Journal of Econometrics 175 (2013) 1–21

Contents lists available at SciVerse ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

The performance of estimators based on the propensity score

Martin Huber, Michael Lechner *, Conny Wunsch

Swiss Institute for Empirical Economic Research (SEW), University of St. Gallen, Varnbüelstrasse 14, CH-9000 St. Gallen, Switzerland

ARTICLE INFO

Article history:
Received 20 October 2010
Received in revised form
13 November 2012
Accepted 15 November 2012
Available online 27 February 2013

JEL classification:
C21

Keywords:
Propensity score matching
Kernel matching
Inverse probability weighting
Inverse probability tilting
Selection on observables
Empirical Monte Carlo study
Finite sample properties

ABSTRACT

We investigate the finite sample properties of a large number of estimators for the average treatment effect on the treated that are suitable when adjustment for observed covariates is required, like inverse probability weighting, kernel and other variants of matching, as well as different parametric models. The simulation design used is based on real data usually employed for the evaluation of labour market programmes in Germany. We vary several dimensions of the design that are of practical importance, like sample size, the type of the outcome variable, and aspects of the selection process. We find that trimming individual observations with too much weight as well as the choice of tuning parameters are important for all estimators. A conclusion from our simulations is that a particular radius matching estimator combined with regression performs best overall, in particular when robustness to misspecifications of the propensity score and different types of outcome variables is considered an important property.

1. Introduction

Semiparametric estimators using the propensity score to adjust in one way or another for covariate differences are now well-established. They are used for estimating causal effects in a selection-on-observables framework with discrete treatments, or for simply purging the means of an outcome variable in two or more subsamples from differences due to observed variables.[1] Compared to (non-saturated) parametric regressions, they have the advantage of including the covariates in a more flexible way without incurring a curse-of-dimensionality problem, and of allowing for effect heterogeneity. The former problem is highly relevant due to the large number of covariates that should usually be adjusted for. It is tackled by collapsing the covariate information into a single parametric function. This function, the so-called propensity score, is defined as the probability of being observed in one of two subsamples conditional on the covariates. The difference to parametric regression is that this parametric function is not directly related to the outcome (as it would be in regression) and thus, additional robustness to misspecification can be expected.[2] These methods originate from the pioneering work of Rosenbaum and Rubin (1983) who show that balancing two samples on the propensity score is sufficient to equalize their covariate distributions.

Although many of these propensity-score-based methods are not asymptotically efficient (see for example Heckman et al., 1998a,b; Hahn, 1998),[3] they are the work-horses in the literature on programme evaluation and are now rapidly spreading to other fields. They are usually implemented as semiparametric estimators: the propensity score is based on a parametric model, but the relationship between the outcome variables and the propensity score is non-parametric. However, despite the popularity of propensity-score-based methods, the issue of which version of the many different estimators suggested in the literature should be used in a particular application is still unresolved,

* Corresponding author. Tel.: +41 712242814.
E-mail address: Michael.Lechner@unisg.ch (M. Lechner).
URL: http://www.michael-lechner.eu (M. Lechner).

[1] See for example the recent surveys by Blundell and Costa Dias (2009), Imbens (2004), and Imbens and Wooldridge (2009) for a discussion of the properties of such estimators as well as a list of recent applications.

[2] The propensity-score could also be non-parametrically estimated for maximum robustness. In practice, this is however avoided because the dimension of covariates is too large for such an estimator to have desirable properties with the samples usually available for such studies.

[3] See the paper by Angrist and Hahn (2004) for an alternative justification of conditioning on the propensity score by using non-standard (panel) asymptotic theory.

0304-4076/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jeconom.2012.11.006

despite recent advances in important Monte Carlo studies by Frölich (2004) and Busso et al. (forthcoming, 2009). In this paper we address this question and add further insights to it. Broadly speaking, the popular estimators can be subdivided into four classes: parametric estimators (like OLS or probit or their so-called double-robust relatives, see Robins et al., 1992), inverse (selection) probability weighting estimators (similar to Horvitz and Thompson, 1952) or to the recently introduced titling version by Graham et al. (2011, 2012), direct matching estimators (Rubin, 1974; Rosenbaum and Rubin, 1983), and kernel matching estimators (Heckman et al., 1998a,b).[4] However, many variants of the estimators exist within each class and several methods combine the principles underlying these main classes.

There are two strands of the literature that are relevant for our research question: First, the literature on the asymptotic properties of a subset of estimators provides some guidance on their small sample properties. In Section 3 we review this literature and discuss the various estimators. Unfortunately, asymptotic properties have not (yet?) been derived for all estimators used in practice, nor is it obvious how well they approximate small sample behaviour. Furthermore, these results are usually not informative for the important choice of tuning parameters on which many estimators critically depend (e.g., number of matched neighbours, bandwidth selection in kernel matching).

