Simple Least Squares Estimator for Treatment Effect Using Propensity Score Residual

(R&R, *Biometrika*)

Myoung-jae Lee

Korea University

February 3, 2017

Mean Difference and PSM

- For a binary treatment D, a response Y and covariates X, let Y^d be the potential response for $D = d$; $Y = (1 - D)Y^0 + DY^1$. If D is randomized,
$$E(Y|D = 1) - E(Y|D = 0) = E(Y^1 - Y^0).$$

- The sample version of $E(Y|D = 1) - E(Y|D = 0)$ equals
$$\text{Slope LSE of } Y \text{ on } (1, D) = \text{LSE of } Y - E(Y) \text{ on } D - E(D).$$

- Suppose D is not randomized and X needs to be controlled. If '$(Y^0, Y^1) \amalg D|X$', then
$$E(Y|D = 1, X) - E(Y|D = 0, X) = E(Y^1 - Y^0|X).$$

- To avoid the dimension problem in controlling X, propensity score matching (PSM) with $\pi(X) \equiv E(D|X)$ is used, as (Rosenbaum & Rubin 1983, BKA)
$$Y^d \amalg D|X \implies Y^d \amalg D|\pi(X) \quad \forall d.$$

Problems with PSM

- PSM requires several decisions on the user, according to which the effect estimate can change much.
- First, how many matched subjects for individual i: one for pair matching, and more for multiple matching.
- Second, whether to use a fixed number of matches M, or an individual-varying number M_i.
- Third, whether to use a caliper (a bound on the deviation between X_i and X of a matched individual) or not; if yes, its value.
- Fourth, matching with replacement or without. And more,...
- Getting standard errors in PSM is hard, despite the asymptotic normality in Abadie & Imbens (2016, ECA) under a parametric $\pi(X)$.
- The variance estimator is complicated, involving

$$V(Y|D=d, \pi(X)=p) \quad \& \quad COV\{X, E(Y|D=d, X)|\pi(X)=p\}.$$

Main Idea of PS-Residual LSE

- Is it possible to bring back the simple LSE of Y on $(1, D)$ while still controlling X nonparametrically? Can this be done without asking the user to make many decisions as in PSM?

- Under $Y^d \amalg D|X$ and the support-overlap condition $0 < \pi(X) < 1$, the answer is positive: do

$$\text{LSE of} \quad Y - E(Y) \quad \text{on} \quad D - \pi(X). \qquad (\text{LSE}^0_{psr})$$

- LSE^0_{psr} includes the simple LSE for randomized D as a special case, because $\pi(X) \equiv E(D|X) = E(D)$; the superscript 0 will be explained shortly.

- It may look puzzling why X does not appear as regressors along with $D - \pi(X)$. *The key point is that X is uncorrelated with $D - \pi(X)$, and thus X can be buried in the error; balancing/matching on X unnecessary.*

- If $\pi(X)$ is estimated nonparametrically, LSE^0_{psr} is nonparametric as well because the X-part not specified. But probit will be used for $\pi(X)$ under $\pi(X) = \Phi(X'\alpha)$ in this paper, which makes LSE^0_{psr} semiparametric.

Generalizing PS-Residual LSE

- Let $\Pi^q(Y|X'\alpha)$ denotes the linear projection of Y on $\{1,\ X'\alpha,\ ...,\ (X'\alpha)^q\}$. A generalized version of LSE^0_{psr} is

$$\text{LSE of} \quad Y - \Pi^q(Y|X'\alpha) \quad \text{on} \quad D - \pi(X).$$

- With the projection coefficient γ_j for $(X'\alpha)^j$, $j = 0, ..., q$, this LSE is

$$\text{LSE of} \quad Y - \sum_{j=0}^{q} \gamma_j (X'\alpha)^j \quad \text{on} \quad D - \pi(X). \qquad (\text{LSE}^q_{psr})$$

- Replace α with the probit $\hat{\alpha}$, and γ_q's with the LSE of Y on $\{1,\ X'\hat{\alpha},\ ...,\ (X'\hat{\alpha})^q\}$ to implement LSE^q_{psr}. Let $\gamma \equiv (\gamma_0, \gamma_1, ..., \gamma_q)'$. Set q at $1 \sim 3$ in practice; or modify $\pi(X)$ until $\text{LSE}^0_{psr} = \text{LSE}^1_{psr} = \text{LSE}^2_{psr} \cdots$
- Since the LSE of Y on 1 is $\bar{Y}$, LSE^q_{psr} includes LSE^0_{psr} as a special case when $q = 0$. To ease referencing LSE^0_{psr} and LSE^q_{psr} with $q > 0$, use the expression LSE^q_{psr} only for $q > 0$ henceforth. 'LSE_{psr}' refers to both LSE^0_{psr} and LSE^q_{psr}.

