DISCUSSION PAPER SERIES

IZA DP No. 4304

A Note on Adapting Propensity Score Matching and Selection Models to Choice Based Samples

James J. Heckman
Petra E. Todd

July 2009

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

A Note on Adapting Propensity Score Matching and Selection Models to Choice Based Samples

James J. Heckman
University of Chicago, University College Dublin,
Cowles Foundation, Yale University,
American Bar Foundation and IZA

Petra E. Todd
University of Pennsylvania,
NBER and IZA

Discussion Paper No. 4304
July 2009

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 4304
July 2009

ABSTRACT

A Note on Adapting Propensity Score Matching and Selection Models to Choice Based Samples*

The probability of selection into treatment plays an important role in matching and selection models. However, this probability can often not be consistently estimated, because of choice-based sampling designs with unknown sampling weights. This note establishes that the selection and matching procedures can be implemented using propensity scores fit on choice-based samples with misspecified weights, because the odds ratio of the propensity score fit on the choice-based sample is monotonically related to the odds ratio of the true propensity scores.

JEL Classification: C52

Keywords: choice-based sampling, matching models, propensity scores, selection models

Corresponding author:

James J. Heckman
Department of Economics
University of Chicago
1126 East 59th Street
Chicago, IL 60637
USA
E-mail: jjh@uchicago.edu

* This research was supported by NSF SBR 93-21-048 and NSF 97-09-873 and NICHD 40-4043-000-85-261.

1 Introduction

The probability of selection into a treatment, also called the propensity score, plays a central role in classical selection models and in matching models (see, e.g., Heckman, 1980; Heckman and Navarro, 2004; Heckman and Vytlacil, 2007; Hirano et al., 2003; Rosenbaum and Rubin, 1983). [1] Heckman and Robb (1986, reprinted 2000), Heckman and Navarro (2004) and Heckman and Vytlacil (2007) show how the propensity score is used differently in matching and selection models. They also show that, given the propensity score, both matching and selection models are robust to choice-based sampling, which occurs when treatment group members are over- or under-represented relative to their frequency in the population. Choice-based sampling designs are frequently chosen in evaluation studies to reduce the costs of data collection and to obtain more observations on treated individuals. Given a consistent estimate of the propensity score, matching and classical selection methods are robust to choice-based sampling, because both are defined conditional on treatment and comparison group status.

This note extends the analysis of Heckman and Robb (1985),Heckman and Robb (1986, reprinted 2000) to consider the case where population weights are unknown so that the propensity score cannot be consistently estimated. In evaluation settings, the population weights are often unknown or cannot

[1] It also plays a key role in instrumental variables models (see Heckman et al., 2006). Heckman and Vytlacil (2007) discuss the different role played by the propensity score in matching IV and selection models.

easily be estimated.[2] For example, for the National Supported Work training program studied in LaLonde (1986), Dehejia and Wahba (1999, 2002) and in Smith and Todd (2005), the population consists of all persons eligible for the program, which was targeted at drug addicts, ex-convicts, and welfare recipients. Few datasets have the information necessary to determine whether a person is eligible for the program, so it would be difficult to estimate the population weights needed to consistently estimate propensity scores.

In this note, we establish that matching and selection procedures can still be applied when the propensity score is estimated on unweighted choice based samples. The idea is simple. To implement both matching and classical selection models, only a monotonic transformation of the propensity score is required. In choice based samples, the odds ratio of the propensity score estimated using misspecified weights is monotonically related to the odds ratio of the true propensity scores. Thus, selection and matching procedures can identify population treatment effects using misspecified estimates of propensity scores fit on choice-based samples.

