Treatment Effects for Profiling Unemployment Insurance Programs: Semiparametric Estimation of Matching Models with Fixed Effects

Jose Galdo*
Carleton University and IZA
jose_galdo@carleton.ca

Abstract

This study assesses matching estimators with fixed effects that first removes space and time effects before proceeding to apply standard matching based on participant characteristics. I use data from a novel profiling unemployment insurance program that allows the identification of both experimental and nonexperimental samples within the same local offices and with comparable measures from common administrative data sources. The results show that matching methods perform consistently well. The estimated bias increases dramatically when the comparison groups are drawn from different local offices within the same state. Kernel-based propensity score models show significant better predictive performance than their counterpart parametric logit models. This feature, however, does not translate into lower bias estimates.

Keywords: propensity score matching, treatment effects, unemployment insurance.

JEL Classification: C13, C14.

* I am indebted to Dan Black and Jeff Smith for many discussions and suggestions on the topics of this paper. This study has also benefited from helpful comments from Ana Dammert, Carlos Flores, Jeff Kubik, Tobias Klein, Jeff Racine, and participants at seminars at SUNY Albany, Emory University, Carleton University, University of Waterloo, ITAM, and York University.

1. Introduction

Conflicting evidence on the performance of econometric estimators that are frequently used in the evaluation of social programs has produced a long-standing debate in the applied labor literature over the last two decades. This research literature is motivated by question of whether social programs can be reliably evaluated without randomized experimental data (Lalonde 1986). The basic finding is that the failure to compare "comparable" individuals, as well as the lack of good data, can explain the lack of reliability of nonexperimental methods in replicating experimental determined treatment impacts (e.g., Lalonde 1986; Friedlander and Robins 1995; Heckman, Ichimura, Smith, and Todd 1998; Dehejia and Wahba 1999, 2002; and Smith and Todd 2005a, b).

While most of this literature is based on data from prototypical job training programs, this paper extends this research to a new evaluation area: profiling unemployment insurance programs. I use experimental data combined with nonexperimental data from the Kentucky Working Profiling and Reemployment Services (hereafter WPRS) to learn which combinations of data and institutions lead particular econometric estimators to work well or poorly. The primary goal of this unemployment insurance (UI) program is to shorten the unemployment spells of those with higher probabilities of exhausting their unemployment insurance benefits by means of low-intensity reemployment services. Black, Smith, Berger, and Noel (2003) provide experimental evidence that this program has had a large impact that led to substantial increases in quarterly earnings ($525), as well as reductions in the amount (-$143) and duration (2.24 weeks) of the UI benefits.

This study focuses on a class of estimators called 'propensity scores matching' that aim to balance distributions of individual characteristics across groups, so that groups are similar except for the treatment they receive. There are three main advantages to this program design and data. First, the data satisfies all the criteria needed for reliable implementation of matching estimators (Heckman et al. 1998). Specifically, the Kentucky WPRS program identifies both experimental and nonexperimental comparison samples within the same local office and with comparable measures from common administrative data sources. Second, where most of the previous analyses have examined voluntary programs in which individuals self-select into the programs, the Kentucky WPRS program is a mandatory one. This feature minimizes the role of selection on unobservables that requires stronger identification assumptions. Third, selection into the program is solely based on observable variables (i.e., profiling scores, local offices and weeks), and thus, the implementation of evaluation strategies that assume "selection on observables" is fully justified.

This favorable assessment of the Kentucky WPRS data differs markedly from widely used out-of-state comparisons where the procedures for bringing the individuals into the samples – along with strong variations in labor-market conditions – causes the failure of the identifying conditions for the identification of the treatment impacts. Lalonde's (1986) famous data sets, for instance, vividly illustrate the case in which the disparate distribution of covariates between the experimental and nonexperimental samples is so severe that an overwhelming number of comparison individuals have virtually no use in an evaluation, leading to a high level of sensitivity to the estimates

along many dimensions.[1] The Kentucky WPRS data also improves over within-state evaluations of welfare programs in which the nonexperimental comparison samples are drawn from earlier cohorts of welfare recipients from the same local welfare offices or welfare recipients from other local offices in the same state (e.g., Friedlander and Robins 1995; Lee 2001; Michalopoulos et al. 2004). A potential problem with these refined comparisons is the difficulty of controlling for local labor market conditions between treated and comparison offices (i.e., counties) or for changes in these conditions over time.[2]

From a methodological standpoint, this study formulates and implements a new variant of matching with fixed effects that first removes space and time effects before proceeding to apply matching based on participant characteristics. The motivation for this approach comes from the observation that matching requires *X* covariates that are good enough to obtain conditional independence between the counterfactual outcomes and the treatment indicator, but that are not "too good" i.e. that predict participation perfectly (Heckman, Ichimura, and Todd 1997). This paradoxical behavior of matching estimators clearly emerges in the Kentucky WPRS data as the full interaction of profiling scores, local offices, and weeks can perfectly predict participation, making standard matching impossible because of the violation of the support and balancing conditions.

This study carefully compares matching estimates with fixed effects to the experimental estimates and conducts a large number of sensitivity analyses. In the course

[1] LaLonde (1986) (along with many others) compares experimental and nonexperimental estimates of the impact of the U.S. National Supported Work (NSW) Demonstration program by combining the experimental data with two comparison groups drawn from two major survey data sets – the current population survey (CPS) and the Panel Study of Income Dynamics (PSID). Smith and Todd (2005a, b) conclude that the identifying conditions for the matching methods do not hold in this context, contrary to the claims of widely cited research by Dahejia and Wahba (1999, 2002).

[2] Hollister and Hill (1995) address in detail the difficulties of controlling for differences in local conditions.

of this analysis, this paper implements the ridge estimator (Seifert and Gasser 1996, 2000) and compares its performance to that of the widely used local linear estimator (Heckman et al. 1998). The ridge estimator shares all asymptotic advantages with the local linear one but also has better finite sample properties. For instance, whenever the design becomes sparse or clustered (a common issue in program evaluation), the ridge estimator leads to more reliable estimation of the regression curve (Frölich 2004). Despite this strong evidence from the broader statistical literature, the ridge estimator does not yet form part of the standard toolkit in the applied treatment effects literature.

This paper also compares matching methods based on multivariate kernels to traditional logit propensity score models to assess the issue of misspecification of the propensity score model. To the best of my knowledge, the study by Li, Racine, and Wooldridge (2005) is the only one to implement kernel-based propensity scores within a class of weighting estimators proposed by Hirano, Imbens, and Ridder (2003). This aspect of the analysis contributes to the small econometric literature on the importance of various details in the implementation of matching estimators.

Four main results emerge from this analysis. First, this study finds strong support for matching models with fixed effects. The estimated treatment effects show low bias for all outcomes of interest independently of which local polynomial estimator, bandwidth-selection method, and empirical overlapping region are used. These results suggest that when matching methods are applied to high-quality data, they perform consistently well. Second, the ability of matching estimators to solve the evaluation problem worsens dramatically when the comparison groups are drawn from different local offices within the same state. This result might explain why matching estimators do not consistently

reduce bias substantially in carefully executed in-state evaluations. Third, kernel-based propensity scores models show significant better predictive performance than their counterpart parametric logit models. This nice feature, however, does not translate into lower bias estimates. This result shows the trade-off between the ability of the models to maximize the probability of successful prediction into treatment and their ability to balance the distribution of covariates. Fourth, the performance of the matching estimators is largely dependent on the outcome of interest. Overall, quarterly earnings – the variable with the highest variance – show the largest bias and the highest sensitivity across different dimensions.

The remainder of this paper is organized as follows. Section 2 describes the program and data. Section 3 discusses the methods used to generate and assess the nonexperimental estimates. Section 4 discusses the empirical findings. Section 5 explores the sensitivity of the estimates to certain robustness specifications. Section 6 offers some conclusions.

2. Research Design, Program Description, and Data

The potentially distortionary incentives that the UI system provides for workers are well known. The incentives motivate UI claimants to extend their unemployment spells beyond what they would be in the absence of UI benefits, either by subsidizing additional job searching or by subsidizing the consumption of leisure.[3] In the 1980s and early 1990s, demonstration projects conducted in New Jersey, Nevada, Minnesota, and Washington showed the efficacy of using statistical methods and administrative data to identify those

[3] See Mortensen (1970) for earlier work on job search models and UI as well as Ashenfelter (1978) and Moffitt and Nicholson (1982) for labor supply models and UI. Meyer (1995) documents spikes in the empirical hazard function as claimants approach the exhaustion of their UI benefits.

who are more likely to exhaust their UI benefits. Further evaluations from these demonstration projects showed that providing more intensive job search assistance to these individuals leads to reductions in the amount and duration of the UI benefits (Eberts, O'Leary, and Wandner 2002). In response to these events, President Clinton signed into law the Unemployment Compensation Amendments in November of 1993, which require that states establish and utilize a system of profiling to identify those claimants that would be likely to exhaust regular UI benefits and refer them to reemployment services.