Advantages of PS-Residual LSE and Remarks

- First, LSE_{psr} is possibly the easiest to implement, with hardly any choice required by the user; it is numerically stable.

- Second, it has a simple asymptotic variance estimator that works also well in small samples.

- Third, as will be seen, it can be easily extended to multiple/multi-valued D by replacing $\pi(X)$ with a 'generalized PS'.

- The motivation to extend LSE^0_{psr} to LSE^q_{psr} is to improve LSE^0_{psr} in case PS is misspecified, although LSE_{psr} proceeds on the premise of the correctly specified PS as PSM does—more on this shortly.

- Simply put, LSE_{psr} brings the "time-tested work horse" LSE back to life for binary or multiple treatment while controlling covariates semiparametrically.

Motivating Semi-Linear Parallel-Shift Model

- For an unknown $\mu(\cdot)$, let (this "parallel shift" will be relaxed later):
$$Y^0 = \mu(X) + U,\ \ Y^1 = \beta + Y^0 \implies Y = \beta D + \mu(X) + U,\ \ E(U|X) = 0.$$

- $Y^d \amalg D|X \implies U \amalg D|X \implies U \amalg D|\pi(X)$. Take $E\{\cdot|\pi(X)\}$ on the Y eq.:
$$E\{Y|\pi(X)\} = \beta\pi(X) + E\{\mu(X)|\pi(X)\}.$$

- Hence,
$$Y - E(Y) = \beta\{D - \pi(X)\} + V \qquad \text{where}$$
$$V \equiv \mu(X) - E\{\mu(X)|\pi(X)\} + E\{Y|\pi(X)\} - E(Y) \ + U.$$

- Since V is determined by U with X given,
$$U \amalg D|X \implies V \amalg D|X \implies V \amalg D|\pi(X);\ \text{the proof on the next slide}$$

- $E\{\cdot|\pi(X)\}$ on $\pi(X) \equiv E(D|X)$ gives $\pi(X) = E\{D|\pi(X)\}$. LSE^0_{psr} works:
$$E[\{D - \pi(X)\}V] = E[\ E\{DV - \pi(X)V|\pi(X)\}\ ] = 0.$$

Implementation and Generalization

- Set $\pi(X) = \Phi(X'\alpha)$ to apply probit for α. LSE^0_{psr} is much easier to implement than PSM.
- When PS is misspecified, $COR\{D - \pi(X), V\} \neq 0$ in general, and the omitted X-dependent terms in V result in biases. This may be alleviated if $E\{Y|\pi(X)\}$ is explicitly accounted for by $\Pi^q(Y|X'\alpha)$ in LSE^q_{psr}.
- Using $X'\alpha$ instead of $\Phi(X'\alpha)$ in $\Pi^q(Y|X'\alpha)$ makes the extension to multiple treatments easier.
- In LSE_{psr}, the only decision to make is specifying the PS regression function $X'\alpha$, which is common for all PS-based estimators. For simplicity, proceed with LSE^2_{psr} henceforth, unless otherwise noted.

The proof for $V \amalg D|X \Longrightarrow V \amalg D|\pi(X)$ comes from the 1st & last terms in

$$\begin{aligned} & E\{D|V, \pi(X)\} = E\{E(D|V, X)|V, \pi(X)\} \\ = \quad & E\{E(D|X)|V, \pi(X)\} = \pi(X) = E\{D|\pi(X)\}. \end{aligned}$$

Asymptotic Distribution

- With $\pi(X) \equiv E(D|X)$ and $E(Y|X)$ nonparametrically estimated in the LSE of $Y - E(Y|X)$ on $D - \pi(X)$, the first-stage errors, $\hat{\pi}(X) - \pi(X)$ and $\hat{E}(Y|X) - E(Y|X)$, are orthogonal to the LSE moment condition.

