The Oaxaca–Blinder Unexplained Component as a Treatment Effects Estimator*

Tymon Słoczyński†

In this paper I use the National Supported Work (NSW) data to examine the finite-sample performance of the Oaxaca–Blinder unexplained component as an estimator of the population average treatment effect on the treated (PATT). Precisely, I follow sample and variable selections from Dehejia and Wahba (1999), and conclude that Oaxaca–Blinder performs better than any of the estimators in this influential paper, provided that overlap is imposed. As a robustness check, I consider alternative sample (Smith and Todd 2005) and variable (Abadie and Imbens 2011) selections, and present a simulation study which is also based on the NSW data.

*I am grateful to two anonymous referees, Arun Advani, Joshua Angrist, Thomas Crossley, Patrick Kline, Paweł Strawiński, and seminar and conference participants in Dublin, Kraków, Odense, and Warsaw for useful comments and discussions. I would like to acknowledge financial support for this research from the Foundation for Polish Science (a START scholarship), the National Science Centre (grant DEC-2012/05/N/HS4/00395), the Warsaw School of Economics (grant 03/BMN/25/11), and the "Weź stypendium – dla rozwoju" scholarship programme. I would also like to thank the Clifford and Mary Corbridge Trust, the Cambridge European Trust, and the Faculty of Economics at the University of Cambridge for financial support which allowed me to undertake graduate studies at the University of Cambridge where this project was started.

†Michigan State University, Warsaw School of Economics, and IZA. E-mail: tymon.sloczynski@gmail.com.

1 Introduction

Recent papers by Barsky et al. (2002), Black et al. (2006), Melly (2006), and Fortin, Lemieux, and Firpo (2011) have noted that the Oaxaca–Blinder decomposition, a popular method used in empirical labour economics to study differentials in mean wages,[1] provides a consistent estimator of the population average treatment effect on the treated (PATT). Precisely, applied researchers in labour economics have often used the Oaxaca–Blinder decomposition to estimate two components of a wage differential: a component attributable to differences in group composition (the explained component) and a component attributable to net effects of group membership (the unexplained component). It is the unexplained component in the most basic version of the Oaxaca–Blinder decomposition which constitutes a consistent estimator of the PATT. In an important contribution, Kline (2011) has recently shown that this method is equivalent to a propensity score reweighting estimator based on a linear model for the treatment odds, and satisfies therefore the "double robustness" property (Robins, Rotnitzky, and Zhao 1994). He has also used the well-known National Supported Work (NSW) data[2] to provide a seminal assessment of the finite-sample performance of the Oaxaca–Blinder decomposition, though he has only used a single non-experimental comparison dataset and a single selection of control variables, and he has compared his result to a relatively small number of alternative estimates.

In this paper I provide a much broader picture of the finite-sample performance of the Oaxaca–Blinder unexplained component as an estimator of the PATT. I also use the NSW data, but I closely follow Dehejia and Wahba (1999) in their sample and variable selections, so that I can reassess their influential claim that methods based on the propensity score compare favourably with other estimators. When overlap is imposed, the Oaxaca–Blinder decomposition is shown to perform superior compared to any of the estimators in Dehejia and Wahba (1999) and to additional methods such as inverse probability weighting, kernel matching, matching on covariates, and bias-corrected matching. To assess the robustness of this

[1] See Blinder (1973) and Oaxaca (1973) for seminal contributions and Fortin et al. (2011) for a comprehensive survey. Over the last two decades, the decomposition framework has also been extended to distributional statistics other than the mean (see, e.g., Juhn, Murphy, and Pierce 1993; DiNardo, Fortin, and Lemieux 1996; Melly 2005).

[2] These data were analysed originally by LaLonde (1986) and subsequently by Heckman and Hotz (1989), Dehejia and Wahba (1999, 2002), Smith and Todd (2001, 2005), Becker and Ichino (2002), Angrist and Pischke (2009), Porro and Iacus (2009), Abadie and Imbens (2011), Diamond and Sekhon (2013), and others.

result, I consider alternative sample and variable selections, and present an "empirical Monte Carlo study" (Huber, Lechner, and Wunsch 2013) which is also based on the NSW data.[3] Generally, the Oaxaca–Blinder decomposition always performs very well, and never significantly worse than any other method. At first, this might be seen as surprising, given the simplicity of this estimator. Note, however, that at least two recent papers, Khwaja et al. (2011) and Huber et al. (2013), have presented simulation studies which are suggestive of very good finite-sample performance of flexible OLS.[4] In both cases the authors have actually applied an estimator which is either equivalent or very similar to Oaxaca–Blinder, although have referred to this method in a different way.[5] In this paper I complement these previous analyses by exploring the connection with the decomposition literature, and focus on the NSW data.

2 The Treatment Effects Framework

Consider a population of N individuals, indexed by $i = 1, \ldots, N$, who are divided into two disjoint groups, 0 and 1.[6] Individuals in group 1 are exposed to regime that is called *treatment*, while individuals in group 0 are exposed to regime that is called *control*. To indicate group membership, a binary variable W_i is used, and $W_i = 0$ ($W_i = 1$) if individual i belongs to group 0 (group 1). A row vector of covariates, X_i, is also observed for each i. Moreover, it is assumed there exist two potential outcomes for each individual i, the treated outcome $Y_i(1)$ and the nontreated outcome $Y_i(0)$. It is the group membership of each individual i which causes one of the potential outcomes to become observable and the other potential outcome to become counterfactual. The realised outcome is denoted by Y_i. Consequently, $Y_i = Y_i(W_i) = Y_i(0)(1 - W_i) + Y_i(1)W_i$.

The main interest in the treatment effects framework lies in determining causal effects of treatment. Such an effect, for each individual i, is defined as the difference between her treated

[3] Since Advani and Słoczyński (2013) have recently demonstrated that the internal validity of empirical Monte Carlo studies might be quite low, this simulation study is only intended to provide a comparison with the previous literature. The choice of simulation design is quite limited anyway, as it is widely accepted that stylised Monte Carlo studies do not have much external validity (Busso, DiNardo, and McCrary 2013; Huber et al. 2013).

[4] A related point has also been made by Kang and Schafer (2007) in the context of incomplete-data estimation.

[5] Generally, various versions of the Oaxaca–Blinder decomposition are equivalent to various versions of flexible OLS in Imbens and Wooldridge (2009). See also Słoczyński (2013) for a discussion.

