Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score

Hirano, Imbens, and Ridder (2003)

Presented by Jason Blevins
Applied Microeconometrics Reading Group
Duke University

October 5, 2006

Context

Previous Work:

- As shown by Rosenbaum and Rubin (1983, 1985), the unconfoundedness assumption implies that adjusting for $p(x)$ removes all bias associated with differences in x.
- Hahn (1998) shows that while this removes all bias, it is not necessarily as efficient as conditioning on the covariates.
- Rosenbaum (1987), Rubin and Thomas (1996), and Robins, Rotnitzky, and Zhao (1995): There can be efficiency gains by using parametric estimates of the propensity score, rather than the true propensity score.

Main Finding:

- Estimators for τ, τ_{wate}, and $\tau_{treated}$ are presented which weight observations by the inverse of nonparametric estimates of $p(x)$. If the estimator for $p(x)$ is sufficiently flexible, this leads to a fully efficient estimator.

Outline

- Model, Objectives, and Assumptions
- Other Approaches: Matching
- Estimation Using the Propensity Score
- Previous Results: Hahn (1998)
- Missing Data Example
- Three Efficient Estimators
  - Population Average Treatement Effect
  - Weighted Average Treatement Effect
  - Average Treatement Effect for the Treated

Model

Population	$(T, X, Y(0), Y(1))$
Missing data	$Y \equiv T \cdot Y(1) + (1 - T) \cdot Y(0)$
Random sample	$\{(T_i, X_i, Y_i)\}_{i=1}^{N}$
Treatment indicator	$T_i \in \{0, 1\}$
Vector of covariates	X_i
Outcomes	$Y_i(0), Y_i(1)$

Assumption 1. *Unconfoundedness:* $T \perp (Y(0), Y(1)) \,|\, X$

Quantities of Interest

Population ATE	$\tau = \mathbb{E}[Y(1) - Y(0)]$
Weighted ATE	$\tau_{wate} = \frac{\int \mathbb{E}[Y(1) - Y(0) \,\vert\, X = x] g(x) dF(x)}{\int g(x) dF(x)}$
ATE on the Treated	$\tau = \mathbb{E}[Y(1) - Y(0) \,\vert\, T = 1]$
Propensity Score	$p(x) = \mathbb{P}(T = 1 \,\vert\, X = x)$

- ATE on the Treated arises when weight is $g(x) = p(x)$.
- Problem: we only observe either $Y_i(0)$ or $Y_i(1)$, never both.
- Straightforward nonparametric estimators are all infeasible!

Estimation by Matching

The unconfoundedness assumption implies that

$$\begin{aligned}
\tau(x) &\equiv \mathbb{E}[Y(1) - Y(0) \,|\, X = x] \\
&= \mathbb{E}[Y(1) \,|\, X = x] - \mathbb{E}[Y(0) \,|\, X = x] \\
&= \mathbb{E}[Y(1) \,|\, T = 1,\ X = x] - \mathbb{E}[Y(0) \,|\, T = 0,\ X = x] \\
&= \mathbb{E}[Y \,|\, T = 1,\ X = x] - \mathbb{E}[Y \,|\, T = 0,\ X = x]
\end{aligned}$$

since

$$\begin{aligned}
\mathbb{E}[Y \,|\, T = 1,\ X = x] &= \mathbb{E}[T \cdot Y(1) + (1 - T) \cdot Y(0) \,|\, T = 1,\ X = x] \\
&= \mathbb{E}[Y(1) \,|\, T = 1,\ X = x]
\end{aligned}$$

$$\tau = \int \tau(x) dF(x)$$

Estimation Using the Propensity Score

Rosenbaum and Rubin (1983, 1985) show that the unconfoundedness assumption $T \perp (Y(0), Y(1)) \mid X$ implies $T \perp (Y(0), Y(1)) \mid p(X)$.

Unconfoundedness gives:

$$\begin{aligned}\mathbb{E}[TY \mid X = x] &= \mathbb{E}[TY(1) \mid X = x] \\ &= \mathbb{E}[T \mid X = x]\,\mathbb{E}[Y(1) \mid X = x]\end{aligned}$$

$$\mathbb{E}[Y(1) \mid X = x] = \frac{\mathbb{E}[TY \mid X = x]}{\mathbb{E}[T \mid X = x]} = \frac{\mathbb{E}[TY \mid X = x]}{p(x)}$$

This suggests using a sample average to nonparametrically estimate

$$\tau = \mathbb{E}[\tau(x)] = \mathbb{E}\left[\mathbb{E}[Y(1) - Y(0) \mid X = x]\right].$$

Previous Results: Hahn (1998)

