

Mostly Harmless Econometrics

Session 3: Propensity Scores and the Efficiency Paradox

Balancing $\rightarrow$ Weighting $\rightarrow$ Efficiency $\rightarrow$ Estimated Scores

Kaiwen Luo

Washington University in St. Louis

October 16, 2026

The Central Puzzle

Suppose we observe a binary treatment D , an outcome Y , and rich baseline covariates X .

Three seemingly conflicting claims

- 1 Controlling for the **scalar** propensity score $p(X)$ is sufficient for causal identification.
- 2 Controlling for the **full vector** X may yield more precise estimates.
- 3 Replacing the **true** propensity score with an **estimated** one can improve an IPW estimator.

Identification $\neq$ Estimand $\neq$ Statistical efficiency

Roadmap: Rosenbaum–Rubin (1983) $\rightarrow$ Hahn (1998) $\rightarrow$ Angrist–Hahn (2004) $\rightarrow$ Hirano–Imbens–Ridder (2003).

Start with Potential Outcomes and Conditional Independence

$$Y = DY_1 + (1 - D)Y_0, \quad D \in \{0, 1\}, \quad p(X) = \mathbb{P}(D = 1 \mid X).$$

Conditional Independence Assumption (CIA)

$$(Y_0, Y_1) \perp D \mid X.$$

Conditional on observed baseline characteristics, treatment is as good as randomly assigned.

With overlap, $0 < p(X) < 1$ almost surely, define

$$m_d(X) = \mathbb{E}[Y \mid D = d, X] = \mathbb{E}[Y_d \mid X], \quad \text{ATE} = \mathbb{E}[m_1(X) - m_0(X)].$$

CIA is a substantive identifying assumption, not a statistical consequence of estimating a score.

Rosenbaum–Rubin: A One-Dimensional Balancing Score

Propensity Score Theorem (Rosenbaum and Rubin, 1983)

$$(Y_0, Y_1) \perp D \mid X \implies (Y_0, Y_1) \perp D \mid p(X).$$

- A high-dimensional control vector can be reduced to one number for *identification*.
- People with the same $p(X)$ need not have the same covariate values.
- Equal scores balance *distributions of covariates* across treatment status, not the outcomes of individual matched pairs.

Example: job training

A person's age, education, employment history, and family background may jointly determine enrollment. For identification, matching on the correctly specified $p(X)$ can replace exact matching on every characteristic.

Why Does the Balancing-Score Theorem Hold?

For $j \in \{0, 1\}$, it suffices to show that conditioning on Y_j does not change the treatment probability once we know $p(X)$:

$$\begin{aligned}\mathbb{P}(D = 1 \mid Y_j, p(X)) &= \mathbb{E}[\mathbb{E}[D \mid Y_j, X, p(X)] \mid Y_j, p(X)] \\ &= \mathbb{E}[\mathbb{E}[D \mid Y_j, X] \mid Y_j, p(X)] \\ &\stackrel{\text{CIA}}{=} \mathbb{E}[\mathbb{E}[D \mid X] \mid Y_j, p(X)] \\ &= \mathbb{E}[p(X) \mid Y_j, p(X)] = p(X).\end{aligned}$$

Takeaway

Conditioning on $p(X)$ eliminates selection on X for identifying potential-outcome means. It does **not** claim that $p(X)$ captures everything that predicts Y .

Two Implementations: Matching and IPW

Propensity-score matching: compare treated and untreated observations with similar scores.

$$ATT = \mathbb{E}[\mathbb{E}[Y \mid D = 1, p(X)] - \mathbb{E}[Y \mid D = 0, p(X)] \mid D = 1].$$

Inverse probability weighting (IPW): reweight the observed outcomes to reconstruct potential-outcome means.

$$\mathbb{E}[Y_1] = \mathbb{E}\left[\frac{DY}{p(X)}\right], \quad \mathbb{E}[Y_0] = \mathbb{E}\left[\frac{(1-D)Y}{1-p(X)}\right].$$

$$\boxed{ATE = \mathbb{E}\left[\frac{DY}{p(X)} - \frac{(1-D)Y}{1-p(X)}\right].}$$

Intuition

A treated observation who was unlikely to be treated gets greater weight, $1/p(X)$; an untreated observation who was unlikely to be untreated gets $1/(1-p(X))$.

What Does Regression Weight?

Consider a population regression of Y on D plus a *saturated* set of controls for X . By regression anatomy,

$$\beta_R = \frac{\mathbb{E}[(D - p(X))Y]}{\mathbb{E}[(D - p(X))^2]}.$$

Write the conditional mean as

$$\mathbb{E}[Y \mid D, X] = m_0(X) + \tau(X)D, \quad \tau(X) = m_1(X) - m_0(X).$$

Then

$$\beta_R = \frac{\mathbb{E}[p(X)(1 - p(X))\tau(X)]}{\mathbb{E}[p(X)(1 - p(X))]}.$$

Key insight from Section 3.3.1

OLS gives the greatest weight to covariate cells with balanced treated/control frequencies: $p(X) \approx 1/2$.