[6] The exposition here is standard and borrows notation from Imbens and Wooldridge (2009). Other surveys of the treatment effects literature include Cobb-Clark and Crossley (2003) and Angrist and Pischke (2009).

and her nontreated outcome, $Y_i(1) - Y_i(0)$. In general, such treatment effects are averaged over various (sub)populations of interest. The average over the subpopulation of treated individuals is called the population average treatment effect on the treated (PATT):

$$\tau_{PATT} = \mathrm{E}[Y_i(1) - Y_i(0) \mid W_i = 1]. \tag{1}$$

Alternatively, one may wish to average individual treatment effects over the whole population to obtain the population average treatment effect (PATE):

$$\tau_{PATE} = \mathrm{E}[Y_i(1) - Y_i(0)]. \tag{2}$$

There are generally two main strands in the treatment effects literature, often referred to as selection on observables and selection on unobservables, and this division is based on assumptions which are used to identify various treatment effects. This paper – and all the analyses of the NSW data in general – is only concerned with selection on observables, a strand whose main assumptions are typically referred to as unconfoundedness and overlap.[7] Under unconfoundedness, it is assumed there do not exist any unobserved variables which would be associated both with the potential outcomes and the treatment status. Consequently:

$$W_i \perp (Y_i(0), Y_i(1) \mid X_i). \tag{3}$$

Under overlap, on the other hand, it is assumed there do not exist such (sets of) values of the control variables which would perfectly predict either of the treatment statuses:

$$0 < \mathrm{pr}(W_i = 1 \mid X_i = x) < 1, \text{for all } x. \tag{4}$$

Under the assumptions of unconfoundedness and overlap both the PATT and the PATE are identified (see Imbens and Wooldridge 2009),[8] and can be estimated using a large number of

[7] As discussed by Smith and Todd (2005), however, the assumption of unconfoundedness is unlikely to hold in the NSW data. For example, NSW participants were generally placed in different local labour markets than comparison group members. Also, the set of observed control variables is relatively poor. Nevertheless, previous studies of the NSW data were implicitly based on unconfoundedness, and this paper follows in this tradition.

[8] In order to identify the PATT, only the second inequality in (4) is required.

alternative estimators. Like previous studies of the NSW data, this paper investigates the finite-sample performance of various estimators of the PATT.

3 Estimators

A recent survey of the alternative estimators of average treatment effects has been provided by Imbens and Wooldridge (2009). Several contributions have also noted (Barsky et al. 2002; Black et al. 2006; Melly 2006; Fortin et al. 2011) that the PATT can be estimated using the Oaxaca–Blinder decomposition.[9] Precisely, let the model for outcomes be linear and let the regression coefficients be flexible, i.e. different for the treated and the nontreated individuals:

$$Y_i = X_i\beta_1 + \upsilon_{1i} \quad \text{if} \quad W_i = 1 \qquad \text{and} \qquad Y_i = X_i\beta_0 + \upsilon_{0i} \quad \text{if} \quad W_i = 0, \tag{5}$$

where $\mathrm{E}[\upsilon_{1i} \mid X_i] = \mathrm{E}[\upsilon_{0i} \mid X_i] = 0$. What follows:

$$\begin{aligned}
&\mathrm{E}[Y_i \mid W_i = 1] - \mathrm{E}[Y_i \mid W_i = 0] = \\
&= \mathrm{E}[X_i \mid W_i = 1] \cdot \beta_1 - \mathrm{E}[X_i \mid W_i = 0] \cdot \beta_0 \\
&= \mathrm{E}[X_i \mid W_i = 1] \cdot (\beta_1 - \beta_0) + (\mathrm{E}[X_i \mid W_i = 1] - \mathrm{E}[X_i \mid W_i = 0]) \cdot \beta_0 \\
&= \mathrm{E}[Y_i(1) - Y_i(0) \mid W_i = 1] + (\mathrm{E}[Y_i(0) \mid W_i = 1] - \mathrm{E}[Y_i(0) \mid W_i = 0]) \\
&= \tau_{PATT} + (\mathrm{E}[Y_i(0) \mid W_i = 1] - \mathrm{E}[Y_i(0) \mid W_i = 0]).
\end{aligned} \tag{6}$$

In other words, any intergroup differential in outcomes can be decomposed into the net effect of treatment (the PATT) and a component attributable to differences in group composition (selection bias). These two components have typically been referred to as the unexplained component and the explained component, respectively, and the former has often been interpreted as "discrimination" in studies of intergroup wage differentials. Such an estimator of the PATT can be applied either as the distance between the two estimated regression functions which is evaluated at the mean values of control variables in the treated subsample or, as noted by Słoczyński

[9] Also, the PATE can be estimated using a version of the so-called "generalised Oaxaca–Blinder decomposition" which has been proposed by Słoczyński (2013).

(2013), as the coefficient on W_i in the regression of Y_i on 1, W_i, X_i, and $W_i \cdot (X_i - \overline{X}_1)$. Recently, Kline (2011) has shown that this estimator is not only consistent for the PATT, but also "doubly robust" (Robins et al. 1994), since it is equivalent to a reweighting estimator based on a linear model for the treatment odds.[10] Standard errors for various components of Oaxaca–Blinder decompositions were derived by Jann (2008).[11]

In this paper I also implement several more sophisticated methods which have received considerable attention in the treatment effects literature. I use three other reweighting (inverse probability weighting) estimators in which the nontreated subsample is reweighted with the inverse of the estimated propensity score (the conditional probability of treatment). These estimators were described in detail by Busso, DiNardo, and McCrary (2009), and referred to as IPW1, IPW2, and IPW3. In IPW1, the sum of weights is stochastic; in IPW2, it is always equal to 1; IPW3 is a linear combination of IPW1 and IPW2 which minimises the asymptotic variance of the resulting estimator (Lunceford and Davidian 2004). As shown by Hirano, Imbens, and Ridder (2003), IPW1 achieves the semiparametric efficiency bound if the propensity score is estimated with a particular series estimator. In practice, however, a logit or probit model is typically used, and inference either follows Lunceford and Davidian (2004) or relies on the bootstrap.

I also use kernel matching, and match on the estimated propensity score using both the Epanechnikov and Gaussian kernels. Large sample properties of this class of estimators were studied by Heckman, Ichimura, and Todd (1998) and kernel-based propensity score matching was shown to be inefficient. Nevertheless, these estimators are generally quite popular, and standard errors are usually bootstrapped.[12]

Another popular estimator is nearest-neighbour (NN) matching which has been studied extensively by Abadie and Imbens (2006, 2008, 2011). NN matching was shown not to be $\sqrt{n}$-

[10] This reweighting interpretation of Oaxaca–Blinder only requires the overlap assumption in its weaker form. "Double robustness" guarantees that estimation is consistent if either the model for each of the potential outcomes or the model for the treatment odds is linear. As explained by Kline (2011), this latter functional form arises naturally whenever the (treatment) assignment error is log-logistic.

[11] I am not aware of any papers which would study specification choice for Oaxaca–Blinder decompositions. Still, since Oaxaca–Blinder is essentially equivalent to a linear regression with a full set of interactions between the treatment and control variables, applied researchers might find it less important to include further interactions. Higher-order terms of certain continuous variables might still be useful, though.

[12] Kernel matching also requires the choice of bandwidth, and I rely on leave-one-out cross-validation (see, e.g., Busso et al. 2009) using a relatively sparse grid of 0.005×1.8^g for $g = 0, 1, \ldots, 5$.

consistent in general, and not to attain the semiparametric efficiency bound in settings where it attains $\sqrt{n}$-consistency (Abadie and Imbens 2006). Therefore, I use both the standard and the bias-adjusted variant of matching (Abadie and Imbens 2011), and match both on covariates and on the estimated propensity score, using 1 and 4 matches. It is important to note that the bootstrap is not valid for matching estimators (Abadie and Imbens 2008), and inference should be based on the analytic standard errors in Abadie and Imbens (2006).