The second strand of the literature provides Monte Carlo evidence on the properties of the estimators of the effects.[5] As one of the first papers investigating estimators from several classes simultaneously, Frölich (2004) found that a particular version of kernel-matching based on local regressions with finite sample adjustments (local ridge regression) performs best. In contrast, Busso et al. (forthcoming, 2009) conclude that inverse probability weighting (IPW) has the best properties (when using normalized weights for estimation). They explain the differences to Frölich (2004) by claiming that he (i) considers unrealistic data generating processes and (ii) does not use an IPW estimator with normalized weights. In other words, they point to the design dependence of the Monte Carlo results as well as to the requirement of using optimized variants of the estimators. Below, we argue that their work may be subject to the same criticism. This provides a major motivation for our study.

We contribute to the literature on the properties of estimators based on adjusting for covariate differences in the following way: firstly, we suggest a different approach to conduct simulations. This approach is based on 'real' data. Therefore, we call our particular implementation of this idea an 'Empirical Monte Carlo Study'.[6] The basic idea is to use the empirical data to simulate realistic 'placebo treatments' among the non-treated. The various estimators then use the remaining non-treated in different ways to estimate the (known) non-treatment outcome of the 'placebo-treated'. Selection into treatment, which is potentially of key importance for the performance of the various estimators, is based on a selection process directly obtained from the data. Moreover, we exploit the actual dependence of the outcome of interest on the covariates on which selection is based in the data rather than making assumptions on this relation when specifying the data generating process. Thus, this approach is less prone to the standard critique of simulation studies that the chosen data generating processes are irrelevant for real applications. Since our model for the propensity score mirrors specifications used in past applied work, it depends on many more covariates compared to the studies mentioned above. Although this makes the simulation results particularly plausible in our context of labour market programme evaluation in Europe, this may also be seen as a limitation concerning its applications to other fields. Therefore, to help generalize the results outside our specific data situation, we modify many features of the data generating process, like the type of the outcome variable and as well as various aspects of the selection process.

Secondly, we consider standard estimators as well as their modified (optimized?) versions based on different tuning parameters such as bandwidth or radius choice. This leads to a large number of estimators to evaluate, but it also provides us with more information on important choices regarding the parameters on which the various estimators depend. Such estimators may also consist of combinations of estimators, like combining matching with weighted regression, which have not been considered in any simulation so far. Finally, we reemphasize the relevance of trimming to improve the finite sample properties of all estimators. The rule we propose is (i) a data driven trimming rule, (ii) easy to implement, (iii) identical for all estimators, and (iv) avoids asymptotic bias. We show that for almost all estimators considered, including the parametric ones, trimming based on this rule effectively improves their performance.

Overall, we find that (i) trimming observations that have 'too large' a weight is important for many estimators; (ii) the choices of the various tuning parameters play an important role; (iii) simple matching estimators are inefficient and have considerable small sample bias; (iv) no estimator is superior in all designs and for all outcomes; (v) particular bias-adjusted radius (or calliper) matching estimators perform best on average, but may have fat tails if the number of controls is not large enough; and finally, (vi) flexible, but simple parametric approaches do almost as well in the smaller samples, because their gain in precision frequently compensates (in part) for their larger bias which, however, dominates when samples become larger. Strictly speaking these properties relate to our particular data generating process (DGP) only. However, at least such a DGP is typical for an important application of matching methods, namely labour market evaluations.

The paper proceeds as follows: in the next section we describe our Monte Carlo design, relegating many details as well as descriptive statistics to online Appendices B and C, where the latter contains a description of the support features of our data. In Section 3 we discuss the basic setup of each of the relevant estimators and their properties, as well as the issue of trimming, while relegating the technical details of the estimators to Appendix. The main results are presented in Section 4, while the full set of results is given in online Appendix D. Section 5 concludes and online Appendix E contains further sensitivity checks. The website of this paper (www.sew.unisg.ch/lechner/matching) will contain additional material that has been removed from the paper for the sake of brevity, in particular Appendices B, C, D, and E as well as the Gauss, Stata, and R codes for the preferred estimators. Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jeconom.2012.11.006.

[4] There is also the approach of stratifying the data along the values of the propensity score ('blocking'), but this approach did not receive much attention in the empirical economic literature and does not have very attractive theoretical properties. It is thus omitted (see for example Imbens, 2004, for a discussion of this approach).

[5] There are several papers not interested in the properties of the estimators of the effects, but merely in the quality of covariate balancing of different matching methods. For example, King et al. (2011) motivate this by not regarding matching as an estimator, but merely as a 'pre-processor' that purges the data from differences related to observed covariates. After this pre-processing step, other estimators are used with the matched data to obtain the final result.

[6] Stigler (1977) is probably the first paper explicitly suggesting a way to do a type of Monte Carlo study with real data (we thank a referee of this journal for this reference). See Section 3.1 for more recent references using the same basic idea of informing the simulations by real data.