Matching As An Econometric Evaluation Estimator

by James J. Heckman, Hidehiko Ichimura, and Petra Todd
Review of Economic Studies (1998)

James J. Heckman

THE UNIVERSITY OF CHICAGO

Econ 312, Spring 2023

1. Introduction

- Matching is a widely-used method of evaluation.

- It is based on the intuitively attractive idea of contrasting the outcomes of programme participants (denoted Y1) with the outcomes of "comparable" nonparticipants (denoted Y0).

- Differences in the outcomes between the two groups are attributed to the programme.

- The estimated gain for each person i in the treated sample is

$$Y_{1i} - \sum_{j \in I_0} W_{N_0,N_1}(i,j) Y_{0j}, \quad (1)$$

- The widely-used evaluation parameter on which we focus in this paper is the mean effect of treatment on the treated for persons with characteristics X

$$E(Y_1 - Y_0 | D = 1, X), \qquad \textbf{(P-1)}$$

where D = 1 denotes programme participation. Heckman (1997) and Heckman and Smit (1998) discuss conditions under which this parameter answers economically interesting questions.

- For a particular domain $\mathcal{H}$ for X, this parameter is estimated by

$$\sum_{i \in I_1} w_{N_0,N_1}(i)[Y_{1i} - \sum_{j \in I_0} W_{N_0,N_1}(i,j) Y_{0j}], \qquad \textbf{(2)}$$

where different values of wN0,N1(i) may be used to select different domains ff or to account for heteroskedasticity in the treated sample. Different matching methods are based on different weighting functions $\{w_{N0,N1}(i)\}$ and $\{W_{N0,N1}(i,j)\}$.

- We show that the fundamental identification condition of the matching method for estimating (P-1) is

$$E(Y_0 | D=1, X) = E(Y_0 | D=0, X),$$

whenever both sides of this expression are well defined.

- In order to meaningfully implement matching it is necessary to condition on the support common to both participant and comparison groups S, where

$$S = \text{Supp}\,(X | D=1) \cap \text{Supp}\,(X | D=0),$$

and to estimate the region of common support.

2. The Evaluation Problem and The Parameters of Interest

- The selection bias that arises from making this approximation is

$$B(X) = \mathrm{E}(Y_0 | D=1, X) - \mathrm{E}(Y_0 | D=0, X).$$

- Averaging the estimators over intervals of X produces a consistent estimator of

$$M(S) = \mathrm{E}(Y_1 - Y_0 | D = 1, X \in S), \qquad \text{(P-2)}$$

with a well-defined $N^{-1/2}$ distribution theory where Sis a subset of Supp $(X|D = 1)$.

3. How Matching Solves The Evaluation Problem

- Using the notation of Dawid (1979) let

$$(Y_0, Y_1) \perp\!\!\!\perp D \mid X, \qquad \text{(A-1)}$$

denote the statistical independence of (Y0 , Y1) and D conditional on X. An equivalent formulation of this condition is

$$\Pr(D=1 \mid Y_0, Y_1, X) = \Pr(D=1 \mid X).$$

- This is a non-causality condition that excludes the dependence between potential outcomes and participation that is central to econometric models of self selection. (See Heckman and Honore (1990).)

- Rosenbaum and Rubin (1983), henceforth denoted RR, establish that, when (A-1) and

$$0 < P(X) < 1, \qquad \text{(A-2)}$$

are satisfied, (Yo, Y1) JLDIP(X), where P(X) =Pr (D= 1 IX).

- When the strong ignorability condition holds, one can generate marginal distribution of the counterfactuals

$$F_0(y_0 | D=1, X) \quad \text{and} \quad F_1(y_1 | D=0, X),$$

- Note that under assumption (A-1)

$$\begin{aligned}\mathrm{E}(Y_0|D=1, X\in S) &= \mathrm{E}[\mathrm{E}(Y_0|D=1, X)|D=1, X\in S] \\ &= \mathrm{E}[\mathrm{E}(Y_0|D=0, X)|D=1, X\in S],\end{aligned}$$

so $E(Y_0|D = 1, X \in S)$ can be recovered from $E(Y_0|D + 0, X)$ by integrating over X using the distribution of X given $D = 1$, restricted to S.

