NEW EVIDENCE ON THE FINITE SAMPLE PROPERTIES OF PROPENSITY SCORE REWEIGHTING AND MATCHING ESTIMATORS

Matias Busso, John DiNardo, and Justin McCrary*

Abstract—Frölich (2004) compares the finite sample properties of reweighting and matching estimators of average treatment effects and concludes that reweighting performs far worse than even the simplest matching estimator. We argue that this conclusion is unjustified. Neither approach dominates the other uniformly across data-generating processes (DGPs). Expanding on Frölich's analysis, this paper analyzes empirical as well as hypothetical DGPs and also examines the effect of misspecification. We conclude that reweighting is competitive with the most effective matching estimators when overlap is good, but that matching may be more effective when overlap is sufficiently poor.

I. Introduction

A common goal of empirical work is to assess the impact of a nonrandomized program on a subpopulation of interest. Estimates of program impacts are often based on reweighting or on matching on covariates or the propensity score. The empirical literature, particularly in economics but also in medicine, political science, sociology, and other disciplines, features an extraordinary number of program impact estimates based on these estimators. Propensity score matching is particularly popular and is described by Smith and Todd (2005) as "the estimator du jour in the evaluation literature."

Frölich (2004) uses simulation to examine the finite sample properties of various propensity score matching estimators and compares them to those of a particular reweighting estimator. To the best of our knowledge, his is the first paper to explicitly compare the finite sample performance of propensity score matching and reweighting.[1] The topic is an important one because large sample theory is currently available only for some matching estimators and because there can be meaningful discrepancies between large and small sample performance.[2]

Summarizing his findings regarding the mean squared error of the various estimators studied, Frölich (2004, p. 86) states that the "the weighting estimator turned out to be the worst of all [estimators considered]. . . . it is far worse than pair matching in all of the designs." This conclusion is at odds with some of the conclusions from the large sample literature. For example, pair matching is well understood to have higher variance than other matching estimators, and Hirano, Imbens, and Ridder (2003) show that reweighting with a nonparametric estimate of the propensity score can be semiparametrically efficient.[3] The seeming juxtaposition of these conclusions motivated us to reexamine the evidence.

We build on the analysis in Frölich (2004) by presenting evidence on the finite sample performance of a broad set of matching and reweighting estimators over a broad set of data-generating processes (DGPs). We consider nearest-neighbor matching on covariates and on the propensity score with and without bias correction, local linear matching on the propensity score, and three types of reweighting estimators. The DGPs we consider are based on hypothetical data, following Frölich (2004), as well as more empirical DGPs based on the National Supported Work (NSW) Demonstration program data previously studied by Dehejia and Wahba (1999), Smith and Todd (2005), and others.

We conclude that reweighting is a much more effective approach to estimating average treatment effects than is suggested by the analysis in Frölich (2004). In particular, in finite samples, an appropriate reweighting estimator nearly always outperforms pair matching and is often competitive with the more sophisticated matching estimators in DGPs where overlap is good.

In DGPs where overlap is poor, however, reweighting tends not to perform as well as some of the more effective matching estimators. One of the most effective of these is bias-corrected matching with a fixed number of neighbors. Because the relative performance of estimators hinges so powerfully on features of the DGP, we suggest that researchers estimate average treatment effects using a variety of approaches; researchers may also want to conduct a small-scale simulation study designed to mimic their empirical context.

The remainder of the paper is organized as follows. Section II defines notation and estimators. In section III, we replicate and extend the findings of Frölich (2004). We consider matching and reweighting on the parametrically estimated propensity score rather than the true propensity score, and we further examine the performance of several estimators not considered in that article, including normalized reweighting and bias-corrected matching. A limitation of this analysis is that the DGPs examined are rather stylized and involve only a single covariate. In section IV, we use a more empirically grounded DGP with multiple covariates to assess estimator

Received for publication Januray 31, 2009. Revision accepted for publication August 14, 2013. Editor: Alberto Abadie.

*Busso: Inter-American Development Bank; DiNardo: University of Michigan and NBER; McCrary: University of California Berkeley and NBER.

For comments that improved the paper, we thank Alberto Abadie, Matias Cattaneo, Keisuke Hirano, Guido Imbens, Patrick Kline, four anonymous referees, and seminar participants at Berkeley, Wisconsin, and the Atlantic Causal Inference Conference, but in particular Bryan Graham, Jack Porter, and Jeff Smith. We also like to thank Markus Frölich for providing us copies of the code used to generate the results in Frölich (2004).

[1] More recently, Huber, Lechner, and Wunsch (2010, 2013) have investigated these issues as well.

[2] Large sample properties of these estimators are studied in Heckman, Ichimura, and Todd (1998), Hirano et al. (2003), Lunceford and Davidian (2004), and Abadie and Imbens (2006), among others.

[3] As we discuss below in greater detail, we study the properties of parametric reweighting, or reweighting with a parametric logit model, for the propensity score, as in Wooldridge (2007). Hirano et al. (2003) focus on semiparametric reweighting, or reweighting with a series logit model for the propensity score where the series approximation is more complex in larger samples.

The Review of Economics and Statistics, December 2014, 96(5): 885–897

doi:10.1162/REST_a_00431

performance in a more realistic setting. Following the recent literature on this topic, we focus on the context of the NSW observational data. The analyses in sections III and IV are based on well-specified estimators only and consequently do not allow an examination of robustness to common forms of misspecification. To examine the extent to which misspecification affects relative estimator performance, we turn in section V to an examination of DGPs in which linear and interaction terms of multiple covariates may affect both the treatment selection process and the regression functions of the counterfactual outcomes. Section VI concludes.

II. Background and Estimators

In this section we provide a brief discussion of the context in which the estimators we evaluate are applied and then define the estimators considered. (For further detail regarding context, see the excellent review by Imbens, 2004.)

The data observed by the researcher are $(Y_i, T_i, X_i)_{i=1}^{n}$, where Y_i is an outcome, T_i is a treatment indicator, and X_i is a vector of covariates. The outcome observed is either $Y_i(0)$ if $T_i = 0$ or $Y_i(1)$ if $T_i = 1$, where $Y_i(0)$ and $Y_i(1)$ are counterfactual outcomes or outcomes that would be observed under the control and treatment regimes, respectively (Rubin, 1974).

The estimators we evaluate are consistent under the assumptions of conditional independence and overlap. Conditional independence asserts that counterfactual outcomes are independent of the treatment indicator conditional on the covariates. Overlap asserts that the propensity score, $p(x)$, or the conditional probability of treatment given the covariates, is strictly between 0 and 1 for all x.[4] Khan and Tamer (2010) establish that the overlap assumption is not sufficient for $\sqrt{n}$-consistency but that strict overlap is. Strict overlap requires that $p(x)$ be strictly between c and $1-c$ for all x and for some $c > 0$.

There are many possible parameters of interest associated with this model. The literature focuses primarily, although not exclusively, on two parameters: the effect of treatment on the treated (TOT), defined as $\mathbb{E}[Y_i(1) - Y_i(0)|T_i = 1] \equiv \theta$, and the average treatment effect (ATE), defined as $\mathbb{E}[Y_i(1) - Y_i(0)]$. We focus on TOT in the interest of space.

Aside from bias-corrected matching, the matching estimators we examine can be written as

$$\widetilde{\theta} = \frac{\sum_i T_i \{Y_i - \widehat{Y}_i(0)\}}{\sum_i T_i}, \tag{1}$$

where the sums are over all of the data, $\widehat{Y}_i(0) = \sum_j (1 - T_j)W(i,j)Y_j \big/ \sum_j (1 - T_j)W(i,j)$ is the out-of-sample forecast for treated unit i based only on control units j, and the function $W(i,j)$ gives the distance between observations i and j in terms of either covariates or propensity scores, depending on the context.[5] For propensity score-based estimators, we use an estimate of the propensity score rather than the true propensity score, since it is unusual to find empirical applications in which the true propensity score is known.[6] We use a parametric approach where the propensity score model is fixed across samples and the complexity of the model is modest relative to the number of observations.

The other matching estimator we study is bias-corrected matching (Abadie & Imbens, 2011). Motivated by the finding that nearest-neighbor matching is no longer $\sqrt{n}$-consistent when matching more than one continuous variable (Abadie & Imbens, 2006), this approach subtracts an estimate of the asymptotic bias of nearest-neighbor matching from the nearest-neighbor matching estimator itself. We follow the suggestions of Abadie & Imbens (2011, sections 4 and 5) and regression-adjust using a linear regression on the relevant covariates among the matched control units.

Matching estimators require the researcher to choose one or more tuning parameters. Nearest-neighbor matching requires choosing a number of neighbors, and local linear matching requires choosing a bandwidth. For nearest-neighbor matching, we focus on a fixed number of matches. In empirical applications, the number is chosen in order to successfully balance features of the covariate distribution between treatment and control units. Although many of our simulation experiments are based on quite small samples (e.g., $n = 100$), four matches performs quite well in terms of covariate balance and mean-squared error. Consequently, we report results for one and four matches.