Moreover, I use stratification on the estimated propensity score as well as a combination of stratification and within-strata regression adjustment. As recommended by Rosenbaum and Rubin (1984), I divide all observations into five strata using the quintiles of the distribution of the estimated propensity score. Then, I either compare mean outcomes of the treated individuals and the nontreated individuals within each stratum or estimate within-strata average treatment effects using linear regression, and average across all strata. In both cases inference should be based on a simple formula in Imbens and Wooldridge (2009).

As a comparison with the previously discussed methods, I also use linear regression (pooled OLS). Of course, this method is similar to the Oaxaca–Blinder decomposition, although it restricts the regression coefficients to be equal for the treated and the nontreated individuals; it is also implicitly based on the assumption of homogeneous treatment effects.

All these estimators are applied in four variants, as I use them both on the full sample and on samples which are restricted in order to improve overlap. Since a weaker version of (4) is required for identification, I discard all the treated individuals whose estimated propensity score is less than the minimum or greater than the maximum estimated propensity score for the nontreated individuals (Rule 1). This rule guarantees that treatment effects are not estimated for those treated individuals for whom no similar counterparts can be found in the nontreated subsample. Following Dehejia and Wahba (1999), I also use an alternative rule, and discard all the nontreated individuals whose estimated propensity score is less than the minimum or greater than the maximum estimated propensity score for the treated individuals (Rule 2). There is a subtle difference between these two rules, as in the latter case I still estimate treatment effects for all the treated individuals, but it is guaranteed that none of the dissimilar nontreated individuals is used to calculate the counterfactual outcome for the treated. Finally, I use a rule of

thumb which has recently been derived by Crump et al. (2009). These authors have developed a systematic approach to select subsamples which diminish sensitivity to the choice of specification, and concluded that the optimal rule can typically be approximated by discarding all the individuals whose estimated propensity score is less than 0.1 or greater than 0.9 (Rule 3). It is important to note that this rule is not designed to remove biases in estimation of average treatment effects; still, it has been used to reduce bias by Angrist and Pischke (2009), so it may be worthwhile to examine whether it is successful in general. Also, note that Rules 1 and 3 implicitly change the estimand. The new estimand is also an average treatment effect on the treated, but only averaged for individuals with appropriate values of the estimated propensity score. In all cases, however, I define biases relative to the "true" PATT, as this estimand seems to be more interesting in applications.

4 An Application of the Oaxaca–Blinder Unexplained Component to the NSW Data

4.1 The National Supported Work (NSW) data

The National Supported Work (NSW) Demonstration was a U.S. employment programme implemented in the mid-1970s to provide work experience to disadvantaged workers. Unlike many similar programmes, the NSW assigned treatment (participation) on random, so the pool of potential participants was exogenously divided into an experimental and a control group, thus allowing for a straightforward, unbiased estimation of average treatment effects (see LaLonde 1986 and Smith and Todd 2005 for detailed descriptions of the NSW).

In an influential paper, LaLonde (1986) examined the finite-sample performance of various non-experimental estimators in a novel way. He discarded the original control group from the NSW data, and created six alternative non-experimental comparison datasets using standard surveys of the U.S. population, the Panel Study of Income Dynamics (PSID) and the Current Population Survey (CPS). His approach was based on a conjecture that a reasonable non-experimental estimator should be able to closely replicate the experimental estimate of

Table 1: Sample means of outcome and control variables for the NSW and comparison datasets

	DW (1999)		ST (2005)							
	Treated	Control	Treated	Control	PSID-1	PSID-2	PSID-3	CPS-1	CPS-2	CPS-3
Number of observations	185	260	108	142	2,490	253	128	15,992	2,369	429
Outcome variable										
Earnings '78	6,349	4,555	7,357	4,609	21,554	9,996	5,279	14,847	10,171	6,984
	(7,867)	(5,484)	(9,027)	(6,032)	(15,555)	(11,184)	(7,763)	(9,647)	(8,852)	(7,294)
Control variables										
Age	25.82	25.05	25.37	26.01	34.85	36.09	38.26	33.23	28.25	28.03
	(7.16)	(7.06)	(6.25)	(7.11)	(10.44)	(12.08)	(12.89)	(11.05)	(11.70)	(10.79)
Education	10.35	10.09	10.49	10.27	12.12	10.77	10.30	12.03	11.24	10.24
	(2.01)	(1.61)	(1.64)	(1.57)	(3.08)	(3.18)	(3.18)	(2.87)	(2.58)	(2.86)
No degree	0.71	0.83	0.71	0.80	0.31	0.49	0.51	0.30	0.45	0.60
	(0.46)	(0.37)	(0.45)	(0.40)	(0.46)	(0.50)	(0.50)	(0.46)	(0.50)	(0.49)
Black	0.84	0.83	0.82	0.82	0.25	0.39	0.45	0.07	0.11	0.20
	(0.36)	(0.38)	(0.38)	(0.39)	(0.43)	(0.49)	(0.50)	(0.26)	(0.32)	(0.40)
Hispanic	0.06	0.11	0.07	0.11	0.03	0.07	0.12	0.07	0.08	0.14
	(0.24)	(0.31)	(0.26)	(0.32)	(0.18)	(0.25)	(0.32)	(0.26)	(0.28)	(0.35)
Married	0.19	0.15	0.20	0.19	0.87	0.74	0.70	0.71	0.46	0.51
	(0.39)	(0.36)	(0.40)	(0.39)	(0.34)	(0.44)	(0.46)	(0.45)	(0.50)	(0.50)
"Earnings '74"	2,096	2,107	3,590	3,858	19,429	11,027	5,567	14,017	8,728	5,619
	(4,887)	(5,688)	(5,971)	(7,254)	(13,407)	(10,815)	(7,255)	(9,570)	(8,968)	(6,789)
"Nonemployed '74"	0.71	0.75	0.50	0.54	0.09	0.23	0.41	0.12	0.21	0.26
	(0.46)	(0.43)	(0.50)	(0.50)	(0.28)	(0.42)	(0.49)	(0.32)	(0.41)	(0.44)
Earnings '75	1,532	1,267	2,596	2,277	19,063	7,569	2,611	13,651	7,397	2,466
	(3,219)	(3,103)	(3,872)	(3,919)	(13,597)	(9,042)	(5,572)	(9,270)	(8,112)	(3,292)
Nonemployed '75	0.60	0.68	0.32	0.47	0.10	0.34	0.61	0.11	0.18	0.31
	(0.49)	(0.47)	(0.47)	(0.50)	(0.30)	(0.47)	(0.49)	(0.31)	(0.38)	(0.46)

NOTE: Standard deviations are in parentheses. Earnings are in 1982 dollars. Education = number of years of schooling; No degree = 1 if no high school degree, 0 otherwise. DW (1999) and ST (2005) refer to subsets of the NSW dataset which were created by Dehejia and Wahba (1999) and Smith and Todd (2005), respectively.

the average treatment effect, while using only the treated subsample and a non-experimental comparison group. LaLonde (1986) concluded that non-experimental estimators were typically unable to replicate the experimental results, and his findings were instrumental in popularising experimental and quasi-experimental designs in labour economics.

