Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Econometrics 125 (2005) 355–364
JOURNAL OF Econometrics
www.elsevier.com/locate/econbase

Practical propensity score matching: a reply to Smith and Todd

Rajeev Dehejia[a,b,*]

[a] *Department of Economics and SIPA, Columbia University, 420W.118th street, Room 1022, New York, NY10027, USA*
[b] *NBER, 1050 Massachusetts Avenue, Cambridge, MA 02138, USAAvailable online 15 June 2004*

Abstract

This paper discusses propensity score matching in the context of Smith and Todd's (Does matching overcome Lalonde's critique of nonexperimental estimators, J. Econom., in press) reanalysis of Dehejia and Wahba (J. Am. Statist. Assoc. 97 (1999) 1053; National Bereau of Economics Research working Paper No. 6829, Rev. Econom. Statist., 2002, forthcoming). Propensity score methods require that a separate propensity score specification be estimated for each treatment group-comparison group combination. Furthermore, a researcher should always examine the sensitivity of the estimated treatment effect to small changes in the propensity score specification; this is a useful diagnostic on the quality of the comparison group. When these are borne in mind, propensity score methods are useful in analyzing all of the subsamples of the NSW data considered in Smith and Todd (Does matching overcome Lalonde's critique of nonexperimental estimators, J. Econom., in press).

Keywords: Causal inference; Non-experimental methods; Program evaluation; Labor training

1. Introduction

This paper discusses propensity score matching in the context of Smith and Todd's (2004) reanalysis of Dehejia and Wahba (1999, 2002). Smith and Todd's paper makes some useful contributions to the literature on propensity score matching.[1] In

*Corresponding author. Department of Economics and SIPA, Columbia University, 420 W. 118th Street, Room 1022, New York, NY 10027, USA. Fax: +1-212-854-8059.
E-mail address: rd247@columbia.edu (R. Dehejia).

[1] The propensity score was first introduced by Rosenbaum and Rubin (1983). See also Rosenbaum and Rubin (1984, 1985), Rubin and Thomas (1992a, b, 1996), Heckman et al. (1997, 1998a, b), and Dehejia and Wahba (1999, 2002). On (non-propensity score) matching methods see Rubin (1973, 1976a, b, 1979) and Abadie and Imbens (2001).

0304-4076/$ - see front matter
doi:10.1016/j.jeconom.2004.04.012

particular, their application of difference-in-differences propensity score matching illustrates an interesting technique. However, their paper also illustrates some of the mistakes that are often made when applying propensity score methods. In this paper, I will address three of these issues.

First, I draw attention to some elements of Dehejia and Wahba (1999, 2002) that are misinterpreted or overlooked by Smith and Todd. Second, I address the issue of propensity score specification. Propensity score methods require that a different specification be selected for each treatment group–comparison group combination. Smith and Todd misapply the specifications Dehejia and Wahba selected for their samples to two samples for which the specifications are not necessarily appropriate. With suitable specifications selected for these alternative samples, more accurate estimates can be obtained.

Third, I address the issue of sensitivity of the results to changes in the specification of the propensity score. Presumably, the goal in using any estimator is to have a sense of the contexts in which it should perform well, and to have diagnostics that will raise a red flag when the technique is not working. Sensitivity of the estimates to small changes in the specification is the most basic check that a researcher should perform. In this sense, propensity score methods work for the National Supported Work Demonstration (NSW) data. For the Dehejia–Wahba sample, these methods produce reliable and robust estimates. For the original Lalonde sample and the Smith–Todd sample, these methods exclude themselves from the running because of sensitivity to changes in the specification.

The plan of the paper is as follows. Section 2 briefly reviews the Dehejia and Wahba articles. Section 3 reexamines the propensity score estimates under new specifications. Section 4 examines sensitivity of the estimates to changes in the specification. Section 5 concludes.

