10-2007

Bandwidth Selection and the Estimation of Treatment Effects with Unbalanced Data

Jose Galdo
Syracuse University

Jeffrey A. Smith
University of Michigan

Dan Black
University of Chicago

DISCUSSION PAPER SERIES

IZA DP No. 3095

Bandwidth Selection and the Estimation of Treatment Effects with Unbalanced Data

Jose Galdo
Jeffrey Smith
Dan Black

October 2007

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Bandwidth Selection and the Estimation of Treatment Effects with Unbalanced Data

Jose Galdo
McMaster University and IZA

Jeffrey Smith
University of Michigan, NBER, IFS, PSI, ZEW and IZA

Dan Black
University of Chicago and NORC

Discussion Paper No. 3095
October 2007

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 3095
October 2007

ABSTRACT

Bandwidth Selection and the Estimation of Treatment Effects with Unbalanced Data*

This paper addresses the selection of smoothing parameters for estimating the average treatment effect on the treated using matching methods. Because precise estimation of the expected counterfactual is particularly important in regions containing the mass of the treated units, we define and implement weighted cross-validation approaches that improve over conventional methods by considering the location of the treated units in the selection of the smoothing parameters. We also implement a locally varying bandwidth method that uses larger bandwidths in areas where the mass of the treated units is located. A Monte Carlo study compares our proposed methods to the conventional unweighted method and to a related method inspired by Bergemann et al. (2005). The Monte Carlo analysis indicates efficiency gains from all methods that take account of the location of the treated units. We also apply all five methods to bandwidth selection in the context of the data from LaLonde's (1986) study of the performance of non-experimental estimators using the experimental data from the National Supported Work (NSW) Demonstration program as a benchmark. Overall, both the Monte Carlo analysis and the empirical application show feasible precision gains for the weighted cross-validation and the locally varying bandwidth approaches.

JEL Classification: C13, C14

Keywords: matching, cross-validation, kernel regression, Monte Carlo simulation

Corresponding author:

Jeffrey Smith
Department of Economics
University of Michigan
238 Lorch Hall
611 Tappan Street
Ann Arbor, MI 48109-1220
USA
E-Mail: econjeff@umich.edu

* We thank participants at the 2005 Canadian Economic Association Meetings at McMaster University, along with Bernd Fitzenberger, Markus Frölich, Pat Kline, two anonymous referees, and seminar audiences at Syracuse University and Goethe Universität Frankfurt, for helpful comments and suggestions.

1. Introduction

One of the fundamental contributions arising from the use of matching estimators in program evaluation over the last 15 years has been a better understanding of how disparate the distribution of covariates may be between treatment and comparison groups. Because many social programs select on criteria such as income, assets, past program participation, or past interaction with the criminal justice system, comparison groups drawn from the population at large, or even from crudely matched sub-populations, may contain an overwhelming number of observations that have virtually no use in an evaluation. Thus, despite a large total number of observations, a comparison group may contain only a few observations relevant to evaluating the program.

In this paper, we examine how the disparate distributions of covariates in the treatment and comparison groups affect the proper choice of the smoothing parameter.[1] Bandwidth selection has always posed a problem for evaluation methods that rely on kernel regression. The broader statistical literature offers some guidance by suggesting the minimization of quadratic loss functions such as the mean integrated squared error (MISE) through cross-validation methods. These data driven methods have the considerable advantage of allowing researchers to avoid arbitrary selection of bandwidths, and they converge to the optimal bandwidth, albeit at a slow rate. At the same time, the conventional cross-validation approach selects the bandwidth using only the distribution of the untreated units while completely neglecting the location of the treated ones. As Figure 1 illustrates, this approach may be inappropriate in the context of estimating the average treatment effect on the treated because the shape of the regression function and distribution of covariates in regions with few treated observations may

[1] Throughout this study we use the terms smoothing parameter and bandwidth interchangeably.

substantially affect the chosen bandwidth. Although this general insight applies to a variety of different econometric estimators applied to the evaluation problem, we focus on propensity score matching estimators that rely on local constant and local linear regression to estimate the counterfactual outcome regression function because of their wide use in the applied literature.

To account for the location of the treated units, we define and implement two weighted versions of the usual cross-validation bandwidth selection method. In the first version, the weighting function gives to untreated units the same weight they receive in the estimation of the counterfactual outcomes. This implies a different set of weights for each bandwidth considered in the bandwidth search grid, which may impose a computational burden. In the second version, the weighting function consists of an estimated density function for the propensity scores of the treated units. Both versions reweight the data to reflect differences in the distributions of propensity scores between the treated and untreated observations.

We also evaluate two alternative procedures. The first one, inspired by Bergemann, Fitzenberger, and Speckesser (2005), we call the "nearest neighbor" method. This procedure accounts for the location of the treated units by using standard cross-validation methods but counting the prediction errors only for the sub-sample of untreated units that are nearest neighbors of the treated units. In the second, we implement a locally varying bandwidth approach that selects a bandwidth for each treated unit according to the local density of the untreated units, with narrower bandwidths in regions dense in untreated units and wider bandwidths in regions with few untreated units.

To the best of our knowledge, Frölich's (2004) study represents the first to address the problem of bandwidth selection in the context of local polynomial matching estimators.[2] He finds that conventional cross-validation bandwidth choice, which does not account for the location of the treated units, performs well in small samples.

We study the finite-sample performance of the various bandwidth selection methods using a Monte Carlo analysis that combines three pairs of propensity score densities with four different regression functions for the untreated outcome. This analysis yields four main conclusions. First, conventional unweighted cross-validation consistently yields larger MSE than any of the four methods that take account of the location of the treated units. Second, of the two weighted cross-validation methods we propose here, the variable weight method does better for local linear rather than local constant kernel matching. Third, the locally varying bandwidth method and the nearest-neighbor approach generally perform better than the other methods, particularly in the most difficult density designs and when using the Epanechnikov kernel. Fourth, the shape of the regression function does not consistently determine the performance of the alternative bandwidth selection procedures.

We also apply the various bandwidth selection methods to the data from LaLonde's (1986) analysis that compares experimental and non-experimental estimates of the impact of the U.S. National Supported Work (NSW) Demonstration program. These data, also analyzed by Dehejia and Wahba (1999, 2002), and Smith and Todd (2005a,b) (and many others) include two different comparison group samples in addition to the experimental treatment and control groups. Three main results emerge from this

[2] Ichimura and Linton (2001) also study the problem of selecting optimal smoothing parameters when estimating average treatment effects. However, they focus on the non-parametric series estimator proposed by Hirano, Imbens, and Ridder (2003).

analysis. First, the variable weight approach and the fixed weights approach based on the density of the treated units both yield non-trivial efficiency gains relative to conventional cross-validation. Second, the nearest neighbor approach generates a lot of variability in the estimated impacts and displays a lot of sensitivity to the choice of kernel function. Third, the locally varying bandwidths do not do as well as in the Monte Carlo analysis.

The remainder of the paper proceeds as follows. Section 2 discusses identification and estimation while Section 3 lays out the general problem of optimal bandwidth selection as well as the conventional solution. Section 4 lays out the various bandwidth selection schemes we examine. Section 5 describes the Monte Carlo analysis and its findings while Section 6 describes our application of the various bandwidth selection methods to the National Supported Work data. Section 7 concludes.

2. Identification and Estimation

2.1 Identification

In recent years, matching estimators have received a lot of attention in economics as a flexible alternative to traditional parametric regression methods when the data contain a sufficiently rich set of observable determinants of treatment and outcomes to justify a "selection on observables" assumption. See, for instance, Heckman, Ichimura and Todd (1997, 1998), Heckman, Ichimura, Smith and Todd (1998), Hirano, Imbens and Ridder (2003), Imbens (2004), and Smith and Todd (2005a, b). This section discusses identification and estimation in the context of the potential outcomes framework commonly used in this literature, with a special focus on matching estimators that rely on

local polynomial regression to estimate the expected counterfactual outcome for each treated unit.

Let Y_1 and Y_0 denote the potential outcomes conditional on participation and non-participation, respectively. Let $T_i \in \{0,1\}$ indicate participation. In many (if not most) evaluation contexts, interest centers on the mean impact of treatment on the treated, given by $\Delta_{TT} = E(Y_1 \mid T=1) - E(Y_0 \mid T=1)$; we focus our analysis on this parameter.

Data on program participants identify $E(Y_1 \mid T=1)$. The mean counterfactual outcome $E(Y_0 \mid T=1)$, however, is missing and cannot be directly identified from the data. Matching proceeds by invoking the Conditional Independence Assumption (CIA),

$$Y_0 \perp T \mid X. \tag{1}$$

Under (1), $E(Y_0 \mid T=1) = E_{X|T=1}(E(Y \mid X, T=1)) = E_{X|T=1}(E(Y \mid X, T=0))$ for all values of *X* that satisfy the common support condition $\Pr(T=1 \mid X) < 1$. This latter condition guarantees the existence (at least in the population) of non-participants with the same values of *X* as all of the participants.[3]

We can think of matching as using predicted values from a regression of Y_0 on *X* to form the expected counterfactual outcome for each treated unit. More formally, $E(Y_0 \mid T=1) = \int m(x) f_x(x \mid T=1) dx$, where $m(x) = E(Y_0 \mid X)$ denotes the conditional mean function given *non-participation* and $f_x(X \mid T=1)$ denotes the density of *X* conditional on *participation*.