Following LaLonde (1986), the NSW data were analysed by many researchers, including Heckman and Hotz (1989), Dehejia and Wahba (1999, 2002), Smith and Todd (2001, 2005), Becker and Ichino (2002), Angrist and Pischke (2009), Porro and Iacus (2009), Abadie and Imbens (2011), Kline (2011), and Diamond and Sekhon (2013). In an influential contribution, Dehejia and Wahba (1999) closely replicated the experimental estimate of the average treatment effect using various methods based on the propensity score.

In this paper I use a version of the NSW data which was created by Dehejia and Wahba (1999), and supplement it with the "early RA" sample from Smith and Todd (2005). These

latter data are generally preferable to those from Dehejia and Wahba (1999), since Dehejia and Wahba (1999) controversially included only those individuals randomised after April 1976 who were not employed in months 13–24 before random assignment. Table 1 presents descriptive statistics for all the subsamples used in the analysis, including the PSID and CPS comparison datasets.[13] There are substantial disparities in means of control and outcome variables between the NSW experimental and control groups and the PSID and CPS comparison groups. It is precisely these disparities that hinder non-experimental replication of the experimental estimate of the average treatment effect. This estimate is equal to $1,794 for Dehejia and Wahba (1999) and $2,748 for Smith and Todd (2005).

4.2 A reanalysis of Dehejia and Wahba (1999)

In this subsection I closely follow Dehejia and Wahba (1999) in their sample and variable selections, so that I can reassess their claim that methods based on the propensity score compare favourably with other estimators. Dehejia and Wahba (1999) used all the six non-experimental comparison datasets (PSID1–3 and CPS1–3), and descriptive statistics in Table 1 in this paper are nearly identical to the values reported in Table 1 in Dehejia and Wahba (1999) and Table 1 in Smith and Todd (2005).[14] In their analysis, Dehejia and Wahba (1999) applied three different selections of control variables, each of them matched to one, two or three non-experimental comparison datasets.[15] As explained by the authors, their variable selections were based on balancing tests, i.e. a specification was accepted whenever the null that all control variables are balanced within each stratum could not be rejected. To make the subsequent estimates of the

[13] As described in LaLonde (1986), PSID-1 includes all men in the original PSID data, except those who were older than 55 or classified as retired; PSID-2 is a subset of PSID-1 which includes those men who were not employed in the spring of 1976; PSID-3 is a subset of PSID-2 which includes those men who were not employed in the spring of 1975. Similarly, CPS-1 includes all men in the original CPS data, except those who were older than 55; CPS-2 is a subset of CPS-1 which includes those men who were not employed in March 1976; CPS-3 is a subset of CPS-2 which includes those men whose income in 1975 was lower than the poverty level.

[14] Unfortunately, this is not the case with LaLonde (1986) whose CPS-2 and CPS-3 subsamples could not be recreated by Dehejia and Wahba (1999). Table 1 in this paper closely replicates, however, descriptive statistics for PSID-1, PSID-2, PSID-3, and CPS-1 which were reported in Table 3 in LaLonde (1986).

[15] For PSID-1, Dehejia and Wahba (1999) selected Age, Age squared, Education, Education squared, Married, No degree, Black, Hispanic, "Earnings '74", "Earnings '74" squared, Earnings '75, Earnings '75 squared, and the product of Black and "Nonemployed '74". For PSID-2 and PSID-3, they also included "Nonemployed '74" and Nonemployed '75, but excluded the product of Black and "Nonemployed '74". For CPS-1, CPS-2, and CPS-3 – as compared with the latter variable selection – they also included Age cubed and the product of Education and "Earnings '74", but on the other hand excluded both "Earnings '74" squared and Earnings '75 squared.

PATT fully comparable with the results reported by Dehejia and Wahba (1999), I apply exactly the same sets of control variables throughout this subsection.[16]

Table A.1 presents mean biases, root mean square errors (RMSEs), and standard deviations (SDs) for a large number of non-experimental estimators which utilise sample and variable selections from Dehejia and Wahba (1999). RMSEs are calculated as:

$$\text{RMSE} = \sqrt{\frac{\sum_{j \in \jmath} (\hat{\tau}_j - \hat{\tau}_{exp})^2}{6}}, \tag{7}$$

where $\jmath$ is a set of comparison datasets and $\hat{\tau}_{exp}$ is the benchmark estimate. Mean biases are calculated analogously. Similar to Becker and Ichino (2002), I have been unable to replicate most of the results in Dehejia and Wahba (1999), so the upper panel of Table A.1 reports values which can be calculated using the estimates in Table 3 in Dehejia and Wahba (1999).[17]

Among new results in Table A.1, Oaxaca–Blinder performs remarkably well. Whenever overlap is improved (Rules 1–3), the Oaxaca–Blinder decomposition performs best in terms of RMSE and very well in terms of mean bias. When overlap is not improved (Full sample), Oaxaca–Blinder is still classified as the third best estimator, both in terms of RMSE and mean bias. Also, for Rules 1 and 2 Oaxaca–Blinder performs better in terms of RMSE than any of the estimators in Dehejia and Wahba (1999); although Oaxaca–Blinder is slightly more biased than the stratification-based estimators in Dehejia and Wahba (1999), it has very small variance, and performs therefore particularly well on RMSE. Still, when I test the statistical significance of the differences between the smallest RMSE (Oaxaca–Blinder, Rule 1) and all other RMSEs, I often cannot reject the null. Especially, Oaxaca–Blinder seems to be only insignificantly better than IPW, kernel matching with the Epanechnikov kernel, some variants of NN matching on

[16] I perform all calculations in Stata and apply the following user-written commands: `nnmatch` (Abadie et al. 2004), `oaxaca` (Jann 2008), and `psmatch2` (Leuven and Sianesi 2003).

[17] It is generally impossible to replicate the results in Dehejia and Wahba (1999) for stratification-based estimators, since the authors did not report the number of strata and their boundaries. Their regression estimates (column 2, Table 3) can be replicated, although the authors reported their variable selection incorrectly; these estimates require including Earnings '75 squared in the reported specification. Using variable selections reported in Dehejia and Wahba (1999), I also obtain very different estimates for NN matching on the propensity score. For PSID-1, I get 560 instead of 1,691; for PSID-2, 871 instead of 1,455; for PSID-3, 1,522 instead of 2,120; for CPS-1, 730 instead of 1,582; for CPS-2, 1,399 instead of 1,788; for CPS-3, –662 instead of 587. At the same time, I have been able to replicate the original estimates for PSID-2 and PSID-3, and this requires excluding No degree from the reported specification, as the authors – again – reported their variable selection incorrectly. Therefore, in general, I might not be applying specifications which were *used* by Dehejia and Wahba (1999), even though I definitely apply their *reported* specifications.

the propensity score, and stratification with regression adjustment.