- We can get by with a weaker condition since our objective is construction of the counterfactual $E(Y_0|X, D = 1)$

$$Y_0 \perp\!\!\!\perp D|X, \tag{A-3}$$

which implies that $\Pr(Y_0 < t|D = 1, X) = \Pr(Y_0 < t|D = 0, X)$ for $X \in S$.

- For identification of the mean treatment impact parameter (P-1), an even weaker mean independence condition suffices

$$E(Y_0 | D=1, X) = E(Y_0 | D=0, X) \quad \text{for } X \in S. \qquad \text{(A-1')}$$

3. Separability and Exclusion Restrictions

- In many applications in economics, it is instructive to partition *X* into two not-necessarily mutually exclusive sets of variables, *(T, Z)*, where the *T* variables determine outcomes

$$Y_0 = g_0(T) + U_0, \tag{3a}$$

$$Y_1 = g_1(T) + U_1, \tag{3b}$$

and the Z variables determine programme participation

$$\Pr(D=1 \mid X) = \Pr(D=1 \mid Z) = P(Z). \tag{4}$$

- The evidence reported in Heckman, Ichimura, Smith and Todd (1996a), reveals that the no-training earnings of persons who chose to participate in a training programme, Y_0, can be represented in the following way

$$E(Y_0 | D=1, X) = g_0(T) + E(U_0 | P(Z)),$$

where Z and T contain some distinct regressors.

- Thus, instead of (A-1) or (A-3), we consider the case where

$$U_0 \perp\!\!\!\perp D \mid X. \tag{A-4a}$$

- Invoking the exclusion restrictions P(X) = P(Z) and using an argument analogous to Rosenbaum and Rubin (1983), we obtain

$$\begin{aligned} \mathrm{E}\{D \mid U_0, P(Z)\} &= \mathrm{E}\{\mathrm{E}(D \mid U_0, X) \mid U_0, P(Z)\} \\ &= \mathrm{E}\{P(Z) \mid U_0, P(Z)\} = P(Z) = \mathrm{E}\{D \mid P(Z)\}, \end{aligned}$$

so that

$$U_0 \perp\!\!\!\perp D \mid P(Z). \tag{A-4b}$$

- Under condition (A-4a) it is not necessarily true that (A-1) or (A-3) are valid but it is obviously true that

$$[Y_0 - g_0(T)] \perp\!\!\!\perp D \mid P(Z).$$

- In order to identify the mean treatment effect on the treated, it is enough to assume that

$$\mathrm{E}(U_0 | D=1, P(Z)) = \mathrm{E}(U_0 | D=0, P(Z)), \qquad \text{(A-4b}')$$

instead of (A-4a) or (A-4b).

- In order to place these results in the context of classical econometric selection models, consider the following index model setup

$$\begin{aligned} Y_0 &= g_0(T) + U_0, \\ D &= 1 \quad \text{if } \psi(Z) - v \geqq 0; \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

- If Z and v are independent, then $P(Z) = F_v(\psi(Z))$ where $F_v(\cdot)$ is the distribution function of v.

- In this case identification condition (A-4b') implies

$$\mathrm{E}[U_0 | D=1, F_v(\psi(Z))] = \mathrm{E}[U_0 | D=0, F_v(\psi(Z))], \qquad (*)$$

or when F_v is strictly increasing,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\psi(Z)} U_0 f(U_0, v | \psi(Z)) dv dU_0 / F_v(\psi(Z))$$
$$= \int_{-\infty}^{\infty}\int_{\psi(Z)}^{\infty} U_0 f(U_0, v | \psi(Z)) dv dU_0 / [1 - F_v(\psi(Z))].$$

If, in addition, $\psi(Z)$ is independent of (U_0, v), and $\mathrm{E}(U_0)=0$, condition $(*)$ implies

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\psi(Z)} U_0 f(U_0, v) dv dU_0 = 0,$$