2. Rereading Dehejia and Wahba

There are two features of Dehejia and Wahba (1999, 2002) that merit emphasis in the context of Smith and Todd (2004). First, Dehejia and Wahba (1999, 2002) nowhere claim that matching estimators provide a "magic bullet" (Smith and Todd, 2004) method for evaluating social experiments. Instead, these papers conclude that:

> …[T]he methods we suggest are not relevant in all situations. There may be important unobservable covariates... However, rather than giving up, or relying on assumptions about the unobserved variables, there is substantial reward in exploring first the information contained in the variables that *are* observed. In this regard, propensity score methods can offer both a diagnostic on the quality of the comparison group and a means to estimate the treatment impact (Dehejia and Wahba 1999, p. 1062).

and

> …[t]he methods that we discuss in this paper should be viewed as a complement to the standard techniques in the researcher's arsenal. By starting with a

propensity score analysis, the researcher will have a better sense of the extent to which the treatment and comparison groups overlap and consequently of how sensitive estimates will be to the choice of functional form (Dehejia and Wahba 2002, p. 106).[2]

Nor do Dehejia and Wahba (1999, 2002) claim that propensity score methods *always* provide a reliable method for estimating the treatment effect in non-experimental studies. Instead, they demonstrate that propensity score methods *can* reliably estimate treatment effects, and they then try to establish some features of situations in which these methods might work. Among these, they identify observing more than 1 year of pre-treatment earnings information as important. This observation is a natural implication of Ashenfelter's (1978) and Ashenfelter and Card's (1985) findings in the training literature, and is also congruent with the findings of Heckman et al. (1998a).

Second, Smith and Todd's observation that propensity score methods do not yield robustly accurate estimates of the treatment effect for Lalonde's original sample is implicit in Dehejia and Wahba (1999, 2002). Dehejia and Wahba create their subsample from Lalonde's data in order to obtain two years of pre-treatment earnings information. For this subset of the data, they then demonstrate that estimates of the treatment effect based on only 1 year of pre-treatment earnings are not robust. It is thus a natural implication that propensity score methods would not work well in Lalonde's sample, where two years of pre-treatment earnings are not available.

3. Re-visiting the propensity score estimates

In this section, we re-examine Smith and Todd's contention that propensity score matching methods are unable to replicate the treatment effect in Lalonde's original sample or in the further subsample that Smith and Todd extract from the NSW. This conclusion is based on applying the propensity score models that Dehejia and Wahba developed for their samples (DW-PSID and DW-CPS) to these alternative samples (Lalonde-PSID, Lalonde-CPS, ST-PSID, and ST-CPS). However, as discussed in Dehejia and Wahba (1999, 2002), a different specification must be considered for each combination of treatment and comparison group.

This is demonstrated in Table 1, which—as in the Smith–Todd paper—applies the propensity score specifications used in Dehejia and Wahba to the Lalonde and Smith–Todd samples. There is no reason to believe that these specifications—selected specifically for Dehejia and Wahba's samples—will balance the covariates in

[2] Smith and Todd's use of quotation marks around the term "magic bullet" might suggest that it is a direct quote from Dehejia and Wahba, which it is not. Likewise when Smith and Todd refer to "recent claims in the literature by Dehejia and Wahba (1999, 2002) and others regarding the general effectiveness of matching estimators relative to more traditional econometric methods" they are again misrepresenting the tone of the articles, as the previous quotations illustrate. Since such statements are not contained in Dehejia and Wahba, they are presumably referring to the "others" whom they do not cite.

Table 1
Checking balance of the covariates using the Dehejia–Wahba propensity score specification for the Lalonde and Smith–Todd Samples

Comparison group/ treatment group	Number treated	Number control	Covariates not balanced
CPS/Lalonde[a]	297	15,992	Education, black
CPS/ST[a]	108	15,992	Black
PSID/Lalonde[b]	297	2,490	Married, black
PSID/ST[b]	108	2,490	Black

Notes: [a]Propensity score specification: constant, age, education, no degree, married, black, hispanic, age^2, education^2, (Re74 = 0), (Re75 = 0), Education × Re74, Age^3. [b]Propensity score specification: constant, age, education, married, no degree, black, hispanic, Re74, Re75, age^2, education^2, Re74^2, Re75^2, (Re74 = 0) × Black.

alternative samples. We test for mean differences in pre-treatment covariates across the treatment and comparison groups using the procedure described in Dehejia and Wahba (2002, Appendix), and find that a number of covariates (education, black, and married) in fact are not balanced.