2 Discussion of the Proposition

Let $D = 1$ if a person is a treatment group member; $D = 0$ if the person is a member of the comparison group. $X = x$ is a realization of X. In the

[2]The methods of Manski and Lerman (1977) and Manski (1986) for adjusting for choice-based sampling in estimating the discrete choice probabilities cannot be applied when the weights are unknown and cannot be identified from the data.

population generated from random sampling, the joint density is

$$g(d,x) = [\Pr(D=1 \mid x)]^d [Pr(D=0 \mid x)]^{1-d} g(x)$$
$$\text{for} \qquad D=d, \qquad d \in \{0,1\},$$

where g is the density of the data. By Bayes's theorem, we have, letting $Pr(D=1)=P$,

(1a) $\quad g(x \mid D=1)P = g(x)Pr(D=1 \mid x)$

and

(1b) $\quad g(x \mid D=0)(1-P) = g(x)Pr(D=0 \mid x).$

Take the ratio of (1a) to (1b)

(2) $\quad \frac{g(x|D=1)}{g(x|D=0)}\left(\frac{P}{1-P}\right) = \frac{Pr(D=1|x)}{Pr(D=0|x)}.$

Assume $0 < Pr(D=1 \mid x) < 1$. From knowledge of the densities of the data in the two samples, $g(x \mid D=1)$ and $g(x \mid D=0)$, one can form a scalar multiple of the ratio of the propensity score without knowing P. The odds ratio is a monotonic function of the propensity score that does not require knowledge of the true sample weights. In a choice-based sample, both the numerator and denominator of the first term in (2) can be consistently estimated. This monotonic function can replace $P(x)$ in implementing both matching and nonparametric selection models.

However, estimating $g(x \mid D=d)$ is demanding of the data when X is of high dimension. Instead of estimating these densities, we can substitute for the left hand side of (2) the odds ratio of the estimated conditional probabilities obtained using the choice-based sample with the wrong weights.

(*i.e.* for example, ignoring the fact that the data are a choice based sample). The odds ratio of the estimated probabilities is a scalar multiple of the true odds ratio. It can therefore be used instead of $Pr(D = 1 \mid X)$ to match or construct nonparametric control functions in selection bias models.

In the choice-based sample, let $\tilde{P}r(D = 1 \mid x)$ be the conditional probability that $D = 1$ and P^* be the unconditional probability of sampling $D = 1$, where $P^* \neq P$, the true population proportion. The joint density of the data from the sampled population is

$$[g(x \mid D = 1)P^*]^d[g(x \mid D = 0)(1 - P^*)]^{1-d}.$$

Using (1a) and (1b) to solve for $g(x \mid D = 1)$ and $g(x \mid D = 0)$ one may write the data density as

$$\left[\frac{Pr(D=1|x)g(x)}{P}P^*\right]^d \left[\frac{Pr(D=0|x)g(x)}{(1-P)}(1 - P^*)\right]^{1-d}$$

so

(3a) $$\tilde{P}r(D = 1 \mid x) = \frac{Pr(D=1|x)g(x)\frac{P^*}{P}}{g(x|D=1)P^*+g(x|D=0)(1-P^*)}$$

and

(3b) $$\tilde{P}r(D = 0 \mid x) = \frac{Pr(D=0|x)g(x)\frac{1-P^*}{1-P}}{g(x|D=1)P^*+g(x|D=0)(1-P^*)}.$$

Under random sampling, the right-hand sides of (3a) and (3b) are the limits to which the choice-based probabilities converge. Taking the ratio of (3a) to (3b), assuming the latter is not zero, one obtains

(4) $$\frac{\tilde{P}r(D=1|x)}{\tilde{P}r(D=0|x)} = \frac{Pr(D=1|x)}{Pr(D=0|x)}\left(\frac{P^*}{1-P^*}\right)\left(\frac{1-P}{P}\right).$$

Thus, one can estimate the ratio of the propensity score up to scale (the scale is the product of the two terms on the right-hand side of (4)). Instead

of estimating matching or semiparametric selection models using $\Pr(D = 1 \mid x)$ (as in, for example, Ahn and Powell (1993); Heckman (1980); Heckman and Hotz (1989); Heckman et al. (1998); Heckman and Robb (1986); Powell (2001), one can, instead, use the odds ratio of the estimate $\tilde{Pr}(D = 1 \mid x)$, which is monotonically related to the true $Pr(D = 1 \mid x)$. In the case of a logit $P(x), P(x) = \exp(x\beta)/(1 + \exp(x\beta))$, the log of this ratio becomes

$$\ln \frac{\tilde{Pr}(D=1|x)}{\tilde{Pr}(D=0|x)} = x\tilde{\beta}$$

where the slope coefficients are the true values and the intercept $\tilde{\beta}_0 = \beta_0 + \ell n(P^*/(1 - P^*)) + \ell n((1 - P)/P))$, where β_0 is the true value.[3]