- Semiparametric efficiency bounds and estimators for τ and $\tau_{treated}$
  - Knowing $p(x)$ does not affect bound for τ.
  - Knowing $p(x)$ decreases the bound for $\tau_{treated}$.
- In general, conditioning only on $p(x)$ and not the covariates does not lead to an efficient estimator (experimental data case).
- Efficient estimator for τ, regardless of whether $p(x)$ is known.
  - Nonparametrically estimate $\mathbb{E}[YT \mid X = x]$, $\mathbb{E}[Y(1-T) \mid X = x]$, and $p(x)$.
  - Impute values for $Y_i(1)$ and $Y_i(0)$ using

$$\hat{Y}_i(1) = \frac{\hat{\mathbb{E}}[YT \mid X_i]}{\hat{p}(X_i)} \quad \text{and} \quad \hat{Y}_i(0) = \frac{\hat{\mathbb{E}}[Y(1-T) \mid X_i]}{1 - \hat{p}(X_i)}$$

Example: Missing Data with Binary Covariates

- Want to estimate $\beta_0 \equiv \mathbb{E}[Y]$ given a random sample $\{(T_i, X_i, T_iY_i)\}_{i=1}^N$.
- T_i and X_i are observed for everyone, Y_i observed only if $T_i = 1$.
- "Unconfoundedness": $T \perp Y \mid X$
- "Propensity score": $p(x) = \mathbb{E}[T \mid X = x] = \mathbb{P}[T = 1 \mid X = x]$
- Suppose $p(x) = 1/2$
- Binary covariates: $x \in \{0, 1\}$, $N_{tx} = \#\{i \mid T_i = t, X_i = x\}$

Example: True Weights Estimator

Normalized variance bound for β_0:

$$V_{bound} = 2\,\mathbb{E}\left[\mathrm{V}(Y \mid X)\right] + \mathrm{V}\left(\mathbb{E}[Y \mid X]\right)$$

True weights estimator:

$$\hat{\beta}_{tw} = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i T_i}{p(X_i)} = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_i T_i}{1/2}$$

$$V_{tw} = 2\,\mathbb{E}\left[V(Y \mid X)\right] + V\left(\mathbb{E}[Y \mid X]\right) + \mathbb{E}\left[\mathbb{E}[Y \mid X]^2\right]$$

Inefficient unless $\mathbb{E}[Y \mid X] = 0$

Example: Estimated Weights Estimator

Estimated "propensity score":

$$\hat{p}(x) = \begin{cases} N_{10}/(N_{00} + N_{10}) & \text{if } x = 0 \\ N_{11}/(N_{01} + N_{11}) & \text{if } x = 1 \end{cases}$$

$$N_{tx} = \#\{i \mid T_i = t,\, X_i = x\}$$

Estimated weights estimator:

$$\hat{\beta}_{ew} = \frac{1}{N} \sum_{i=1}^{N} \frac{Y_i T_i}{\hat{p}(X_i)}$$

$$V_{ew} = 2\,\mathbb{E}\left[V(Y \mid X)\right] + V\left(\mathbb{E}[Y \mid X]\right) = V_{bound}$$

Fully Efficient!

GMM Interpretation: True Weights Estimator

True weights estimator β_{tw} is GMM estimator with moment

$$\psi_1(y,\ t,\ x,\ \beta) = \frac{yt}{p(x)} - \beta = \frac{yt}{1/2} - \beta$$

corresponding to

$$\mathbb{E}\left\{\mathbb{E}[YT \mid X] - \mathbb{E}[Y \mid X]\,\mathbb{E}[T \mid X]\right\} = 0$$

Ignores information about T, not necessarily efficient.

GMM Interpretation: Estimated Weights Estimator

The propensity score provides additional information:

$$\mathbb{E}\left\{\mathbb{E}[T \mid X] - p(X)\right\} = \mathbb{E}[T - 1/2] = 0$$

With a binary covariate, we have

$$\psi_2(y,\ t,\ x,\ \beta) = \begin{bmatrix} x(t - 1/2) \\ (1 - x)(t - 1/2) \end{bmatrix}$$

GMM with moment conditions ψ_1 and ψ_2 is fully efficient and corresponds to the estimated weights estimator.