In June 1994, the Commonwealth of Kentucky was selected as a prototype state for implementing the WPRS program. After identifying potential exhaustees of the UI benefits among new initial claimants, this program offers them mandatory reemployment services, such as job-training and job-search workshops, early in their spell. The services themselves can be viewed either as a valuable opportunity to learn new employment-related skills, or as an in-kind tax on the leisure of the UI claimants (see Black et al. 2003). Hence, the Kentucky WPRS program combines aspects of two prototypical UI reforms that aim to reduce the incentives for excess benefit receipt without either punishing workers for whom a longer search is optimal (e.g., Illinois bonus experiment) or by enforcing job search requirements (e.g., Connecticut experiment).[4]

The Center for Business and Economic Research (CBER) at the University of Kentucky took responsibility for developing and predicting the fraction of their 26 weeks of UI benefits that claimants would use up. The model was estimated by employing five years of claimant data obtained from the Kentucky unemployment insurance mainframe

[4] See Woodbury and Spiegelman (1987) for a detailed analysis of the Illinois Bonus experiment as well as Ashenfelter, Ashmore, and Deschenes (1999) for evidence about work search enforcement programs.

computer databases, supplemented with data from other administrative data sources.[5] Two main features distinguish the Kentucky model from prior prototypical profiling models implemented in other states (see Eberts et al. 2002). First, the dependent variable is not represented by a dichotomous variable indicating whether the claimant exhausted UI benefits but rather uses the fraction of benefits received as a continuous variable. Second, the Kentucky WPRS model relies on more than 140 covariates, including past earnings, schooling, past job characteristics, prior UI benefit receipt, prior welfare receipt, industry and occupation, and local economic and labor market conditions.[6] With these data, a double-limit tobit model was implemented resulting in monotonic increases in the weekly benefit amount, months of job experience, and the previous year's earnings as the fraction of benefits exhausted increases. Most important, the richness of the data yields significant gains in predictive power with respect to profiling models from other states (Berger, Black, Chandra, and Allen 1997).

This profiling score is collapsed into a discrete score ranging from 1 to 20. Claimants predicted by the model to exhaust between 95 and 100 percent of their unemployment benefits receive a score of 20; those predicted to exhaust between 90 and 95 percent of their unemployment benefits receive 19; and so on. For each local employment office in each week, claimants starting new spells are ranked by their assigned scores. Those individuals with the highest scores are the first to be selected for reemployment services, and this process continues until the number of slots available for each office in each week is reached. Those claimants selected to receive reemployment services are contacted via mail to inform them about their rights and responsibilities

[5] Enhanced National Data System (ENDS), U.S. Department of Labor ES-202 database, U.S. 1990 Census.
[6] It is against the law to profile based on ethnicity, age, sex, or veteran status.

under the program. Due to the fact that many selected claimants may leave the UI system before receiving services but after being required to receive services, the Kentucky WPRS treatment can be thought of as the requirement to receive reemployment services (see Black et al. 2003 for further discussion).

If the maximum number of claimants who will receive reemployment in a given local office and in a given week is reached, and if there are two or more claimants with the same discrete profiling score, then a random number generator assigns the appropriate number of claimants to treatment. Therefore, only claimants with marginal profiling scores – the one at which the capacity constraint is reached in a given week and in a given local office – are randomly assigned into experimental treated and control groups. Black et al. (2003) call these marginal sets of claimants "profiling tie groups", or PTGs. This design differs from typical experimental evaluations wherein all eligible program applicants are randomly assigned. Those claimants with scores below the marginal scores are, by design, denied treatment, and they represent the nonexperimental comparison group.

From June 1994 to October 1996, the period for which we currently have data, 1,236 and 745 claimants are in the experimental treated and control groups, representing 286 PTGs ranging in size from 2 to 54.[7] For the same period, 9,032 claimants fell into the nonexperimental comparison group. I then combine the treated individuals from the PTGs to untreated individuals from the nonexperimental comparison group to form the nonexperimental data.

[7] The combination of 87 weeks and 32 local offices give 2,742 potential PTGs. Empty cells, however, for many weeks and local offices give a final number of 286.

It is important to highlight that this program allows us to identify both experimental and nonexperimental samples without the need for resorting to "external" comparable groups. In this sense, one of the contributions of this paper is its reliance on high-quality data that put all individuals in the same local labor market; moreover, all socioeconomic, demographic, and labor information comes from comparable measures from common (administrative) data sources. Furthermore, the availability of administrative data minimizes the risk of randomization bias and attrition bias. These features of the data overcome one of the main criticisms of matching incomparable nonexperimental samples to experimentally determined samples (Smith and Todd 2005a).

Table 1 presents descriptive statistics for key pre-treatment covariates for each one of the samples after discarding individuals with missing information for any covariate of interest. The continuous profiling scores are 0.83 and 0.80 for the experimental treated and control individuals respectively and 0.58 for nonexperimental comparison ones. These results show the ability of the profiling model to select UI claimants into treatment. The large difference in quarterly earnings between the experimental and nonexperimental samples is remarkable. In particular, the nonexperimental comparison group individuals present lower quarterly earnings with respect to the other groups. This result seems counterintuitive since individuals with lower predicted probabilities of benefit exhaustion are supposedly individuals with relatively better labor market attachment. A plausible explanation is that poor individuals who work enough to qualify for UI do not stay unemployed very long. In terms of schooling and age, all groups are similar, having on average 12 years of schooling and an average age of 37. In addition, the table indicates

some variation in the percentage of females and blacks among the groups, although we do not observe statistically significant differences.

In order to determine whether the individuals from the experimental sample were drawn from the same population, we present in columns 4 and 5 the *p*-values for test of differences in means. Since there are as many experiments as PTGs, the test is based on a linear regression that includes a treatment dummy variable and PTGs. The *p*-values do not reject the null for all covariates. On the other hand, when the test is applied to the nonexperimental sample after conditioning on local office and week variables instead of PTGs, the null hypothesis is rejected for almost all covariates. Looking at the standardized differences in the last column, which show systematic differences in covariate distributions between treatment and comparison units, reinforces this result.

3. The Econometric Framework

3.1 Identification

Let $T_i \in \{1,0\}$ denote the treatment indicator that takes the value one when the individual is treated and is otherwise zero. Let Y_{1i} denote the potential outcome in the treated state and Y_{0i} the potential outcome in the untreated state. For the i^{th} individual, one wishes to know the treatment effect $\Delta_i = Y_{1i} - Y_{0i}$. The fundamental problem of evaluation is one of missing data because Y_{0i} is not observed for the treated individual. Alternatively, one might focus on the average treatment effect on the treated (*ATT*), $\Delta_{TT} = E(Y_{1i} - Y_{0i} \mid T = 1)$. Data on program participants identify $E(Y_{1i} \mid T = 1)$. The mean counterfactual outcome $E(Y_{0i} \mid T = 1)$, however, is missing and cannot be directly identified without invoking

further assumptions. Somehow, one has to rely on a comparison group to obtain information about the counterfactual outcome of the treated in the untreated state. A simple replacement of $E(Y_{0i} \mid T = 1)$ by $E(Y_{0i} \mid T = 0)$ does not solve the evaluation problem in an observational study because *T* may not be independent of Y_{0i}. This is where one needs to present a statistical model.

Let the potential outcomes model be written in an additively separable structure: $Y_{1i} = g_1(X_i) + U_{1i}$ and $Y_{0i} = g_0(X_i) + U_{0i}$, where X_i is a vector of observed random variables and (U_{1i}, U_{0i}) are not observed (by the analyst) random variables. The functions $g_1(.)$ and $g_0(.)$ are assumed to be sufficiently well-behaved in that the first two moments exist. Additive separability is not strictly required in conventional matching, but it plays an important role in the variant of matching with fixed effects later. I now suppress the subscript *i* except where necessary for clarity.

The decision rule for program participation follows the index function framework, $T^* = Z\gamma + U_T$, where $T = \mathbb{1}[T^* > 0]$, $\mathbb{1}[.]$ is the indicator function, *Z* is a set of observable random variables, and U_T is an unobserved random variable. The decision to participate in the program may be determined by a prospective participant, by a program administrator, or by both. No distinction between Z and X is needed in standard matching.[8] It is also assumed that (U_1, U_0, U_T) is unobserved *i.i.d.* with zero conditional means, and the random variable U_T may be a function of U_1 and U_0.

[8] The information set may include covariates that are normally considered "endogenous" when implementing other estimators. Heckman et al. (1998), for instance, uses labor force transitions when estimating the probability of participation in a training program. These covariates are considered endogenous in the regression framework.

For the average treatment effect on the treated, matching proceeds by invoking the conditional independence assumption, $Y_0 \perp T \mid \{Z, X\}$, which assumes that selection into treatment is "on observables" (Rubin 1973) and can therefore be eliminated by conditioning on a rich set of observed covariates (Z, X). Matching also needs to invoke the common support condition, $\Pr(T = 1 \mid \{Z, X\}) < 1$, which guarantees the existence (at least in the population) of non-participants with the same values of (Z, X) as all of the participants. The inclusion of a high dimensional covariate set, however, can be impractical and lead to extremely slow convergence rates for any nonparametric estimator of $E(Y_0 \mid Z, X)$ (Pagan and Ullah 1999).

Rosenbaum and Rubin (1983) show that if the information sets justify matching on (Z, X), then they also justify matching on the propensity score $\rho = \Pr(T = 1 \mid X, Z)$, the probability of exposure to treatment conditional on observed covariates. Thus, the identifying assumption becomes $Y_0 \perp T \mid \rho$ that states that treatment exposure is unrelated to the counterfactual outcome for individuals sharing the same propensity score. This implies that $(Z, X) \perp T \mid \rho$, so individuals from either treatment group with the same propensity score are 'balanced' in that the distribution of (Z, X) is the same regardless of the treatment status.

