EVALUATING THE PERFORMANCE OF NON-EXPERIMENTAL ESTIMATORS: EVIDENCE FROM A RANDOMIZED UI PROGRAM

Jose Galdo*
Center for Policy Research and Department of Economics
Maxwell School of Citizenship and Public Affairs
426 Eggers Hall - Syracuse University

Abstract

One of the lessons of the treatment effects literature is the lack of consensus about the ability of statistical and econometric methods to replicate experimental estimates. In this paper, we provide new evidence using an unusual unemployment insurance experiment that allows the identification of discontinuities in the assignment mechanism. In particular, we use a set of regression functions and matching estimators based on kernel methods with mixed categorical and continuous data. A crucial issue with the kernel approach is the choice of the smoothing parameters. We develop a leave-one-out cross-validation algorithm that minimizes the mean square error of the average treatment effect on the treated weighting each comparison unit according to their distribution of covariates in the support region. Two main findings emerge. First, local constant and nearest-neighbor matching on kernel-based propensity score with mixed categorical and continuous data produces a closer approximation to the experimental estimates than traditional parametric propensity score models do. Second, the regression-discontinuity design emerges as a promising method for solving the evaluation problem. When restricted to sample observations in the neighborhood of the discontinuity points, the estimates are close approximation to the experimental estimates and are robust across different subsamples and estimators.

KEYWORDS: Treatment Effects, Kernel Regressions, Kernel Distributions, Cross-Validation, Regression-Discontinuity Design.

---

* Ph.D. Candidate. Preliminary version. The research presented herein could not have been carried out without the advice and encouragement of Dan Black, Jeff Racine, Jeff Smith and Ana Dammert. Dan Black generously provided the data used in this study. All errors are my own, comments are welcome to jcgaldo@maxwell.syr.edu.

Introduction

Since LaLonde's (1986) influential paper, several studies have tried to address the effectiveness of non-experimental methods to replicate experimental evaluation estimates. The main advantage of having access to experimental data is that it solves the evaluation problem after balancing the distribution of observable and unobservable variables in the treatment and control group. Under the assumptions that randomization does not alter both the pool of participants (randomization bias) and their behavior (Hawthorne effect), and that close substitutes for the experimental treatment are unavailable (substitution bias), it follows that the evaluator can estimate the impact of the treatment on the treated through simple difference in outcome means. Therefore, experimentally determined impacts allow a unique opportunity to calibrate non-experimental estimators, visualize strategies for choosing competing alternative non-experimental estimators, and evaluate the underlying identification assumptions justifying the estimators.

LaLonde's critique of the use of non-experimental techniques in the evaluation of social programs relies on the premise that successful econometric models intend to reproduce the experimental estimates. He use the experimental estimates from the National Support Work Demonstration Program (NSW) as the benchmark and then, setting the controls aside, wed the treated units from the experiment to different sets of non-experimental samples – called comparison groups - extracted from national public surveys. He concludes that non-experimental estimators would not have yielded accurate estimates of the impact of the NSW program. Since then, the search for new answers to the old problem of estimating unbiased treatment effects in non-experimental samples has produced an impressive development of statistical and econometric methods (see e.g. Heckman, LaLonde and Smith, 2000 and Imbens 2003). The power of alignment tests (Heckman and Hotz, 1989), semiparametric kernel regressions (Heckman, Ichimura, Smith and Todd, 1998), simple matching estimators (Dahejia and Wahba, 1998), bias-adjusted matching (Abadie and Imbens, 2002) and nonparametric series estimators (Hahn 1998, Hirano, Imbens and Ridder, 2003) has yield a rich set of new econometric tools long away from simple parametric linear methods.

One characteristic that may summarize this literature is the lack of consensus in the ability of statistical and econometric methods in replicating experimentally determined estimates.[1] A firm result, however, is that non-experimental estimators can not prevail over bad quality data. If the evaluation aim is to compare experimental and non-experimental samples the starting point is compare comparable units which supposes having geographically-aligned samples (Heckman et al. 1998), under the same definition of outcome and pre-treatment variables and survey instruments (Smith, 1995).

In this paper, we provide new evidence about the ability of several new econometric estimators in replicating average treatment effects from an unusual UI experiment, the Kentucky Working Profiling Reemployment Services (KWPRS), that provides low intensity reemployment services to treated claimants. The peculiarity of this experiment is the profiling mechanism used to select the treated and control units. The randomization occurs only to satisfy capacity constrains and at the margin. For each local office and each week, claimants starting new spells are ranked by their profiling scores and those who have the highest scores receive automatically the reemployment services until the exhaustion of the budget for that particular office and week. Only claimants with marginal profiling scores are assigned to the experimental treatment or control group when they exceed the number of available slots. Importantly, as a consequence of this "tie-breaking experiment" a non-experimental comparison group is automatically form by those who were not assigned into the experimental sample because of their lower profiling scores than that for the marginal score group in each week and local office. The underlying idea of this experimental program is to reduce the duration of unemployment spells for those with higher probabilities of exhausting the 26 weeks of UI benefits and, therefore, to reduce the costs of the UI system. Black, Smith, Berger and Noel (2003) find experimental estimates of –2.24 for weeks receiving UI benefits, $-143 for amount of UI benefits received and $1,054 for annual earnings.

We use a set of semiparametric and nonparametric regression functions and matching estimators in order to replicate these experimental estimates. In particular, the regression approach estimates the functions $E(Y_1 \mid X, T=1)$ and $E(Y_0 \mid X, T=0)$ using

[1] Dahejia and Wahba (1999, 2002), and Smith and Todd (2003) vividly illustrate two different answers to the same problem.

local linear kernel methods with mixed categorical and continuous data. Traditional nonparametric kernel regression limits their application to continuous variables only. The presence of discrete variables is handling by frequency estimation method that splits the sample into cells, which implies a loss in finite-sample efficiency. Racine and Li (2003) show that kernel regression estimation with mixed data has a rate of convergence that depends only on the number of continuous variables involved, it does not split the sample into cells and nicely handles interactions among the discrete and continuous variables. One promising application of the proposed regression functions is the estimation of regression adjusted matching that is carry out without imposing the stronger assumptions of the partially linear model (Heckman et al.,1998) or the linear regression functions (Rubin, 1973b).

It is documented the usefulness of nonparametric propensity scores for estimating average treatment effects in a regression framework. Hahn (1998), and Hirano, Imbens and Ridder (2003) show that weighting by the inverse of a nonparametric – series logit - propensity score leads to efficient estimates of average treatment effects. We estimate the propensity score using a kernel-based probability density function with mixed categorical and continuous data. In addition to the natural advantages that offer the interaction of mixed data, this nonparametric propensity score can be incorporated smoothly in the treatment effect estimators proposed by Hahn (1998) and Hirano et al. (2003) without need for using series logit propensity score estimators.

For the kernel-based regression and kernel-base matching estimators, a crucial issue is the choice of the smoothing parameter. The treatment effects literature, however, has been largely silent concerning the optimal choice of the soothing parameters. The standard approach is to use a fixed smoothing parameter (e.g. Heckman et al. 1998) with variations in the selected parameter to determine the sensitivity of the treatment estimates (Smith and Todd, 2003). In general, the kernel-based literature has developed some methods to find the optimal smoothing parameters using data-driven methods. Hall, Racine and Li (2002), Racine and Li (2002) propose least square cross-validation to obtain the smoothing parameters of regression and distributional functions with mixed categorical and continuous data. It has the ability to automatically remove irrelevant regressors by smoothing out such variables, a property not shared by other bandwidth

selection rules (e.g., plug-in). We use this approach when estimating the regression-based estimators and the nonparametric propensity score. In the case of kernel-based matching (local constant and local linear matching), the problem of finding optimal smoothing parameters is more complicated because the mean-squared-error criteria weights evenly all units even those with distribution of covariates in regions that are not important in the estimation of the average treatment effects. We develop an algorithm base on cross-validation methods and mean-squared-error criteria weighting each unit by its role in the determination of the average treatment effects.

The unique design of the KWPRS allows us to identify discontinuities in the assignment mechanism that makes it similar to a quasi-experimental design originally introduced by Thistlethwaite and Campbell (1960) and named "tie-breaking" experiment. We exploit the discontinuity assignment to the experimental and non-experimental samples using a regression-discontinuity approach in order to solve the evaluation problem. Previous evidence shows its usefulness in identifying causal relationships in self-selected samples (e.g. Angrist and Lavy, 1996; van der Klauuw 2001), although there is no empirical evidence about its effectiveness in replicating experimental estimates.

The paper proceeds as follows: Section1 explains some key characteristics of the KWPRS program and gives a description of the experimental and non-experimental data. Section 2 reviews the methodological aspects of the program evaluation. Section 3 describes the set of parametric, semiparametric, and nonparametric estimators that intend to replicate the experimental estimates. Section 4 shows the empirical results, and Section 5 concludes.

1. The Kentucky UI Experiment: Working Profiling and Reemployment Services

In November 1993, President Clinton signed into law the Unemployment Compensation Amendments of 1993 that offered a variety of low-intensity reemployment services to UI claimants that were identified through a profiling statistical model as potential exhaustees of UI benefits. In June of 1994, the Commonwealth of Kentucky was selected as a prototype state for implementing the Working Profiling and Reemployment Services

(KWPRS).[2] The underlying idea of this program is to reduce the duration of unemployment spells for those with higher probabilities of exhausting the 26 weeks of UI benefits and, thereby, to reduce the costs of the UI system.

A great concern with the UI benefits is the potential distortionary behavior of the UI claimants that may extend the unemployment spell beyond what it would be in the absence of UI benefits, either by subsidizing additional job search or by subsidizing the consumption of leisure.[3] In order to deter these undesirable effects some policies, such as the UI Bonus experiments offer cash rewards to UI beneficiaries who find a job rapidly, and, thereby reduce the incentives for excess benefit receipt without punishing workers for whom a longer search is optimal.[4] Likewise, alternative UI experiments used "sticks" instead of "carrots" by enforcing the job search requirements together with job search assistance. The KWPRS program combines aspects of both types of UI reforms. One week after the UI claimants receive the first UI check, the experimental treatment group received a notification letter that informed them about the mandatory reemployment services.[5] Conditional on receiving these services they collect the next checks until the point of benefit exhaustion. Therefore, the treatment is not the services themselves but the notification of mandatory participation in such services that may cause changes in their behavior. Depending on each individual, this UI program can be a "carrot" or a "stick". For those who see this program as an opportunity to increase their human capital skills, this program increase the value of being unemployed prior to the start of the services as they anticipated higher wage offers. For those who see this program a "leisure tax", the program lowers the value of remains unemployed before and during the period of service receipt. The net effect of this program depends on the signs and magnitudes of these two effects.

---

[2] The Kentucky profiling statistical model was estimated thorough a double-limit Tobit model using 140 covariates, including variables representing characteristics of the local and state economy as well as the workers characteristics such as past earnings, participation in welfare programs, past UI recipiency, past job characteristics, education, etc. It is against the law to profile based on gender, veteran status, ethnicity, and age.

[3] See Mortensen (1970) and McCall (1970) for earlier works in job search models and UI; Ashenfelter (1978a) and Moffitt and Nicholson (1982) for labor supply models and UI.

[4] See Meyer (1995) for a comprehensive survey, Woodbury and Spiegelman (1987) for a detailed analysis of Illinois Bonus experiment.

[5] Employment counseling, job search workshops, labor market information, job referral and placement, relocation assistance, education and training opportunities of modest duration.

The Experimental Sample

The peculiarity of the KWPRS program is the random assignment that occurs at the margin. The statistical profiling model assigns to each UI claimant a continuous score based on the probability of exhausting UI benefits. These continuous estimates are collapsed into a discrete profiling score ranging from 1 to 20 such that potential participants predicted by the profiling system to exhaust between 95 and 100 percent of their unemployment benefits received a score of 20; potential participants predicted to exhaust between 90 and 95 percent of their unemployment benefits receive 19 and so on. For each local office and each week, claimants starting new spells are ranked by their scores, those who have the highest scores receive mandatory reemployment services automatically until the budget for that particular office, and week is exhausted. Only the claimants with marginal profiling scores are assigned to the experimental treatment and control groups when they exceed the number of available slots in a given week and local office. Because of this "tie-breaking experiment" (Campbell, 1969), 1,236 and 745 claimants are in the treatment and control groups. Black, Smith, Berger and Noel (2003) call these sets of claimants "profiling tie groups" (PTG's).

The Non-Experimental Sample

The UI claimants with profiling scores below the (week/local office specific) marginal scores are not profiled into the KWPRS treatment and, therefore, they constitute the non-experimental sample. From June 1994 to October 1996, 8629 claimants fall in this category. It is important to highlight that the experimental and non-experimental samples are composed of individuals living in the same local labor market at the same time and who provided all the socio-economic, demographic, and labor information following the same battery of pre-program and follow-up instruments. Furthermore, we only use administrative data that minimizes the risk of randomization bias and attrition bias. This high quality data is very appropriate for econometric methods that intend to correct selection on observables because the mandatory assignment into treatment is base on a score derived from observable characteristics for each individual.