An Estimator for τ

Quantities of interest:

$$\tau^* \equiv \mathbb{E}[Y(1) - Y(0)] \qquad p(x) \equiv \mathbb{P}[T = 1 \,|\, X = x]$$

Conditional moments:

$$\mu_t(x) \equiv \mathbb{E}[Y(t) \,|\, X = x] \qquad \sigma_t^2(x) \equiv \mathrm{V}\left(Y(t) \,|\, X = x\right)$$

τ^* satisfies $\mathbb{E}\left[\psi(Y, T, X, \tau^*, p^*(X))\right] = 0$ where

$$\psi(t, t, x, \tau, p(x)) = \frac{yt}{p(x)} - \frac{y(1-t)}{1-p(x)} - \tau$$

Given an estimate of p,

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Y_i T_i}{\hat{p}(X_i)} - \frac{Y_i(1-T_i)}{1-\hat{p}(X_i)}\right)$$

Series Logit Estimator

Vector of functions: $R^K(x) = (r_{1K}(x), r_{2K}(x), \ldots, r_{KK}(x))^\top$

Multi-index: $\lambda = (\lambda_1, \ldots, \lambda_r)^\top$, $\lambda_j \in \mathbb{N}$, $r = \dim(x)$

Norm: $|\lambda| \equiv \sum_{j=1}^r \lambda_j$

Sequence of distinct multi-indices: $\{\lambda(k)\}_k$ with $|\lambda(k)| \leq |\lambda(k+1)|$

Power series elements: $x^\lambda = \prod_{j=1}^r x_j^{\lambda_j} = x_1^{\lambda_1} x_2^{\lambda_2} \cdots x_r^{\lambda_r}$

Take the sequence $\{r_{kK}(x)\}_k$ where $r_{kK}(x) = x^{\lambda(k)}$

Series Logit Estimator

Example ($r = 3$):

$$\lambda(1) = (0,\ 0,\ 0), \quad \lambda(2) = (1,\ 0,\ 0), \quad \lambda(3) = (0,\ 1,\ 0),$$
$$\lambda(4) = (0,\ 0,\ 1), \quad \lambda(4) = (2,\ 0,\ 0), \ldots$$

$$R^1 = 1 \quad R^2 = \begin{bmatrix} 1 \\ x_1 \end{bmatrix} \quad R^3 = \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} \quad R^4 = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} \quad R^5 = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \\ x_1^2 \end{bmatrix} \quad \ldots$$

Series Logit Estimator

Logistic CDF: $L(a) = \frac{\mathrm{e}^a}{1+\mathrm{e}^a}$

Series Logistic Estimator for $p^*(x)$ is $\hat{p}(x) = L(R^K(x)^\top \hat{\pi}_K)$ with

$$\hat{\pi}_K = \arg\max_{\pi} \sum_{i=1}^{N} \left[T_i \ln L(R^K(X_i)^\top \pi) + (1 - T_i) \ln(1 - L(R^K(X_i)^\top \pi)) \right]$$

More Assumptions

Assumption 2. *Distribution of X:*

i. *Support of X is a compact subset of $\mathbb{R}^r$.*

ii. *Density of X is bounded and bounded away from 0.*

Assumption 3. *Distribution of $(Y(0), Y(1))$:*

i. $\mathbb{E}[Y(0)^2] < \infty$ *and* $\mathbb{E}[Y(1)^2] < \infty$.

ii. $\mu_0(x)$ *and* $\mu_1(x)$ *are continuously differentiable.*

Assumption 4. *Selection probability:*

i. $p^*(x)$ *is continuously differentiable of order s with $s \geq 7r$.*

ii. $0 < p^*(x) < 1$

Assumption 5. *The Series Logit Estimator of $p^*(x)$ uses a power series with $K = N^\nu$ for some $\frac{1}{4(s/r-1)} < \nu < \frac{1}{9}$.*

Asymptotic Properties of $\hat{\tau}$

Theorem 1. *Suppose assumptions 1–5 hold. Then:*

i. $\hat{\tau} \xrightarrow{p} \tau^*$.

ii. $\sqrt{N}(\hat{\tau} - \tau^*) \xrightarrow{d} \mathrm{N}(0, V)$ *with*

$$V = \mathbb{E}\left[(\tau(x) - \tau)^2 + \frac{\sigma_1^2(X)}{p^*(X)} + \frac{\sigma_0^2(X)}{1 - p^*(X)} \right].$$

iii. $\hat{\tau}$ *reaches the semiparametric efficiency bound.*

Asymptotic Properties of $\hat{\tau}$

$\hat{\tau}$ is asymptotically linear:

$$\hat{\tau} = \tau^* + \frac{1}{N}\sum_{i=1}^{N} [\psi(Y_i, T_i, X_i, \tau^*, p^*(X_i)) + \alpha(T_i, X_i)] + o_p(1/\sqrt{N})$$

where

$$\alpha(t, x) = -\left(\frac{\mu_1(x)}{p^*(x)} + \frac{\mu_0(x)}{1 - p^*(x)}\right)(t - p^*(x))$$

and so

$$V = \mathbb{E}[(\psi + \alpha)^2]$$

Known weights estimator is asymptotically linear with influence function ψ.

Consistent estimator for V is found using another Series Logit Estimator.