5. Estimating The Mean Effect Of Treatment: Should One Use The Propensity Score Or Not?

Under (A-1′) with $S = \text{Supp}(X|D=1)$ and random sampling across individuals, if one knew $E(Y_0|D=0, X=x)$, a consistent estimator of (P-2) is

$$\hat{\Delta}_X = N_1^{-1} \sum_{i \in I_1} [Y_{1i} - E(Y_0|D=0, X=X_i)],$$

where I_1 is the set of i indices corresponding to observations for which $D_i = 1$. If we assume

$$E(Y_0|D=1, P(X)) = E(Y_0|D=0, P(X)) \text{ for } X \in \text{Supp}(P(X)|D=1), \qquad \text{(A-1″)}$$

which is an implication of (A-1), and $E(Y_0|D=0, P(X)=p)$ is known, the estimator

$$\hat{\Delta}_P = N_1^{-1} \sum_{i \in I_1} [Y_{1i} - E(Y_0|D=0, P(X)=P(X_i))]$$

is consistent for $E(\Delta|D=1)$.

Theorem 1. *Assume*:

(i) (A-1′) *and* (A-1″) *hold for* $S = \text{Supp}(X \mid D=1)$;
(ii) $\{Y_{1i}, X_i\}_{i \in I_1}$ *are independent and identically distributed*;

and

(iii) $0 < \text{E}(Y_0^2) \cdot \text{E}(Y_1^2) < \infty$.

Then $\hat{\Delta}_X$ *and* $\hat{\Delta}_P$ *are both consistent estimators of* (P-2) *with asymptotic distributions that are normal with mean* 0 *and asymptotic variances* V_X *and* V_P, *respectively, where*

$$V_X = \text{E}[\text{Var}(Y_1 \mid D=1, X) \mid D=1] + \text{Var}[\text{E}(Y_1 - Y_0 \mid D=1, X) \mid D=1],$$

and

$$V_P = \text{E}[\text{Var}(Y_1 \mid D=1, P(X)) \mid D=1] + \text{Var}[\text{E}(Y_1 - Y_0 \mid D=1, P(X)) \mid D=1].$$

- The theorem directly follows from the central limit theorem for iid sampling with finite second moment and for the sake of brevity its proof is deleted.

- Observe that

$$E[\mathrm{Var}\,(Y_1 | D=1, X) | D=1] \leqq E[\mathrm{Var}\,(Y_1 | D=1, P(X)) | D=1],$$

because *X* is in general a better predictor than *P(X)* but

$$\mathrm{Var}\,[E(Y_1 - Y_0 | D=1, X) | D=1] \geqq \mathrm{Var}\,[E1(Y_1 - Y_0) | D=1, P(X)) | D=1],$$

6. Asymptotic Distribution Theory for Kernel-based Matching Estimators

- The general class of estimators of (P-2) that we analyse are of the form

$$\hat{\Delta} = \frac{N_1^{-1} \sum_{i \in I_1} [Y_{1i} - \hat{g}(T_i, \hat{P}(Z_i))] I(X_i \in \hat{S})}{N_1^{-1} \sum_{i \in I_1} I(X_i \in \hat{S})} \tag{6}$$

where $I(A) = 1$ if A holds and $= 0$ otherwise and $\hat{S}$ is an estimator of S, the region of overlapping support, where $S = \text{Supp}\,\{X \mid D=1\} \cap \text{Supp}\,\{X \mid D=0\}$.

Definition 1. An estimator $\hat{\theta}(x)$ of $\theta(x)$ is an asymptotically linear estimator with trimming $I(x \in \hat{S})$ if and only if there is a function $\psi_n \in \Psi_n$, defined over some subset of a finite-dimensional Euclidean space, and stochastic terms $\hat{b}(x)$ and $\hat{R}(x)$ such that for sample size n:

(i) $[\hat{\theta}(x) - \theta(x)] I(x \in \hat{S}) = n^{-1} \sum_{i=1}^{n} \psi_n(X_i, Y_i; x) + \hat{b}(x) + \hat{R}(x)$;
(ii) $\mathrm{E}\{\psi_n(X_i, Y_i; X) \mid X = x\} = 0$;
(iii) $\mathrm{plim}_{n \to \infty} n^{-1/2} \sum_{i=1}^{n} \hat{b}(X_i) = b < \infty$;
(iv) $n^{-1/2} \sum_{i=1}^{n} \hat{R}(X_i) = o_p(1)$.