- But for LSE^2_{psr}, the error $\hat{\alpha} - \alpha$ matters, and it holds that

$$\sqrt{N}(\hat{\beta} - \beta) \rightsquigarrow N(0, \Omega) \quad \text{where} \quad \hat{\Omega} \equiv (\frac{1}{N}\sum_i \hat{\varepsilon}_i^2)^{-2} \cdot \frac{1}{N}\sum_i (\hat{V}_i\hat{\varepsilon}_i + \hat{L}\hat{\eta}_i)^2$$

and

$$\begin{aligned}
\hat{\varepsilon}_i &\equiv D_i - \Phi(X_i'\hat{\alpha}), \qquad \hat{V}_i \equiv Y_i - \{\hat{\gamma}_0 + \hat{\gamma}_1 X_i'\hat{\alpha} + \hat{\gamma}_2 (X_i'\hat{\alpha})^2\} - \hat{\beta}\hat{\varepsilon}_i, \\
\hat{\eta}_i &\equiv (\frac{1}{N}\sum_i \hat{s}_i\hat{s}_i')^{-1}\hat{s}_i \quad \text{with} \quad \hat{s}_i \equiv \frac{\hat{\varepsilon}_i \phi(X_i'\hat{\alpha})}{\Phi(X_i'\hat{\alpha})\{1 - \Phi(X_i'\hat{\alpha})\}} X_i, \\
\hat{L} &\equiv -\frac{1}{N}\sum_i \hat{V}_i \phi(X_i'\hat{\alpha}) X_i'.
\end{aligned}$$

- If more polynomial terms of $X'\alpha$ are used for $\Pi^q(Y|X'\alpha)$, the modification needed is adding the extra terms into $\hat{V}$; $\hat{V}_i \equiv Y_i - \bar{Y} - \hat{\beta}\hat{\varepsilon}_i$ in LSE^0_{psr}.

Efficiency Question

- The simulation section will demonstrate that $\hat{\Omega}$ works well in small samples. If desired, one may use nonparametric bootstrap, resampling from the original sample with replacement.

- Hahn (1998, ECA, p. 323) showed that the LSE of $Y - E(Y|X)$ on $D - E(D|X)$ is *not* semiparametrically efficient. This suggests that, with α further estimated, LSE$_{psr}$ would not be semiparametrically efficient.

- Despite the inefficiency, it will be shown by a simulation study that, in finite samples, LSE$_{psr}$ is far more efficient as well as less biased than supposedly efficient estimators.

- This holds despite no user-interventions on LSE$_{psr}$, such as using a caliper in matching or excluding extreme observations with $\pi(X) \simeq 0, 1$ in weighting.

General Model with Heterogeneous Effect

- To relax parallel shift, let, for unknown $\mu(X)$ & $\mu_D(X)$ and errors U^0 & U^1,

$$Y^0 = \mu(X) + U^0, \;\; Y^1 = \mu(X) + \mu_D(X) + U^1, \;\; E(U^d|X) = 0$$
$$\Longrightarrow \;\; Y = \mu(X) + \mu_D(X)D + U, \; U \equiv (1-D)U^0 + DU^1, \; E(U|X,D) = 0.$$

- $E(Y^1 - Y^0|X) = \mu_D(X)$; parallel shift if $\mu_D(X) = \beta, U^0 = U^1$. Omitting U

$$Y - E\{Y|\pi(X)\} = \mu(X) - E\{\mu(X)|\pi(X)\} + \mu_D(X)D - E\{\mu_D(X)D|\pi(X)\}.$$

- Since $D - \pi(X)$ has slope 0, LSE$_{psr}$ '$\hat{\beta}_{psr}$' is consistent for the omitted variable bias due to $\mu_D(X)D - E\{\mu_D(X)D|\pi(X)\}$ that is

$$\beta_\omega \equiv E\{\omega(X)\mu_D(X)\} = E\{\omega(X)E(Y^1 - Y^0|X)\}, \; \omega(X) \equiv \frac{V(D|X)}{E\{V(D|X)\}}.$$

- If interested in the X-conditional effect to model it as $\beta D + \beta_x' XD$ ($\Longrightarrow E(Y^1 - Y^0|X) = \beta + \beta_x' X$), estimate the Y model with OLS and obtain $E\{\omega(X)(\beta + \beta_x' X)\}$: comparing this to LSE$_{psr}$, check the Y model

Why the Weighted Effect is Good

- When the X-conditional effect is $\mu_D(X)$, for the population, it is a matter of how to average X out. In a weighted averaging, higher weights are given to individuals deemed to be more important for the purpose.

- This importance is gauged by f_X in $E\{\mu_D(X)\}$, and by the proximity of $\pi(X)$ to 0.5 in $E\{\omega(X)\mu_D(X)\}$ because $V(D|X) = \pi(X)\{1 - \pi(X)\}$.