Standard matching methods do not apply in the Kentucky WPRS context because the nature of the profiling system leads to the breakdown of support and balancing conditions. In this program, the treatment assignment is solely based on profiling scores, time (weeks), and space (local offices). Because there is no random overlap in the distribution of profiling scores for the experimental and nonexperimental samples for a given week in a given local office, the inclusion of these three variables along with a full

set of interactions among them in the estimation of the propensity score would predict *T* perfectly. Therefore, implementing standard matching methods would be possible only if the specification of the propensity scores model does not follow a flexible approach (e.g., omitting time-space variables or interactions between them). Heckman et al. (1997) as well as Dehejia and Wahba (1999) show that omitting important variables can significantly increase bias in resulting estimates.

Introducing the distinction between *X* and *Z* makes it possible to overcome the problem arising from perfect classification of treatment assignment for some values of (*X, Z*) if there are some variables *Z* not in *X*. The idea is to first remove space and time effects before proceeding to apply standard matching based on participants' characteristics. Let the potential outcomes be rewritten as partially linear regressions,

$$Y_{1jt} = g_1(\rho_{jt}) + \delta_{1j} + \beta_{1t} + U_{1jt},$$

$$Y_{0jt} = g_0(\rho_{jt}) + \delta_{0j} + \beta_{0t} + U_{0jt},$$

where (ρ, δ, β) is the information set that satisfies the conditional independence assumption, $\rho = P(T = 1 \mid X)$ is the propensity score based only on individuals' characteristics, δ_j ($j = 1, 2, ...m$) is the local office and β_t ($t = 1, 2, ..., n$) is the specific week in which each individual is selected into treatment. The model is inspired by the idea of differencing the effect of some variables, under the premise that ρ 's that are close will have corresponding values of the regression function that are also close.[9] Since the parameter of interest is the average treatment effect on the treated, I proceed by

[9] See Powell (1987) and Ahn & Powell (1993) for censored selection models as well as Honoré (1992) and Kyriazidou (1997) for panel data sample extensions.

differencing out the local office and week effects from Y_0 by using a pairwise difference approach in three steps.

In the first step, the local office with the largest number of observations, $\delta_{01} = 0$, is normalized, and the estimation of the *k* office effects, $\hat{\delta}_{0k}$, proceeds by taking weighted averages over outcomes for observable similar individuals within the subsample with *T=0* in local offices 1 and *k* who have approximately equal propensity scores in the same week,

$$Y_{0kt} - Y_{01t} = [g_0(\rho_{kt}) - g_0(\rho_{1t})] + [\delta_{0k} - \delta_{01}] + [\beta_{0t} - \beta_{0t}] + [U_{0kt} - U_{01t}] \qquad (1)$$

$$\cong \delta_k + U_{0kt} - U_{01t}.$$

The degree of similarity between two individuals is determined by the distance, based on some metric, between the observed covariates that constitute the matching variables. As there are m offices, the remaining *m-2* parameters are identified using the same approach. The local office effect $(\hat{\delta}_0)$ is obtained by averaging the $m-1$ individual effects, which is then differenced out from the outcome equation Y_0.[10]

The second step follows the same approach. After normalizing the week with the largest number of observations, $\beta_{01} = 0$, the ϖ -week effect is estimated by taking weighted averages over the new outcome $\tilde{Y}_0 = Y_0 - \hat{\delta}_0$ for observably similar individuals within the subsample with *T=0* in weeks 1 and ϖ who have approximately equal propensity scores,

$$\tilde{Y}_{0\varpi} - \tilde{Y}_{01} = [g_0(\rho_\varpi) - g_0(\rho_1)] + [\beta_{0\varpi} - \beta_{01}] + [\tilde{U}_{0\varpi} - \tilde{U}_{01}] \qquad (2)$$

[10] Although $E[Y_{0kt}-Y_{01t}]=0$, there is no reason to believe that the realization of U_{0kt} should be close to that of U_{01t}. It is only after averaging that these differences go away.

$$\cong \beta_{\varpi} + \tilde{U}_{0\varpi} - \tilde{U}_{01}.$$

As there are *n* different weeks, the time effect $(\hat{\beta}_0)$ is obtained by averaging the $n-1$ individual effects, which is then differenced out from the outcome equation $\tilde{Y}_0$.

In the third step, a new model free of time and local office effects emerges: $\tilde{\tilde{Y}}_0 = g_0(\rho) + \tilde{\tilde{U}}_0$, where $\tilde{\tilde{Y}}_0 = \tilde{Y}_0 - \hat{\beta}_0$ is the new outcome variable and $\tilde{\tilde{U}}_0 = \delta_{0j} + \beta_{0t} + U_{0jt} - \hat{\delta}_0 - \hat{\beta}_0$ is the error term with $E(\tilde{\tilde{U}}_0) = 0$. Standard matching based on participants' characteristics are implemented to form the expected counterfactual outcome for each treated unit. More formally, $\hat{g}_{0|T=1}(\rho) = \int g_0(\rho).f_{\rho|T=1}(\rho)d\rho$, where $g_0(\rho) = E(\tilde{\tilde{Y}}_0 \mid \rho)$ denotes the conditional mean function given non-participation and $f_{\rho|T=1}(\rho)$ denotes the distribution of ρ conditional on participation. Finally, the counterfactual outcome for each treated unit is defined by $\hat{Y}_0 = \hat{g}_{0|T=1} + \hat{\delta}_0 + \hat{\beta}_0$, which allows one to estimate the sample analog of the average treatment effect on the treated:

$$\hat{\Delta}_{TT} = (1/n_1)\sum_{i=1}^{n_1}(Y_1 - \hat{Y}_0).$$

Three features of the model need to be highlighted. First, this approach assumes that both local offices and time variables enter linearly in the model. If that is not the case, the proposed estimator may perform poorly. I then use the experimental Kentucky WPRS data as a benchmark against which to judge the performance of the proposed estimator. Second, this model implicitly allows both $\hat{\delta}$ and $\hat{\beta}$ to differ across *T=1* and *T=0*, although they are not estimated directly. Likewise, the estimated parameter is

equivalent to $\Delta_{TT} = g_1(\rho) + \delta_1 + \beta_1 - g_0(\rho) - \delta_0 - \beta_0$, which implies that $g(.), \delta$ and β can differ by treatment status.

Third, this variant of matching with fixed effects applies not only to situations in which the analyst has covariates (or combinations of them) that perfectly predict participation in the program but more generally when he or she has covariates that prevent fulfilling the balancing property of the propensity score matching methods. This situation is not uncommon in program evaluation. For instance, Eichler and Lechner (2002) report balancing problems in the distribution of a specific covariate (gender) when implementing the propensity score method. Michalopoulos et al. (2004) point outs that for some out-of-state comparisons balancing is not achieved and no attempt is made to use propensity score matching methods.

3.2 Estimation

3.2.1 Conditional Mean Functions

The literature suggests a wide variety of ways to estimate the conditional mean functions (semi) nonparametrically (see Heckman et al. 1998; Imbens 2004; Hirano Imbens and Ridder 2003). The focus of this paper is on propensity score matching methods.

In step 1, nearest neighbor with replacement estimates the conditional mean function (equation 1). A Euclidian metric on the propensity scores is used along with matching on exact weeks. This approach is less efficient (i.e., larger variance) than local polynomial matching (Frölich 2004), but it is pursued because of its simplicity in the estimation of the "nuisance" parameters. In step 2, nearest-neighbor matching on the propensity scores is also implemented to estimate equation 2. In step 3, local polynomial

matching is used in the estimation of the average treatment effects on the treated (Heckman et al. 1998; Smith and Todd 2005a, b). The conditional mean $[\text{of } \tilde{\tilde{Y}}_0]$ at a point $\rho = \rho_i$ readily follows as a weighted average of the data points in the comparison sample $\{\rho = \rho_j, \tilde{\tilde{Y}}_0 = \tilde{y}_{0j}\}$, where the weights depend upon ρ_j and the point ρ_i at which the conditional mean is evaluated. The conditional mean function $\hat{g}_{0|T=1}(\rho)$ equals θ_0 from the solution to the optimization problem:

$$\min_{\theta_0,\ldots,\theta_p} \sum_{j=1}^{n_0} (\tilde{\tilde{Y}}_{0j} - \sum_{m=0}^{p} \hat{\theta}_m (\rho_j - \rho_i)^m)^2 K(\frac{\rho_j - \rho_i}{h}),$$

where $(\theta_0, \theta_1, ..., \theta_p)$ denotes a vector of regression coefficients, *p* denotes the order of the local polynomial, $K(.)$ denotes a symmetric kernel function satisfying some standard properties, and *h* a bandwidth parameter. Fan, Gasser, Gijbels, Brockmann and Engel (1997) present the general solution to this problem. When $p = 0$, the resulting estimator corresponds to local constant kernel regression (called the Nadaraya-Watson estimator in statistics),

$$g^{LC}(\rho_i, \rho_j) = \frac{T_0}{S_0} \qquad (3)$$

where $T_r = \sum \tilde{\tilde{Y}}_{0j} K(\frac{\rho_j - \rho_i}{h})(\rho_j - \rho_i)^r$, and $S_r = \sum K(\frac{\rho_j - \rho_i}{h})(\rho_j - \rho_i)^r$. The corresponding local linear regression ($p = 1$) equals

$$g^{LL}(\rho_j, \rho_i) = \frac{T_0}{S_0} + (\rho_j - \rho_i)\frac{T_1}{S_2}. \qquad (4)$$