Choosing the bandwidth for local linear matching is more challenging. Whereas for nearest-neighbor matching there is always the conservative option of a single match (pair matching), there is no such conservative option for choosing a bandwidth. We follow the suggestion in Frölich (2004) of cross-validation for choosing the bandwidth, a common choice in empirical applications (e.g., Black & Smith, 2004).[7]

[4] The dual assumptions of conditional independence and overlap are referred to as strongly ignorable by Rosenbaum and Rubin (1983).

[5] For example, for nearest-neighbor matching with m matches, $W(i,j)$ is $1/\widetilde{m}$ for the control observations j that are as close to a treated observation i as the mth closest control observation, where $\widetilde{m} \geq m$ is the number of such controls. For details, see Abadie et al. (2004). For local linear matching, $W(i,j) = K_{ij}/\sum_\ell(1 - T_\ell)K_{i\ell} + K_{ij}\Delta_j\Delta_i/((\sum_\ell(1 - T_\ell)K_{i\ell}\Delta_\ell^2) + rh|\Delta_i|)$, where $K_{ij} = K((p_j - p_i)/h)$ for $K(\cdot)$ a kernel function and h a bandwidth, $\Delta_i = p_i - \overline{p}_i$, $\Delta_j = p_j - \overline{p}_i$ for $j \neq i$, $\overline{p}_i = \sum_j(1 - T_j)K_{ij}p_j/\sum_j(1 - T_j)K_{ij}$, and $r = 0.3125$ is an adjustment factor suggested by Seifert and Gasser (2000).

[6] This point is noted in Abadie and Imbens (2012) and Lunceford and Davidian (2004), among many others. Note that while it is relatively efficient to use the estimated propensity score rather than the true propensity score when estimating the ATE, this does not carry through for estimating the TOT. On this point, see Hirano et al. (2003, section 4.3) for the case of semiparametric reweighting and Abadie and Imbens (2012, section III) for the case of matching.

[7] This procedure chooses a bandwidth, h, to minimize $Q(h) = \sum_j (1 - T_j)(Y_j - \widetilde{Y}_{-j,h})^2$ where $\widetilde{Y}_{-j,h}$ is the out-of-sample forecast for control unit j based only on control units $\ell \neq j$. We evaluate $Q(h)$ for $h = 0.01 \times 1.2^{g-1}$ for $g \in \{1, 2, \ldots, 28, 29, \infty\}$. An emerging literature (Flossmann, 2007; 2008; Galdo, Smith, & Black, 2008) considers cross-validation routines specialized to this context, but we leave a full consideration of competing cross-validation proposals to future research.

We are unaware of any theoretical support for cross-validation of matching estimators, but consider the performance of local linear matching, as it exhibited the lowest mean squared error (MSE) of the estimators considered in Frölich (2004).

In summary, we report results for nine matching estimators: nearest-neighbor matching on the propensity score and on covariates with one and four matches, with and without bias correction, and local linear matching on the propensity score with the bandwidth chosen by cross-validation.[8]

In addition to matching estimators, we study unnormalized reweighting, normalized reweighting, and a specific variety of normalized reweighting due to Graham, Pinto, and Egel (2012), which we term GPE reweighting. Unnormalized and normalized reweighting estimators are given by

$$\widehat{\theta}_U = \frac{\sum_i T_i Y_i}{\sum_i T_i} - \frac{\sum_j (1-T_j) W_j Y_j}{\sum_j T_j}, \tag{2}$$

$$\widehat{\theta}_N = \frac{\sum_i T_i Y_i}{\sum_i T_i} - \frac{\sum_j (1-T_j) W_j Y_j}{\sum_j (1-T_j) W_j}, \tag{3}$$

respectively, where $W_j = \widehat{p}(X_j)\big/(1-\widehat{p}(X_j))$ and $\widehat{p}(X_j) = \Lambda(Z_j'\widehat{\beta})$ is the estimated propensity score for unit j based on a logit model, where $\Lambda(v) = 1/(1+\exp(-v))$ and Z_i is a vector of functions of X_i, including a constant term. We choose a small number of functions of X_i where that number is fixed as the sample size grows, as noted above. By focusing on parametric reweighting rather than the semiparametric reweighting that was the focus of Hirano et al. (2003), we hope to approximate the standard practice of applied researchers, most of whom estimate a parsimonious propensity score model based on prior considerations rather than estimating the propensity score nonparametrically. However, this brings up the issue of specifying the propensity score model, which in a sense is analogous to selecting tuning parameters in covariate matching. In sections III and IV, we consider well-specified propensity score models, and in section V, we investigate the consequences of misspecification of the propensity score model.

GPE reweighting is given by equation (3), but $\widehat{p}(X_j)$ is not based on a logit model. To explain the approach, note that if the true propensity score is of the form $\Lambda(Z_i'\beta_0)$ for some parameter β_0, then $0 = \mathbb{E}[(T_i - \Lambda(Z_i'\beta_0))g(Z_i)]$ for any function $g(\cdot)$. This suggests by the analogy principle a class of moment-based estimators for the propensity score indexed by $g(\cdot)$. The logit model uses $g(Z_i) = Z_i$. GPE reweighting with logit functional form uses $g(Z_i) = Z_i/(1-\Lambda(Z_i'\beta_0))$. This leads to exact finite-sample balance, or $\sum_i T_i Z_i / \sum_i T_i = \sum_j (1-T_j) W_j Z_j / \sum_j (1-T_j) W_j$, and thus regression adjustment for Z_i is redundant in the reweighted sample.[9] This redundancy makes GPE reweighting double-robust in the sense of Robins, Rotnitzky, and Zhao (1994): the estimator is consistent if either the propensity score model is correctly specified or if $\mathbb{E}[Y_i(0)|X_i]$ is linear in Z_i.

These three variations of reweighting differ in their prominence in the literature. Asymptotically, GPE has the smallest bias to second order in a class of double robust reweighting estimators for estimating the ATE (Graham et al. 2012), but it has been proposed only recently and so has not been extensively studied or used in empirical work. However double-robust estimators more broadly have been studied extensively in the recent theoretical statistics literature (see Tan, 2010, for a recent review). The unnormalized reweighting estimator dates at least to Horvitz and Thompson (1952) and features prominently in the theoretical statistics and econometrics literatures. The normalized reweighting estimator receives less attention in the theoretical literature but features prominently in empirical work.[10] In our context, normalized reweighting is of particular interest because nearest-neighbor and local linear matching estimators can be interpreted as normalized reweighting estimators.[11] Consequently, a meaningful comparison of matching and reweighting requires that the normalized reweighting estimator be considered. In Frölich (2004), the only reweighting estimator considered is the unnormalized version based on the true propensity score.

III. Previous Results

We turn now to a reexamination of the performance of propensity score reweighting and matching estimators in the context of the DGPs that Frölich (2004) used. Those DGPs can be expressed as

$$Y_i(0) = m(Z_i) + \sigma\varepsilon_i, \tag{4}$$

$$T_i^* = \alpha + \beta Z_i - U_i, \tag{5}$$

where $Z_i = \Lambda(\sqrt{2}X_i)$ is a function of the single standard normal covariate X_i, the error term ε_i is independent and identically distributed (i.i.d.) uniform with a mean of 0 and a variance of 1 and is independent of X_i, the regression function $m(\cdot)$ is one of a list of functions specified in Frolich (2004, table A1), $Y_i(0)$ is the counterfactual outcome under control, $Y_i(1) = Y_i(0)$ is the counterfactual outcome under treatment, the error term U_i is i.i.d. standard uniform and is independent of ε_i and X_i, T_i^* is the latent variable corresponding to treatment ($T_i \equiv \mathbf{1}[T_i^* > 0]$ is the treatment indicator), and α and β are parameters given in table 1 of Frölich (2004).[12] Given this

[8] For matching on covariates, we use the normalized Euclidean metric in light of the sample size.

[9] This is easy to see from

$$\begin{aligned} E[(T_i - p(X_i))\, g(Z_i)] &= E[(1/(1-p(X_i)))(T_i - T_i p(X_i) \\ &\quad + T_i p(X_i) - p(X_i))Z_i] \\ &= E[(T_i - (1-T_i)W_i)\, Z_i]. \end{aligned}$$

[10] A brief list of studies discussing the unnormalized estimator, but not the normalized estimator, includes Rosenbaum (1987, equation [3.1]), Dehejia & Wahba (1997, proposition 4), Wooldridge (2002, Equation [18.22]), and Hirano et al. (2003). The normalized reweighting estimator is discussed in Lunceford and Davidian (2004), Imbens (2004), and Robins et al. (2007), for example.

[11] See, for example, equations (3) and (4) of Abadie and Imbens (2006).

[12] Strictly speaking, Frölich (2004) does not use a model for $Y_i(1)$ at all. This omission is motivated by the recognition that the DGP for $Y_i(1)$ does not affect the relative performance of estimators for TOT. We prefer to be able to discuss the results in terms of traditional notation and models, however, and so we let $Y_i(1) = Y_i(0)$.

