forthcoming in: Econometric Reviews (2004), 23, 167-174

A Note on the Role of the Propensity Score for Estimating Average Treatment Effects

Markus Frölich
Department of Economics, University of St. Gallen

Last changes: October 30, 2003

Abstract:

Hahn (1998) derived the semiparametric efficiency bounds for estimating the average treatment effect (ATE) and the average treatment effect on the treated (ATET). The variance of ATET depends on whether the propensity score is known or unknown. Hahn attributes this to 'dimension reduction'. In this paper an alternative explanation is given: Knowledge of the propensity score improves upon the estimation of the distribution of the confounding variables.

Keywords: Evaluation, matching, causal effect, semiparametric efficiency bound
JEL classification: C13, C14

The author is also affiliated with the Institute for the Study of Labor (IZA), Bonn. Financial support from the Swiss National Science Foundation (Project NSF 4043-058311) and the Grundlagenforschungsfonds HSG (project G02110112) is gratefully acknowledged. I would like to thank Michael Lechner, an associate editor and two anonymous referees for helpful comments and suggestions. Address for correspondence: Markus Frölich, Swiss Institute for International Economics and Applied Economic Research (SIAW), Department of Economics, University of St. Gallen, Dufourstrasse 48, CH-9000 St.Gallen, Switzerland; markus.froelich@unisg.ch, www.siaw.unisg.ch/froelich

1 Introduction

Propensity score matching is a technique widely used in biometrics, econometrics and other social sciences to estimate average treatment effects of medical treatments, active labour market programmes, training programmes etc. Its popularity stems from the fact that, instead of controlling for all confounding factors X, it suffices to control for a one-dimensional propensity score (the conditional probability of treatment receipt) to remove all selection bias (Rosenbaum and Rubin 1983). This reduces the dimension of the estimation problem substantially. However, in many applications the propensity score is unknown and needs to be estimated. This gave rise to a debate on how much efficiency is lost by not knowing the true propensity score. Hahn (1998) derived the semiparametric efficiency bounds for the average treatment effect and the average treatment effect on the treated. He found that the variance bound for the average treatment effect (ATE) is completely unaffected by knowledge of the propensity score, whereas knowing the propensity score reduces the variance for the average effect on the treated (ATET). This leads to the question, *why* these bounds are affected differently. Hahn argues that the reduction in the variance of the treatment effect on the treated "can be solely attributed to the 'dimension reduction' feature of the propensity score".

This paper provides a different explanation for the role of the propensity score. It argues that the reason for why knowledge of the propensity score affects the variance bound is not the dimension reduction, but the information it provides for estimating the *distribution function of the confounding variables X in the treated subpopulation.* This distribution function $F_{X|treated}$ is used as weighting function for the ATET. When the propensity score is unknown, $F_{X|treated}$ is identified by the X observations of the treated individuals; non-treated individuals are not informative for estimating the distribution of X among the treated. On the other hand, if the propensity score is known, also the non-treated individuals are helpful for estimating $F_{X|treated}$, because $F_{X|untreated}$ and $F_{X|treated}$ are related via the propensity score.

The variance of the ATE is unaffected, because it is obtained through weighting by the distribution of X in the full population, which in any case is estimated from all the treated and the non-treated observations together.

2 Efficiency bounds and the propensity score

Define Y_i^0, Y_i^1 as the *potential outcomes* of individual i: Y_i^0 is the outcome that individual i would realize if not receiving the treatment and Y_i^1 the outcome if receiving the treatment. The average causal impact of the treatment can be measured by the average treatment effect

$$\alpha = E[Y^1 - Y^0] \tag{1}$$

and by the average treatment effect on the treated

$$\alpha_T = E[Y^1 - Y^0 | D = 1], \tag{2}$$

where D_i indicates whether an individual received treatment ($D_i = 1$) or not ($D_i = 0$). Whereas α measures the impact of treatment for the full population, α_T represents the effect for the subpopulation of individuals who actually received treatment. If treatment is unconfounded (Rubin 1974), $Y^0, Y^1 \perp\!\!\!\perp D \,|X$ and the treatment effects are identified as

$$\begin{aligned} \alpha &= \int (m_1(x) - m_0(x)) \cdot dF_X \\ \alpha_T &= \int (m_1(x) - m_0(x)) \cdot dF_{X|D=1}, \end{aligned} \tag{3}$$

where $m_d(x) = E[Y|X = x, D = d]$ is the conditional mean function, $f_X = dF_X$ is the density of X in the population and $f_{X|D=1} = dF_{X|D=1}$ is the density of X in the treated subpopulation.[1][2]