As discussed in Fan (1992), the local linear estimator converges faster near boundary points (a potentially important property in contexts with many estimated propensity

scores near zero or one) and appears more robust to different data designs. At the same time, the local linear estimator demands more of the data because it estimates one additional parameter in every local regression. Moreover, when ρ_i is far away from ρ_j or the data design becomes sparse or clustered, then the local linear estimator is unstable, leading to a high variance.[11]

To overcome the poor finite sample properties of the local linear estimator, Seifert and Gasser (2000) propose adding a ridge parameter, $R = \frac{5}{16} h \mid \rho_j - \rho_i \mid$, in the denominator of the local linear estimator such that the variance of the estimator becomes finite,

$$g^{RIDGE}(\rho_j, \rho_i) = \frac{T_0}{S_0} + (\rho_j - \rho_i)\frac{T_1}{S_2 + R}. \qquad (5)$$

The ridge estimator can be thought of as a weighted average of the local constant and local linear estimators:

$$g^{RIDGE}(\rho_j, \rho_i) = (1-\alpha)\frac{T_0}{S_0} + \alpha\left(\frac{T_0}{S_0} + (\rho_j - \rho_i)\frac{T_1}{S_2}\right),$$

where the weight $\alpha = S_2/(S_2 + R)$. When *R=0* then $\alpha = 1,$ and the ridge estimator becomes the local linear one. When $R \to \infty$ then $\alpha = 0,$ and the ridge estimator becomes the local constant estimator. Thus, the ridge estimator provides a compromise between the local constant estimator with finite variance but encounters problems with boundary bias and the local linear estimator with nice bias behavior but unbounded variance (Seifert and Gasser 2000). So far, the only evidence about the superior performance of

[11] Seifert and Gasser (1996) show that one needs at least four observations in the smoothing window to obtain a finite unconditional variance of the local linear estimator, but even with more points, there is no upper bound for the conditional variance.

the ridge estimator in the context of matching estimators constitutes Frölich's (2004) Monte Carlo analysis. Given the lack of a clear choice between these estimators in many applied contexts, I consider all of them in the empirical analysis later on.

The greater flexibility associated with local polynomial estimation of the conditional mean functions comes with a price: bandwidth selection. To avoid the excesses of bias or variance associated with a poor bandwidth choice, this paper implements a locally varying bandwidth approach that improves over conventional cross-validation methods by taking into account the location of the treated individuals in the selection of the optimal bandwidths. This approach selects a bandwidth for each treated individual according to the local density of the untreated ones, with narrower bandwidths in regions that are dense with untreated individuals and wider bandwidths in regions with few untreated individuals. Thus, this estimator adjusts itself to changes in the shape of the regression function, providing a good estimation of the counterfactual outcomes for high values of the propensity scores (see Galdo et al. 2007 for more details).[12]

3.2.2 Propensity Scores, Common Support Region, and Balancing Conditions

This study adopts a flexible logit parametric specification for the propensity scores, thus changing the overall procedure from a nonparametric to a semiparametric one. Balancing tests, as described in Dehejia and Wahba (2002), and Smith and Todd (2005b), for instance, guides the selection of the parametric specification for a given set of conditioning variables thought to satisfy the conditional independence assumption. Covariates influencing both the decision to participate in the Kentucky WPRS program

[12] The formula for the locally varying bandwidth is given by $h(\rho) = h_{cv}(\rho/(1-\rho))^{1/5}$, where h_{cv} denotes the bandwidth from conventional cross-validation and ρ the estimated propensity scores.

and potential outcomes of UI claimants should be included in the specification of the propensity scores. 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 analyses to approximate these categories include sex, education, age, and race. I also include earnings measures for the first and fourth quarters before the start of the UI spell to control for a potential earnings dip (Ashenfelter's Dip) and employment transitions between the first and fourth quarters before the start of the spell: Employed→ Employed, Employed→ Not Employed, Not Employed→ Employed, and Not Employed→ Not Employed, as suggested by Heckman and Smith (1999).

To analyze how the inclusion of week and local office covariates affect the support and balancing conditions of this baseline model, two alternative parametric logit models are estimated in which office and week dummy variables are added successively to the baseline model. Table 2 shows the overall classification rates for these models. The probability of successfully predicting treatment increases from 73 percent to 79 percent and then to 84 percent when successively adding local offices and weeks covariates to the estimation of the propensity score. This result is strongly related to the distribution of the propensity scores across the models. For the baseline model, the mean propensity score is 0.261 and 0.101 for treated and comparison samples respectively. Adding local office variables has the immediate effect of increasing (decreasing) the mean propensity score to 0.335 (0.091) for the treated (untreated) samples. Finally, the inclusion of dummy

variables for weeks i.e. the full model, further increases the mean score values for treated observations (0.45) and further decreases the mean score values (0.075) for the comparison observations. It is worth noticing that a full set of interactions among profiling scores-offices-weeks is not used because of the complete failure of the support and balancing conditions.

These particular data impose a clear trade-off between the richness of the set of covariates that may be included in the propensity score model and the balance of the pre-treatment covariates between the treated and comparison observations conditional on the propensity score. As Table 2 reveals, the baseline model shows the best balance across all parametric models when implementing the balancing test proposed by Rosenbaum and Rubin (1985).[13] The sample covariate differences between treatment and matched comparison groups range from 0.31 percent for education to 9 percent for profiling scores, with an absolute median value of 0.61 percent for all covariates appearing in the propensity scores. Successively adding offices and weeks into the estimation of the propensity score model dramatically worsens the quality of the covariate balance. The same patterns are observed when implementing Smith and Todd's (2005) parametric balancing test. Indeed, Table 2 provides an example of the conflicting relationship between the ability of the models to maximize the probability of successful prediction and their ability to balance the distribution of covariates between the treatment and comparison samples conditional on the propensity score model.

[13] The standardized difference in percent is $100*[\bar{x}_1 - \bar{x}_{0M}]/[(s_1^2 + s_0^2)/2]^{1/2}$, where for each covariate, $\bar{x}_1$ and $\bar{x}_{0M}$ are the sample means in the treated and matched comparison sample and s_1^2 and s_0^2 are the sample variances in the raw, treated, and comparison groups.

4. Matching Estimates

Panels A and B in Table 3 show the bias estimates associated to matching with fixed effects (on the baseline model) and standard matching (on the full model), respectively. Within each panel, the rows correspond to different dependent variables: weeks receiving UI benefits measured over the 52-week period starting in the first week of the UI claim, employment in the first quarter after the start of the UI spell, and earnings in the first quarter after the start of the UI spell. Each cell of the table is devoted to one estimate and shows the following information (from the top down): the estimated treatment effects, the bootstrapped standard error over 100 simulations of the data (in parentheses), and the bias (in brackets) measured as a percentage of the experimental program effects.[14] All estimates are based on observations within the support region defined by the 2-percent trimming method (Heckman et al. 1998). The first column presents the experimental estimates that follow from Black et al. (2003) after imposing the common support condition.[15]

Five main patterns emerge from Table 3. First, matching with fixed effects performs well on this particular data. The size of the biases is relatively small for all outcomes of interest. By looking at weeks receiving UI benefits, for instance, one observes that the biases range from 1 to 7 percent. The corresponding biases for employment range from 6 to 28 percent and from 11 to 25 percent for quarterly earnings. These results clearly show that comparing experimental and nonexperimental samples

[14] The bias is defined as $\left((\Delta_{TT_i}^{No-Exp} - \Delta_{TT}^{Exp}) / \Delta_{TT}^{Exp}\right) * 100$

[15] While not shown, using the full sample of experimental treated individuals yields similar qualitative conclusions.

within the same local office and with comparable measures from common data sources yield relatively small biases when implementing matching methods.[16]

Second, standard matching on the full propensity score model performs poorly. The size of the biases in most cases is substantially larger than those obtained from matching with fixed effects. For example, the bias for weeks receiving UI benefits ranges from 14 to 20 percent and from 44 to 52 percent for quarterly earnings. Likewise, the variance of the treatment effects estimates are also larger than those from matching with fixed effects, leading to the loss of statistical significance of the quarterly earnings pointwise estimates. The inclusion of the full set of covariates (i.e., scores, weeks, offices) in the propensity score specification causes problems in terms of higher variance, since a higher proportion of treated individuals are discarded from the analysis and the poor match quality (see Augurzky and Schmidt 2001 for a Monte Carlo analysis). This result reminds us that the main purpose of the propensity score approach is not to maximize the probability of successful prediction into treatment but to balance all covariates.[17]

Third, Table 3 reveals that in the Kentucky WPRS data, the estimated biases are not sensitive to the selection of the local polynomial matching estimator. In particular, one observes that the ridge estimator does not improve over the local linear estimator. Two factors explain this result. The estimates are based on relatively large sample sets

[16] Heckman et al. (1998), for example, report biases ranging from 88 to 670 percent in the analysis of the JTPA program. Smith and Todd (2005a) estimate biases of over 400 percent in the analysis of the NSW program. Michalopoulos et al. (2004) report short-run bias of 75 percent for in-state comparisons and 267 percent for out-of-state comparisons in the analysis of the Welfare-to-Work Strategies (NEWWS) using the propensity score subclassification estimator.