In order to determine if the observations from the experimental sample were drawn from the same population, we perform a kernel-based nonparametric test for equality of distributions. It is more stringent than simple tests for equality of means, it

can be applied to both categorical and continuous data, and it allows obtain p-values using bootstrap methods (Li and Racine 2003b). The test is based on an empirical test statistic (I^*) the integrated squared density differences between two distribution functions $I = \int [f(x) - g(x)]^2 dx$. It uses the empirical distribution of B bootstrap statistics $\left\{I_i^*\right\}_{i=1}^{B}$ to approximate the null distribution of I.

Table 1 presents the first two moments of the distributions of some pre-treatment covariates together with the p-value statistics for differences in the distributions for both the experimental and non-experimental samples. It is clear from column (4) that the null hypothesis of equality of distributions for the experimental sample is not rejected, which suggests that the treatment and control groups were drawn from the same population. Unsurprisingly, column (5) shows that for the non-experimental treatment and comparison sample the nonparametric test for equality of distributions rejects the null hypothesis for most of the covariates. We reinforce this result by observing the standardized differences between the covariates in column (6). It reveals systematic differences between treatment and comparison units. The huge difference in annual and quarterly earnings before the program between the experimental and non-experimental sample is remarkable.

2) The Evaluation Problem

The main goal of evaluating social programs is find consistent and unbiased estimates of the program impacts on the treatment group. Yet in a world where it is extremely difficult implement experiments within social programs (see Heckman and Smith, 1995) the construction of adequate counterfactuals is the Gordian knot of the evaluation problem. The problem arises because the evaluators observed mutually exclusive states for the individuals: treatment (T=1, associate to outcome Y_1) or non-treatment (T=0, associate to outcome Y_0), but not both states at the same time. Therefore, estimating the outcomes that would have been observed for participants in the program had they not participated, Y_0, is the evaluator's task. Denoting Δ_i as the individual gain of moving from state 0 to state 1, we cannot identify for anyone the impact of participating in the program $\Delta_i = Y_{1i} - Y_{0i}$ because of the missing data problem.

We can only identify mean or distributional gains under some exogeneity assumptions. In this paper, we focus in the mean impact of treatment on the treated (subsequently, Δ_{TT}) that estimates the average impact among those participating in the program:

$$\Delta_{TT} = E(Y_1 - Y_0 \mid X, T=1) = E(Y_1 \mid X, T=1) - E(Y_0 \mid X, T=1). \qquad (1)$$

While $E(Y_1 \mid X, T=1)$ may be estimated from the observed treated sample, the right-hand side of the equation (1) contains the missing data $E(Y_0 \mid T=1, X)$. If we know for certain the outcome that would have been observed for participants in the program had they not participated, $Y_0 \mid T=1$, we have solved the evaluation problem. In this context, using non-participants outcomes, $Y_0 \mid T=0$, to approximate the counterfactual missing participant's outcomes originates the mean selection bias because those who participated in the program may have different levels of Y_0 even in the absence of receiving any program services,

$$SB = E(Y_0 \mid D=1, X) - E(Y_0 \mid D=0, X). \qquad (2)$$

Having access to an experimental control group solves the problem of mean selection bias, under certain behavioral and statistical assumptions.[6] Denote $T^*=1$ for individuals who would participate in the random assignment, $T^*=0$ for everyone else. Also, define $r=1$ for randomization into the treatment group and $r=0$ for randomization into the control group. The crucial assumption for identifying the mean impact of treatment on the treated is,

$$E(Y_1 - Y_0 \mid T=1, X) = E(Y_1 \mid T^*=1, r=1, X) - E(Y_0 \mid T^*=1, r=0, X). \qquad (3)$$

If this condition holds, the mean selection bias is equal to zero because the control group outcomes are unbiased estimates of the outcomes that would have been observed for participants in the program had they not participated. Under this condition and assuming that the outcome functions are represented by a general functional form:

$$\begin{aligned} Y_{1i} &= j_1(X_i) + x_{1i}. \\ Y_{0i} &= j_0(X_i) + x_{0i}. \end{aligned} \qquad (4)$$

[6] Under the assumptions that randomization does not alter the pool of participants (randomization bias) , or their behavior (Hawthorne effect), and that close substitutes for the experimental treatment are unavailable (substitution bias); see Heckman and Smith (1995).

where $E(x_{1it}) = E(x_{0it}) = 0$, the observed outcome equation can be represented by:

$$Y_i = D_i Y_i + (1 - D_i) Y_{0i} = j_0(X_i) + T_i d + e_i \ . \qquad (5)$$

where $d = j_1(X_i) - j_0(X_i) + x_{1i} - x_{0i}$ is the treatment effect and $e_i = (1-T)x_{0i} + Tx_{1i}$ the error term.[7] The random assignment guarantees that $f(X, x \mid T = 1) = f(X, x \mid T = 0)$ from which two important results emerge: $E(j_j(X_i) \mid T = 1) = E(j_j(X_i) \mid T = 0)$ and $E(x_{ji} \mid T = 1) = E(x_{ji} \mid T = 0)$ for the treated and untreated states j. Therefore, the fundamental result for experimental samples is obtained straightforwardly:

$$E(Y_j \mid T = 1, X) = E(Y_j \mid T = 0, X) = E(Y_j \mid X), \ j = 1,0 \ . \qquad (6)$$

Random assignment does not remove selection bias by setting $E(e \mid T) = 0$, but instead balances the bias in the treatment and control samples such that it cancels out when estimating the mean impact estimate. Now the treatment effect on the treated can be estimated using simple mean differences: $\Delta_{TT} = \overline{Y}_1 - \overline{Y}_0$.[8] Moreover, the experimental sample solves one of the main source of bias in the non-experimental samples, the lack of common support. The experimental data guarantee a full common support by balancing the distribution of observable and unobservable variables between the treated and control groups.

Selectivity bias may be a major problem whenever the assignment to treatment and comparison groups is not random. The realization that estimates based on selected samples are troublesome can be traced to earlier work of Gronau (1974) and Heckman (1976). The selection bias arises from the dependence between T and e that under the separable framework is represented by,

$$B(X) = E(x_0 \mid X, T = 1) - E(x_0 \mid X, T = 0) \ . \qquad (7)$$

The conventional econometric approach to solve this endogeneity problem considers the partition of X into two not necessarily disjoint sets: $X = \{X_1, X_2\}$ where X_1 is the set of

[7] This is a random effect model where $j_1(X_i) - j_0(X_i)$ is the treatment effect common to all units with a given value of X_i and $x_{1i} - x_{0i}$ is the random component of the treatment effect. If we assume that $x_{1i} = x_{0i}$, a common effects model of treatment effect emerges. See Heckman and Robb (1985) and Robinson (1989) for a detailed discussion.

[8] Heckman , LaLonde and Smith (1999) point out that even if assumption (3) does not hold, we still can get unbiased estimates in two special cases: (1) under the "common effect" case. (2) if individual decisions about participating in the program are not affected by the forcasted gain from participating in the program.

covariates in the outcome equation and X_2 is the set of covariates in the participation equation that includes some exclusion restrictions. The latent index model (Heckman and Robb, 1986) further restricts the model so that the bias only depends on X_2 through a scalar index (the probability of participation). Denote $T_i^* = H(X_{2i}) + m_i$ where T_i^* is a latent index variable such as $T_i(X_{2i}) = 1$ if $T_i^* \geq 0$ for participants in the program and $T_i(X_{2i}) = 0$ if $T_i^* < 0$ for non-participants in the program; $H(X_{2i}) = X_{2i}g$ is the mean difference in utilities between the treatment and non-treatment states; and m_i is the unobservable white noise error variable independent of X_2 with a distribution denoted by $P(T = 1 \mid X_2) = F(H(X_2 g))$. If F is strictly monotonic, $F^{-1}(P(T = 1 \mid X_2)) = H(X_2 g)$ and, therefore, the bias depends only on P:

$$B(P(X_2)) = E(x_0 \mid P(X_2), T = 1) - E(x_0 \mid P(X_2), T = 0). \quad (8)$$

The classical selection model assumes that (X_1, X_2) is independent of e and the dependence between T and e arises through the correlation between the outcome equation unobservable (e) and the participation equation unobservable (m). Heckman and Hotz (1989) refer to this case as selection on unobservables. For instance, individuals with high unobservables in their participation equation are most likely to participate in the program. If the unobservable in earnings and participation equations are negatively correlated, these individuals are likely to have relatively low earnings, even after conditioning on X_2. On the other hand, dependence between T and e, that arises through the correlation between X_2 and T, is called selection on observable (Barnow, Cain and Goldberger, 1980). Non-experimental estimators that intend to estimate unbiased and consistent treatment effect estimates invoke different identification assumptions depending on the nature of the selection process they are dealing with. In the case of selection on unobservables, it is necessary to form assumptions about the distributions of unobserved variables e and u as well as the functional form relating e and m to X_2. In the case of selection on observables, it is necessary to identify, quantify, and include the variables that determine both participation in the program and outcomes in the absence of treatment. The empirical failure of the assumptions justifying any non-experimental

estimator results in estimates that strongly differ from the corresponding experimental estimates.

3. Non-Experimental Estimators for Average Treatment Effects on the Treated.

Assumption 1. (Exogeneity)

$$Y_0 \perp T \mid X .$$

This assumption is known in the literature as the ignorable treatment assignment (Rosenbaum and Rubin, 1983) or conditional independence assumption (Lechner, 1999). It refers to the independence of the counterfactual outcome from program participation conditional on a set of observable variables. Assumption 1 implies that systematic differences in outcomes between the treatment and comparison groups are attributable to the treatment once some observable variables are held constant. It assures the identification of the regression functions: $E[Y(T) \mid X] = E[Y(T) \mid T, X) = E[Y \mid T, X]$ and, therefore, the average treatment effects on the treated for a subpopulation with covariates X : $\Delta_{TT} = \underset{X|T=1}{E}[\Delta_{TT}] = E[Y_1 \mid X, T=1] - \underset{X|T=1}{E}(E[Y_0 \mid X, T=1])$. This treatment effect can only be estimated in the common support of the X covariates because for any particular $X = x$ out of the support there would be either treated or only comparison units. Therefore, it is necessary to invoke the second fundamental assumption:

Assumption 2. (Common Support)

$$\Pr(T = 1 \mid X) < 1 .$$

Rosenbaum and Rubin (1993) shows that if the exogeneity assumption holds, the problem of "curse of dimensionality" inherent in the dimension of X can be simplified. Let define the propensity score as $P(x) = \Pr(T = 1 \mid X = x)$ and let $b(X)$ be a function of attributes at least as "fine" as the propensity score. They show that the assignment and potential outcomes are independents once the balancing score (instead of the finest X) is held constant. Then, assumptions 1 and 2 are replaced by $Y_0 \perp T \mid b(X) = b(x)$ and $\Pr(T = 1 \mid b(X) = b(x)) < 1$.

The relevance of the second assumption varies across different estimators. In the context of average treatment effect on the treated, the existence of treated units and comparison units with "singular" covariates (propensity score) values have asymmetric

consequences for the precision of the estimates. Adding "singular" treated units have more adverse consequences for the precision of the estimates than adding "singular" comparison units since these last units are irrelevant for the average treatment effect for the treated. How considerable is this lack of precision depends on the type of estimator used. In the context of parametric regression functions that rely on extrapolation the presence of out of support treated units lead to an increase of variance although comparison units out of the support lead to an spurious increase in the precision of the estimates.[9] The nonparametric regression functions are better equipped to deal with limited overlapping since the out of support observations received smaller weights. Simple matching and propensity score matching are sensitive to the presence of treated units with outlying values. The quality of the matching can be severely affected leading to possible biased estimates. In fact, Heckman et al. (1998) find that limited overlap region produces biased matching treatment effects.

3.1 Matching Estimators

Matching estimators have been widely studied in the program evaluation literature (e.g., Heckman et al. 1997; Heckman et al. 1998; Dahejia and Wahba 1998; Abadie and Imbens 2002; Smith and Todd 2003). To estimate the treatment effect on the treated the matching estimator imputes the counterfactual for each treatment unit using a weighted average of the comparison units outcomes over the common support region and, then, estimates a simple means difference between the two samples. In that sense, matching resembles an experiment, no functional assumptions for the outcome equation is required:

$$\Delta_{TT}^{M} = \frac{1}{N_T}\sum_{i_T=1}^{N_T}\{Y_{1i}I^{CS} - (\sum_{i_C=1}^{N_c} W(i,j)Y_{0j}I^{CS})\}. \qquad (9)$$

where Y_{1i} is the outcome for treated units; $\sum_{jc} W(i,j)Y_{0j}$ is the estimated counterfactual mean with $W(i,j)$ representing the weights and Y_0 the outcome for comparison units; I^{CS} is an indicator function that takes the value 1 if the unit is in the common support region, 0 otherwise; N_T and N_C are the sample of treated and comparison units.

