On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects

Author(s): Jinyong Hahn

Source: *Econometrica*, Mar., 1998, Vol. 66, No. 2 (Mar., 1998), pp. 315-331

Published by: The Econometric Society

Stable URL: https://www.jstor.org/stable/2998560

Econometrica, Vol. 66, No. 2 (March, 1998), 315–331

ON THE ROLE OF THE PROPENSITY SCORE IN EFFICIENT SEMIPARAMETRIC ESTIMATION OF AVERAGE TREATMENT EFFECTS

BY JINYONG HAHN[1]

In this paper, the role of the propensity score in the efficient estimation of average treatment effects is examined. Under the assumption that the treatment is ignorable given some observed characteristics, it is shown that the propensity score is ancillary for estimation of the average treatment effects. The propensity score is not ancillary for estimation of average treatment effects on the treated. It is suggested that the marginal value of the propensity score lies entirely in the "dimension reduction." Efficient semiparametric estimators of average treatment effects and average treatment effects on the treated are shown to take the form of relevant sample averages of the data completed by the nonparametric imputation method. It is shown that the projection on the propensity score is not necessary for efficient semiparametric estimation of average treatment effects on the treated even if the propensity score is known. An application to the experimental data reveals that conditioning on the propensity score may even result in a loss of efficiency.

KEYWORDS: Treatment effect, propensity score, semiparametric efficiency bound.

1. INTRODUCTION

THE CENTRAL PROBLEM IN EVALUATION STUDIES is that any potential outcome that program participants would have received in the absence of the program is not observed. Let D_i denote a dummy variable such that $D_i = 1$ when treatment is given to the ith individual, and $D_i = 0$ otherwise. Let Y_{0i} and Y_{1i} denote potential outcomes when $D_i = 0$ and $D_i = 1$, respectively. We can then say that the treatment *causes* the outcome variable of the ith individual to increase by $Y_{1i} - Y_{0i}$. Thus, $Y_{1i} - Y_{0i}$ can be called the treatment effect for the ith individual. Individual treatment effects cannot be observed, though, because we only observe D_i and $Y_i \equiv D_i Y_{1i} + (1 - D_i)Y_{0i}$. Because of this *missing data* problem, attention has been focused on some parameters which can summarize the impact of the program in a meaningful way. Usually, the parameter of interest is formulated in terms of conditional means, presumably because the case for social experimentation implicitly assumes that the mean gain from program participation is the primary object of interest. See Heckman (1992), Clements,

[1] Previous versions of this paper have been circulated under the title "Efficient Semiparametric Estimation of the Average Treatment Effects from the Experimental Data." I appreciate helpful comments from Joshua Angrist, Gary Chamberlain, Whitney Newey, James Powell, Petra Todd, Fannie Tseng, Yoon Jae Whang, a co-editor, two anonymous referees, and seminar participants at Lehigh University, Northwestern University, Penn State University, and the 1996 North American Summer Meeting of the Econometric Society. Guido Imbens inspired this research and deserves a lot more than a usual thank you. Financial support has been provided by the Institute for Economic Research and the Research Foundation of the University of Pennsylvania.

Heckman, and Smith (1993), and Heckman and Smith (1995) for related discussion. Thus, the *average treatment effects*

$$\beta \equiv E[Y_{1i} - Y_{0i}],$$

and the *average treatment effects on the treated*

$$\gamma \equiv E[Y_{1i} - Y_{0i} | D_i = 1],$$

have received a lot of attention in the literature. For example, Heckman, Ichimura, Smith, and Todd (1995) and Todd (1995) considered the mean impact of job training (for the program participants) on earnings. Angrist (1995a) considered the mean impact of military service (for veterans) on civilian earnings. In a related context, Imbens and Angrist (1994) reinterpreted the IV estimator as the estimator of some local average treatment effects.

Problems of sample selection are common in evaluation studies. Traditionally, two main approaches have been used in the literature to control for the bias: regression-based "control function" methods, predominantly used in econometrics, and "matching" methods, mainly used in statistics. A common feature of both approaches is that the conditional probability of program participation given some observed characteristics, often called the *propensity score*, plays a crucial role in controlling bias to obtain the estimator of the impact of the program. Many estimators proposed in the econometric literature for evaluating the impact of a social program rely on estimates of this propensity score to control for systematic differences between treatment and comparison groups. Examples include Heckman, Ichimura, Smith, and Todd (1995), Todd (1995), and Angrist (1995a, b). The critical role played by the propensity score in the literature is often motivated by Rosenbaum and Rubin's (1983, 1984) argument. They showed that if (i) there exists a variable X_i (which is always observed) such that D_i is *ignorable* given X_i, i.e., D_i and (Y_{0i}, Y_{1i}) are independent of each other given X_i; and (ii) $0 < P[D_i = 1 | X_i] < 1$ for all X_i; then D_i and (Y_{0i}, Y_{1i}) are independent of each other given the propensity score

$$p(x) \equiv P[D_i = 1 | X_i = x].$$

This in particular implies that $E[Y_{ji} | p(X_i)] = E[Y_i | D_i = 1, p(X_i)]$ for $j = 0, 1$, and hence,

$$\beta = E\{E[Y_i | D_i = 1, p(X_i)] - E[Y_i | D_i = 0, p(X_i)]\}.$$

Also observe that conditioning on X_i has the same effect:

$$\beta = E\{E[Y_i | D_i = 1, X_i] - E[Y_i | D_i = 0, X_i]\}.$$

These observations suggest that a consistent estimator of β may be constructed as a sample average of

$$\hat{E}[Y_i | D_i = 1, p(X_i)] - \hat{E}[Y_i | D_i = 0, p(X_i)] \quad \text{or}$$

$$\hat{E}[Y_i | D_i = 1, X_i] - \hat{E}[Y_i | D_i = 0, X_i],$$

where $\hat{E}[Y_i|D_i, p(X_i)]$ and $\hat{E}[Y_i|D_i, X_i]$ denote some nonparametric estimators of Y_i given $(D_i, p(X_i))$ and (D_i, X_i), respectively. Similar observation suggests that a consistent estimator of γ may be constructed as a sample average of the same object over the subsample where $D_i = 1$. But because conditioning on the *univariate* propensity score fully controls for the bias, and because the estimation of conditional distribution is more difficult when the dimension of the conditioning variable is large due to the curse of dimensionality, this "dimension reduction" has led many to focus on more reliable estimation of the propensity score.