- Since $\{1 - \pi(X)\}\pi(X)$ attains its maximum at $\pi(X) = 0.5$ and decreases toward 0 as $\pi(X) \to 0, 1$, those with $\pi(X) \simeq 0.5$ get higher weights (& those with $\pi(X) \simeq 0, 1$ get lower weights). Why is this good?

- First, those with $\pi(X) \simeq 0.5$ are close to being randomized, thus less susceptible to confounding by unobservables; they deserve high weights.

- Second, other estimators have an arbitrary feature to downweight extreme observations with $\pi(X) \simeq 0, 1$, but the $\omega(X)$-weighting of LSE$_{psr}$ is a built-in, non-arbitrary feature to downweight observations with $\pi(X) \simeq 0, 1$.

Non-Continuous Response

- LSE_{psr} works for any response Y, not just continuously distributed Y.

- E.g., suppose $Y = 1[X'\psi + \beta D + N(0,1) > 0]$. Then,

$$\begin{aligned}\mu(X) &= \Phi(X^T\psi), \quad U^0 = Y - \Phi(X^T\psi), \\ \mu_D(X) &= \Phi(X^T\psi + \beta) - \Phi(X^T\psi), \quad U^1 = Y - \Phi(X^T\psi + \beta).\end{aligned}$$

- $\hat{\beta}_{psr} \to^p E[\omega(X)\{\Phi(X'\psi + \beta) - \Phi(X'\psi)\}]$, while typically $E[\{\Phi(X'\psi + \beta) - \Phi(X'\psi)\}]$ is presented as a marginal effect.

- For Y probit, estimating $E[\{\Phi(X'\psi + \beta) - \Phi(X'\psi)\}]$ requires an extra work. In contrast, LSE_{psr} gives $\hat{\beta}_{psr} \to^p E[\omega(X)\{\Phi(X'\psi + \beta) - \Phi(X'\psi)\}]$ directly, with an extra work done for the D probit instead.

- This is fine as long as misspecifications in $\pi(X)$ are less worrisome than those in the Y-model, which is the stance taken in the PS matching literature, as it has chosen to specify $\pi(X)$, instead of $E(Y^d|X)$.

Weighted PS-Residual LSE

- There is a weighted version of LSE_{psr} that is consistent for $\beta = E(Y^1 - Y^0)$.

- Rewrite the general $Y - E\{Y|\pi(X)\}$ equation as, omitting $U/\omega(X)$,

$$\frac{Y - E(Y|X^T\alpha)}{\omega(X)} = \frac{\mu(X) - E\{\mu(X)|\pi(X)\}}{\omega(X)} + \frac{\mu_D(X)D - E\{\mu_D(X)D|\pi(X)\}}{\omega(X)}$$

- Let $\hat{\beta}^{\omega}_{psr}$ denote the (weighted) LSE to this; $\hat{\beta}^{\omega}_{psr} \to^p \beta$ because $\omega(X)^{-1}$ in the omitted variable bias cancels $\omega(X)$ in $E\{\omega(X)E(Y^1 - Y^0|X)\}$.

- Unless $\hat{\pi}(X)$ is well bounded within $(0,1)$, however, the finite sample performance of $\hat{\beta}^{\omega}_{psr}$ is poor due to $\hat{\pi}(X) \simeq 0, 1$ in $\hat{\omega}(X)^{-1}$.

- This can be overcome by using only observations with $\hat{\pi}(X)$ away from 0 and 1, but this brings in arbitrariness. If desired, use $\hat{\beta}^{\omega}_{psr}$ as a reference, discarding observations with $\hat{\pi}(X) \simeq 0, 1$

Multiple LSE for Multiple Treatment

- Suppose D takes on $0, 1, ..., J$. Let $D_d \equiv 1[D = d]$ to consider parallel-shift:

$$Y = \mu(X) + \sum_{d=1}^{J} \beta_d D_d + U \qquad \text{where} \quad E(U|X) = 0.$$

- With $\pi_d(X) \equiv E(D_d|X)$ and $\pi(X) \equiv \{\pi_1(X), ..., \pi_J(X)\}'$,

$$Y - E(Y|\pi(X)) = \sum_{d=1}^{J} \beta_d \{D_d - \pi_d(X)\} + V.$$

- The analog for LSE^0_{psr} is

$$\text{LSE of } Y - \bar{Y} \quad \text{on} \quad D_d - \pi_d(X), \quad d = 1, ..., J.$$

- The analog for LSE^q_{psr} is

$$\text{LSE of } Y - \Pi^q(Y|X'\alpha) \quad \text{on} \quad D_d - \pi_d(X), \quad d = 1, ..., J$$

where $X'\alpha$ can be uni- or multi-dimensional; examples next.