Table 1.—Simulation Results
DGP 1 (Frölich): Linear Models with One Covariate Correctly Specified

		Covariate Matching				Propensity Score Matching					Reweighting		
	Design	NN ($k = 1$)	NN ($k = 4$)	BCM ($k = 1$)	BCM ($k = 4$)	NN ($k = 1$)	NN ($k = 4$)	BCM ($k = 1$)	BCM ($k = 4$)	LL (CV)	Not Normalized	Normalized	GPE
$\sigma^2 = 0.01$													
Average	1	9.43	22.61	0.27	0.40	8.43	21.02	0.29	0.41	7.87	4.61	5.39	14.27
|Bias × 1000|	2	3.62	7.75	0.14	0.03	2.82	6.55	0.17	0.02	2.23	0.25	0.58	1.90
	3	1.06	3.81	0.42	0.02	0.68	2.35	0.40	0.05	1.30	0.08	0.09	0.22
	4	0.28	0.95	0.15	0.04	0.19	0.60	0.13	0.05	2.16	1.33	0.67	0.55
	5	8.72	17.07	0.40	0.23	8.25	16.22	0.35	0.24	9.69	18.05	7.64	10.44
Average		4.62	10.44	0.28	0.14	4.07	9.35	0.27	0.15	4.65	4.87	2.87	5.47
Average rank		8.03	10.77	4.63	2.00	7.00	9.30	4.33	2.87	8.80	7.20	6.00	7.07
Average	1	0.17	0.14	0.17	0.11	0.17	0.14	0.16	0.11	0.14	1.70	0.25	0.29
(Var × n)	2	0.10	0.08	0.09	0.06	0.10	0.08	0.09	0.06	0.07	0.23	0.12	0.11
	3	0.08	0.06	0.08	0.05	0.08	0.06	0.08	0.05	0.05	0.10	0.08	0.07
	4	0.13	0.09	0.13	0.09	0.13	0.09	0.13	0.08	0.08	0.08	0.08	0.07
	5	0.16	0.12	0.17	0.12	0.16	0.12	0.17	0.12	0.13	1.76	0.16	0.21
Average		0.13	0.10	0.13	0.09	0.13	0.10	0.13	0.09	0.09	0.77	0.14	0.15
Average rank		8.83	4.70	9.20	4.20	8.70	4.17	8.80	3.20	3.67	9.97	6.90	5.67
$\sigma^2 = 0.1$													
Average	1	9.63	22.88	0.48	0.92	8.65	21.31	0.50	0.95	15.53	4.13	5.47	14.15
|Bias × 1000|	2	3.66	7.80	0.75	0.17	2.86	6.61	0.84	0.14	6.58	0.35	0.69	1.73
	3	1.16	3.76	0.79	0.07	0.89	2.36	0.72	0.07	4.32	0.21	0.25	0.37
	4	0.31	0.94	0.26	0.07	0.27	0.60	0.26	0.07	6.07	1.17	0.56	0.46
	5	9.04	16.93	1.21	0.52	8.53	16.04	1.13	0.57	15.66	18.32	7.94	10.48
Average		4.76	10.46	0.70	0.35	4.24	9.39	0.69	0.36	9.63	4.83	2.98	5.44
Average rank		7.30	10.10	5.33	2.47	7.10	9.00	5.03	2.47	10.03	6.83	5.43	6.90
Average	1	1.39	0.88	1.64	1.05	1.40	0.89	1.64	1.05	1.01	3.02	1.10	1.30
(Var × n)	2	0.90	0.63	0.93	0.65	0.90	0.63	0.92	0.65	0.65	0.75	0.63	0.63
	3	0.77	0.54	0.78	0.55	0.77	0.53	0.77	0.54	0.51	0.52	0.50	0.49
	4	1.33	0.86	1.33	0.86	1.33	0.85	1.33	0.85	0.75	0.72	0.72	0.72
	5	1.50	0.97	1.74	1.16	1.51	0.97	1.74	1.15	1.04	2.93	1.12	1.69
Average		1.18	0.78	1.28	0.85	1.18	0.78	1.28	0.85	0.79	1.59	0.82	0.97
Average rank		9.30	4.17	10.27	5.90	9.70	3.53	10.27	4.83	4.20	7.67	3.67	4.50

Each entry shows the average bias/variance for each design estimator. The data-generating process follows Frölich (2004) with a sample size of n=100. The variance of the error term of the outcome equation (σ^2) is assumed to be 0.01 and 0.1. See section III for datails. NN = nearest-neighbor matching, BCM = bias-corrected matching, LL = local linear matching. For matching estimators, tuning parameter choices specified in parentheses. CV = cross-validation. The propensity score model and the bias adjustment models are correctly specified. Simulation estimates based on 10,000 replications. Estimand is the TOT. Last two lines in each panel show the average of absolute value of bias and the average rank of |bias|.

setup, $p(X_i) = \alpha + \beta\Lambda(\sqrt{2}X_i)$ is the true propensity score. Frölich (2004) sets $\sigma = \sqrt{0.01}$, but we additionally consider larger values of σ. There are five combinations of α and β (selection equation "designs") and six functional forms for $m(\cdot)$ (outcome "curves"), for a total of thirty DGPs.

An issue with these DGPs is that, as originally posed in Frölich (2004), the five designs accord with a standard logit model for treatment only for the special case of design 1, which sets $\alpha = 0$ and $\beta = 1$.[13] In empirical work, logit models are the most common approach used for estimating the propensity score. While we consider misspecification below, we want to begin our analysis by placing all estimators on equal footing in the sense of allowing standard implementations to be well specified. Consequently, in equation (5), we change the distribution of U_i to be standard logistic and replace $\alpha + \beta Z_i$ with the Fourier approximation $k\,(X_i) \equiv \beta_0 + \sum_{\ell=1}^{L}\left\{\beta_\ell^S \sin(\ell x_{i,n}) + \beta_\ell^C \cos(\ell x_{i,n})\right\}$ where $x_{i,n}$ is a rescaled version of X_i that ranges from $-\pi$ to π in each sample.[14] With this specification, the true propensity score is given by $p(X_i) = \Lambda(k(X_i))$.[15]

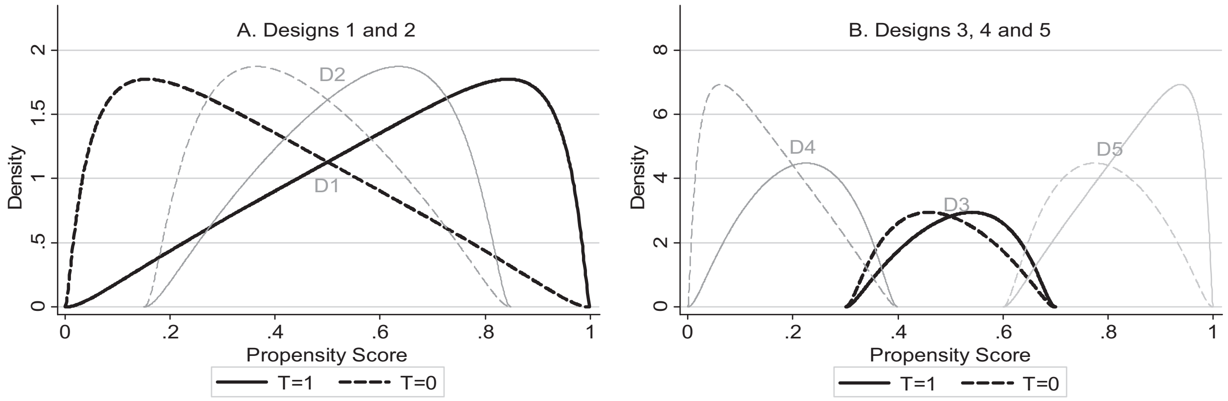
Turning to the implementation of estimators, note that we have several options for estimating a well-specified propensity score model. For example, both a low-dimensional

[13] We note that one could estimate the propensity score by maximum likelihood under the assumption of a uniform error. However, since logit and probit models dominate empirical research, we prefer not to pursue that approach.

[14] The approximation varies by design. In particular, we take $L = 5$ for design 3 and $L = 3$ for other designs, and we set $(\beta_1^S, \beta_2^S, \beta_3^S, \beta_4^S, \beta_5^S, \beta_1^C, \beta_2^C, \beta_3^C, \beta_4^C, \beta_5^C, \beta_0)$ to 3.804583, −1.0764500, 0.2052452, 0, 0, 0.0357783, −0.020493, 0.0052849, 0, 0, −0.0208564 (design 1); 1.6826, −0.159906, 0.07393, 0, 0, 0.0077514, −0.0056401, 0.0016085, 0, 0, −0.0036519 (design 2); 0.8935404, −0.1201665, 0.107236, −0.0283796, 0.0066526, 0.0561901, −0.0460486, 0.0295689, −.0151546, 0.0053763, −0.0291041 (design 3); 2.107084, −0.4891177, 0.0894879, 0, 0, 0.6159659, 0.2563425, −0.0641854, 0, 0, −2.195924 (design 4); and 2.081543, −0.4648486, 0.081096, 0, 0, −0.5374329, −0.2979904, 0.0745325, 0, 0, 2.152262 (design 5).

