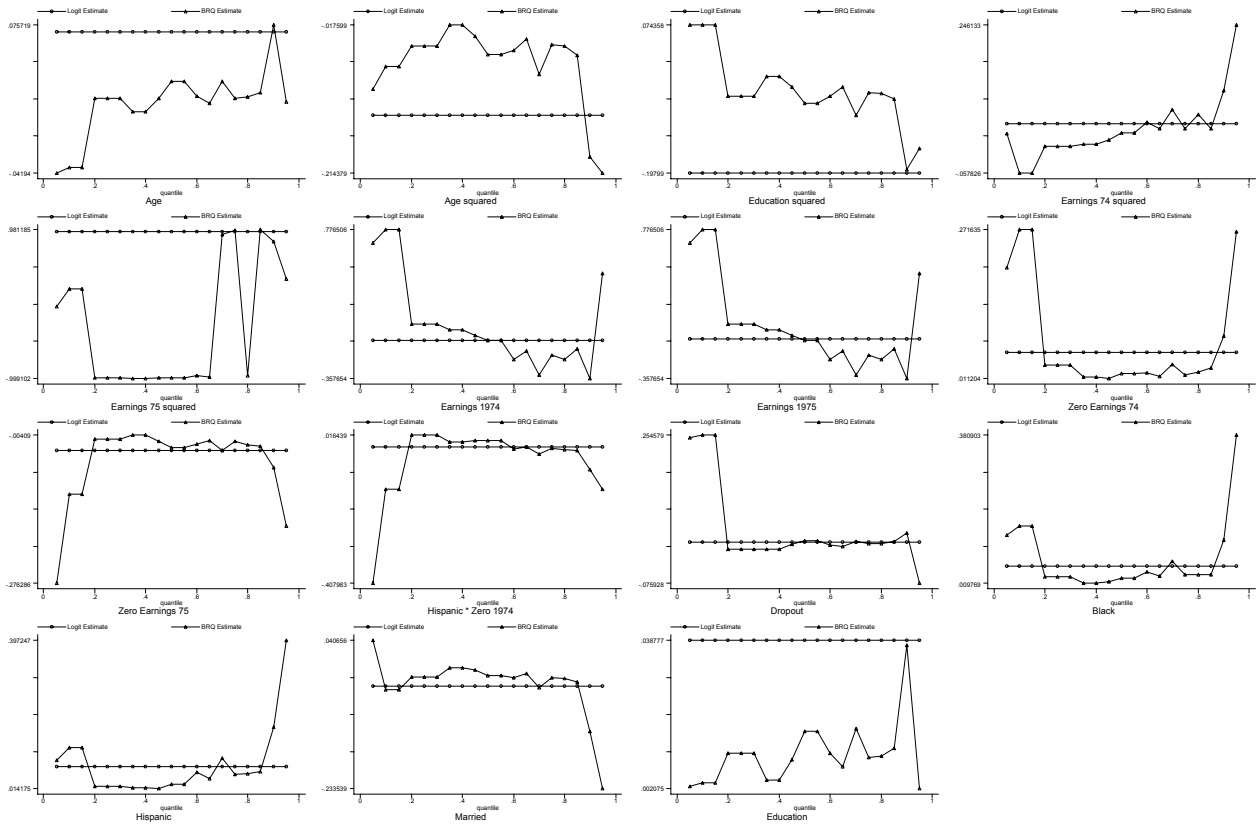

Figure 1: Normalized Binary Regression Quantile and Logit Coefficient Estimates

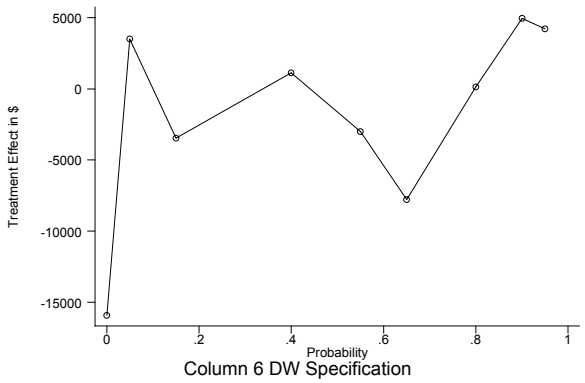

Figure 2: Quantile Treatment Effects Column 6 of Table 4

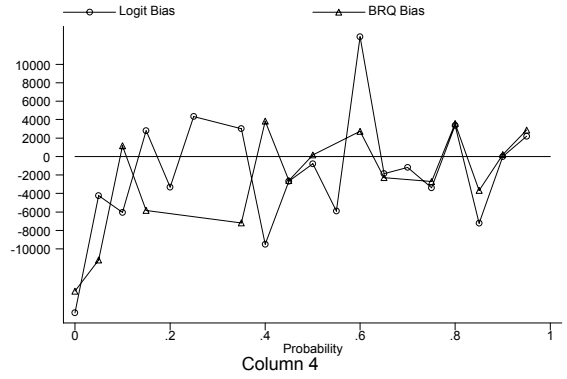

Figure 3: Estimated Bias Parametric and Semiparametric Estimates Column 3 of Table 4

Table 1: Treatment Effects Estimates with Semiparametric Propensity Scores using Stratification Matching

	SPECIFICATION ONE						**SPECIFICATION TWO**					
	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
Experiment Impact	886.32	886.32	1794.34	1794.34	2748.49	2748.49	886.32	886.32	1794.34	1794.34	2748.49	2748.49
20 Bins	85.44 (618.77)	665.55 (900.59)	1925.19 (678.11)	1390.79 (675.37)	2779.06 (1047.0)	2697.48 (1624.8)	96.62 (587.43)	652.04 (874.25)	2039.67 (831.12)	1406.82 (705.95)	2556.02 (1196.2)	2418.79 (1619.5)
10 Bins	114.13 (508.81)	773.91 (886.37)	1391.97 (935.18)	1206.23 (684.84)	2051.61 (1242.6)	2922.78 (1329.7)	-201.02 (584.42)	-179.87 (1234.5)	2145.43 (866.29)	1441.15 (674.73)	2750.53 (1202.6)	2989.07 (1326.6)
5 Bins	221.76 (667.10)	126.80 (945.53)	2400.79 (961.73)	1258.74 (669.76)	2635.53 (1903.0)	963.14 (1933.2)	-34.71 (593.42)	-568.94 (1102.3)	2205.51 (873.32)	1436.19 (668.59)	2640.13 (1226.4)	2852.39 (1143.1)
Including Lowest Probability Bin												