Multiple Treatment Cases

- First, the treatments are *ordered* to be generated by

$$D_i = \sum_{d=1}^{J} 1[\zeta_d \le X_i'\alpha + \varepsilon_i], \quad \zeta_1 = 0 < \zeta_2 <, \cdots, < \zeta_J.$$

  - E.g., D is schooling years. Under $\varepsilon \sim N(0,1) \amalg X$, apply ordered probit to estimate the 'single index' $X'\alpha$. Then use $\Pi^q(Y|X'\alpha)$.

- Second, the treatments are *partly ordered* as in

$$\begin{aligned} D_{0i} &\equiv 1[0 \le X_{0i}'\alpha_0 + \varepsilon_{0i}], \quad D_{ri} \equiv 1 + \sum_{d=1}^{J-1} 1[\zeta_d \le X_{ri}'\alpha_r + \varepsilon_{ri}], \\ \zeta_1 &= 0 < \zeta_2 <, \cdots, < \zeta_{J-1}, \quad D_i \equiv (1 - D_{0i})D_{ri} \text{ taking on } 0, 1, 2, ..., J. \end{aligned}$$

  - E.g., $D_0 = 1$ if not joining military, and $D_r = 1, 2, ..., J$ is military rank. (D_0, D_r) depends on X through $(X_0'\alpha_0, X_r'\alpha_r)$. Use $\Pi^q(Y|X_0'\alpha_0, X_r'\alpha_r)$.

- Third, if D is *multinomial*, J linear indices appear; e.g., D represents job categories.

Other Estimators: Regression Imputation (RI) and PSM

- With $\hat{\pi}(X) \equiv \Phi(X'\hat{\alpha})$, a PS-based 'regression imputation' (RI) estimator is

$$\hat{\beta}_{ri} \equiv \frac{1}{N}\sum_{i=1}^{N}\{\hat{E}(Y|\hat{\pi}(X_i), D=1) - \hat{E}(Y|\hat{\pi}(X_i), D=0)\};$$

$\hat{E}(Y|\hat{\pi}(X_i), D=d)$ is a nonparametric estimator for $E(Y^d|\pi(X_i))$.

- A PS pair-matching estimator for $E(Y^1 - Y^0)$ is

$$\hat{\beta}_{m1} \equiv \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i^1 - \hat{Y}_i^0) \quad \text{with} \quad \hat{Y}_i^1 \equiv D_i Y_i + (1-D_i)Y_{t(i)}$$

$$\hat{Y}_i^0 \equiv (1-D_i)Y_i + D_i Y_{c(i)};$$

$t(i)$ is the matched treated for control i; $c(i)$ matched control for treated i.

- If $Y_{c(i)}$ is replaced by the average of the four nearest controls and if $Y_{t(i)}$ is replaced by the average of the four nearest treated, then 'PS four-multiple-matching estimator' $\hat{\beta}_{m4}$ is obtained.

Other Estimators: Bias-Corrected PSM

- Whereas the above RI and PSM specify $\pi(X)$, not $E(Y^d|X) = E(Y|X, D = d)$, there are estimators specifying $E(Y|X, D = d) = X'\beta_d$ (and $\pi(X)$).

- A bias-corrected version of $\hat{\beta}_{m1}$ (Abadie and Imbens 2011, JBES) is

$$\hat{\beta}_{mbc} \equiv \frac{1}{N}\sum_{i=1}^{N}(\tilde{Y}_i^1 - \tilde{Y}_i^0), \;\; \tilde{Y}_i^1 \equiv D_i Y_i + (1 - D_i)(Y_{t(i)} + X_i'\hat{\beta}_1 - X_{t(i)}'\hat{\beta}_1),$$
$$\tilde{Y}_i^0 \equiv (1 - D_i)Y_i + D_i(Y_{c(i)} + X_i'\hat{\beta}_0 - X_{c(i)}'\hat{\beta}_0).$$

- Matching is not exact (i.e., $X_{t(i)} \neq X_i$ or $X_{c(i)} \neq X_i$) to cause a bias, and adding $X_i'\hat{\beta}_1 - X_{t(i)}'\hat{\beta}_1$ and $X_i'\hat{\beta}_0 - X_{c(i)}'\hat{\beta}_0$ avoids the bias.