While improving overlap using Rules 1 and 2 does not seem, on average, to make much difference,[18] Rule 3 (Crump et al. 2009) has a clear negative effect on the performance of the estimators, and it increases both their bias and variance. Intuitively, if treatment effects are heterogeneous, then removal of a large fraction of treated individuals (28–52%) will typically bias the resulting estimate of the PATT. Clearly, this rule has not been designed to reduce biases when estimating average treatment effects, and one should generally acknowledge that its application changes the estimand. Still, it has been used to reduce bias by Angrist and Pischke (2009), and this has warranted an examination of its performance.

4.3 Robustness checks

To assess the robustness of the very good performance of Oaxaca–Blinder, in this subsection I consider alternative sample and variable selections. First, I continue using the Dehejia and Wahba (1999) version of the NSW data, but change the variable selection, and utilise a specification from a recent paper by Abadie and Imbens (2011).[19] These results are presented in Table A.2. Second, I use the "early RA" sample from Smith and Todd (2005), but maintain the variable selection from the previous subsection. These results are presented in Table A.3.

Under the new variable selection (Table A.2), biases and variances of the estimators are generally higher. Oaxaca–Blinder continues, however, to perform very well. In terms of RMSE, it is only outperformed by inverse probability weighting, but this difference is not significant. In terms of mean bias, Oaxaca–Blinder performs relatively worse, although it continues to be one of the best-performing estimators. Stratification and NN matching with a small number of neighbours ($k = 1$) generally perform significantly worse than IPW. Rule 3 (Crump et al. 2009) continues to increase both bias and variance of the estimators.

As reported by Smith and Todd (2005), it is very difficult to replicate the experimental benchmark using their "early RA" sample, and this is evident in Table A.3 where biases and variances are again much higher. Still, Oaxaca–Blinder with no overlap improvement performs

[18] If anything, Rule 1 (Rule 2) seems to be slightly unsuccessful (successful) in improving the finite-sample performance of the estimators.

[19] This selection of control variables is identical for all the comparison datasets, and it includes Age, Education, Married, Black, Hispanic, "Earnings '74", Earnings '75, "Nonemployed '74", and Nonemployed '75.

best in terms of RMSE among all the estimators, and it also performs very well – especially in terms of RMSE, but also in terms of mean bias – within each class of overlap improvement rules. Many of these differences in RMSEs are again not significant, but Oaxaca–Blinder seems to consistently outperform regression, stratification, and several variants of NN matching. Rules 1 and 3 increase bias and variance of the estimators.

4.4 An empirical Monte Carlo study

In this subsection I provide a further robustness check, and present an "empirical Monte Carlo study" which is also based on the NSW data. It is a difficult decision to choose an appropriate design for a simulation study, since it is now widely accepted that traditional ("stylised") Monte Carlos do not have much external validity (Busso et al. 2013; Huber et al. 2013) and a recent contribution has questioned the internal validity of empirical Monte Carlo studies, i.e. their ability to replicate the true ranking of estimators for a given dataset (Advani and Słoczyński 2013). This robustness check is therefore primarily intended to provide a comparison with the recent literature.

The design of this simulation exercise follows a recent paper by Huber et al. (2013). In the first step, I estimate a logit model for the propensity score using the Dehejia and Wahba (1999) subset of the treated subsample and the CPS-1 comparison dataset. My variable selection follows Abadie and Imbens (2011). I calculate the linear prediction from this model for each individual in the nontreated subsample ($X_i\hat{\beta}$), and discard all the treated. Next, in each replication I draw a sample of size N from the remaining data (with replacement). For each unit in this sample, I then draw an iid logistic error, ϵ_i, and assign the status of "placebo treated" using $W_i = \mathbf{1}(W_i^* > 0)$ where $W_i^* = \hat{\alpha} + X_i\hat{\beta} + \epsilon_i$ and $\hat{\alpha}$ is a constant which is chosen to ensure that the proportion of "placebo treated" is equal to the desired value. Clearly, such a simulation design guarantees that the true effect of treatment is always zero by construction, and does not rely therefore on artificial data-generating processes.

To shed some light on the data features which codetermine the relative performance of the Oaxaca–Blinder decomposition, I vary N and $\hat{\alpha}$, and run four simulation exercises in total: *(i)* with $N = 300$ and $\text{pr}(W_i = 1) = 0.5$, *(ii)* with $N = 1,200$ and $\text{pr}(W_i = 1) = 0.1$, *(iii)*

Table 2: Regression analysis of the Monte Carlo results: The dependent variable is the root mean square error of an estimator

	Model 1		Model 2		Model 3	
	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
Constant	1,947***	(255)	1,947***	(254)	1,967***	(267)
Small dataset ($N = 300$)	1,151***	(161)	1,151***	(163)	1,170***	(175)
Small pr. of treatment ($p = 10\%$)	–127	(161)	–127	(161)	–118	(173)
Large pr. of treatment ($p = 90\%$)	671***	(186)	671***	(186)	667***	(201)
Improving overlap: Rule 1	–1,121***	(178)	–1,121***	(178)	–1,134***	(192)
Improving overlap: Rule 2	8	(126)	8	(131)	–8	(135)
Improving overlap: Rule 3	–1,811***	(181)	–1,811***	(181)	–1,884***	(193)
Oaxaca–Blinder	138	(271)	138	(268)	–146	(335)
Stratification	72	(296)	72	(293)	72	(308)
IPW1	4,962***	(735)	4,962***	(728)	4,962***	(742)
IPW2	1,143***	(280)	1,143***	(277)	1,143***	(288)
IPW3	845***	(260)	845***	(258)	845***	(269)
Kernel matching, Epanechnikov	902***	(314)	902***	(312)	889***	(332)
Kernel matching, Gaussian	810**	(321)	810**	(319)	797**	(335)
NN matching on covariates, $k = 1$	681**	(272)			681**	(282)
NN matching on covariates, $k = 1$ (bias-adj.)	974***	(274)			974***	(283)
NN matching on the score, $k = 1$	1,496***	(301)			1,496***	(309)
NN matching on the score, $k = 1$ (bias-adj.)	1,126***	(279)			1,126***	(288)
NN matching on covariates, $k = 4$	850***	(282)			850***	(291)
NN matching on covariates, $k = 4$ (bias-adj.)	656**	(265)			656**	(274)
NN matching on the score, $k = 4$	691***	(266)			691**	(276)
NN matching on the score, $k = 4$ (bias-adj.)	988***	(287)			988***	(296)
NN matching			923***	(256)		
NN matching on the score			285**	(116)		
NN matching, $k = 4$			–273**	(116)		
NN matching (bias-adj.)			7	(116)		
Oaxaca–Blinder × Small dataset ($N = 300$)					–266	(189)
Oaxaca–Blinder × Small pr. of treatment ($p = 10\%$)					–118	(313)
Oaxaca–Blinder × Large pr. of treatment ($p = 90\%$)					49	(372)
Oaxaca–Blinder × Improving overlap: Rule 1					180	(336)
Oaxaca–Blinder × Improving overlap: Rule 2					235	(314)
Oaxaca–Blinder × Improving overlap: Rule 3					1,059***	(406)
Observations	232		232		232	
R^2	0.721		0.715		0.725	

