Busso, Matias; DiNardo, John; McCrary, Justin

Working Paper

New evidence on the finite sample properties of propensity score matching and reweighting estimators

IZA Discussion Papers, No. 3998

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Busso, Matias; DiNardo, John; McCrary, Justin (2009) : New evidence on the finite sample properties of propensity score matching and reweighting estimators, IZA Discussion Papers, No. 3998, Institute for the Study of Labor (IZA), Bonn, https://nbn-resolving.de/urn:nbn:de:101:1-20090304687

This Version is available at:
https://hdl.handle.net/10419/35376

DISCUSSION PAPER SERIES

IZA DP No. 3998

New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators

Matias Busso
John DiNardo
Justin McCrary

February 2009

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators

Matias Busso
Inter-American Development Bank and IZA

John DiNardo
University of Michigan and NBER

Justin McCrary
University of California at Berkeley and NBER

Discussion Paper No. 3998
February 2009

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. 3998
February 2009

ABSTRACT

New Evidence on the Finite Sample Properties of Propensity Score Matching and Reweighting Estimators*

Currently available *asymptotic* results in the literature suggest that matching estimators have higher variance than reweighting estimators. The extant literature comparing the *finite sample* properties of matching to specific reweighting estimators, however, has concluded that reweighting performs far worse than even the simplest matching estimator. We resolve this puzzle. We show that the findings from the finite sample analyses are not inconsistent with asymptotic analysis, but are very specific to particular choices regarding the implementation of reweighting, and fail to generalize to settings likely to be encountered in actual empirical practice. In the DGPs studied here, reweighting typically outperforms propensity score matching.

JEL Classification: C14, C21, C52

Keywords: treatment effects, propensity score, semiparametric efficiency

Corresponding author:

Matias Busso
Inter-American Development Bank
1300 New York Avenue, NW
Washington, DC 20577
USA
E-mail: mbusso@iadb.org

* For comments that improved the paper, we thank Alberto Abadie, Matias Cattaneo, Bryan Graham, Keisuke Hirano, Guido Imbens, Pat Kline, and Jack Porter, but in particular Jeff Smith. We would also like to thank Markus Frölich for providing us copies of the code used to generate the results in his paper.

I. Introduction

A common goal of empirical work is to assess the impact of a non-randomized program on a subpopulation of interest. Empirical estimates of program impacts are often based on matching or reweighting using an estimate of the propensity score, or the conditional probability of treatment given baseline characteristics.[1] Empirical literatures, particularly in economics, but also in medicine, sociology and other disciplines, feature an extraordinary number of program impact estimates based on these estimators. Propensity score matching is particularly popular and has been described by Smith and Todd (2005) as "the estimator *du jour* in the evaluation literature."

Perhaps surprisingly, large sample properties of these estimators have only recently been documented (e.g., Heckman, Ichimura and Todd 1998, Hirano, Imbens and Ridder 2003, Lunceford and Davidian 2004, Abadie and Imbens 2006). Because there are many competing estimators, all of which are consistent, the theoretical literature has also considered which estimators are efficient, in the sense of achieving the efficiency bound established by Hahn (1998) for this problem.

Among other important findings, the large sample literature has established two results that are relevant here. First, a suitable reweighting estimator is asymptotically efficient (Hirano, Imbens and Ridder 2003). Second, pair matching is asymptotically inefficient (Abadie and Imbens 2006).[2]

In a recent article in the *Review of Economics and Statistics*, Frölich (2004) extends the large sample work on this topic and examines the finite sample properties of several propensity score matching and reweighting estimators. To the best of our knowledge, this is the only paper in the literature explicitly comparing reweighting and propensity score matching.[3] The focus of this note is a puzzling feature of Frölich (2004): in the data generating processes (DGPs) he studies, he finds the *reverse* of what is suggested by the large sample results. Summarizing his findings, Frölich (2004) states that the "the weighting estimator turned out to be the worst of all [estimators considered in terms of mean-squared error]... it is far worse than pair matching in all of the designs" (p. 86).

In this note, we resolve this puzzle. We show that the negative conclusions of Frölich (2004) regarding reweighting stem from three specific choices, each of which we argue are undesirable. First, a correct implementation of reweighting normalizes the weights involved so that they sum to one. This is the standard empirical implementation; software typically normalizes weights to sum to one automatically.[4]

[1] Imbens (2004) provides a review of these methods.

[2] Distributional results are available for kernel-based matching estimators, but efficiency has not been considered in the literature.

[3] The dim view Frölich (2004) takes of reweighting, however, has been echoed recently by Freedman and Berk (2008).

[4] For representative empirical applications using normalized weights see DiNardo, Fortin and Lemieux (1996), Bell and Pitt

Frölich (2004) leaves the weights unnormalized. Second, reweighting using the estimated propensity score is more efficient than reweighting using the true propensity score (Hirano et al. 2003), and the resulting efficiency loss can be practically important. Frölich (2004) uses the true propensity score. Third, the consequences of these two choices for the relative MSE of reweighting and pair matching are magnified by the small variance of the outcome equation error used by Frölich (2004) in his simulations. We argue that this variance is too small to be of relevance to empirical practice.

We show that these three choices drive the conclusion of Frölich (2004) that reweighting performs worse than pair matching. Indeed, we show a stronger result: in DGPs more repesentative of the microeconomic settings in which these estimators are typically used than the ones considered in Frölich (2004), a suitable version of reweighting performs at least as well as and usually better than *all* the propensity score matching estimators considered in Frölich (2004).

The remainder of the paper is organized as follows. In Section II, we define notation, estimands, efficiency bounds, and estimators, and we review and extend the large sample theory of reweighting and pair matching estimators. In Section III we use large sample theory to provide intuition for the finite sample results of Frölich (2004). Section IV replicates the main findings of Frölich (2004) and presents new finite sample evidence on the topic. Section V concludes.

II. Background

A. Notation, Estimands and Identification

The starting point for much of the traditional program evaluation literature (e.g., Maddala 1983, Section 9.2, Heckman and Robb 1985, Maddala 1986, and Heckman, Ichimura and Todd 1998, Section 4) is the following DGP for the latent variables $(Y_i(1), Y_i(0), T_i^*)$:

$$Y_i(1) = \mu_1(X_i) + \varepsilon_i \tag{1}$$

$$Y_i(0) = \mu_0(X_i) + \varepsilon_i \tag{2}$$

$$T_i^* = \mu_T(X_i) - u_i \tag{3}$$

where X_i is a vector of baseline characteristics, and u_i and ε_i are mean zero and independent of X_i. Here, $Y_i(1)$ denotes the outcome that would obtain under treatment and $Y_i(0)$ the outcome that would obtain under control. If the latent variable T_i^* exceeds zero, then the unit is assigned to treatment and otherwise is

(1998), Budd and McCall (2001), Biewen (2001), and McCrary (2007).

assigned to control: $T_i = \mathbf{1}(T_i^* > 0)$. The researcher observes $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$, but never the pair $(Y_i(0), Y_i(1))$. The data observed to the researcher are $(Y_i, T_i, X_i)_{i=1}^n$ and are assumed to be independent and identifically distributed (iid) across i.[5] Define the propensity score, or the conditional probability of treatment, as $p(x) \equiv P(T_i = 1 | X_i = x)$. Under equations (1) through (3), we obtain $p(x) = F(\mu_T(x))$, where $F(\cdot)$ is the distribution function for u_i.

In this framework, there are many possible parameters of interest. Frölich (2004) focuses on the effect of treatment on the treated, or TOT $= \mathbb{E}[Y_i(1) - Y_i(0) | T_i = 1] = \mathbb{E}[\mu_1(X_i) - \mu_0(X_i) | T_i = 1]$, and we maintain that focus here. Traditionally, researchers interested in estimating TOT focused on modeling $\mu_0(X_i)$ and $\mu_1(X_i)$ directly using separate regressions for treatment and control units.[6] At the present time, this type of approach is not in wide use in the empirical literature. However, this may soon change; econometric analysis of this approach is the subject of an emerging literature (e.g., Chen, Hong and Tarozzi 2008).

In the framework outlined in equations (1) through (3), propensity score matching and reweighting estimators are $\sqrt{n}$-consistent for TOT and asymptotically normal when u_i and ε_i are independent of one another conditional on the covariates, and when the distribution of the propensity score satisfies a condition known as *strict overlap*.[7] Strict overlap maintains that there exists a constant $c > 0$ such that $c < p(x) < 1 - c$ for almost every x in the support of X_i. This assumption limits the predictability of treatment: no value of the covariates can assure or preclude treatment. The distinction between strict overlap and the weak overlap assumption—that $0 < p(x) < 1$ for almost every x in the support of X_i—is subtle, but important for understanding some aspects of the finite sample performance of these estimators (See Busso, DiNardo and McCrary 2008).

B. Efficiency

Hahn (1998) establishes the semiparametric efficiency bound (SEB) for TOT under conditional independence and weak overlap. The class of estimators to which this bound pertains is the class of regular estimators which are $\sqrt{n}$-consistent for TOT. This efficiency bound can be understood as the supremum of the Crámer-Rao lower bounds associated with regular parametric submodels.[8] If $\ddot{\theta}$ is an estimator that is regular, $\sqrt{n}$-consistent for TOT, and semiparametrically efficient, then $\sqrt{n}(\ddot{\theta} - \theta) \xrightarrow{d} N(0, \text{SEB})$. If $\dot{\theta}$ is

[5] The iid assumption can be relaxed. We assume it here to maintain the connection to Frölich (2004).

[6] See, for example, Blinder (1973), Oaxaca (1973), and Maddala (1983, Section 9.2).

[7] Weaker conditions also suffice. Confusingly, the independence of u_i and ε_i is called different things in the literature. Heckman and Robb (1985) refer to this assumption as selection on observables; Maddala (1986) refers to it as exogeneity of switching; and Rosenbaum and Rubin (1983) refer to it as unconfoundedness.

[8] A regular parametric submodel consists of a parametric specification of the DGP. As noted in Hahn (1998), in the context of average treatment effects, for a parameter vector η and a set of functions $f_t(y|x,\eta)$, $p(x,\eta)$, and $f(x,\eta)$ corresponding to the conditional density of $Y_i(t)$ given $X_i = x$, the propensity score, and the marginal density of X_i, the data (Y_i, T_i, X_i) are

an estimator that is regular, $\sqrt{n}$-consistent for TOT, and does not utilize (correct) parametric knowledge of the joint density for (Y_i, T_i, X_i), then $\sqrt{n}(\dot{\theta} - \theta) \xrightarrow{d} N(0, V)$ with $V \geq \text{SEB}$.[9]

The functional form of Hahn's bound, assuming that the propensity score is not known, is given by

$$\text{SEB} = \mathbb{E}\left[\frac{\sigma_1^2(X_i)p(X_i)}{p^2} + \frac{\sigma_0^2(X_i)p(X_i)^2}{p^2(1-p(X_i))} + \frac{p(X_i)}{p^2}\left(\tau(X_i) - \theta\right)^2\right] \tag{4}$$

where $p = P(T_i = 1)$ and $\tau(X_i) = \mu_1(X_i) - \mu_0(X_i)$ is the covariate-specific treatment effect.

C. Matching Estimators

Frölich (2004) considers many matching estimators: (1) pair matching, (2) kernel matching, (3) local linear matching, (4) ridge matching, and (5) nearest neighbor matching. Kernel, local linear, and ridge matching are implemented using a Gaussian and an Epanechnikov kernel. All take the form

$$\widetilde{\theta} = \frac{\sum_{i=1}^n T_i \left\{Y_i - \widehat{Y}_i(0)\right\}}{\sum_{i=1}^n T_i} \tag{5}$$

where $\widehat{Y}_i(0) = \sum_{j=1}^n (1 - T_j)W(i,j)Y_j$ is the imputed outcome for unit i, based only on observations in the control group (cf., Heckman, Ichimura and Todd 1998, Smith and Todd 2005, Abadie and Imbens 2006).