Thus, it is not surprising that the Dehejia and Wahba propensity score specifications do not replicate the experimental treatment effects for the alternative samples considered in Smith and Todd. Likewise, it is not surprising that the specification used in Lalonde (1986) does not produce reliable estimates, since this specification was not chosen to balance the covariates in the context of propensity score estimation.

The first step in implementing propensity methods is to estimate the propensity score. Dehejia and Wahba (2002, Appendix) discuss how this can be done. Essentially, one searches for a specification that balances the pre-program covariates between the treatment and comparison groups conditional on the propensity score.[3] Since this procedure does not rely on looking at the outcomes, it does not constitute data mining in any way. Note that a different specification typically is required for each treatment group–comparison group combination. Thus, it is not surprising that Smith and Todd find that Dehejia and Wahba's specification for the DW-CPS (DW-PSID) sample does not perform well in the Lalonde-CPS or ST-CPS (Lalonde-PSID or ST-PSID) samples. Table 2 presents the specifications that are used in this paper. The specifications were selected on the basis of balancing pre-treatment

[3]This procedure is based on Theorem 1 in Rosenbaum and Rubin (1983), which states that conditional on the propensity score the distribution of (pre-program) covariates is independent of assignment to treatment. Note that this subsumes one of the specification tests used in Lalonde (1986) and developed in Heckman and Hotz (1989), namely the pre-program alignment test. Since one of the pre-program covariates we use is earnings, this test is implicit in our specification test for the propensity score.

Table 2
Propensity score specification

Comparison group	Treatment group	Propensity score specification
CPS	Lalonde	Constant, Re75, married, black, hispanic, age, education, married*u75, no degree*Re75, age^2
CPS	DW	Constant, Re74, Re75, married, black, hispanic, education, age, black *age
CPS	ST	Constant, Re74, Re75, married, black, hispanic, education, age, no degree*1(Re75 = 0), Re74*Re74
PSID	Lalonde	Constant, Re75, married, black, hispanic, age, education, black *education, hispanic*Re75, no degree *education
PSID	DW	Constant, Re74, Re75, married, black, hispanic, education, age, married*1(Re75 = 0), no degree*1(Re74 = 0).
PSID	ST	Constant, Re74, Re75, married, black, hispanic, education, age, hispanic*education, Re74^2

Notes: The propensity score specification is selected for each treatment–comparison group combination to balance pre-treatment covariates. Re74 and Re75 refer to earnings one and two years prior to the treatment. 1 (Re74 = 0) and 1 (Re75 = 0) are indicators for zero earnings.

Table 3
Estimated treatment effects, nearest-neighbor matching

Comparison group	Treatment group		
	Lalonde sample	Dehejia–Wahba sample	Smith–Todd sample
Experimental controls	886 (472)	1,794 (633)	2,717 (956)
CPS	880 (817)	1,589 (897)	2,705 (1474)
PSID	863 (1020)	1,869 (951)	2,711 (1402)

Notes: Each cell uses a different propensity score specification, corresponding to Table 2. The treatment effect for the experimental controls is computed using a difference in means. For the non-experimental comparison groups, the treatment effect is computed using nearest-neighbor matching; standard errors, in parentheses, are computed using the bootstrap.

covariates in their respective samples, using the method described in Dehejia and Wahba (2002, Appendix).