NOTE: The estimation sample consists of the results of all Monte Carlos. All coefficients are expressed in 1982 dollars. Robust standard errors are in parentheses. *Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

with $N = 1,200$ and $\text{pr}(W_i = 1) = 0.5$, and *(iv)* with $N = 1,200$ and $\text{pr}(W_i = 1) = 0.9$.[20] Similar to Huber et al. (2013), I use 16,000 replications for $N = 300$ and 4,000 replications for $N = 1,200$. Also, I follow Huber et al. (2013) in summarising the results of these simulations using regression analysis, i.e. root mean square errors of the estimators are regressed on binary variables which represent these estimators as well as data features, overlap improvement rules,

[20]These combinations of N and $\hat{\alpha}$ follow Huber et al. (2013) who have also considered a larger sample of $N = 4,800$.

and selected interactions. These results are presented in Table 2.[21]

Stratification with regression adjustment (omitted category) performs best in terms of RMSE, and there are only two estimators which do not perform significantly worse: stratification and Oaxaca–Blinder.[22] IPW1 (unnormalised reweighting) and NN matching on covariates with a small number of matches perform particularly badly. On the other hand, matching on covariates is generally better than matching on the propensity score (Model 2); also, if one uses NN matching, then it seems to make sense to choose a larger number of matches, while bias adjustment does not make much difference. Intuitively, RMSEs are larger for small datasets and whenever the ratio of treated to control units is very large (9:1).

Unlike in the previous applications, Rules 1 and 3 improve the finite-sample performance of the estimators. This difference can be interpreted as an effect of the simulation design which restricts treatment effects to be homogeneous. In such a setting it might always be helpful to discard all the individuals which do not have good matches in the other subsample, as the true effect of treatment can still be estimated using the remaining data.

Also, this simulation study does not seem to have uncovered any data features which would determine the relative performance of Oaxaca–Blinder. Its relative performance improves in small datasets, but this effect is not significant. Rule 3 (Crump et al. 2009) has a relatively small effect on the performance of Oaxaca–Blinder, compared to other estimators.

5 Summary and Conclusions

In this paper I use the NSW data to examine the finite-sample performance of the Oaxaca–Blinder decomposition as an estimator of the population average treatment effect on the treated (PATT). I utilise the same sample and variable selections which were used in an influential paper by Dehejia and Wahba (1999), and conclude that Oaxaca–Blinder performs better, on average, than any of the estimators in this original paper. To assess the robustness of this result,

[21] Because of computational burden I exclude kernel matching from simulations with $N = 1,200$. This estimator is computationally intensive, as it requires cross-validation of the bandwidth in each replication. Also, I do not report simulation results for regression, since this method has an unfair advantage in a design which implicitly assumes that treatment effects are homogeneous. On average, regression performed best in terms of RMSE, and such a result is clearly not believable in general.

[22] Note that neither stratification nor stratification with regression adjustment has been considered by Huber et al. (2013), while Oaxaca–Blinder has been referred to in a different way.

I explore alternative variable (Abadie and Imbens 2011) and sample (Smith and Todd 2005) selections, and perform an "empirical Monte Carlo study" (Huber et al. 2013) which is also based on the NSW data. I conclude that the very good performance of Oaxaca–Blinder is indeed a robust result which holds in all these cases.

More generally, however, I do not wish to claim that this result will inevitably hold in every setting. The programme evaluation literature acknowledges that there exists no estimator which performs very well in every circumstance, and in my view rightly so. Also, although I use a dataset which has received remarkable attention in this literature, it can still be argued that it is not clear whether this result should hold for other datasets. Empirical researchers are usually advised to apply several estimators as a form of a robustness check. This paper might encourage them to consider Oaxaca–Blinder as an easily applicable counterpart of more sophisticated semiparametric and nonparametric methods.

Table A.1: A comparison of Dehejia and Wahba (1999) with other estimates of the PATT using Dehejia and Wahba (1999) dataset and variable selections

Improving overlap?	Full sample			Rule 1			Rule 2			Rule 3		
	Mean bias	RMSE	SD	Mean bias	RMSE	SD	Mean bias	RMSE	SD	Mean bias	RMSE	SD
Dehejia and Wahba (1999):												
Regression on a quadratic in the score	–	–	–	–	–	–	–1,191	1,218	253	–	–	–
Stratification	–	–	–	–	–	–	–18	378	378	–	–	–
Stratification and regression	–	–	–	–	–	–	75	289	279	–	–	–
NN matching on the score, $k = 1$	–	–	–	–	–	–	–257	538	472	–	–	–
NN matching on the score and regression, $k = 1$	–	–	–	–	–	–	–403	521	329	–	–	–
New estimates:												
Regression	–921	1,008*	408	–852	949*	418	–1,127	1,231*	495	–1,742	1,983***	948
Oaxaca–Blinder	91	414	403	–97	<u>211</u>	188	–130	282	250	–1,301	1,640	999
Stratification	–1,897	2,479***	1,596	–2,170	2,462***	1,164	–1,228	1,670**	1,132	–1,838	2,252***	1,302
Stratification and regression	–880	1,039	553	–973	1,122	559	–775	1,316	1,063	–1,919	2,611	1,772
IPW1	–556	765	526	–1,055	1,163	491	–475	720	542	–1,194	1,727**	1,248
IPW2	215	623	585	193	615	584	255	635	581	–1,813	2,426***	1,612
IPW3	–34	324	322	–244	444	370	20	332	332	–1,868	2,511***	1,679
Kernel matching, Epanechnikov	–545	584	209	–835	892	311	–416	489	257	–1,984	2,535***	1,579
Kernel matching, Gaussian	–898	968*	360	–1,058	1,219**	606	–598	652	260	–1,986	2,497***	1,515
NN matching on covariates, $k = 1$	–588	1,149**	988	–658	1,087**	866	–557	1,077**	922	–1,515	1,844**	1,051
NN matching on covariates, $k = 1$ (bias-adj.)	–694	1,131*	894	–667	1,132**	915	–574	1,075*	908	–1,087	1,663	1,258
NN matching on the score, $k = 1$	–1,037	1,240**	679	–1,177	1,341**	643	–1,058	1,276**	714	–2,650	3,036***	1,483
NN matching on the score, $k = 1$ (bias-adj.)	–1,066	1,409	922	–1,019	1,418	987	–1,069	1,440	964	–1,964	2,833	2,042
NN matching on covariates, $k = 4$	–567	1,049**	883	–732	981*	654	–553	995*	827	–1,484	1,825**	1,063
NN matching on covariates, $k = 4$ (bias-adj.)	–484	1,018*	895	–608	982*	771	–505	960	816	–1,382	1,745	1,065
NN matching on the score, $k = 4$	10	303	303	–335	546	431	13	309	309	–1,784	2,274**	1,411
NN matching on the score, $k = 4$ (bias-adj.)	–298	511	415	–354	566	442	–280	503	418	–1,518	2,255	1,668