[3] As noted in Heckman, Ichimura and Todd (1997), Assumption (1) can be weakened to conditional mean independence.

As discussed in, e.g., Pagan and Ullah (1999), the number of covariates included in X generally determines the rate of convergence for nonparametric estimators of the regression function. Thus, including a rich covariate set X in the hope of satisfying the CIA can lead to extremely slow convergence rates. Rosenbaum and Rubin (1983) show that if the CIA holds for X then it also holds for the conditional probability of participation $P(X) = \Pr(T = 1 \mid X)$ or propensity score. Replacing X with $P(X)$, the CIA becomes $Y_0 \perp T \mid P(X)$. Matching on the scalar propensity score reduces the dimensionality of the problem of estimating the conditional mean function for the untreated outcome from $\dim(X)$ to one. Of course, this does not really solve the problem, but instead pushes it back to the level of estimating the probability of participation.[4]

2.1 Estimation

The sample analog to the integral above constitutes the estimator for the counterfactual mean given matching on the propensity score,

$$\hat{E}(Y_0 \mid T = 1) = (1/n_1)\sum_{i=1}^{n_1} \hat{m}(\rho_i),$$

where $\hat{P}(x_i) = \rho_i$, $\hat{m}(\rho_i)$ indicates a regression estimator of $m(\rho)$ evaluated at the covariate values of participant *i*, and n_1 denotes the size of the participant sample. The literature suggests a wide variety of ways to estimate the conditional mean function non-parametrically. We focus our attention on local constant and local linear matching

[4] In the matching context, researchers typically adopt a flexible parametric specification for the propensity score, thus changing the overall procedure from a non-parametric to a semi-parametric one. Balancing tests, as described in, e.g., Smith and Todd (2005b) and Lee (2006), then guide the selection of the flexible parametric specification for a given set of conditioning variables thought to satisfy the CIA. Todd (2002) and Kordas and Lehrer (2004) examine semi-parametric estimation of the propensity score. Hirano, Imbens and Ridder (2003) propose a non-parametric series estimator for the propensity score.

estimators of the conditional mean function (Heckman et al. 1997) but, as noted in the introduction, our general point applies to all non-parametric and semi-parametric estimators that require a bandwidth choice or its equivalent, as with the number of terms in an expansion, the number of strata or the number of nearest neighbors.

The general form of the matching estimator for the impact of treatment on the treated is given by

$$\Delta_{TT}^{M} = \frac{1}{n_{1c}} \sum_{i \in C} \{Y_{1i} - \hat{m}(\rho_i)\} = \frac{1}{n_{1c}} \sum_{i \in C} \left\{ Y_{1i} - \left\{ \sum_{j \in C} W_h(\rho_i, \rho_j) Y_{0j} \right\} \right\}, \quad (2)$$

where Y_{1i} and Y_{0j} indicate the outcome for treated unit "*i*" and untreated unit "*j*", *C* denotes the region of common support, n_{1c} denotes the number of treated units in the common support, respectively, and $W_h(\rho_i, \rho_j)$ indicates the weight that untreated observation "*j*" receives in the construction of the estimated expected counterfactual outcome for treated unit "*i*".[5] The weights depend on the particular kernel function employed, the smoothing parameter *h*, and the choice of local constant or local linear regression.

The local polynomial regression estimator of the conditional mean function equals $\hat{\beta}_0$ from the solution to the optimization problem

$$\min_{\beta_0, \ldots \beta_p} \sum_{j=1}^{n_0} (Y_{0j} - \sum_{m=0}^{p} \hat{\beta}_m (\rho_j - \rho_i)^m)^2 K(\frac{\rho_j - \rho_i}{h}),$$

where $(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)$ denotes a vector of regression coefficients, *p* denotes the order of the

[5] The literature typically distinguishes only casually between the region of common support in the population and that in the sample at hand. We assume full common support in the population and then impose a more restrictive common support condition in the sample as described below.

local polynomial, and $K(\cdot)$ denotes a symmetric kernel function satisfying some assumptions. 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), with the implied weights,

$$W_h(\rho_i, \rho_j) = \frac{K_{ij}}{\sum_{g \in T=0} K_{ig}}, \qquad (3)$$

where $K_{ij} = K((\rho_j - \rho_i)/h)$. The corresponding weights for the local linear regression are given in equations (2.2)-(2.4) of Fan (1992).

As discussed in Fan (1992), several issues arise in choosing between the local linear and local constant kernel regression estimators. 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. Intuitively, the local linear estimator should perform better in contexts with the untreated units distributed asymmetrically around the treated units and a relatively steep conditional mean function. At the same time, the local linear estimator demands more of the data because it estimates one additional parameter in every local regression. This suggests the possibility that the local constant estimator might have lower mean squared error in finite samples. Given the lack of a clear choice between the two in many applied contexts, we consider both estimators in our Monte Carlo and empirical analyses later on.

3. Optimal Bandwidths for Average Treatment Effects

3.1 The standard approach

The greater flexibility associated with non-parametric estimation of the conditional mean function comes with a price: bandwidth selection. Choosing too narrow a bandwidth leads to under-smoothing. This means an unstable estimated function that confuses noise in the data for features of the population regression function. In contrast, choosing too wide a bandwidth leads to over-smoothing. This means that potentially interesting and important features of the population regression function get smoothed away in the estimation. 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 the result.

To avoid the excesses of bias or variance associated with a poor bandwidth choice, the standard approach chooses the bandwidth based on some measure of fit, typically the Mean Integrated Squared Error (MISE), given by

$$\text{MISE} = E(\int [m(\rho) - \hat{m}(\rho, h)]^2 d\rho) = \int [\text{bias}^2(\hat{m}(\rho)) + \text{var}(\hat{m}(\rho))] d\rho \,.$$

The MISE criterion embodies a particular trade-off between bias and variance; Pagan and Ullah (1999) discuss alternative fit criteria. Calculation of the MISE requires the pointwise bias and variance of the regression function. The derivation of analytical formulae for these quantities builds on the following standard assumptions:

(A-1) Sampling of $\{X_i, Y_i\}$ is i.i.d., with $\text{Var}(Y_i) < \infty$;

(A-2) $h = h_n \to 0$ and $nh \to \infty$ as $n \to \infty$;

(A-3) $m(\rho)$ has a bounded and continuous second derivative and $f_{\rho|T=0}(\rho)$ is continuous and bounded away from zero in the neighborhood of ρ where ρ is a point in the interior of the support;

(A-4) $K(\cdot)$ is symmetric density function satisfying the following properties:

(i) $\int K(z)dz = 1$, (ii) $\int zK(z)dz = 0$, (iii) $\int z^2 K(z)dz = \kappa_2 < \infty$, (iv) $\int K^2(z_i)dz_i = \kappa^2$,

where κ_2 refers to the second-order kernel and κ^2 refers to the square of the kernel function.

Assumptions (A-1) through (A-2) guarantee consistency of the MISE for both the local constant and local linear estimators. Assumption (A-3) allows us to evaluate the formulae for the bias and variance given below. Assumption (A-4) assumes a second order kernel with bounded and nonzero first and second moments. We can easily generalize this assumption to the multivariate case through higher order kernels, with the order determined by the largest nonzero moment. Assumption (A-4) also implies symmetry around zero for the kernel functions; a large class of kernel functions, including the normal and Epanechnikov kernels, satisfies (A-4).

Fan (1992) proves that under assumptions (A-1) to (A-4) the (asymptotic) pointwise bias and variance of the local constant and local linear regressions estimators are approximated by

$$\text{bias}_{\text{LCR}}(\hat{m}(\rho)) \approx \frac{h^2}{2f_{\rho|T=0}(\rho)} \int \left\{ (m''(\rho) f_{\rho|T=0}(\rho)\kappa_2 + 2m'(\rho) f'_{\rho|T=0}(\rho)\kappa_2 \right\} d\rho,$$

$$\text{bias}_{\text{LLR}}(\hat{m}(\rho)) \approx \frac{h^2}{2} \int \left\{ (m''(\rho)\kappa_2 \right\} d\rho,$$

$$\text{var}_{\text{LCR}}(\hat{m}(\rho)) \approx \text{var}_{\text{LLR}}(\hat{m}(\rho)) = \frac{1}{n_0 h} \int \frac{\kappa^2 \sigma_0^2(\rho)}{f_{\rho|T=0}(\rho)} d\rho,$$

where $\sigma_0^2(\rho)$ denotes the conditional variance of the untreated outcome.[6] In conventional nonparametric regression, cross-validation methods are often used to minimize the MISE criterion. Hall (1983) and Stone (1984), among others, have shown that bandwidths selected by cross-validation converge to the MISE-minimizing bandwidth.[7]

3.2 Problems with the standard approach

The standard approach has problems in the context of matching estimators. First, and most obviously, the object of interest in the matching case consists of the estimated average treatment effect rather than the regression function for the untreated outcome. Therefore, we are interested in minimizing the mean squared error (MSE) of the matching estimator rather than the MISE of the regression function.[8] Given the additional averaging involved in constructing the matching estimate, we should not expect that a bandwidth that minimizes the MISE for the regression function also minimize the MSE of the matching estimator.