The purpose of this paper is to consider the efficient estimation of β and γ when the treatment is ignorable given observed characteristics, and to examine the role of the propensity score from an efficiency point of view. This problem is not a standard parametric problem because the distribution of (Y_{0i}, Y_{1i}) is not parametrically specified. The semiparametric efficiency bound, introduced by Stein (1956), and developed by Begun, Hall, Huang, and Wellner (1983) and Bickel, Klaassen, Ritov, and Wellner (1993), among others, provides the semiparametric analog of the Cramer-Rao lower bound. See Newey (1990), for example, for a review on this subject. I calculate the semiparametric efficiency bounds under various assumptions and develop estimates whose asymptotic variances achieve these bounds. It turns out that the propensity score $p(x)$ is *ancillary* for the estimation of β: the efficiency bound for β under the knowledge of the propensity score is the same as the one without knowledge of the propensity score. The knowledge of the propensity score does decrease the asymptotic variance bound for γ, though. I provide a heuristic argument that this added information can be solely attributed to the "dimension reduction" feature of the propensity score.

I show that conditioning on the propensity score is not necessary and may even be harmful for the efficient estimation of β and γ. For the case where the propensity score is not known, I construct efficient estimators which take the forms of some relevant sample averages of the data completed by the nonparametric *imputation* method based on the nonparametric regression X_i. Even when the propensity score is known, in which case the asymptotic variance bound for γ is smaller when compared to the case where the propensity score is not known, it is found that the projection on the propensity score is not necessary to achieve the semiparametric efficiency bound. It is then found that conditioning on the propensity score results in a loss of efficiency in the case of *experimental data*.

2. EFFICIENCY BOUNDS

In this section, I calculate the semiparametric efficiency bounds of β and γ, and examine the role of the propensity score in efficient estimation. Knowledge of the propensity score is shown to add no additional information for estimation of β, and hence, the propensity score is *ancillary* for β. For the estimation of γ,

I argue that the marginal value of the propensity score is concentrated solely on the dimension reduction feature.

Assume that the treatment is ignorable give some covariates X_i. Our data set consists of (D_i, Y_i, X_i) $i = 1, \ldots, n$, where $Y_i \equiv D_i Y_{1i} + (1 - D_i) Y_{0i}$. Notice that we observe only one of Y_{0i} and Y_{1i}. Our objects of interest are the average treatment effects β and the average treatment effects on the treated γ. The asymptotic variance bounds for β and γ are calculated in the following theorem. The semiparametric asymptotic variance bound provides the semiparametric analog of the Cramer-Rao lower bound: no regular estimator sequence has a smaller asymptotic variance.

THEOREM 1: *Under the assumption that* $(Y_{0i}, Y_{1i}) \perp D_i | X_i$, *the asymptotic variance bounds for* β *and* γ *are*

$$E\left[\frac{\sigma_1^2(X_i)}{p(X_i)} + \frac{\sigma_0^2(X_i)}{1-p(X_i)} + (\beta(X_i) - \beta)^2\right],$$

and

$$E\left[\frac{p(X_i)\sigma_1^2(X_i)}{p^2} + \frac{p(X_i)^2\sigma_0^2(X_i)}{p^2(1-p(X_i))} + \frac{(\beta(X_i) - \gamma)^2 p(X_i)}{p^2}\right], \tag{1}$$

respectively, where

$$\begin{aligned}
&\beta_1(X_i) = E[Y_{1i}|X_i],\\
&\beta_0(X_i) = E[Y_{0i}|X_i],\\
&\beta(X_i) = \beta_1(X_i) - \beta_0(X_i),\\
&\sigma_1^2(X_i) = \mathrm{var}(Y_{1i}|X_i),\\
&\sigma_0^2(X_i) = \mathrm{var}(Y_{0i}|X_i),\\
&p = E[p(X_i)].
\end{aligned}$$

(Proof in Appendix.)

To examine the role of the propensity score in efficient estimation of β and γ, consider the hypothetical situation where the propensity score $p(\cdot)$ is known while maintaining the assumption that $(Y_{0i}, Y_{1i}) \perp D_i | X_i$. The reduction in the asymptotic variance bounds due to this additional assumption would then indicate the role of propensity score from the efficiency point of view.

THEOREM 2: *Assume that* $(Y_{0i}, Y_{1i}) \perp D_i | X_i$. *Furthermore, assume that the propensity score* $p(\cdot)$ *is known. The asymptotic variance bounds for* β *and* γ *are then equal to*

$$E\left[\frac{\sigma_1^2(X_i)}{p(X_i)} + \frac{\sigma_0^2(X_i)}{1-p(X_i)} + (\beta(X_i) - \beta)^2\right],$$

and

$$E\left[\frac{p(X_i)\sigma_1^2(X_i)}{p^2}+\frac{p(X_i)^2\sigma_0^2(X_i)}{p^2(1-p(X_i))}+\frac{(\beta(X_i)-\gamma)^2p^2(X_i)}{p^2}\right],$$

respectively.

(Proof in Appendix.)

A comparison of asymptotic variance bounds in Theorems 1 and 2 shows that the propensity score does not play any role in the estimation of β: the knowledge of the propensity score does not decrease the asymptotic variance bound. In this sense, the propensity score is *ancillary* for the estimation of β. On the other hand, knowledge of the propensity score clearly plays some role for the estimation of γ: it reduces the asymptotic variance bound by

$$E\left[\frac{(\beta(X_i)-\gamma)^2p(X_i)(1-p(X_i))}{p^2}\right], \tag{2}$$

which can be interpreted as the marginal value of the propensity score. Because the propensity score is not known in many realistic circumstances, this marginal value can only tell us the hypothetical marginal benefit.

One might also ask the marginal value of the "dimension reduction" due to the propensity score. To be more specific, suppose that X_i has a continuous distribution, and the support $\mathscr{X}$ of X_i is a union of the equivalence classes $\mathscr{X}_\alpha$ such that the propensity score is equal to α on each $\mathscr{X}_\alpha$. Suppose that we can identify such equivalence classes, although we do not know the propensity score itself. Observe that the knowledge of such equivalence classes amounts to the "dimension reduction" often associated with Rosenbaum and Rubin (1983, 1984). What is the marginal value of such knowledge? It is clear that knowledge of the equivalence classes should not add any information in estimation of β: the marginal value (in terms of asymptotic variance bound) of the propensity score itself was zero. For the estimation of γ, I do not yet know how to compute the efficiency bound under this generality. Instead, I consider a simple case which suggests that the marginal value of the propensity score entirely consists of the "dimension reduction." I consider an extreme example where the propensity score is constant over $\mathscr{X}$. This is the case of random treatment assignment. Observe that $\beta=\gamma$ in this case.