20 Bins	-627.70 (574.16)	476.12 (841.31)	223.27 (833.20)	1108.07 (681.35)	592.24 (1127.3)	1834.56 (1635.9)	-881.44 (535.94)	389.32 (874.25)	684.29 (736.18)	1014.49 (710.99)	1070.34 (1057.5)	1664.49 (1509.2)
10 Bins	-850.68 (649.83)	516.66 (837.60)	253.79 (804.42)	720.96 (669.84)	-116.96 (1092.5)	1769.43 (1337.1)	-1267.4 (550.35)	-561.83 (1284.1)	470.36 (779.58)	825.42 (702.81)	621.01 (1068.0)	1816.20 (1291.6)
5 Bins	-1157.1 (622.46)	-838.17 (854.67)	73.13 (891.22)	488.72 (735.58)	-668.71 (1075.1)	-391.72 (1903.0)	-1447.6 (510.77)	-1455.2 (1036.5)	-206.31 (820.30)	577.36 (700.04)	156.02 (1042.2)	1274.53 (1153.3)

Note: Bootstrapped standard errors in parentheses. 1000 Bootstrap replications.

Table 2: Balancing Test Results

	SPECIFICATION ONE						SPECIFICATION TWO					
Quantile	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
0↔0.05	11	2	14	4	12	4	7	3	10	3	10	3
0.05↔0.10	0	0	0	0	1	0	0	0	3	0	0	0
0.10↔0.15	0	0	0	0	0	0	0	0	0	1	0	0
0.15↔0.20	0	0	1	0	0	2	0	0	3	0	0	0
0.20↔0.25	0	0	0	0	0	0	0	2	0	0	0	0
0.25↔0.30	0	2	0	0	0	0	0	0	0	0	0	0
0.30↔0.35	0	2	0	0	0	0	2	0	2	0	0	0
0.35↔0.40	3	1	0	0	0	0	0	0	0	0	1	0
0.40↔0.45	0	0	0	0	0	1	0	0	0	0	3	0
0.45↔0.50	0	0	1	0	0	0	0	0	2	0	0	0
0.50↔0.55	1	0	0	1	0	0	0	0	0	0	0	0
0.55↔0.60	0	0	0	0	0	0	1	0	6	0	0	0
0.60↔0.65	0	0	0	0	0	0	0	0	0	0	0	0
0.65↔0.70	0	0	0	0	0	0	4	0	0	0	2	0
0.70↔0.75	0	1	0	0	0	0	2	1	1	0	2	0
0.75↔0.80	0	2	0	0	0	0	0	0	0	0	0	0
0.80↔0.85	0	0	0	0	0	1	0	0	0	0	0	0
0.85↔0.90	0	0	0	0	0	0	1	0	1	0	0	0
0.90↔0.95	0	0	0	0	0	0	0	0	0	0	0	0
0.95↔1.00	0	0	0	0	0	0	0	0	0	0	0	0