Four factors determine the methodological differences among matching estimators in theory and practice: weights, sample repetition, metric, and common

[9] The increase in the precision is not spurious if the functional form is correct.

support, all of which may lead to differences in judging the effectiveness of matching in solving the evaluation problem. The weights determine the type of matching estimator to be use by the evaluator. Dahejia and Wahba (1998) use the nearest–neighbor matching that match each treated unit to the closest – in terms of a parametric propensity score – unit in the comparison group. In this case the weights take only two values $W(i, j) = 1$ for the nearest-neighbor unit and $W(i, j) = 0$ otherwise. They find that this approach yield estimates very similar to the experimental ones using a subsample of LaLonde's data for the NSW program.[10] Three caveats have to be address when using this approach. First, in the presence of multiple ties in the propensity score, the estimates could be very sensitive to the seed used for the random number generator to break ties (Smith and Todd, rejoinder 2003). Second, the literature does not have any solid indication about the optimal number of neighbors to use, which may affect the robustness of these results. Abadie and Imbens (2002), show that bias-corrected matching estimates are more robust to different number of neighbors than simple matching. Third, matching with and without replacement may have important consequences in the estimation of the treatment effects. Matching with replacement allows comparison units be reused in the matching process which allows many treated units to be matched to the same comparison unit. This process has two potential effects. First, the precision of the estimates is improved (less variance) because of the larger sample but at the cost of greater variability (higher bias). Second, it is possible, however, that matching with replacement improves (worsens) the quality of the matching if a large (small) overlapping region is present, which leads to a lower (higher) bias and higher (lower) variance. What of these two effects will dominate is an empirical problem.

The kernel matching reuses the comparison units differently for each treated unit by a "smooth" kernel density function. Heckman, et al. (1998) propose the use a local linear kernel matching approach that define the weight function as,

[10] The NSW was a federal employment program developed in the mid-1970s that assigned qualified applicants to temporal training positions randomly. It was carrying out in ten cities: Atlanta, Chicago, Hartford, Jersey City, Newark, New York, Oakland, Philadelphia, San Francisco and Wisconsin.

$$W(i,j)=\frac{K_{ij}\sum_{k\in C}^{N_c}K_{ij}(P_k-P_i)^2-[K_{ij}(P_j-P_i)][\sum_{k\in C}^{N_c}K_{ij}(P_k-P_i)]}{\sum_{j\in C}^{N_c}K_{ij}\sum_{k\in C}^{N_c}K_{ij}(P_k-P_i)^2-\{\sum_{k\in C}^{N_c}K_{ij}(P_k-P_i)\}^2}. \qquad (10)$$

where $K_{ij}=K[(P_j-P_i)/h]$ is the kernel function that depends on the distribution of the propensity score P_j and P_i in the comparison and treated units; and h is the smoothing parameter that plays the same role as the number of neighbors in the nearest-neighbor matching case.[11] They reject empirically the assumption justifying matching when evaluating the JTPA training program.[12]

The caveat when evaluating this result is the choice of the smoothing parameters. The statistical literature (e.g., Marron, 1988; Jones, Marron and Sheather, 1996; Silverman, 1986) shows there is a price to be paid for the greater flexibility of the kernel-based method: the selection of the smoothing parameter h. When insufficient smoothing is done (h quite narrow) there are not enough observations appearing in each window for stability of the counterfactual average, resulting in matching estimates too rough and subject to sample variability. When excessive smoothing is done (h quite wide), the window is so large that the conterfactual average includes outcomes of distant comparison units that important features of the underlying structure is smoothed away. Thus, the implicit problem in the selection of the bandwidth is the trade-off between bias and variance. The evaluation literature, however, has been silent in the selection of the smoothing parameter in the context of average treatment effects.[13]

One difficulty in dealing with the integrated mean square error criteria is the presence of regions with irrelevant comparison units – those that receive zero or low weights in the matching process – that may mislead the results. We develop a leave-one-out cross-validation algorithm that minimize the mean square error of the average

[11] The local linear matching can be though as a kernel weighted regression of Y_{0j} on an intercept and a linear term in P_i where the estimated intercept is the missing counterfactual. This estimator is base on Fan's (1992) work that shows superior properties under boundaries bias with respect to the local constant kernel that is estimated without the linear term in P_i.`

[12] The JTPA is a federal funded program that provides a variety of employment, training, and services designed for economically disadvantaged adults and youth.

[13] An exception is Ichimura and Linton (2001) who derived a theoretically optimum smoothing parameter for the estimator proposed by Hirano, Imbens and Ridder (2003) in a kernel framework.

treatment effect on the treated weighting each comparison unit according to their distribution of covariates in the support region. The leave-one-out validation drops the ith unit in the comparison group and forms the counterfactual $\hat{Y}_{0-i}$ for that unit using the remaining $N_c - 1$ observations. The process is repeated N_c times until a counterfactual for each comparison unit is find. As each estimation does not include the ith unit, it represents an "out-of-sample" forecast that replicates well the essential features of the estimation problem.[14] Then, we minimize the mean square error of the average treatment effect on the (leave-one-out counterfactual) treated, weighting each unit with the overall kernel weights they would receive in the matching process. Following Frölich (2004), and Black and Smith (2003) we use in all the estimations the Epanechnikov kernel function because of its limited support that causes a faster rate of convergence than the Gaussian kernel and implicitly imposes the common support condition depending on the size of the smoothing parameter.

The metric used in the nearest-neighbor matching process can play a role in the size of the treatment effect estimates. Dahejia and Wahba (1998), for instance, use the standard Euclidean distance $d^E = \| P_j - P_i \|$ based on a parametric propensity score. In an earlier study, Rosenbaum and Rubin (1985) find that incorporating additional information in the metric yields a better balance in the distribution of observable covariates between the treated and comparison units. In particular, they propose to use the Mahalanobis distance metric with caliper, $d^M = (x_j - x_i)' \sum_X^{-1} (x_j - x_i)$, where $\sum_X^{-1}$ is the covariance matrix of the vector x that incorporates the propensity score.[15] This metric has the nice property of balancing the covariates in all directions within matched pairs. In the presence of high correlation among the covariates, however, this metric can give misleading results (Imbens, 2003). Abadie and Imbens (2002) modify the Mahalanobis metric to $d^{AI} = (x_j - x_i)' diag(\sum_X^{-1})(x_j - x_i)$ where $diag(\sum_X^{-1})$ is the diagonal of the inverse of the covariance matrix, and it may overcome the problems cause by having high correlations between the covariates. Something missing in the evaluation of non-

[14] Hall et al. (2003) proves that cross validation without leave-one-out produces an optimal h=0.
[15] The calipers are defined by the propensity score and they are selected following the method proposed by Cochran and Rubin (1973).

experimental estimators in replicating experimental estimates is the robustness of the results to the adoption of different distance metrics. We use alternative metrics to analyze the sensitivity of the treatment effects in finite samples.

Similarly, there are differences in the way the common support region is constructed. The predominant method is based on an estimate parametric propensity score. The advantage of using the scalar propensity score instead of a higher dimensional vector *X* is that allows to detect and to depict problems with the common support easily. Regions with limited support can be difficult to identify in high dimensional space as it can be masked for any single covariate. Heckman et al. (1998) and Smith and Todd (2003) proposes a trimming method base on the estimation of a kernel density function for a parametric propensity score,

$$CS^{HIST} = \{S : \hat{f}(P \mid T = 1) > C_q \ \& \ \hat{f}(P \mid T = 0) > C_q\}. \qquad (11)$$

where $\hat{f}(P \mid T = t) = \frac{1}{Nh} \sum_{k \in N} K[(P_k - P)/h]$ is the kernel density function evaluated at all observed data points for both the treated and comparison units, and C_q is the density cut-off level below 5%. The determination of the smoothing parameter is critical here. It is widely documented that the distribution of the empirical common support is highly skew to the right for the comparison units and highly skew to the left for the treated units. Therefore, assuming a normal distribution for the selection of the smoothing parameter may mislead the finding of the empirical common support region. In fact, using the Silverman's rule-of-thumb, based on the normal assumption, or using a least-square cross-validation process for selecting the optimal smoothing parameter yield two very different empirical support regions in our data that strongly affect the estimation of the treatment effects. Dahejia and Wahba (1998) use a simpler approach. They drop all units with parametric propensity scores below the maximum of the minima and above the minimum of the maximums:

$$CS^{DW} = \{S : i, j > \max\{\min(P_i), \min(P_j)\} \ \& \ i, j < \min\{\max(P_i), \max(P_j)\}\}. \qquad (12)$$

A potential problem with this approach is the existence of one comparison unit with high propensity score and a treated unit with low propensity score that make impossible to

drop any unit even when many comparison and treated units are concentrated in the left and right extremes of the propensity distribution.

Lechner (2000) focus only in the distribution of the treated units and define the common support as the region where at least k potential comparison units are available, dropping all the treated units with parametric propensity scores higher than that for the kth larger comparison unit,

$$CS^L = \{S : i, j = [i < \max^k(P_j), j]\}. \quad (13)$$

All of these approaches drop observations in the tails of the propensity score distributions although the trimming method also allows gaps in the empirical support region. Large differences in the treatment estimates based on these empirical support regions may reveal lack of robustness of non-experimental estimates in replicating experimental estimates. In addition, any misspecification in the estimation of a parametric propensity score can mislead the identification of the support.

Propensity Score Model and the Common Support Region

One of the most intriguing results in the evaluation literature is the specification of the parametric propensity score and its effect on the estimation of average treatment effects. It is show that different specifications lead to different magnitudes of the treatment effects (Heckman et al. 1998; Lechner 2000) which, may affect the relevance of matching in solving the evaluation problem (Dahejia 2003; Smith and Todd 2003). Parametric models that pass standard balancing tests are regard as valid because they balance the distribution of pre-treatment covariates between matched units conditional on the propensity score (Rosenbaum et al. 1985; Lechner 2000). One potential problem, however, is that different standard balancing tests may yield different answers (Smith and Todd 2003). Henceforth, the adoption of a nonparametric approach gives a robust approach to dealing with the misspecification problem: it allows a clean identification of the support region; and, as it is show by Hahn (1998), and Hirano et al. (2003), it leads to efficient estimators.

Following the work of Li and Racine (2002) we estimate a nonparametric kernel conditional probability density function, $f(T \mid X) = f(T, X) / f_1(X)$, where $f(T, X)$ is

the joint density function and $f_1(X)$ is the marginal density function of X. The joint PDF is estimated by,

$$\hat{f}(T, X) = \hat{f}(Z) = \frac{1}{n}\sum_{i=1}^{n} K_m(z_i, z, h). \qquad (14)$$

where n is the number of observations and $K_{m,iz}$ is the multivariate kernel constructed by "hybrid" product kernels where each univariate kernel corresponds to each data type:

$$K_m(z_i, z, h) = \prod_{j=1}^{d} K_d(z_{ij}, z_j, h_j) \prod_{j=d+1}^{d+c} K_c(z_{ij}, z_j h_j). \qquad (15)$$

where K_d and K_c are the univariate kernel function for categorical and continuous data; d and c are the number of categorical and continuous covariates; and h is the smoothing parameter.[16] We use the Epanechnikov kernel function for continuous data:

$$K_c(z_i, z, h) = \begin{cases} \frac{1}{h}\left(\frac{3}{4\sqrt{5}}\left(1-\frac{1}{5}(\frac{z_i - z}{h})^2\right)\right), \left|\frac{z_i - z}{h}\right| < \sqrt{5}. \\ 0, \text{otherwise}. \end{cases} \qquad (16)$$

and the Aitchison and Aitken (1976) kernel function for c-categorical data:

$$K_d(z_i, z, h) = \begin{cases} 1-h, z_i = z. \\ \frac{h}{c-1}, \text{otherwise}. \end{cases} \qquad (17)$$

One advantage of this method over the conventional frequency estimator is that it does not split the data into cells avoiding a potential small number of observations in each cell that may cause inaccurate nonparametric estimation of the PDF of the remaining continuous variables in finite sample applications.[17] The identification of the optimal smoothing parameters is carry out by least-square minimization of a cross-validation function with mixed data,

$$CV(h) = \int \{\frac{1}{n}\sum_{i=1}^{n} K_m(z_i, z, h) - f(z)\}^2. \qquad (18)$$

[16] The marginal probability density function for X is derived following exactly the same definitions.
[17] It is very common to find small datasets in the evaluation of labor market programs. For instance, the JTPA dataset use in Heckman et al. (1998) has less than 1,000 observations.

that has a rate of convergence that depends only on the number of continuous variables involved and smooth away irrelevant regressors.[18]

Given that the identification of the treatment effect relies on Assumption 1, covariates influencing both the decision to participate in the KWPRS program and future potential outcomes of the UI beneficiaries should be included in the estimation of the conditional density function. As the participation decision is entirely based on the profiling score, this variable is a crucial component of the model. Standard human capital and search models suggest that when predicting future outcomes of UI beneficiaries, it is important to take into account opportunity costs such as lost earnings and lost leisure that differ across individuals according to tastes, socioeconomic factors, and personal labor market history. Common variables that have been used in empirical analysis and can approximate these categories are sex, education, age, race, region of residence. Following Ashenfelter (1978), we include past earnings (annual earnings) as a key variable for predicting participation. In addition, Heckman and Smith (1995b) emphasize the importance of considering labor force status transitions in the participation model. Unfortunately, we do not have information to identify unemployed or inactive persons, only participation status. Consequently, we include a set of dummy variables to indicate transitions between four and one quarter before the program: Employed→ Employed, Employed→ No Employed, No Employed→ Employed, and No Employed→ No Employed. Also, we consider a set of dummy variables indicating each one of the 32 local offices where the individuals receive the UI benefits. Its inclusion, however, worsens the support region and it is correlated to residence dummy variables.[19] For these reasons, we decide to drop this variable from the participation model. It is worth notice that we have only one continuous covariate, therefore, we get a consistent nonparametric propensity score having one-dimensional rate of convergence.