Table 3 presents the results of propensity score matching. These results are obtained using nearest-neighbor matching, the technique discussed in Dehejia and Wahba (2002). Columns (1)–(3) of the table represent, respectively, the alternative treatment groups: Lalonde's sample, Dehejia and Wahba's subsample, and Smith and Todd's subsample from Dehejia and Wahba. The first row corresponds to the

experimental control groups, and accordingly presents essentially unbiased estimates of the treatment effect for each sample. The next two rows correspond to the CPS and PSID comparison groups. Note that a different propensity score specification is used for each cell (i.e., the specifications listed in Table 2).

From the first column we can see that the propensity score matching estimate of the treatment effect in Lalonde's sample using the CPS comparison group is \$880, compared with the experimental benchmark of \$886, and that the estimated treatment effect using PSID comparisons for this sample is \$863. When Smith and Todd compute their bias estimates, they use a propensity score specification that was not specifically selected to balance the covariates for the Lalonde sample. For the Smith–Todd sample, column (3), the benchmark experimental estimate is \$2717. The propensity score matching estimates are \$2705 using the CPS and \$2711 using the PSID. Column 2 presents estimates for the DW sample. The experimental benchmark estimate is \$1794. The estimates from the CPS and PSID are \$1589 and \$1869, respectively.

Thus, propensity score methods are able to replicate the experimental benchmark estimates for all six treatment group–comparison group combinations. However, before we accept these estimates, we must check their sensitivity to changes in the specification, a diagnostic that is particularly important in the absence of an experimental benchmark estimate.

4. Sensitivity to changes in the specification

The final diagnostic that must be performed is to check the sensitivity of the estimated treatment effect to small changes in the specification of the propensity score (for example, the inclusion or deletion of higher-order terms). If the results are robust, then the estimates in Table 3 can reasonably be labeled as estimates from "the propensity score method." If, instead, the results are highly sensitive to changes in the specification, then a careful researcher would consider the propensity score method to have removed itself from the running.[4]

We perform two sensitivity checks. In Table 4, we apply the specifications chosen for a given treatment–comparison group combination to the other five treatment–comparison combinations. Note that there is no reason to expect that these specifications balance the pre-treatment covariates in the alternative samples; we perform this exercise purely as a sensitivity check.

In column (2), for Dehejia and Wahba's sample, we see that the estimated treatment effect is not as accurate for the alternative specifications. However, the estimated treatment effect is reasonably robust, ranging from \$1283 to \$1867 for the CPS and \$1313 to \$1991 for the PSID. In columns (1) and (3), we note that the

[4] Unless, of course, these alternative specifications all achieve a demonstrably worse balance of the pre-treatment covariates. Another diagnostic and reason why propensity score methods could exclude themselves from the running is a failure of overlap of the treatment and comparison groups in terms of the estimated propensity score; see Rubin (1977).

Table 4
Estimated treatment effects, nearest-neighbor matching

Comparison group	Propensity score specification	Treatment group		
		Lalonde sample	Dehejia–Wahba sample	Smith–Todd sample
Experimental controls		886 (472)	1,794 (633)	2,717 (956)
CPS	CPS/Lalonde	- -	1867 (932)	−77 (1448)
CPS	CPS/DW	−1,419 (701)	- -	184 (1171)
CPS	CPS/Smith-Todd	−473 (813)	1,283 (900)	- -
PSID	PSID/Lalonde	- -	1,313 (1689)	417 (2029)
PSID	PSID/DW	−37 (873)	- -	1,494 (1569)
PSID	PSID/Smith–Todd	247 (799)	1,991 (1017)	- -

Notes: Each row uses a different propensity score specification, corresponding to Table 2. The treatment effect for the experimental controls is computed using a difference in means. For the non-experimental comparison groups, the treatment effect is computed using nearest-neighbor matching; standard errors, in parentheses, are computed using the bootstrap.

estimates for the Lalonde and the Smith and Todd samples are not as robust. For the Lalonde sample, estimates range from -$473 to –$1419 for the CPS and from -$37 to $247 for the PSID. For the Smith–Todd sample, the estimates range from -$77 to $184 for the CPS and $417 to $1494 for the PSID. Even in the absence of a benchmark estimate from a randomized trial, one would hesitate to adopt estimates that are demonstrably sensitive to the specification of the propensity score.