Note: Number of unbalanced covariates at the 5% level reported.

Table 3: Evaluation Bias Estimates with Semiparametric Propensity Scores using Stratification Matching

	SPECIFICATION ONE						**SPECIFICATION TWO**					
	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
Experiment Impact	886.32	886.32	1794.34	1794.34	2748.49	2748.49	886.32	886.32	1794.34	1794.34	2748.49	2748.49
20 Bins	-336.54 (480.81)	-139.93 (810.83)	161.28 (741.64)	1673.97 (746.29)	-589.35 (461.54)	515.52 (641.74)	436.09 (640.18)	1916.99 (704.56)	-639.12 (680.63)	153.73 (972.87)	2556.02 (1196.2)	2418.79 (1619.5)
10 Bins	-335.03 (594.49)	241.27 (853.37)	384.81 (771.63)	1465.84 (812.25)	-908.48 (447.11)	-903.74 (23.12)	642.51 (640.09)	1648.89 (755.05)	-536.74 (716.03)	674.81 (995.20)	2750.53 (1202.6)	2989.07 (1326.6)
5 Bins	-405.13 (556.55)	-735.40 (861.99)	1138.15 (781.37)	1666.38 (649.94)	-983.01 (413.25)	-1195.69 (1124.9)	732.39 (652.39)	1586.93 (590.43)	39.20 (805.45)	555.40 (777.46)	2640.13 (1226.4)	2852.39 (1143.1)
Including Lowest Probability Bin												
20 Bins	-319.93 (482.99)	-26.65 (872.13)	271.95 (701.25)	1486.15 (767.14)	340.51 (817.92)	-214.48 (1000.8)	-1387.62 (458.99)	292.88 (666.71)	-17.52 (589.98)	1630.89 (716.22)	-1442.96 (641.02)	-614.39 (1008.24)
10 Bins	-1215.47 (560.22)	-91.68 (833.26)	-573.74 (624.82)	863.00 (831.63)	-1371.57 (713.60)	376.71 (911.09)	-1648.51 (430.82)	-1150.33 (1153.4)	-418.66 (635.01)	896.64 (719.66)	-1727.59 (690.56)	-506.14 (981.67)
5 Bins	-1568.71 (556.55)	-1359.71 (845.89)	-650.74 (678.53)	744.53 (698.07)	-1843.34 (669.55)	-829.00 (976.36)	-2003.59 (423.29)	-1988.88 (1101.5)	-948.47 (621.08)	729.50 (636.01)	-2077.87 (725.08)	-914.14 (866.48)

Note: Bootstrapped standard errors in parentheses. 1000 Bootstrap replications.