Different matching estimators involve different choices for the function $W(i,j)$. For example, pair matching on the propensity score sets $W(i,j) = 1$ if control observation j has the propensity score closest to that of treatment observation i, and sets $W(i,j) = 0$ otherwise. Table 1 provides the weighting functions for the matching estimators studied in Frölich (2004).[10] Kernel, local linear, and ridge matching all require selection of a bandwidth, which is done using cross-validation among control observations.[11] Cross-validation is also used to select the number of neighbors for nearest neighbor matching.

---

assumed to be a set of n realizations from a distribution with joint density function $q(y, t, x, \eta_0)$, where

$$q(y, t, x, \eta) = \left[f_1(y|x, \eta)p(x, \eta)\right]^t \left[f_0(y|x, \eta)(1 - p(x, \eta))\right]^{1-t} f(x, \eta)$$

The supremum is taken over $q(\cdot)$ and is finite under strict overlap and conditional independence (Khan and Tamer 2007).

[9] For further discussion of the concept of semiparametric efficiency, see Newey (1990) and references therein.

[10] The notation in the table is as follows: $\mathcal{J}_m(i)$ is the set of m estimated propensity scores among the control observations that are closest to $\hat{p}(X_i)$, where m denotes the number of "neighbors", $K_{ij} = K\left(\left(\widehat{p}(X_j) - \widehat{p}(X_i)\right)/h\right)$, where $K(\cdot)$ is a kernel function and h is a bandwidth, $\hat{\Delta}_i = \widehat{p}(X_i) - \overline{p}_i$ and $\hat{\Delta}_j = \widehat{p}(X_j) - \overline{p}_i$, where $\overline{p}_i = \sum_j (1 - T_j)K_{ij}\widehat{p}(X_j) / \sum_j (1 - T_j)K_{ij}$ is a kernel average of the propensity scores in the control group that are near $\widehat{p}(X_i)$, and r is an adjustment factor suggested by Seifert and Gasser (2000). For a Gaussian kernel, $r = 0.3535$ and for an Epanechnikov kernel, $r = 0.3125$.

[11] There is a small error in Frölich (2004)'s implementation of cross-validation for ridge matching. See Appendix II.

TABLE 1. WEIGHTS USED FOR MATCHING ESTIMATORS

Estimator	Weighting Function, $W(i,j)$
Nearest Neighbor	$\frac{1}{m}\mathbf{1}(\hat{p}(X_j)\in\mathcal{J}_m(i))$
Kernel	$K_{ij}\Big/\sum_j(1-T_j)K_{ij}$
Local Linear	$K_{ij}/\sum_j(1-T_j)K_{ij} + K_{ij}\hat{\Delta}_j\hat{\Delta}_i/\left(\sum_j(1-T_j)K_{ij}\hat{\Delta}_j^2\right)$
Ridge	$K_{ij}/\sum_j(1-T_j)K_{ij} + K_{ij}\hat{\Delta}_j\hat{\Delta}_i/\left(\sum_j(1-T_j)K_{ij}\hat{\Delta}_j^2 + rh\lvert\hat{\Delta}_i\rvert\right)$

D. Reweighting Estimators

The reweighting estimator studied in Frölich (2004) is

$$\widehat{\theta}_F = \frac{\sum_{i=1}^n T_iY_i}{\sum_{i=1}^n T_i} - \frac{\sum_{j=1}^n (1-T_j)W_jY_j}{\sum_{i=1}^n T_j} \tag{6}$$

where $W_j = T_j + (1-T_j)p(X_j)\big/(1-p(X_j))$.[12] As noted, Frölich's version of reweighting is different from the standard empirical implementation of the reweighting estimator, which is instead given by

$$\widehat{\theta} = \frac{\sum_{i=1}^n T_iY_i}{\sum_{i=1}^n T_i} - \frac{\sum_{j=1}^n (1-T_j)\widehat{W}_jY_j}{\sum_{i=1}^n (1-T_j)\widehat{W}_j} \tag{7}$$

where $\widehat{W}_j = T_j + (1-T_j)\widehat{p}(X_j)\big/(1-\widehat{p}(X_j))$.

There are two important differences between equations (6) and (7). First, the weighting function in the counterfactual mean in equation (6), $(1-T_j)W_j\big/\sum_{j=1}^n T_j$, does not sum to one, while that in equation (7), $(1-T_j)\widehat{W}_j\big/\sum_{j=1}^n(1-T_j)\widehat{W}_j$ does. As discussed in the literature, it is preferable to normalize the weights so that they sum to one (e.g., Imbens 2004). Second, the propensity score in equation (6) is the true propensity score, while that in equation (7) is an estimate of the propensity score. This makes an investigation of the behavior of equation (6) less practical than an investigation regarding equation (7). Moreover, as emphasized in Heckman, Ichimura and Todd (1998) and Hirano et al. (2003), there can be efficiency gains associated with using the estimated propensity score, even when the propensity score is known.

Thus, in addition to being somewhat exotic, $\widehat{\theta}_F$ is specifically not recommended. Reflecting these judgements, we refer to $\widehat{\theta}_F$ as "Frölich reweighting" and to $\widehat{\theta}$ as "correct reweighting".

Although reweighting and propensity score matching estimators seem quite different, they share a common structure as weighted least squares estimators. In particular, for weights $\widehat{V}_j$, all of the matching

[12] This is similar to the estimator Hirano et al. (2003, p. 1176) refer to as $\widehat{\tau}_{te}$, if their series logit first-step estimated propensity score had been replaced by the known propensity score.

estimators discussed in Frölich (2004) can be written as

$$\widetilde{\theta} = \frac{\sum_{i=1}^{n} T_i Y_i}{\sum_{i=1}^{n} T_i} - \frac{\sum_{j=1}^{n} (1-T_j)\widehat{V}_j Y_j}{\sum_{j=1}^{n} (1-T_j)\widehat{V}_j}$$

where $\widehat{V}_j = \sum_{i=1}^{n} T_i W(i,j) \big/ \sum_{i=1}^{n} T_i$ is the average weight received by control observation j, on average across all treatment observations i. For details on this result, see Appendix I. Careful inspection of the weights used for matching reveals that they often approximate the weighting function used by reweighting, in a large sample sense.[13] This common structure is consistent with the sense of many applied researchers that, in many applications, propensity score matching and reweighting estimators yield roughly comparable estimates of program impacts. This similarity highlights another reason why the claims of poor peformance of reweighting in Frölich (2004) are puzzling.

E. Distribution Theory for Pair Matching and Reweighting for TOT

Frölich (2004) uses pair matching as a benchmark for the mean-squared error of reweighting and propensity score matching estimators. It is thus instructive to compare the large sample properties of pair matching to those of reweighting, particularly with respect to the DGPs studied in Frölich (2004). Using unpublished results from Abadie and Imbens (2006) and derivations in Appendix I, we have

$$\begin{aligned}
\sqrt{n}(\widetilde{\theta}_{PM} - \theta) &\xrightarrow{d} N\left(0, \text{SEB} + G_{PM}\right) \\
\sqrt{n}(\widehat{\theta} - \theta) &\xrightarrow{d} N\left(0, \text{SEB} + G - G_1 + H_1 - H_2\right) \\
\sqrt{n}(\widehat{\theta}_F - \theta) &\xrightarrow{d} N\left(0, \text{SEB} + G_F + H_1\right)
\end{aligned}$$

where H_1 and H_2 pertain to treatment effect heterogeneity and are zero under homogeneity, and

$$G_{PM} = \frac{1}{2}\mathbb{E}\left[\frac{\sigma_0^2(X_i)}{p^2}\left\{\frac{1}{1-p(X_i)} - (1-p(X_i))\right\}\right] \tag{8}$$

$$G = \mathbb{E}\left[\frac{(\mu_0(X_i) - \mathbb{E}[\mu_0(X_i)|T_i=1])^2}{p^2}\frac{p(X_i)}{1-p(X_i)}\right] \tag{9}$$

$$G_F = \mathbb{E}\left[\frac{\mu_0(X_i)^2}{p^2}\frac{p(X_i)}{1-p(X_i)}\right] \tag{10}$$

are positive terms which can prevent, even under homogenous treatment effects, these estimators from

[13] This can easily be seen, for example, for kernel and nearest neighbor matching.

being fully efficient.[14,15] The term G_1 is also positive and is given by

$$G_1 \quad = \quad \mathbb{C}[\mu_0(X_i), \nu_i Z_i | T_i = 1]\mathbb{E}[\nu_i^2 p(X_i)(1-p(X_i))Z_i Z_i']^{-1}\mathbb{C}[\nu_i Z_i, \mu_0(X_i)|T_i = 1] \tag{11}$$

where $Z_i = (1, X_i')'$, $\nu_i = F'(Z_i'\beta)/(p(X_i)(1-p(X_i)))$, and $F(\cdot)$ is the distribution function associated with the first-step propensity score model.[16] This is a matrix weighted average of squares and cross-products of covariances between the covariates included in the propensity score model and the conditional expectation of the counterfactual outcome under control, or $\mu_0(X_i)$. This term is related to the famous result of Hirano et al. (2003), that a nonparametric first-step estimate of the propensity score can lead to semiparametric efficiency asymptotically. Intuitively, including variables in the propensity score model that are related to $\mu_0(X_i)$ apparently can, under suitable conditions, help the G_1 term to "knock out" the G term that stands in the way of efficiency.[17]

These results can be intuitively summarized as follows:

Result 1. *Pair matching is not efficient, in the sense that its asymptotic variance exceeds the SEB.*

Result 2. *A sufficient condition for the efficiency of correct reweighting is that $\mu_0(X_i)$ does not depend on X_i for units in the treatment group, that is, there is no selection problem. Under homogenous treatment effects, this condition is also necessary.*

Result 3. *A sufficient condition for the efficiency of Frölich reweighting is $\mu_0(X_i) = 0$ for every unit in the treatment group. Under homogenous treatment effects, this condition is also necessary.*

Result 4. *The asymptotic distribution of correct reweighting is invariant to additive shifts of the outcome, while that of Frölich reweighting is not.*

The first result follows from the machinery developed in Abadie and Imbens (2006) and is analogous to their result for the population average treatment effect. The second and third results follow from algebra, and the fourth result is implied by the second and third results.

[14] In the main text, Abadie and Imbens (2006) provide explicit large sample characterizations for the case of the population average treatment effect. To derive results for TOT, see their equation (13) in the main text and equation (A.34) in the unpublished proofs. Note that while their results pertain to matching on covariates, they can be applied to pair matching with an estimated propensity score in the context of Frölich (2004)'s study, because X_i is scalar and hence can be derived from knowledge of $\widehat{p}(X_i)$ alone.

[15] The terms H_1 and H_2 are given by $H_1 = 2\frac{1}{p}\mathbb{C}[\tau(X_i), \mu_0(X_i)|T_i = 1]$ and

$$H_2 = \mathbb{C}[\tau(X_i), (1-p(X_i))\nu_i Z_i | T_i = 1]\mathbb{E}[\nu_i^2 p(X_i)(1-p(X_i))Z_i Z_i']^{-1}\mathbb{C}[\nu_i Z_i, \mu_0(X_i)|T_i = 1]$$

Under homogenous treatment effects, $\tau(X_i)$ is constant and both of these terms are zero. Generally, however, H_1 and H_2 are nonzero and can be either positive or negative.

[16] Here, β is the probability limit of the first-step coefficients, i.e., $p(X_i) = F(Z_i'\beta)$. Standard practice in empirical work is to use a logit model, in which case $\nu_i = 1$.