THEOREM 3: *Assume that* $(Y_{0i},Y_{1i})\perp D_i|X_i$. *Furthermore, assume that the propensity score* $p(\cdot)$ *is equal to some unknown constant* p. *The asymptotic variance bound for* $\beta=\gamma$ *is equal to*

$$E\left[\frac{\sigma_1^2(X_i)}{p}+\frac{\sigma_0^2(X_i)}{1-p}+(\beta(X_i)-\beta)^2\right].$$

(Proof in Appendix.)

Now, consider the variance bounds in Theorem 1 for the case where $p(\cdot) = p$. We can see that the bound for β equals

$$E\left[\frac{\sigma_1^2(X_i)}{p} + \frac{\sigma_0^2(X_i)}{1-p} + (\beta(X_i) - \beta)^2\right], \tag{3}$$

and that for γ equals

$$E\left[\frac{\sigma_1^2(X_i)}{p} + \frac{\sigma_0^2(X_i)}{1-p} + \frac{(\beta(X_i) - \beta)^2}{p}\right]. \tag{4}$$

These are the bounds if we do not know that the data are generated by the random treatment assignment. A comparison of (3) with the bound in Theorem 3 suggests that the bound for β does not change even if we know that the data are generated by random treatment assignment. This is hardly surprising when viewed against Theorem 2: the marginal value of the propensity score, which in this case is the knowledge that the data are generated by the random treatment assignment *and* the knowledge of the probability of treatment, is zero for β. Thus, the marginal value of the former knowledge should also be zero. Now, compare (4) with the bound in Theorem 3. The difference between them,

$$E\left[\frac{1-p}{p}(\beta(X_i) - \beta)^2\right],$$

indicates the marginal value (in the estimation of γ) of the knowledge that the data are generated by the random treatment assignment, or the marginal value of the dimension reduction. It turns out that this marginal value equals (2) when $p(\cdot) = p$. In other words, the marginal value (in the estimation of γ) of the knowledge of the propensity score entirely consists of the marginal value of dimension reduction.

3. EFFICIENT ESTIMATION

Having calculated efficiency bounds for β and γ, it is of interest to develop estimators which achieve these bounds. The estimators take the forms of some relevant sample averages from the data completed by the nonparametric imputation method based on the projection on X_i. I then consider estimation of γ when the propensity score is known, in which case the asymptotic variance bound is decreased, and argue that conditioning on the propensity score is not necessary for efficient estimation. Finally, I argue that conditioning on the propensity score may even be harmful in efficient estimation by considering the random treatment assignment where the propensity score is constant, under which case projection on the propensity score is equivalent to taking the marginal expectation.

Notice that the original data set contains some missing values because only one of Y_{1i} and Y_{0i} are observed. If both were observed, then the sample average of the difference $Y_{1i} - Y_{0i}$ would consistently estimate β, and the sample average of the difference $Y_{1i} - Y_{0i}$ over a subsample where $D_i = 1$ would consistently estimate γ.

The nonparametric imputation method *imputes* the missing values of Y_{1i} and Y_{0i} using their conditional expectation given X_i. In general, these conditional expectations are not identified. But the ignorability of D_i given X_i helps us to identify them. Because we have

$$E[D_i Y_i | X_i] = E[D_i Y_{1i} | X_i] = E[D_i | X_i] E[Y_{1i} | X_i] = E[D_i | X_i] \beta_1(X_i),$$

we can identify $\beta_1(X_i)$ by $E[D_i Y_i | X_i]/E[D_i | X_i]$. Similarly, we can identify $\beta_0(X_i)$. Even though $E[D_i Y_i | X_i]$, $E[(1 - D_i) Y_i | X_i]$, and $E[D_i | X_i]$ are not exactly known in the sample, we can use various nonparametric regression techniques to consistently estimate them. Let $\hat{E}[D_i Y_i | X_i]$, $\hat{E}[(1 - D_i) Y_i | X_i]$, and $\hat{E}[D_i | X_i]$ denote the corresponding nonparametric regression estimators. We can then fill in the missing values of Y_{1i} and Y_{0i} by

$$\hat{\beta}_1(X_i) \equiv \frac{\hat{E}[D_i Y_i | X_i]}{\hat{E}[D_i | X_i]} \quad \text{and} \quad \hat{\beta}_0(X_i) \equiv \frac{\hat{E}[(1 - D_i) Y_i | X_i]}{1 - \hat{E}[D_i | X_i]},$$

respectively. With this "nonparametric imputation," we have a "complete" data set, where we "observe" $\hat{Y}_{1i} \equiv D_i Y_i + (1 - D_i) \hat{\beta}_1(X_i)$ under "treatment," and $\hat{Y}_{0i} \equiv (1 - D_i) Y_i + D_i \hat{\beta}_0(X_i)$ under "control." Our "complete" data set thus consists of $(\hat{Y}_{1i}, \hat{Y}_{0i}, D_i, X_i)$, $i = 1, \ldots, n$, and we can estimate β and γ by

$$\hat{\beta} = \frac{1}{n} \sum_i \left(\hat{Y}_{1i} - \hat{Y}_{0i} \right) \quad \text{and} \quad \hat{\gamma} = \frac{(1/n) \Sigma_i D_i \cdot \left(\hat{Y}_{1i} - \hat{Y}_{0i} \right)}{(1/n) \Sigma_i D_i}.$$

Notice that we may consistently estimate β and γ by the sample averages of $\beta_1(X_i) - \beta_0(X_i)$ over the entire sample and over the subsample where $D_i = 1$, respectively, if $(\beta_1(X_i), \beta_0(X_i))$ were observed. Because they are not, we may use

$$\tilde{\beta} = \frac{1}{n} \sum_i \left(\hat{\beta}_1(X_i) - \hat{\beta}_0(X_i) \right) \quad \text{and}$$

$$\tilde{\gamma} = \frac{(1/n) \Sigma_i D_i \cdot \left(\hat{\beta}_1(X_i) - \hat{\beta}_0(X_i) \right)}{(1/n) \Sigma_i D_i}$$

instead. Because these estimators are based on the data set where the missing values of $\beta_1(X_i)$ and $\beta_0(X_i)$ are imputed by the nonparametric regression method, they can also be interpreted as nonparametric imputation based estimators.