A typical estimator of a parametric regression function $m(x;\beta)$ takes the form $m(x;\hat{\beta})$, where m is a known function and $\hat{\beta}$ is an asymptotically linear estimator, with $\hat{\beta}-\beta=n^{-1}\sum_{i=1}^{n}\psi(X_i, Y_i)+o_p(n^{-1/2})$. In this case, by a Taylor expansion,

$$\begin{aligned}\sqrt{n}[m(x,\hat{\beta})-m(x,\beta)] &= n^{-1/2}\textstyle\sum_{i=1}^{n}[\partial m(x,\beta)/\partial\beta]\psi(X_i, Y_i)\\ &\quad +[\partial m(x,\bar{\beta})/\partial\beta-\partial m(x,\beta)/\partial\beta]n^{-1/2}\textstyle\sum_{i=1}^{n}\psi(X_i, Y_i)+o_p(1),\end{aligned}$$

where $\bar{\beta}$ lies on a line segment between β and $\hat{\beta}$. When $\mathrm{E}\{\psi(X_i, Y_i)\}=0$ and $\mathrm{E}\{\psi(X_i, Y_i)\psi(X_i, Y_i)'\}<\infty$, under iid sampling, for example, $n^{-1/2}\sum_{i=1}^{n}\psi(X_i, Y_i)=O_p(1)$ and $\mathrm{plim}_{n\to\infty}\hat{\beta}=\beta$ so that $\mathrm{plim}_{n\to\infty}|\partial m(x,\bar{\beta})/\partial\beta-\partial m(x,\beta)/\partial\beta|=o_p(1)$ if $\partial m(x,\beta)/\partial\beta$ is Hölder continuous at β.[14]

Under these regularity conditions

$$\sqrt{n}[m(x,\hat{\beta})-m(x,\beta)]=n^{-1/2}\textstyle\sum_{i=1}^{n}[\partial m(x,\beta)/\partial\beta]\psi(X_i, Y_i)+o_p(1).$$

(a) *Asymptotic linearity of the kernel regression estimator*

We now establish that the more general kernel regression estimator for nonparametric functions is also asymptotically linear. Corollary 1 stated below is a consequence of a more general theorem proved in the Appendix for local polynomial regression models used in Heckman, Ichimura, Smith and Todd (1998) and Heckman, Ichimura and Todd (1997). We present a specialized result here to simplify notation and focus on main ideas. To establish this result we first need to invoke the following assumptions.

Assumption 1. Sampling of $\{X_i, Y_i\}$ is i.i.d., X_i takes values in R^d and Y_i in R, and $\text{Var}(Y_i) < \infty$.

When a function is p-times continuously differentiable and its p-th derivative satisfies Hölder's condition, we call the function p-smooth. Let $m(x) = \text{E}\{Y_i | X_i = x\}$.

Assumption 2. $m(x)$ is $\bar{p}$-smooth, where $\bar{p} > d$.

We also allow for stochastic bandwidths:

Assumption 3. Bandwidth sequence a_n satisfies $\text{plim}_{n\to\infty} a_n/h_n = \alpha_0 > 0$ for some deterministic sequence $\{h_n\}$ that satisfies $nh_n^d/\log n \to \infty$ and $nh_n^{2\bar{p}} \to c < \infty$ for some $c \geqq 0$.

This assumption implies $2\bar{p} > d$ but a stronger condition is already imposed in Assumption 2.[15]

Assumption 4. Kernel function $K(\cdot)$ is symmetric, supported on a compact set, and is Lipschitz continuous.

Assumption 5. Trimming is $\bar{p}$-nice on S.

In order to control the bias of the kernel regression estimator, we need to make additional assumptions. Certain moments of the kernel function need to be 0, the underlying Lebesgue density of X_i, $f_X(x)$, needs to be smooth, and the point at which the function is estimated needs to be an interior point of the support of X_i. It is demonstrated in the Appendix that these assumptions are not necessary for $\bar{p}$-th order local polynomial regression estimator.