[17] For further discussion of the intuition behind the Hirano et al. (2003) result, see Graham (2008). We pause to note that in the hybrid case of reweighting with normalized weights that sum to one, but using a known propensity score, the asymptotic variance is simply SEB $+ G + H_1$ (see Appendix I), and hence inefficient unless $\mu_0(X_i)$ and $\mu_1(X_i)$ covary in particular ways.

To get a quick sense of the magnitude of the differences between the variances of the different varieties of reweighting estimators, Table 2 presents asymptotic variances for six varieties of reweighting, based on (i) whether the weights are normalized to sum to one (columns (4) through (6)) or are left unnormalized (columns (1) through (3)) and (ii) whether the first-step propensity score is the true propensity score ("Known"), estimated parametrically by a correctly specified maximum likelihood routine ("Estd."), or estimated nonparametrically using a series logit ("Overfit"). As noted, $\widehat{\theta}_F$ leaves the weights unnormalized and uses the known propensity score. The Hirano et al. (2003) estimator leaves the weights unnormalized and uses a nonparametric estimate of the propensity score. Our preferred version of reweighting, $\widehat{\theta}$, normalizes the weights to sum to one and utilizes a parsimonious logit model.

Table 2. Illustrative Variances, Different Varieties of Reweighting

Outcome Equation Parameters				Weights Left Unnormalized			Weights Normalized		
				Known, $\widehat{\theta}_F$	Estd.	Overfit, $\widehat{\theta}_{HIR}$	Known	Estd., $\widehat{\theta}$	Overfit
Intercept	Slope	σ^2	SEB	(1)	(2)	(3)	(4)	(5)	(6)
0	0	0.1	1.1	1.1	1.1	1.0	1.1	1.1	1.0
0	1	0.1	1.1	8.4	3.8	1.6	1.7	1.4	1.1
0	2	0.1	1.1	31.6	12.7	3.1	3.7	2.4	1.3
10	0	0.1	1.1	1,085	286.3	68.5	1.0	1.0	1.0
10	1	0.1	1.1	1,287	322.0	67.1	1.7	1.4	1.1
10	2	0.1	1.1	1,466	479.2	87.6	3.6	2.4	1.3

The asymptotic variances displayed were obtained by simulation using 5,000 estimator replications, with each estimate based on 1,000 observations. The DGP is based on equations (1) through (3), with $\mu_0(X_i)$ an affine function of X_i ("Intercept", "Slope"), $\mu_0(X_i) = \mu_1(X_i)$, $\mu_T(X_i) = \sqrt{2}X_i$, and u_i distributed standard logistic. In light of the sample size, the overfit propensity score model was taken to be a fifth order polynomial in X_i.

The results in Table 2 show plainly that leaving the weights unnormalized performs terribly, even with the series logit model suggested by Hirano et al. (2003). This variety of reweighting is particularly susceptible to the nuisance parameter of the location of the outcome. Normalizing the weights so that they sum to one eliminates this deficiency. However, both varieties of reweighting suffer from increased variance when there is a selection problem, i.e., when the slope parameter exceeds zero in this DGP. Using an overfit logit model reduces the variance in such a situation. However, overfitting also worsens the bias of the estimator (results not shown).

III. Large Sample Intuition

A. Data Generating Process

Frölich (2004) considers thirty DGPs in his study. To simplify the discussion we focus on one of the DGPs ("Frölich's baseline DGP") which is a specialized version of equations (1) through (3), with

$$Y_i(1) = \theta + 0.15 + 0.7p(X_i) + \varepsilon_i \tag{12}$$

$$Y_i(0) = 0.15 + 0.7p(X_i) + \varepsilon_i \tag{13}$$

$$T_i^* = \sqrt{2}X_i - u_i \tag{14}$$

where X_i is distributed standard normal, ε_i is distributed uniform with mean zero and variance $\sigma^2 = 0.01$, and u_i is distributed standard logistic, implying $F(u) = 1/(1+\exp(-u))$ and $p(X_i) = F(\sqrt{2}X_i)$. The treatment effect, θ, is taken to be constant in the population and equal to zero.[18] Qualitatively, our conclusions do not change when we consider other DGPs, as will become clear in Section IV.

This DGP has a homoskedastic outcome equation error and homogenous treatment effects. Thus, the efficiency bound in equation (4) simplifies to $\text{SEB} = \frac{\sigma^2}{p^2}\mathbb{E}\left[\frac{p(X_i)}{1-p(X_i)}\right]$ and by standard integration we have $\text{SEB} = \sigma^2 e/p^2$, where $\ln(e) = 1$. This proves a useful benchmark, both conceptually and numerically.

B. Variance Decompositions

To get a sense of the magnitudes of the variances associated with pair matching and reweighting estimators, Table 3 presents a decomposition of the variance expressions in the context of Frölich's baseline DGP. Like all of the DGPs studied in Frölich (2004), the baseline DGP is homoskedastic and sets $\sigma^2 = 0.01$. In our view, such a choice for the variance of the outcome equation error limits the relevance of the simulation results to the microeconomic applications that have motivated the econometric program evaluation literature. In the context of the DGP in equations (12) through (14), choosing an error variance of $\sigma^2 = 0.01$ would be equivalent to a situation where R^2 from a regression of Y_i on T_i and $p(X_i)$ would be approximately 0.77 when the treatment is ineffective ($\theta = 0$). If the treatment is effective (say, $\theta = 0.15$), then the R^2 from this regression would be 0.85. In our experience, outcome variables in microeconomic applications—e.g., labor earnings—are dominated by factors unavailable to the researcher and difficult to predict. We are unaware of situations in empirical practice where the outcome is so predictable that a

[18] Strictly speaking, Frölich (2004) does not specify the DGP for equation (1). This is due to his focus on the success of various estimators in estimating the counterfactual mean under treatment, or $\mathbb{E}[Y_i(0)|T_i = 1]$. We prefer to specify the entire DGP. This amounts to changing the units in which variance is measured. See Appendix II for details and discussion.

researcher running such a simple regression would achieve an R^2 of such a high magnitude. The R^2 values for similar regressions reported in Dehejia (2005), for example, range from approximately 0.1 to 0.3. Taking the larger of these R^2 values as a reference point corresponds to a value of roughly $\sigma^2 = 0.1$ in Frölich's baseline DGP, when the treatment is effective.

TABLE 3. DECOMPOSING VARIANCE OF ESTIMATORS: FRÖLICH'S BASELINE DGP

Estimator	σ^2	SEB	G_{PM}	G_F	G	G_1	H_1, H_2	n Variance
Pair Matching	0.01	0.11	0.06	-	-	-	-	0.17
Frölich Reweighting	0.01	0.11	-	5.83	-	-	0	5.94
Correct Reweighting	0.01	0.11	-	-	0.35	0.17	0	0.28
Pair Matching	0.10	1.09	0.64	-	-	-	-	1.73
Frölich Reweighting	0.10	1.09	-	5.83	-	-	0	6.92
Correct Reweighting	0.10	1.09	-	-	0.35	0.17	0	1.26

Table 3 shows that for the small error variance of $\sigma^2 = 0.01$, pair matching has a much smaller asymptotic variance (0.17) than Frölich reweighting (5.94). This result provides a large sample interpretation for the simulation evidence presented in Frölich (2004). The reason for the enormous difference in variances is that G_F is much larger than G_{PM}. In particular, returning to the characterization of these terms in equations (8) through (10), we see that G_{PM} is proportional to σ^2, whereas G_F is not. Thus, when σ^2 is small enough, pair matching performs best, but when σ^2 is large enough, reweighting performs best.

Table 3 also presents a decomposition for the empirically more relevant case of $\sigma^2 = 0.1$. In that case, Frölich reweighting has larger asymptotic variance than pair matching, which in turn has larger asymptotic variance than correct reweighting. Pair matching has larger asymptotic variance than correct reweighting as long as $\sigma^2 > 0.028$. Table 3 also clarifies the extent to which correct reweighting is preferred to Frölich reweighting. In Frölich's baseline DGP, regardless of the value of σ^2, the discrepancy between Frölich reweighting and correct reweighting is a large 5.65.

C. A Graphical View of Efficiency

This background clarifies some conceptual distinctions between matching and reweighting approaches to estimating average treatment effects. We are now in a position to graphically illustrate how Frölich's conclusions about the superiority of matching *not* at odds with the asymptotic results, but are highly context-specific.

We begin this discussion by noting that the search for efficient estimators of average treatment effects can be understood as the search for an appropriate intercept and slope in a figure such as Figure 1. Figure

1 presents the asymptotic variance of average treatment effect estimators in a homoskedastic DGP, as a function of σ^2, the homogenous variance of the outcome equation error. An efficient estimator is one which has a variance curve on top of the SEB, which here is a straight line going through the origin. In a case with homogenous treatment effects and homoskedasticity of the outcome equation error, matching estimators tend to have variances that are zero at the origin, but have a steeper slope than that of the SEB. In such settings, reweighting estimators tend to have variances that are *positive* at the origin, but have a slope equal to that of the SEB.

Figure 1 makes this point for the special case of pair matching and Frölich reweighting, in the context of Frölich's baseline DGP. As we saw in equations (8) through (10) and then concretely in Table 3, the intercept for Frölich reweighting is positive and large (5.83), whereas the intercept for pair matching is zero. In contrast, the slope for Frölich reweighting is that of the SEB, whereas the slope for pair matching is strictly above that of the SEB. This figure makes it plain that reweighting has the wrong intercept and that pair matching has the wrong slope.

Figure 2 revisits this picture, but replacing Frölich reweighting with correct reweighting. The intercept for reweighting is now much smaller (0.18 rather than 5.83). It is tempting to conclude that correct reweighting is efficient for all practical purposes. However, this conclusion must be tempered by the recognition that for very small values of σ^2, correct reweighting will have larger variance than pair matching.[19]

IV. Finite Sample Results

A. Finite Sample Performance

As noted, Frölich (2004) considers thirty DGPs, corresponding to all possible combinations of five density "designs" and six outcome "curves". The five designs pertain to the distribution of propensity scores among treatment and control units, and the six curves pertain to the nonlinearity of the relationship between the covariates and the outcome.

We turn now to a replication of the main results in Frölich (2004), which pertain to $n = 100$. Table 4 presents simulation estimates of the bias and variance of pair matching, correct, reweighting, and Frölich's preferred matching estimator, ridge matching, for each of the thirty DGPs using 10,000 simulation replications, as in Frölich (2004). Following our discussion in Section III, we set the variance of the outcome equation error term to be 0.1.[20] For reference, we present the SEB for each DGP, as well as the asymptotic

[19] Empirical researchers may find it worthwhile to engage in simulation studies tailored to the properties of the data they study. One could imagine an applied paper where the data were characterized by very strong selection and very high predictability of the outcome. In such a setting, matching might be expected to outperform correct reweighting.

[20] See Appendix II for a detailed description of these DGPs. There, we replicate the results of Frölich (2004). We also

variance for pair matching and correct reweighting.[21]

Two broad features of these DGPs are relevant to the results in Table 4. First, designs 1 and 5 satisfy the weak overlap condition, but fail the strict overlap condition. In these two designs, the propensity score cannot be strictly bounded away from 1. In the remaining designs, the propensity score is strictly bounded away from 0 and 1. Nonetheless, in all five designs, the SEB can be shown to be finite by direct integration. Second, in design 1, but in no other design, the propensity score can be reliably estimated using a parametric maximum likelihood routine. This is because design 1 corresponds to a standard latent variable model for treatment. Designs 2 through 5 cannot be written in this way. For these designs, even taking advantage of the knowledge of the DGP, estimation of the propensity score would entail estimating a maximum likelihood model based on a uniform density whose parameters are in the boundary of the parameter space, which is unlikely to work well. Instead of estimating a uniform binary choice model, we choose to use a logit model with a second order polynomial.