[17] Matching with fixed effects also outperforms conventional OLS estimation with fixed effects. The OLS specification relies on the same specification of the baseline propensity score model, plus a set of dummy variables for offices and weeks, and is estimated over the common support region. The resulting bias is 13 percent for weeks receiving UI benefits, 36 percent for employment in 1st quarter after UI spell, and 31 percent for earnings in 1st quarter after UI spell.

(about ten thousand observations) and, most importantly, high-quality data sets. This minimizes the risk of sparse or clustered designs that cause unreliable estimation of the regression curve with local linear estimators (Seifert and Gasser 1996). Moreover, kernel estimation of the conditional mean functions is based on locally varying bandwidths, which accounts for the location of the treated observations. This method selects larger bandwidths in areas of high propensity scores (sparse data regions), which also minimizes potential differences between the ridge and the local linear estimators.

Fourth, the largest differences between matching with fixed effects and standard matching occur when implementing the nearest-neighbor estimator. In fact, this particular estimator is very unstable and yields, as expected, the largest standard errors. Fifth, the finite sample performance of the matching estimators is affected by the variance of the outcome of interest. In general, one observes that the larger the variance of the outcomes, the larger the sensitivity of the estimated treatment effects along some dimensions. By looking at employment and quarterly earnings variables, for instance, one observes in the outcomes that have the highest variances that the size of the bias is the largest and varies greatly between the nearest neighbor and polynomial-matching estimators.

5. Sensitivity Analysis

Specification of the Propensity Scores

For these propensity score methods to produce consistent estimates, it is very important that the statistical model is correctly specified. To evaluate the sensitivity of the results to the specification of the propensity score, a kernel-based counterpart to the baseline parametric logit model is estimated following the work of Li, Racine, and Wooldridge

(2005). These authors show higher classification rates for the kernel-based propensity score method relative to the frequency approach, and argue efficiency gains in the estimation of treatment impacts within a class of weighting estimators proposed by Hirano, Imbens, and Ridder (2003).[18]

As the propensity scores can be thought of as a conditional probability density function, let define $\rho = \hat{f}(T \mid X) = \hat{f}(T_i, X_i) / \hat{f}(X_i)$, where $\hat{f}(T_i, X_i)$ is the joint density of T_i, X_i, and $\hat{f}(X_i)$ is the marginal density of X_i. The nonparametric propensity score estimator is defined by

$$\rho = \hat{f}(T \mid X) = \frac{\sum_{i=1}^{n} K(T_i, X_i, T, X, h_T, h_x)}{\sum_{i=1}^{n} K(X_i, X, h_x)} \qquad (6)$$

where $K(.)$ is a well-behaved kernel function that depends on the distribution of T_i, X_i and bandwidth parameters h_T, h_x. If *c* and *d* are the number of continuous and discrete regressors respectively then the numerator in equation (6) can be written as

$$K(T_i, X_i, T, X, h_T, h_X) = K(T_i, T, h_T) \prod_{j=1}^{c} K(X_{ij}, X_j, h_{xj}) \times \prod_{j=c+1}^{d+c} K(X_{ij}, X_j, h_{xj}),$$

and similarly the denominator can be written as

$$K(X_i, X, h_x) = \prod_{j=1}^{c} K(X_{ij}, X_j, h_{xj}) \times \prod_{j=c+1}^{d+c} K(X_{ij}, X_j, h_{xj}),$$

The Epanechnikov weight-assigning kernel function is used for the continuous covariates. The discrete covariates are considered categorical unordered variables and

[18] Todd (2002) as well as Kordas and Lehrer (2004) examine semiparametric estimation of the propensity scores.

follow the univariate Aitchison and Aitken's (1976) weight-assigning kernel function.[19] The selection of the optimal values $h_T, h_{1x}, \ldots h_{cx}, h_{(c+1)x}, \ldots h_{(c+d)x}$ follows cross-validation methods, which minimizes the sample version of the mean squared error of the prediction of participation into the program. Hall, Racine, and Li (2004) show that this method, applied in the context of multivariate "hybrid" product kernels, has the property of smoothing away categorical conditioning variables that are irrelevant in the estimation of the density function by assigning them large bandwidth parameters and, consequently, reducing them toward the uniform distribution.[20]

Figure 1 compares the distribution of scores between the baseline logit model and its kernel-based nonparametric version. One can observe that the distribution of the propensity scores values is somewhat different even though the correlation between the predicted probabilities from the two models is high (0.9). For both parametric and nonparametric models, the distribution of propensity score values has support over the entire [0, 1] distribution. The nonparametric model, however, exhibits higher probability mass in its tails than the parametric one. An examination of the in-sample predictions for both models (Table 2) reveals the nonparametric estimator does a better job of predicting for both participants and nonparticipants. In particular, the kernel-based model gives a rate of 83 percent correct predictions for treatment, whereas the parametric model correctly predicts 73 percent. These results indicate that the parametric logit model may suffer from misspecification.

---

[19] $K(x_i, x, h_x) = \begin{cases} 1 - h_x, & \text{if } x_i = x_j \\ \dfrac{h_x}{p-1}, & \text{if } x_i \neq x_j \end{cases}$. The range of h_x is [0,1], where h_x =0 represents an indicator function and h_x =1 a uniform weight function.

[20] The kernel-based propensity score model is estimated by using five discrete regressors (sex, age, schooling, race, and employment transitions) and three continuous regressors (profiling scores, 1st and 4th quarter earnings before program).

Does this better predictive performance translate into lower bias estimates? Table 4 shows the estimated bias with kernel-based propensity score. By looking at the local polynomial estimates, one observes that the estimated bias for both weeks receiving UI benefits (5 percent) and employment on first quarter after UI spell (30 percent) are somewhat similar to those emerging from parametric-based propensity scores models. On the other hand, the bias for quarterly earnings (65 percent) is consistently higher than the parametric one (25 percent). It is noteworthy that matching estimators show higher levels of sensitivity along the parametric-nonparametric dimension for the outcome with the highest variance (i.e., quarterly earnings).

Overall, the bias estimates show that, in the Kentucky WPRS data, nonparametric propensity score models do not improve over flexible logit models (either in terms of bias or standard errors) even though their higher predictive performance. This result is driven by the trade-off between the ability of the models to maximize the probability of successful prediction into treatment and their ability to balance the distribution of covariates. As columns 1 and 4 in Table 2 show, the match quality is better for the baseline logit model than for its counterpart kernel-based model. The absolute median value for the sample covariate differences between treatment and matched comparison groups increases from 0.61 (logit model) to 2.79 percent (kernel-based model). Likewise, the kernel-based propensity score model yields higher variances for the estimated treatment effects because a higher proportion of treated individuals are discarded from the analysis relative to the parametric-based propensity score model.

Furthermore, the selection of optimal bandwidths in the estimation of high-dimensional density functions is not a trivial issue (see Pagan and Ullah 1999). In

particular, the cross-validation property of smoothing away categorical conditioning variables that are irrelevant in the estimation of the density function can be counterproductive in the context of matching estimation because there is no support for the rule of selecting matching variables by choosing the set of variables that maximizes the probability of successful prediction into treatment or by including variables in conditioning sets that are statistically significant in binary choice models. Finally, the reader should bear in mind that this method requires nonparametric estimation of high-dimensional objects, which leads back to the "curse of dimensionality" problem.

Mismatch of Geography

One intriguing result in the evaluation literature is that in-state evaluations of training and welfare programs do not consistently reduced matching bias substantially even when one observes large improvements with respect to out-of-state comparisons (e.g., Michalopoulos 2004). The Kentucky WPRS data allows one to investigate further this issue because of the program's random design. I match experimental treated individuals from the manufacturing-based local offices in 'northern' Kentucky (i.e., Greater Louisville and Blue Grass regions) to two geographically misaligned nonexperimental comparison groups. The first comparison group is based on local offices located in the 'western' region of Kentucky where the economy is also driven by manufacturing businesses. The second comparison group is based on local offices from the less-developed Appalachian region that largely depends on mining. [21]

[21] Louisville, Fern Valley, Covington, Frankfort, Lexington, and Georgetown local offices compose the experimental subsample, which together represents 47 percent of the full experimental sample. The first nonexperimental comparison group comes from Bardstown, Elizabethtown, Glasgow, Bowling Green,

Table 5 shows the effect of geography on estimated bias using the baseline logit propensity score matching model with (weeks) fixed effects. The geographically aligned comparison group (column 2) shows the smallest bias with estimates close to those from the full sample set. On the contrary, the geographically misaligned comparison groups (columns 3 and 4) yield consistently large bias. By looking at the local constant matching estimates, one observes that the bias for week receiving UI benefits increases to 69 and 135 percent for the 'Western' and 'Appalachian' comparison groups, respectively.

Likewise, the bias for the employment outcome increases to 58 and 168 percent, respectively. This evidence indicates that the ability of matching estimators to solve the evaluation problem worsens dramatically even when the comparison groups are drawn from the same state but from different local welfare offices. It highlights the difficulty of controlling for local labor market conditions between treated and comparison offices and, at the same time, it might help to understand the relative poor performance of matching estimators in careful executed in-state evaluations of welfare programs (e.g., Michalopoulos 2004).