In implementing nearest-neighbor matching estimators, matching on the log odds ratio gives identical estimates to matching on the (unknown) $\Pr(D = 1 \mid x)$, because the odds ratio preserves the ranking of the neighbors. In application of either matching or classical selection bias correction methods, one must account for the usual problems of using estimated log odds ratios instead of true values.[4]

[3]See Manski and McFadden (1981, p. 26).

[4]For discussion related to using estimated propensity scores, see Hahn (1998); Heckman et al. (1998); Heckman et al. (1998); Hirano et al. (2003).

References

Ahn, H. and J. Powell (1993, July). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics 58*(1-2), 3–29.

Dehejia, R. and S. Wahba (1999, December). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association 94*(448), 1053–1062.

Dehejia, R. and S. Wahba (2002, February). Propensity score matching methods for nonexperimental causal studies. *Review of Economics and Statistics 84*(1), 151–161.

Hahn, J. (1998, March). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica 66*(2), 315–31.

Heckman, J. J. (1980). Addendum to sample selection bias as a specification error. In E. Stromsdorfer and G. Farkas (Eds.), *Evaluation Studies Review Annual*, Volume 5. Beverly Hills: Sage Publications.

Heckman, J. J. and V. J. Hotz (1989, December). Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of Manpower Training. *Journal of the American Statistical Association 84*(408), 862–874. Rejoinder also published in Vol. 84, No. 408, (Dec. 1989).

Heckman, J. J., H. Ichimura, J. Smith, and P. E. Todd (1998, September). Characterizing selection bias using experimental data. *Econometrica 66*(5), 1017–1098.

Heckman, J. J., H. Ichimura, and P. E. Todd (1998, April). Matching as an econometric evaluation estimator. *Review of Economic Studies 65*(223), 261–294.

Heckman, J. J. and S. Navarro (2004, February). Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and Statistics 86*(1), 30–57.

Heckman, J. J. and R. Robb (1985, October-November). Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics 30*(1-2), 239–267.

Heckman, J. J. and R. Robb (1986). Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In H. Wainer (Ed.), *Drawing Inferences from Self-Selected Samples*, pp. 63–107. New York: Springer-Verlag. Reprinted in 2000, Mahwah, NJ: Lawrence Erlbaum Associates.

Heckman, J. J., S. Urzua, and E. J. Vytlacil (2006). Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics 88*(3), 389–432.

Heckman, J. J. and E. J. Vytlacil (2007). Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effects in new environments. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics*, Volume 6B, pp. 4875–5144. Amsterdam: Elsevier.

Hirano, K., G. W. Imbens, and G. Ridder (2003, July). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

LaLonde, R. J. (1986, September). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review 76*(4), 604–620.

Manski, C. F. (1986, February). Semiparametric analysis of binary response from response-based samples. *Journal of Econometrics 31*(1), 31–40.

Manski, C. F. and S. R. Lerman (1977, November). The estimation of choice probabilities from choice based samples. *Econometrica 45*(8), 1977–1988.

Manski, C. F. and D. McFadden (1981). Statistical analysis of discrete probability models. In C. F. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*, pp. 2–49. Cambridge, MA: MIT Press.

Powell, J. L. (2001). Semiparametric estimation of bivariate latent variable models. In C. Hsiao, K. Morimune, and J. L. Powell (Eds.), *Nonlinear*

Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya. New York: Cambridge University Press.

Rosenbaum, P. R. and D. B. Rubin (1983, April). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

Smith, J. A. and P. E. Todd (2005, March-April). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics 125*(1-2), 305–353.