Panel A of Figure 1 shows the distribution of the kernel-based propensity scores. For the treated units it has distribution values over the entire support [0,1] ranging from 0.033 to 0.991 with a mode close to 0.15. The comparison units present a thinner support

[18] Li and Racine (2002) also show the consistency and asymptotic normality of this estimator.
[19] It affects adversely the quality of the matches: there are some comparison units that are used as matches up to 56 times, whereas without include this variable, some comparison units are use as matches up to 16 times.

with values ranging from 0.008 to 0.831 and a mode close to the minimum value. It implies that 25% of the comparison units have propensity scores less than the minimum for the treated units and 8.4% of the treated units have propensity scores higher than the maximum for those in the comparison group. One of the features of smoothing categorical variables using Hall et al.'s (2003) data-driven cross-validation method is its ability to remove irrelevant covariates. In equation (17) we obtain a uniform weight function when h attains its upper bound value $h^{+} = \frac{c-1}{c}$. It means that those covariates that have their cross-validated smoothing parameters equal to their upper bound are in fact irrelevant predictors and will be automatically removed.[20] In Table 2 we observe that for age the upper bound coincides with the cross-validated optimal smoothing parameter, which implies that this covariate is irrelevant for the conditional probability estimator. Only sex, race, region of residence, labor market transitions, profiling scores and past earnings are relevant predictors for participation and to a lesser extent schooling. This is similar to Ashenfelter's (1978) and Heckman et al.'s (1995b) findings about the importance of including past earnings and labor market transitions as key predictors for participation in labor market programs. In order to assess the misspecification issue, we consider the in-sample predictions for the kernel method and its parametric counterpart model (Table 3) that is estimated using the same covariates – including higher order terms for age and profiling score – and pass balancing tests as those describe in Smith and Todd (2003).[21] Table 4 shows the confusion matrix for kernel and probit models. $\hat{P}(X) > E(T)$ and $\hat{P}(X) \leq E(T)$ are use to predict $T = 1$ and $T = 0$. The nonparametric model gives an overall rate of 78% correct predictions whereas the parametric model correctly predicts 69%. The predictions results indicate that kernel-based approach with mixed data makes a better job in predicting the probability of participation (non-

[20] $h \to \infty$ is the upper bound in the case of irrelevant continuous covariates.

[21] The parametric model pass the standardized differences test (Rosenbaum and Rubin 1985) that considers the size of the weighted difference in means of pre-treatment covariates between the treated and matched comparison units, using the standard deviation of each covariate in the raw data as the scale. The weighted differences range from 0.15% for labor market transitions to -8.9% for region of residence (metropolitan area) with a median value of 2.85% for all covariates. Also, it pass for 6 out of 12 variables the regression-based parametric test proposed by Smith and Todd (2003) that examine whether T provides any information for each covariate conditional on a quartic in the estimate propensity score.

participation) in the program for those that participate (non-participate); and it indicates that the parametric model may suffer misspecification.

This better specification performance, however, implies a slightly thinner support region than that for the parametric model. Panel B in Figure1 show the distribution of the propensity score values for the probit participation models. In the case of the parametric model, the propensity score values are not over the entire support [0,1] but in the range [0.003, 0.77] yielding a thicker overlapping region than in the case of the kernel-based model. Any misspecification in this model, however, masks the true overlapping region. Panel C and D in Figure 1 show the distributions after imposing an empirical common support region using the trimming method described in Heckman et al. (1997) and least-square cross-validation for selecting the smoothing parameters. The imposition of the Silverman's rule-of-thumb for selecting the smoothing parameters may mislead the empirical overlapping region if the estimated propensity score (Panel A) does not follow a normal distribution.[22] In fact, under the kernel propensity model and 2% trimming method, the least squares cross-validation shrinks 24% and 27% of treated and comparison units with 64% of out-of-support treated units having $\hat{P}(X) > 0.75$ and 97% of out-of-support comparison units having $\hat{P}(X) < 0.05$. On the other hand, using the Silverman's rule-of thumb shrinks 19% and 3% of the treated and comparison sample with 76% of out-of-support treated units having $\hat{P}(X) > 0.75$ and 83% of out-of-support comparison units having $\hat{P}(X) < 0.05$. In the case of the probit model, that imposes a normal distribution of the errors terms, the differences between least-square cross-validation and the Silverman's rule-of-thumb yield almost similar results. Panel E and F in Figure 1 show the distribution of the propensity score after imposing the kth larger comparison unit criterion, which drops 18% of treated units with $\hat{P}(X) > 0.752$ in the kernel model; and 21% of treated units with $\hat{P}(X) > 0.54$ in the probit model.

[22] Silverman's rule-of-thumb is define as: $h = 1.06An^{-1/5}$ with $A = \min(R, s)$, where R is the interquartile range/ 1.34 and s is the standard deviation.

3.2. Regression-Approach Estimators

The regression-based approach for estimating the average treatment effect for the treated relies on the consistency of the counterfactual regression $E[Y_0 \mid T=1, X] = m_0(X)$. Invoking Assumption 1 we can identify this conditional expectation and estimate the treatment effects by averaging the difference between the actual outcome for the treated and their estimated counterfactual outcomes:

$$\Delta_{TT}^{REG} = \frac{1}{n_1}\sum_{i=1}^{n_1} T_i(Y_i - \hat{m}_0(X)) \cdot \qquad (19)$$

We use a training/evaluation framework to estimate the regression-based treatment effects. First, we estimate the conditional expectation $\hat{m}_0(X)$ using only the comparison units (training data) with local linear kernel regression function with mixed categorical and continuous data. Then, we predict the counterfactual outcomes for each treated unit evaluating the estimated conditional expectation function in the sample realization of the treated units (evaluation data). Finally, a simple means difference between the observed outcomes for the treated units and the estimated counterfactuals outcomes gives the treatment effect on the treated. The obvious advantage of this method over the classical parametric approach (e.g., Rubin 1977) is its greater flexibility that helps to solve some problems caused by parametric assumptions in sample-selectivity models (Moffitt, 1999). For instance, if the differences between the treated and comparison mean covariates are large, the predictions based on parametric linear models can be very sensitive to changes in the specification. Furthermore, the adoption of the "hybrid" product kernels proposed by Racine and Li (2001) allows a nice interaction between discrete and continuous covariates without need for splitting the sample into cells. We denote the conditional expectation $\hat{m}_0(X)$ by a local linear regression function,

$$\hat{m}_0(X) = \frac{\sum_{i=1}^{n_0} Y_i[K_{im}\sum_{i=1}^{n_0} K_{im}(x_i - x)^2 - K_{im}(x_i - x)\sum_{i=1}^{n_0} K_{im}(x_i - x)]}{\sum_{i=1}^{n_0} K_{im}\sum_{i=1}^{n_0} K_{im}(x_i - x)^2 - [\sum_{i=1}^{n_0} K_{im}(x_i - x)]^2} = \sum_{i=1}^{n_0} Y_i W(x_i, x, h). \qquad (20)$$

where $K_m(x_i, x, h)$ is the "hybrid" product kernel defined in (15), and n_0 is the number of comparison units. Likewise, to the case of the kernel density function, the smoothing

parameters are chosen by least-square minimization of a cross-validation function with mixed data,

$$CV(h) = \frac{1}{n_0}\sum_{j=1}^{n_0}[Y_j - \sum_{i=1, i\neq j}^{n_0} Y_i W(x_i, x_j, h)]^2 \ . \qquad (21)$$

that has a rate of convergence that depends only on the number of continuous variables involved, and, as in the case of kernel distribution estimation, smooth away irrelevant covariates.

A promising application of the kernel regression-approach with mixed data is the estimation of regression-adjusted matching. Abadie and Imbens (2002) show that the bias of the nearest-neighbor matching can dominate the variance if the dimension of the covariates is large, so additional bias correction through regression can be very useful. Heckman et al. (1998) show that regression-adjusted local linear matching helps in reducing the size of the selection bias when estimating treatment effects with non-experimental data. The basic idea is to remove Xb_0 from Y_0 where $\hat{b}_0$ is estimated using a parametric linear model (Rubin 1973b) or semiparametric linear model (Heckman et al. 1998) defined by $Y_i - E(Y_i \mid P_i, T_i) = [X - E(X \mid P_i, T_i)]b_0 + e_i$, where $E(Y_i \mid P_i, T_i)$ and $E(X_i \mid P_i, T_i)$ are univariate kernel conditional expectations on the parametric propensity score. The outcomes Y_{1i} and Y_{0i} in the equation (9) are replaced by $Y_{1i} - X_i\hat{b}_0$ and $Y_{01} - X_i\hat{b}_0$. This process assumes, however, that the slope coefficients b_0 come from an adjusted linear model and are constant across all units and across time.[23] It imposes, also, the linearity of the adjusted factor $X_i\hat{b}_0$. We relax these assumptions using a multivariate kernel regression with mixed data for the comparison units, which allow us to estimate unit-varying b_{0i} without imposing both the linearity of the regression function and the linearity of the adjusted factor; and without relying on a parametric propensity score. Then, we evaluate the results in the treated sample and replace the outcomes Y_{1i} and Y_{0i} in the equation (9) by $Y_{1i} - \hat{E}(Y_{1i} \mid X_i)$ and $Y_{0i} - \hat{E}(Y_{0i} \mid X_i)$.

Although adjusting only for differences in the propensity score removes bias, it needs not be as efficient as adjusting for differences in all covariates. Hahn (1998), and

[23] For further details, see Robinson (1988).

Hirano et al. (2003) show that the estimated nonparametric propensity score is a valuable source of information for estimating average treatment effects in a regression framework. They advocate the use of all comparison units to adjust for covariate imbalances by weighting the estimate treatment effects by the probability of participating in the program. Units with low (high) $f_0(X_i)$ density relative to $f_1(X_i)$ are under (over) represented in the comparison sample. Hence, weighting the comparison units by $f_1(X_i)/f_0(X_i)$ corrects this unbalance. We follow this approach by using a nonparametric propensity score as the weight variable in a regression-framework for average treatment effect on the treated:

$$\Delta_{TT}^{HIR} = [\frac{1}{n_1}\sum_{i=1}^{n_1} Y_i] - [\sum_{i=1}^{n_0} Y_i \frac{P(X_i)}{1-P(X_i)} / \sum_{i=1}^{n_0} \frac{P(X_i)}{1-P(X_i)}]. \qquad (22)$$

It achieves the semiparametric efficiency bound, which only requires a conditional mean estimation $E(P \mid X)$. Our estimation approach, however, is different to Hahn's and Hirano's et al. To estimate the propensity score, we use kernel-based approach with mixed categorical and continuous data instead of the proposed series logit estimator. The optimal smoothing parameters are found by least-square cross validation.

3.3 The Regression-Discontinuity Approach

The design of the Kentucky UI program is based on an experiment where the random assignment occurs at the margin only to satisfy capacity constraints. This assignment rule has discontinuities that naturally identify a non-experimental group: those with profiling scores below the marginal score. This idiosyncratic feature in the participation process is similar to a quasi-experimental design originally introduced by Thistlethwaite and Campbell (1960) and named "tie-breaking" experiment. Under some conditions, this design can be use to estimate unbiased and consistent non-experimental results without imposing arbitrary exclusion restrictions, index assumptions on the selection process, functional forms, and distributional assumptions on errors through a regression-discontinuity approach. This method combines distinctive features of social experiments (there is a known rule that assign persons in or out of the treatment) with features of non-experimental designs (the rule is not random by nature) that makes it a powerful and interesting method of evaluation.

Earlier parametric applications of this method exploit discontinuities in the relationship between two endogenous variables in order to identify parameters of interest. Thistlethwaite and Campbell (1960) estimate the effect of receiving a National Merit Scholarship Award on student's scholar success. Angrist and Lavy (1996) measure the effect of classroom size on student test scores using a nonlinear and nonmonotonic relationship between grade enrollment and class size in Israeli public schools. We owe to van der Klaauw (1986, 2001), and Hahn, Todd, and van der Klaauw (1999) the integration of the regression-discontinuity design in a kernel-based nonparametric estimation. They take advantage of the discontinuities in the relationship between average aid offer and student's ability index to estimate the effect of colleges' financial aid offers on student enrollment decisions, and features (discontinuities) in the law to estimate the effect of firm size on minority employment using one-side kernel regressions.