Alternative Empirical Support Regions and Bandwidth Selection Methods

The violation of the common support condition is a major source of evaluation bias in observational studies (Heckman et al. 1998). Thus, the definition of the empirical overlapping region is an important step in the implementation of matching estimators. As several analysts may have experienced, however, the number of units in or out of the overlapping region can vary depending on the definition employed, which adds a source

---

Campbellsville, Paducah, and Mayfield local offices; whereas the second one from Pikeville, Whitesburg, Harlan, Middlesboro, Corbin, Hazard, Prestonsburg, and Jackson local offices.

of sensitivity to the matching estimators that is not fully appreciated in the empirical literature. To address this potential issue, I impose a new empirical overlapping region that follows the "minima and maxima" approach (Dehejia and Wahba 1999).[22] Column 3 in Table 6 shows the new estimates for local constant matching using the baseline logit propensity score model. By comparing these estimates to those of the baseline estimation (column 2), one observes that the size of the treatment estimates change somewhat without affecting any of the qualitative conclusions of the paper. Once again, it is the bias associated with the outcome with the highest variance (i.e., quarterly earnings) that changes the most (15 percentage points).

Likewise, bandwidth selection has posed a problem for evaluation methods that rely on kernel regression (Frölich 2004). In the particular case of matching, the bandwidth affects the number of untreated units used to estimate the expected counterfactual outcome for each treated unit. Too large a bandwidth means including untreated units quite different from each treated unit in the estimation while too small a bandwidth means using only one or two untreated units for each treated unit, with noisy estimates being the result. Column 4 in Table 6 shows local constant matching estimates with a fixed (global) bandwidth rather than locally varying bandwidths, which is selected by standard cross-validation methods (Black and Smith 2004).[23] For all outcomes of interest, the treatment estimates are not sensitive to the imposition of alternative bandwidth-selection approaches. This result is just a manifestation of the good covariate

[22] The basic criterion of this approach is to drop all observations whose propensity score is smaller than the minimum and larger than the maximum in the opposite group. In the Kentucky WPRS data, the common support lies within the interval [0.004, 0.773].
[23] The grid for the bandwidth search equals [0.01, 0.04, 0.07,…, 0.49].

balance in the treated and comparison samples, conditional on the baseline parametric propensity scores model.

Alternative Fixed Effects Estimation

The estimation of the space and time effects may be sensitive to which covariate is first eliminated. To investigate this potential issue of order, column 5 in Table 6 re-examines the local constant matching estimates by reversing the order in which the space and time effects are eliminated. Now, the time effect is first removed, followed by the local offices effect. By looking at the new estimates, one concludes that matching with fixed effects is robust to the selection of which variable is eliminated first.

6. Concluding Remarks

The main implication of this study for evaluation research is clear. Matching methods perform consistently well in the Kentucky WPRS data because the program identifies both experimental and nonexperimental comparison samples within the same local office and with comparable measures from common data sources. These features largely improve over comparisons based on geographic mismatch or situations where the dependent variables are measured in different ways in the treated and comparison groups.

Overall, this study found strong support for matching with fixed effects that first removes space and time effects before proceeding to apply matching based on participant characteristics. The estimated treatment effects show much lower bias for all outcomes of interest than those emerging from standard matching approaches. Matching with fixed effects can be used not only to address situations in which the analyst has covariates (or

combinations of them) that perfectly predict participation in the program but more generally when he or she has some particular covariates that prevent fulfilling the balancing property of the propensity score matching methods.

Three main lessons emerge from the sensitivity analyses. First, the ability of the matching estimators to solve the evaluation problem worsens dramatically when the comparison groups are drawn from different local welfare offices within the same state. This result might explain why matching estimators do not consistently reduce bias substantially in carefully executed in-state evaluations. Second, the Kentucky WPRS treatment effects are very sensitive to the parametric/nonparametric specification of the propensity scores; in general, kernel-based propensity scores models show significant better predictive performance than their counterpart parametric logit models. This nice feature, however, does not translate into lower bias estimates. This result reinforces the vision that the main purpose of the propensity score estimation is not to maximize the probability of successful prediction into treatment but to balance all covariates. Third, the performance of the matching estimators is largely dependent on the outcome of interest. Overall, quarterly earnings – the variable with the highest variance – show the largest bias and the highest sensitivity across different dimensions. Unfortunately, there is little work on assessing the performance of matching methods to the changes in the distribution of the outcomes of interest. More research in this direction will be welcome.

References

Ahn, H. and J. Powell. 1993. "Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism," *Journal of Econometrics*, 58, 3-29.

Aitchison, J. and C. Aitken. 1976. "Multivariate Binary Discrimination by the Kernel Method," *Biometrika* 63(3): 413-420

Ashenfelter, O., D. Ashmore, O. Deschenes. 1999. "Do Unemployment Insurance Recipients Actively Seek Work? Randomized Trials in Four U.S. States," NBER WP No. W6982.

Ashenfelter, O. 1978. "Estimating the Effect of Training on Earnings," *Review of Economics and Statistics,* 1978, 60: 47-57.

Augurzky B. and C. Schmidt. 2001. "The Propensity Score: a Means to an End. Discussion Paper No. 271, IZA.

Berger, M., D. Black, A. Chandra, and S. Allen. 1997. "Kentucky's Statistical Model of Working Profiling for Unemployment Insurance," Kentucky Journal of Economics and Business, 16: 1-18.

Black. D., J. Galdo, and J. Smith. 2006. "Estimating the Bias of the Regression Discontinuity design Using Experimental Data," Unpublished Manuscript.

Black D., and J. Smith. 2004. "How robust is the evidence on the effects of the college quality? Evidence from Matching." *Journal of Econometrics*, 121(1):99-124.

Black, D., 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.

Dehejia, R. and S. Wahba. 2002. "Propensity Score Matching Methods for Nonexperimental Causal Studies," *Review of Economics and Statistics* 84(1): 151-161.

Dehejia, R. and S. Wahba. 1999. "Causal effects in Nonexperimental Studies: Re-evaluating the Evaluation of Training Programs," *Journal of the American Statistical Association,* 94: 1053-1062.

Eberts, R., C. O'Leary, and S. Wandner eds. 2002. "Targeting Employment Services," Kalamazoo, MI: W.E Upjohn Institute for Employment Research.

Eichler M. and M. Lechner. 2002. "An Evaluation of Public Employment Programmes in the East German State of Sachsen-Anhalt. *Labour Economics* 9: 143-186.

Fan, J. T. Gasser, I. Gijbels, M. Brockmann, and J. Engel. 1997. "Local Polynomial Regression: Optimal Kernels and Asymptotic Minimax Efficiency," *Annals of the Institute of Mathematical Statistics* 49: 79-99.

Fan, J. 1992. "Design-Adaptive Nonparametric Regression," *Journal of the American Statistical Association,* 87: 998-1004.

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

Frölich, M. 2004. "Finite Sample Properties of Propensity-Score Matching and Weighting Estimators," *The Review of Economics and Statistics*, 86(1): 77-90.

Galdo, J., J. Smith, and D. Black. 2007. "Bandwidth Selection and the Estimation of Treatment Effects with Unbalanced Data," IZA WP#3095.

Hall, P., J. Racine and Q. Li. 2004. "Cross-Validation and the Estimation of Conditional Probability Densities," *Journal of the American Statistical Association*, 99(486):1051-26.

Heckman, J. and J. Smith. 1999. "The Pre-Programme Earnings Dip and the Determinants of Participation in a Social Programme. Implications for Simple Programme Evaluation Strategies," *The Economic Journal,* 109 (457), 313-348.

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

Heckman, J., 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

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

Hollister, R. and J. Hill. 1995. "Problems in the Evaluation of Community-Wide Initiatives," in J. Connell, A. Kubisch, L. Schorr, and C. Weiss (Eds.), New Approaches to Evaluating Community Initiatives: Concepts, Methods, and Contexts, Aspen Institute.

Honore, B. 1992. "Trimmed Lad and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica,* Vol. 60.

Imbens, G. 2004. "Nonparametric Estimation of Average Treatment Effects under Exogeneity," *The Review of Economics and Statistics*, 86(1): 4-29

Kordas, G. and S. Lehrer. 2004. "Matching Using Semiparametric Propensity Scores," Unpublished manuscript, Queen's University.

Kyriazidou, E. 1997. "Estimation of a Panel Data Sample Selection Model," *Econometrica*, 65, 1335-1364.

LaLonde, R. 1986. "Evaluating the Econometric Evaluation of Training Programs with Experimental Data," *The American Economics Review*, 76(4) , 604-620

Lee, W. 2001. "Propensity Score Matching on Commonly Available Nonexperimental Comparison Groups," Abt Associates, Unpublished Manuscript.

Li, Q., J. Racine, and J. Wooldridge. 2005. "Efficient Estimation of Average Treatment Effects with Mixed Categorical and Continuous Data," *Journal of Business and Economic Statistics* (forthcoming).

Meyer, B. 1995. "Lessons from the U.S Unemployment Insurance Experiments," *Journal of Economic Literature*, 33: 91-131.

Michalopoulos, C., H. Bloom, and C. Hill. 2004. "Can Propensity-Score Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs?," *The Review of Economics and Statistics*, 86(1): 156-179.