Same Conditional Effects, Different Target Populations

Under CIA and overlap, let $\tau(X) = \mathbb{E}[Y_1 - Y_0 \mid X]$.

Estimand	Formula	Relative weight on X
ATE	$\mathbb{E}[\tau(X)]$	1
ATT	$\frac{\mathbb{E}[p(X)\tau(X)]}{\mathbb{E}[p(X)]}$	$p(X)$
Saturated OLS	$\frac{\mathbb{E}[p(X)(1 - p(X))\tau(X)]}{\mathbb{E}[p(X)(1 - p(X))]}$	$p(X)(1 - p(X))$

- If treatment effects are constant, all three targets coincide.
- With heterogeneous effects, different results can reflect *different estimands*, not estimator failure.

Paradox I: Sufficient for Identification, Not for Efficiency

Question: If $p(X)$ suffices for identification, why retain the full X ?

Job-training example

Let $X = (X_1, X_2)$:

- X_1 : prior unemployment predicts participation in training;
- X_2 : ability strongly predicts earnings, but does *not* predict participation after controlling for X_1 .

$$\mathbb{P}(D = 1 \mid X_1, X_2) = p(X_1), \quad \mathbb{E}[Y_d \mid X_1, X_2] \text{ varies with } X_2.$$

Matching on $p(X_1)$ is sufficient for identification. Yet adjusting for X_2 can reduce unexplained outcome variation.

A balancing score is not an outcome-sufficient statistic.

Hahn (1998): discarding outcome-predictive covariates can carry an asymptotic efficiency cost.

An Extreme Illustration: Precision Without Confounding

Suppose $D \perp X_2$, and outcomes satisfy

$$Y = \tau D + 10X_2 + \varepsilon, \quad \mathbb{E}[\varepsilon \mid D, X_2] = 0.$$

Because X_2 does not predict treatment, omitting it need not bias the treatment coefficient.

Without outcome adjustment

$$Y = \alpha + \beta D + u,$$

u includes $10X_2$; residual noise can be large.

With outcome adjustment

$$Y = \alpha + \beta D + \gamma X_2 + v.$$

Less residual noise can deliver a smaller standard error.

No bias reduction is needed for an efficiency gain. This motivates efficient methods that combine treatment and outcome information.

Paradox II: Why Compress Covariates in a Finite Sample?

Question: If full-covariate adjustment can be more efficient asymptotically, why use a propensity score? Suppose there are 20 binary pre-treatment covariates:

$$\underbrace{2^{20}}_{1,048,576} \text{ potential covariate cells.}$$

- Exact matching on X may leave many sparse or empty cells.
- A parsimonious model $p(X) = \Lambda(X'\gamma)$ replaces multidimensional cells by a scalar index.
- With a credible specification, smoothing can decrease finite-sample estimation noise.

Angrist and Hahn (2004): a finite-sample perspective

Dimension reduction can improve precision when excluded predictors contribute little to outcome variation relative to the burden of estimating their effects.

Trade-off: potentially useful outcome information versus complexity and sampling noise (plus misspecification risk).

Paradox III: How Can Estimating a Known Score Help?

Question: Why can IPW with $\hat{p}(X)$ outperform IPW with the *true* $p(X)$? Start with the Horvitz–Thompson estimator:

$$\hat{\tau}_{HT}(p) = \frac{1}{n} \sum_{i=1}^n \left[\frac{D_i Y_i}{p(X_i)} - \frac{(1 - D_i) Y_i}{1 - p(X_i)} \right].$$

Hirano, Imbens, and Ridder (2003)

Under suitable conditions, an IPW estimator using a *nonparametrically estimated* propensity score can achieve the semiparametric efficiency bound for the ATE.

This is **not** a claim that knowing less is better. It compares *specific estimation procedures*: a simple HT estimator with a known score versus an estimator that exploits sample treatment-balance information.

A Transparent Example: Randomized Treatment with $p = 1/2$

Assume Bernoulli random assignment with $\mathbb{P}(D = 1) = 1/2$. Let $n_1 = \sum_i D_i$, $n_0 = n - n_1$. Using the *known* assignment probability:

$$\hat{\tau}_{HT}(1/2) = \frac{2}{n} \sum_i D_i Y_i - \frac{2}{n} \sum_i (1 - D_i) Y_i.$$

Using the estimated assignment share $\hat{p} = n_1/n$:

$$\begin{aligned} \hat{\tau}_{HT}(\hat{p}) &= \frac{1}{n} \sum_i \frac{D_i Y_i}{n_1/n} - \frac{1}{n} \sum_i \frac{(1 - D_i) Y_i}{n_0/n} \\ &= \boxed{\bar{Y}_1 - \bar{Y}_0}. \end{aligned}$$

Intuition

Estimated probabilities automatically normalize by the *realized* group sizes, eliminating avoidable fluctuations associated with random treatment counts.