If the estimators are $\sqrt{n}$-consistent and asymptotically normal, we can use Newey's (1994) argument to show that the asymptotic variances of $\sqrt{n}(\hat{\beta}-\beta)$ and $\sqrt{n}(\tilde{\beta}-\beta)$ are equal to each other and equal to

$$E\left[\frac{\sigma_1^2(X_i)}{p(X_i)}+\frac{\sigma_0^2(X_i)}{1-p(X_i)}+(\beta(X_i)-\beta)^2\right].$$

Similarly, we can show that the asymptotic variances of $\sqrt{n}(\hat{\gamma}-\gamma)$ and $\sqrt{n}(\tilde{\gamma}-\gamma)$ are equal to each other and equal to

$$E\left[\frac{p(X_i)\sigma_1^2(X_i)}{p^2}+\frac{p(X_i)^2\sigma_0^2(X_i)}{p^2(1-p(X_i))}+\frac{(\beta(X_i)-\gamma)^2p(X_i)}{p^2}\right].$$

From Theorem 1, it follows that $\hat{\beta}$ and $\tilde{\beta}$ are efficient for β, and $\hat{\gamma}$ and $\tilde{\gamma}$ are efficient for γ.

PROPOSITION 4: *Assume that* $(Y_{0i},Y_{1i})\perp D_i|X_i$. *Then,* $\hat{\beta}$ *and* $\tilde{\beta}$ *are efficient semiparametric estimators for* β, *and* $\hat{\gamma}$ *and* $\tilde{\gamma}$ *are efficient semiparametric estimators for* γ.

(Proof in Appendix.)

Proposition 4 does not provide any regularity conditions. Neither does it tell us any specific nonparametric regression estimation to be used. In the case where X_i has a finite support, it is trivial to fill the gap. Notice that, if we can take

$$\hat{E}[D_iY_i|X_i=x]=\frac{\Sigma_iD_iY_i\cdot 1(X_i=x)}{\Sigma_i1(X_i=x)},$$

$$\hat{E}[(1-D_i)Y_i|X_i=x]=\frac{\Sigma_i(1-D_i)Y_i\cdot 1(X_i=x)}{\Sigma_i1(X_i=x)},$$

and

$$\hat{E}[D_i|X_i=x]=\frac{\Sigma_iD_i\cdot 1(X_i=x)}{\Sigma_i1(X_i=x)},$$

the usual argument will establish the asymptotic distribution.

THEOREM 5: *Assume that* $(Y_{0i},Y_{1i})\perp D_i|X_i$. *Furthermore, assume that* X_i *has a known finite support. Then,* $\hat{\beta}$ *and* $\tilde{\beta}$ *are efficient semiparametric estimators for* β, *and* $\hat{\gamma}$ *and* $\tilde{\gamma}$ *are efficient semiparametric estimators for* γ.

(Proof omitted.)

When X_i has a continuous distribution, we can choose a variety of nonparametric estimators. When these estimators are computed by the series estimation, we can find some regularity conditions under which the nonparametric imputa-

tion based estimators are asymptotically normal.[2] To obtain a series estimator of $E[Y_i|X_i=x]$, for example, we take

$$p^K(x)=(p_{1K}(x),\ldots,p_{KK}(x))',$$
$$y=(Y_1,\ldots,Y_n)',$$
$$p^K=\left[p^K(X_1),\ldots,p^K(X_n)\right]', \quad \text{and}$$
$$\hat{E}[Y_i|X_i=x]=p^K(x)'\hat{\pi}, \quad \hat{\pi}=(p^{K\prime}p^K)^{-}p^{K\prime}y.$$

Let

$$u_{1i}=D_iY_i-p(X_i)\beta_1(X_i),$$
$$u_{2i}=(1-D_i)Y_i-(1-p(X_i))\beta_0(X_i),$$
$$u_{3i}=D_i-p(X_i).$$

THEOREM 6: *Assume that* $(Y_{0i},Y_{1i})\perp D_i|X_i$. *Furthermore, assume that*:
(i) $E[u_{ki}^2|X_i]$ *is bounded for* $k=1,2,3$;
(ii) *the support of* X_i *is a Cartesian product of compact intervals* $\prod_{j=1}^{r}[x_{lj},x_{uj}]$;
(iii) *the density of* X_i *is bounded below by* $C\prod_{j=1}^{r}[(x-x_{lj})(x_{uj}-x)]^{\nu}$ *for some* $C>0$, *and* $p_{kK}(x)$ *are the products of polynomials that are orthonormal with respect to* $\prod_{j=1}^{r}[(x-x_{lj})(x_{uj}-x)]^{\nu}$;
(iv) $p(x)$, $\beta_1(x)$, $\beta_0(x)$ *are continuously differentiable of all orders*;
(v) $K=n^{\varepsilon}$ *for some* $\varepsilon>0$, *and* $K^{7+4\nu}/n\to 0$.
Then, $\hat{\beta}$ *and* $\tilde{\beta}$ *are efficient semiparametric estimators for* β, *and* $\hat{\gamma}$ *and* $\tilde{\gamma}$ *are efficient semiparametric estimators for* γ.

(Proof in Appendix.)

It seems that imputation is unavoidable even for the experimental data case. Consider regressing $Y_i-\hat{E}[Y_i|X_i]$ on $D_i-\hat{E}[D_i|X_i]$, where $\hat{E}[Y_i|X_i]$ and $\hat{E}[D_i|X_i]$ are some nonparametric estimators of $E[Y_i|X_i]$ and $E[D_i|X_i]$. Call this estimator $\hat{\beta}_{SL}$. The probability limit of $\hat{\beta}_{SL}$ equals

$$\frac{E[(Y_i-E[Y_i|X_i])(D_i-E[D_i|X_i])]}{E\left[(D_i-E[D_i|X_i])^2\right]}=\beta.$$

This is an estimator due to Robinson (1988) for the partially linear semiparametric regression model. The asymptotic variance $\text{var}_a(\hat{\beta}_{SL})$ of $\hat{\beta}_{SL}$, computed using Newey's (1994) machinery, equals

$$E\left[\frac{\sigma_1^2(X_i)}{p}\right]+E\left[\frac{\sigma_0^2(X_i)}{1-p}\right]+\left(\frac{1}{p(1-p)}-3\right)\cdot\text{var}(\beta(X_i)).$$

[2] In practice, it can be extremely difficult to construct a series $p_{K1}(\cdot),\ldots,P_{KK}(\cdot)$ such that Condition 3 in Theorem 6 is satisfied. This condition should thus be viewed as a "high level" assumption. I thank an anonymous referee who pointed it out.

Because

$$\text{var}_a(\hat{\beta}_{SL}) - \text{var}_a(\hat{\beta}) = \left(\frac{1}{p(1-p)} - 4\right) \cdot \text{var}(\beta(X_i)) \geq 0,$$

$\hat{\beta}_{SL}$ is not an efficient estimator.