Five main results arise from these Monte Carlo simulations. First, correct reweighting and ridge matching have a variance that is 60% to 80% of pair matching.[22] Second, the variance of reweighting is similar to the variance of ridge matching.[23] Third, the variances of pair matching and reweighting are close to their asymptotic variances in designs 2, 3 and 4, but are different for designs 1 and 5. This seems to be a general phenomenon, as we have discussed elsewhere (Busso, DiNardo and McCrary 2008). In DGPs violating strict overlap, finite sample performance can often be quite different from that suggested by the large sample theory. Fourth, and relatedly, the bias for all three estimators is much larger in designs 1 and 5 than in designs 2, 3, and 4.[24] Fifth, in most DGPs the bias of ridge matching is the largest of the three estimators under consideration.

In order to analyze further the performance of correct reweighting and ridge matching, we explore the robustness of the results to different values of the variance of the outcome equation error. Figure 3 displays the results of our analysis for a sample size of 100 and Frölich's baseline DGP (i.e., design 1 and curve 1). The top half of the figure graphs the variance. For comparison, we also plot the SEB and the variance of pair matching. The analysis points again to the specificity of Frölich's results and the peculiarities of

expand those results in two directions. First, we consider estimation of the TOT rather than the counterfactual mean of the outcome under treatment. Second, we consider larger values of σ^2 than are considered in Frölich (2004). See Appendix Table A.1.

[21] Large sample properties of ridge matching are not yet available in the literature.

[22] The asymptotic variances presented in this table are estimated with a great deal of precision, and have standard errors based on Wishart (Wishart 1928, Muirhead 2005) and bootstrap approximations of about 0.015 or less.

[23] The differences observed are nonetheless significant in nearly all the DGPs.

[24] To the best of our knowledge, no finite sample result regarding unbiasedness exists for these estimators. For all three estimators, the null hypothesis of zero bias is strongly rejected for nearly all DGPs.

the DGP. The variance of both reweighting and ridge matching are below the SEB, although reweighting seems to be always closer to the SEB. The unusual efficiency of both estimators in these DGPs, however, comes at price: both are biased. For ridge matching, the problem of bias is severe, particularly for DGPs with noisy outcome measures. Even for the smallest (empirically least plausible) values of the variance of the outcome equation error, the bias in all estimators exceeds what might be expected for an effective treatment (say, $\theta = 0.015$). The bias in ridge matching grows the most quickly and approximately triples in value going from a variance of the outcome equation error of near 0 to 1. Intuitively, the problems with bias for ridge matching seem likely to arise because of the cross-validation algorithm. When the outcome is difficult to predict, cross-validation may well choose the largest bandwidth considered, in which case ridge matching reduces to the raw difference in means between treatment and control units. In larger samples, or with a more predictable outcome, ridge matching could potentially perform better, as the performance of cross-validation improves.

V. Conclusion

The existing finite sample literature on semiparametric estimation of average treatment effects is generally critical of the performance of reweighting and tends to favor matching. The leading paper on this topic, Frölich (2004), finds that in small samples reweighting estimators tend to perform much worse than many of the most popular matching estimators (namely, pair, nearest-neighbor, kernel, local linear or ridge matching). This conclusion is at odds with the findings of the large sample literature. We resolve this puzzle in this paper and show that reweighting performs much better than suggested in Frölich (2004).

We derive large sample results for reweighting that complement those of Hirano et al. (2003). These results demonstrate the wisdom of normalizing the weights to sum to one and of using an estimated propensity score. Frölich (2004) leaves the weights unnormalized and uses the true propensity score. This skews his findings towards the conclusion that reweighting is not effective. The consequences of these two choices for the relative MSE of reweighting and pair matching are magnified by the small variance of the outcome equation error used in his simulations. We argue that this error variance is sufficiently small that the DGPs studied in Frölich (2004) are of limited relevance to empirical microeconomic practice.

We show that in DGPs with only slightly larger values of the variance of the outcome equation error than are considered in Frölich (2004), an appropriate implementation of reweighting has much lower variance than pair matching. This variance reduction appears to come without a cost: the bias of reweighting and pair matching is near zero in the DGPs we have analyzed here and elsewhere (Busso et al. 2008).

We also contrast the performance of reweighting and ridge matching—the preferred matching estimator in Frölich (2004). We find that in small samples the variance of ridge matching and reweighting is usually comparable. However, the bias of ridge matching is small for some DGPs, but quite large for other DGPs. The bias of reweighting, in contrast, is small uniformly across DGPs, especially when strict overlap is satisfied. Generally, the bias of reweighting seems equivalent to that of pair matching, the matching estimator with the best performance in terms of bias.

If preferences over bias and variance are not lexicographic, then some of the biased matching estimators may be preferred to reweighting. We caution, however, that the DGPs considered in this paper may not adequately span those likely to confront empirical researchers. In general, the bias of these estimators in any given DGP could be of lesser or greater magnitude than documented here. In such a case, the researcher's preference ranking over estimators could be different than that suggested by a literal interpretation of the simulation evidence.[25] Our own preference is for estimators that minimize the maximum bias over possible DGPs (e.g., unbiased estimators), and among those we prefer low variance. The small sample evidence presented in this paper suggests that reweighting is better than both pair and ridge matching in that sense.

Finally, reweighting has two practical advantages over ridge matching. First, reweighting is easy to compute, because it is a difference in weighted means by treatment status. Ridge matching is hard to compute. It requires looping over observations and estimating many different local linear ridge regressions, even when the bandwidth is known. Since the bandwidth is not known, ridge matching further entails selection of a bandwidth using cross-validation, which can be extremely time-intensive. Second, accurate standard errors for reweighting are readily obtained.[26] To date, no valid inference procedure for ridge matching has been proposed. In light of the smoothness of ridge matching, an appropriate bootstrap algorithm is likely to work, but bootstrapping an estimator that uses cross-validation is unlikely to be practical in empirical work, particularly in applications with more than 1,000 observations.

[25] For example, a researcher seeking to minimize the maximum mean squared error across *all* DGPs would not be comforted with the knowledge that the bias was small relative to the variance in the DGPs studied here.

[26] Busso (2008) notes that a sequential GMM approach is highly effective and that, if the sample size is sufficiently large ($n > 500$), robust regression standard errors that ignore the estimation error in the weights work well.

Appendix I

A. Propensity Score Matching is a Weighted Least Squares Estimator

Result. *If the weighting function $W(i,j)$ satisfies the property $\sum_{j=1}^{n}(1-T_j)W(i,j)=1$, we have*

$$\widetilde{\theta}=\frac{\sum_{i=1}^{n}T_iY_i}{\sum_{i=1}^{n}T_i}-\frac{\sum_{j=1}^{n}(1-T_j)\widehat{V}_jY_j}{\sum_{j=1}^{n}(1-T_j)\widehat{V}_j}$$

where $\widetilde{\theta}$ is defined as in equation (5) and $\widehat{V}_j=\sum_{i=1}^{n}T_iW(i,j)/\sum_{i=1}^{n}T_i$ is the average weight received by control observation j, on average across all treatment observations i.

Proof. *Define $\sum_{i=1}^{n}T_i=n_1$ and write*

$$\begin{aligned}\widetilde{\theta} &= \frac{1}{n_1}\sum_{i=1}^{n}T_iY_i-\frac{1}{n_1}\sum_{i=1}^{n}T_i\widehat{Y}_i(0) &(15)\\ &= \frac{1}{n_1}\sum_{i=1}^{n}T_iY_i-\frac{1}{n_1}\sum_{i=1}^{n}T_i\sum_{j=1}^{n}(1-T_j)W(i,j)Y_j &(16)\\ &= \frac{1}{n_1}\sum_{i=1}^{n}T_iY_i-\sum_{j=1}^{n}(1-T_j)Y_j\frac{1}{n_1}\sum_{i=1}^{n}T_iW(i,j) &(17)\\ &\equiv \frac{1}{n_1}\sum_{i=1}^{n}T_iY_i-\sum_{j=1}^{n}(1-T_j)Y_j\widehat{V}_j &(18)\end{aligned}$$

It remains to show that $\sum_{j=1}^{n}(1-T_j)W(i,j)=1$ implies $\sum_{j=1}^{n}(1-T_j)\widehat{V}_j=1$. Write

$$\sum_{j=1}^{n}(1-T_j)\widehat{V}_j=\sum_{j=1}^{n}(1-T_j)\frac{1}{n_1}\sum_{i=1}^{n}T_iW(i,j)=\frac{1}{n_1}\sum_{i=1}^{n}T_i\sum_{j=1}^{n}(1-T_j)W(i,j) \quad (19)$$

and the result follows. □

The key adding-up property $\sum_{j=1}^{n}(1-T_j)W(i,j)=1$ is satisfied by all of the matching estimators studied in Frölich (2004). We note that $\sum_{i=1}^{n}T_iW(i,j)$ is the $K_M(j)$ function studied by Abadie and Imbens (2006). Their results can be used to show that nearest neighbor matching entails a $\widehat{V}_j$ function that approximates $\widehat{p}(X_j)/(1-\widehat{p}(X_j))$. Finally, the $W(i,j)$ function used by kernel matching implies, for a symmetric kernel, that $\widehat{V}_j$ is a ratio of kernel regression estimators and also approximates $\widehat{p}(X_j)/(1-\widehat{p}(X_j))$. There is thus a sense in which some matching estimators approximate reweighting in specific ways.

B. Derivation of Large Sample Results: Unnormalized True Weights

To economize on space, we adopt the following notations. First, we drop all i subscripts, trusting the reader to remain aware of which objects are stochastic and which are not. Second, for treatment assignments $t=0,1$, we write $\mu_t=\mu_t(X_i)$ and $\sigma_t^2=\sigma_t^2(X_i)$, and we let $\tau=\mu_1-\mu_0$ denote the covariate-specific treatment effect and $e=p(X_i)$ the propensity score.

Using this notation, we can define the non-stochastic terms $\theta=\mathbb{E}[\tau|T=1]$, $\alpha=\mathbb{E}[\mu_0|T=1]$, and $p=\mathbb{E}[e]$, and we can write Frölich (2004)'s version of reweighting as

$$\widehat{\theta}_F=\overline{h}/\overline{T} \quad (20)$$

where $\overline{h}$ is the sample mean of $YT-Y(1-T)e/(1-e)$ and $\overline{T}$ is the sample mean of T. We will show that $\sqrt{n}\left(\overline{h}-p\theta,\overline{T}-p\right)$ converges in distribution to a bivariate normal distribution with a particular variance matrix Ω.

We then use Slutsky's theorem to compute the first order asymptotic distribution of $\widehat{\theta}_F$.

We will make repeated use of iterated expectations over X and in particular the following facts, valid for treatment assignments $t = 0, 1$:

$$\begin{aligned} p\mathbb{E}[Y(t)|T=1] &= \mathbb{E}[\mu_t e] \quad \text{and} \quad (1-p)\mathbb{E}[Y(t)|T=0] = \mathbb{E}[\mu_t(1-e)] \\ p\mathbb{E}[Y(t)^2|T=1] &= \mathbb{E}[(\mu_t^2+\sigma_t^2)e] \quad \text{and} \quad (1-p)\mathbb{E}[Y(t)^2|T=0] = \mathbb{E}[(\mu_t^2+\sigma_t^2)(1-e)] \end{aligned}$$

This type of reasoning shows that the probability limit of $\overline{h}$ is

$$\mathbb{E}[YT - Y(1-T)\frac{e}{1-e}] = \mathbb{E}[\mu_1 e] - \mathbb{E}[\mu_0 e] = p\theta \tag{21}$$

The probability limit of $\overline{T}$ is $\mathbb{E}[T] = p$, so by continuity of probability limits, $\widehat{\theta}_F$ is consistent for θ.