Table 4: Average Absolute Bias Error

Matching Algoritm	SPECIFICATION ONE						SPECIFICATION TWO					
	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
Nearest Neighbor 1 W/O Common Support	6531.57 (5746.5)	6515.78 (8011.1)	5043.09 (5375.8)	5231.05 (5190.2)	5956.38 (6444.2)	5834.49 (6005.6)	6495.40 (5719.1)	6065.43 (6199.8)	6073.86 (5848.0)	5027.43 (6518.4)	6200.70 (6322.6)	5657.73 (5847.6)
Nearest Neighbor 10 W/O Common Support	6540.42 (5840.8)	6515.78 (8011.1)	5141.34 (5426.0)	5231.05 (5190.2)	6100.74 (6488.0)	5834.49 (6005.6)	6466.16 (5740.2)	6065.43 (6199.8)	5953.66 (5963.3)	5027.43 (6518.4)	6298.87 (6309.0)	5657.73 (5847.6)
Nearest Neighbor 1 W. Common Support	6544.23 (5779.5)	6515.78 (8011.1)	4996.16 (5450.1)	5231.05 (5190.2)	5983.31 (6502.3)	5834.49 (6005.6)	6561.55 (5785.3)	6065.43 (6199.8)	5867.80 (5834.5)	5027.43 (6518.4)	6500.32 (6476.5)	5657.73 (5847.6)
Nearest Neighbor 10 W. Common Support	6597.81 (5820.5)	6515.78 (8011.1)	5109.14 (5587.4)	5231.05 (5190.2)	6077.74 (6564.0)	5834.49 (6005.6)	6551.98 (5769.2)	6065.43 (6199.8)	5989.54 (5978.8)	5027.43 (6518.4)	6077.74 (6564.0)	5657.73 (5847.6)
Kernel (Bandwidth 0.04)	5155.23 (3738.3)	5146.12 (4086.8)	4536.32 (3878.1)	4746.40 (4204.6)	5329.49 (4378.1)	5051.71 (4588.3)	5711.01 (4405.8)	7105.71 (5955.1)	4819.80 (4509.0)	6596.85 (5487.0)	5415.97 (4940.4)	6055.22 (5164.6)
Kernel (Bandwidth 0.01)	4987.96 (3839.0)	5216.04 (4335.5)	4307.99 (3962.5)	4788.36 (4339.8)	5034.23 (4671.2)	5300.18 (4720.5)	5664.64 (4572.7)	6553.06 (8027.5)	4841.42 (4720.5)	6012.27 (5287.2)	5484.68 (5198.4)	6209.80 (5337.7)
Local Linear (Bandwidth 0.04)	4675.54 (3345.0)	4845.76 (3749.6)	4134.49 (3524.6)	4298.21 (3636.9)	4670.90 (3857.7)	4871.93 (4038.7)	4754.06 (3459.7)	5051.76 (3891.8)	4254.59 (3546.9)	4427.15 (3908.4)	4831.42 (3790.4)	4900.41 (4223.0)
Local Linear (Bandwidth 0.01)	4705.60 (3379.1)	4995.25 (4105.4)	4145.24 (3498.1)	4353.05 (3895.5)	4743.96 (3883.0)	4813.54 (4356.2)	4680.02 (3464.6)	5059.29 (4100.5)	4150.35 (3613.5)	4460.52 (3917.0)	4743.96 (3883.0)	5046.30 (4247.1)
Caliper (0.01)	6505.52 (5758.3)	6600.18 (8138.6)	4978.08 (5365.1)	4945.54 (5098.1)	5928.61 (6492.5)	5529.58 (5879.6)	6791.76 (5844.4)	6340.82 (6428.5)	6128.93 (5760.1)	5168.24 (6612.4)	6357.82 (6499.1)	5769.70 (6022.4)
Caliper (0.001)	6860.53 (5981.8)	6175.00 (6132.1)	4942.18 (5057.3)	4434.86 (4838.2)	6841.91 (6621.5)	5976.01 (6788.6)	6619.27 (5791.0)	7073.58 (7189.0)	6448.08 (6387.0)	5683.33 (8489.1)	6380.06 (6218.1)	5092.21 (5162.5)
Caliper (0.0001)	6079.16 (5766.4)	6166.09 (6749.7)	5268.85 (5375.1)	4349.57 (5403.5)	5904.92 (6546.1)	6359.16 (8353.7)	6812.13 (6073.7)	8286.05 (10196.)	5788.78 (5185.5)	8622.76 (14083.)	7946.84 (6395.2)	4429.00 (5506.3)
Stratification 20 Bins Logit	2131.75 (1879.6)	2272.02 (1884.1)	1798.77 (1825.0)	3206.00 (2961.3)	2953.57 (2382.5)	3519.43 (3140.3)	2360.50 (1879.6)	2356.05 (2890.4)	1798.77 (1825.0)	3206.00 (2961.3)	2953.57 (2382.5)	3519.43 (3140.3)
Stratification 20 Bins BRQ	2054.54 (2192.5)	2210.44 (2312.4)	1741.31 (1833.8)	2864.02 (1997.1)	2929.14 (2032.3)	2234.41 (2063.6)	2451.74 (2633.4)	2585.78 (1968.3)	2372.92 (1611.9)	2596.23 (2174.3)	2511.21 (2099.7)	3133.87 (3444.3)
Stratification 10 Bins BRQ	1768.95 (2364.5)	2113.56 (2833.4)	2654.56 (2412.1)	2445.20 (2844.3)	2716.92 (2137.9)	2713.09 (2757.8)	2262.78 (2626.1)	1356.82 (2577.7)	2834.35 (2043.8)	2414.72 (2954.8)	2675.21 (2954.8)	3536.80 (3624.5)
Stratification 20 Bins BRQ No Zeros	1669.03 (1116.5)	2064.17 (2100.5)	1352.69 (2306.5)	2726.98 (1549.5)	2412.47 (1468.4)	2612.59 (1917.5)	1770.34 (1878.5)	2028.09 (1557.5)	1981.78 (1210.4)	2376.78 (1281.4)	1874.07 (1437.1)	2480.27 (1578.6)
Stratification 20 Bins Logit No Zeros	1971.97 (1218.8)	2119.36 (2352.0)	1421.64 (1164.9)	2991.63 (2430.7)	2458.90 (2052.7)	3137.16 (2219.0)	1971.97 (1218.8)	2119.36 (2352.0)	1421.64 (1164.9)	2991.63 (2430.7)	2458.90 (2052.7)	3137.16 (2219.0)