Assumption 6. Kernel function $K(\cdot)$ has moments of order 1 through $\bar{p}-1$ that are equal to zero.

Assumption 7. $f_X(x)$ is $\bar{p}$-smooth.

Assumption 8. A point at which $m(\cdot)$ is being estimated is an interior point of the support of X_i.

The following characterization of the bias is a consequence of Theorem 3 that is proved in the Appendix.

Corollary 1. *Under Assumptions* 1–8, *if* $K(u_1, \ldots, u_d) = k(u_1) \cdots k(u_d)$ *where* $k(\cdot)$ *is a one dimensional kernel, the kernel regression estimator* $\hat{m}_0(x)$ *of* $m(x)$ *is asymptotically linear with trimming, where, writing* $\varepsilon_i = Y_i - \mathrm{E}\{Y_i | X_i\}$, *and*

$$\psi_n(X_i, Y_i; x) = (n\alpha_0 h_n^d)^{-1} \varepsilon_i K((X_i - x)/(\alpha_0 h_n)) I(x \in S)/f_X(x),$$

$$\hat{b}(x) = (\alpha_0 h_n)^{\bar{p}} \cdot \left[f_X(x) \cdot \int K(u)du \right]^{-1} \textstyle\sum_{s=1}^{\bar{p}} [s!(\bar{p}-s)!]^{-1}$$

$$\times \textstyle\sum_{k=1}^{d} \left[\left[\int u_k^{\bar{p}} K(u)du \right] \left[\partial^s m(x)/(\partial x_k)^s \right] \cdot \left[\partial^{(\bar{p}-s)} f_X(x)/(\partial x_k)^{(\bar{p}-s)} \right] \right] I(x \in \hat{S}).$$

(*b*) *Extensions to the case of local polynomial regression*

In the Appendix, we consider the more general case in which the local polynomial regression estimator for $\hat{g}(t, p)$ is asymptotically linear with trimming with a uniformly consistent derivative. The latter property is useful because, as the next lemma shows, if both $\hat{P}(z)$ and $\hat{g}(t, p)$ are asymptotically linear, and if $\partial\hat{g}(t, p)/\partial p$ is uniformly consistent, then $\hat{g}(t, \hat{P}(z))$ is also asymptotically linear under some additional conditions. We also verify in the Appendix that these additional conditions are satisfied for the local polynomial regression estimators.

Let $\bar{P}_t(z)$ be a function that is defined by a Taylor's expansion of $\hat{g}(t, \hat{P}(z))$ in the neighbourhood of $P(z)$, i.e. $\hat{g}(t, \hat{P}(z)) = \hat{g}(t, P(z)) + \partial\hat{g}(t, \bar{P}_t(z))/\partial p \cdot [\hat{P}(z) - P(z)]$.

Lemma 1. *Suppose that:*

(i) *Both* $\hat{P}(z)$ *and* $\hat{g}(t,p)$ *are asymptotically linear with trimming where*

$$[\hat{P}(z) - P(z)]I(x \in \hat{S}) = n^{-1}\sum_{j=1}^{n} \psi_{np}(D_j, Z_j; z) + \hat{b}_p(z) + \hat{R}_p(z),$$

$$[\hat{g}(t,p) - g(t,p)]I(x \in \hat{S}) = n^{-1}\sum_{j=1}^{n} \psi_{ng}(Y_j, T_j, P(Z_j); t, p) + \hat{b}_g(t,p) + \hat{R}_g(t,p);$$

(ii) $\partial\hat{g}(t,p)/\partial p$ *and* $\hat{P}(z)$ *are uniformly consistent and converge to* $\partial g(t,p)/\partial p$ *and* $P(z)$, *respectively and* $\partial g(t,p)/\partial p$ *is continuous;*

(iii) $\operatorname{plim}_{n\to\infty} n^{-1/2}\sum_{j=1}^{n} \hat{b}_g(T_i, P(Z_i)) = b_g$ *and* $\operatorname{plim}_{n\to\infty} n^{-1/2}\sum_{i=1}^{n} \partial g(T_i, P(Z_i))/\partial p \cdot \hat{b}_p(T_i, P(Z_i)) = b_{g_p}$;