[15] The coefficients in the $k\,(X_i)$ are chosen to match the distribution of the original propensity score $\alpha + \beta\Lambda(\sqrt{2}X_i)$ in Frölich's (2004) original implementation. This approximation is highly accurate. To substantiate this point, we drew 10,000 samples of various sample sizes ($n = 100$, $n = 1,000$, $n = 10,000$, and $n = 20,000$). For each sample size, we computed the rate at which the Kolmogorov-Smirnov test rejects at the 95% level the null hypothesis of equal distributions for $\alpha + \beta\Lambda(\sqrt{2}X_i)$ and our approximation to it. The rejection rate exceeded the nominal size of the test only for the largest sample size. Consequently, in the sample sizes under discussion here, our approximation is observationally equivalent to Frölich's (2004) original DGP.

FIGURE 1.—FROLICH PROPENSITY SCORE CONDITIONAL DENSITIES



Note: Each selection equation design is made following Frölich (2004) designs. Equation (5) in section III is specified as follows: Design 1 ($\alpha = 0, \beta = 1$), design 2 ($\alpha = 0.15, \beta = 0.7$), design 3 ($\alpha = 0.3, \beta = 0.4$), design 4 ($\alpha = 0, \beta = 0.4$), design 5 ($\alpha = 0.6, \beta = 0.4$).

logit model with the single covariate $k(X_i)$ and a medium-dimensional logit model with the $2L$ covariates that comprise $k(X_i)$ are properly specified parametric models for the propensity score. Relatedly, for the bias adjustment proposed in Abadie and Imbens (2011) to be well specified, we could either regression-adjust using the single covariate $m(\Lambda(k(X_i)))$ or, since $m(\Lambda(k(X_i)))$ is a linear combination of several underlying functions of X_i, we could regression-adjust using those functions.[16] We conducted the simulations using low and medium dimensionality for estimating the propensity score and bias adjusting and found generally similar results. We mention below when the results seem to depend on the dimension of the covariates, but in the interest of space, we present the results that use the single covariate of $k(X_i)$ for estimating the propensity score and $m(\Lambda(k(X_i)))$ for bias adjusting.[17]

Figures 1 and 2 provide a visual assessment of these DGPs. Figure 1 presents population overlap plots for the five different designs, and figure 2 presents the curves used for $m(\cdot)$. Figure 1 shows that designs 1 and 5 violate strict overlap but not overlap and that designs 2, 3, and 4 satisfy the more stringent strict overlap condition.[18] Figure 2 shows the range of shapes taken on by the curves used, from approximately low-order polynomial (e.g., curves 1 and 4) to highly nonlinear (e.g., curves 2 and 6).

Below, we discuss the somewhat surprising result that the variance of ε_i affects the relative performance of matching and reweighting estimators. We show this in a simple way by presenting results based on the exact DGP used in Frölich (2004), which sets $\sigma = \sqrt{0.01}$, and then conducting the same analysis with $\sigma = \sqrt{0.1}$. The tendency of matching to outperform reweighting for very small values of the outcome equation variance is more extreme if the true propensity score is used, as in Frölich (2004), rather than the estimated propensity score.

Our primary aim in this section of the paper is to demonstrate that reweighting is an estimator worth considering in the context of the designs studied in Frölich (2004). This is the opposite of the conclusion in that paper, as we noted in section I. Because of this focus, we do not seek to justify these DGPs. We tend to agree with the position emphasized in Huber et al. (2013) that it is preferable to study DGPs that are empirically relevant, and we do so in the subsequent section. For now, however, our focus is simply assessing whether in the context of these DGPs, it is correct that reweighting performs worse than pair matching and worst among all estimators.

For each of the thirty DGPs outlined, we construct 10,000 samples of size $n = 100$ taken randomly with replacement from the population model described above.[19] Schematically, each sample is constructed in six steps: (a) draw i.i.d. standard normals X_i; (b) draw iid standard logistic errors U_i; (c) construct $T_i^* = k(X_i) - U_i$; (d) assign $T_i = \mathbf{1}(T_i^* > 0)$; (e) draw i.i.d. uniform errors ε_i with mean 0 and variance 1, and (f) construct $Y_i(0) = Y_i(1) = m(\Lambda(k(X_i))) + \sigma\varepsilon_i$. With a sample size this small, many of the estimators we consider may not perform well, particularly in terms of bias. For example, all of the reweighting estimators are consistent but are presumably finite sample biased, even when the propensity score is correctly specified. Similarly, nearest-neighbor matching on covariates or on the propensity score is presumably finite sample biased.

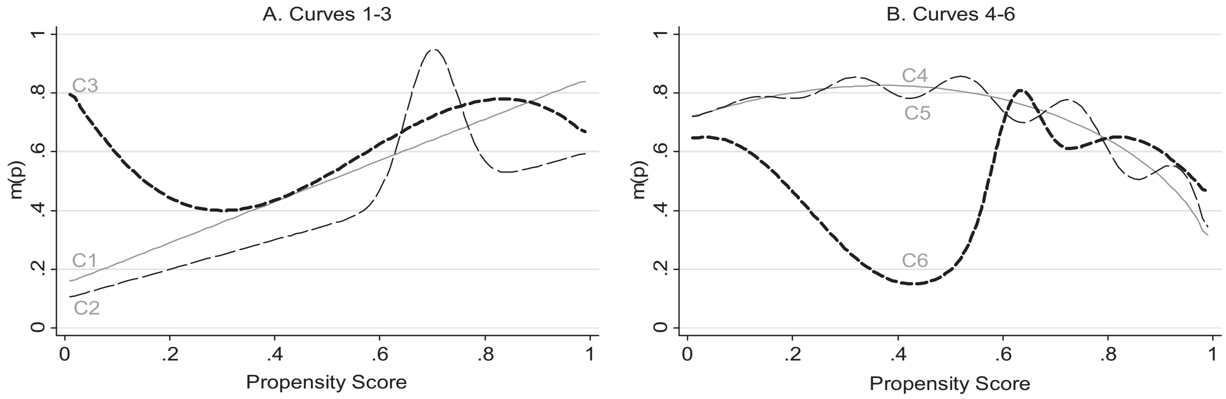
[16] For ease of exposition, let $p_i = p(X_i)$. Then these functions are p_i (curve 1), p_i and $\exp(-200(p_i - 0.7)^2)$ (curve 2), $(p_i - 0.9)^2$, $(p_i - 0.7)^3$, and $(p_i - 0.6)^{10}$ (curve 3), $(p_i - 0.9)^2$ and $\sqrt{1 - p_i}$ (curve 4), $(p_i - 0.9)^2$, $\sqrt{1 - p_i}$, and $p_i \cos(30 p_i)$ (curve 5), and $\sin(8p_i - 5)$ and $\exp(-16(4p_i - 2.5)^2)$ (curve 6).

[17] Full results are available from the authors on request.

[18] It can be shown that for these DGPs, the asymptotic variance for all estimators (adopting either a parametric or semiparametric perspective) is finite, with the exception of local linear matching, where to date the asymptotic variance has not been computed. Consequently the analysis of Khan and Tamer (2010) does not imply a lack of $\sqrt{n}$-consistency for the estimators we study in these DGPs, although it may continue to shed light on some of the differences in results across designs.

[19] Programming of estimators and construction of hypothetical data sets was performed in Stata, version 11.0.

FIGURE 2.—FROLICH OUTCOME CURVES



Note: Each panel displays the conditional expectation of $Y(0)$ given the propensity score p for the curve in question. See the text for details and Frölich (2004, table A1).

Using these samples, we construct simulation estimates of the absolute value of the bias ("absolute bias") and variance.[20] The results are summarized in table 1, which presents estimates of the average absolute bias and average variance for each design across curves for $\sigma = \sqrt{0.01}$ (top two panels) as well as $\sigma = \sqrt{0.1}$ (bottom two panels). This economized presentation of averages is preferred to a presentation of results for all thirty DGPs, as the results are largely similar across curves.[21] For readability, we scale all estimates of the absolute bias by 1,000 and all estimates of the variance by $n = 100$.

We turn first to the results on the absolute bias for the $\sigma = \sqrt{0.01}$ case. Several features stand out. First, bias-corrected matching performs extremely well in terms of bias. Standard errors for the bias estimates are suppressed to economize the presentation, but for entries in the table corresponding to the $\sigma = \sqrt{0.01}$ case, the standard error of the average absolute bias is roughly 0.2 to 0.3 when scaled by 1,000. Consequently, none of the average absolute bias estimates are statistically distinct from 0. Indeed, in unreported results, none of the 30 absolute bias estimates are statistically distinct from 0, as expected, since the regression function is here well specified.