- $\hat{\beta}_{mbc}$ differs from Abadie and Imbens (2011): $\hat{\beta}_{mbc}$ uses linear models for $E(Y^d|X)$ while Abadie and Imbens used nonparametric estimators, and $\hat{\pi}(X)$ is used in selecting $t(i)$ and $c(i)$ while X is used in Abadie and Imbens.

Other Estimators: Doubly Robust (DR)

- An inverse-probability-weighted estimator (IPW) is

$$\frac{1}{N}\sum_i\{\frac{D_i}{\hat{\pi}(X_i)} - \frac{1-D_i}{1-\hat{\pi}(X_i)}\}Y_i.$$

- 'Doubly robust' (DR) estimators are consistent if either $\pi(X)$ or $E(Y^d|X)$ is correctly specified, not necessarily both. There are many versions of DR estimator.

- A canonical DR estimator modifying IPW is

$$\hat{\beta}_{dr} \equiv \hat{E}(Y^1) - \hat{E}(Y^0),\ \hat{E}(Y^1) \equiv \frac{1}{N}\sum_i\{\frac{D_iY_i}{\hat{\pi}(X_i)} - \frac{D_i-\hat{\pi}(X_i)}{\hat{\pi}(X_i)}X_i'\hat{\beta}_1\},$$

$$\hat{E}(Y^0) \equiv \frac{1}{N}\sum_i\{\frac{(1-D_i)Y_i}{1-\hat{\pi}(X_i)} - \frac{\hat{\pi}(X_i)-D_i}{1-\hat{\pi}(X_i)}X_i'\hat{\beta}_0\}.$$

Estimators Compared in Simulation

- RI1 & RI2 denote 2 RI estimators with 2 bandwidths. M# denotes PSM with pair or 4 matching; Mbc is the bias corrected version. Let $\hat{\beta}^0_{lse}$, $\hat{\beta}^2_{lse}$ & $\hat{\beta}^4_{lse}$ be LSE^q_{psr} with $q = 0, 2, 4$.

- Abadie and Imbens (2016) noted that Mbc would be DR; the simulation study supports this.

- In total, 9 estimators are compared:

$$\text{RI1 } \hat{\beta}_{ri1}, \quad \text{RI2 } \hat{\beta}_{ri2}, \quad \text{M1 } \hat{\beta}_{m1}, \quad \text{M4 } \hat{\beta}_{m4} : \pi(X) \text{ should be correct;}$$

$$\text{Mbc } \hat{\beta}_{mbc}, \quad \text{DR } \hat{\beta}_{dr} : \text{either } \pi(X) \text{ or } E(Y^d|X) \text{ should be correct;}$$

$$\text{LSE}^0_{psr}\ \hat{\beta}^0_{lse}, \quad \text{LSE}^2_{psr}\ \hat{\beta}^2_{lse}, \quad \text{LSE}^4_{psr}\ \hat{\beta}^4_{lse} : \pi(X) \text{ should be correct.}$$

Simulation Study 1

- The basic simulation design is: with the simulation repetition 10000,

$$D = 1[0 < \alpha_1 + \alpha_2 X_2 + \alpha_3 X_3 + \varepsilon], \qquad \varepsilon \sim N(0,1) \amalg (X_2, X_3),$$
$$(X_2, X_3) \text{ is jointly standard normal with } COR(X_2, X_3) = \sqrt{0.5} \simeq 0.71,$$

-

$$Y = \beta_d D + \beta_1 + \beta_2 X_2 + \beta_3 X_3 + U, \qquad U \sim N(0,1) \amalg (X_2, X_3, \varepsilon),$$
$$\alpha_1 = 0,\ \alpha_2 = 1,\ \alpha_3 = \pm 1,\ \ \beta_1 = 0,\ \beta_d = \beta_2 = \beta_3 = 1,\ \ N = 400, 800.$$

- $E(D) \simeq 0.5$. When $\alpha_3 = -1$, (X_2, X_3) averages around $(-0.2, 0.2)$ and $(0.2, -0.2)$ for the two groups, but when $\alpha_3 = 1$, much further away, around $(-0.7, -0.7)$ and $(0.7, 0.7)$; X overlaps much better in the former.