(iv) $\operatorname{plim}_{n\to\infty} n^{-1/2}\sum_{j=1}^{n} [\partial\hat{g}(T_i, \bar{P}_{T_i}(Z_i))/\partial p - \partial g(T_i, P(Z_i))/\partial p] \cdot \hat{R}_p(Z_i) = 0$;

(v) $\operatorname{plim}_{n\to\infty} n^{-3/2}\sum_{i=1}^{n}\sum_{j=1}^{n} [\partial\hat{g}(T_i, \bar{P}_{T_i}(Z_i))/\partial p - \partial g(T_i, P(Z_i))/\partial p] \cdot \psi_{np}(D_j, Z_j; Z_i) = 0.$

then $\hat{g}(t, \hat{P}(z))$ *is also asymptotically linear where*

$$[\hat{g}(t, \hat{P}(z)) - g(t, P(z))]I(x \in \hat{S}) = n^{-1}\sum_{j=1}^{n} [\psi_{ng}(Y_j, T_j, P(Z_j); t, P(z)) + \partial g(t, P(z))/\partial p \cdot \psi_{np}(D_j, Z_j; z)] + \hat{b}(x) + \hat{R}(x),$$

and $\operatorname{plim}_{n\to\infty} n^{-1/2}\sum_{i=1}^{n} \hat{b}(X_i) = b_g + b_{g_p}$.

Assumption 9. $K(\cdot)$ is 1-smooth.

Lemma 1 implies that the asymptotic distribution theory of $\hat{\Delta}$ can be obtained for those estimators based on asymptotically linear estimators with trimming for the general nonparametric (in P and g) case. Once this result is established, it can be used with lemma 1 to analyze the properties of two stage estimators of the form $\hat{g}(t, \hat{P}(z))$.

Theorem 2. *Under the following conditions:*

(i) $\{Y_{0i}, X_i\}_{i \in I_0}$ *and* $\{Y_{1i}, X_i\}_{i \in I_1}$ *are independent and within each group they are* i.i.d. *and* Y_{0i} *for* $i \in I_0$ *and* Y_{1i} *for* $i \in I_1$ *each has a finite second moment;*

(ii) *The estimator* $\hat{g}(x)$ *of* $g(x) = \mathrm{E}\{Y_{0_i} | D_i = 1, X_i = x\}$ *is asymptotically linear with trimming, where*

$$[\hat{g}(x) - g(x)] I\{x \in \hat{S}\} = N_0^{-1} \sum_{i \in I_0} \psi_{0N_0N_1}(Y_{0i}, X_i; x) + N_1^{-1} \sum_{i \in I_1} \psi_{1N_0N_1}(Y_{1i}, X_i; x) + \hat{b}_g(x) + \hat{R}_g(x)$$

and the score functions $\psi_{dN_0N_1}(Y_d, X; x)$ *for* $d = 0$ *and* 1, *the bias term* $\hat{b}_g(x)$, *and the trimming function satisfy:*

(ii-a) $\mathrm{E}\{\psi_{dN_0N_1}(Y_{di}, X_i; X) | D_i = d, X, D = 1)\} = 0$ *for* $d = 0$ *and* 1, *and* $\mathrm{Var}\{\psi_{dN_0N_1}(Y_{di}, X_i; X)\} = o(N)$ *for each* $i \in I_0 \cup I_1$;

(ii-b) $\mathrm{plim}_{N_1 \to \infty} N_1^{-1/2} \sum_{i \in I_1} \hat{b}(X_i) = b$;