Turning to the asymptotic variance, we first note that by the Lindeberg-Levy central limit theorem and the Crámer-Wold device, $\sqrt{n}\left(\overline{h} - p\theta, \overline{T} - p\right)$ converges in distribution to a normal distribution with variance matrix

$$\Omega = \begin{bmatrix} \mathbb{V}[YT - Y(1-T)e/(1-e)] & \mathbb{C}[YT - Y(1-T)e/(1-e), T] \\ \mathbb{C}[YT - Y(1-T)e/(1-e), T] & \mathbb{V}[T] \end{bmatrix} \tag{22}$$

where we have $\mathbb{V}[T] = p(1-p)$, $\mathbb{C}[YT - Y(1-T)e/(1-e), T] = \mathbb{E}[\mu_1 e] - p^2\theta$, and

$$\mathbb{V}[YT - Y(1-T)e/(1-e)] = \mathbb{E}\left[\sigma_1^2 e + \sigma_0^2 \frac{e^2}{1-e}\right] + \mathbb{E}\left[\mu_1^2 e + \mu_0^2 \frac{e^2}{1-e}\right] - (p\theta)^2 \tag{23}$$

Then define $\widehat{\theta}_F = r(\overline{h}, \overline{T})$ where $r(h,p) = h/p$ has gradient evaluated at $h \equiv p\theta$ and p of $R \equiv p^{-2}(p, -h)$. Then by Slutsky's theorem, to first order we have

$$\begin{aligned} p^2 n\mathbb{V}[\widehat{\theta}_F] &= p^2 R'\Omega R = \mathbb{V}[YT - Y(1-T)e/(1-e)] - 2\theta\mathbb{C}[YT - Y(1-T)e/(1-e), T] + \theta^2\mathbb{V}[T] && (24) \\ &= \mathbb{E}\left[\sigma_1^2 e + \sigma_0^2 \frac{e^2}{1-e}\right] + \mathbb{E}\left[\mu_1^2 e + \mu_0^2 \frac{e^2}{1-e}\right] - 2\theta\mathbb{E}\left[\mu_1 e\right] + p\theta^2 && (25) \\ &= \mathbb{E}\left[\sigma_1^2 e + \sigma_0^2 \frac{e^2}{1-e}\right] + \mathbb{E}\left[\mu_1^2 e + \mu_0^2 \frac{e^2}{1-e}\right] - \frac{1}{p}\mathbb{E}\left[\mu_1 e\right]^2 + \frac{1}{p}\mathbb{E}\left[\mu_0 e\right]^2 && (26) \\ &= \mathbb{E}\left[\sigma_1^2 e + \sigma_0^2 \frac{e^2}{1-e}\right] + \mathbb{E}\left[\mu_0^2 \frac{e^2}{1-e}\right] + \frac{1}{p}\mathbb{E}\left[\mu_0 e\right]^2 + p\mathbb{V}\left[\mu_1|T=1\right] && (27) \\ &= \mathbb{E}\left[\sigma_1^2 e + \sigma_0^2 \frac{e^2}{1-e}\right] + \mathbb{E}\left[\mu_0^2 \frac{e}{1-e}\right] + p\mathbb{V}\left[\mu_1|T=1\right] - p\mathbb{V}\left[\mu_0|T=1\right] && (28) \end{aligned}$$

where we use the fact that $\mathbb{E}[(\tau-\theta)^2 e] = p\mathbb{V}[\tau|T=1] = p\mathbb{V}[\mu_1|T=1] + p\mathbb{V}[\mu_0|T=1] - 2p\mathbb{C}[\mu_0, \mu_1|T=1]$ and $e^2/(1-e) = e/(1-e) - e$. Recall that $\text{SEB} = \mathbb{E}\left[\frac{\sigma_1^2 e}{p^2}\right] + \mathbb{E}\left[\frac{\sigma_0^2(1-e)^2}{p^2(1-e)}\right] + \frac{1}{p}\mathbb{V}[\tau|T=1]$. Thus, to first order we have

$$n\mathbb{V}[\widehat{\theta}_F] = \text{SEB} - 2\frac{1}{p}\left\{\mathbb{V}[\mu_0|T=1] - \mathbb{C}[\mu_0, \mu_1|T=1]\right\} + \mathbb{E}\left[\frac{\mu_0^2 e}{p^2(1-e)}\right] \tag{29}$$

C. Derivation of Large Sample Results: Normalized True Weights

A reweighting estimator using true weights can be rewritten as

$$\ddot{\theta} = \overline{g}/\overline{T} - \overline{f}/\overline{S} \tag{30}$$

where $\overline{g}$ is the sample mean of YT, $\overline{f}$ is the sample mean of $Y(1-T)e/(1-e)$, and $\overline{S}$ is the sample mean of $(1-T)e/(1-e)$. The probability limit of $\overline{g}$ is $g \equiv \mathbb{E}[\mu_1 e]$, the probability limit of $\overline{T}$ is p, the probability limit of $\overline{f}$ is

$f \equiv E[\mu_0 e]$, the probability limit of $\overline{S}$ is p, and continuity of probability limits then implies that $\ddot{\theta}$ is consistent for θ.

For the asymptotic variance, note that $\mathbb{V}[\ddot{\theta}] = \mathbb{V}[\overline{g}/\overline{T}] + \mathbb{V}[\overline{f}/\overline{S}]$. Consider first $\mathbb{V}[\overline{g}/\overline{T}]$. By the Lindeberg-Levy central limit theorem and the Crámer-Wold device, $\sqrt{n}\left(\overline{g} - g, \overline{T} - p\right)$ converges in distribution to a normal distribution with variance matrix

$$\Omega = \left[\begin{array}{cc} \mathbb{V}[YT] & \mathbb{C}[YT, T] \\ \mathbb{C}[YT, T] & \mathbb{V}[T] \end{array} \right] \tag{31}$$

where $\mathbb{V}[T] = p(1-p)$ as before and $\mathbb{C}[YT, T] = \mathbb{E}[\mu_1 e](1-p) = g(1-p)$ and

$$\mathbb{V}[YT] = \mathbb{E}\left[\left(\sigma_1^2 + \mu_1^2\right) e\right] - \mathbb{E}\left[\mu_1 e\right]^2 = \mathbb{E}\left[\left(\sigma_1^2 + \mu_1^2\right) e\right] - g^2 \tag{32}$$

Define $\overline{g}/\overline{T} = r(\overline{g}, \overline{T})$ where $r(g,p) = g/p$ has gradient evaluated at g and p of $R \equiv p^{-2}(p, -g)$. Then by Slutsky's theorem, to first order we have

$$\begin{aligned} p^2 n \mathbb{V}\left[\overline{g}/\overline{T}\right] &= \mathbb{V}[YT] - 2\frac{g}{p}\mathbb{C}[YT, T] + \frac{g^2}{p^2}\mathbb{V}[T] && (33) \\ &= \mathbb{E}\left[\sigma_1^2 e\right] + \mathbb{E}\left[\mu_1^2 e\right] - g^2/p = \mathbb{E}\left[\sigma_1^2 e\right] + \mathbb{E}\left[\mu_1^2 e\right] - \frac{1}{p}\mathbb{E}\left[\mu_1 e\right]^2 && (34) \\ &= \mathbb{E}\left[\sigma_1^2 e\right] + p\mathbb{V}[\mu_1 | T = 1] && (35) \end{aligned}$$

Consider next $\overline{f}/\overline{S}$. By the Lindeberg-Levy central limit theorem and the Crámer-Wold device, $\sqrt{n}\left(\overline{f} - f, \overline{S} - p\right)$ converges in distribution to a normal distribution with variance matrix

$$\Omega = \left[\begin{array}{cc} \mathbb{V}[Y(1-T)e/(1-e)] & \mathbb{C}[Y(1-T)e/(1-e), (1-T)e/(1-e)] \\ \mathbb{C}[Y(1-T)e/(1-e), (1-T)e/(1-e)] & \mathbb{V}[(1-T)e/(1-e)] \end{array} \right] \tag{36}$$

where $\mathbb{V}[(1-T)e/(1-e)] = \mathbb{E}\left[e^2/(1-e)\right] - p^2$, and

$$\begin{aligned} \mathbb{V}[Y(1-T)e/(1-e)] &= \mathbb{E}\left[\left(\sigma_0^2 + \mu_0^2\right)\frac{e^2}{1-e}\right] - \mathbb{E}\left[\mu_0 e\right]^2 = \mathbb{E}\left[\left(\sigma_0^2 + \mu_0^2\right)\frac{e^2}{1-e}\right] - f^2 && (37) \\ \mathbb{C}[Y(1-T)e/(1-e), (1-T)e/(1-e)] &= \mathbb{E}\left[\mu_0 \frac{e^2}{1-e}\right] - fp && (38) \end{aligned}$$

Then redefine $\overline{f}/\overline{S} = r(\overline{f}, \overline{S})$ where $r(f,p) = f/p = \alpha$ has gradient evaluated at f and p of $R \equiv p^{-2}(p, -f)$. Then by Slutsky's theorem, to first order we have

$$\begin{aligned} p^2 n \mathbb{V}\left[\overline{f}/\overline{S}\right] &= \mathbb{V}[Y(1-T)e/(1-e)] - 2\alpha\mathbb{C}[Y(1-T)e/(1-e), (1-T)e/(1-e)] + \alpha\mathbb{V}[(1-T)e/(1-e)] && (39) \\ &= \mathbb{E}\left[\sigma_0^2 \frac{e^2}{1-e}\right] + \mathbb{E}\left[\mu_0^2 \frac{e^2}{1-e}\right] - 2\alpha\mathbb{E}\left[\mu_0 \frac{e^2}{1-e}\right] + \alpha^2\mathbb{E}\left[\frac{e^2}{1-e}\right] && (40) \\ &= \mathbb{E}\left[\sigma_0^2 \frac{e^2}{1-e}\right] + \mathbb{E}\left[\left(\mu_0 - \alpha\right)^2 \frac{e}{1-e}\right] - p\mathbb{V}[\mu_0 | T = 1] && (41) \end{aligned}$$

Putting these results together, we have

$$p^2 n \mathbb{V}[\ddot{\theta}] = \mathbb{E}\left[\sigma_1^2 e + \sigma_0^2 \frac{e^2}{1-e}\right] + p\mathbb{V}[\mu_1 | T = 1] - p\mathbb{V}[\mu_0 | T = 1] + \mathbb{E}\left[\left(\mu_0 - \alpha\right)^2 \frac{e}{1-e}\right] \tag{42}$$

which implies that to first order

$$n\mathbb{V}[\ddot{\theta}] = \text{SEB} - 2\frac{1}{p}\left\{\mathbb{V}[\mu_0 | T = 1] - \mathbb{C}[\mu_0, \mu_1 | T = 1]\right\} + \mathbb{E}\left[\frac{\left(\mu_0 - \alpha\right)^2 e}{p^2(1-e)}\right] \tag{43}$$

D. Derivation of Large Sample Results: Normalized Estimated Weights

For the case of the reweighting estimator with a parametric estimate of the propensity score, we use a method of moments framework to derive large sample properties. Define $h = (a, q, b')'$, $Z = (1, X')'$, and $F(\cdot)$ a parametric distribution function such as the logistic, and define the moment functions

$$r(h) = \left[\begin{array}{c} (Y - a - qT)\, W(b) \\ (Y - a - qT)\, W(b) T \\ (T - F(Z'b))\, \nu(b) Z \end{array} \right] \tag{44}$$

where $W \equiv W(b) = T + (1 - T)F(Z'b)/(1 - F(Z'b))$ and where we assume that $\mathbb{E}[r(\ddot{h})] = \mathbb{E}[r(\dot{h})]$ if and only if $\ddot{h} = \dot{h}$ and $0 = \mathbb{E}[r(\eta)]$, where $\eta = (\alpha, \theta, \beta')'$. This condition ensures that $\alpha = \mathbb{E}[\mu_0 | T = 1] = \mathbb{E}[\mu_0 e]/p$, $\theta = \mathbb{E}[\mu_1 - \mu_0 | T = 1] = \mathbb{E}[(\mu_1 - \mu_0)e]/p$, and $e = F(Z'\beta)$ are defined as before and guarantees unique identification. The first two moments are scalar and are implied by a weighted regression of Y on T using weights W; the third moment is actually a $K + 1$ vector of moments and incorporates the estimation of e using a binary choice model. Generally, $\nu \equiv \nu(b) = F'(Z'b)/(F(Z'b)(1 - F(Z'b)))$. For the logit, $\nu = 1$, and among distributions with $F(0) = 1/2$, the logit distribution is the only distribution for which $\nu = 1$. This can be shown by solving the differential equation implied by $\nu = 1$.