Simulation Study 2

Table 1. *Base design; both* $\pi(X)$ *&* $E(Y^d\|X)$ *correctly specified*	Good X Overlap ($\alpha_3 = -1$) bias, sd, rmse ($N = 400$)	Poor X overlap ($\alpha_3 = 1$) bias, sd, rmse ($N = 400$)
$\hat{\beta}_{ri1}$	0.00, 0.13, 0.13	0.58, 0.20, 0.61
$\hat{\beta}_{ri2}$	0.00, 0.13, 0.13	0.91, 0.16, 0.92
$\hat{\beta}_{m1}$	0.00, 0.23, 0.23	0.33, 0.33, 0.47
$\hat{\beta}_{m4}$	0.00, 0.17, 0.17	0.47, 0.23, 0.52
$\hat{\beta}_{mbc}$	0.00, 0.15, 0.15	0.00, 0.32, 0.32
$\hat{\beta}_{dr}$	0.00, 0.14, 0.14	0.01, 0.66, 0.66
$\hat{\beta}^0_{psr}$	0.00, 0.12, 0.12	0.00, 0.16, 0.16
$\hat{\beta}^2_{psr}$	0.00, 0.12, 0.12	0.00, 0.15, 0.15
$\hat{\beta}^4_{psr}$	0.01, 0.12, 0.12	0.00, 0.15, 0.15
$\overline{sd}$	*0.12, 0.12, 0.12*	*0.16, 0.15, 0.15*

$\overline{sd}$: average of asymptotic sd estimates for $\hat{\beta}^0_{psr}$, $\hat{\beta}^2_{psr}$, $\hat{\beta}^4_{psr}$.

Simulation Study 3

Table 2. *Poor X-overlap design with $N = 400$ and "tuning"*

	(1) base design bias, sd, rmse	(2) $\pi(X)$ wrong bias, sd, rmse	(3) heterogeneity bias, sd, rmse	(4) binary Y sd=rmse
$\hat{\beta}_{ri1}$	0.47, 0.26, 0.54	0.28, 0.27, 0.39	0.11, 0.15, 0.19	0.056
$\hat{\beta}_{ri2}$	0.66, 0.21, 0.70	0.48, 0.24, 0.54	0.23, 0.14, 0.26	0.050
$\hat{\beta}_{m1}$	0.21, 0.30, 0.36	0.02, 0.19, 0.19	0.02, 0.18, 0.18	0.061
$\hat{\beta}_{m4}$	0.01, 0.17, 0.17	0.00, 0.15, 0.15	0.00, 0.14, 0.14	0.048
$\hat{\beta}_{mbc}$	0.00, 0.32, 0.32	0.01, 0.28, 0.28	0.00, 0.18, 0.18	0.062
$\hat{\beta}_{dr}$	0.00, 0.22, 0.22	0.00, 0.24, 0.24	0.00, 0.16, 0.16	0.053
$\hat{\beta}^{0}_{psr}$	0.00, 0.16, 0.16	0.24, 0.14, 0.28	0.00, 0.13, 0.13	0.044
$\hat{\beta}^{2}_{psr}$	0.00, 0.15, 0.15	-0.01, 0.13, 0.13	0.00, 0.13, 0.13	0.044
$\hat{\beta}^{4}_{psr}$	0.00, 0.15, 0.15	-0.11, 0.13, 0.17	0.00, 0.13, 0.13	0.043
$\overline{sd}$	*0.16, 0.15, 0.15*	*0.21, 0.13, 0.13*	*0.13, 0.12, 0.12*	*0.043*

$\overline{sd}$, average of asymptotic sd estimates for $\hat{\beta}^{0}_{psr}$, $\hat{\beta}^{2}_{psr}$, $\hat{\beta}^{4}_{psr}$;
"tuning" means $\hat{\beta}_{ri1}$ & $\hat{\beta}_{ri4}$ with 4 times smaller bandwidths,
$\hat{\beta}_{m1}$ & $\hat{\beta}_{m1}$ with caliper 0.05, and $\hat{\beta}_{dr}$ only with $0.01 < \hat{\pi}(X) < 0.09$.