An Efficient Estimator for τ_{wate}

$$\tau_{wate} = \frac{\int \mathbb{E}[Y(1) - Y(0) \mid X = x] g(x) dF(x)}{\int g(x) dF(x)}$$

By choosing a weighting function g appropriately, we can obtain average treatment effects for a subpopulation defined by X.

Note that $g = p^*$ yields $\tau_{treated}$.

$$\psi(y,\, t,\, x,\, \tau_{wate},\, p(x)) = g(x) \left(\frac{yt}{p(x)} - \frac{y(1-t)}{1-p(x)} - \tau_{wate} \right)$$

$$\hat{\tau}_{wate} = \sum_{i=1}^{N} g(X_i) \left[\frac{Y_i T_i}{\hat{p}(X_i)} - \frac{Y_i(1-T_i)}{1-\hat{p}(X_i)} \right] \Bigg/ \sum_{i=1}^{N} g(X_i)$$

Asymptotic Properties of $\hat{\tau}_{wate}$

Theorem 3. *Suppose assumptions 1–5 hold, $|g(x)|$ is bounded, and $\mathbb{E}[g(x)] > 0$. Then:*

i. $\hat{\tau}_{wate} \xrightarrow{p} \tau^*_{wate}$.

ii. $\sqrt{N}(\hat{\tau}_{wate} - \tau^*_{wate}) \xrightarrow{d} \mathrm{N}(0, V)$ *with*

$$V = \frac{1}{\mathbb{E}[g(X)]^2}\mathbb{E}\Big[g(X)^2\left(\tau(X) - \tau^*_{wate}\right)^2 + \frac{g(X)^2}{p^*(X)}\sigma_1^2(X) + \frac{g(X)^2}{1 - p^*(X)}\sigma_0^2(X)\Big].$$

iii. $\hat{V}$ *is consistent for* V.

Theorem 4. $\hat{\tau}_{wate}$ *reaches the semiparametric efficiency bound for* τ_{wate}.

An Estimator for $\tau_{treated}$ with p^* Known

Take $g(x) = p^*(x)$ and apply the estimator for τ_{wate}.

$$\psi(y,\, t,\, x,\, \tau_{wate},\, p(x)) = p^*(x) \left(\frac{yt}{p(x)} - \frac{y(1-t)}{1-p(x)} - \tau_{wate} \right)$$

The estimator $\hat{\tau}_{treated}$ is the solution to

$$0 = \sum_{i=1}^{N} p^*(X_i) \left(\frac{Y_i T_i}{\hat{p}(X_i)} - \frac{Y_i(1-T_i)}{1-\hat{p}(X_i)} - \tau_{treated} \right)$$

Notice that p^* is used as the weighting function while $\hat{p}$ weights observations.

Hahn (1998) showed that knowing p^* reduces the variance bound for $\tau_{treated}$.

From Theorems 3 and 4, this estimator is $\sqrt{N}$-consistent, asymptotically normal, and efficient.

An Estimator for $\tau_{treated}$ with p^* Unkown

If p^* is unknown, the efficiency bound for $\tau_{treated}$ is higher.

We need a new estimator since $\hat{\tau}_{treated}$ used p^*.

Let $\hat{\tau}_{te}$ be the solution to

$$0 = \sum_{i=1}^{N} \hat{p}(X_i) \left(\frac{Y_i T_i}{\hat{p}(X_i)} - \frac{Y_i(1 - T_i)}{1 - \hat{p}(X_i)} - \tau_{te} \right).$$

Asymptotic Properties of $\hat{\tau}_{te}$

Theorem 5. *Suppose that assumptions 1–5 hold. Then:*

i. $\hat{\tau}_{te} \xrightarrow{p} \tau^*_{treated}$.

ii. $\sqrt{N}(\hat{\tau}_{te} - \tau^*_{treated}) \xrightarrow{d} \mathrm{N}(0, V)$ *with*

$$V = \frac{1}{\mathbb{E}[p^*(X)]^2}\mathbb{E}\Big[p^*(X)^2\left(\tau(X) - \tau^*_{treated}\right)^2 + p^*(X)\sigma_1^2(X) + \frac{p^*(X)^2}{1 - p^*(X)}\sigma_0^2(X)\Big].$$

iii. τ_{te} *reaches the semiparametric efficiency bound for estimation of* $\tau_{treated}$ *when the propensity score is not known.*

Conclusion

Results:

- Hahn (1998) showed that conditioning on the true propensity score does not, in general, yield an efficient estimator.

- Weighting by the *true* propensity score does not yield efficient estimators, however, using the *estimated* propensity score does.

- The proposed estimators require nonparametric estimation of fewer functions than other estimators.

Open Questions:

- Finite sample properties

- Computational properties