NOTE: All statistics are expressed in 1982 dollars. Propensity scores are estimated using a logit model. Rules 1–3 are explained in the text. Rule 1 discards 6, 38, 31, 6, 5, and 8 treated individuals for PSID1–3 and CPS1–3, respectively. Rule 2 discards 1,344, 136, 68, 12,136, 1,182, and 108 nontreated individuals for PSID1–3 and CPS1–3, respectively. Rule 3 discards 96, 97, 97, 52, 61, and 57 treated individuals as well as 2,369, 170, 69, 15,764, 2,190, and 300 nontreated individuals for PSID1–3 and CPS1–3, respectively. Also, Rule 1 (Rule 3) changes the experimental benchmark to $1,894, $1,255, $1,090, $1,894, $1,873, and $1,863 ($703, –$18, $572, $2,363, $1,485, and $1,339) for PSID1–3 and CPS1–3, respectively. Underline denotes the smallest RMSE. Stars refer to a bootstrap test of equality between the given RMSE and the smallest RMSE (100 replications). *Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

Table A.2: A robustness check: Using an alternative variable selection (Abadie and Imbens 2011)

Improving overlap?	Full sample			Rule 1			Rule 2			Rule 3		
	Mean bias	RMSE	SD	Mean bias	RMSE	SD	Mean bias	RMSE	SD	Mean bias	RMSE	SD
Regression	–997	1,046	318	–1,001	1,059	345	–750	758	110	–1,055	1,321*	795
Oaxaca–Blinder	–476	632	414	–601	702	363	77	636	632	–684	902	588
Stratification	–1,940	2,408***	1,426	–2,134	2,488***	1,279	–1,518	1,603**	514	–1,125	1,590	1,124
Stratification and regression	–851	913	331	–880	1,056	583	–1,184	1,249	399	–1,131	1,566	1,083
IPW1	–618	706	342	–1,234	1,417	697	–556	662	359	–232	713	674
IPW2	–53	622	620	–121	538	524	–42	624	623	–1,028	1,377	917
IPW3	–234	470	407	–482	572	309	–212	<u>461</u>	409	–1,114	1,497	999
Kernel matching, Epanechnikov	–1,023	1,036	163	–1,297	1,353*	387	–1,013	1,078	370	–1,455	1,782*	1,029
Kernel matching, Gaussian	–803	917	443	–1,105	1,253*	591	–670	809	453	–1,473	1,792*	1,021
NN matching on covariates, $k = 1$	–565	1,489***	1,378	–817	1,445**	1,192	–583	1,552***	1,438	–1,069	1,637**	1,240
NN matching on covariates, $k = 1$ (bias-adj.)	–544	1,480***	1,376	–781	1,465**	1,240	–430	1,531***	1,469	–1,083	1,709**	1,322
NN matching on the score, $k = 1$	–1,224	1,473**	819	–1,749	2,283***	1,467	–1,228	1,457**	784	–2,031	2,528**	1,506
NN matching on the score, $k = 1$ (bias-adj.)	–573	1,147*	994	–946	1,342**	952	–591	1,130*	963	–1,037	1,383	916
NN matching on covariates, $k = 4$	–313	910	855	–511	833	658	–354	918	848	–878	1,100	663
NN matching on covariates, $k = 4$ (bias-adj.)	–135	786	774	–281	761	708	–186	766	744	–698	1,009	728
NN matching on the score, $k = 4$	–545	695	431	–929	1,221*	792	–550	697	428	–1,432	1,783*	1,062
NN matching on the score, $k = 4$ (bias-adj.)	–361	802	715	–489	826	666	–347	784	703	–865	1,072	633

NOTE: All statistics are expressed in 1982 dollars. Propensity scores are estimated using a logit model. Rules 1–3 are explained in the text. Rule 1 discards 3, 34, 50, 5, 0, and 5 treated individuals for PSID1–3 and CPS1–3, respectively. Rule 2 discards 1,215, 74, 51, 10,552, 860, and 56 nontreated individuals for PSID1–3 and CPS1–3, respectively. Rule 3 discards 87, 87, 91, 44, 27, and 9 treated individuals as well as 2,362, 155, 58, 15,679, 2,108, and 270 nontreated individuals for PSID1–3 and CPS1–3, respectively. Also, Rule 1 (Rule 3) changes the experimental benchmark to \$1,672, \$1,576, \$954, \$1,853, \$1,799, and \$1,830 (\$1,418, \$1,307, \$969, \$2,001, \$2,038, and \$1,783) for PSID1–3 and CPS1–3, respectively. Underline denotes the smallest RMSE. Stars refer to a bootstrap test of equality between the given RMSE and the smallest RMSE (100 replications). *Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

Table A.3: A robustness check: Using an alternative dataset selection (Smith and Todd 2005)

Improving overlap?	Full sample			Rule 1			Rule 2			Rule 3		
	Mean bias	RMSE	SD	Mean bias	RMSE	SD	Mean bias	RMSE	SD	Mean bias	RMSE	SD
Regression	–1,726	1,788*	466	–1,914	1,980**	507	–2,197	2,253**	500	–2,500	2,685**	980
Oaxaca–Blinder	–886	1,022	511	–1,551	1,596	376	–1,455	1,485	299	–1,592	1,794	827
Stratification	–2,799	3,193***	1,537	–3,444	3,646***	1,196	–1,958	2,195*	993	–1,757	2,027	1,011
Stratification and regression	–2,619	2,894	1,232	–2,454	2,660**	1,027	–1,912	2,178	1,043	–935	2,046	1,821
IPW1	–1,420	1,495	467	–2,752	3,233	1,697	–1,296	1,396	519	–1,920	2,906**	2,181
IPW2	–1,081	1,343	797	–1,493	1,844	1,081	–1,102	1,362	801	–1,668	1,967	1,042
IPW3	–1,316	1,443	592	–1,828	2,024*	870	–1,271	1,410	612	–1,545	1,908	1,119
Kernel matching, Epanechnikov	–2,185	2,589*	1,388	–2,803	3,114***	1,357	–1,477	1,582	565	–1,568	1,887	1,050
Kernel matching, Gaussian	–1,833	1,960*	694	–2,381	2,522**	832	–1,524	1,610	518	–1,757	2,052	1,059
NN matching on covariates, $k = 1$	–1,867	1,926*	472	–2,428	2,537**	736	–1,956	2,003**	432	–2,475	2,598**	789
NN matching on covariates, $k = 1$ (bias-adj.)	–1,916	2,083	819	–2,458	2,744*	1,220	–2,091	2,246	821	–2,297	2,690	1,400
NN matching on the score, $k = 1$	–2,027	2,379*	1,245	–2,925	3,411***	1,754	–2,065	2,406*	1,234	–2,298	3,009*	1,942
NN matching on the score, $k = 1$ (bias-adj.)	–1,943	2,427	1,454	–2,654	3,109	1,620	–2,374	3,511	2,586	–2,782	3,492	2,111
NN matching on covariates, $k = 4$	–1,649	1,680	321	–2,296	2,326**	372	–1,603	1,620	231	–1,943	2,048	647
NN matching on covariates, $k = 4$ (bias-adj.)	–1,677	1,731	429	–2,259	2,376**	736	–1,702	1,780	519	–1,692	1,991	1,048
NN matching on the score, $k = 4$	–1,250	1,322	429	–2,041	2,227*	890	–1,257	1,332	441	–1,794	2,116	1,122
NN matching on the score, $k = 4$ (bias-adj.)	–1,748	1,937	836	–2,088	2,219	750	–1,840	2,069	945	–1523	1,755	873