Since nonparametric estimation of $m_d(x)$ can be difficult in finite samples if the dimension of X is high, Rosenbaum and Rubin (1983) suggested to reduce the dimension of the estimation problem by making use of the balancing property of the propensity score: The unconfoundedness assumption implies $Y^0, Y^1 \perp\!\!\!\perp D \,|p(X)$, where $p(x) = P\,(D = 1|X = x)$ is the propensity score and the average treatment effects are also identified as

$$\begin{aligned} \alpha &= \int (\mathfrak{m}_1(\rho) - \mathfrak{m}_0(\rho)) \cdot dF_p(\rho) \\ \alpha_T &= \int (\mathfrak{m}_1(\rho) - \mathfrak{m}_0(\rho)) \cdot dF_{p|D=1}(\rho), \end{aligned} \tag{4}$$

[1] A further condition for identification is that $Supp(X|D = 1) = Supp(X|D = 0)$, or equivalently that $0 < p(x) < 1 \quad \forall x \in Supp(X)$, where $p(x) \equiv P\,(D = 1|X = x)$ is the propensity score.

[2] Hirano, Imbens, and Ridder (2003) also consider a weighted average treatment effect $\int (m_1(x) - m_0(x))\, g(x) dF_X / \int g(x) dF_X$ for a known weighting function $g(x)$ and derive its efficiency bound. The weighted average treatment effect contains α_T as a special case, for $g(x) = p(x)$.

where $\mathfrak{m}_d(\rho) = E[Y|p(X) = \rho, D = d]$ is the mean outcome conditional on the propensity score, F_p is the distribution of $p(x)$ in the population and $F_{p|D=1}$ is the distribution of the propensity score in the treated subpopulation. Since the propensity score $p(x)$ is *one-dimensional*, nonparametric estimation of $\mathfrak{m}_d(\rho)$ is usually substantially less demanding than nonparametric estimation of $m_d(x)$. In this sense, propensity score matching circumvents the dimensionality problem of nonparametric regression on X and is therefore widely used in applied evaluation studies. On the other hand, the propensity score is often unknown and estimation of the average treatment effects must make do with an estimated propensity score.

To analyze the role of the propensity score, Hahn (1998) derived the *semiparametric efficiency bounds* for known and for unknown propensity score. To make the following discussion more intuitive, the variance bounds are henceforth scaled by the number of observations, i.e. divided by $n_0 + n_1$, to reflect the approximate variance for a given number of observations. (n_1, n_0 are the number of treated/non-treated observations, respectively.) In addition, in the following expressions, $P(D = 1)$ is approximated by $n_1/(n_0+n_1)$. The scaled variance bound for α is then

$$\frac{1}{n_0}\underset{f_0}{E}\left[\sigma_0^2(X)\frac{f_X^2(X)}{f_{X|D=0}^2(X)}\right] + \frac{1}{n_1}\underset{f_1}{E}\left[\sigma_1^2(X)\frac{f_X^2(X)}{f_{X|D=1}^2(X)}\right] + \frac{1}{n_0+n_1}E\left[(m_1(X) - m_0(X) - \alpha)^2\right], \tag{5}$$

where $\sigma_d^2(x) = Var\,(Y|X = x, D = d)$, and $E_{f_1}[\cdot] = \int \cdot f_{X|D=1}(x)dx$ refers to the expected value in the treated subpopulation and $E_{f_0}[\cdot]$ to the expected value in the non-treated subpopulation (see appendix). This variance bound is the same for known and for unknown propensity score, i.e. knowledge of the true propensity score is not informative for estimating α.