The regression-discontinuity approach is based on the total or partial dependence of the treatment assignment on an observed variable (S_i) such that the probability of participating in the program is a discontinuous function of this variable at the cutoff score ($\overline{S}_i$). Depending on the nature of the observed variable, S_i, the literature distinguish between sharp and fuzzy designs. In the first case, individuals are assigned to treatment solely on the basis of a known and quantifiable observed measure of S_i – selection on observables -. On the other hand, if the assignment to treatment is based on known and unknown variables (i.e., variables observed by the administrator but not for the evaluator) we are in a fuzzy design world that entails both selection on observables and selection on unobservables.[24]

Our case in one of sharp design where the selection rule (S_i) is the profiling score (r_i) and the cutoff score ($\overline{S}_i$) is the (week/local office specific) marginal profiling score ($\overline{r}$) which assigns into the treatment to those units with equal or higher profiling scores than the marginal profiling score: $T_i = T_i(r_i) = 1\{r_i \geq \overline{r}\}$. The selection bias comes from the potential relationship between r_i and the outcome variable Y_i that causes a correlation between the treatment indicator $T_i(r_i)$ and the unobservables. We can solve the selection

[24] van der Klaauw (1986, 2001) provides an excellent discussion and empirical application of the fuzzy design.

issue if we assume there is a locally continuous relationship between the outcome and the selection variable (see van der Klaauw 2001; Hahn et al. 1999):

Assumption 3.

$E(e \mid r)$ is continuous at $\bar{r}$.

It is equivalent to assuming that for those individuals just above and just below the marginal profiling score, we expect similar average outcomes because they share almost identical observable characteristics. Therefore, for any arbitrary small bandwidths (Ω) we expect that $E(\Delta_{TT} \mid r + \Omega) = E(\Delta_{TT} \mid r - \Omega)$. It implies that the treatment effect on the treated will be identify by mean differences of the outcomes for those units immediately above and below of the marginal profiling scores,

$$\Delta_{TT}^{RDD} = \lim_{r \uparrow \bar{r}} (E(Y_i \mid r)) - \lim_{r \downarrow \bar{r}} (E(Y_i \mid r)) . \qquad (23)$$

In this sense, the regression-discontinuity approach resembles matching by balancing the distribution of observable covariates in the selected units. A key difference, however, is the absence of the region of common support since by construction $\Pr(T = 1 \mid r) \in \{0,1\}$.

The estimator (19) can be estimated by extrapolation through parametric fixed-effect models or by flexible kernel-based approaches. In the parametric regression-approach, it is possible to purge any correlation between T and ε if the variables that determined the assignment are known, quantify and include in the regression (Barnow, Cain and Goldberger, 1980; Heckman and Robb, 1985; Heckman and Hotz, 1989). It is worth noticing, the Kentucky UI program does not have a simple discontinuity region. Since the data includes 32 local offices and 87 weeks of program's length, it counts for 2,784 potential cutoff scores. Given that for some week/offices interactions we have empty cells, the actual data includes 286 "'discontinuity" groups with at least one treated and comparison units, and an average of 36.8 units per group ranging in size from 2 to 134. Therefore, a parametric approach relies on the following specification,

$$\Delta_{TT}^{PRDD} = b : Y_i = a + bT_i + F(r_i) + h_i + x_i . \qquad (24)$$

where Y_i is the outcome variable, T_i is a dummy variable 1/0 for selection into the treatment, $F(r_i)$ is a control function, h_i is a week/offices fixed effect variable and $x_i = Y_i - E(Y_i \mid T_i, r_i)$ is the error term. We consider different specifications for the control

function assuming that the true conditional mean function $E(e \mid r)$ belongs to the class of polynomial functions (constant, linear or quadratic). Following van der Klaauw (2001), we also adopt a cross-validation method to select the optimal number of terms in the control function based on a power series approximation $F(r) \cong \sum_{i=1}^{J} l_j (r)^j$ where the size of the polynomial is determined by a data-driven process.

A more flexible method to estimate the treatment effects is provided by the kernel-based approach that avoids any risk of misspecification of the functional form in the outcome equation. Hahn et al (1999) proposed estimate $\lim_{r \uparrow \bar{r}}(E(Y_i \mid r))$ and $\lim_{r \downarrow \bar{r}}(E(Y_i \mid r))$ by one-sided kernel or local linear kernel regression which is proved to be numerically equivalent to a local Wald estimator under some conditions.[25] Although the identification of the treatment effects only requires Assumption 3, there is a cost of using this nonparametric approach: we cannot identify the treatment effect over the full support of r but for values just above and below the cutoff score $\bar{r}$. The estimation of separate one-side kernels regressions for each one of the 286 "discontinuity-groups" is not possible because of the small sample size for many of them. Therefore, we adopt a simpler non-parametric approach by comparing treated units (directly above the cutoff score) and comparison units (directly below the cutoff score) through a weighted sample mean difference,

$$\Delta_{TT}^{NPRDD} = \sum_{j=1}^{286} w_j \{ \frac{1}{n_1} \sum_{i=1}^{n_1} Y_{1i} I^D(h) - \frac{1}{n_0} \sum_{i=1}^{n_0} Y_{0i} I^D(h) \} . \qquad (25)$$

where w_j are the cell-weighs; Y_{1i} and Y_{0i} are the treated and comparison outcomes; $I^D(h)$ is an indicator function that depends on a selected bandwidth (h) and takes the value 1 for those units in the vicinity of the marginal profiling scores, 0 otherwise.

[25] It assume a uniform kernel function over the subsample $r_0 - h^- \prec p_i \prec r_0 + h^+$, where h is the bandwidth.

4. Empirical Estimates

Two direct measures of the KWPRS treatment effects are the spell duration and the amount of UI subsidizes received. Black's et al. (2003) experimental estimates show that the threat of mandatory reemployment services can be more effective than the services themselves. Likewise, they estimate annual and quarterly earnings with results suggesting that the shorter spells induced by the KWPRS program does not result in less favorable match between workers and jobs. It is important to mention that these experimental results are not estimated by simple mean differences of outcomes between the treated and control samples as it usually happens in the context of simple random assignment. Indeed the KWPRS experiment ensures a random assignment only within each PTG. Therefore, the identification of the parameters of interest needs additional structure such as the "common effect" assumption or heterogeneous impacts across persons but only as a function of observed covariates $b = b(X_i)$. Under the common effect world, the experimental treatment effects are estimated by unweighted least squares using a vector of PTG fixed effects to control for differences in expected outcomes in the absence of treatment across PTGs. Consequently, the experimental treatment estimates we intent to replicate are –2.24 for weeks of UI benefits; -$143 for amount of UI benefits received; $1054 for annual earnings; $525 for 1st quarter earnings; $344 for second quarter earnings; $220 for 3rd quarter earnings; and $-35 for 4th quarter earnings.

4.1 Matching Estimates

As described in section 4, we present the nearest-neighbor matching estimates using both kernel-based and parametric (probit) propensity score models. We select the Mahalanobis distance as the nearest-neighbor metric because it has the nice property of balancing the covariates (including the propensity score) in all directions within matched pairs. In fact, our data show lower standardized differences between the treated and matched comparison units for all the pre-treatment covariates when using Mahalanobis instead of the Euclidian distance on the propensity score. We use the *kth* larger comparison unit criterion (Lechner 2000) to define the empirical support region in all the estimations. Two reasons explain this approach. First, it drops almost the same number of units in both kernel-based and probit propensity score models, which allow us to perform a

comparability analysis between both models under a common overlapping region.[26] Second, in the context of treatment effects on the treated, "singular" comparison units are less relevant than "singular" treated units.

Table 5 show the estimates using the kernel-based propensity score. Each row depicts the treatment effects for each outcome variable and each column represents a different matching estimator. Bootstrap standard errors appear in parenthesis below each estimate and the number in brackets represents the estimate non-experimental bias.[27] The first important result that comes forward is the ability of the nearest-neighbor matching estimator with common support region to approximate the experimental estimates. With the exception of 4th quarter earnings, all the outcomes present low bias, ranging from -7% (weeks receiving benefits) to -34% (3rd quarter earnings). It is worth noticing that the imposition of an empirical common support improves the estimates for all the variables. Likewise, the regression-adjusted nearest-neighbor matching produces the best results among all the matching estimators. It combines matching with nonparametric regression-adjusted functions without imposing both the linearity of the regression function and the linearity of the adjusted factor. The estimate bias drops to -4% for the weeks receiving UI benefits; -16% for amount of UI benefits; -12% for annual earnings; -13% for 1st quarter earnings; -28% for 2nd quarter earnings, and -33% for 3rd quarter earnings. Columns 6 to 9 present the local constant and local linear kernel matching estimates using the Epanechnikov kernel function with cross-validated optimal smoothing parameters. Comparing with the nearest-neighbor estimates, they produce somewhat similar results for weeks receiving UI benefits (-7% to -3% bias), amount of UI benefits (0% to 15%), and 1st quarter earnings (-26% to -23%); they improve over the nearest-neighbor estimate in the case of 4th quarter earnings (-42%); and they get worse for annual, 2nd and 3rd quarter earnings. It is important to mention that local constant kernel shows less bias than local linear kernel for all the outcome variables. This result is consistent with Frölich's (2004) Monte Carlo analysis.

[26] Under the trimming method, both models depict different overlapping regions. Hence, this feature can mask any potential difference between the models when analyzing their robustness to alternative number of neighbors or metrics.

[27] The bias is define as $[(\Delta_{TT}^{NoExp} - \Delta_{TT}^{Exp}) / \Delta_{TT}^{Exp}]*100$

Table 6 shows matching estimates with a probit propensity score as the only methodological difference with respect to Table 5. In general, the estimate treatment effects present higher bias that its counterpart kernel-based propensity score. Moreover, the local linear kernel matching estimates yield for most of the outcome variables treatment effects far away from the experimental estimates. Taken together the estimates in Table 5 and 6 three basic patterns emerge. First, using a kernel-based propensity score with mixed categorical and continuous data and cross-validated smoothing parameters produce a much better approximation to the experimental estimates than the traditional probit parametric propensity score. This result holds for both the nearest-neighbor and the kernel matching. Second, the biggest difference in favor of the kernel-based propensity score specification is in the estimation of the earnings treatment effects. Third, for both approaches, the nearest-neighbor matching estimator produces lower bias estimates than the local linear kernel estimates.

In order to investigate the robustness of this result, we estimate treatment effects for the earning variables using a different metric (the Euclidian distance) and without imposing an empirical common support. Figure 2 shows the new estimates and clearly reinforce our previous findings. An additional sensitivity check is performed using alternative number of neighbors. We use five sets of neighbors ranging from 1 to 5.[28] We estimate treatment effects for each neighbor set and then we average the absolute value of the resulting bias across the five different sets. Table 7 shows the average bias for each earning variable with their respective coefficient of variation. Again, we obtain for the kernel-based propensity score a lower bias (global average of 51%) than that for the parametric probit model (global average of 88%), although the coefficient of variation shows a higher dispersion for the kernel-based nearest-neighbor estimates. A final sensitivity test considers a different specification for the parametric model. Following Heckman et al. (1998) we consider the best-predictor parametric model build under two restrictions: (1) Minimization of the classification error where $\hat{P}(x) > E(T)$ predicts participation and $\hat{P}(x) < E(T)$ predicts nonparticipation; and (2) Inclusion of only statistical significant covariates. Table 8 shows treatment effect estimates for the kernel-

[28] Abadie and Imbens (2002) suggest choose in practice a fairly small number. Their simulations show that four matches perform well in terms of mean square error.

based and the best predictor parametric model using nearest-neighbor matching on the propensity score and the *kth* larger comparison unit criterion to define the empirical support region. The estimates show again a better approximation of the kernel-based propensity score to the benchmark estimates. Therefore, our results hold independently of the metric, number of neighbors, empirical support region, and specification of the parametric model.

4.2 Regression-Based Estimates

In Table 9, we explore the ability of regression-based functions in replicating the experimental estimates. We use a parametric and nonparametric training/evaluation framework based on the same set of covariates –age, schooling, sex, race, region of residence -. The number in parenthesis gives the standard errors and the number in brackets the estimated bias. Three major patterns emerge. First, the parametric model makes a good job in replicating the earning outcomes. In particular, annual, 1st and 3rd quarter earnings present low bias (less than 10%). These estimates, however, are very sensitive to changes in the specification of the regression function, and therefore should be taken with reserve. Second, the nonparametric regression function gives estimates close to the experimental ones for most of the earnings outcomes. For instance, annual earnings present -7% bias; 1st quarter earnings -12%; 2nd quarter earnings -41%; and 3rd quarter earnings -16%. It is worth noticing that these estimates do not suffer from misspecification and in that sense are robust. Likewise, the weighted regression (Hirano et al. 2003) use the kernel-based propensity score and gives somewhat similar bias for the earnings variables and produces the best approximation to the experimental estimate for weeks receiving UI benefits (-16% bias). Third, all the regression-based estimates perform poorly for amount of UI benefits and 4th quarter earnings. Not only the bias exceeds somehow acceptable regions but also the estimates are qualitatively different to the experimental estimates (opposite sign). In this sense, a clear result comes forward when contrasting the regression-based estimates with those for the nearest-neighbor matching. The implicit "common support" condition inherent in the pair matching method increases the ability of non-experimental estimators in replicating experimental

estimates. In particular, the amount of UI benefits and 4th quarter earnings variables are much better estimate with matching than with regression methods.