Second, the chosen bandwidth does not depend on the location of the treated units. As illustrated in Figure 1 and discussed in the introduction, when using observational data, imbalance in the distributions of conditioning variables between the treated and untreated samples may lead to poor bandwidth choice. Intuitively, things will go wrong if the optimal bandwidth in the region of low propensity scores, where most

[6] Fan's (1992) proofs also require a known propensity score. Given a parametric propensity score model, the variance component from the propensity score estimation converges faster than the variance component from the non-parametric regression, and so does not matter for the (asymptotic) results.

[7] The rate of convergence of cross-validation is glacial, of the order of $n^{-1/10}$ (Pagan and Ullah 1999). See the related discussion, references and simulation results in Loader (1999).

[8] As discussed in detail in, e.g., Li and Racine (2007), MSE and MISE represent different, but closely related, concepts. MSE applies to a specific point (as in our locally varying bandwidth scheme) or estimator (as when minimizing the MSE of the matching estimator of the treatment effect. MISE constitutes a global criterion defined over an entire regression function; the standard approach uses cross-validation in an attempt to minimize the MISE of the estimated regression function for the untreated outcome.

untreated units lie, differs from the optimal bandwidth in the region of high propensity scores where most of the treated units lie. The more numerous untreated units with low propensity scores will dominate the bandwidth choice under the standard approach, which may lead to substantively important pointwise biases where it matters – in the region of high propensity scores.

To address both problems in the context of matching estimators, Frölich (2005) derives an asymptotic linear approximation to the MSE of the expected counterfactual outcome (the part of the matching estimator that relies on the non-parametric estimate of the conditional mean function) and then uses the approximation to guide bandwidth selection. Frölich (2005) demonstrates that under assumptions (A-1) to (A-4) above the second-order linear approximations to the bias and variance of the expected counterfactual outcome for the local constant and local linear estimators depends on the location of the treated units.

Though potentially promising, this approach has (at least) three problems. First, Frölich's (2005) own Monte Carlo analysis suggests a lack of sensitivity of the approximate MSE, which turns out to be quite flat, to the bandwidth choice. In particular, he finds that his approach tends to pick bandwidths that substantially under-smooth the conditional mean function. Second, from a practical standpoint, the approximate bias and variance depend on several unknowns, such as the population regression and density functions. Estimating these unknowns notably increases both required research time and the computational burden of the overall estimation. Third, estimation of these unknowns involves the selection of additional smoothing parameters.

4. Weighted Cross-Validation

4.1 Basic idea

We propose a weighted leave-one-out cross-validation bandwidth selection approach that accounts for the location of the treated units, and thus may improve over the conventional cross-validation algorithm. As outlined in Stone (1974), the conventional approach estimates the MISE associated with any given candidate bandwidth using leave-one-out cross-validation. See, e.g. Black and Smith (2004) for an application.

Formally, the standard approach chooses the bandwidth *h* to minimize the approximation to the MISE (of the estimated counterfactual mean regression function) associated with a particular bandwidth given by

$$\text{MISE}_{\text{C}}(h) = \arg\min_{h}\left(\frac{1}{n_0}\sum_{j=1}^{n_0}(Y_{0j} - \hat{m}_{-j}(\rho_j, h))^2\right), \tag{4}$$

where $\hat{m}_{-j}(\rho_j, h)$ denotes the estimated conditional mean function for the untreated outcome evaluated at ρ_j using all of the untreated units except unit "*j*". The omission of unit "*j*" avoids a minimum of zero at a bandwidth small enough that only observation "*j*" receives positive weight in estimating the conditional mean function at ρ_j. The cost of omitting unit "*j*" is that the cross-validation proceeds with a sample size one smaller than the sample actually used in the estimation of the treatment effect. The benefit comes from using out-of-sample forecasts rather than in-sample fit to guide the bandwidth choice. This approach implicitly weights the MISE calculation by the distribution of estimated propensity scores in the untreated sample. Operationally, the traditional approach proceeds via a grid search.

Our method proceeds along the same path as the conventional method just described, but instead weights the MISE criterion using the distribution of estimated propensity scores in the treated sample rather than the distribution in the untreated sample. In this way, the selected bandwidth should provide a lower local mean squared error for the regression function in the regions dense with treated units, typically regions of relatively high propensity scores, rather than in regions dense with untreated units. While our scheme is not fully efficient (relative to selecting the bandwidth by minimizing the MSE at each point) at the usual task of minimizing the MISE of the regression function, it should lead to a lower MSE for the matching estimator than naïve cross-validation.[9]

In notation, we replace the MISE criterion given in equation (4) above with the alternative MISE criterion

$$\text{MISE}_{\text{W}} = \arg\min_{h} \left(\frac{1}{n_0} \sum_{j=1}^{n_0} (Y_{0j} - \hat{m}_{-j}(\rho_j, h))^2 W_j(\rho_j, h) \right), \tag{5}$$

where $W_j(\cdot)$ denotes a weighting function that depends on the relative density of treated units in the vicinity of ρ_j. In this paper, we implement three alternative definitions of $W_j(\cdot)$, which we define in the next two sub-sections.

4.2 Variable Weights

Under the first definition of the weighting function, each untreated unit receives the same weight that it receives in the estimation of the expected counterfactual outcome, a quantity that clearly varies with *h* (i.e. that given in equation (2) for the local constant

[9] Intuitively, choosing a single bandwidth by minimizing the MISE via conventional cross-validation will be less efficient than pointwise bandwidth selection. We expect the conventional approach to pick a bandwidth that oversmooths in the region of low propensity scores and undersmooths in the region of high propensity scores. Our method works against this tendency in the region of high propensity scores and so, while not fully efficient, should yield a lower MSE for the matching estimand.

case and that given in Fan (1992) for the local linear case). For instance, in the local constant case with bandwidth *h*, observation "*j*" receives the following total weight in constructing the counterfactual mean from the n_1 treated observations:

$$W_j(\cdot) = \sum_{i=1}^{n_1} \frac{K((\rho_j - \rho_i)/h)}{\sum_{l=1}^{n_0} K((\rho_j - \rho_l)/h)}.$$

Each term in the sum represents the weight on untreated observation "*j*" in constructing the estimated expected counterfactual for treated observation "*i*". For kernel functions like the Epanechnikov, which has compact support, if any treated unit falls outside the support region spanned by *h*, the corresponding term equals zero. This feature of the weights can lead to odd behavior in finite samples; in particular, it can lead to discontinuous jumps in the estimated MISE of the regression function (and MSE of the matching estimator) as treated units move in and out of the support region as the bandwidth changes.

A quick inspection of the equation reveals that untreated units located near the mass of the treated units (typically those with higher scores) receive on average higher weights in the construction of the estimated MISE than untreated units located at a distance from the mass of the treated units.

4.3 Fixed Weights

The second definition of the weighting function defines the weights as proportional to the estimated density of the propensity scores among the treated units. Under this definition, the weights do not vary with the bandwidth. We propose estimating this density using standard non-parametric estimators as in Silverman (1986). Doing so requires an

additional bandwidth choice (which, of course, the first definition of the weights does not). We use least squares cross-validation as in Hall, Racine and Li (2004).

More formally, we estimate the density as

$$\hat{f}_{T=1}(\rho \mid T=1,\rho_j) = \frac{1}{n_1}\sum_{i\in\{T=1\}} K((\rho_i - \rho_j)/h_d)$$

for each value of ρ_j present in the untreated sample, where h_d denotes the bandwidth (different, in general, from *h*) used in the density estimation. We then define the weights by dividing each density estimate by the sum of the density estimates so that the weights sum to one. Like the first definition of the weighting function, this one implies a larger weight on the mean squared error associated with untreated units near the mass of the treated units when constructing the MISE of the regression function than does the conventional approach.

The third definition avoids the estimation of the density function by weighting comparison group observation "*i*" by the re-scaled odds ratio, given by

$$W_j(\square) = \frac{\rho_j}{(1-\rho_j)}\left[\sum_{l\in\{T=0\}}\frac{\rho_l}{(1-\rho_l)}\right]^{-1}.$$

This weighting scheme causes the comparison group have the same distribution of profiling scores as the treatment group. Those familiar with propensity score weighting methods will recognize this as the same weights used to estimate the impact of treatment on the treated in Hirano, Imbens, and Ridder (2003); see also Horvitz and Thompson (1952), DiNardo, Fortin and Lemiexu (1996) and Imbens (2004). Under this weighting scheme, small deviations in the estimates of Y_0 for values of ρ near one get penalized much more heavily than those for values of ρ near zero.

4.4 A nearest-neighbor approach

Bergemann et al. (2005) propose an alternative weighted bandwidth selection scheme similar in spirit to our own.[10] Their approach minimizes the MSE of the matching estimator by selecting the smoothing parameter by cross-validation on the sample of nearest-neighbor untreated units. More specifically, their scheme minimizes

$$\left(\frac{1}{n_1} \sum_{i=1}^{n_1} \left[Y^0_{nn(i)} - \hat{m}_{-nn(i)}(\rho_{nn(i)}, h) \right] \right)^2,$$

where $nn(i)$ denotes the index of the untreated nearest neighbor of treated unit "i". We do not adopt their method as they define it because their version allows positive and negative prediction errors to cancel out – a very unattractive feature in our view.[11] Instead, we square the prediction errors but retain the idea of counting the prediction errors only for the sub-sample of untreated nearest neighbors to the treated observations.