(ii-c) $\mathrm{plim}_{N_1 \to \infty} \mathrm{Var}\{\mathrm{E}[\psi_{0N_0N_1}(Y_{0i}, X_i; X) | Y_{0i}, D_i = 0, X_i, D = 1] | D = 1\} = V_0 < \infty$
$\mathrm{plim}_{N_1 \to \infty} \mathrm{Var}\{\mathrm{E}[\psi_{1N_0N_1}(Y_{1i}, X_i; X) | Y_{1i}, D_i = 1, X_i, D = 1] | D = 1\} = V_1 < \infty$,
and

$$\lim_{N_1 \to \infty} \mathrm{E}\{[Y_{1i} - g(X_i)] I(X_i \in S) \cdot \mathrm{E}[\psi_{1N_0N_1}(Y_{1i}, X_i; X) | Y_{1i}, D_i = 1, X_i, D = 1] | D = 1\} = \mathrm{Cov}_1;$$

(ii-d) S *and* $\hat{S}$ *are* $\bar{p}$*-nice on* S, *where* $\bar{p} > d$, *where* d *is the number of regressors in* X *and* $\hat{f}(x)$ *is a kernel density estimator that uses a kernel function that satisfies Assumption* 6.

Then under (A-1′) the asymptotic distribution of

$$N_1^{1/2}\left[\frac{N_1^{-1}\sum_{i\in I_1}[Y_{1i}-\hat{g}(X_i)]I(X_i\in\hat{S})}{N_1^{-1}\sum_{i\in I_1}I(X_i\in\hat{S})}-\mathrm{E}_S(Y_1-Y_0|D=1)\right]$$

is normal with mean $(b/\Pr(X\in S|D=1))$ *and asymptotic variance*

$$\begin{aligned}\Pr(X\in S|D=1)^{-1}\{\mathrm{Var}_S[\mathrm{E}_S(Y_1-Y_0|T,P(Z),D=1)|D=1]\\ +\mathrm{E}_S[\mathrm{Var}_S(Y_1|T,P(Z),D=1)|D=1]\}\\ +\Pr(X\in S|D=1)^{-2}\{V_1+2\cdot Cov_1+\theta V_0\}.\end{aligned}$$

Proof. See the Appendix. ‖

7. Answers to The Three Questions of Section 1 and More General Questions Concerning The Value of A Priori Information

- In this case the score function $\psi 1N0N1\ (Y_{1i}, X_i; x)$ and

$$\psi_{0N_0N_1}(Y_{0i}, X_i; x) = \frac{\varepsilon_i K((X_i - x)/a_{N_0}) I(x \in S)}{a_{N_0}^d f_X(x \mid D=0) \int K(u)du},$$

where $\varepsilon_i = Y_{0i} - \mathrm{E}\{Y_{0i} \mid X_i, D_i = 0\}$ and we write $f_X(x \mid D=0)$ for the Lebesgue density of X_i given $D_i = 0$. (We use analogous expressions to denote various Lebesgue densities.) Clearly V_1 and Cov_1 are zero in this case. Using the score function we can calculate V_0 when we match on X. Denoting this variance by V_{0X},

$$\begin{aligned} V_{0X} &= \lim_{N_0 \to \infty} \mathrm{Var}\,\{\mathrm{E}[\psi_{0N_0N_1}(Y_{0i}, X_i, X) \mid Y_{0i}, D_i = 0, X_i, D = 1] \mid D = 1\} \\ &= \lim_{N_0 \to \infty} \mathrm{Var}\left\{\mathrm{E}\left[\frac{\varepsilon_i K((X_i - X)/a_{N_0}) I(X \in S)}{a_{N_0}^d f_X(X \mid D=0) \int K(u)du} \mid Y_{0i}, D_i = 0, X_i, D = 1\right] \mid D = 1\right\}. \end{aligned}$$

- Now observe that conditioning on X_i and Y_{0i} is given, so that we may write the last expression as

$$\mathrm{Var}\left\{\varepsilon_i \mathrm{E}\left[\frac{K((X_i - X)/a_{N_0}) I(X \in S)}{a_{N_0}^d f_X(X \mid D=0) \int K(u)du} \mid D_i = 0, X_i, D=1\right] \mid D=1\right\}.$$