4.3 Regression-Discontinuity Estimates

In this section we present the regression-discontinuity estimates. In particular, Table 10 presents the nonparametric treatment effects for those units just above and below the week/local offices marginal profiling scores. The rows represent each of the outcome variables of interest and the columns represent alternatives distances (bandwidth) from the marginal profiling score. The numbers in parenthesis are the standard errors, and the non-experimental bias is shown in brackets. It is clear that for most of the earning estimates, the regression-discontinuity approach show a close approximation to the experimental treatment effects. In particular, 1st and 2nd quarter earning present estimates with low bias (around 5%). The treatment effects for annual and 3rd quarter earnings are still reasonably close approximations to the experimental estimates with a bias around 40%. A different picture, however, takes place for amount of UI benefits and 4th quarter earnings that present highly biased treatment effects. This result is consistent with the regression-based estimates that also show a poor performance in replicating the experimental estimates for these two variables. An interesting, result observed for all outcome variables, is the stability of the treatment effect estimates to alternative bandwidth sizes. The variation in the estimates is less than 5% across three different definitions for the bandwidths.

Table 11 shows the fixed-effect parametric estimates using the same sets of observations as Table 10. As before, the treatment effects for most of the earnings variables have a relative small bias. In particular, the treatment effects for annual and 3rd quarter earnings improve over the nonparametric case. For instance, the treatment effect for annual earnings with a +/- 3 bandwidth replicates almost exactly the experimental estimates (1% bias). On the other hand, the treatment effects for 4th quarter earnings and amount of UI benefits are again badly biased respect to the benchmark experimental estimates. The assumption of a correct specification for the control function in a parametric setting allows the expansion of the sample size beyond the units just above or below the cutoff points. Taking all units in the non-experimental sample and using

week/local office fixed-effects, we obtain parametric estimates for a constant, linear, and quadratic specification of the profiling score. As we observe in Table 12, the treatment effects do not improve over the estimates founded using only the units in the neighborhood of the discontinuity points. In addition, the treatment effects reinforce the previous results about the inability of the regression-discontinuity design to replicate the experimental estimates for amount of UI benefits and 4th quarter earnings. They present opposite signs and bias estimates over 200%.

Taking together all these results, we can conclude that the regression-discontinuity design is a promissory method for solving the evaluation problem. When restricted to sample observations in the neighborhood of the discontinuity points, this method does not only replicate well the experimental estimates for weeks receiving UI benefits and most of the earnings categories, but also they show a strong consistency across different subsamples and across different estimators. In this sense, the regression-discontinuity design shows more stable results than matching estimates that present higher variability depending on the number of neighbors, the specification of the propensity score, empirical support region, etc.

5. Conclusions

Our analysis of the data from the Kentucky Working Profiling Reemployment Services (KWPRS) yields five main conclusions. First, a fully nonparametric matching approach using kernel methods with mixed categorical and continuous variables and cross-validation methods for selecting the optimum bandwidths gives closer approximations to experimental estimates than semiparametric matching estimators do. This result is consistent across different metrics, empirical support region, number of neighbors and different specification for the parametric propensity score model. Second, we find that nonparametric adjusted-regression with nearest-neighbor matching produces the best results among all matching estimators. It relaxes the linearity of the regression function and the linearity of the adjusted factor implicit in other applications (e.g., Heckman 1998; Abadie and Imbens 2002). Third, the regression-based estimators provide mixed evidence about their ability to replicate experimentally determine results. They fail completely to replicate the treatment effects for two out of seven outcome variables. Four, the

regression-discontinuity approach proves the usefulness of econometric methods with underlying identification properties beyond the traditional methods that are widely used in the program evaluation literature. Finally, our results confirm the importance of having high quality data that allows the evaluator to construct "comparable" comparison groups using the same local labor market and the same survey instruments for both the treatment and the comparison groups.

References

A. Abadie, G. Imbens (2002): "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects", NBER technical Working Paper 283.

A. Abadie, D. Drukker, J. Leber, G. Imbens (2001): Implementing Matching Estimators for Average Treatment Effects in Stata, The Stata Journal 1, 1-18.

J. Angrist and V. Lavy (1999): "Using Maimondes Rule to Estimate the Effect of Class Size on Scholastic Achievement", *The Quarterly Journal of Economics*, CXIV: 533-569

B. Barnow, G. Cain, A. Goldberger (1980): "Issues in the Analysis Of Selectivity Bias", Institute for Research on Poverty, Discussion Paper, University of Wisconsin-Madison.

L. Bassi, (1984): "Estimating The Effects of Training Programs with Non-Random Selection", *The Review of Economics and Statistics*, Volume 66(1) 36-43.

E. Battistin and E. Rottore (2002) "Another Look at the Regression Discontinuity Design" (Manuscript)

M. Berger, D. Black, J. Smith (2000): "Evaluating Profiling as a Mean of Allocating Government Services" (Manuscript)

R. Berk and D. Rauma (1983): "Capitalizing on Nonrandom Assignment to Treatments: A Regression-Discontinuity Evaluation of a Crime-Control Program", *Journal of the American Statistical Association*, 78(381):21-27

D. Black, J. Smith (2003): "How Robust is the Evidence on the Effects of College Quality? Evidence from Matching. Journal of Econometrics (forthcoming).

D. Black, J. Smith, M. Berger, B. Noel (2003): "Is the Threat of Reemployment Services more effective than the services themselves? Experimental evidence from the UI system", *American Economic Review*, 93 (4) : 1313-1327.

S. Black (1999): "Do Better Schools Matter? Parental Evaluation of Elementary Education", *The Quarterly Journal of Economics*, CXIV: 577-599

L. Christofides, Q. Li , Z. Liu and I. Min (2003): "Recent Two-Stage Sample Selection Procedures with an Application to the Gender Wage Gap", *Journal of the American Statistical Association*, 21(433): 401-407.

R. Dehejia and S. Wahba (1998): "Propensity Score matching Methods for Non-Experimental Causal Studies", NBER Working Paper No 6829.

R. Dahejia and S. Wahba (1999): "Causal effects in Non-Experimental Studies: Re-evaluating the Evaluation of Training Programs", Journal of the American Statistical Association, 94, 1053-1062.

R. Dahejia (2003): "Practical Propensity Score Matching: A Reply to Smith and Todd", unpublished manuscript.

M. Frolich (2000): "Treatment Evaluation: Matching versus Local Polynomial Regression", Discussion Paper 2000-17, Universitat St Gallen.

M. Frolich (2004): "Finite Sample Properties of Propensity-Score Matching and Weighting Estimators" Review of Economics and Statistics, (Forthcoming).

T. Fraker and R. Maynard (1987): "The Adequacy of Comparison Group for Evaluations of Employment-Related Programs", *The Journal of Human Resources*, 22(2): 194-227.

D. Friedlander and P. Robins (1995): "Evaluating Program Evaluations: New Evidence on Commonly Used Non-experimental Methods", *The American Economic Review*, 85(4): 923-937.

J. Hahn, P. Todd and W. van der Klaauw (1999): "Evaluating the Effect of an Antidiscrimination Law Using a Regression- Discontinuity Design", NBER Working Paper No. 7131.

J. Hahn, (1998): "On the Role of The Propensity Score In Efficient Semiparametric Estimation of Average Treatment Effects", Econometrica 66(2) 315-331.

J. Heckman, H. R. LaLonde, J. Smith, (1999): "The Economics and Econometrics of Active Labor Programs", in O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics* Volume 3A, Amsterdam, 1865-2097.

J. Heckman, H. Ichimura, J. Smith, P. Todd (1998): "Characterizing Selection Bias Using Experimental Data", *Econometrica* 66 (5), 1017-1098.

J. Heckman, H. Ichimura, J. Smith, P. Todd (1998):"Matching as a econometric evaluation estimator", *The Review of Economics Studies*, Volume 65 (2), 261-694

J. Heckman, H. Ichimura, P. Todd (1997): "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme", *The Review of Economics Studies*, Volume 64 (4) , 605-654

J. Heckman, J. Hotz (1989): "Choosing Among Alternative Non-Experimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training", *Journal of the American Statistical Association*, 84 (408), 862-874

J. Heckman and R. Robb (1986): "Alternative methods for solving the problem of selection bias in evaluating the impact of treatment on outcomes" in R. Wainer ed. Drawing inferences from self-selected samples, 1986 Springer-Verlag New York Inc.

J. Heckman and J. Smith (1995): "Assessing the Case for Social Experiments", *The Journal of Economic Perspectives*, 9 (2), 85-100.

J. Heckman, J. Tobias and E. Vytlacil (2000): "Simple Estimators For Treatment Parameters In A Latent Framework With An Application To Estimating The Returns To Schooling", NBER Working Paper No 7950.

K. Hirano, G. Imbens, G. Ridder (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score", Econometrica 71(4) 1161-1189.

G. Imbens (2003): "Nonparametric Estimation of Average Treatment Effects under Exogeneity", NBER Technical Working Paper 294.

H. Ichimura, O. Linton (2001): "Asymptotic Expansions for Some Semiparametric Program Evaluation Estimators", The Institute of Fiscal Studies, Cemmap Working Paper CWP04/01.

G. Imbens (2003): "Sensitivity to Exogeneity Assumptions in Program Evaluation", American Economic Review Papers and Proceedings 93(2) 126-132.

P. Hall, J. Racine and Q. Li (2003): "Cross-Validation and the estimation of Conditional Distributions with both Categorical and Continuous Data, Working Paper.

J. Horowitz, W. Hardle (1996): "Direct Semiparametric Estimation of Single-Index Models with Discrete Covariates", Journal of the American Statistical Association, 91(436) 1632-1640.

M. Jones, J. Marron, S. Sheather (1996): "A Brief Survey of Bandwidth Selection for Density Estimation" *Journal of the American Statistical Association*, 91(433): 401-407.

W. van der Klaauw (2001) "Estimating the Effect of Financial Aid Offers on *College Enrollment: A Regression-Discontinuity Approach"*, *International* Economic Review (forthcoming)

R. Klein and R. Spady (1993) "An Efficient Semiparametric Estimator for Binary Response Models", *Econometrica*, 61(2): 387-421

R. LaLonde, (1986): "Evaluating the econometric evaluation of training programs with experimental data", *The American Economics Review*, 76(4) , 604-620

M. Lechner, (2000): "A Note on the Common Support Problem in Applied Evaluation Studies", SIAW, University of St. Gallen.

M. Lechner, (2000): "An Evaluation of Public-Sector-Sponsored Continuous Vocational Training Programs in East Germany", The Journal of Human Resources, 35(2) 347-375.

Q. Li, J. Racine (2003): "Cross-Validated Local Linear Nonparametric Regression", Statistica Sinica, (forthcoming).

Q. Li , J. Racine (2003): " Nonparametric Estimation of Distributions with Both Categorical and Continuous Data " , Journal of Multivariate Analysis, 86, 266-292.

Q. Li, J. Racine (2003): A Nonparametric Test for Equality of Distributions with Mixed categorical and Continuous Data " , Working Paper.

J. Marron (1988): "Automatic Smoothing Parameter Selection: A Survey", *Empirical Economics* Vol 13, pag187-208.

W. Newey (1988): "Two Step Estimation of Sample Selection Models" (Manuscript)

W. Newey, J. Powell and J. Walker (1990): "Semiparametric Estimation of Selection Models: Some Empirical Results", *The American Economic Review*, 80(2):324-328

P. Pettersson-Lindbom (2001): "Do Parties Matter for Fiscal Policy Choices? A Regression Discontinuity Design Approach" (Manuscript) .

J. Racine, Q. Li (2003): " Nonparametric Regression with Both Categorical Continuous Variables, Journal of Econometrics (forthcoming).

P. Robinson (1988): Root-N-Consistent Semiparametric Regression, Econometrica, 56(4), 931-954.

C. Robinson (1989): "The Joint Determination of Union Status and Union Wage Effects: Some Tests of Alternative Models", The Journal of Political Economy, 97(3) 639-667.

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

P. Rosenbaum and D. Rubin (1983): "The Central Role of The Propensity Score In Observational Studies For Causal Effects" *Biometrika* 70 (1): 41-55.

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

D. Rubin (1979): "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies", *Journal of the American Statistical Association*, 74(366): 318-328

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

D. Rubin and N. Thomas (1992): "Characterizing the Effects of Matching Using Linear Propensity Scores with Normal Distributions", *Biometrika*, 79(4):797-809.

D. Rubin and N. Thomas (1996): "Matching Using Estimated Propensity Scores: Relating Theory to Practice", *Biometrics*, 52(1):249-264

B. Silverman (1986): "Density Estimation for Statistics and Data Analysis", London-New York, Chapman and Hall.

J. Smith, (2000): "A Critical Survey of Empirical Methods for Evaluating Active Labor Market Policies", Manuscript.

J. Smith, P. Todd (2003): "Does Matching Overcome LaLonde's Critique of Non-Experimental Estimators?, *Journal of Econometrics* (forthcoming)

J. Smith, P. Todd (2003): A Rejoinder, Unpublished Manuscript.

F. Vella (1992): “Simple Tests for Sample Selection Bias in Censored and Discrete Choice Model” , Journal of Applied Econometrics, 7, 413-421

F. Vella (1998): “Estimating Models with Sample Selection Bias: A Survey”, Journal of Human Resources, 33, 127-169.

J. Wooldridge (1994): “Selection Corrections for Panel Data Models under Conditional Mean Independent Assumptions”, Journal of Econometrics, 68, 115-132.