Figs. 1 and 2 present an alternative sensitivity analysis. We consider all specifications up to four squares or interactions of the covariates. From these, we consider the specifications with the 10 highest Schwarz model selection criteria among those specifications that substantially balance the covariates within six equally spaced strata on the propensity score. In Figs. 1 and 2, we see that, for the Lalonde sample, the estimates from the 10 models selected by this method are neither close to the experimental benchmark estimate nor on average clustered around that estimate. Instead, for the Dehejia and Wahba sample, the models selected by this procedure produce estimates that are clustered around the experimental benchmark estimate for both the PSID and CPS comparison groups. Finally, for the ST sample, although the range of estimates is more focused than the estimates produced by the sensitivity analysis shown in Table 4, they still underestimate the treatment effect by about $1000.

Combining the two sets of sensitivity checks, in the absence of an experimental benchmark, a careful researcher would not be led to adopt the propensity score

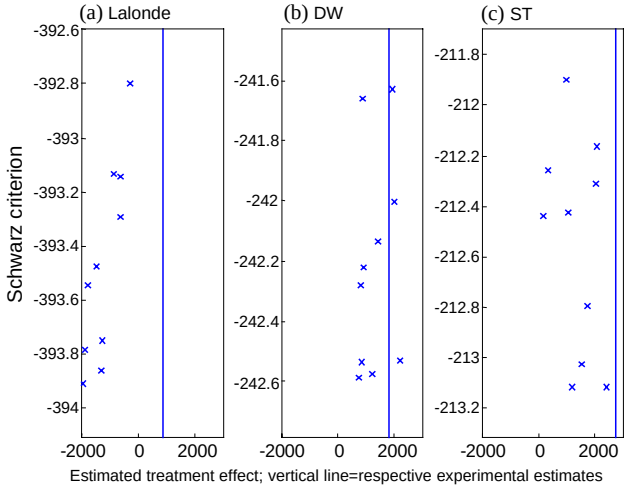

Fig. 1. Sensitivity of non-experimental estimates, PSID.

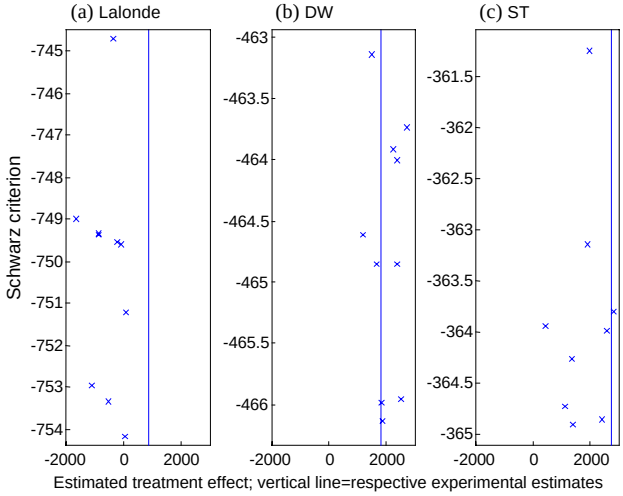

Fig. 2. Sensitivity of non-experimental estimates, CPS.

estimates for the Lalonde and Smith–Todd samples, but would adopt them for the Dehejia and Wahba sample.

5. Conclusion

Two points should be borne in mind in light of Smith and Todd's reanalysis of Dehejia and Wahba (1999, 2002) and more generally when applying propensity score

methods. First, a suitable and distinct propensity score specification must be estimated for each treatment group–comparison group combination. Second, one must examine sensitivity of the estimates to small changes in the specification. For the NSW data, accurate estimates are obtained for each of the three samples considered. However, for Lalonde's sample, and to a lesser extent for the Smith–Todd sample, these estimates are sensitive to small changes in the propensity score specification.