Second, of the remaining estimators, normalized reweighting has the smallest bias, particularly when overlap is good, as in designs 2, 3, and 4. For designs 1 and 5, where overlap is more problematic, normalized reweighting has a larger bias, but it still outperforms pair matching on covariates and on the propensity score. Normalized reweighting performs best in regard to bias among all reweighting estimators. Both unnormalized and GPE reweighting seem to deteriorate faster than normalized reweighting as overlap worsens.[22]

Third, and turning to the variance results for the $\sigma = \sqrt{0.01}$ case, bias-corrected matching on the propensity score performs best, followed by local linear matching and bias-corrected matching on covariates. Normalized reweighting does not perform as well regarding variance in this context, although it outperforms unnormalized reweighting and is at least competitive with GPE reweighting. Consistent with the conclusions in Frölich (2004), unnormalized reweighting performs worst in terms of variance; this result is all the more extreme when reweighting is done using the true propensity score rather than a parametric estimate of the propensity score (results not shown).

Turning to the $\sigma = \sqrt{0.1}$ case, we continue to see good performance from bias-corrected matching and normalized reweighting in regard to the absolute bias. However, perhaps surprisingly in light of the conclusions in Frölich (2004), normalized reweighting now emerges as among the best estimators in terms of variance, particularly when overlap is good. For example, in designs 2, 3, and 4, normalized reweighting (and GPE reweighting) have the smallest average variance. In designs 1 and 5, the variance of normalized reweighting is somewhat higher than that of nearest-neighbor matching with $k = 4$ matches.[23]

Why does normalized reweighting seem to perform relatively better for the $\sigma = \sqrt{0.1}$ case than it does for the $\sigma = \sqrt{0.01}$ case? A simple explanation can be found in

[20] We assume finite first and second moments of these estimators. Examination of QQ-plots for these estimators shows that only unnormalized reweighting, which has rather fat tails, has a distribution that departs from normal.

[21] For example, taking the results for the $\sigma = \sqrt{0.01}$ case and regressing the thirty absolute biases on four dummies for design and five dummies for curves, the between R^2 for designs ranges from 45% to 95% for the case of bias and from 70% to 99% for the case of variance. Curves play no role in the bias of bias-corrected matching in this context, but nonlinearity does lead to greater variance for that estimator. A different pattern holds for nearest-neighbor matching, with curves playing a somewhat important role for bias for designs 1 and 5, where strict overlap is violated, but curves play essentially no role for the variance. For local linear matching, curves play something of a role for bias for all designs but affect variance to a much lesser extent. For reweighting, curves play a role only for designs 1 and 5, but for those designs, curves affect both the bias and the variance. The dominant role of designs relative to curves is somewhat stronger for the $\sigma = \sqrt{0.1}$ case. Full results for all thirty DGPs available from the authors on request.

[22] We note that GPE reweighting is sometimes not computable when normalized reweighting is. In the DGPs being described here, GPE reweighting was computable for 9,664 of the 10,000 simulation runs. GPE reweighting is computable less frequently as the dimensionality of Z_i increases and computation becomes problematic when we use three powers of X_i as elements of Z_i.

[23] On the other hand, nearest-neighbor matching with $k = 4$ matches is notably biased for designs 1 and 5.

asymptotic approximations. In the context of these DGPs, the asymptotic variance of matching estimators is proportional to σ^2 (Abadie & Imbens, 2006, 2011), implying that a small value of σ leads to a small value of the variance of matching estimators. Parametric reweighting estimators, on the other hand, have a constant term that does not involve σ and goes away only as the propensity score model is estimated and indeed overfit (Wooldridge, 2007).

Overall, a researcher with a strong distaste for bias would prefer bias-corrected matching, at least if the regression function were properly specified, as it is here. All of the other estimators have nonnegligible biases. Among the estimators built around a properly specified parametric model for the propensity score, normalized reweighting exhibits the smallest bias and one of the smaller variances, particularly when overlap is good. In the next two sections, we investigate a more empirically relevant DGP with greater dimensionality of the covariates, and we examine the consequences of misspecification of the propensity score model and the regression function.

IV. Results from the National Supported Work Demonstration

In this section, we focus on DGPs based on the data from the National Supported Work (NSW) Demonstration. These data are described in some detail in Dehejia and Wahba (1999) and have been further studied by Smith and Todd (2005), among others. These data have also been the basis for some previous simulation studies (Abadie & Imbens, 2011).

We focus on the African American subsample of those in the experimental group and those in a comparison group taken from the PSID. African Americans comprise roughly 85% of the NSW experimental data. Our study sample consists of 780 individuals (156 experimental, 624 comparison). The covariates we condition on are age, years of education, an indicator for being a high school dropout, an indicator for being married, an indicator for 1974 unemployment, an indicator for 1975 unemployment, 1974 earnings in thousands of dollars and its square, 1975 earnings in thousands of dollars and its square, and interactions between the 1974 and 1975 unemployment indicators and between 1974 and 1975 earnings. Define X_i to be the unique list of variables (i.e., the covariates excluding interactions and square terms), and let Z_i denote the full set of covariates including the interactions and square terms described above. Following the literature, the outcome of interest Y_i is 1978 earnings, again measured in thousands of dollars.

As before, we draw 10,000 hypothetical samples. This time, however, to mimic the original NSW data, we draw $n = 780$ observations for each such sample rather than 100. Schematically, each sample is constructed in eight steps: (a) draw covariates X_i from a population model specified below; (b) draw i.i.d. logistic errors U_i; (c) construct T_i^* according to equation (5), using the full set of covariates Z_i and using in place of α and β the coefficients from a logit model estimated on the original NSW data; (d) assign $T_i = \mathbf{1}(T_i^* > 0)$; (e) draw i.i.d. normal errors ε_{0i} with mean zero and variance σ_0^2 defined below; (f) construct $Y_i(0) = \delta_0' Z_i + \varepsilon_{0i}$, using in place of δ_0 the coefficients from a regression model estimated using the control observations in the original NSW data, where the root mean squared error of the regression is assigned to σ_0^2; (g) construct $Y_i(1)$ analogously, but using the treated units from the original NSW data; and (h) construct $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$.

In order to generate the covariates X_i in each simulation sample, we construct a population model by proceeding in three steps: (a) draw indicators for married, unemployed in 1974, and unemployed in 1975 (a "group") from the empirical distribution of the observed measures in the original study sample; (b) draw age, education, earnings in 1974, and earnings in 1975 from a group-specific multivariate normal distribution; and (c) take the integer part of age and education and impose group-specific minima and maxima on 1974 and 1975 earnings consistent with those in the original study sample. For each group, the parameters of the multivariate normal distribution are taken to be the empirical means of and covariances among age, education, and 1974 and 1975 earnings estimated from the original study sample. The population treatment effect on the treated is $2,334.

Figure 3A presents a sample overlap plot from this DGP.[24] There is very little overlap in the NSW data and therefore in our DGP. Most of the mass for the treatment group is above $p(X_i) = 0.8$, whereas the control group has only five observations in this range. In order to produce data with better overlap, we divide the coefficients in equation (5) by a constant c. The benchmark case is $c = 1$ ("bad overlap"), and we also consider $c = 5$ ("good overlap"). The sample overlap plot for the latter DGP is shown in Figure 3B.

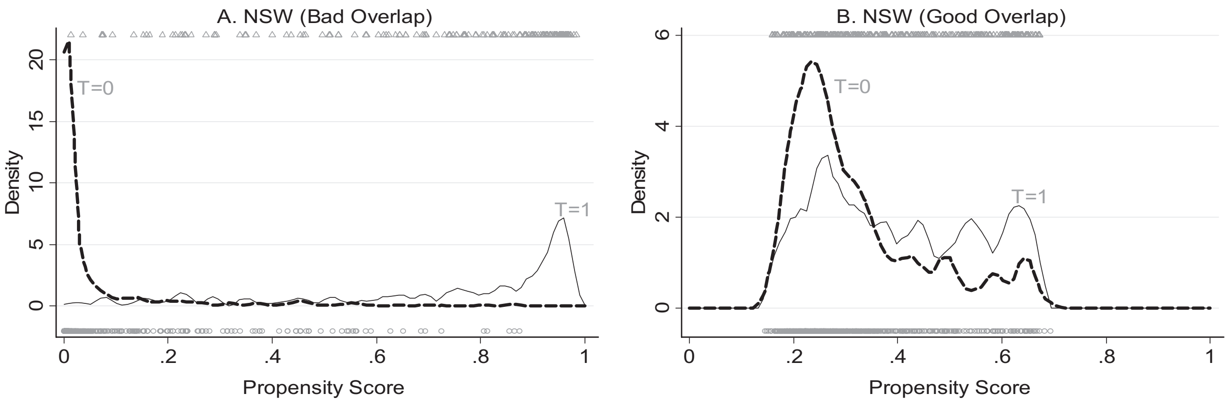
In this section, we are focusing on estimator performance in the context of a more empirically relevant DGP than in those from section III. We consider estimators that are properly specified here and defer consideration of misspecification to section V, below. Consequently, we estimate the propensity score using a logit model and covariates Z_i, and we bias-adjust using a linear regression on the full set of covariates Z_i, estimated using only the matched control units.[25] Simulation estimates of the absolute bias and variance are presented in table 2. Since earnings are measured in thousands of dollars, the scaled bias estimates are in units of dollars.[26] Estimators are given in rows. We first focus on the

[24] A graphical display of the population overlap plot is uninformative here because of the nature of the design. For example, ignoring ties, the distribution of the population propensity score based on the empirical distribution of the covariates is uniform over the sample values for X_i in the NSW study sample, as transformed by $p(\cdot)$.