Note: Standard deviation in parentheses

Table 5: Number of Individuals Assigned to a Bin by Parametric and Semiparametric Estimates Using PSID and Early Random Assignment Experimental Sample via Specification 2

Logit Bins -> BRQ Bins ↓	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Total
[0-0.05%)	2178	58	17	5	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2266
[.05-0.1%)	42	42	23	21	19	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	153
[0.1-0.15%)	0	1	2	2	4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	10
[0.15-0.2%)	0	0	0	5	6	4	3	1	1	0	0	0	0	0	0	0	0	0	0	0	20
[0.2-0.25%)	1	1	0	1	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	5
[0.25-0.3%)	0	0	0	0	2	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
[0.3-0.35%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[0.35-0.4%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
[0.4-0.45%)	0	1	2	0	2	1	4	0	4	5	2	2	6	0	0	0	0	0	0	0	27
[0.45-0.5%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[0.5-0.55%)	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	2
[0.55-0.6%)	0	0	1	2	1	1	5	2	1	2	4	5	7	4	7	2	1	1	0	0	45
[0.6-0.65%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[0.65-0.7%)	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	2	0	1	0	0	6
[0.7-0.75%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[0.75-0.8%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[0.8-0.85%)	0	0	0	0	0	0	0	0	2	3	3	5	10	7	10	2	5	10	3	1	61
[0.85-0.9%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[0.9-0.95%)	0	0	0	0	0	0	0	0	0	0	0	1	2	5	5	3	2	5	12	32	62
[0.95-1.0%)	0	0	0	0	0	0	1	0	2	0	1	0	0	1	1	7	20	8	14	24	79
Total	2221	103	45	36	42	12	13	6	12	10	11	18	20	14	23	16	28	24	29	57	2740

Appendix Table 1: Summary Statistics

Sample	LaLonde Treated	LaLonde Controls	DW Treated	DW Controls	Early Assignment Treated	Early Assignment Control	CPS	PSID
Sample Size	297	425	185	260	108	142	15992	2490
Age	24.626 (6.686)	24.447 (6.590)	25.816 (7.155)	25.054 (7.058)	25.370 (6.251)	26.014 (7.108)	33.225 (11.045)	34.851 (10.441)
Years of Education	10.380 1.818)	10.188 (1.619)	10.346 (2.011)	10.088 (1.614)	10.491 (1.643)	10.275 (1.572)	12.028 (2.871)	12.117 (3.082)
Hispanic	0.094	0.113	0.059	0.108	0.074	0.113	0.072	0.032
Black	0.801	0.80	0.843	0.827	0.824	0.817	0.074	0.251
Married	0.168	0.158	0.189	0.154	0.204	0.190	0.712	0.866
Dropout	0.731	0.814	0.708	0.835	0.713	0.803	0.296	0.305
Zero Earnings in 1974	0.441	.461	0.708	0.75	0.50	0.542	0.120	0.086
Zero Earnings in 1975	0.374	0.419	0.60	0.685	0.324	0.472	0.109	0.100
Real Earnings in 1974	3571.00 (5773.13)	3672.49 (6521.53)	2095.57 (4886.62)	2107.03 (5687.91)	3589.64 (5970.74)	3857.94 (7254.27)	14016.8 (9569.80)	19428.8 (13406.9)
Real Earnings in 1975	3066.10 (4874.89)	3026.68 (5201.25)	1532.06 (3219.25)	1266.91 (3102.98)	2596.03 (3871.68)	2276.96 (3919.28)	13650.8 (9270.40)	19063.3 (13596.9)
Real Earnings in 1978	5976.35 (6923.80)	5090.05 (5718.09)	6349.14 (7867.40)	4554.80 (5483.84)	7357.41 (9027.18)	4608.92 (6031.96)	14846.66 (9647.39)	21553.9 (15555.4)