Our variant of their method starts by finding the nearest neighbor untreated unit for each treated unit based on absolute distances in propensity scores. A given untreated unit may get selected more than once if it represents the nearest neighbor to multiple treated units. It then chooses the bandwidth by minimizing the MISE based on the sum of squared prediction errors from leave-one-out cross-validation for the set of nearest neighbor untreated units. Formally, the selected bandwidth minimizes

$$\frac{1}{n_1} \sum_{i=1}^{n_1} \left[Y^0_{nn(i)} - \hat{m}_{-nn(i)}(\rho_{nn(i)}, h) \right]^2,$$

[10] Flossmann (2006) suggests a bandwidth choice algorithm that builds on Ruppert's (1997) Empirical Bias Bandwidth Selection (EBBS) method. His ongoing work show efficiency gains and increased stability relative to conventional cross-validation approaches. We do not study his method here as it remains in development.

[11] A limited Monte Carlo analysis using the Bergemann et al. (2005) method as defined in their paper confirms that it yields larger MSE for the matching estimator than our variant of it.

The main difference between this selection method and the other three proposed directly above lies in how they respond to increases in the size of the comparison group, holding the size of the treatment group fixed. The nearest neighbor method continues to evaluate the prediction errors only for the nearest neighbor observations, though these will get closer, on average, to the treated observations as the size of the comparison sample increases. In contrast, the three methods we propose evaluate the prediction error at all of the comparison observations (with some, of course, receiving more weight than others). As a result, we expect the relative performance of our methods to improve as the number of comparison observations increases.

4.5 Locally varying bandwidths

It is well known – see, e.g., Herrmann (1997) – that nonparametric kernel regression estimators exhibit increased bias around peaks in the regression curve and increased variance in regions with a low density of the explanatory variable. Bandwidth selection schemes that select a separate bandwidth for each point attempt to overcome these problems with the standard fixed bandwidth estimator. Employing such locally varying bandwidths in the context of sparse and/or rough data has generated a large literature in statistics; see e.g. Müller and Stadtmüller (1987), Fan and Gijbels (1995), and Fan, Hall, Martin, Patil (1996).[12]

In this paper, we implement locally varying bandwidths using a method inspired by the standard "plug-in" approach in the literature. As described in, e.g., Song et al. (1995) and Loader (1999) the "plug-in" arises by solving for the bandwidth that

[12] Nearest-neighbor matching can be thought of as a kernel smoother with the uniform kernel and a data-dependent bandwidth.

minimizes a second-order Taylor series expansion of the asymptotic MSE of a regression function for a generic data generating process at a given point as a function of the sample size and some parameters. The bandwidth that solves this problem is given by

$$h(\rho) = \left[\frac{\sigma_0^2(\rho)}{n_0 \left[m^k(\rho) \right] f_{T=0}(\rho)} c(K) \right]^{\frac{1}{5}},$$

where

$$c(K) = \frac{(k!) \int [K(\rho)]^2 d\rho}{2k \int \rho^2 K(\rho) d\rho},$$

and $K(\cdot)$ is a k^{th} order kernel, $\sigma_0^2(\rho)$ is the conditional variance, n_0 is the size of the comparison sample, $m^k(\rho)$ is the k^{th} derivative of $m(\rho)$, and $f_{T=0}(\rho)$ is the density of the propensity scores for the untreated units.

We avoid the computational burden of calculating the densities and the derivatives by using the (admittedly somewhat atheoretic) approximation

$$\hat{h}(\rho_i) = h_{CV} \left\{ \frac{\rho_i}{1-\rho_i} \right\}^{1/5},$$

where h_{CV} denotes the bandwidth from conventional cross-validation. Our approximation draws inspiration from the (very) similar approximation in equation (2.7) of Song et al. (1995), which we modify by adding reweighting based on the distribution of the propensity scores of the treated units. The reweighting relies on the fact that, as noted in Heckman and Todd (1995),

$$\frac{f_{\rho|T=0}(\rho)}{f_{\rho|T=1}(\rho)} \propto \frac{1-\rho}{\rho}.$$

Using this method, we proceed to estimate the same parameter but with $W_h(\rho_i, \rho_j)$ in equation (2) replaced with $W_{h(\rho_i)}(\rho_i, \rho_j)$. In regions rich in comparison units (i.e. the low propensity score regions), the estimation of the expected counterfactual outcome in the matching estimator relies on relatively narrow bandwidths, whereas in regions with few comparison units (i.e. the high propensity score regions) the estimation relies on wider bandwidths.

5. Monte Carlo Analysis

5.1 Design of the Monte Carlo analysis

In this section we examine the finite sample properties of local constant and local linear matching estimators when choosing bandwidths using the conventional method, our three proposed fixed bandwidth selection methods that weight the criterion function based on the distribution of treated units, the nearest neighbor method inspired by Bergemann et al. (2005), and our locally varying bandwidth approach.

Our Monte Carlo design consists of twelve different settings, where each setting corresponds to a combination of two propensity score densities, $f_{T=1}(\rho)$ for the treated units and $f_{T=0}(\rho)$ for the untreated units, and a conditional mean function $E(Y_0 \mid \rho) = m(\rho)$. We follow Frölich's (2004) specification of the propensity score and use $\hat{P}(X) = \rho = \alpha + \beta X$, where α is the parameter that controls the number of treated units relative to the number of untreated units, β is the parameter that controls the spread of the propensity score values, and X is a univariate covariate drawn from the Johnson S_B distribution (Johnson, Kotz, and Balakrishnan, 1994). Because the support of the

propensity scores is $(\alpha, \alpha+\beta)$, we re-scaled the scores by $(\rho-\alpha)/\beta$ to ensure that its support always lies in (0,1). As in Frölich (2004), we use known rather than estimated propensity scores.

Figure 2A displays three different density designs corresponding to three different combinations for (α, β): {(0,1), (0.25, 0.5), (0.1, 0.3)}. These parameter values generate combinations of density functions that differ in both the amount of separation between the treated and untreated distributions and the ratio of control to treated observations. The appendix provides the formulae for the densities. The first design, D1, has the strongest separation among the three, design D2 somewhat less separation and design D3 the least separation. Both designs D1 and D2 have a ratio of treated to control units equal to approximately 1:1; in contrast, design D3 has a ratio of treated to control units of approximately 1:3. The exact sample sizes equal 202 treated and 198 untreated for D1 and D2, and 108 treated and 292 untreated for D3. These represent small sample sizes indeed relative to what one would want when applying non-parametric methods; we use small sample sizes here to accentuate the performance differences between the different methods. Regardless of the bandwidth selection method and estimator employed, we expect the third design to have the smallest MSE for the matching estimator.

Figure 2B illustrates the four different conditional mean functions for the untreated outcome that we consider. These differ in their monotonicity and their degree of non-linearity. In particular, the first regression curve (M1) is linear, the second (M2) is concave and free of any local roughness, the third (M3) is highly nonlinear and the fourth (M4) has a bimodal shape with the largest "bump" placed in a region with relatively high propensity scores and thus dense in treated units. The appendix provides

the exact formulae for the regression functions. The first and second conditional mean functions represent the most realistic cases in most contexts; the others serve to test the bandwidth selection methods under relatively extreme circumstances.

We sample observations from each one of the four regression curves and then add a lognormal error, which when transformed to a normal random variable has mean zero and standard deviation of 3. This error distribution implies a large variance for the outcome variables, consistent with the usual situation when using earnings as a dependent variable in program evaluations.[13]

The estimation of the MSE proceeds as follows: First, we draw a sample of size $k = n_0 + n_1$ where n_0 indicates the number of comparison group observations and n_1 indicates the number of treated observations. We then compute the (true) expected values of the counterfactual outcomes for the given propensity scores and conditional mean function. Next, we simulate the bias and variance of the matching estimator using Monte Carlo samples of size $k = n_0 + n_1$. Within a simulation, the process for each Monte Carlo sample proceeds as follows. First, we draw $\{\rho_i, Y_{i0}\}_{i=1}^{n_1}$ for the treated units, using the conditional mean impact function given in the appendix to obtain the corresponding values of Y_{i1}.[14] Then we draw $\{\rho_j, Y_{j0}\}_{j=1}^{n_0}$ for the untreated units. Next we determine the bandwidth choice implied by each selection method for the current Monte Carlo

[13] The non-zero mean of the error term, which equals about 90, has no effect on the results; in the parametric analogue to our non-parametric regressions it would get absorbed in the intercept.