NOTE: All statistics are expressed in 1982 dollars. Propensity scores are estimated using a logit model. Rules 1–3 are explained in the text. Rule 1 discards 4, 28, 31, 0, 3, and 5 treated individuals for PSID1–3 and CPS1–3, respectively. Rule 2 discards 1,516, 147, 56, 12,718, 1,473, and 157 nontreated individuals for PSID1–3 and CPS1–3, respectively. Rule 3 discards 21, 25, 32, 29, 29, and 27 treated individuals as well as 2,380, 173, 73, 15,810, 2,233, and 326 nontreated individuals for PSID1–3 and CPS1–3, respectively. Also, Rule 1 (Rule 3) changes the experimental benchmark to \$2,801, \$1,361, \$1,661, \$2,748, \$2,293, and \$2,368 (\$2,600, \$1,375, \$1,662, \$3,376, \$2,522, and \$2,601) for PSID1–3 and CPS1–3, respectively. Underline denotes the smallest RMSE. Stars refer to a bootstrap test of equality between the given RMSE and the smallest RMSE (100 replications). *Statistically significant at the 10% level; **at the 5% level; ***at the 1% level.

References

[1] Abadie, Alberto, David Drukker, Jane Leber Herr, and Guido W. Imbens. 2004. Implementing matching estimators for average treatment effects in Stata. *Stata Journal* 4:290–311.

[2] Abadie, Alberto, and Guido W. Imbens. 2006. Large sample properties of matching estimators for average treatment effects. *Econometrica* 74:235–67.

[3] Abadie, Alberto, and Guido W. Imbens. 2008. On the failure of the bootstrap for matching estimators. *Econometrica* 76:1537–57.

[4] Abadie, Alberto, and Guido W. Imbens. 2011. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics* 29:1–11.

[5] Advani, Arun, and Tymon Słoczyński. 2013. Mostly harmless simulations? On the internal validity of empirical Monte Carlo studies. Unpublished manuscript, Department of Economics, University College London.

[6] Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly harmless econometrics: An empiricist's companion*. Princeton and Oxford: Princeton University Press.

[7] Barsky, Robert, John Bound, Kerwin Kofi Charles, and Joseph P. Lupton. 2002. Accounting for the black-white wealth gap: A nonparametric approach. *Journal of the American Statistical Association* 97:663–73.

[8] Becker, Sascha O., and Andrea Ichino. 2002. Estimation of average treatment effects based on propensity scores. *Stata Journal* 2:358–77.

[9] Black, Dan, Amelia Haviland, Seth Sanders, and Lowell Taylor. 2006. Why do minority men earn less? A study of wage differentials among the highly educated. *Review of Economics and Statistics* 88:300–13.

[10] Blinder, Alan S. 1973. Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources* 8:436–55.

[11] Busso, Matias, John DiNardo, and Justin McCrary. 2009. Finite sample properties of semiparametric estimators of average treatment effects. Unpublished manuscript, School of Law, University of California, Berkeley.

[12] Busso, Matias, John DiNardo, and Justin McCrary. 2013. New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics* (forthcoming).

[13] Cobb-Clark, Deborah A., and Thomas Crossley. 2003. Econometrics for evaluations: An introduction to recent developments. *Economic Record* 79:491–511.

[14] Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96:187–99.

[15] Dehejia, Rajeev H., and Sadek Wahba. 1999. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94:1053–62.

[16] Dehejia, Rajeev H., and Sadek Wahba. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics* 84:151–61.

[17] Diamond, Alexis, and Jasjeet S. Sekhon. 2013. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics* 95:932–45.

[18] DiNardo, John, Nicole M. Fortin, and Thomas Lemieux. 1996. Labor market institutions and the distribution of wages, 1973–1992: A semiparametric approach. *Econometrica* 64:1001–44.

[19] Fortin, Nicole, Thomas Lemieux, and Sergio Firpo. 2011. Decomposition methods in economics. In *Handbook of labor economics*, vol. 4A, ed. Orley Ashenfelter and David Card. San Diego and Amsterdam: Elsevier.

[20] Heckman, James J., and V. Joseph Hotz. 1989. Choosing among alternative non-experimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American Statistical Association* 84:862–74.

[21] Heckman, James J., Hidehiko Ichimura, and Petra Todd. 1998. Matching as an econometric evaluation estimator. *Review of Economic Studies* 65:261–94.

[22] Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71:1161–89.

[23] Huber, Martin, Michael Lechner, and Conny Wunsch. 2013. The performance of estimators based on the propensity score. *Journal of Econometrics* 175:1–21.

[24] Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature* 47:5–86.

[25] Jann, Ben. 2008. The Blinder–Oaxaca decomposition for linear regression models. *Stata Journal* 8:453–79.

[26] Juhn, Chinhui, Kevin M. Murphy, and Brooks Pierce. 1993. Wage inequality and the rise in returns to skill. *Journal of Political Economy* 101:410–42.

[27] Kang, Joseph D. Y., and Joseph L. Schafer. 2007. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* 22:523–39.

[28] Khwaja, Ahmed, Gabriel Picone, Martin Salm, and Justin G. Trogdon. 2011. A comparison of treatment effects estimators using a structural model of AMI treatment choices and severity of illness information from hospital charts. *Journal of Applied Econometrics* 26:825–53.

[29] Kline, Patrick. 2011. Oaxaca–Blinder as a reweighting estimator. *American Economic Review: Papers & Proceedings* 101:532–37.

[30] LaLonde, Robert J. 1986. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review* 76:604–20.

[31] Leuven, Edwin, and Barbara Sianesi. 2003. PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. URL http://ideas.repec.org/c/boc/bocode/s432001.html. This version 4.0.6.

[32] Lunceford, Jared K., and Marie Davidian. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statistics in Medicine* 23:2937–60.

[33] Melly, Blaise. 2005. Decomposition of differences in distribution using quantile regression. *Labour Economics* 12:577–90.

[34] Melly, Blaise. 2006. Applied quantile regression. PhD diss., University of St. Gallen.

[35] Oaxaca, Ronald. 1973. Male-female wage differentials in urban labor markets. *International Economic Review* 14:693–709.

[36] Porro, Giuseppe, and Stefano Maria Iacus. 2009. Random recursive partitioning: A matching method for the estimation of the average treatment effect. *Journal of Applied Econometrics* 24:163–85.

[37] Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1994. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89:846–66.

[38] Rosenbaum, Paul R., and Donald B. Rubin. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79:516–24.

[39] Słoczyński, Tymon. 2013. Population average gender effects. IZA Discussion Paper no. 7315, Institute for the Study of Labor, Bonn.

[40] Smith, Jeffrey A., and Petra E. Todd. 2001. Reconciling conflicting evidence on the performance of propensity-score matching methods. *American Economic Review: Papers & Proceedings* 91:112–18.

[41] Smith, Jeffrey A., and Petra E. Todd. 2005. Does matching overcome LaLonde's critique of non-experimental estimators? *Journal of Econometrics* 125:305–53.