Moffitt, R. and W. Nicholson. 1982. "The Effect of Unemployment Insurance on Unemployment: The Case of Federal Supplemental Benefits," *Review of Economics and Statistics,* 64:1-11.

Mortensen, D. 1970. "Job Search, The Duration of Unemployment, and the Phillips Curve," *American Economic Review,* 60: 505-517.

Pagan, A. and A. Ullah. 1999. "Nonparametric Econometrics," Cambridge University Press.

Powell, J.L. 1987. "Semiparametric Estimation of Bivariate Latent Variable Models," Working Paper #8704, Social System research Institute, University of Wisconsin.

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

Rosenbaum, P. 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.

Rubin, D. 1973. "Matching to Remove Bias in Observational Studies," *Biometrics*, 29: 159-183.

Seifert, B. and T. Gasser. 1996. "Finite-Sample Variance of Local Polynomials: Analysis and Solutions," *Journal of the American Statistical Association* 91: 267-275.

Seifert, B. and T. Gasser. 2000. "Data Adaptive Ridging in Local Polynomial Regression," *Journal of Computational and Graphical Statistics*, 9(20): 338-360.

Smith J., and P. Todd 2005a. "Does Matching Overcome Lalonde's Critique of Non-Experimental Estimators?," *Journal of Econometrics,* 125(1-2) 305-353.

Smith, J. and P. Todd. 2005b. "Rejoinder," *Journal of Econometrics* 125 (1-2) 365-375

Todd, Petra. 2002. "Local Linear Approaches to Program Evaluation Using a Semiparametric Propensity Score," Unpublished manuscript, University of Pennsylvania.

Woodbury, S. and R. Spiegelman. 1987. "Bonuses to Workers and Employers to Reduce Unemployment: Randomized Trials in Illinois," *American Economic Review,* 77:513-530.

Table 1: Summary Statistics
Kentucky WPRS Program, October 1994 to June 1996

	Sample Sets			Test for Differences In Means		
	Treated *(1)*	Control *(2)*	Comparison (3)	p-values *(1)-(2)*	p-values *(1)-(3)*	Standardized difference (%) *(1)-(3)*
Profiling score	0.83	0.80	0.58	0.91	0.00	108
	(0.22)	(0.21)	(0.23)			
1st quarter earnings	$4560	$5011	$3971	0.81	0.00	15.8
	(3821)	(4074)	(3647)			
2nd quarter earnings	$4456	$4683	$3846	0.90	0.00	16.6
	(3829)	(3744)	(3578)			
3rd quarter earnings	$4901	$4976	$4174	0.84	0.00	19.9
	(3795)	(3515)	(3521)			
4th quarter earnings	$5130	$5103	$4496	0.43	0.00	17.9
	(3735)	(3609)	(3396)			
Years of schooling	12.5	12.3	12.4	0.22	0.21	6.6
	(2.1)	(2.0)	(1.9)			
Less than high school (%)	1.9	1.3	1.1	0.88	0.00	4.5
	(0.1)	(0.1)	(0.1)			
Bachelor degree (%)	5.4	5.2	4.3	0.31	0.20	5.0
	(0.2)	(0.2)	(0.2)			
Graduate studies (%)	1.30	0.94	0.81	0.44	0.16	4.8
	(0.11)	(0.0)	(0.08)			
Age	37.0	37.0	36.6	0.77	0.31	3.9
	(11)	(10.8)	(11.4)			
Percent females	43.1	39.6	40.8	0.06	0.26	4.5
	(0.4)	(0.4)	(0.4)			
Percent whites	88.9	91.7	90.2	0.76	0.43	-4.3
	(0.3)	(0.2)	(0.2)			
Percent blacks	10.7	7.9	9.4	0.84	0.44	4.3
	(0.3)	(0.2)	(0.2)			
N	1236	745	9,002			

Notes: Standard deviations are given in parenthesis. Means are unweighted. Test for differences in means for the experimental sample (1 versus 2) are based on a linear regression that conditions on a treatment dummy variable and on the PTGs. Test for differences in means for the nonexperimental sample (1 versus 3) are based on a linear regression that conditions on a treatment dummy variable and on local office and week. The standardized difference is the mean difference as a % 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
Balancing Test, Empirical Support Region, and Classification Rates
Kentucky WPRS Program, October 1994 to June 1996

	parametric propensity scores models [a]			kernel-based[e]
	$\rho = P(T=1 \mid X)$	$\rho = P(T=1 \mid X, office)$	$\rho = P(T=1 \mid X, office, week)$	$\rho = P(T=1 \mid X)$
	(1)	(2)	(3)	(4)
Standardized Differences[b]				
sex	0.65	0.32	-1.80	2.62
schooling	-0.31	-2.70	-2.89	-0.26
age	1.65	0.89	5.33	1.71
white	-0.53	-0.26	-0.53	-1.06
profiling scores	9.01	17.02	23.15	4.71
1st quarterly earnings	3.96	3.59	4.87	2.96
4th quarterly earnings	4.39	4.00	5.61	4.54
Employed→ Employed	-0.49	-0.49	0.00	-3.99
Employed→ Not Employed	0.56	0.56	0.00	4.50
Not Employed→ Not Employed	0.00	0.00	0.00	0.00
Classification Rates (%)[c]				
Treatment	73	79	84	83
Control	68	74	79	74
Treated units out of CS (%)[d]				
trimming	13	15	15	21
max of min	2	3	3	13
kth larger comparison unit	8	11	16	18

[a] The specification for the baseline propensity scores model follows: $T_i = f(\alpha_1 \text{age} + \alpha_2 \text{sex} + \alpha_3 \text{white} + \alpha_4 \text{white*sex} + \alpha_5 \text{educ} + \alpha_6 score + + \alpha_7 \text{score*educ} + \alpha_8 \text{score*age} + \alpha_9 \text{past 1st quarterly earnings} + \alpha_{10} \text{past 4th quarterly earnings} + + \alpha_{11} \text{work-work} + \alpha_{12} \text{nowork-work} + + \alpha_{13} \text{nowork-nowork} + \varepsilon)$.

[b] The standardized differences are defined by $100*(\bar{x}_T - \bar{x}_C)/[(s_T^2 + s_C^2)/2]^{1/2}$ where $\bar{x}_T$ and $\bar{x}_C$ are the sample weighted means of covariates in the treatment and matched comparison groups and, s_T^2 and s_C^2 are the sample variances in the raw data for both groups. Nearest-neighbor matching with replacement is estimated using the Mahalanobis metric distance.

[c] Classification rates are based on P(x)>0.12 to predict T=1 and P(x) $\leq$ 0.12 to predict T=0.

[d] "trimming" uses the 2 percent trimming method developed in Heckman et al. (1998). "max of min" drops all units with propensity scores smaller than the minimum and larger than the maximum in the opposite group. "*kth* larger comparison observation" defines the empirical common support as the region where at least *k* comparison units with the highest propensity scores are available to match. All treatment units with propensity scores higher than that for the *kth* larger comparison unit are dropped. *k=15* is the cutoff point.

[e] The kernel-based propensity scores model is based on multivariate density function estimation with cross-validated bandwidth selection and mixed data type (Hall, Racine, and Li 2004) . The estimated model is:
$T_i = g$(age, sex, white, educ, scores, past 1st quarterly earnings, past 4th quartely earnings, employment transitions).

Table 3
Estimated Bias for Parametric Propensity Score Models
Kentucky WPRS Program, October 1994 to June 1996.

Outcomes	PANEL A					PANEL B				
	experimental estimates	matching with fixed effects				experimental estimates	standard matching			
		nearest-neighbor	local constant	local linear	ridge estimator		nearest-neighbor	local constant	local linear	ridge estimator
Weeks receiving UI benefits	-2.23	-2.39	-2.20	-2.15	-2.15	-2.25	-1.80	-1.93	-1.92	-1.92
	(0.52)	(0.62)	(0.35)	(0.36)	(0.36)	(0.52)	(0.65)	(0.41)	(0.44)	(0.44)
		[7]	[-1]	[-4]	[-4]		[-20]	[-14]	[-14]	[-14]
Employment (%) on 1st quarter after UI spell	8.53	7.99	6.25	6.17	6.17	9.48	2.61	7.13	6.83	6.83
	(2.73)	(2.80)	(1.71)	(1.80)	(1.80)	(2.71)	(3.40)	(1.80)	(2.35)	(2.35)
		[-6]	[-27]	[-28]	[-28]		[-72]	[-25]	[-28]	[-28]
Earnings on 1st quarter after UI spell	488	543	365	383	383	526	268	295	251	251
	(201)	(167)	(130)	(128)	(128)	(200)	(222)	(177)	(193)	(193)
		[11]	[-25]	[-22]	[-22]		[-49]	[-44]	[-52]	[-52]
N	1810	9938	9938	9938	9938	1791	9929	9929	9929	9929

Bootstrap standard errors are shown in parenthesis. They are based on 100 repetitions. Nonexperimental bias defined as $[(\Delta_{TT}^{non-\exp} - \Delta_{TT}^{\exp})/\Delta_{TT}^{\exp}]*100$ is shown in brackets.

The experimental treatment effects are estimated as in Black et al. (2003) after imposing the common support condition using the 2 percent trimming method developed in Heckman et al. (1998).