[14] Given our focus on the mean effect of treatment on the treated parameter, the choice of fixed versus heterogeneous treatment effects and of the particular form of the treatment effects within these classes, has no effect on the relative performance of the alternative bandwidth selection methods, which depend only on the distributions of the untreated units used to estimate the expected counterfactual mean outcome of the treated units.

sample.[15] Finally, we compute the matching estimates corresponding to each of the selected bandwidths and save them. Once we have completed this process for all of the Monte Carlo samples, we compute the estimated bias and variance associated with each bandwidth selection algorithm by comparing the estimated expected counterfactual outcomes they produce to the true values. In the case of the locally varying bandwidth algorithm we use the mean (over the Monte Carlo simulation samples) of the bandwidths from conventional cross-validation as our value of h_{CV} in the formula for $\hat{h}(\rho)$ when computing the matching estimates. We repeat the entire exercise for both the local constant and local linear matching estimators using both the Epanechnikov and Gaussian kernel functions.[16]

5.2 Results from the Monte Carlo analysis

Tables 1 and 2 present the estimated MSEs for the counterfactual outcomes of the treated observations (denoted MSETT in the tables) and the means of the selected bandwidths for the Epanechnikov and Gaussian kernels, respectively. Panel A of each table presents results using the local constant kernel while Panel B of each table presents results using the local linear kernel. Within each panel, the first five rows correspond to design D1, the second five rows correspond to design D2, and the last five rows correspond to design D3. Because we want to compare the relative performance of alternative bandwidth selection methods rather than examining their performance relative to an ideal standard, we do not present MSE estimates based on the optimal bandwidth in the population.

[15] The grid for the bandwidth search in the local constant case equals [0.01, 0.03,…, 0.51] while in the local linear case it equals [0.03,0.05,…,0.61].

[16] As Silverman (1986) notes, both kernel functions are nonnegative everywhere and almost equally efficient on the basis of MISE.

Four general patterns appear in the Monte Carlo results whether or not we take account of the location of the treated units in bandwidth selection. First, the MSE of the matching estimators increases when the extent of the overlap between the treated and untreated propensity score densities decreases. To see this, compare the most difficult density design, D1, in which most of the mass of the treated units lies in the region with little mass for the comparison group, with the design with the greatest overlap, D3.

Second, the Gaussian kernel performs better than the Epanechnikov kernel, particularly in the most difficult designs, D1 and D2. This result holds for local constant and local linear matching and all five bandwidth selection approaches. Taking distant observations into account in the estimation helps with the small sample sizes. On the other hand, when the distribution of propensity scores between the treated and untreated units is not so disparate and the number of untreated units is much higher than the treated ones, as in D3, the difference between the Gaussian and the Epanechnikov kernel disappears.

Third, local constant matching consistently performs better than local linear matching, particularly in the most difficult density designs with many treated and untreated observations near the boundary at zero. As discussed in Seifert and Gasser (1996), in regions of sparse data, the denominator of the weights implicit in the local linear estimator can end up quite small, leading to very large values of the ratio and thereby very large values of the MSE for particular observations. As the simulations make clear, the Gaussian kernel partially ameliorates this problem by drawing on distant observations ruled out by the compact support implicit in the Epanechnikov kernel.[17] As

[17] Alternatively, one can modify the local linear estimator by adding a ridging term to its denominator. The resulting ridge regression estimator represents a weighted average of the local constant and the local linear

a result, the MSEs for the local linear estimator end up smaller with the Gaussian kernel than the Epanechnikov kernel, often substantially so.

Fourth, the mean bandwidths selected by all of the methods behave in expected ways. The algorithms select, on average, larger bandwidths for the Epanechnikov kernel (due to its compactness) than for the Gaussian kernel. They also select larger bandwidths for local linear regression than for local constant regression because the former requires the estimation of more local parameters. Particularly in designs D1 and D2, we find that all of the methods generally select narrower bandwidths for the regression functions with the non-linearities located in regions with many treated units, namely M3 and M4. In contrast, we tend to observe the largest bandwidths for the linear and convex regression functions, M1 and M2.

Now consider the relative performance of the various bandwidth selection methods. In general, the methods that take account of the location of the treated units select larger bandwidths than conventional cross-validation. This comes as no surprise; by focusing on regions with more treated units, these methods also focus on regions with fewer untreated units, and so select larger bandwidths.

In addition to this general pattern, we observe important differences in the mean bandwidths selected by the different methods. The variable-weights method yields larger bandwidths than those emerging from the fixed-weight methods, where the latter typically differ only modestly from those selected by conventional cross-validation. On the other hand, the nearest-neighbor approach yields the largest bandwidth values for most combinations of regression functions and density designs. This tendency appears

---

regression estimators. See Seifert and Gasser (1996) for the statistical details and Frölich (2004) for an application in a matching context.

most strongly for local linear with the Epanechnikov kernel, where this method sometimes chooses bandwidths almost twice as large (on average) as those selected by the fixed weight and variable weight methods.

Do these differences in selected bandwidths imply substantial efficiency differences among the alternative bandwidth selection models? Four key patterns emerge from Tables 1 and 2. First, in general we observe rather modest efficiency gains when accounting for the location of the treated units in the selection of the bandwidths. Such gains appear most often for the most difficult density design, D1, where conventional cross-validation almost never has the lowest average MSE over our 500 simulations and when using the Epanechnikov kernel.

Second, of the five methods that we implemented, the locally varying bandwidth selection method nearly always shows the lowest MSE. This result holds for local constant and local linear kernel matching, for all three density combinations and for both kernels. Using larger bandwidths in areas where the distribution of the untreated units has low density and smaller bandwidths in high-density areas yields significant reductions in the MSE of the matching estimators.

Third, the relative efficiency of the alternative bandwidth selection methods depends strongly on the choice of kernel functions. In both Table 1 and Table 2 the locally varying and nearest neighbor approaches almost always produce the lowest average MSE when using the Epanechnikov kernel but less frequently when using the Gaussian kernel. The larger bandwidths selected (on average) by the nearest neighbor approach appear to provide a real benefit in the case of the Epanechnikov kernel, as it helps the matching estimator to avoid small denominators even when the kernel assigns

zero weight to all relatively distant observations. The Gaussian kernel avoids this problem by assigning a non-zero weight to distant observations; as a result, for this kernel the large bandwidths the nearest neighbor method selects make its performance relatively worse.

Fourth, perhaps surprisingly, the relative performance of the various bandwidth selection estimators does not vary in any systematic way with the shape of the regression function, with the exception of the locally varying bandwidths, which have trouble with highly non-linear regression function M3. This finding stands in sharp contrast to the result for the kernel functions just described.

Overall, the Monte Carlo evidence suggests the value, in many contexts, of taking account of the location of the treated units when choosing a bandwidth for a local constant or local linear kernel matching estimator. In general, the largest benefits accrue when the data feature disparate propensity score (and thus covariate) distributions in the treatment and comparison samples. In contrast, in contexts with little separation between the treated and untreated units, accounting for the location of the treated units in the selection of the bandwidth becomes less relevant.

6. Empirical Application

No paper on matching methods would be complete without an analysis of the data from LaLonde's (1986) famous paper. He combines experimental data from the U.S. National Supported Work Demonstration (hereafter NSW) with non-experimental comparison groups drawn from two major survey data sets – the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID) – in order to examine the performance

of alternative non-experimental evaluation estimators. LaLonde shows that simply conditioning on a handful of variables in a linear regression context, or doing differences-in-differences, does not suffice to solve the selection problem in these data.[18]

Dehejia and Wahba (1999, 2002) apply propensity score matching methods to a subset of the data on one of the two demographic groups examined by LaLonde (1986) and find low biases (with large standard errors). Smith and Todd (2005a,b) revisit the Dehejia and Wahba (1999, 2002) analyses and demonstrate a high level of sensitivity to the estimates obtained from matching in these data along many dimensions. They find this sensitivity unsurprising given the application of semi-parametric methods to very small samples. They also conclude that the CIA does not hold in this context; this conclusion also seems unsurprising when one realizes that the NSW program served ex-convicts, ex-addicts, long-term welfare recipients and high school dropouts while the data at hand contain no measures of crime or punishment, no measures of welfare receipt, no measures of current or past drug use and no measure of ability.

Despite the small number of treated units and the likely failure of the CIA, the ease of use of these data and the general familiarity with them among applied researchers has led to their wide use in papers, like this one, that examine methodological innovations in matching. Thus, for comparability with the existing literature and because we want to consider a real data environment in addition to our Monte Carlo analysis, in this section we examine the performance of all bandwidth selection methods proposed here, using the data from Dehejia and Wahba (1999, 2002).

[18] As noted in Smith and Todd (2005a) LaLonde (1986) also applies Heckman's (1979) two-step estimator of the bivariate normal selection model. Due to multiple issues of implementation, his study provides no information about the performance of that estimator. See Heckman and Hotz (1986) for further analysis of the LaLonde data and Hollister et al. (1984) for an overview of the NSW experiment.