Illustration, not the general proof of the HIR efficiency theorem.

The Influence-Function Explanation

Write $m_d(X) = \mathbb{E}[Y \mid D = d, X]$ and $\tau = \mathbb{E}[m_1(X) - m_0(X)]$. The influence function of simple IPW with the true score is

$$\phi_{HT}(O) = \frac{DY}{p(X)} - \frac{(1-D)Y}{1-p(X)} - \tau.$$

The efficient influence function in the nonparametric observational-data model is

$$\phi_{eff}(O) = m_1(X) - m_0(X) - \tau + \frac{D(Y - m_1(X))}{p(X)} - \frac{(1-D)(Y - m_0(X))}{1-p(X)}.$$

Subtract:

$$\phi_{HT}(O) - \phi_{eff}(O) = (D - p(X)) \left[\frac{m_1(X)}{p(X)} + \frac{m_0(X)}{1-p(X)} \right].$$

Meaning

The raw HT estimator carries extra treatment-assignment fluctuation. Proper score estimation can implicitly remove this component, much like calibration or regression adjustment.

Why the “Paradox” Does Not Violate Information Theory

- **Known score vs. unknown score:** information about assignment cannot make the optimal achievable precision worse.
- **Known-score HT vs. estimated-score IPW:** these are different *estimators*; the former does not generally use all available information efficiently.
- **Hirano–Imbens–Ridder:** a sufficiently rich and carefully estimated score can make IPW asymptotically efficient under their regularity conditions.

A useful alternative

Augmented IPW (AIPW) explicitly combines assignment and outcome models:

$$\hat{\tau}_{AIPW} = \frac{1}{n} \sum_i \left[\hat{m}_1(X_i) - \hat{m}_0(X_i) + \frac{D_i(Y_i - \hat{m}_1(X_i))}{\hat{p}(X_i)} - \frac{(1 - D_i)(Y_i - \hat{m}_0(X_i))}{1 - \hat{p}(X_i)} \right].$$

AIPW is a modern connection, beyond the core presentation in Section 3.3.2.

A Subtle Point: Nonparametric Scores May Not Reduce Dimension

Suppose X is discrete and we estimate a separate propensity score in every cell:

$$\hat{p}(x) = \frac{n_{1x}}{n_x}.$$

Then, for each cell with both treated and control observations,

$$\frac{1}{n} \sum_{i: X_i=x} \frac{D_i Y_i}{\hat{p}(x)} = \frac{n_x}{n} \bar{Y}_{1x}, \quad \frac{1}{n} \sum_{i: X_i=x} \frac{(1 - D_i) Y_i}{1 - \hat{p}(x)} = \frac{n_x}{n} \bar{Y}_{0x}.$$

So the plug-in IPW ATE equals

$$\hat{\tau}_{IPW} = \sum_x \frac{n_x}{n} (\bar{Y}_{1x} - \bar{Y}_{0x}),$$

exactly the fully stratified covariate-matching estimator on common-support cells. **Angrist's point:** the actual payoff comes from meaningful smoothing, restrictions, and prior knowledge—not merely renaming X as $\hat{p}(X)$.

Three Papers, Three Different Efficiency Comparisons

Paper	Comparison	Main lesson
Hahn (1998)	Score-only versus outcome-informed adjustment	Score is sufficient for identification, not necessarily precision.
Angrist & Hahn (2004)	Rich controls versus restricted-score methods in finite samples	Useful restrictions may reduce estimation noise.
Hirano, Imbens & Ridder (2003)	True-score HT versus nonparametric estimated-score IPW	Estimating scores can generate efficiency gains.

Different comparisons $\implies$ no genuine contradiction.

Takeaways: A Practical Decision Sequence

- 1 **Identification:** Is $(Y_0, Y_1) \perp D \mid X$ credible? Is there overlap?
- 2 **Estimand:** Do we want ATE, ATT, or a variance-weighted regression average?
- 3 **Modeling:** What do we know about treatment assignment? Which covariates predict outcomes?
- 4 **Estimation:** Does score compression help with sparse cells? Is score specification credible?
- 5 **Inference:** Account for score estimation and potentially extreme inverse-probability weights.

The overarching message

**Use propensity scores to understand and model assignment,
not because they are automatically superior to regression.**

References and Scope

- Angrist, J. D., and J.-S. Pischke. *Mostly Harmless Econometrics*, §3.3.2, pp. 80–86.
- Rosenbaum, P. R., and D. B. Rubin (1983). “The Central Role of the Propensity Score in Observational Studies for Causal Effects.”
- Hahn, J. (1998). “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects.”
- Angrist, J. D., and J. Hahn (2004). “When to Control for Covariates? Panel Asymptotics for Estimates of Treatment Effects.”
- Hirano, K., G. W. Imbens, and G. Ridder (2003). “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score.”

Source organization follows Section 3.3.2. The Bernoulli-randomization example, influence-function decomposition, and AIPW link are added pedagogical explanations.