It was seen that the propensity score is ancillary for estimation of β. On the other hand, the propensity score is not ancillary for γ, but its value is solely concentrated on the "dimension reduction" feature. Thus, it is of interest to ask whether the projection on the propensity score instead of X_i is *necessary* to attain the efficiency bound in the estimation of γ. Although the propensity score is unknown in many realistic situations, many estimators in the literature use the nonparametric regression estimation of some conditional expectation on the propensity score to exploit the "dimension reduction" feature of the propensity score. I argue that an efficient estimator for γ does not have to use the projection on the propensity score even when the propensity score is known. Because the sole value of the propensity score seems to be its "dimension reduction" feature, it can be inferred that the "dimension reduction" does not imply the necessity of the projection on the propensity score.

PROPOSITION 7: *Assume that* $(Y_{0i}, Y_{1i}) \perp D_i | X_i$. *Furthermore, assume that the propensity score* $p(\cdot)$ *is known. Then, the estimator*

$$\frac{1}{n}\sum_i p(X_i) \cdot \left(\frac{\hat{E}[D_i Y_i | X_i]}{\hat{E}[D_i | X_i]} - \frac{\hat{E}[(1-D_i) Y_i | X_i]}{1 - \hat{E}[D_i | X_i]}\right) \Big/ \frac{1}{n}\sum_i p(X_i)$$

is efficient for the estimation of γ.

(Proof in Appendix.)

I now argue that the projection on the propensity score may even be harmful for the estimation of $\beta = \gamma$ by considering the experimental data case. As for efficient estimation, we would want to use the estimator which is efficient when the propensity score is known, because the marginal role of the propensity score is purely contained in the "dimension reduction." Observe that $\tilde{\beta}$, which is an efficient estimator for β with or without the knowledge of the propensity score, is still efficient for β. As for the estimation of γ with the knowledge of the propensity score, we observe that the estimator developed in Proposition 7 reduces to $\tilde{\beta}$ when the propensity score is constant. Note that we would not want to use $\tilde{\gamma}$, because it is efficient only when the propensity score is unknown and does not make use of the "dimension reduction."

Now, consider the projection on the propensity score. Because the propensity score is a constant, the projection on the propensity is equivalent to the marginal expectation. Thus, the idea of conditioning on the propensity score leads us to consider the difference of the sample averages as our estimator. Call such an estimator $\hat{\beta}_{OLS}$. It can easily be shown that the asymptotic variance

$\text{var}_a(\hat{\beta}_{OLS})$ of $\hat{\beta}_{OLS}$ equals

$$E\left[\frac{\sigma_1^2(X_i)}{p}\right]+E\left[\frac{\sigma_0^2(X_i)}{1-p}\right]+\frac{\text{var}(\beta_1(X_i))}{p}+\frac{\text{var}(\beta_0(X_i))}{1-p}.$$

Comparing this with the asymptotic variance of $\hat{\beta}$ or $\tilde{\beta}$, we find that

$$\text{var}_a(\hat{\beta}_{OLS})-\text{var}_a(\hat{\beta})=\text{var}\left(\sqrt{\frac{1-p}{p}}\,\beta_1(X_i)+\sqrt{\frac{p}{1-p}}\,\beta_0(X_i)\right)$$
$$\geq 0,$$

and thus $\hat{\beta}_{OLS}$ is not an efficient estimator.

Comparison of the asymptotic variance of $\hat{\beta}$ (or $\tilde{\beta}$) and $\hat{\gamma}$ (or $\tilde{\gamma}$) suggests that knowledge of the propensity score can help in a subtler way than the mere projection on the propensity score. Consider again the experimental data case where $\beta=\gamma$. We observe that $\hat{\beta}$ is efficient whereas $\hat{\gamma}$ is not. This is due to the fact that we essentially throw away observations with $D_i=0$ in the "complete" data analysis. With the knowledge that $\beta=\gamma$, we can avoid this loss of information. It is natural to conjecture that this observation would generalize to the situation where the propensity score is not necessarily constant. Suppose that $p(X_i)=p_0$ for $X_i\in\mathscr{X}_0$ and $p(X_i)=p_1$ for $X_i\in\mathscr{X}_1$, where $p_0\neq p_1$ and $\mathscr{X}_0\cup\mathscr{X}_1=\mathscr{X}$. Suppose that we classify the observations according to the known propensity score. On each subgroup where the propensity score is equal to p_0, say, we can efficiently estimate

$$E[Y_{1i}-Y_{0i}|D_i=1,p(X_i)=p_0]=E[Y_{1i}-Y_{0i}|p(X_i)=p_0]$$

by $\hat{\beta}$.

Dept. of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104-6297, U.S.A.; hahn@econ.sas.upenn.edu

Manuscript received February, 1995; final revision received April, 1997.

TECHNICAL APPENDIX

Proof of Theorem 1

In calculating the variance bounds of β and γ, I follow the approach of Bickel, Klaassen, Ritov, and Wellner (1993, Section 3.3). First, the tangent space is characterized. The density of (Y_0,Y_1,D,X) (with respect to some σ-finite measure) is given by

$$\bar{q}(y_0,y_1,d,x)=f(y_0,y_1|x)p(x)^d(1-p(x))^{1-d}f(x),$$

where $f(y_0,y_1|x)$ and $f(x)$ denote the conditional distribution of (Y_0,Y_1) given X, and the marginal distribution of X, respectively. The density of (Y,D,X) is then equal to

$$q(y,d,x)=[f_1(y|x)p(x)]^d[f_0(y|x)(1-p(x))]^{1-d}f(x),$$

where $f_1(\cdot|x) = \int f(y_0, \cdot|x)\,dy_0$, and $f_0(\cdot|x) = \int f(\cdot, y_1|x)\,dy_1$. Consider a regular parametric submodel

$$[f_1(y|x,\theta)p(x,\theta)]^d[f_0(y|x,\theta)(1-p(x,\theta))]^{1-d}f(x,\theta), \tag{5}$$

which equals $q(y,d,x)$ when $\theta = \theta_0$. The corresponding score is given by

$$\begin{aligned} s(d,y,x|\theta) \equiv{}& d\cdot s_1(y|x,\theta) + (1-d)\cdot s_0(y|x,\theta) \\ &+ \frac{d-p(x,\theta)}{p(x,\theta)(1-p(x,\theta))}\cdot \dot{p}(x,\theta) + t(x,\theta), \end{aligned} \tag{6}$$

where

$$\begin{aligned} s_1(y|x,\theta) &= \frac{d}{d\theta}\log f_1(Y|X,\theta), \\ s_0(y|x,\theta) &= \frac{d}{d\theta}\log f_0(Y|X,\theta), \\ \dot{p}(x,\theta) &= \frac{d}{d\theta}p(x,\theta), \\ t(x,\theta) &= \frac{d}{d\theta}\log f(X,\theta). \end{aligned}$$

From (6), we obtain

$$\mathscr{S} = \{d\cdot s_1(y|x) + (1-d)\cdot s_0(y|x) + a(x)\cdot(d-p(x)) + t(x)\}$$

as the tangent space of this model, where $\int s_j(y|x)f_j(y|x)\,dy = 0\ \forall x, j = 0,1$, $\int t(x)f(x)\,dx = 0$, and $a(x)$ is any square-integrable measurable function of x.