Next define $\widehat{\eta} = (\widehat{\alpha}, \widehat{\theta}, \widehat{\beta}')'$ by $\overline{r}(\widehat{\eta}) = 0$, where $\overline{r}(h)$ is the sample mean of $r(h)$. By the Lindeberg-Levy central limit theorem and the Crámer-Wold device, $\sqrt{n}\overline{r}(\eta) \xrightarrow{d} N(0, \Omega)$, where $\Omega = \mathbb{V}[r(\eta)]$. A Taylor approximation to $\overline{r}(\widehat{\eta})$ centered about η, together with continuity of probability limits and Slutsky's theorem, then shows that

$$\sqrt{n}(\widehat{\eta} - \eta) \xrightarrow{d} N\left(0, R^{-1}\Omega R^{-1'}\right) \tag{45}$$

where R is the expectation of the derivative matrix of $r(h)$ with respect to h, evaluated at $h = \eta$. One can show that

$$R = -p \left[\begin{array}{ccc} 2 & 1 & -c' \\ 1 & 1 & 0_K' \\ 0_K & 0_K & D \end{array} \right] \quad \text{and} \quad R^{-1} = -\frac{1}{p} \left[\begin{array}{ccc} 1 & -1 & c'D^{-1} \\ -1 & 2 & -c'D^{-1} \\ 0_K & 0_K & D^{-1} \end{array} \right] \tag{46}$$

where $c' \equiv \mathbb{C}[\mu_0, \nu Z | T = 1]$ is a $K + 1$ row vector, $pD \equiv \mathbb{E}\left[\nu^2 e(1 - e) Z Z'\right]$ is Fisher's information matrix for the binary choice model, and 0_K is a column vector of K zeroes. Next, note that

$$p^2 R^{-1}\Omega R^{-1'} = \left[\begin{array}{ccc} \Omega_{11} - \Omega_{21} + c'D^{-1}\Omega_{31} & \Omega_{21} - \Omega_{22} + c'D^{-1}\Omega_{32} & \Omega_{31}' - \Omega_{32}' + c'D^{-1}\Omega_{33} \\ -\Omega_{11} + 2\Omega_{21} - c'D^{-1}\Omega_{31} & -\Omega_{21} + 2\Omega_{22} - c'D^{-1}\Omega_{32} & -\Omega_{31}' + 2\Omega_{32}' - c'D^{-1}\Omega_{33} \\ D^{-1}\Omega_{31} & D^{-1}\Omega_{32} & D^{-1}\Omega_{33} \end{array} \right] \left[\begin{array}{ccc} 1 & -1 & 0_K' \\ -1 & 2 & 0_K' \\ D^{-1}c & -D^{-1}c & D^{-1} \end{array} \right]$$

where the $(2, 2)$ element is proportional to the first order approximation to the variance of $\widehat{\theta}$. To first order, we have

$$\begin{aligned} p^2 n \mathbb{V}[\widehat{\theta}] &= \Omega_{11} + 4\Omega_{22} - 4\Omega_{21} + c'D^{-1}\Omega_{33}D^{-1}c + c'D^{-1}\Omega_{31} + \Omega_{31}'D^{-1}c - 2c'D^{-1}\Omega_{32} - 2\Omega_{32}'D^{-1}c & (47) \\ &= \Omega_{11} + 4\Omega_{22} - 4\Omega_{21} + (2\Omega_{31}' - 4\Omega_{32}' + c'D^{-1}\Omega_{33})D^{-1}c & (48) \end{aligned}$$

Iterated expectations shows that

$$\begin{aligned}
\Omega_{11} &= \mathbb{E}[Y^2W^2] + \alpha^2\mathbb{E}[W^2] + \theta^2\mathbb{E}[TW^2] - 2\alpha\mathbb{E}[YW^2] - 2\theta\mathbb{E}[YTW^2] + 2\alpha\theta\mathbb{E}[TW^2] && (49)\\
&= \mathbb{E}\left[(\sigma_1^2+\mu_1^2)e + (\sigma_0^2+\mu_0^2)\frac{e^2}{1-e}\right] + \frac{1}{p^2}\mathbb{E}[\mu_0 e]^2\mathbb{E}\left[\frac{e}{1-e}\right] + \frac{1}{p^2}\mathbb{E}[(\mu_1-\mu_0)e]^2 p && (50)\\
&- 2\frac{1}{p}\mathbb{E}[\mu_0 e]\mathbb{E}\left[\mu_1 e + \mu_0\frac{e^2}{1-e}\right] - 2\frac{1}{p}\mathbb{E}[(\mu_1-\mu_0)e]\mathbb{E}[\mu_1 e] + 2\frac{1}{p^2}\mathbb{E}[\mu_0 e]\mathbb{E}[(\mu_1-\mu_0)e]p && (51)\\
&= \mathbb{E}\left[\sigma_1^2 e + \sigma_0^2\frac{e^2}{1-e}\right] + p\mathbb{V}[\mu_1|T=1] + \mathbb{E}\left[(\mu_0-\alpha)^2\frac{e^2}{1-e}\right] && (52)\\
&= \mathbb{E}\left[\sigma_1^2 e + \sigma_0^2\frac{e^2}{1-e}\right] + p\mathbb{V}[\mu_1|T=1] + \mathbb{E}\left[(\mu_0-\alpha)^2\frac{e}{1-e}\right] - p\mathbb{V}[\mu_0|T=1] && (53)
\end{aligned}$$

Since $TW^2 = T$, we have $\Omega_{22} = \Omega_{21}$, which means we do not need to calculate either term to approximate $\mathbb{V}[\widehat{\theta}]$. Finally,

$$\begin{aligned}
\Omega'_{31} &= \mathbb{E}[(Y-\alpha-\theta T)W(T-e)\nu Z] = p\mathbb{C}[\mu_1, (1-e)\nu Z|T=1] - p\mathbb{C}[\mu_0, e\nu Z|T=1] && (54)\\
\Omega'_{32} &= \mathbb{E}[(Y-\alpha-\theta T)TW(T-e)\nu Z] = p\mathbb{C}[\mu_1, (1-e)\nu Z|T=1] && (55)\\
\Omega_{33} &= \mathbb{E}[(T-e)^2\nu^2 ZZ'] = \mathbb{E}\left[e(1-e)\nu^2 ZZ'\right] \equiv pD && (56)
\end{aligned}$$

Putting these results together, we have

$$\begin{aligned}
2\Omega'_{31} - 4\Omega'_{32} + c'D^{-1}\Omega_{33} &= -2p\mathbb{C}[\mu_1, \nu Z|T=1] + 2p\mathbb{C}[\tau, e\nu Z|T=1] + p\mathbb{C}[\mu_0, \nu Z|T=1] && (57)\\
&= -p\mathbb{C}[\mu_0, \nu Z|T=1] - 2p\mathbb{C}[\tau, (1-e)\nu Z|T=1] && (58)
\end{aligned}$$

and thus to first order

$$\begin{aligned}
n\mathbb{V}[\widehat{\theta}] &= \frac{1}{p^2}\left\{\Omega_{11} + \left(2\Omega'_{31} - 4\Omega'_{32} + c'D^{-1}\Omega_{33}\right)D^{-1}c\right\} && (59)\\
&= \text{SEB} - 2\frac{1}{p}\left\{\mathbb{V}[\mu_0|T=1] - \mathbb{C}[\mu_0, \mu_1|T=1]\right\} + \mathbb{E}\left[\frac{(\mu_0-\alpha)^2 e}{p^2(1-e)}\right] && (60)\\
&- \mathbb{C}[\mu_0, \nu Z|T=1]\mathbb{E}[\nu^2 e(1-e)ZZ']^{-1}\mathbb{C}[\nu Z, \mu_0|T=1] && (61)\\
&- 2\mathbb{C}[\tau, (1-e)\nu Z|T=1]\mathbb{E}[\nu^2 e(1-e)ZZ']^{-1}\mathbb{C}[\nu Z, \mu_0|T=1] && (62)
\end{aligned}$$

Appendix II

The conclusions reached in Section II, based on what we called "Frölich's baseline DGP", are in fact valid for all the DGPs considered in Frölich (2004). These DGPs can be written as

$$\begin{aligned}
Y_i(0) &= m(F(\sqrt{2}X_i)) + \varepsilon_i && (63)\\
p(X_i) &= \alpha + \beta F(\sqrt{2}X_i) && (64)\\
T_i &= \mathbf{1}(p(X_i) \leq v_i) && (65)
\end{aligned}$$

where X_i is distributed standard normal, $F(\cdot)$ is a logistic distribution function, v_i is distributed standard uniform, and ε_i distributed uniform with mean zero and variance $\sigma^2 = 0.01$. The distribution of $F(\sqrt{2}X_i)$ is known as the Johnson S_B distribution.

Frölich (2004) considers thirty DGPs, corresponding to all possible combinations of five density "designs" and six outcome "curves". A design refers to the distribution of the propensity score in the treated relative to the nontreated population. This is manipulated by the parameters α and β in equation (64) which are defined in Frölich (2004)'s

Table 1. An outcome curve, on the other hand, refers to the function $m(\cdot)$ that controls the dependence of the outcome on the rescaled propensity score $Z_i = F(\sqrt{2}X_i)$. The first outcome function is linear in Z_i with a positive slope, while the rest are highly nonlinear. The six functions in question are defined in Frölich (2004)'s Table A1. For easy reference, we reproduce here both Table 1 and Table A1 of Frölich (2004):

TABLE 1 OF FRÖLICH (2004): DENSITY DESIGNS

Design	α	β	Control-treated Ratio
1	0	1	1:1
2	0.15	0.7	1:1
3	0.3	0.4	1:1
4	0	0.4	4:1
5	0.6	0.4	1:4

TABLE A1 OF FRÖLICH (2004): OUTCOME CURVES

Curve	Functional Form of $m(z_i)$, $z_i = F(\sqrt{2}X_i)$
1	$0.15 + 0.7z_i$
2	$0.1 + \frac{z_i}{2} + \frac{1}{2}\exp\left[-200\left(z_i - 0.7\right)^2\right]$
3	$0.8 - 2\left(z_i - 0.9\right)^2 - 5\left(z_i - 0.7\right)^3 - 10\left(z_i - 0.6\right)^{10}$
4	$0.2 + \sqrt{1 - z_i} - 0.6\left(0.9 - z_i\right)^2$
5	$0.2 + \sqrt{1 - z_i} - 0.6\left(0.9 - z_i\right)^2 + -0.1z_i\cos\left(30z_i\right)$
6	$0.4 + 0.25\sin\left(8z_i - 5\right) + 0.4\exp\left[-16\left(4z_i - 2.5\right)^2\right]$

We replicate the main results of Frölich (2004). In Appendix Table A.1, we focus on three estimators based on the true propensity score: pair matching, ridge matching and Frölich's version of reweighting. The first two columns present the mean squared error (MSE) of reweighting and ridge-matching *relative* to that of pair-matching, using the true propensity score, as they were published.[27] The first six rows present results for the first density design, the second six rows those for the second design, and so on. Within each block, each row corresponds to a DGP based on outcome curves 1 to 6. We are able to replicate these results in columns 3 and 4. The differences between these columns are small and generally consistent with simulation error.