- Now

$$\mathrm{E}\left[\frac{K((X_i - X)/a_{N_0}) I(X \in S)}{a_{N_0}^d f_X(X \mid D=0) \int K(u)du} \mid D_i = 0, X_i, D=1\right],$$

can be written in the following way, making the change of variable $\frac{X_i - X}{a_{n0}} = w$:

$$\int \frac{K(w) I([X_i - a_{N_0} w] \in S)}{\int K(u)du} \frac{f(X_i - a_{N_0} w \mid D=1)}{f(X_i - a_{N_0} w \mid D=0)} dw.$$

- Taking limits as $N_0 \to \infty$, and using assumptions 3, 4 and 7, so we can take limits inside the integral

$$\lim_{N_0 \to \infty} \mathrm{E}\left[\frac{K((X_i - X)/a_{N_0}) I(X \in S)}{a_{N_0}^d f_X(X \mid D=0) \int K(u)\,du} \,\middle|\, D_i = 0, X_i, D = 1\right] = \frac{f(X_i \mid D=1)}{f(X_i \mid D=0)} I(X_i \in S),$$

since $a_{N_0} \to 0$ and $\int K(w)\,dw / \int K(u)\,du = 1$. Thus, since we sample the X_i for which $D_i = 0$,

$$V_{0X} = \mathrm{E}_S\left[\frac{\mathrm{Var}\,(Y_{0i} \mid X_i, D_i = 0) f_X^2(X_i \mid D_i = 1)}{f_X^2(X_i \mid D_i = 0)} \,\middle|\, D_i = 0\right] \Pr\,\{X_i \in S \mid D_i = 0\}.$$

Hence the asymptotic variance of $\hat{\Delta}_X$ is, writing $\lambda = \Pr\{X \in S | D=0\} / \Pr(X \in S | D=1)$,

$$\Pr(X \in S | D=1)^{-1} \{ \mathrm{Var}_S [\mathrm{E}_S(Y_1 - Y_0 | X, D=1) | D=1] + \mathrm{E}_S[\mathrm{Var}_S(Y_1 | X, D=1) | D=1]$$
$$+ \lambda\theta \mathrm{E}_S [\mathrm{Var}(Y_0 | X, D=0) f_X^2(X | D=1) / f_X^2(X | D=0) | D=0] \}.$$

Similarly for $\hat{\Delta}_P$, V_{0P} is

$$\Pr(X \in S | D=1)^{-1} \{ \mathrm{Var}_S [\mathrm{E}_S(Y_1 - Y_0 | P(X), D=1) | D=1]$$
$$+ \mathrm{E}_S [\mathrm{Var}_S(Y_1 | P(X), D=1) | D=1]$$
$$+ \lambda\theta \mathrm{E}_S [\mathrm{Var}(Y_0 | P(X), D=0)$$
$$\times f_P^2(P(X) | D=1) / f_P^2(P(X) | D=0) | D=0] \}.$$

To show this first note that in this case $\mathrm{Var}(D|X) = \mathrm{Var}(D|Z)$. Thus

$$
\begin{aligned}
&[V_{2X} - V_{2Z}] \cdot \Pr\{X \in S\} \\
&\quad = \mathrm{E}_S\{\mathrm{Var}(D|Z)[\partial g(P(Z))/\partial p]^2 \cdot [[f_X^2(X|D=1)/f_X^2(X)] - [f_Z^2(Z|D=1)/f_Z^2(Z)]]\} \\
&\quad = \mathrm{E}_S\left\{\mathrm{Var}(D|Z)[\partial g(P(Z))/\partial p]^2 \cdot \left[\frac{f_Z^2(Z|D=1)}{f_Z^2(Z)}\right] \cdot \left[\mathrm{E}\left(\frac{f_X^2(X|Z, D=1)}{f_X^2(X|Z)}|Z\right) - 1\right]\right\} \\
&\quad \geqq \mathrm{E}_S\left\{\mathrm{Var}(D|Z)[\partial g(P(Z))/\partial p]^2 \cdot \left[\frac{f_Z^2(Z|D=1)}{f_Z^2(Z)}\right] \cdot \left[\mathrm{E}\left(\frac{f_X(X|Z, D=1)}{f_X(X|Z)}|Z\right)^2 - 1\right]\right\} \\
&\quad = 0.
\end{aligned}
$$

8. Summary and Conclusion