In contrast, the variance bound for the treatment effect on the treated α_T depends on knowledge of the propensity score. If the propensity score is *unknown*, the scaled variance bound of α_T is

$$\frac{1}{n_0}\underset{f_0}{E}\left[\sigma_0^2(X)\frac{f_{X|D=1}^2(X)}{f_{X|D=0}^2(X)}\right] + \frac{1}{n_1}\underset{f_1}{E}\left[\sigma_1^2(X)\right] + \frac{1}{n_1}\underset{f_1}{E}\left[(m_1(X) - m_0(X) - \alpha_T)^2\right], \tag{6}$$

while it is lower when the propensity score is *known*:

$$\frac{1}{n_0}\underset{f_0}{E}\left[\sigma_0^2(X)\frac{f_{X|D=1}^2(X)}{f_{X|D=0}^2(X)}\right] + \frac{1}{n_1}\underset{f_1}{E}\left[\sigma_1^2(X)\right] + \frac{1}{n_0+n_1}\underset{f_1}{E}\left[\frac{f_{X|D=1}(X)}{f_X(X)}(m_1(X) - m_0(X) - \alpha_T)^2\right], \tag{7}$$

see appendix. Hahn (1998) attributes this reduction of the variance from (6) to (7) to the *'dimension reducing' property of the propensity score*. This interpretation, however, cannot explain why the variance of α is not affected.

Hirano, Imbens, and Ridder (2003) analyze the estimation of α and α_T through weighting by the propensity score, noting that $\alpha = E\left[\frac{YD}{p(X)} - \frac{Y(1-D)}{1-p(X)}\right]$ and $\alpha_T = E\left[\frac{p(X)}{P(D=1)}\left(\frac{YD}{p(X)} - \frac{Y(1-D)}{1-p(X)}\right)\right]$. If the propensity score were known, α could be estimated as $\hat{\alpha}(p) = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_iD_i}{p(X_i)} - \frac{Y_i(1-D_i)}{1-p(X_i)}$ or as $\hat{\alpha}(\hat{p}) = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_iD_i}{\hat{p}(X_i)} - \frac{Y_i(1-D_i)}{1-\hat{p}(X_i)}$. Potential estimators for α_T are $\hat{\alpha}_T(p,\hat{p})$, $\hat{\alpha}_T(p,p)$ and $\hat{\alpha}_T(\hat{p},\hat{p})$, where

$$\hat{\alpha}_T(p,\hat{p}) = \frac{\sum p(X_i)\left(\frac{Y_iD_i}{\hat{p}(X_i)} - \frac{Y_i(1-D_i)}{1-\hat{p}(X_i)}\right)}{\sum p(X_i)}$$

and $\hat{\alpha}_T(p,p)$ and $\hat{\alpha}_T(\hat{p},\hat{p})$ defined analogously. If the propensity score is unknown, only $\hat{\alpha}(\hat{p})$ and $\hat{\alpha}_T(\hat{p},\hat{p})$ are feasible. Hirano, Imbens, and Ridder (2003) show that $\hat{\alpha}(\hat{p})$ is efficient while $\hat{\alpha}(p)$ is not.[3][4] With known propensity score, $\hat{\alpha}_T(p,\hat{p})$ is efficient while $\hat{\alpha}_T(p,p)$ and $\hat{\alpha}_T(\hat{p},\hat{p})$ are not. With unknown propensity score, $\hat{\alpha}_T(\hat{p},\hat{p})$ is efficient. The efficiency bounds are the same as in Hahn (1998). Since the propensity score enters the average treatment effect α in the weighting estimator only through a single channel, while it enters α_T through two different channels, knowledge of the propensity score affects α differently than α_T.

An alternative explanation to why knowing the propensity score is useful for α_T but not for α is developed below. The basic insight is that knowing the propensity score helps in estimating the distribution $F_{X|D=1}$ whereas it does not help for estimating F_X. The central difference between α and α_T is that α is identified in (3) by weighting $m_1(x) - m_0(x)$ by the distribution

[3] Under certain regularity and smoothness conditions.