[25] Bias adjustment could also be done using only the covariates X_i. We prefer in this section to keep all estimators on equal footing in regard to specification.

[26] For each estimator, the standard error on the bias (variance) estimate is about 25 (75).

FIGURE 3.—NSW PROPENSITY SCORE CONDITIONAL DENSITIES



Note: (A) An overlap plot for the original NSW (bad overlap). (B) An overlap plot for the NSW DGP in which the selection equation coefficients were divided by 5 (good overlap). In each case, the solid line is a kernel density estimate of the conditional density of the propensity score among treated units for a representative data set. The dashed line is for the conditional density among control units. Solid triangles at top of figure give propensity score values for treated units, and open circles at the bottom of the figure give propensity score values for control units. See text for details.

TABLE 2.—SIMULATION RESULTS
DGP 2 (NSW): Linear Models with Many Covariates Correctly Specified

			Baseline		Good Overlap	
			|Bias| × 1000 (1)	Variance ×n (2)	|Bias| × 1000 (3)	Variance ×n (4)
Covariate matching	NN	$k = 1$	28.6	4,040.2	60.5	829.3
	NN	$k = 2$	73.7	3,183.8	52.2	716.0
	NN	$k = 3$	120.6	2,771.2	43.0	677.3
	NN	$k = 4$	168.6	2,507.9	33.9	663.0
	BCM	$k = 1$	40.2	5,549.9	7.7	839.0
	BCM	$k = 2$	27.1	4,596.4	12.1	717.2
	BCM	$k = 3$	20.4	4,179.9	14.0	676.5
	BCM	$k = 4$	19.6	3,949.4	12.9	661.3
Propensity score matching	NN	$k = 1$	35.8	5,840.5	24.2	966.5
	NN	$k = 2$	22.0	4,580.6	11.3	781.8
	NN	$k = 3$	1.1	4,055.6	7.5	716.1
	NN	$k = 4$	12.0	3,720.4	2.5	682.5
	BCM	$k = 1$	8.9	6,029.8	9.1	862.3
	BCM	$k = 2$	2.8	5,048.2	14.3	720.3
	BCM	$k = 3$	0.4	4,655.6	12.1	670.6
	BCM	$k = 4$	12.1	4,430.9	13.7	648.9
	Local linear	*CV*	213.6	4,519.2	123.1	663.5
Reweighting	Unnormalized	–	57.6	4,907.1	8.3	585.0
	Normalized	–	31.5	4,061.6	9.8	588.7
	GPE	–	2.4	5,819.8	10.6	580.0

The data-generating process is based on the National Supported Work (NSW) and the PSID data sets. Sample size is $n = 780$. See Section IV for details. CV = cross-validation. k specifies the number of neighbors. The propensity score model and the bias adjustment models are correctly specified parametric model on age, years of education, dropout, married, unemployed in 1974, unemployed in 1975, earnings in 1974 linear and square, earnings in 1975 linear and square, unemployed-in-1974 × unemployed-in-1975, earnings-in-1974 × earnings-in-1975. Simulations based on 10,000 replications. Estimand is the TOT.

case of bad overlap presented in the first two columns of the table.

Two patterns stand out regarding bias. First, despite the difficulties with overlap in this DGP, nearly all estimators show absolute bias of less than $60 (2.5% of the real treatment effect). The estimators that perform badly in terms of bias are nearest-neighbor matching on covariates with many matches and local linear matching. The defects of nearest-neighbor matching on covariates with a large number of matches seem to be cured by either matching on the correctly specified propensity score or bias adjustment. Second, normalized reweighting performs well in terms of bias, but GPE reweighting performs particularly well, closely matching the performance of bias-corrected matching on the propensity score.

In terms of variance, we have a number of interesting results. We focus on estimators with bias smaller than $60. First, nearest-neighbor matching on the propensity score with $k = 4$ matches exhibits the lowest variance. Close competitors are nearest-neighbor matching on covariates with $k = 1$ matches and normalized reweighting. Second, GPE reweighting has a very large variance, among the worst of all estimators considered. This is consistent with the results in section III, where GPE reweighting was notably variable in settings of poor overlap. Third, bias-corrected matching improves on the bias properties of nearest-neighbor matching, but this comes at the expense of added variance. Nearest-neighbor matching declines in variance with additional matches, but this contributes greatly to its bias due to lower match quality. Bias adjustment involves a greater variance than nearest-neighbor matching, but additional matches lead to lower variance for bias-corrected matching without increasing bias.

Turning next to the case of good overlap (columns 3 and 4), nearly all estimators show biases lower than 60, the exception being local linear matching. In this context, it is not as difficult to find good matches, and consequently nearest-neighbor matching on covariates does not suffer as much from increasing the number of matches. Bias-corrected matching on covariates continues to improve on the bias of nearest-neighbor matching on covariates. Matching on the propensity score involves small bias for both nearest-neighbor matching and bias-corrected matching. Reweighting also performs well in terms of bias in the context of good overlap.

In terms of variance, an interesting pattern is that in the context of matching on the propensity score, bias-corrected matching exhibits smaller variance than nearest-neighbor matching. In the context of matching on covariates, bias-corrected matching has somewhat higher variance than nearest-neighbor matching. Reweighting estimators perform best in terms of variance in this DGP, with a variance that is 12% smaller than the other leading estimators. Finally, it is interesting to compare the relative performance of GPE reweighting in the case of good and bad overlap. GPE reweighting is among the most variable estimators when overlap is bad, but its relative performance improves significantly in the case of good overlap.

Overall, the results for the NSW DGP agree with the conclusions reached using Frölich's (2004) DGP. In terms of bias, bias-corrected matching and reweighting tend to perform well when overlap is good, with the latter displaying a smaller variance than the former. When overlap is bad, both bias-corrected matching and reweighting show a relatively small bias, with the former having smaller bias and larger variance than the latter. In the NSW DGPs generally, local linear matching and nearest-neighbor matching on covariates tend to perform worse in terms of bias and variance. To some extent, this is expected since reweighting and bias-corrected matching are implemented here using properly specified parametric models, whereas nearest-neighbor on covariates is fully nonparametric and local linear matching requires the selection of a tuning parameter by cross-validation.

V. Misspecification

The results presented so far have considered only cases in which models of the propensity score and bias adjustment were correctly specified. In this section, we investigate the effects of misspecification on the bias and variance of matching and reweighting estimators.[27] To focus on these issues, we introduce a third set of DGPs.[28]

We draw $n = 400$ observations on four covariates X_i, where X_i is distributed i.i.d. and jointly uniform with mean zero and a block diagonal variance matrix Σ.[29] The block diagonal structure means that X_{1i} and X_{2i} are correlated (as are X_{3i} and X_{4i}) but that X_{1i} is uncorrelated with X_{3i} and X_{4i} (as is X_{2i}). We then generate a latent treatment variable following equation (5), taking Z_i to be a function (specified below) of the covariates X_i and U_i to be distributed i.i.d. standard logistic and independent of X_i. We draw observations on counterfactual outcomes using equation (4) with $m(Z_i)$ a linear function of Z_i and with the new equation, $Y_i(1) = T_i + Y_i(0)$, implying a constant treatment effect of 1. We take ε_i in equation (4) to be iid standard normal and independent of X_i and U_i and set σ in equation (4) to 1.

We draw samples from four DGPs by varying the selection equation and the regression function. The first DGP sets the true selection index and regression function to be a linear combination of the four individual elements of the X_i vector. The second DGP sets the true selection index and regression function to be a linear combination of the six interaction terms of the X_i vector ($X_{1i}X_{2i}, X_{1i}X_{3i}$, and so on). The third DGP sets the true selection index and regression function to be a linear combination of the ten individual and interaction terms. The fourth DGP sets the true selection index to be a linear combination of the four individual terms and the regression function to be a linear combination of the six interaction terms.[30] For all four DGPs, all the coefficients in the selection index and the regression function are 1. [31]

The results of these investigations are presented in table 3. The first two columns describe the estimators, and the next two report on aspects of estimator implementation. In the case of matching, we mainly report matching on $k = 4$ neighbors, except in the case of nearest-neighbor matching without bias correction, which we continue to use as a benchmark.[32] For each estimator, we report results obtained by estimating the propensity score in four different ways: using the true index as a single covariate ("True," column 3), using the four individual elements of X_i ("Linear"), using the six interaction terms ("Interactions"), and using all ten individual and

[27] Kang and Schafer (2007) study the effects of misspecification on reweighting, stratification, and regression estimators of average treatment effects. They find that unnormalized reweighting is severely biased and imprecise when models are misspecified. Drake (1993) finds that treatment effect estimators that misspecify the regression functions have much larger biases than those for estimators that misspecify the propensity score.