Table 1. Sample Means of Earnings and Demographic Characteristics
Kentucky WPRS Experiment, October 1994 to June 1996

	Mean Values for Covariates[A]			Nonparametric test for differences in distributions for experimental sample[B]	Nonparametric test for differences in distributions for non-experimental sample[B]	
Covariates	**(1) Treatment Group**	**(2) Control Group**	**(3) Comparison Group**	**(1) – (2) p-value**	**(1) – (3) p-value**	**(1) – (3) Standardized difference[C] (%)**
Age	37.07	36.99	36.67	0.86	0.62	3.47
	(11.05)	(10.86)	(11.51)			
Education level	12.40	12.28	11.80	0.78	0.00	19.5
	(2.62)	(2.28)	(3.51)			
Annual earnings before program	19,046	19,758	16,553	0.001	0.00	19.0
	(13,636)	(13,676)	(12,754)			
One quarter before program	4,555	5,008	3,984	0.85	0.02	15.2
	(3,815)	(4,072)	(3,653)			
Two quarters before program	4,461	4,680	3,862	0.16	0.00	16.1
	(3,832)	(3,745)	(3,586)			
Three quarters before program	4,898	4,967	4,179	0.26	0.00	19.6
	(3,789)	(3,514)	(3,519)			
Four quarters before program	5,131	5,102	4,506	0.28	0.49	17.4
	(3731)	(3,608)	(3,404)			
White female (%)	37.21	35.16	36.40	0.37	0.63	1.68
White male (%)	51.77	56.37	53.93	0.04	0.18	-4.30
Black female (%)	5.50	4.29	4.40	0.24	0.08	5.04
Black male (%)	5.09	3.89	4.86	0.21	0.72	1.08

[A] Sample standard deviation in parenthesis.

[B] It uses the empirical distribution of 499 bootstrap statistics I_i to approximate the null distribution of I where $I = \int [\, f(x) - g(x)]^2 \, dx$

[C] The standardized difference is the mean difference as a percentage of the average standard deviation: $100(\bar{x}_T - \bar{x}_C) / [(s_T^2 + s_C^2)/2]^{1/2}$ where $\bar{x}_T$ and $\bar{x}_C$ are the sample means for each variable in the treatment group and comparison group and, s_T^2 and s_C^2 are the sample variances in both groups

Table 2.
Least-Square Cross-Validation Selected Smoothing Parameters[A]
Kentucky WPRS Experiment, October 1994 to June 1996

Covariate	h^{opt}	*Upper Bound[B]*
Sex	0.245	0.5
Schooling	0.905	1.00
Race	0.141	0.5
Region of residence	0.018	0.75
Profiling scores	0.001	0.94
Age	1.000	1.00
Labor market transitions	0.485	0.75
Past annual earnings	6871	∞

[A]The optimal smoothing parameter is obtained by least-square minimization of the cross-validation function $CV(h) = \int \{\frac{1}{n}\sum_{i=1}^{n} K_m(z_i, z, h) - f(z)\}^2 \cdot$

[B]The upper bound for categorical data is defined by $h^+ = (c-1)/c$.

Table 3.
Coefficient Estimates and p Values from Participation Probit Model
Dependent Variable:1 for Treated Units, 0 for Comparison Units
Kentucky WPRS Experiment, October 1994 to June 1996

Variable	Coefficient	Std. Error	p-Value[A]
Intercept	-1.460	0.325	0.000
Profiling scores	-0.199	0.037	0.000
Profiling scores2	0.016	0.001	0.000
Appalachian Area[B]	-0.387	0.066	0.000
Metropolitan Area	-0.101	0.054	0.060
RMSA Area	0.415	0.079	0.000
Sex	0.004	0.038	0.907
Age	0.004	0.009	0.664
Age2	-0.000	0.000	0.645
High School & Vocational[B]	0.381	0.234	0.104
High School & GED	-0.053	0.127	0.677
Some College	0.448	0.140	0.001
Bachelor	0.206	0.086	0.017
Master	0.266	0.199	0.182
Ph.D.	0.655	0.366	0.074
Race (1=white, 0 otherwise)	-0.108	0.059	0.066
Previous Annual earnings	-0.000	0.000	0.052
Employed $\rightarrow$ Employed[B]	-0.072	0.114	0.523
No Employed $\rightarrow$ No employed	-0.133	0.416	0.749
Employed $\rightarrow$ No Employed	-0.154	0.128	0.228

[A]Maximum likelihood probit estimation. Reported p-values are for two-tailed test of the null hypothesis that the true coefficient equals zero.
[B]The omitted category for region of residence is west; the omitted schooling category is less than high school; the omitted category for labor market transitions is No Employed $\rightarrow$ Employed.

Table 4.
Confusion Matrix and Classification rates for Propensity Score Models[A]
Kentucky WPRS Experiment, October 1994 to June 1996

Kernel Propensity Score Model		
	Treated	Comparison
Treated	927	295
Comparison	1798	6831
% Correct	78.7%	
% CCR (0)	79.1%	
% CCR (1)	75.8%	

Parametric Propensity Score Model		
	Treated	Comparison
Treated	874	375
Comparison	2603	6026
% Correct	69.9%	
% CCR (0)	69.8%	
% CCR (1)	71.1%	

[A] $P(x)>0.14$ is used to predict T=1 and $P(x) \leq 0.14$ is used to predict T=0.

Table 5 .
Comparison of Average Treatment Effects on the Treated
Under Alternative Matching Estimators and Kernel Propensity Score[A]
Kentucky Experiment, October 1994 to June 1996.

Variables	Experimental Estimates	Nearest-neighbor matching without CS[B,C]	Nearest-neighbor matching with CS[B,C]	Regression-adjusted nearest neighbor matching with CS[B,C]	Local constant kernel matching with CS[C,D]	Regression-adjusted local constant kernel matching with CS[C,D]	Local linear kernel matching with CS[C,D]	Regression-adjusted local linear kernel matching with CS[C,D]
Weeks receiving UI	-2.24 (0.50)	-1.90 (0.56) [-15]	-2.08 (0.58) [-7]	-2.14 (0.56) [-4]	-2.14 (0.51) [-4]	-2.13 (0.49) [-4]	-2.07 (0.57) [-7]	-2.17 (0.57) [-3]
Amount of UI benefits	-143.1 (100)	-90.1 (106) [-37]	-117 (101) [-18]	-119 (101) [-16]	-143 (87.0) [0]	-153 (79.8) [7]	-144 (99) [0]	-165 (91.1) [15]
Annual earnings	1054 (588)	762 (646) [-27]	908 (490) [-13]	920 (469) [-12]	704 (621) [33]	645 (609) [-38]	373 (587) [-64]	386 (574) [-63]
1st quarter earnings	525 (192)	422 (223) [-19]	453 (166) [-13]	454 (162) [-13]	388 (202) [-26]	399 (198) [-24]	394 (170) [-24]	402 (166) [-23]
2nd quarter earnings	344 (161)	109 (189) [-68]	255 (144) [-25]	246 (138) [-28]	61.2 (174) [-82]	70.3 (174) [-79]	38.0 (163) [-88]	43.0 (154) [-87]
3rd quarter earnings	220 (181)	164 (175) [-25]	145 (153) [-34]	147 (149) [-33]	88.3 (195) [-59]	71.8 (193) [-67]	13.4 (200) [-93]	8.10 (193) [-96]
4th quarter earnings	-35.6 (176)	66 (173) [-285]	54.7 (180) [-253]	59.7 (28.7) [-267]	-91 (192) [160]	-53 (193) [51]	-15.0 (196) [-57]	-20.5 (192) [-42]

[A]Bootstrap standard errors are shown in parenthesis. They are based on 50 repetitions. Non-experimental bias $[(\Delta_{TT}^{non-\exp} - \Delta_{TT}^{\exp})/\Delta_{TT}^{\exp}]100$ is shown in brackets

[B]The nearest-neighbor matching is estimated using Mahalanobis metric including the propensity score.

[C]The empirical common support region is defined by the *kth* criterion: we drop all the treated units with propensity scores higher that the 15th highest score for the comparison units.

[D]A weighted MSE with leave-one-out cross-validation is used to find the optimal smoothing parameters.

.

Table 6 .
Comparison of Average Treatment Effects on the Treated
Under Alternative Matching Estimators and Parametric Propensity Score[A]
Kentucky Experiment, October 1994 to June 1996.

Variables	Experimental Estimates	Nearest-neighbor matching without CS[B,C]	Nearest-neighbor matching with CS[B,C]	Regression-adjusted nearest neighbor matching with CS[B,C]	Local constant kernel matching with CS[C,D]	Regression-adjusted local constant kernel matching with CS[C,D]	Local linear kernel matching with CS[C,D]	Regression-adjusted local linear kernel matching with CS[C,D]
Weeks receiving UI	-2.24 (0.50)	-2.04 (0.48) [-9]	-1.99 (0.51) [-11]	-2.01 (0.50) [-10]	-2.17 (0.48) [-3]	-1.91 (0.42) [-59]	-0.76 (0.49) [-66]	-0.71 (0.47) [-68]
Amount of UI benefits	-143.1 (100)	-97.4 (77.6) [-32]	-145 (91.9) [1]	-140 (91.6) [-2]	-179 (97.1) [25]	-154 (86.2) [6]	-482 (97.6) [236]	-363 (95.5) [153]
Annual earnings	1054 (588)	949 (452) [-10]	627 (550) [-40]	629 (536) [-40]	168 (513) [-84]	120 (512) [-88]	31.3 (578) [-97]	-14.0 (553) [-101]
1st quarter earnings	525 (192)	490 (156) [-6]	439 (152) [-16]	436 (151) [-17]	285 (151) [-45]	268 (149) [-49]	141 (165) [-73]	115 (161) [-78]
2nd quarter earnings	344 (161)	174 (135) [-49]	100 (179) [-71]	90.9 (175) [-73]	-2.53 (158) [-100]	-20.3 (157) [-106]	52 (163) [-84]	60.7 (156) [-82]
3rd quarter earnings	220 (181)	157 (147) [-28]	53.8 (170) [-75]	54.3 (167) [-75]	-51.6 (167) [-123]	-67.3 (167) [-130]	-36.4 (177) [-116]	-63.4 (-58.5) [-128]
4th quarter earnings	-35.6 (176)	126 (173) [-453]	33.9 (187) [-195]	59.7 (166) [-267]	-111 (147) [217]	-77 (170) [120]	-95.6 (187) [165]	-58.0 (172) [64]

[A]Bootstrap standard errors are shown in parenthesis. They are based on 50 repetitions. Non-experimental bias $[(\Delta_{TT}^{non-\exp} - \Delta_{TT}^{\exp})/\Delta_{TT}^{\exp}]100$ is shown in brackets

[B]The nearest -neighbor matching is estimated using Mahalanobis metric including the propensity score.

[C]The empirical common support region is defined by the *kth* criterion: we drop all the treated units with propensity scores higher that the 15th highest score for the comparison units.

[D]A weighted MSE with leave-one-out cross-validation is used to find the optimal smoothing parameters.

.

Table 7 .
Sensitivity of the Non-Experimental Bias under Variations in the Nearest-Neighbor Set
Propensity Score Matching Average Treatment Effect on the Treated
Kentucky Experiment, October 1994 to June 1996.

Variables	Kernel Propensity Score		Parametric Propensity Score	
	Average Bias (%)[A]	Coefficient of Variation[B]	Average Bias (%)[A]	Coefficient of Variation[B]
Annual Earnings	23	21	60	6
1st quarter earnings	17	5	26	3
2nd quarter earnings	31	32	123	11
3rd quarter earnings	33	31	60	4
4th quarter earnings	151	45	174	59

[A]The non-experimental average bias is define as $(1/5)\sum_{j=1}^{5}[(\Delta_{TT_j}^{non\text{-}exp}-\Delta_{TT}^{exp})/\Delta_{TT}^{exp}]100$, where j represents the number of neighbors used in the estimation of the average treatment effect on the treated.

[B]The coefficient of variation is the sample standard deviation of j non-experimental bias.

Table 8.
Comparison of Average Treatment Effects on the Treated
Under Alternative Specification for the Probit Model and Nearest-Neighbor Matching[A]
Kentucky Experiment, October 1994 to June 1996.

Variables	Experimental estimates	Kernel Propensity Score Model[B]	Best Parametric Score Model[B,C]
Annual earnings	1054	669	533
	(588)	(665)	(565)
		[-36]	[-47]
1st quarter earnings	525	476	375
	(192)	(186)	(196)
		[-9]	[-29]
2nd quarter earnings	344	202	95
	(161)	(179)	(165)
		[-41]	[-72]
3rd quarter earnings	220	-12.5	82.8
	(181)	(219)	(180)
		[-105]	[-62]
4th quarter earnings	-35.6	-40	-29
	(176)	(176)	(192)
		[14]	[-17]

[A]Bootstrap standard errors are shown in parenthesis. They are based on 50 repetitions. Non-experimental bias $[(\Delta_{TT}^{non-\exp} - \Delta_{TT}^{\exp})/\Delta_{TT}^{\exp}]100$ is given in brackets.

[B]The nearest-neighbor matching is estimated using the Euclidian metric including the propensity score. The empirical common support region is obtained using the kth criterion: we drop all the treated units with propensity scores higher that the 15th highest score for the comparison units.