[4] Hirano, Imbens, and Ridder (2003) give an intuition why $\hat{\alpha}(\hat{p})$ but not $\hat{\alpha}(p)$ is efficient for estimating the average treatment effect α. For developing the intuition, they consider a single binary covariate $x \in \{a,b\}$ and examine $\frac{1}{n}\sum\frac{Y_iD_i}{\hat{p}(X_i)}$ as an estimator of $E[Y^1]$. This estimator can then be written as

$$\frac{1}{n}\sum_i \frac{Y_iD_i}{\hat{p}(X_i)} = \frac{n_a}{n}\bar{\mu}_a + \frac{n_b}{n}\bar{\mu}_b,$$

where n_a is the number of observations with $X_i = a$ and

$$\bar{\mu}_a = \frac{1}{n_a}\sum_{X_i=a}\frac{Y_iD_i}{\hat{p}(a)}$$

is an estimator of $E[Y|X=a,D=1]$. Dividing by the propensity score $\hat{p}(a)$ in $\bar{\mu}_a$ accounts for that only a fraction of the observations with $X=a$ are informative for estimating $E[Y|X=a,D=1]$. With a binary covariate, the propensity score can be estimated by the proportion of observations with $D_i=1$ for given X, i.e. $\hat{p}(a) = n_{1a}/n_a$ where n_{1a} is the number of observations with $D_i=1$ and $X_i=a$. The estimator $\bar{\mu}_a$ with the estimated propensity score, $\bar{\mu}_a = \frac{1}{n_{1a}}\sum_{X_i=a,D_i=1} Y_i$, is more efficient for estimating $E[Y|X=a,D=1]$ than the estimator with the true propensity score: $\bar{\mu}_a = \frac{1}{n_{1a}}\sum_{X_i=a,D_i=1} Y_i\frac{n_{1a}/n_a}{p(a)}$. In essence, weighting by the estimated instead of the true propensity score thus leads to more efficient estimators of $m_1(x)$ and $m_0(x)$.

F_X, whereas α_T is obtained by weighting by $F_{X|D=1}$.

Without knowledge of the propensity score, the distribution $F_{X|D=1}$ can be estimated by the empirical distribution function of X among the n_1 treated individuals. The X values of the non-treated observations contain no information about $F_{X|D=1}$. However, when the propensity score is known, in addition also the n_0 non-treated individuals become informative for estimating the distribution of X among the treated, because the distributions $F_{X|D=0}$ and $F_{X|D=1}$ are related through the propensity score by Bayes' theorem:

$$\frac{p(x)}{1-p(x)} = \frac{f_{X|D=1}(x)}{f_{X|D=0}(x)} \frac{P(D=1)}{P(D=0)}. \tag{8}$$

The propensity score enables using the X observations of one subpopulation (the non-treated) to estimate the distribution of X in a different subpopulation (the treated). Hence all $n_0 + n_1$ observations can be used to estimate $F_{X|D=1}$. On the other hand, the distribution F_X is in any case identified by all $n_0 + n_1$ observations, regardless of knowledge on the propensity score.

By this explanation it is also obvious why the variance bounds (5) and (7) coincide in the case of random treatment assignment ($p(x) = p$), and why (6) does not (Hahn 1998, Theorem 3). With treatment randomly assigned, the distribution of X is identical among the treated and the non-treated: $F_{X|D=1} = F_{X|D=0}$. If it is known that treatment was randomly assigned, estimation of α and α_T can proceed by estimating $m_1(x)$ and $m_0(x)$ separately from the respective subsamples and weighting $\hat{m}_1(x) - \hat{m}_0(x)$ by the distribution of X in the full sample. If it is unknown that treatment assignment was random, only the n_1 treated observations can be used for weighting $\hat{m}_1(x) - \hat{m}_0(x)$ to obtain α_T.

This is reflected in the variance bounds (5) to (7). Each expression consists of three terms: The first term in each bound captures the variance due to estimating $m_0(x)$, re-weighted by the density of X in the relevant population. This term vanishes at rate $\frac{1}{n_0}$ since only the n_0 non-treated observations are informative for estimating m_0. Analogously, the second term in each bound represents the variance due to estimating $m_1(x)$, which vanishes at rate $\frac{1}{n_1}$. The third term stems from estimating the distribution F_X in (5) and $F_{X|D=1}$ in (6) and (7). This term vanishes either at rate $\frac{1}{n_0+n_1}$ or $\frac{1}{n_1}$. If treatment assignment is random, the first term and also the second term is identical in all three bounds. The third term differs only in the scaling factor. In (5), for estimating α, the third term is scaled by $\frac{1}{n_0+n_1}$ since the full sample is used for estimating F_X. For α_T, the third term is also scaled by $\frac{1}{n_0+n_1}$ if it is known that assignment was random (7), and by $\frac{1}{n_1}$ if this is unknown (6). For example, if random assignment is with

probability $p = 0.5$, this variance component reduces by half when random-assignment is known, since the number of observations that are useful for estimating $F_{X|D=1}$ increases from n_1 to $2n_1$.