[28] We elect not to adapt the DGPs described in previous sections to study misspecification. The DGP used in Frölich (2004) is a function of only one linear covariate. Studying misspecification in the context of the NSW DGP from the previous section has the potential to conflate the issues of misspecification and overlap. We prefer to use a setting where we can focus on the important issue of misspecification in isolation of other considerations.

[29] The upper left and lower right blocks of Σ are given by $\frac{1}{3}\left(\begin{smallmatrix} 1 & -1 \\ -1 & 2 \end{smallmatrix}\right)$.

[30] The obvious fifth DGP, which is analogous to the fourth, but with reversed roles for the selection index and the regression function, shows similar results to the fourth and is omitted in the interest of space.

[31] For the first DGP, the constant in the selection index and the regression function is 0. For the other DGPs, we set the constant in the selection index to 0.65, as this maintains an equal ratio of treated to control units across DGPs.

[32] We also computed matching estimators using $k = 1, 2, 3$ neighbors. To save space, we decided to report $k = 4$ (which minimizes the MSE among most of these estimators). Results for $k = 1, 2, 3$ are available from the authors on request.

TABLE 3.—SIMULATION RESULTS
DGP 3: Linear and Nonlinear Models with Multiple Covariates

		Propensity	Bias Adjustment	DGP: Linear Selection and Outcome Equations		DGP: Interactions Selection and Outcome Equations		DGP: Linear + Interactions Selection and Outcome Equations		DGP: Linear Selection Equation and Interactions Outcome Equation	
Estimator		Score Model	Model	|Bias| × 1000	Var. × n	|Bias| × 1000	Var. × n	|Bias| × 1000	Var. × n	|Bias| × 1000	Var. × n
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Covariate matching	NN ($k = 1$)	–	–	134.6	8.2	78.1	7.4	15.5	7.1	12.7	7.9
	NN ($k = 4$)	–	–	207.5	6.2	113.5	5.7	27.5	5.2	0.8	5.9
	BCM ($k = 4$)	–	True	**0.3**	**6.4**	**0.7**	**5.9**	**0.4**	**5.1**	**0.3**	**5.5**
		–	Linear	**0.3**	**6.4**	114.5	5.7	28.7	5.5	14.1	7.1
		–	Interactions	205.2	6.5	**1.0**	**5.9**	57.0	5.4	0.1	5.6
		–	Linear + Interactions	0.1	6.7	1.1	6.0	**0.4**	**5.6**	0.1	6.7
Propensity score matching	NN ($k = 1$)	True	–	**5.0**	**9.2**	**3.3**	**8.5**	**0.2**	**9.6**	**3.3**	**13.4**
		Linear	–	**3.8**	**9.7**	398.6	11.2	106.1	8.6	**3.5**	**13.7**
		Interactions	–	591.8	11.8	5.4	8.8	171.5	8.9	1.8	8.3
		Linear + Interactions	–	6.5	10.2	4.4	9.2	**2.5**	**9.3**	5.3	12.0
Propensity score matching	NN ($k = 4$)	True	–	**12.6**	**6.5**	**10.9**	**5.9**	**1.0**	**6.6**	**11.2**	**9.1**
		Linear	–	**12.5**	**6.7**	399.5	7.8	104.9	5.9	**9.4**	**9.4**
		Interactions	–	590.8	8.2	**11.6**	**6.0**	169.0	6.0	4.8	5.3
		Linear+Interactions	–	13.3	7.0	11.7	6.3	**2.6**	**6.0**	2.4	7.9
	BCM ($k = 4$)	True	True	**0.7**	**6.6**	**1.1**	**5.9**	**0.4**	**5.5**	**0.7**	**6.5**
		True	Linear	**0.1**	**6.7**	17.9	6.1	5.0	6.0	2.8	9.3
		True	Interactions	30.7	7.1	**1.1**	6.1	12.0	6.0	**0.3**	**6.7**
		True	Linear+Interactions	0.6	7.0	1.2	6.3	**0.6**	**5.9**	0.6	7.0
		Linear	True	**0.2**	**6.7**	1.7	6.0	0.8	5.1	**0.2**	**6.7**
		Linear	Linear	**0.2**	**6.7**	398.4	7.7	109.6	5.9	6.4	9.3
		Linear	Interactions	29.3	7.2	2.0	6.2	12.7	5.9	**0.4**	**6.8**
		Linear	Linear+Interactions	0.4	7.0	2.0	6.2	1.2	5.8	0.4	7.0

TABLE 3.—(CONTINUED)

Estimator		Propensity Score Model	Bias Adjustment Model	DGP: Linear Selection and Outcome Equations		DGP: Interactions Selection and Outcome Equations		DGP: Linear + Interactions Selection and Outcome Equations		DGP: Linear Selection Equation and Interactions Outcome Equation	
				|Bias| × 1000	Var. × n	|Bias| × 1000	Var. × n	|Bias| × 1000	Var. × n	|Bias| × 1000	Var. × n
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Propensity score matching	BCM ($k = 4$)	Interactions	True	1.4	6.7	**0.8**	**6.0**	0.1	5.1	0.7	4.9
		Interactions	Linear	1.4	6.8	18.1	6.2	21.9	5.9	86.6	8.9
		Interactions	Interactions	589.5	8.1	**0.9**	**6.0**	171.1	5.9	0.6	4.9
		Interactions	Linear+Interactions	1.2	7.2	1.2	6.2	0.1	5.9	1.2	7.2
		Linear+Interaction	True	0.0	6.9	1.3	6.2	**1.4**	**5.7**	0.1	6.8
		Linear+Interaction	Linear	0.2	6.9	16.1	6.2	4.2	5.8	0.5	7.9
		Linear+Interaction	Interactions	22.6	7.1	1.3	6.2	11.3	5.9	0.2	6.8
		Linear+Interaction	Linear+Interaction	0.3	7.0	1.6	6.2	**1.5**	**5.8**	0.3	7.0
	LL (CV)	True	–	**14.7**	**5.9**	**10.8**	**5.3**	**13.2**	**6.2**	**26.9**	**9.0**
		Linear	–	**14.6**	**6.1**	426.3	6.7	95.5	5.5	**22.4**	**9.1**
		Interactions	–	597.5	7.1	**9.8**	**5.4**	173.2	5.4	19.3	4.8
		Linear+Interaction	–	14.4	6.3	10.8	5.7	**7.4**	**5.6**	16.2	7.6
Reweight	Unnormalized	True	–	**0.4**	**6.9**	**1.0**	**5.7**	**0.3**	**5.8**	**0.7**	**8.7**
		Linear	–	**0.1**	**7.1**	448.7	5.9	109.9	4.9	**0.1**	**8.8**
		Interactions	–	591.0	6.9	**1.7**	**5.8**	174.9	5.2	0.3	4.2
		Linear+Interaction	–	1.7	7.1	2.0	6.0	**1.1**	**5.1**	2.8	6.8
	Normalized	True	–	**1.5**	**6.6**	**0.5**	**5.5**	**0.3**	**5.8**	**1.3**	**8.7**
		Linear	–	**1.2**	**6.8**	448.5	5.9	110.6	4.9	**0.5**	**8.8**
		Interactions	–	591.0	6.9	**1.0**	**5.7**	174.8	5.2	0.3	4.2
		Linear+Interaction	–	2.5	6.8	1.2	5.9	**1.2**	**5.1**	2.4	6.8
	GPE	True	–	**0.7**	**5.7**	**1.2**	**5.1**	**0.8**	**5.8**	**2.2**	**8.5**
		Linear	–	**0.4**	**5.8**	446.9	5.9	121.1	5.0	**2.2**	**8.4**
		Interactions	–	590.9	6.9	**1.2**	**5.2**	173.3	5.1	**0.0**	**4.2**
		Linear+Interaction	–	0.1	6.1	1.4	5.4	**0.6**	**5.1**	0.1	6.1

See the note to table 1. The data-generating process follows a DGP as specified in section V with $n = 400$. NN = nearest-neighbor matching, BCM = bias-corrected matching, LL = local linear matching. For matching estimators, tuning parameter choices are specified in italics. CV = cross-validation. There are four versions of the propensity score and the outcome equation model: "True Index" means that the index was used (imposing the true parameters); "Linear" means that the models include only linear terms of the covariates; "Interactions" refers to models that only include interactions terms. "Linear + Interact." refers to models that include linear and interaction terms. Each entry in the table is marked to highlight whether the underlying propensity score and bias adjustment models are (in any given DGP) well specified, misspecified, or overspecified: bold font (all models on which the estimator is based are correctly specified), "underline" (at least one model on which the estimator is based is misspecified), "regular font" (no underspecified model is used but at least one model is overspecified).Simulation estimates based on 10,000 replications. Estimand is the TOT. Columns show results for different DGPs in terms of its selection and outcome equation. See text for details.

interaction terms ("Linear + Interact"). We proceed analogously when specifying the bias adjustment model for the bias-corrected matching estimator (column 4). All matching estimators based on covariates are obtained by matching on the individual elements of X_i.[33] The other eight columns of table 3 report absolute bias and variance estimates for each of the four DGPs considered.