Their sub-sample, which we denote by "DW", includes only individuals with a valid value for earnings in 1975 (just before the program) and only a zero value for earnings in "1974" (actually months 13-24 before random assignment). We focus on their sample as the CIA has the greatest plausibility for this group. For simplicity, we employ the same (logit) propensity score specifications as in Dehejia (2005); see the notes to Table 3 for details. After estimating the scores, we impose the common support condition using the trimming method developed in Heckman, Ichimura, Smith and Todd (1998), which estimates separate densities for the scores in the treated and untreated samples and then drops all observations whose score implies a zero estimated density in either distribution as well as the observations with the lowest five percent of the non-zero estimated density values. This procedure leads us to drop about 10 percent of the treated units, in addition to dropping a fraction of the comparison group sample similar to that dropped in Dehejia and Wahba (1999). We apply the various bandwidth selection algorithms to this reduced sample.[19]

Table 3 presents the MSE of the treatment effects over 100 bootstrap simulations of the DW data. We obtain untreated outcomes for the treated units for use in calculating our MSEs by subtracting off the experimental impact estimated over the empirical common support region. Panel A of Table 3 presents impact estimates obtained using local constant matching while Panel B presents estimates obtained using local linear matching. The first column indicates the combination of treatment group and comparison

[19] The bandwidths for the grid search equal [0.05, 0.10, …, 2.00] for the local constant kernel matching and [0.40, 0.45, …, 4.00] for the local linear kernel matching.

group, while the second column indicates the kernel function. The corresponding experimental impact estimates appear in the notes to Table 3.[20]

Three main patterns emerge from Table 3. First, as in Smith and Todd (2005a,b) the details of the estimation procedure, in this case the bandwidth selection algorithm, matter for the obtained estimates in the NSW data because of the small sample size and the large variance of the earnings outcome variable. The estimated MSEs depict large variability across the bandwidth selection methods, matching estimators, and even kernel used. Indeed, the Epanechnikov kernel for the local linear regression estimates using the simulated DW-CPS and DW-PSID data yields such bizarre results for all bandwidth selection approaches that we do not report these results.

Second, the choice among the conventional, variable weights, fixed weights and locally varying bandwidth selection algorithms becomes much clearer in the simulated NSW data than it was in the simulations reported in Tables 1 and 2. In general, the variable weights approach and the fixed weights approach based on the density of the treated units yield the lowest MSEs for both the DW-CPS and DW-PSID datasets. In contrast, the conventional approach and the nearest neighbor approach imply the largest MSEs. Relative to the Monte Carlo simulations, the locally varying bandwidth does not do especially well compared to the other methods, as it falls in the middle of the five approaches.

Third, Table 3 reveals that in the NSW data, the nearest neighbor method performs much less well than the other four bandwidth choice schemes in MSE terms. In general, it substantially oversmooths relative to the other approaches, as a comparison of

[20] We do not employ the second of the two fixed weights methods with the NSW data given that we match on the log-odds ratio.

the selected bandwidths clearly illustrates. Thus, despite its reasonable performance in the simulations in Tables 1 and 2, we suggest avoiding the nearest neighbor approach in small data sets with highly variable outcome measures.

Overall, the lessons from our foray into the NSW data are similar to those of our Monte Carlo analysis. Both the varying-weighting cross-validation and the locally varying bandwidths perform relatively well when applied to the NSW data and, at the same time, the conventional cross-validation approach does relatively poorly. The desirability of the nearest neighbor approach remains ambiguous given its poor performance in the context of the NSW data.

7. Conclusions

In estimating the counterfactual mean regression function in an evaluation context, the choice of smoothing parameter should reflect the density of the untreated observations that look like the treated observations as well as the smoothness of the regression function in regions of high treated unit density. Although this insight applies to a variety of estimators for the counterfactual mean, in this paper we focus on the use of local constant and local linear matching estimators. We propose three alternative methods for incorporating the location of the treated units into the bandwidth choice process. We then compare among our three methods, a related alternative method inspired by Bergemann et al. (2005), a version of locally varying bandwidths, and conventional cross-validation by conducting a Monte Carlo analysis and by applying them to the oft-examined NSW data.

The Monte Carlo analysis suggests that taking account of the location of the treated units has enough of a payoff in terms of the MSE to make it worth doing in applied work as a general rule, particularly in contexts with dissimilar covariate distributions in the treatment and comparison groups. The NSW data lead to largely similar conclusions, but cast doubt on the value of the nearest neighbor method and the locally varying bandwidths in small, highly variable samples. Overall, the variable weight bandwidth selection method and (subject to the caveat just noted) the locally varying bandwidths display the best performance. We thus recommend these methods along with the Gaussian kernel.

References

Bergemann, Annette, Bernd Fitzenberger and Stefan Speckesser. 2005. "Evaluating the Dynamic Effects of Training Programs in East Germany Using Conditional Difference-in-Differences." IZA Discussion Paper No. 1848.

Black, Dan and Jeffrey Smith. 2004. "How Robust is the Evidence on the Effects of College Quality? Evidence from Matching?" *Journal of Econometrics* 121(1): 99-124.

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

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

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

DiNardo, John, Nicole Fortin and Thomas Lemieux. 1996. "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach." *Econometrica* 64(5): 1001-1044.

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

Fan, Jianqing, Theo Gasser, Irene Gijbels, Michael Brockmann, and Joachim Engel. 1997. "Local Polynomial Regression: Optimal Kernels and Asymptotic Minimax Efficiency" *Annals of the Institute of Mathematical Statistics* 49: 79-99.

Fan, Jianqing, P. Hall, M. Martin, and P. Patil. 1996. "On the Local Smoothing of Nonparametric Curve Estimation". *Journal of the American Statistical Association*, 91(433), pp. 258-266.

Fan, Jianqing and Irene Gijbels. 1995. "Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation" *Journal of the Royal Statistical Society*. 57(2): 371-394

Flossman, Anton. 2006. "Empirical Bias Bandwidth Choice for Local Polynomial Matching Estimators." Unpublished manuscript, Universität Kontstanz.

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

Frölich, Markus. 2005. "Matching estimators and Optimal Bandwidth Choice." *Statistics and Computing* 15: 197-215.

Hall, Peter. 1983. "Large Sample Optimality of Least-Squares Cross-Validation in Density Estimation." *Annals of Statistics* 11: 1156-1174.

Hall, Peter, Jeffrey Racine and Qi Li. 2004. "Cross-Validation and the Estimation of Conditional Probability Densities." *Journal of the American Statistical Association* 99(486): 1015-1026.

Heckman, James. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47(1): 153-161.

Heckman, James and Joseph Hotz. 1989. "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association* 84(408): 862-880.

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

Heckman, James, Hidehiko Ichimura, and Petra Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economics Studies* 64(4): 605-654.

Heckman, James, Hidehiko Ichimura, and Petra Todd. 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 65(2): 261-294.

Heckman, James and Petra Todd. 1995. "Adapting Propensity Score Matching and Selection Models to Choice-Based Samples." Unpublished manuscript, University of Chicago.

Herrmann, Eva. 1997. "Local Bandwidth Choice in Kernel Regression Estimation." *Journal of Computational and Graphical Statistics*. 6(1): 35-54.

Hirano, Keisuke, Guido Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score." *Econometrica* 71: 1161-1189.

Hollister, Robinson, Peter Kemper and Rebecca Maynard. 1984. *The National Supported Work Demonstration*. Madison: University of Wisconsin Press.

Horvitz, D. and D. Thompson. 1952. "A Generalization of Sampling Without Replacement from a Finite Universe." *Journal of the American Statistical Association*. 47(260): 663-685.

Ichimura, Hidehiko and Oliver Linton. 2001. "Asymptotic Expansions for some Semiparametric Program Evaluation Estimators." CEMMAP Working Paper CWP04/01.

Imbens, Guido. 2004. "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review." *Review of Economics and Statistics* 86:4-30.

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

Johnson, Norman, Samuel Kotz, and N. Balakrishnan. 1994. *Continuous Univariate Distributions. Vol.1, 2nd Edition*. New York: Wiley.

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

LaLonde, Robert. 1986. “Evaluating the Econometric Evaluation of Training Programs with Experimental Data.” *American Economics Review* 76(4): 604-620.

Lee, Wang-Sheng. 2006. “Propensity Score Matching and Variations on the Balancing Test.” Unpublished manuscript, University of Melbourne.

Li, Qi and Jeffrey Racine. 2007. *Nonparametric Econometrics. Theory and Practice*. Princeton: Princeton University Press.

Loader, Clive. 1999. “Bandwidth Selection: Classical or Plug-in?” *Annals of Statistics* 70(2): 415-438.

Müller, Hans-Georg and Ulrich Stadtmüller. 1987. “Variable Bandwidth Kernel Estimators of Regression Curves.” *Annals of Statistics*. 15(1):182-201

Pagan, Adrian and Aman Ullah. 1999. *Nonparametric Econometrics*. New York: Cambridge University Press.

Rosenbaum, Paul and Donald Rubin. 1983. “The Central Role of the Propensity Score In Observational Studies For Causal Effects.” *Biometrika* 70(1): 41-55.

Ruppert, David. 1997. “Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation.” *Journal of the American Statistical Association* 92: 1049-1062.

Seifert, Burkhardt and Theo Gasser. 1996. “Finite-Sample Variance of Local Polynomials: Analysis and Solutions.” *Journal of the American Statistical Association* 91: 267-275.

Silverman, Bernard. 1986. *Density Estimation for Statistics and Data Analysis*. London and New York: Chapman and Hall.

Smith, Jeffrey and Petra Todd. 2005a. “Does Matching Overcome LaLonde’s Critique of Non-Experimental Estimators?” *Journal of Econometrics* 125(1-2): 305-353.

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

Stone, M. 1974. “Cross-Validatory Choice and Assessment of Statistical Predictions.” *Journal of the Royal Statistic Society, Series B* 36(2): 111-147

Stone, C. 1984. “An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates.” *Annals of Statistics* 12: 1285-1297.