[C]The best-predictor parametric model is build using the hit-and–miss strategy: (1) Minimization of the classification error where $\hat{P}(x) > E(T)$ predicts participation and $\hat{P}(x) < E(T)$ predicts nonparticipation; and (2) Inclusion of only statistical significant covariates.

Table 9.
Average Treatment Effects on the Treated
Regression-Based Estimators[A]
Kentucky Experiment, October 1994 to June 1996.

Outcomes	*Experimental Estimates*	*Parametric*	*Nonparametric with mixed data*	
		Training/ Evaluation	*Training/ Evaluation*	*Weighting by Propensity*
Weeks receiving UI benefits	-2.24 (0.50)	-1.66 ( ) [-25]	-1.44 ( ) [-35]	-1.87 ( ) [-16]
Amount of UI benefits.	-143.1 (100)	180 ( ) [-225]	155 ( ) [-208]	14.9 ( ) [-110]
Annual earnings.	1054 (588)	1043 ( ) [-1]	976 ( ) [-7]	792 ( ) [-24]
1st quarter earnings.	525 (192)	470 ( ) [-10]	458 ( ) [-12]	465 ( ) [-11]
2nd quarter earnings.	344 (161)	211 ( ) [-38]	200 ( ) [-41]	119 ( ) [-65]
3rd quarter earnings.	220 (181)	206 ( ) [-6]	184 ( ) [-16]	120 ( ) [-45]
4th quarter earnings.	-35.6 (176)	155 ( ) [-542]	117 ( ) [-434]	87.3 ( ) [-349]
n	1981	9851	9851	9851

[A]Bootstrap standard errors are shown in parenthesis. Non-experimental bias $[(\Delta_{TT}^{non-\exp} - \Delta_{TT}^{\exp})/\Delta_{TT}^{\exp}]100$ is given in brackets.

Table 10
Average Treatment Effects on the Treated
Regression-Discontinuity Design: Weighted Mean Differences[A]
Kentucky Experiment, October 1994 to June 1996.

Outcomes	***Experimental***	***Bandwidth***		
	Estimates	***+/- 1***	***+/- 2***	***+/- 3***
Weeks receiving UI benefits.	-2.24 (0.50)	-1.68 (0.58) [-25]	-1.71 (0.51) [-23]	-1.67 (0.49) [-25]
Amount of UI benefits.	-143.1 (100)	182 (117) [-227]	183 (104) [-227]	216 (99) [-251]
Annual Earnings.	1054 (588)	1495 (653) [41]	1490 (587) [41]	1445 (558) [37]
1st quarter earnings.	525 (192)	532 (220) [1]	586 (187) [11]	567 (178) [8]
2nd quarter earnings.	344 (161)	361 (179) [5]	332 (175) [-3]	320 (165) [-7]
3rd quarter earnings.	220 (181)	325 (201) [47]	340 (178) [54]	310 (170) [40]
4th quarter earnings.	-35.6 (176)	274 (195) [-882]	231 (175) [-760]	247 (166) [-805]
n	1981	2306	3012	3465

[A]Standard errors are shown in parenthesis. Non-experimental bias $[(\Delta_{TT}^{non-exp} - \Delta_{TT}^{exp})/\Delta_{TT}^{exp}]100$ is given in brackets .

Table 11.
Average Treatment Effects on the Treated
Regression-Discontinuity Design: Parametric Fixed-Effect Model[A]
Kentucky Experiment, October 1994 to June 1996.

Outcomes	***Experimental***	***Bandwidth***		
	Estimates	***+/- 1***	***+/- 2***	***+/- 3***
Weeks receiving UI benefits.	-2.24 (0.50)	-1.74 (0.42) [-22]	-1.72 (0.37) [-23]	-1.63 (0.35) [-27]
Amount of UI benefits.	-143.1 (100)	177 (82.6) [-223]	209 (71.7) [-246]	255 (67.8) [-278]
Annual Earnings.	1054 (588)	1236 (478) [17]	1140 (425) [8]	1045 (403) [0.8]
1st quarter earnings.	525 (192)	477 (166) [-9]	493 (137) [-6]	457 (128) [-12]
2nd quarter earnings.	344 (161)	320 (131) [-6]	262 (129) [-23]	227 (122) [-34]
3rd quarter earnings.	220 (181)	258 (148) [17]	246 (129) [11]	207 (123.3) [-6]
4th quarter earnings.	-35.6 (176)	180 (143) [-614]	137 (128) [-491]	153 (121) [-537]
n	1981	2306	3012	3465

[A]Standard errors are shown in parenthesis. Non-experimental bias $[(\Delta_{TT}^{non-\exp} - \Delta_{TT}^{\exp})/\Delta_{TT}^{\exp}]100$ is given in brackets .

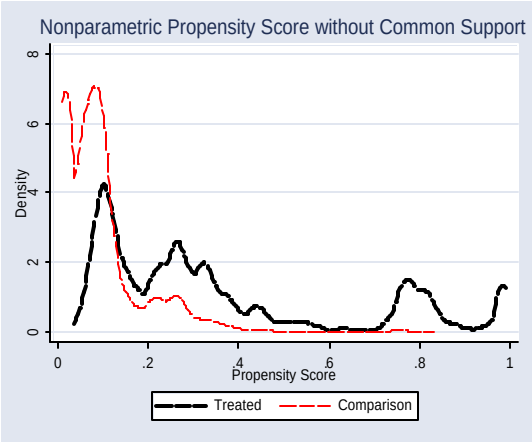
Table 12.
Average Treatment Effects on the Treated
Regression-Discontinuity Design: Parametric Fixed-Effect Model for all Units[A]
Kentucky Experiment, October 1994 to June 1996.

Outcomes	*Experimental Estimation*	*Specification for the Profiling Score*		
		Constant	*Linear*	*Quadratic*
Weeks receiving UI benefits.	-2.24 (0.50)	-1.65 (0.33) [-26]	-1.51 (0.36) [-32]	-1.47 (0.37) [-34]
Amount of UI benefits.	-143.1 (100)	245 (65.4) [-271]	171 (70.9) [-219]	126 (74.2) [-188]
Annual Earnings.	1054 (588)	1343 (369) [27]	904 (400) [-14]	708 (419) [-32]
1st quarter earnings.	525 (192)	502 (103) [-4]	441 (112) [-16]	378 (117) [-28]
2nd quarter earnings.	344 (161)	313 (111) [-9]	188 (120) [-45]	171 (126) [-50]
3rd quarter earnings.	220 (181)	264 (114) [20]	159 (124) [-27]	76.2 (130) [-65]
4th quarter earnings.	-35.6 (176)	262 (113) [-848]	114 (112) [-425]	82.2 (128) [-334]
n	1981	9851	9851	9851

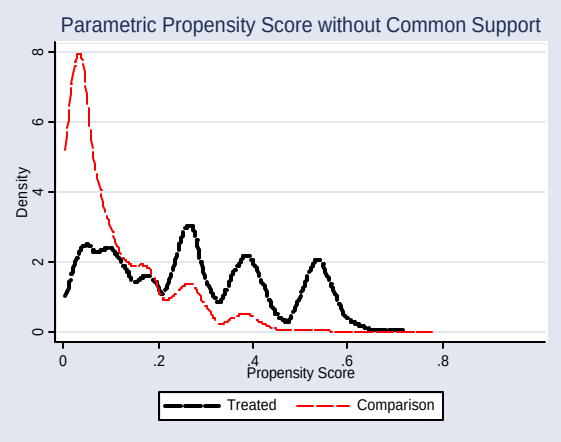
[A]Standard errors are shown in parenthesis. Non-experimental bias $[(\Delta_{TT}^{non-\exp} - \Delta_{TT}^{\exp})/\Delta_{TT}^{\exp}]100$ is given in brackets .

Figure 1

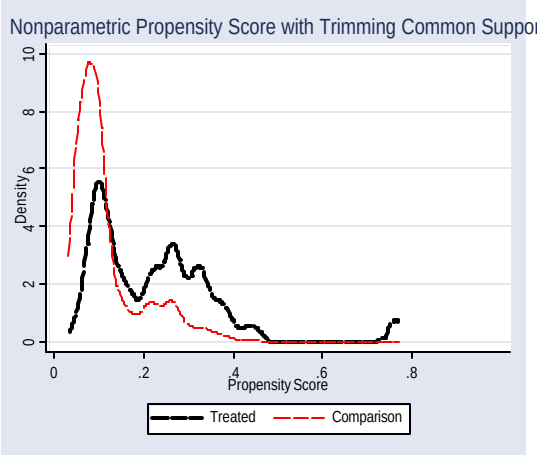
Panel A



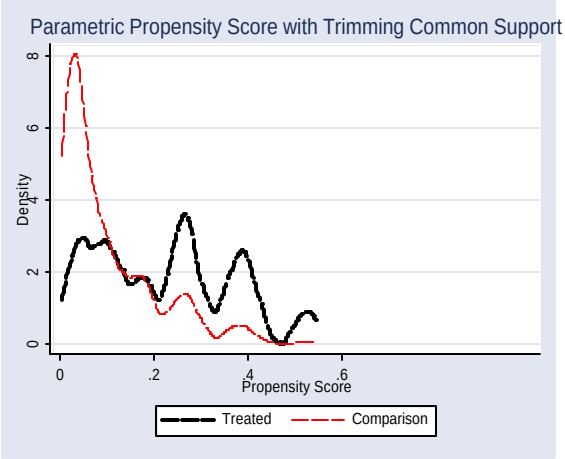
Panel B



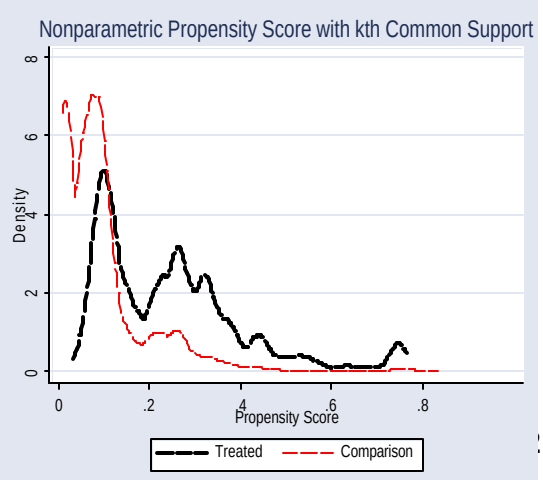
Panel C



Panel D



Panel E



Panel F

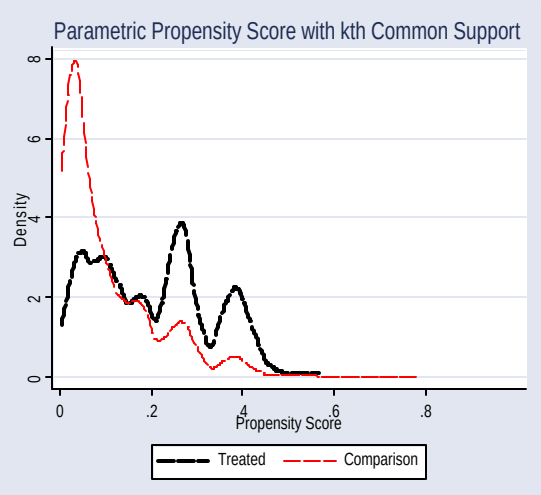
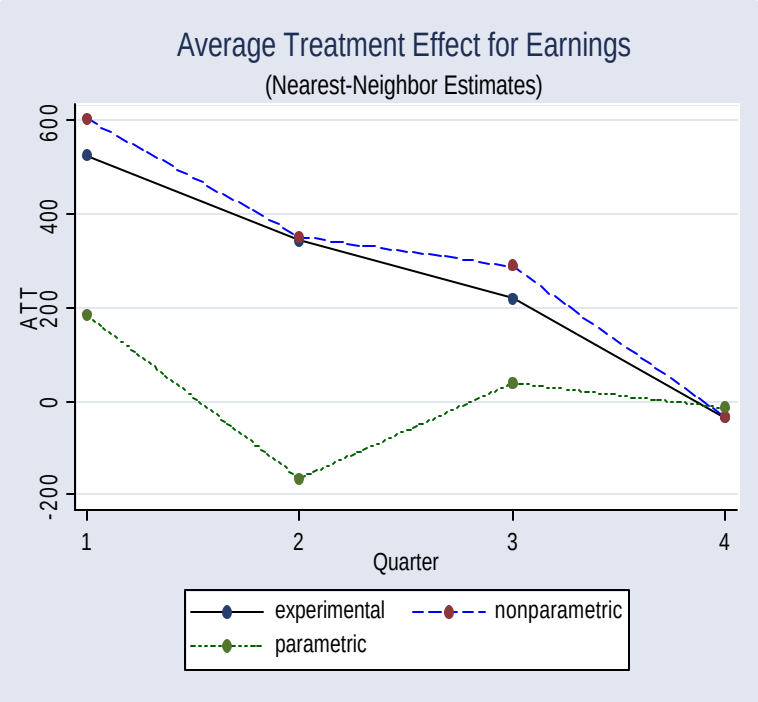


Figure 2


Average Treatment Effect for Earnings
(Nearest-Neighbor Estimates)
ATT
-200
0
200
400
600
1
2
3
4
Quarter
experimental
nonparametric
parametric