A judgment-free method for dealing with problems of sample selection bias is the Holy Grail of the evaluation literature, but this search reflects more the aspirations of researchers than any plausible reality. In practice, the best one can hope for is a method that works in an identifiable set of circumstances, and that is self-diagnostic in the sense that it raises a red flag if it is not functioning well. Propensity score methods are applicable when selection is based on variables that are observed. In the context of training programs, Dehejia and Wahba (1999, 2002), following on a suggestion from the training program literature (Ashenfelter, 1978; Ashenfelter and Card, 1985), suggest that two or more years of pre-treatment earnings are necessary. In terms of the self-diagnosis, the method and its associated sensitivity checks successfully identify the contexts in which it succeeds and those in which it does not succeed, at least for the NSW data.

Propensity score matching does not provide a silver-bullet, black-box technique that can estimate the treatment effect under all circumstances; neither the developers of the technique nor Dehejia and Wahba have claimed otherwise. However, with input and judgment from the researcher, it can be a useful and powerful tool.

Acknowledgements

I am grateful to Han Hong, Bo Honoré, Jennifer Hill, Kei Hirano, Guido Imbens, Don Rubin, and Barbara Sianesi for detailed discussions and suggestions, and to the Princeton Econometrics Reading Group for valuable comments. Remaining errors and omissions are my own.

References

Abadie, A., Imbens, G., 2001. Simple and bias-correct matching estimators for average treatment effects, unpublished.

Ashenfelter, O., 1978. Estimating the effects of training programs on earnings. Review of Economics and Statistics 60, 47–57.

Ashenfelter, O., Card, D., 1985. Using the longitudinal structure of earnings to estimate the effect of training programs. Review of Economics and Statistics 67, 648–660.

Dehejia, R., Wahba, S., 1999. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. Journal of the American Statistical Association 94, 1053–1062.

Dehejia, R., Wahba, S., 2002. Propensity Score Matching Methods for Nonexperimental Causal Studies. National Bureau of Economics Research Working Paper No. 6829, forthcoming Review of Economics and Statistics.

Heckman, J., Hotz, J., 1989. Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training. Journal of the American Statistical Association 84, 862–880.

Heckman, J., Ichimura, H., Todd, P., 1997. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Review of Economic Studies 64, 605–654.

Heckman, J., Ichimura, H., Smith, J., Todd, P., 1998a. Characterizing selection bias using experimental data. Econometrica 66, 1017–1098.

Heckman, J., Ichimura, H., Todd, P., 1998b. Matching as an econometric evaluation estimator. Review of Economic Studies 65, 261–294.

Lalonde, R., 1986. Evaluating the econometric evaluations of training programs. American Economic Review 76, 604–620.

Rosenbaum, P., Rubin, D., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Rosenbaum, P., Rubin, D., 1984. Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association 79, 516–524.

Rosenbaum, P., Rubin, D., 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity. American Statistician 39, 33–38.

Rubin, D., 1973. Matching to remove bias in observational studies. Biometrics 29, 159–183.

Rubin, D., 1976a. Multivariate matching methods that are equal percent bias reducing. I: some examples. Biometrics 32, 109–120.

Rubin, D., 1976b. Multivariate matching methods that are equal percent bias reducing. II: maximums on bias reduction for fixed sample sizes. Biometrics 32, 121–132.

Rubin, D., 1977. Assignment to a treatment group on the basis of a covariate. Journal of Educational Statistics 2, 1–26.

Rubin, D., 1979. Using multivariate matched sampling and regression adjustment to control bias in observation studies. Journal of the American Statistical Association 74, 318–328.

Rubin, D., Thomas, N., 1992a. Affinely Invariant Matching Methods with Ellipsoidal Distributions. The Annals of Statistics 20, 1079–1093.

Rubin, D., Thomas, N., 1992b. Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Covariates. Biometrika 79, 797–809.

Rubin, D., Thomas, N., 1996. Matching using estimated propensity scores: relating theory to practice. Biometrics 52, 249–264.

Smith, J., Todd, P., 2004. Does matching overcome Lalonde's critique of nonexperimental estimators. Journal of Econometrics (in press).