The specification for the baseline logit propensity scores model follows: $T_i = f(\alpha_1 \text{age} + \alpha_2 \text{sex} + \alpha_3 \text{white} + \alpha_4 \text{white*sex} + \alpha_5 \text{educ} + \alpha_6 score + + \alpha_7 \text{score*educ} + \alpha_8 \text{score*age} + \alpha_9 \text{past 1st quarterly earnings} + \alpha_{10} \text{past 4th quarterly earnings} + + \alpha_{11} \text{work-work} + \alpha_{12} \text{nowork-work} + + \alpha_{13} \text{nowork-nowork} + \varepsilon)$. Local polynomial matching estimation is based on Epanechnikov kernel functions. The bandwidth-selector is based on locally varying bandwidths: $h(\rho) = h_{cv}(\rho/(1-\rho))^{1/5}$, where h_{cv} denotes the bandwidth from conventional cross-validation and ρ the estimated propensity scores.

Standard matching is based on the full logit model that includes local office and week dummy variables in the specification of the propensity scores model.

Table 4
Estimated Bias for Kernel-Based Propensity Score Models
Kentucky WPRS Program, October 1994 to June 1996.

	experimental estimates	Matching with Fixed Effects			
		nearest-neighbor	local constant	local linear	ridge estimator
	(1)	(2)	(3)	(4)	(5)
Weeks receiving UI benefits	-2.28 (0.54)	-2.43 (0.78) [7]	-2.04 (0.40) [-11]	-2.17 (0.42) [-5]	-2.17 (0.42) [-5]
Employment (%) on 1st quarter after UI spell	8.60 (2.81)	2.67 (3.48) [-69]	5.93 (2.12) [-31]	6.01 (1.96) [-30]	6.01 (1.96) [-30]
Earnings on 1st quarter after UI spell	416 (162)	155 (200) [-63]	136 (113) [-67]	155 (106) [-63]	155 (106) [-63]
N	1714	9954	9954	9954	9954

Bootstrap standard errors are shown in parenthesis. They are based on 100 repetitions. Nonexperimental bias defined as $[(\Delta_{TT}^{non-\exp} - \Delta_{TT}^{\exp})/\Delta_{TT}^{\exp}]*100$ is shown in brackets.

The experimental treatment effects are estimated as in Black et al. (2003) after imposing the common support condition using the 2 percent trimming method developed in Heckman et al. (1998).

Kernel-based propensity score model is based on multivariate density function estimation with cross-validated bandwidth selection and mixed data type (Hall, Racine, and Li 2004) . The estimated model is:

$T_i = g$(age, sex, white, educ, scores, past 1st quarterly earnings, past 4th quartely earnings, employment transitions).

Local polynomial matching estimation is based on Epanechnikov kernel functions over the empirical common support region, which is defined by the 2 percent trimming method developed in Heckman et al. (1998). The bandwidth-selector is based on locally varying bandwidths: $h(\rho) = h_{cv}(\rho/(1-\rho))^{1/5}$, where h_{cv} denotes the bandwidth from conventional cross-validation and ρ the estimated propensity scores.

Table 5
The Effect of Geography on Estimated Bias
Kentucky WPRS Program, October 1994 to June 1996.

		Matching with Fixed Effects					
	experimental estimates	'northern' local offices		'western' local offices		'appalachian' local offices	
		LC	LL	LC	LL	LC	LL
Weeks receiving UI benefits	-2.01	-1.93	-1.95	-3.39	-3.00	-4.73	-4.84
	(0.86)	(0.50)	(0.51)	(1.31)	(1.42)	(0.56)	(0.69)
		[4]	[3]	[69]	[49]	[135]	[140]
Employment (%) on 1st quarter after UI spell	5.89	4.34	4.28	9.30	7.91	15.79	15.62
	(4.30)	(2.98)	(3.04)	(4.15)	(5.75)	(3.22)	(3.81)
		[26]	[27]	[58]	[34]	[168]	[166]
Earnings on 1st quarter after UI spell	535	279	289	872	735	559	645
	(356)	(213)	(208)	(188)	(260)	(232)	(269)
		[48]	[46]	[63]	[37]	[5]	[21]
N	900	4528		1368		1639	

Bootstrap standard errors are shown in parenthesis. They are based on 100 repetitions. Nonexperimental bias defined as $[(\Delta_{TT}^{non-\exp} - \Delta_{TT}^{\exp})/\Delta_{TT}^{\exp}]*100$ is shown in brackets.

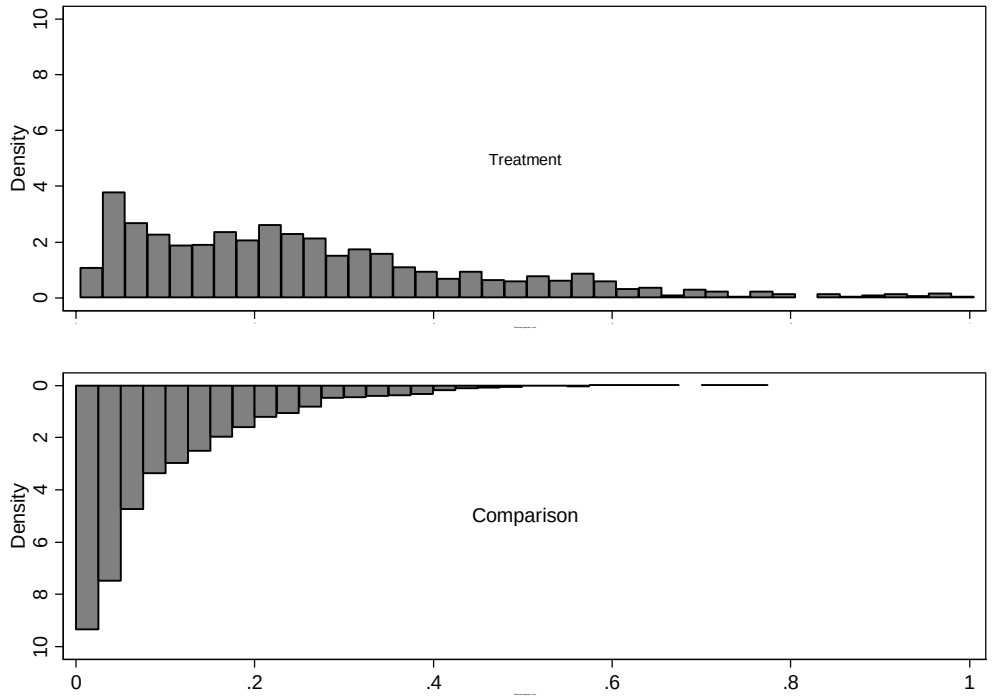
The experimental treatment effects are estimated as in Black et al. (2003) after imposing the common support condition using the 2 percent trimming method developed in Heckman et al. (1998) and for the subsample of 'northern' local offices. Local polynomial matching estimation is based on Epanechnikov kernel functions over the empirical common support region. The bandwidth-selector is based on locally varying bandwidths: $h(\rho) = h_{cv}(\rho/(1-\rho))^{1/5}$, where h_{cv} denotes the bandwidth from conventional cross-validation and ρ the estimated propensity scores.

Table 6
Sensitivity Checks for Local Constant Matching Estimation
Kentucky WPRS Program, October 1994 to June 1996.

	experimental estimates	Local Constant Matching with Fixed Effects			
		Baseline estimation	'max of min' support region	cross-validated bandwidth	time and space fixed effects order
	(1)	(2)	(3)	(4)	(5)
Weeks receiving UI benefits	-2.23	-2.20	-2.05	-2.17	-2.20
	(0.52)	(0.35)	(0.37)	(0.36)	(0.36)
		[-1]	[-8]	[-3]	[-1]
Employment (%) on 1st quarter after UI spell	8.53	6.25	6.44	6.32	6.22
	(2.73)	(1.71)	(1.97)	(1.81)	(1.80)
		[-27]	[-25]	[-26]	[-27]
Earnings on 1st quarter after UI spell	488	365	292	370	365
	(201)	(130)	(131)	(127)	(128)
		[-25]	[-40]	[-24]	[-25]
N	1810	9938	9917	9938	9938

Bootstrap standard errors are shown in parenthesis. They are based on 100 repetitions. Nonexperimental bias defined as $[(\Delta_{TT}^{non-exp} - \Delta_{TT}^{exp})/\Delta_{TT}^{exp}]*100$ is shown in brackets. The experimental treatment effects are estimated as in Black et al. (2003) after imposing the common support condition using the 2 percent trimming method developed in Heckman et al. (1998).
The baseline model (column 2) is implemented using the 2 percent trimming method and locally varying bandwidths: $h(\rho) = h_{cv}(\rho/(1-\rho))^{1/5}$, where h_{cv} denotes the bandwidth from conventional cross-validation and ρ the estimated propensity scores. Column (3) defines the overlapping support region by the 'max of min' criterion that drops all observations whose propensity score is smaller than the minimum and larger than the maximum in the opposite group. The common support lies within the interval [0.004, 0.773]. Column (4) selects the optimal bandwidths for standard leave-one-out cross-validation techniques. The resulting bandwidths are 0.07, 0.04, and 0.04 for weeks receiving UI benefits, employment, and quarterly earnings, respectively. Column (5) removes first time effects followed by local offices effects in the estimation of matching with fixed effects.

Figure 1
Parametric Propensity Score Model Distribution



Kernel Propensity Score Model Distribution