Now, the average treatment effect is shown to be pathwise differentiable. For the parametric submodel under consideration, we find that

$$\beta(\theta) = \iint y f_1(y|x,\theta)f(x,\theta)\,dy\,dx - \iint y f_0(y|x,\theta)f(x,\theta)\,dy\,dx,$$

and

$$\gamma(\theta) = \frac{\int\int y p(x,\theta)f_1(y|x,\theta)f(x,\theta)\,dy\,dx - \int\int y p(x,\theta)f_0(y|x,\theta)f(x,\theta)\,dy\,dx}{\int p(x,\theta)f(x,\theta)\,dx}.$$

Thus,

$$\begin{aligned} \frac{\partial\beta(\theta_0)}{\partial\theta} ={}& \iint y s_1(y|x,\theta_0)f_1(y|x)f(x)\,dy\,dx + \int \beta_1(x)t(x,\theta_0)f(x)\,dx \\ &- \iint y s_0(y|x,\theta_0)f_0(y|x)f(x)\,dy\,dx - \int \beta_0(x)t(x,\theta_0)f(x)\,dx, \end{aligned}$$

and

$$\begin{aligned} \frac{\partial\gamma(\theta_0)}{\partial\theta} ={}& \frac{\int\int y p(x) s_1(y|x,\theta_0)f_1(y|x)f(x)\,dy\,dx}{p} - \frac{\int\int y p(x) s_0(y|x,\theta_0)f_0(y|x)f(x)\,dy\,dx}{p} \\ &+ \frac{\int(\beta(x)-\gamma)\dot{p}(x,\theta_0)f(x)\,dx}{p} + \frac{\int(\beta(x)-\gamma)p(x)t(x,\theta_0)f(x)\,dx}{p}. \end{aligned}$$

Let

$$\begin{aligned} F_\beta(Y,D,X) &= \frac{D}{p(X)}\cdot(Y-\beta_1(X)) - \frac{1-D}{1-p(X)}\cdot(Y-\beta_0(X)) + \beta(X) - \beta, \\ F_\gamma(Y,D,X) &= \frac{D}{p}\cdot(Y-\beta_1(X)) - \frac{1-D}{p}\frac{p(X)}{1-p(X)}\cdot(Y-\beta_0(X)) \\ &\quad + \frac{\beta(X)-\gamma}{p}\cdot(D-p(X)) + \frac{\beta(X)-\gamma}{p}\cdot p(X). \end{aligned}$$

For the parametric submodel whose score is given by (6), we have

$$\frac{\partial\beta(\theta_0)}{\partial\theta} = E[F_\beta(Y,D,X)\cdot s(D,Y,X|\theta_0)],$$

$$\frac{\partial\gamma(\theta_0)}{\partial\theta} = E[F_\gamma(Y,D,X)\cdot s(D,Y,X|\theta_0)],$$

from which we conclude that β and γ are differentiable parameters.

The variance bounds are the expected squares of the projections of F_β and F_γ on $\mathscr{S}$. Because $F_\beta, F_\gamma \in \mathscr{S}$, the projections on $\mathscr{S}$ are themselves, and the variance bounds are the expected squares of the projections of F_β and F_γ.

Proof of Theorem 2

Now the parametric submodel under consideration changes from (5) to

$$[f_1(y|x,\theta)p(x)]^d[f_0(y|x,\theta)(1-p(x))]^{1-d}f(x,\theta).$$

Because the score now equals

$$s(d,y,x|\theta) \equiv d\cdot s_1(y|x,\theta) + (1-d)\cdot s_0(y|x,\theta) + t(x,\theta),$$

the tangent space changes to

$$\mathscr{S} = \{d\cdot s_1(y|x) + (1-d)\cdot s_0(y|x) + t(x)\}.$$

We find that

$$\frac{\partial\beta(\theta_0)}{\partial\theta} = E[F_\beta(Y,D,X)\cdot s(D,Y,X|\theta)],$$

and

$$\frac{\partial\gamma(\theta_0)}{\partial\theta} = E[F_\gamma(Y,D,X)\cdot s(D,Y,X|\theta)],$$

for

$$F_\beta(Y,D,X) = \frac{D}{p(X)}(Y-\beta_1(X)) - \frac{1-D}{1-p(X)}(Y-\beta_0(X)) + \beta(X) - \beta,$$

$$F_\gamma(Y,D,X) = \frac{D}{p}\cdot(Y-\beta_1(X)) - \frac{1-D}{p}\frac{p(X)}{1-p(X)}\cdot(Y-\beta_0(X)) + \frac{\beta(X)-\gamma}{p}\cdot p(X).$$

Because $F_\beta, F_\gamma \in \mathscr{S}$ again, the variance bounds are the expected squares of the projections of F_β and F_γ.

Proof of Theorem 3

The regular parametric submodel under consideration now changes to

$$q(y,d,x) = [f_1(y|x,\theta)p(\theta)]^d[f_0(y|x,\theta)(1-p(\theta))]^{1-d}f(x,\theta).$$

The tangent space is thus equal to

$$\mathscr{S} = \{d\cdot s_1(y|x) + (1-d)\cdot s_0(y|x) + a\cdot(d-p) + t(x)\},$$

where a is real number. We find that

$$\frac{\partial\beta(\theta_0)}{\partial\theta} = E[F_\beta(Y,D,X)\cdot s(D,Y,X|\theta)]$$

for

$$F_\beta(Y,D,X)=\frac{D}{p}\cdot(Y-\beta_1(X))-\frac{1-D}{1-p}\cdot(Y-\beta_0(X))+\beta(X)-\beta.$$

Because $F_\beta\in\mathscr{S}$, we obtain $E[F_\beta^2]$ as the variance bound for β.