Frölich provided us with a copy of the code that produces his result. Upon inspecting this code carefully, we found a small mistake regarding ridge matching. Specifically, the denominator of the second term specified in Table 1, above, should go to infinity as the bandwith h goes to infinity. Instead, $rh|\hat{\Delta}_i|$ is set to 0 when $h \to \infty$, where r is the ridge parameter, h is the bandwidth, and $\hat{\Delta}_i$ is as defined in Table 1. This causes the ridge estimator to be different from the sample mean in cases in which $h \to \infty$. Column 5 shows our replication of the results for the ridge matching estimator using a similar code as the one used in Frölich (2004). The results of column 5 are closer to the published version than those of column 4. For the rest of the paper we use a ridge estimator that lets $rh|\hat{\Delta}_i| \to \infty$ when $h \to \infty$. This should improve somewhat the performance of ridge matching, relative to that documented in Frölich (2004), but in these DGPs seems to slightly worsen the MSE.

In light of the conclusions of Section II, we then change the variance of the outcome error to $\sigma^2 = 0.1$. In such a DGP, columns 6 and 7 show that the performance of reweighting relative to pair matching improves significantly in small samples as it did in large samples (although ridge-matching is still better than Frölich reweighting). For designs 3 and 4 Frölich reweighting dominates pair matching in terms of MSE.

The DGP (64)-(63) does not specify an outcome equation for the observations that received treatment. This is because Frölich (2004) focuses on the estimation of the counterfactual mean under treatment, or $\mathbb{E}[Y_i(0)|T_i = 1]$.

[27] In particular, the first column corresponds to the tenth column of Table 2, and the second column corresponds to the eleventh column of Table 4 of Frölich (2004). We present here the results of ridge matching using an Epanechnicov kernel; the conclusions do not change when utilizing a Gaussian kernel.

As we have already argued, we think it is more natural to analyze the TOT directly. Thus, we need to specify the (potential) outcome equation under treatment: $Y_i(1) = \theta + Y_i(0) + \varepsilon_i$. As before, $\theta = 0$. Columns 8-11 of Appendix Table 1 show the relative MSE of Frolich's reweighting and ridge matching increase when we change the estimand. The broad conclusion of Frölich (2004) that pair-matching performs better than reweighting still holds. As we discussed in Section II this conclusion is basically driven by the choice of a DGP that has an outcome error term with a small enough σ^2.

References

Abadie, Alberto and Guido W. Imbens, "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, January 2006, *74* (1), 235–267.

Bell, B.D. and M.K. Pitt, "Trade Union Decline and the Distribution of Wages in the UK: Evidence from Kernel Density Estimation," *Oxford Bulletin of Economics and Statistics*, 1998, *60* (4), 509–528.

Biewen, M., "Measuring the Effects of Socio-Economic Variables on the Income Distribution: An Application to the East German Transition Process," *Review of Economics and Statistics*, 2001, *83* (1), 185–190.

Blinder, Alan, "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, 1973, *8* (4), 436–455.

Budd, John W. and Brian P. McCall, "The Grocery stores wage distribution: A semi-parametric analysis of the role of retailing and labor market institutions," *Industrial and Labor Relations Review*, March 2001, *54* (2), 484–501.

Busso, Matias, "A Sequential Method of Moments Variance Estimator of Weighting Estimators of Average Treatment Effects," Unpublished manuscript, University of Michigan 2008.

____ , John DiNardo, and Justin McCrary, "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects," *Unpublished manuscript, University of Michigan and University of Californa–Berkeley*, 2008.

Chen, Xiaohong, Han Hong, and Alessandro Tarozzi, "Semiparametric Efficiency in GMM Models with Auxiliary Data," *Annals of Statistics*, forthcoming 2008.

Dehejia, Rajeev H., "Program evaluation as a decision problem," *Journal of Econometrics*, 2005, *125* (1-2), 141–173.

DiNardo, John, Nicole Fortin, and Thomas Lemieux, "Labor Market Institutions and The Distribution of Wages, 1973-1993: A Semi-Parametric Approach," *Econometrica*, September 1996, *64* (5), 1001–1045.

Freedman, David A. and Richard A. Berk, "Weighting Regressions by Propensity Scores," *Evaluation Review*, 2008, *32.*

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

Graham, Bryan S., "Efficient estimation of missing data models using moment conditions and semiparametric restrictions," 2008. Unpublished manuscript. U.C. Berkeley.

Hahn, Jinyong, "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, March 1998, *66* (2), 315–331.

Heckman, James J. and R. Robb, "Alternative Methods for Evaluating the Impact of Interventions," in James J. Heckman and R. Singer, eds., *Longitudinal Analysis of Labor Market Data*, Cambridge University Press Cambridge 1985.

____ , Hidehiko Ichimura, and Petra Todd, "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, April 1998, *65* (2), 261–294.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder, "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, July 2003, *71* (4), 1161–1189.

Imbens, Guido W., "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," *Review of Economics and Statistics*, February 2004, *86* (1), 4–29.

Khan, Shakeeb and Elie Tamer, "Irregular Identification, Support Conditions, and Inverse Weight Estimation," Unpublished manuscript, Northwestern University 2007.

Lunceford, Jared K. and Marie Davidian, "Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study," *Statistics in Medicine*, 15 October 2004, *23* (19), 2937–2960.

Maddala, G.S., *Limited-dependent and qualitative variables in econometrics* number 3. In 'Econometric Society monographs in quantitative economics.', Cambridge [Cambridgeshire] ; New York: Cambridge University Press, 1983.

——— , "Disequilibrium, Self-Selection, and Switching Models," in Zvi Griliches and Michael D. Intriligator, eds., *Handbook of Econometrics*, Vol. III, Amsterdam: Elsevier, 1986, chapter 28, pp. 1633–1688.

McCrary, Justin, "The Effect of Court-Ordered Hiring Quotas on the Composition and Quality of Police," *American Economic Review*, March 2007, Unpublished manuscript, University of Michigan.

Muirhead, Robb J., *Aspects of Multivariate Statistical Theory*, Hoboken: John Wiley and Sons, 2005.

Newey, Whitney, "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, April-June 1990, *5* (2), 99–135.

Oaxaca, R., "Male-female wage differentials in urban labor markets," *International Economic Review*, 1973, *14*, 693–709.

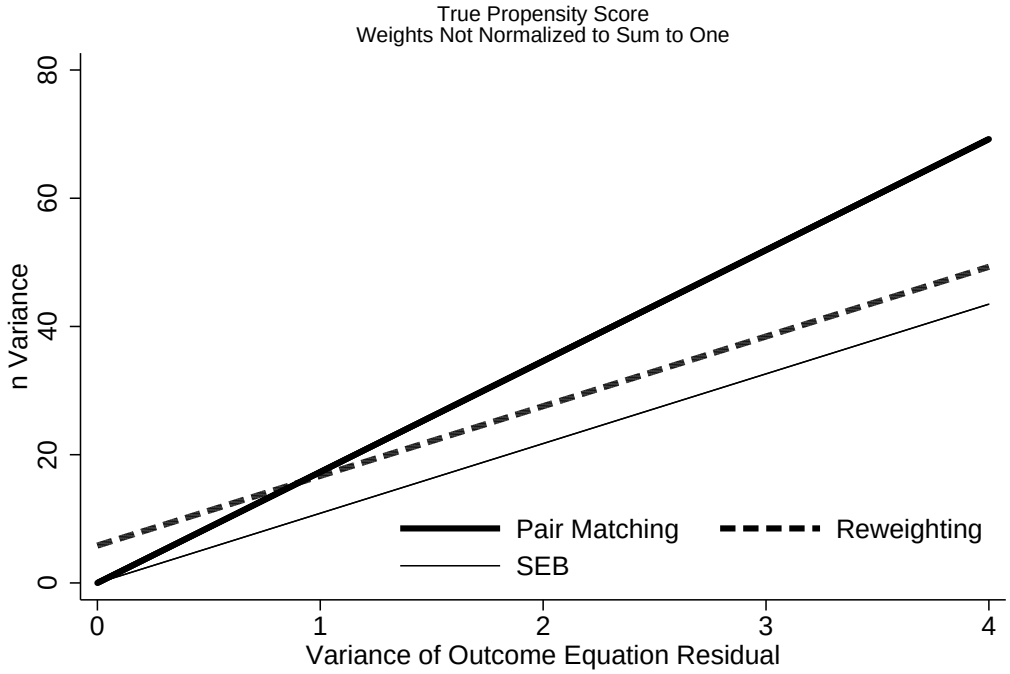
Rosenbaum, Paul R. and Donald B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, April 1983, *70* (1), 41–55.

Seifert, Burkhardt and Theo Gasser, "Data Adaptive Ridging in Local Polynomial Regression," *Journal of Computational and Graphical Statistics*, June 2000, *9* (2), 338–360.

Smith, Jeff and Petra Todd, "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?," *Journal of Econometrics*, September 2005, *125* (1–2), 305–353.

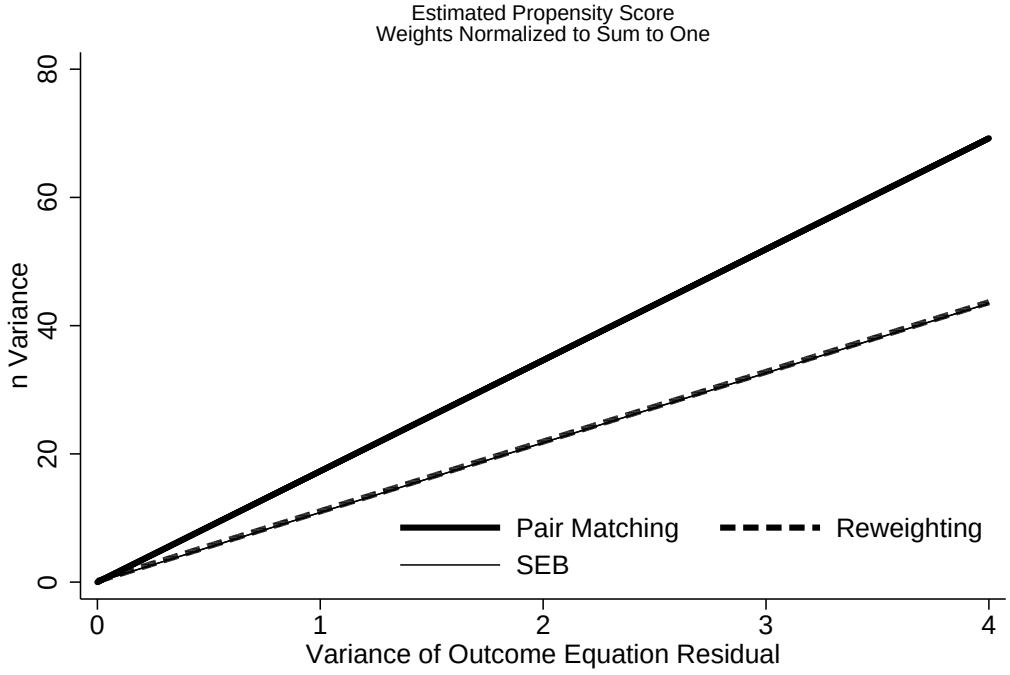
Wishart, John, "The Generalised Product Moment Distribution in Samples from a Normal Multivariate Population," *Biometrika*, July 1928, *20A* (1/2), 32–52.