Song, K., H. Muller, A. Clifford, H. Furr and J. Olson. 1995. “Estimating Derivatives of Pharmacokinetic Response Curves with Varying Bandwidths.” *Biometrics.* 51(1):12-20

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

. **Figure 1: Treated and Untreated Density Distributions**

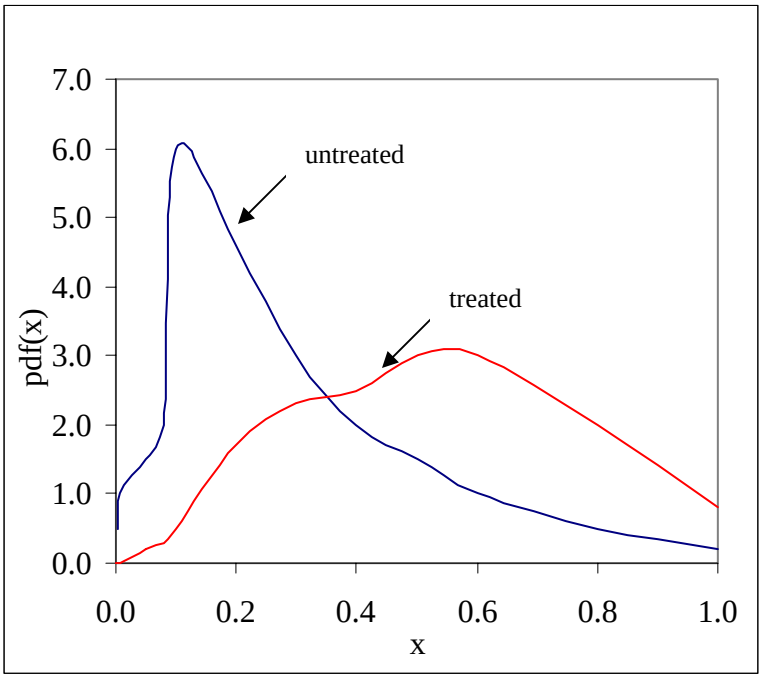


Figure 2A: Monte Carlo Density Distributions

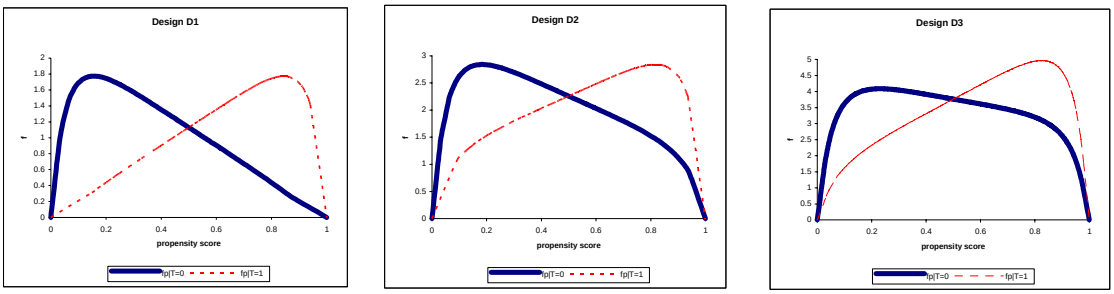
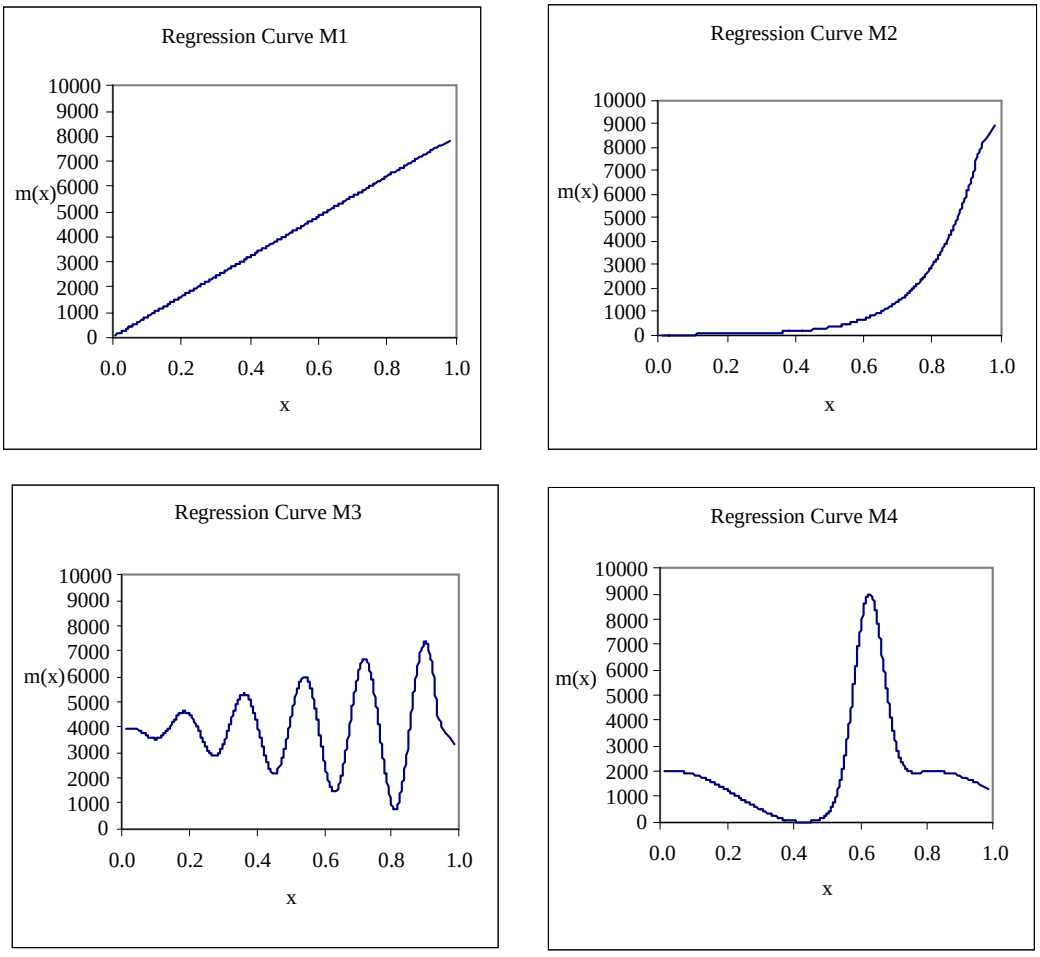


Figure 2B: Monte Carlo Regression Curves



Note: The figures are based on the density design D1

Table 1: Mean Squared Error of Estimated Counterfactual Outcomes (MSECTT /1000)

Panel A: Local Constant – Epanechnikov Kernel

		Conventional		Variable Weights		Fixed Weights				NN		Locally varying
						$f_{T=1}(\rho)$		$\rho/(1-\rho)$				
		h	*MSECTT*	*h*	*MSECTT*	*h*	*MSECTT*	*h*	*MSECTT*	*h*	*MSECTT*	*MSECTT*
Design	M1	0.038	61	0.065	47	0.040	57	0.040	57	0.069	35	35
D1	M2	0.035	241	0.041	236	0.036	241	0.034	251	0.047	192	153
	M3	0.030	30	0.032	33	0.030	30	0.030	30	0.042	57	58
	M4	0.030	39	0.032	41	0.029	43	0.031	37	0.039	35	30
	Mean	**0.033**	**93**	**0.043**	**89**	**0.034**	**93**	**0.034**	**94**	**0.049**	**80**	**69**
Design	M1	0.044	25	0.041	24	0.052	24	0.043	25	0.078	21	24
D2	M2	0.035	52	0.038	50	0.035	52	0.033	55	0.048	37	31
	M3	0.029	32	0.032	35	0.029	32	0.029	32	0.037	45	46
	M4	0.029	34	0.031	33	0.028	34	0.031	33	0.041	24	31
	Mean	**0.034**	**36**	**0.036**	**36**	**0.036**	**36**	**0.034**	**36**	**0.051**	**32**	**33**
Design	M1	0.028	39	0.041	36	0.034	39	0.025	39	0.062	42	39
D3	M2	0.017	43	0.034	41	0.018	43	0.012	43	0.047	43	46
	M3	0.027	24	0.026	24	0.024	25	0.010	29	0.045	18	22
	M4	0.026	45	0.028	45	0.024	46	0.017	47	0.048	45	44
	Mean	**0.025**	**38**	**0.032**	**37**	**0.025**	**38**	**0.016**	**40**	**0.051**	**37**	**38**

Panel B: Local Linear – Epanechnikov Kernel

		Conventional		Variable Weights		Fixed Weights				NN		Locally Varying
						$f_{T=1}(\rho)$		$\rho/(1-\rho)$				
		h	*MSECTT*	*h*	*MSECTT*	*h*	*MSECTT*	*h*	*MSECTT*	*h*	*MSECTT*	*MSECTT*
Design	M1	0.065	156	0.106	90	0.069	152	0.073	50	0.126	23	23
D1	M2	0.075	396	0.089	204	0.074	397	0.074	287	0.100	57	41
	M3	0.085	515	0.091	840	0.084	520	0.091	659	0.120	457	459
	M4	0.059	372	0.092	95	0.059	160	0.063	159	0.111	40	45
	Mean	**0.071**	**360**	**0.095**	**307**	**0.072**	**307**	**0.075**	**289**	**0.114**	**144**	**142**
Design	M1	0.087	280	0.081	25	0.092	280	0.096	280	0.152	31	23
D2	M2	0.091	272	0.074	39	0.092	272	0.085	272	0.130	30	30
	M3	0.074	89	0.059	62	0.075	92	0.064	74	0.101	144	126
	M4	0.065	336	0.075	41	0.065	336	0.071	336	0.102	40	42
	Mean	**0.079**	**244**	**0.072**	**42**	**0.081**	**245**	**0.079**	**241**	**0.121**	**61**	**55**
Design	M1	0.072	39	0.094	38	0.080	40	0.077	40	0.132	45	36
D3	M2	0.056	43	0.084	44	0.057	43	0.050	43	0.148	52	39
	M3	0.063	38	0.108	37	0.071	36	0.055	38	0.101	44	41
	M4	0.050	53	0.066	52	0.049	53	0.053	53	0.102	61	54
	Mean	**0.060**	**43**	**0.088**	**43**	**0.064**	**43**	**0.059**	**44**	**0.121**	**51**	**43**

Notes: The first two columns indicate the density and the regression curve. In each column, the averages over the regression curves in each design appear in bold.