Military Rank Effects on Wage: Mean (SD) and LSE

Table 6: Mean (SD) of Variables ($N = 3172$) and LSE

	1356 Non-Veterans	1816 Veterans	LSE (t-value)
1974 wage ($\exp(Y)$)	15,941 (8,083)	15,374 (7,472)	
1974 schooling years	14.5 (2.42)	13.6 (1.93)	0.038 (8.39)
1957 parent wage	6,458 (6,111)	6,330 (5,513)	0.083 (6.36)
1957 # activities	1.40 (1.50)	1.38 (1.47)	0.014 (1.96)
1957 IQ	103 (16.0)	100 (14.5)	0.395 (6.25)
1957 father alive	0.952	0.951	-0.095 (-2.89)
1957 mother alive	0.975	0.977	-0.042 (-1.00)
1957 any religion	0.789	0.758	
1957 friend military	0.097	0.219	
1974 single	0.073	0.059	-0.190 (-3.00)
1974 married	0.875	0.895	0.104 (2.33)
private	.....	0.376	-0.020 (-0.84)
corporal	.....	0.349	0.009 (0.45)
sergeant	.....	0.202	0.008 (0.29)
officer	.....	0.073	0.165 (3.07)
For LSE: $Y = \ln(\text{wage})$, $\ln(\text{parent wage})$, IQ/100 used; $R^2 = 0.131$			

Military Rank Effect on Wage: Estimate (t-value)

Table 7: Military Rank Effect on Wage: $\hat{\beta}$ (tv)

	Private	Corporal	Sergeant	Officer
LSE	-0.020 (-0.84)	0.009 (0.45)	0.008 (0.29)	0.165 (3.07)
LSE^0_{psr}	0.003 (0.031)	-0.025 (-0.69)	-0.047 (-0.43)	0.302 (0.52)
LSE^1_{psr}	*-0.019 (-0.82)*	*0.007 (0.34)*	*0.007 (0.26)*	*0.174 (3.25)*
LSE^2_{psr}	*-0.017 (-0.74)*	*0.009 (0.42)*	*0.009 (0.33)*	*0.171 (3.20)*
LSE^3_{psr}	*-0.016 (-0.70)*	*0.011 (0.50)*	*0.012 (0.43)*	*0.169 (3.15)*
M1	-0.014 (-0.56)	-0.002 (-0.08)	0.033 (1.16)	0.410 (1.48)
M3	-0.012 (-0.47)	0.004 (0.16)	0.023 (0.79)	0.349 (1.11)
M5	-0.007 (-0.29)	0.008 (0.34)	0.029 (0.99)	0.102 (0.37)
M7	-0.009 (-0.37)	0.010 (0.43)	0.020 (0.69)	0.218 (0.90)
RI0.5	-0.006 (-0.28)	-0.010 (-0.35)	-0.015 (-0.52)	0.235 (0.92)
RI1	-0.009 (-0.40)	-0.009 (-0.35)	-0.014 (-0.50)	0.173 (0.98)
RI2	-0.020 (-0.84)	-0.016 (-0.63)	-0.016 (-0.58)	0.266 (2.56)
RI3	-0.033 (-1.37)	-0.025 (-1.06)	-0.019 (-0.68)	0.221 (2.98)

Conclusions

- PS matching is popular in finding the effect of a binary treatment D. But it requires several arbitrary decisions, and the asymptotic inference is difficult.
- This paper brought LSE back to life in finding the effect of D on Y while controlling covariates X semiparametrically.
- LSE_{psr} uses the projection residual of D on PS, and it reduces to the LSE of Y on $(1, D)$ if D is randomized. Extended to multiple treatments.
- First, do the probit of D on X to find $\hat{\alpha}$ for $\Phi(X'\alpha)$. Second, do the LSE of Y on a polynomial function of $X'\hat{\alpha}$, to get the linear projection $\Pi^q(Y|X'\hat{\alpha})$. Third, do the LSE of $Y - \Pi^q(Y|X'\hat{\alpha})$ on $D - \Phi(X'\hat{\alpha})$ for the desired effect.
- LSE_{psr} works far better than other estimators; set q at $1 \sim 3$, or modify $\Phi(X'\alpha)$ until q does not matter. The asymptotic variance estimator is easy to compute and works well in small samples.
- LSE_{psr} is the easiest to use, and it works well in all aspects that matter in practice—*Simplicity is a virtue, not a "sin"*.

Epilogue

- "Mostly Harmless Econometrics" by Angrist and Pischke (2009, Princeton U. Press) is popular among practitioners—for a good reason.

- In 2016, John Rust published an essay "Mostly Useless Econometrics? Assessing the Causal Effect of Econometric Theory" in little known journal *Foundations and Trends in Accounting*.

- There are many messages in the paper, but the main message is "let's do useful econometrics"; otherwise, econometrics may become marginalized, alienating practitioners to become an irrelevant science.

- One example cited is partial identification, which led to empirical helplessness of "Nothing in, Nothing out".

- Imposing a little assumption can go a long way toward providing informative and useful scientific findings that matter to our daily life. Let's do simple & sensible things, instead of "nobody-but-a-few-can-understand" things.