Figure 1. Variance of Pair Matching and Reweighting



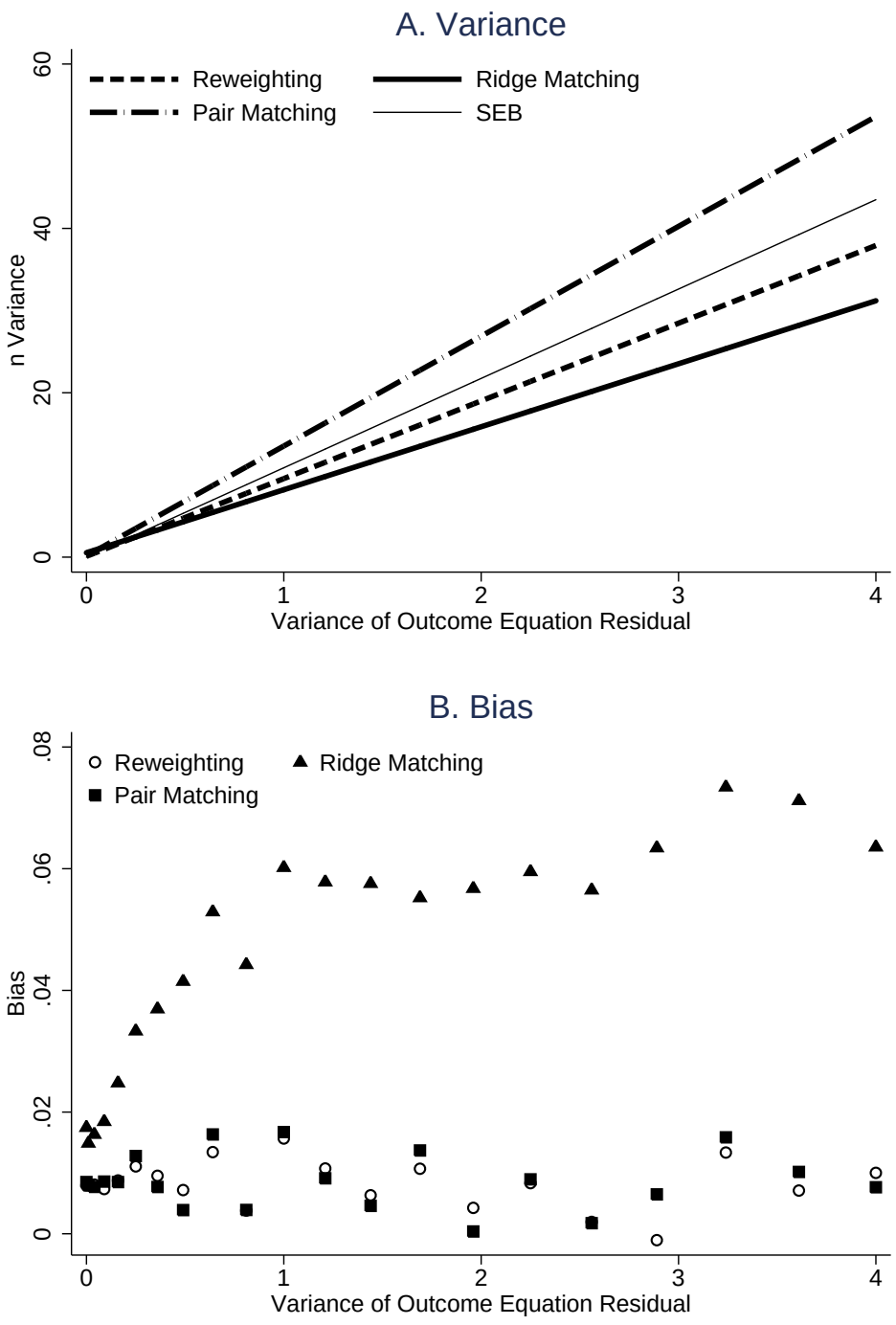
Note: SEB is the semiparametric efficiency bound. Variance refers to the asymptotic variance. See text for details.

Figure 2. Variance of Pair Matching and Reweighting



Note: SEB is the semiparametric efficiency bound. Variance refers to the asymptotic variance. See text for details.

Figure 3. Variance and Bias of Reweighting,
Ridge Matching, and Pair Matching: Sample Size 100



Note: Variance and bias were calculated by simulation with sample size 100. See text for details.

Table 4: Variance and Bias

DGP assumes Var[e]= 0.1 and n=100

Design	Curve		Limiting Variance		n Variance			1000 x |Bias|		
		SEB	Pair Match	Correct Reweight	Pair Match	Ridge Match	Correct Reweight	Pair Match	Ridge Match	Correct Reweight
		[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
1	1	1.09	1.73	1.41	1.36	0.91	1.08	9.95	19.78	9.73
	2	1.09	1.73	1.31	1.42	1.11	1.10	4.45	1.45	0.62
	3	1.09	1.73	1.20	1.35	1.07	1.03	1.45	6.74	2.41
	4	1.09	1.73	1.58	1.40	0.95	1.16	18.78	35.54	11.59
	5	1.09	1.73	1.48	1.41	0.99	1.12	11.74	32.37	10.29
	6	1.09	1.73	1.28	1.40	1.05	1.11	7.74	5.78	0.71
2	1	0.55	0.92	0.68	0.93	0.63	0.65	2.48	9.33	2.29
	2	0.55	0.92	0.75	0.98	0.73	0.76	1.34	6.61	4.59
	3	0.55	0.92	0.63	0.93	0.68	0.65	2.69	0.82	2.45
	4	0.55	0.92	0.66	0.94	0.67	0.68	4.37	13.86	1.31
	5	0.55	0.92	0.65	0.93	0.68	0.65	1.95	11.26	0.19
	6	0.55	0.92	0.72	0.94	0.68	0.75	3.42	3.71	3.01
3	1	0.44	0.76	0.55	0.78	0.50	0.48	1.40	4.90	1.38
	2	0.44	0.76	0.62	0.78	0.58	0.57	0.14	3.60	3.59
	3	0.44	0.76	0.51	0.78	0.53	0.50	1.25	1.40	2.50
	4	0.44	0.76	0.50	0.77	0.51	0.50	2.47	4.27	0.24
	5	0.44	0.76	0.50	0.75	0.51	0.49	2.50	4.54	0.65
	6	0.44	0.76	0.60	0.79	0.57	0.60	0.39	1.98	5.44
4	1	0.68	1.27	0.82	1.36	0.79	0.76	3.94	7.61	0.37
	2	0.68	1.27	0.95	1.37	0.87	0.91	1.22	3.24	2.85
	3	0.68	1.27	0.78	1.31	0.79	0.77	0.06	1.74	2.12
	4	0.68	1.27	0.79	1.31	0.80	0.73	2.70	3.69	1.99
	5	0.68	1.27	0.78	1.33	0.81	0.75	1.30	3.04	0.81
	6	0.68	1.27	0.92	1.36	0.82	0.92	1.05	3.26	3.70
5	1	1.30	2.01	1.89	1.62	1.29	1.23	14.44	34.09	17.53
	2	1.30	2.01	1.72	1.85	1.62	1.41	10.74	14.75	3.80
	3	1.30	2.01	1.51	1.60	1.17	1.17	2.65	21.57	9.72
	4	1.30	2.01	1.90	1.56	1.03	1.16	21.85	33.43	21.45
	5	1.30	2.01	1.79	1.61	1.06	1.20	17.45	28.64	16.94
	6	1.30	2.01	1.61	1.63	1.37	1.37	5.10	6.62	4.28

Note: Results based on 10,000 reps. SEB is the semiparametric efficiency bound. Formulas for the estimators and the limiting variances are given in the text. The limiting variance for ridge matching is unknown in the literature. Ridge uses an Epanechnicov Kernel and the bandwidth was selected by leave 1 out CV. Correct reweighting is an estimator whose weights are normalize to one and based on an estimated propensity score. For design 1 we used the correct model for the propensity score. For designs 2-5 we approximated it by using an overfit logit model (with square terms).

Table A.1: MSE Relative to Pair-Matching

N=100. Estimation using known p(X).

Design	Curve	Estimand=E[Y0|T=1] Var[e]=0.01 (Published)		Estimand=E[Y0|T=1] Var[e]=0.01 (Replication)			Estimand=E[Y0|T=1] Var[e]=0.1 (Extension)		Estimand=TOT Var[e]=0.01 (Extension)		Estimand=TOT Var[e]=0.1 (Extension)	
		Frolich Reweight	Ridge Match*	Frolich Reweight	Ridge Match	Ridge Match*	Frolich Reweight	Ridge Match	Frolich Reweight	Ridge Match	Frolich Reweight	Ridge Match
		[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
1	1	2555	76.2	2603	80.8	76.8	439	62.9	4162	80.5	539	67.9
	2	922	70.1	872	74.7	73.9	248	73.8	1854	71.4	344	78.2
	3	2528	78.4	2516	78.6	78.8	381	72.9	4283	78.9	492	77.4
	4	958	84.6	934	94.0	83.6	209	71.4	2070	98.7	349	76.7
	5	1033	87.9	1017	96.8	88.1	221	69.0	2239	102.5	366	74.6
	6	1111	81.8	1144	82.7	81.6	261	69.4	2403	80.7	350	75.1
2	1	812	77.2	784	76.7	76.2	171	60.1	2714	72.3	327	67.3
	2	435	83.5	433	84.1	83.1	130	68.6	1849	85.7	261	76.0
	3	885	73.2	883	73.8	71.7	162	67.9	2946	74.4	337	74.9
	4	449	72.9	463	75.3	73.5	106	66.1	2679	79.3	342	75.7
	5	475	77.2	475	78.9	75.8	108	67.3	2826	87.0	341	75.7
	6	452	79.8	448	79.9	79.5	128	64.3	1878	78.8	253	73.0
3	1	326	77.7	334	76.5	76.3	100	56.4	2395	67.8	288	65.2
	2	224	82.5	227	82.2	81.4	86	63.3	1774	79.3	236	73.8
	3	371	70.2	354	68.5	69.4	91	58.0	2614	69.0	318	69.8
	4	157	70.3	155	71.7	70.6	59	58.8	3039	74.6	365	70.6
	5	163	74.1	162	75.6	75.1	61	60.4	3267	82.4	393	72.2
	6	210	76.5	198	77.2	77.1	79	59.1	1828	73.5	244	69.4
4	1	198	72.9	201	72.2	72.0	63	38.2	2927	58.4	337	57.6
	2	106	76.6	109	76.4	75.9	52	44.0	2309	66.6	279	64.4
	3	226	62.4	227	62.8	63.5	62	36.2	3236	61.6	375	60.0
	4	126	65.7	125	63.9	65.1	40	37.4	3467	63.0	400	61.1
	5	131	62.3	127	64.0	64.1	42	37.7	3450	67.3	413	62.5
	6	107	75.8	105	75.3	76.0	46	42.2	2129	65.8	257	61.7
5	1	2243	91.5	2035	101.9	94.0	381	82.3	3167	104.5	475	83.8
	2	667	75.1	678	81.5	78.4	247	81.7	1088	84.1	312	84.4
	3	2156	97.9	2277	97.8	95.1	325	77.5	3572	101.2	471	79.7
	4	771	93.6	738	102.2	93.3	182	68.3	2149	104.6	391	70.9
	5	755	89.8	815	100.1	90.4	197	66.8	2088	102.9	398	69.6
	6	920	94.1	881	85.3	94.4	240	78.4	1630	87.3	342	81.6

Note: Results based on 10,000 reps. Ridge uses an Epanechnicov Kernel and the bandwidth was selected by leave 1 out CV. Correct reweighting is an estimator whose weights are not normalized to one. Formulas for the estimators are given in the text. All estimators are based on the true propensity score. * As mentioned in Appendix II we found a mistake in the cross validation procedure used in Frolich (2004). A * means that the CV was implemented as in the published version. No * means that the CV was properly implemented.