Proof of Proposition 4

For the general case, we can use Newey's (1994) argument for a heuristic proof. I only consider $\tilde{\beta}$ and $\tilde{\gamma}$. The asymptotic variance of $\hat{\beta}$ and $\hat{\gamma}$ can be similarly obtained. First, consider $\tilde{\beta}$. This estimator takes the form $(1/n)\Sigma_i m(Z_i,\hat{h}_1,\hat{h}_2,\hat{h}_3)$, where

$$h_{01}(x)=E[D_iY_i|X_i=x],$$
$$h_{02}(x)=E[(1-D_i)Y_i|X_i=x],$$
$$h_{03}(x)=E[D_i|X_i=x],$$

and $\hat{h}_1,\hat{h}_2,\hat{h}_3$ are their estimators. Z_i denotes the observation for individual i. Let $h_1(\theta),h_2(\theta),h_3(\theta)$ denote the corresponding functions under some parametric submodel which equals the true model at $\theta=\theta_0$. Because $m(X_i,h_1,h_2,h_3)$ depends on h_1,h_2,h_3 only through their values $h_1(X_i),h_2(X_i),h_3(X_i)$, it follows that

$$\frac{\partial}{\partial\theta}E[m(X_i,h_1(\theta),h_2(\theta),h_3(\theta))]=\frac{\partial}{\partial\theta}E\left[\sum_{j=1}^{3}h_j(\theta)\cdot\delta_j(X_i)\right],$$

for

$$\delta_j(x)=\frac{\partial}{\partial h_j}m(x,h_1(\theta),h_2(\theta),h_3(\theta))|_{\theta=\theta_0}.$$

Notice that

$$\delta_1(x)=\frac{1}{p(x)},$$
$$\delta_2(x)=-\frac{1}{1-p(x)},\qquad\text{and}$$
$$\delta_3(x)=-\frac{\beta_1(x)}{p(x)}-\frac{\beta_0(x)}{(1-p(x))}.$$

Newey's (1994) Proposition 4 then suggests that the above estimator has the asymptotic influence function equal to

$$\frac{D_i(Y_{1i}-\beta_1(X_i))}{p(X_i)}-\frac{(1-D_i)(Y_{0i}-\beta_0(X_i))}{1-p(X_i)}+(\beta_1(X_i)-\beta_0(X_i)-(\beta_1-\beta_0)),$$

so that its asymptotic variance equals the efficiency bound. For $\tilde{\gamma}$, Newey's (1994) Proposition 4 suggests that the numerator has the asymptotic influence function equal to

$$D_i(Y_{1i}-\beta_1(X_i))-\frac{p(X_i)}{1-p(X_i)}(1-D_i)(Y_i-\beta_0(X_i))+D_i\cdot(\beta_1(X_i)-\beta_0(X_i))-p\gamma.$$

Because the denominator has the asymptotic influence function equal to D_i-p, the asymptotic influence function of the ratio can be computed by the delta method as equal to

$$\frac{D_i}{p}(Y_{1i}-\beta_1(X_i))-\frac{1-D_i}{p}\frac{p(X_i)}{1-p(X_i)}(Y_{0i}-\beta_0(X_i))+\frac{D_i}{p}(\beta_1(X_i)-\beta_0(X_i)-\gamma),$$

and the asymptotic variance of this estimator equals the efficiency bound.

PROOF OF THEOREM 6

I only consider $\tilde{\beta}$. The rest can be shown similarly. First, I introduce some notation. For a vector of function $f(x)$, let

$$|f(x)|_d = \max_{|\lambda| \le d} \max_{x \in \mathscr{X}} |\partial^\lambda f(x)|,$$

where $\mathscr{X}$ is the support of X_i. We also let

$$\zeta_d(K) = \sup_{|\lambda| = d, x \in \mathscr{X}} |\partial^\lambda p^K(x)|.$$

In what follows, C denotes a generic constant.

To show that the series estimation based imputation estimator is efficient, it suffices to show that the following conditions taken from Newey (1994) are satisfied:

1. $E[u_{ki}^2 | X_i]$ is bounded for $k = 1, 2, 3$.

2. (i) The smallest eigenvalue of $E[p^K(X_i)p^K(X_i)']$ is bounded away from zero uniformly in K; (ii) $p^K(x)$ is a subvector of $p^{K+1}(x)$ for all K; (iii) for each K, there is some nonzero $\bar{\pi}$ such that $\bar{\pi}' p^K(x)$ is a nonzero constant on the support of X_i.

3. For each nonnegative integer d, if $|g(x)|_d$ is finite then there are constants $C, \alpha_d > 0$ such that for all K there is π with $|g(x) - p^K(x)'\pi|_d \le CK^{-\alpha_d}$.

4. (i) There is a function $D(z, h)$ linear in h such that for all h with $|h - h_0|_0$ sufficiently small,

$$|m(z, h) - m(z, h_0) - D(z, h - h_0)|_0 \le b(z)|h - h_0|_0^2;$$

(ii) $$E[b(Z_i)]\zeta_d(K)[(K/n)^{1/2} + K^{-\alpha}] \to 0$$

and

$$E[b(Z_0)]\sqrt{n}\,\zeta_d(K)^2[K/n + K^{-2\alpha}] \to 0.$$

5. There is $b(z)$, $d > 0$ such that $E[b(Z_i)^2] < \infty$, $|D(z, g)|_0 \le b(z)|g|_d$, and

$$(\Sigma_{k=1}^K |p_{kK}|_d^2)^{1/2} \times [(K/n)^{1/2} + K^{-\alpha}] \to 0.$$

6. (i) There is $\delta(x)$ such that $E[D(z, g)] = E[\delta(X_i)g(X_i)]$ for all g; (ii) for each K there are π_K such that

$$n \cdot E[|\delta(X_i) - \xi_K p^K(X_i)|_0^2] \cdot E[|g(X_i) - \pi_K p^K(X_i)|_0^2] \to 0,$$

$$\zeta_0(K)^4 K/n \to 0, \quad \zeta_0(K)^2 E[|g(X_i) - \pi_K p^K(X_i)|_0^2] \to 0,$$

$$E[|\delta(X_i) - \xi_K p^K(X_i)|_0^2] \to 0.$$

Except for condition 4 (i), these conditions are equivalent to Newey's (1994) Assumptions 6.1–6.6. Newey's Assumption 6.4 (i) is replaced by his Assumption 5.1 (i), because the former is stronger than necessary for obtaining asymptotic normality: he makes the assumption to make the proof of the consistency of the asymptotic variance estimator easier. Newey's (1994, Theorem 6.1) shows that the semiparametric estimator is asymptotically normal under these conditions.