With this many possibilities for DGPs and for estimator implementation, it is tedious to keep track of which estimator is well specified. To aid the discussion, we display the absolute bias and variance estimates in bold if the estimator in question is well specified in the given DGP and in black if the estimator is overspecified. We underline the numbers when the model is misspecified.

Turning to the results, we see several interesting patterns. First, as expected, whenever the models for the propensity score or bias adjustment model are well specified, the bias is small, and, conversely, when there is misspecification, the bias can be quite large. Overspecification, on the other hand, does not seem to significantly affect the bias.

Second, in the well-specified case, the most biased estimators are nearest-neighbor matching on the propensity score with four neighbors and local linear matching. Bias-corrected matching and reweighting show the lowest biases. In terms of variance, bias-corrected matching on the covariates and normalized and GPE reweighting tend to perform best, with a possible role for local linear matching, which is, however, rather biased as noted.

Third, consistent with the findings from the NSW design, the dimensionality of the covariates is not pivotal for the relative bias of these estimators but does affect the relative variances somewhat.

Fourth, nearest-neighbor matching on covariates, shown in the first two rows of table 3, tends to be the most biased and most variable. This is explained in part because while reweighting, propensity-score matching estimators, and estimators with a parametric bias adjustment are based on parametric estimation techniques, covariate matching estimators are nonparametric.

Fifth, in the overspecified case, there seems to be little cost in terms of bias of including additional covariates. Depending on the DGP, however, including extraneous covariates can either increase or decrease the variance of the estimators considered here. Typically, including extraneous covariates increases the variance, but the opposite is true for the third DGP considered for several estimators.

Sixth, misspecified estimators have very bad bias, as expected. Bias-corrected matching is more robust to misspecification than the other estimators we consider, in that correct specification of either the list of covariates to be matched on or the regression function suffices for relatively good performance. Consider, for instance, column (5). All three reweighting estimators, local linear matching, and nearest-neighbor matching exhibit bias of about 590, double that of nearest-neighbor matching on covariates with $k = 4$ matches. Bias-corrected matching of covariates performs equal to nearest-neighbor matching on covariates in the case of a misspecified regression function, but it plainly dominates nearest-neighbor matching on covariates for proper specification or overspecification of the regression function. A similar pattern holds for columns 7 and 9. GPE reweighting exhibits the same robustness in the sense that it shows very small bias if the covariates included in the propensity score model are either those that comprise the selection index or those that comprise the regression function. Interestingly, however, when GPE is misspecified, it exhibits a bias roughly twice as large as that of bias-corrected matching on covariates.

Finally, the last two columns of the table display an interesting property of many of these estimators. If the propensity score model is misspecified but includes the covariates of the regression function, then estimators built around the propensity score tend to perform well in terms of bias. This pattern is expected for GPE reweighting, as noted, but may be surprising for the other estimators.

VI. Conclusion

We have presented simulation evidence on the finite sample properties of a variety of matching and reweighting estimators across several DGPs. We considered three DGPs: those studied in Frölich (2004) that have a single covariate, a more empirical DGP based on the NSW data that involves many covariates, and a third DGP that allows us to address the effects of misspecification on these estimators.

In broad strokes, nearest-neighbor matching tends to have small bias, especially with a small number of neighbors, but it can be rather variable, particularly for data sets where the outcome is hard to predict. One approach to variance reduction is to include additional matches. This can lead to problems with worse covariate balance, especially in the presence of many covariates. A possible solution to this problem is bias correction. Bias-corrected matching appears to provide the researcher with insurance in the sense that even in the case of a misspecified regression function, the bias is no worse than with nearest-neighbor matching, but the bias is dramatically reduced when the regression function is properly specified. A researcher with a strong distaste for bias is likely to be interested in bias-corrected matching for these reasons.

Normalized reweighting also exhibits small bias when the propensity score model is correctly specified. Moreover while the bias is usually larger than that of bias-corrected matching, the variance is usually smaller. As a way of guarding against the consequences of misspecification, researchers using estimators built around the propensity score should include in the propensity score model covariates believed to influence the treatment selection process as well as any covariates believed

[33] In all cases, we condition on the four covariates. Conditioning only on one, two, or three out of the four covariates increases the bias significantly for all estimators.

to influence the outcome variable. Doing so provides a type of insurance against bad bias, but this may come at the expense of added variance.

In addition to implementation details, the relative performance of estimators also depends on specific features of the DGP in question. Normalized and GPE reweighting perform well in terms of both bias and variance when strict overlap is satisfied, but they deteriorate as overlap worsens. Bias-corrected matching on covariates often has a higher variance than reweighting when strict overlap is satisfied, but it is less affected by the degree of overlap. In terms of recommendations for empirical practice, our results suggest the wisdom of conducting a small-scale simulation study tailored to the features of the data at hand. At a minimum, we recommend that researchers estimating average treatment effects present results from a variety of approaches, particularly when there is evidence that overlap is poor.

REFERENCES

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

——— "Bias-Corrected Matching Estimators for Average Treatment Effects," *Journal of Business and Economic Statistics* 2 (2011), 1–11.

———- "Matching on the Estimated Propensity Score," Harvard University working paper (2012).

Abadie, Alberto, David Drukker, Jane Leber Herr, and Guido Imbens, "Implementing Matching Estimators for Average Treatment Effects in STATA," *Stata Journal* 4 (2004), 290–311.

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

Dehejia, Rajeev H., and Sadek Wahba, "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," in Rajeev H. Dehejia, ed., *Econometric Methods for Program Evaluation* (Cambridge, MA: Harvard University, 1997).

——— "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association* 448a (1999), 1053–1062.

Drake, Christiana, "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect," *Biometrics* 49 (1993), 1231–1236.

Flossmann, Anton, "Empirical Bias Bandwidth Choice for Local Polynomial Matching Estimators," University of Konstanz working paper (2007).

——— "Optimal Bandwidth Choice for Matching Estimators by Double Smoothing," University of Konstanz working paper (2008).

Frölich, Markus, "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators," this REVIEW 86 (2004), 77–90.

Galdo, Jose, Jeffrey Smith, and Dan Black, "Bandwidth Selection and the Estimation of Treatment Effects with Unbalanced Data," *Annals of Economics and Statistics* 91–92 (2008), 189–216.

Graham, Bryan S., Cristine Campos de Xavier Pinto, and Daniel Egel, "Inverse Probability Tilting for Moment Condition Models with Missing Data," *Review of Economic Studies* 79 (2012), 1053–1079.

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

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

Horvitz, Daniel G., and Donovan J. Thompson, "A Generalization of Sampling without Replacement from a Finite Universe," *Journal of the American Statistical Association* 47 (1952), 663–685.

Huber, Martin, Michael Lechner, and Conny Wunsch, "How to Control for Many Covariates? Reliable Estimators Based on the Propensity Score," IZA discussion paper 5268 (2010).

——— "The Performance of Estimators Based on the Propensity Score," *Journal of Econometrics* 175 (2013), 1–21.

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

Kang, Joseph D. Y., and Joseph L. Schafer, "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," *Statistical Science* 22 (2007), 523–539.

Khan, Shakeeb, and Elie Tamer, "Irregular Identification, Support Conditions and Inverse Weight Estimation," *Econometrica* 78 (2010), 2021–2042.

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* 23 (2004), 2937–2960.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao, "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association* 89 (1994), 846–866.

Robins, James, Mariela Sued, Quanhong Lei-Gomez, and Andrea Rotnitzky, "Comment: Performance of Double-Robust Estimators When Inverse Probability Weights Are Highly Variable," *Statistical Science* 22 (2007), 544–559.

Rosenbaum, Paul R., "Model-Based Direct Adjustment," *Journal of the American Statistical Association* 82 (1987), 387–394.

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

Rubin, Donald B., "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology* 66 (1974), 688–701.

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

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

Tan, Zhiqiang, "Bounded, Efficient, and Doubly Robust Estimation with Inverse Weighting," *Biometrika* 97 (2010), 661–682.

Wooldridge, Jeffrey M., *Econometric Analysis of Cross Section and Panel Data* (Cambridge, MA: MIT Press, 2002).

——— "Inverse Probability Weighted Estimation for General Missing Data Problems," *Journal of Econometrics* 141 (2007), 1281–1301.