In the case of non-random assignment ($p(x) \neq p$), the bounds (6) and (7) for α_T still differ only in the third term. Since the first two terms are unaffected, the only channel through which knowledge of the propensity score can influence the variance of α_T is through the estimation of $F_{X|D=1}$. As in the case of random assignment, this third variance term is scaled by $\frac{1}{n_1}$ when the propensity score is unknown, and it is scaled by $\frac{1}{n_0+n_1}$ when it is known, because treated as well as non-treated observations contribute then to the identification of $F_{X|D=1}$. However, the non-treated observations are now less 'efficient' in estimating $F_{X|D=1}$, because the density mass of the non-treated observations may be concentrated in different regions than the mass of the treated observations. This is embodied in the correction term $f_{X|D=1}/f_X$ in the third term in (7).

This relationship between the propensity score and the estimation of $F_{X|D=1}$ becomes even more apparent when examining the variance bounds for estimating the distribution function $F_{X|D=1}(x)$. With unknown propensity score the scaled variance bound for estimating $F_{X|D=1}(x)$ is

$$\frac{1}{n_1} \cdot \underset{f_1}{E}\left[\left(1\left(X \leq x\right) - F_{X|D=1}(x)\right)^2\right], \tag{9}$$

and for known propensity score it is

$$\frac{1}{n_0+n_1} \cdot \underset{f_1}{E}\left[\frac{f_{X|D=1}(X)}{f_X(X)}\left[1\left(X \leq x\right) - F_{X|D=1}(x)\right]^2\right]. \tag{10}$$

(Proof available on request.)

These variance bounds have the same structure as the third terms in (6) and (7): If the propensity score is unknown, the variance (9) vanishes at rate $\frac{1}{n_1}$, whereas it vanishes at rate $\frac{1}{n_0+n_1}$ for known propensity score (10), because the non-treated observations assist in estimating $F_{X|D=1}$, with the same correction factor $f_{X|D=1}/f_X$ as as in (7).

Hence it is this additional information on $F_{X|D=1}$, and not any dimension reduction, that makes knowledge of the propensity score informative for the average treatment effect on the treated α_T and ancillary for the estimation of the average treatment effect α.

A Appendix - Derivation of the variance bounds (5) to (7)

The variance bounds in the notation of Hahn (1998) are the following:

The variance bound for α is

$$E\left[\frac{\sigma_1^2(X)}{p(X)}+\frac{\sigma_0^2(X)}{1-p(X)}+\left(m_1(X)-m_0(X)-\alpha\right)^2\right].$$

The variance bound for α_T with unknown propensity score is

$$\frac{1}{P^2}E\left[\sigma_1^2(X)p(X)+\frac{\sigma_0^2(X)p(X)^2}{1-p(X)}+p(X)\left(m_1(X)-m_0(X)-\alpha_T\right)^2\right],$$

where $P=P\left(D=1\right)=\lim\frac{n_1}{n_0+n_1}$ is the fraction of treated individuals.

The variance bound for α_T with known propensity score is

$$\frac{1}{P^2}E\left[\sigma_1^2(X)p(X)+\frac{\sigma_0^2(X)p(X)^2}{1-p(X)}+p^2(X)\left(m_1(X)-m_0(X)-\alpha_T\right)^2\right].$$

The expressions (5) to (7) follow from these bounds by dividing by the number of observations n_0+n_1, approximating P by $\frac{n_1}{n_0+n_1}$ and noting that

$$p(x)=f_{X|D=1}(x)P(D=1)/f_X(x)$$

by Bayes' theorem.

References

Hahn, J. (1998): "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica*, 66, 315–331.

Hirano, K., G. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189.

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

Rubin, D. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.