I will verify that these conditions are satisfied. Condition 1 is satisfied by hypothesis. By Lemma A.15 of Newey (1995), condition 2 is satisfied by $p_{kK}(x)$ equal to the products of polynomials that are orthonormal with respect to $\prod_{j=1}^{r}[(x - x_{lj})(x_{uj} - x)]^{\nu}$ with $\zeta_d(K) \le CK^{1+\nu+2d}$ and $|p_{kK}(x)|_d \le CK^{.5+\nu+2d}$. By Lemma. A.13 of Newey (1995), condition 3 is satisfied with α_d equal to any positive constant. Because $p(x)$ is bounded away from 0 and 1, for h sufficiently close to h_0, condition 4(i) is satisfied with

$$D(z, h - h_0) = \frac{h_1 - h_{01}}{h_{03}} - \frac{h_2 - h_{02}}{h_{03}} - \left[\frac{h_{01}}{h_{03}^2} + \frac{h_{02}}{(1 - h_{03})^2}\right](h_3 - h_{03})$$

and

$$b(z) = C(1 + |\beta_1(x)| + |\beta_0(x)|).$$

Also observe that

$$\zeta_0(K)\cdot\left[\left(\frac{K}{n}\right)^{1/2} + K^{-\alpha}\right] = O\left(\left(\frac{K^{3+\nu}}{n}\right)^{1/2} + K^{1+\partial-\alpha}\right) \to 0, \qquad \text{and}$$

$$\sqrt{n}\,\zeta_0(K)^2\left[\frac{K}{n} + K^{-2\alpha}\right] = O\left(\frac{K^{3+\nu}}{n^{1/2}} + K^{2+2\nu-2\alpha}\right) \to 0$$

if we take $\alpha > 1 + \nu$. For condition 5, note that

$$|D(z,g)| \le C(1 + |\beta_1(x)| + |\beta_0(x)|)|g|_1, \qquad \text{and}$$

$$\left(\sum_{k=1}^{K} |p_{kK}|_1^2\right)^{1/2}\left[\left(\frac{K}{n}\right)^{1/2} + K^{-\alpha}\right] = CK^{2.5+\nu}\cdot\left[\left(\frac{K}{n}\right)^{1/2} + K^{-\alpha}\right]$$

$$= O\left[\left(\frac{K^{7+2\nu}}{n^{1/2}}\right)^{1/2} + K^{1+\nu-\alpha}\right] \to 0.$$

For condition 6, note that

$$E[D(Z_i,g)] = E[\delta(X_i)g(X_i)] \qquad \text{for}$$

$$\delta(x) = \left(\frac{1}{p(x)}, -\frac{1}{1-p(x)}, -\frac{\beta_1(x)}{p(x)} - \frac{\beta_0(x)}{1-p(x)}\right)'.$$

Because $\delta(x)$ is continuously differentiable of all orders, by Lorentz (1986, Theorem 8), there exist π_K and ξ_K such that

$$E[|\delta(X_i) - \xi_K p^K(X_i)|_0^2] = O(K^{-2\alpha}) \to 0 \quad \text{and}$$

$$E[|g(X_i) - \pi_K p^K(X_i)|_0^2] = O(K^{-2\alpha}) \to 0.$$

We have

$$n\cdot E[|\delta(X_i) - \xi_K p^K(X_i)|_0^2]\cdot E[|g(X_i) - \pi_K p^K(X_i)|_0^2] = O(nK^{-4\alpha}) \to 0,$$

$$\zeta_0(K)^4\frac{K}{n} = O\left(\frac{K^{5+4\nu}}{n}\right) \to 0,$$

$$\zeta_0(K)^2 E[|g(X_i) - \pi_K p^K(X_i)|_0^2] = O(K^{2+2\nu-2\alpha}) \to 0$$

if we take α sufficiently large.

Proof of Proposition 7

Newey's (1994) Proposition 4 suggests that the numerator has the asymptotic influence function equal to

$$D_i(Y_{1i} - \beta_1(X_i)) - \frac{p(X_i)}{1-p(X_i)}(1-D_i)(Y_i - \beta_0(X_i)) + p(X_i)\cdot(\beta_1(X_i) - \beta_0(X_i)) - p\gamma.$$

Because the denominator has the asymptotic influence function equal to $p(X_i) - p$, we can use the delta method to obtain the asymptotic influence function,

$$\frac{D_i}{p}(Y_{1i} - \beta_1(X_i)) - \frac{1-D_i}{p}\frac{p(X_i)}{1-p(X_i)}(Y_{0i} - \beta_0(X_i)) + \frac{p(X_i)}{p}(\beta_1(X_i) - \beta_0(X_i) - \gamma).$$

The asymptotic variance of this estimator equals the efficiency bound.

REFERENCES

ANGRIST, J. (1995a): "Using Social Security Data on Military Applicants to Estimate the Effect of Voluntary Military Service on Earnings," National Bureau of Economic Research, Working Paper No. 5192.

——— (1995b): "Conditioning on the Probability of Selection to Control Selection Bias," forthcoming in *Economics Letters*.

BEGUN, J. M., W. J. HALL, W. M. HUANG, AND J. A. WELLNER (1983): "Information and Asymptotic Efficiency in Parametric-Nonparametric Models," *Annals of Statistics*, 11, 432–452.

BICKEL, P., C. A. J. KLAASSEN, Y. RITOV, AND J. A. WELLNER (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.

CLEMENTS, N., J. HECKMAN, AND J. SMITH (1993): "Making the Most out of Social Experiments: Reducing the Intrinsic Uncertainty in Evidence from Randomized Trials with an Application to the National JTPA Experiment," Unpublished manuscript, University of Chicago.

HECKMAN, J. (1992): "Randomization and Social Policy Evaluation," in *Evaluating Welfare and Training Programs*, ed. by C. Manski and I. Garfinkel. Cambridge: Harvard University Press.

HECKMAN, J., H. ICHIMURA, J. SMITH, AND P. TODD (1995): "Nonparametric Characterization of Selection Bias Using Experimental Data: A Study of Adult Males in JTPA," Unpublished manuscript, University of Chicago.

HECKMAN, J., AND J. SMITH (1995): "Assessing the Case for Randomized Social Experiments," *Journal of Economic Perspectives*, 9, 85–110.

IMBENS, G. W., AND J. D. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

LORENTZ, G. (1986): *Approximation of Functions*. New York: Chelsea Publishing Company.

NEWEY, W. K. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135.

——— (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

——— (1995): "Convergence Rates for Series Estimators," in *Advances in Econometrics and Quantitative Economics: Essays in Honor of Professor C. R. Rao*, ed. by G. Maddala, P. Phillips, and T. Srinivasan. Cambridge: Basil Blackwell.

ROBINSON, P. (1988): "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

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

——— (1984): "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79, 516–524.

STEIN, C. (1956): "Efficient Nonparametric Testing and Estimation," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1. Berkeley: University of California Press.

TODD, P. (1995): "Matching and Local Linear Regression Approaches to Solving the Evaluation Problem with a Semiparametric Propensity Score," Unpublished manuscript, University of Chicago.