Table 2: Mean Squared Error of Estimated Counterfactual Outcomes (MSECTT /1000)

Panel A: Local Constant – Gaussian Kernel

		Conventional		Variable Weights		Fixed Weights				NN		Locally Varying
						$f_{T=1}(\rho)$		$\rho/(1-\rho)$				
		h	*MSECTT*	*h*	*MSECTT*	*h*	*MSECTT*	*h*	*MSECTT*	*h*	*MSECTT*	*MSECTT*
Design	M1	0.013	21	0.026	24	0.015	21	0.015	21	0.029	23	20
D1	M2	0.011	70	0.019	86	0.011	70	0.010	69	0.021	84	70
	M3	0.010	96	0.021	95	0.010	96	0.010	96	0.010	107	96
	M4	0.010	29	0.010	31	0.010	29	0.010	29	0.010	31	29
	Mean	**0.011**	**54**	**0.019**	**59**	**0.012**	**54**	**0.011**	**54**	**0.018**	**61**	**54**
Design	M1	0.017	23	0.015	23	0.021	23	0.017	23	0.033	21	23
D2	M2	0.011	28	0.014	30	0.011	27	0.010	28	0.020	28	27
	M3	0.010	56	0.010	56	0.010	56	0.010	56	0.010	55	52
	M4	0.010	31	0.010	31	0.010	31	0.010	33	0.014	26	32
	Mean	**0.012**	**35**	**0.012**	**35**	**0.013**	**34**	**0.012**	**35**	**0.019**	**33**	**34**
Design	M1	0.028	39	0.041	36	0.034	39	0.014	38	0.025	42	39
D3	M2	0.010	42	0.018	41	0.011	41	0.010	42	0.014	42	45
	M3	0.010	23	0.010	23	0.010	23	0.010	23	0.010	23	24
	M4	0.010	46	0.010	46	0.010	46	0.010	46	0.015	50	45
	Mean	**0.015**	**38**	**0.020**	**37**	**0.016**	**37**	**0.011**	**37**	**0.016**	**39**	**38**

Panel B: Local Linear – Gaussian Kernel

		Conventional		Variable Weights		Fixed Weights				NN		Locally Varying
						$f_{T=1}(\rho)$		$\rho/(1-\rho)$				
		h	*MSECTT*	*h*	*MSECTT*	*h*	*MSECTT*	*h*	*MSECTT*	*h*	*MSECTT*	*MSECTT*
Design	M1	0.031	19	0.040	19	0.033	19	0.034	19	0.051	22	20
D1	M2	0.034	30	0.036	32	0.034	30	0.034	30	0.045	34	37
	M3	0.039	434	0.059	446	0.039	433	0.040	445	0.047	406	380
	M4	0.031	38	0.034	38	0.031	38	0.031	38	0.044	37	36
	Mean	**0.034**	**130**	**0.042**	**134**	**0.034**	**130**	**0.035**	**133**	**0.047**	**125**	**118**
Design	M1	0.038	26	0.036	25	0.039	26	0.039	26	0.062	31	23
D2	M2	0.037	33	0.036	35	0.038	33	0.036	33	0.053	29	29
	M3	0.041	108	0.011	112	0.042	110	0.037	102	0.045	132	143
	M4	0.032	39	0.036	39	0.032	39	0.033	39	0.042	39	41
	Mean	**0.037**	**52**	**0.030**	**53**	**0.038**	**52**	**0.036**	**50**	**0.051**	**58**	**59**
Design	M1	0.035	40	0.046	40	0.037	40	0.036	40	0.052	45	36
D3	M2	0.033	46	0.043	46	0.033	46	0.033	46	0.060	51	41
	M3	0.045	35	0.048	33	0.047	34	0.030	40	0.041	37	43
	M4	0.032	53	0.036	52	0.033	53	0.033	53	0.047	59	53
	Mean	**0.036**	**44**	**0.043**	**43**	**0.038**	**43**	**0.033**	**45**	**0.050**	**48**	**43**

Notes: The first two columns indicate the density and the regression curve. In each column, the averages over the regression curves for each design appear in bold.

Table 3: Mean Squared Error for the National Support Work Demonstration Program (MSETT/1000)

Sample	Kernel	Conventional		Variable Weights		Fixed Weights (estimated density)		NN		Locally Varying
		h	*MSETT*	*h*	*MSETT*	*h*	*MSETT*	*h*	*MSETT*	*MSETT*
Panel A: Local Constant										
DW-PSID	Epanechnikov	0.22	1351	0.68	812	0.67	1100	0.63	1304	931
	Gaussian	0.14	798	0.20	698	0.31	847	1.60	6744	858
DW-CPS	Epanechnikov	0.15	1054	0.39	844	0.18	1038	1.43	1054	1038
	Gaussian	0.11	774	0.19	715	0.12	781	1.42	3362	1066
Panel B: Local linear										
DW-PSID	Gaussian	0.41	2293	0.44	756	0.92	814	3.60	1464	876
DW-CPS	Gaussian	0.35	262	0.53	197	0.37	253	3.00	578	903

Notes: The dependent variable is real earnings in 1978. We match on the log odds ratio due to the choice based sampling; see the discussion in Heckman and Todd (1995). Bandwidth selection takes place after imposing the common support condition using the five percent trimming method developed in Heckman, Ichimura and Todd (1998). The experimental impact estimates equal \$1864 for the DW sample with CPS common support and \$2056 for the DW sample with PSID common support. We estimate logit models of participation using the specifications in Dehejia (2005), the experimental treatment group and the comparison group. The DW-PSID model includes married, black, Hispanic, age, education, real earnings in 1975 (RE75), real earnings in "1974" (RE74), married*1(RE75 = 0), and nodegree*1(RE74 = 0), and the DW-CPS model includes married, black, Hispanic, education, age, RE74, RE75, and black*age. Simulations are based on 100 replications. The locally varying bandwidth is based on the rule $h(\rho) = h_{cv}[\hat{f}_{\rho|T=0}(\rho)]^{-1/5}$, which differs from the version in the text because we are matching on the log odds ratio rather than directly on the propensity score. Ties are broken at random.

Appendix

Details of the Monte Carlo Analysis

The propensity score model is specified as $\hat{P}(X) = \rho = \alpha + \beta X$, where the univariate X is drawn from the Johnson S_B distribution defined in Johnson et al. (1994),

$$f_x(x) = \frac{1}{2\sqrt{\pi}(1-x)} \exp\left[-\frac{1}{4} \ln^2\left(\frac{x}{1-x}\right) \right], \; 0 < x < 1.$$

As discussed in Frölich (2004), the resulting density functions for the treatment and comparison groups are given by

$$f_{\rho|T=1}(\rho) = \frac{\rho}{\beta s_1} . f_x\left(\frac{\rho-\alpha}{\beta}\right) \quad \text{and} \quad f_{\rho|T=0}(\rho) = \frac{1-\rho}{\beta s_0} . f_x\left(\frac{\rho-\alpha}{\beta}\right),$$

where $s_1 = E(\rho)$ is the share of treated units and $s_0 = 1 - s_1$.

The first pair of density functions (design D1) is based on $\alpha = 0$ and $\beta = 1$, the second pair (design D2) is based on $\alpha = 0.25$ and $\beta = 0.5$, and the third pair (design D3) is based on $\alpha = 0.1$ and $\beta = 0.3$.

The outcome equations are as follows:

M1: $m_0(\rho) = 0.01 + 8000\rho$

M2: $m_0(\rho) = 0.02 + 10\exp(6\rho)(3\rho)$

M3: $m_0(\rho) = 3900 + 3900\rho\cos(35\rho)$

M4: $m_0(\rho) = 1010 + 1000\sin(8\rho - 5) + 8000\exp(-16(4\rho - 2.5)^2)$

The outcomes for the treated units are given by

$$m_1(\rho) = m_0(\rho) + T(\rho)$$

where $T(\rho) = (10\rho)\exp(6\rho)$.