Note: Standard Deviation in Parentheses

Appendix Table 2: Treatment Effect Estimates with Parametric Propensity Scores using Stratification Matching

	SPECIFICATION ONE						SPECIFICATION TWO					
	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
Experiment Impact	886.32	886.32	1794.34	1794.34	2748.49	2748.49	886.32	886.32	1794.34	1794.34	2748.49	2748.49
20 Bins Excluding 0-0.05	-277.54 (603.86)	-401.81 (788.77)	1327.11 (860.30)	1631.55 (1157.9)	1386.91 (1108.3)	2242.87 (1265.9)	-373.80 (556.89)	18.44 (706.37)	1361.09 (794.66)	1940.58 (911.70)	2437.87 (1063.0)	1380.82 (1303.5)
20 Bins no exclusion	-1067.65 (596.06)	-611.86 (759.50)	14.76 (816.72)	1151.39 (1094.7)	211.18 (959.90)	1441.67 (1351.0)	-1312.4 (544.38)	-301.82 (707.88)	171.88 (776.22)	1354.01 (947.75)	591.07 (1060.0)	465.81 (1237.7)
20 Bins DW exclusion	-819.79 (582.53)	-632.00 (800.09)	1096.57 (800.63)	1394.41 (1060.4)	1255.21 (1032.5)	1297.61 (1125.5)	-803.78 (489.33)	-268.06 (688.45)	902.33 (731.83)	1513.39 (920.17)	1634.18 (938.87)	694.29 (1245.2)

Note: Bootstrapped standard errors in parentheses. 1000 Bootstrap replications. DW exclusion drops all individuals in the treatment group with estimated propensity scores above the maximum propensity score in the control group and drops all control individuals whose estimated propensity score is less than the minimum propensity score of the treatment group.

Appendix Table 3: Evaluation Bias Estimates with Parametric Propensity Scores using Stratification Matching

	SPECIFICATION ONE						SPECIFICATION TWO					
	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
Experiment Impact	886.32	886.32	1794.34	1794.34	2748.49	2748.49	886.32	886.32	1794.34	1794.34	2748.49	2748.49
20 Bins	-1318.41 (530.93)	-1011.45 (696.50)	62.14 (673.98)	726.59 (779.96)	-1216.63 (769.63)	-302.93 (911.30)	-1303.93 (427.43)	-685.21 (594.36)	-399.41 (547.29)	679.02 (760.47)	-1275.2 (634.42)	-1862.5 (1108.8)
10 Bins	-1148.28 (539.09)	-1070.33 (783.31)	-124.51 (675.83)	84.97 (926.07)	-1135.18 (849.69)	-825.08 (952.01)	-1248.07 (437.63)	-985.08 (634.63)	-229.11 (566.49)	248.66 (1056.9)	-1129.4 (641.74)	-1156.1 (876.66)
Including Lowest Probability Bin												
20 Bins	-1749.99 (517.42)	-1269.00 (695.99)	-411.11 (644.94)	455.02 (770.67)	-1851.27 (737.83)	-765.03 (947.28)	-1890.1 (399.59)	-1091.06 (600.09)	-912.84 (528.85)	270.58 (805.11)	-1939.3 (630.75)	-2286.0 (1097.8)
10 Bins	-2004.45 (491.63)	-1593.87 (741.56)	-977.55 (643.09)	-505.34 (911.10)	-2282.38 (762.06)	-1471.57 (863.60)	-2051.4 (398.61)	-1421.89 (650.42)	-1275.7 (511.59)	-375.59 (1057.2)	-2526.2 (627.09)	-2186.9 (934.91)

Note: Bootstrapped standard errors in parentheses. 1000 Bootstrap replications.