DOES MATCHING OVERCOME LALONDE'S CRITIQUE OF NON-EXPERIMENTAL ESTIMATORS? A POSTSCRIPT

Rajeev Dehejia*
Department of Economics and SIPA
Columbia University
and
NBER

1. Introduction

Jeffrey Smith and Petra Todd's comments on Dehejia and Wahba (1999, 2002) – Smith and Todd (2004a) and the subsequent exchange (Dehejia 2004 and Smith and Todd 2004b) – have been useful in highlighting some features of our original work and of propensity score methods more generally. However, like their original comment, their rejoinder illustrates common sources of confusion in applying these methods. In these remarks, I focus on three issues: the choice of research sample, the choice of propensity score specification, and the use of balancing tests.

2. The Choice of Sample

Smith and Todd contend that the choice of sample in Dehejia and Wahba is arbitrary and unmotivated, because we present our main results for a subset of Lalonde's (1986) data. On the contrary, in that paper we provide a strong economic motivation for why we focus on a subsample of Lalonde's data. The evaluation literature (Ashenfelter 1978, Ashenfelter and Card 1985, Card and Sullivan 1988, and Heckman, Ichimura, Smith, and Todd 1998) underlines the importance of controlling for a sufficiently rich set of pre-

* Department of Economics and SIPA, Columbia University, 420 W. 118th Street, Room 1022, New York, NY, 10027; e-mail: rd247@columbia.edu.

treatment covariates and in the context of labor training programs of controlling for more than one year of pre-treatment earnings.

For the subset of the data that we examine, information is available on two years of pre-treatment earnings, whereas Lalonde's full sample only contains information on one year of pre-treatment earnings. The difficulty lies matching earnings reports from the experiment to the earnings reported for the comparison groups (the former are expressed in terms of experimental time, whereas the latter are in calendar time). For individuals assigned to the treatment sufficiently early or individuals that had zero earnings, it is possible to make this match.

Furthermore, in Dehejia and Wahba (1999) we provide results both for Lalonde's original sample and for our sub-sample, and we show that the additional year of earnings information is indeed necessary. Hence, our choice of sample is clearly motivated, and our results are transparent to the choices we make.

3. Choice of Propensity Score Specification

A consistent mistake that Smith and Todd (2004a, 2004b) make in applying propensity methods is failing to select a propensity score specification for each treatment group-comparison group combination. Consequently, their specifications are biased or inefficient. In Smith and Todd (2004a), they incorrectly apply specifications selected for Lalonde's sample and our sample to the alternate sample that they examine. In Smith and Todd (2004b), they commit a related, albeit more subtle, error. They apply the propensity score specifications that were used in Dehejia (2004) – selected for various subsets of the NSW treatment group and the two non-experimental comparison groups – to other

samples (in particular the full set of treatment and control experimental units plus the relevant comparison group units). Though the NSW experimental control units are randomly sampled from the same population as the NSW treated units, and in that sense presumably share the same underlying (population) propensity score, the estimated propensity score should also account for sampling variation in the selection of a particular treatment group.

The distinction here relates to the fact that the estimated propensity score accounts not only for factors that determine selection into the treatment at a population level, but also sampling variability in the selection of a particular treatment group. Indeed, Hirano, Imbens, and Ridder (2004) have shown that propensity score methods are efficient *only* when the estimated propensity score is used, not when using the true propensity score even if it were available. This implies that propensity score methods can be used even when a random experiment is conducted (see for example Hill, Rubin, and Thomas 2000), and that each treatment group-comparison group combination requires its own propensity score specification.

Thus, it is not surprising that the bias estimates reported by Smith and Todd (2004b) for the specifications used in Dehejia (2004) are low when applied to the NSW-treatment group but are larger when applied to the two other samples. Another set of propensity score specifications should be selected for these alternative samples before estimating the bias or the treatment effect.

Finally it should be noted that Smith and Todd inflate their bias estimates by failing to impose the common support condition (i.e., restricting attention to comparison group members that fall within the support of the propensity score distribution of the

treatment group). This is makes their bias calculations very misleading, since the relevant question is the extent of bias for the estimated treatment effect and the treatment effect would be estimated using observations in the interval of common support.

4. Balancing Tests

As suggested by Rosenbaum and Rubin (1983), and illustrated in Dehejia and Wahba, balancing tests are useful diagnostics on the suitability of a propensity score specification for a particular treatment group-comparison group combination.

As in their bias calculations, Smith and Todd seem to use the full set of comparison observations in their balancing tests. This is misleading because the objective of propensity score methods is to focus attention precisely on the subset of the comparison group that is most comparable to the treatment group. As illustrated in Dehejia and Wahba (1999, 2002), there are many (indeed, most of the) observations in the PSID and CPS comparison groups are not comparable to the treatment group.

Nonetheless, Smith and Todd's observation that there is no consensus on which balancing test to use is useful, and points to the value of ongoing research on this and related topics.

5. Conclusion

In conclusion, it is notable that fundamentally there is more agreement than disagreement between Smith and Todd and myself. We all agree that propensity score methods are a valuable tool in the researcher's arsenal and that these methods are not a silver bullet fix to all evaluation problems.

REFERENCES

Ashenfelter, Orley (1978). "Estimating the Effects of Training Programs on Earnings," *Review of Economics and Statistics,* 60, 47-57.

------ and D. Card (1985). "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics*, 67, 648-660.

Card, David, and Daniel Sullivan (1988). "Measuring the Effect of Subsidized Training Programs on Movements in and out of Employment," *Econometrica*, 56, 497-530.

Dehejia, Rajeev (2004), "Practical Propensity Score Matching: A Reply to Smith and Todd," forthcoming *Journal of Econometrics*.

------ and Sadek Wahba (1999). "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053-1062.

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

Hill, Jennifer, Rubin, D. B., and N. Thomas. (2000) "The Design of the New York School Choice Scholarship Program Evaluation," in Leonard Bickman (ed.), *Research Design: Donald Campbell's Legacy*. New York: Sage Publications.

Lalonde, Robert (1986). "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review*, 76, 604-620.

Rosenbaum, P., and D. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41-55.

Smith, Jeffrey, and Petra Todd (2004a), "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators," forthcoming *Journal of Econometrics*.
------ and ------ (2004b), "Rejoinder," forthcoming *Journal of Econometrics*.