True versus estimated propensity scores with discrete controls: a finite sample analysis

ANDREW HERREN[1,a] and P. RICHARD HAHN[2,b]

[1] *University of Texas,* [a] *andrew.herren@mccombs.utexas.edu*
[2] *Arizona State University,* [b] *prhahn@asu.edu*

The propensity score – the conditional probability of being assigned to treatment – is a key quantity for estimating causal effects from observational data and is a sufficient control covariate for doing so. Existing asymptotic theory shows that estimated propensity scores are more efficient than true propensity scores in the context of inverse propensity weight (IPW) estimators. This paper derives an elementary, but non-trivial, finite-sample analogue of this result in the case of discrete control variables with finite support. Explicit variance calculations help to separate the fundamental statistical challenges of efficient deconfounding, even in the "trivial" case of stratification estimators, from the idiosyncratic behavior of the true-propensity IPW estimator.

Keywords: Causal Inference; Propensity Score; Regression Adjustment; Variable Selection

1. Introduction

1.1. Background

Rosenbaum and Rubin (1983) showed that if a set of covariates satisfy conditional unconfoundedness, the univariate propensity score based on these controls (the conditional expectation of treatment assignment), also satisfies conditional unconfoundedness. Several classic papers have since shown that *even when the true propensity function is known*, the treatment effect can be estimated with greater efficiency using nonparametric estimates of the propensity function (Robins, Mark and Newey (1992); Hahn (1998); Hirano, Imbens and Ridder (2003)). This paper derives a finite-sample analogue of these results in the case of discrete control variables with finite support. On its face, these results would seem to imply that, with a suitably flexible model of treatment given controls, knowledge of the true propensity function is useless. While this is narrowly true for inverse-propensity weighting (IPW), variance analysis of a modified IPW shows that incorporating knowledge of the true propensity score can improve statistical precision in some cases.

Our analysis of this famous puzzle differs from earlier investigations (e.g. Lunceford and Davidian (2004) or Graham (2011)) in providing exact finite sample results in the case of discrete controls whereas earlier work considers asymptotic efficiency. We obtain elementary formulae by applying an elegant probability result from Chao and Strawderman (1972), thereby concretely demonstrating the finite-sampling inferiority of the true-propensity IPW estimator in the intuitive setting of discrete control variables[1] .

[1] An earlier draft of this manuscript was posted to ArXiV with a slightly different title (Herren and Hahn, 2023).

1.2. Notation

Let Z denote a binary treatment, X denote a discrete covariate taking k unique values, $\mathcal{X} = \{1, \ldots, k\}$, and Y denote the outcome of interest. We are interested in the average causal effect of Z on Y, which can be expressed in terms of two counterfactual random variables, Y^1 and Y^0, the outcomes in which the treatment Z is set to 1 and 0, respectively. We make several standard identifying assumptions:

- Stable unit treatment value assumption (SUTVA)
  - *Consistency*: $Y = Y^1 Z + Y^0(1 - Z)$
  - *No Interference*: For any two indices $1 \leq i, j \leq n$ in a dataset of size n, $(Y_i^1, Y_i^0) \perp\!\!\!\perp Z_j$
- Positivity: $0 < \mathbb{P}(Z = 1 \mid X = x) < 1$ for all $x \in \mathcal{X}$
- Conditional unconfoundedness: $(Y^1, Y^0) \perp\!\!\!\perp Z \mid X$

These assumptions allow us to identify the average treatment effect (ATE) as an estimable contrast:

$$\mathbb{E}[Y^1 - Y^0] = \mathbb{E}_X[\mathbb{E}[Y \mid X, Z = 1] - \mathbb{E}[Y \mid X, Z = 0]].$$

The propensity score, $p(X) = \mathbb{P}(Z = 1 \mid X)$, likewise satisfies conditional unconfoundedness:

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

Many estimators target this estimand; in this paper we focus on inverse-propensity weight estimators for arbitrary weighting function w:

$$\bar{\tau}^w_{\text{ipw}} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{Y_i Z_i}{w(X_i)} - \frac{Y_i(1 - Z_i)}{1 - w(X_i)} \right). \tag{1}$$

Note that the weight function itself may be estimated from the sample data.

2. Variance comparison of true versus estimated IPW

Throughout it will be assumed that every strata of $\mathcal{X}$ has at least one treated-control pair: $N_{x,1} > 0$ and $N_{x,0} > 0$ for all x. While this makes deconfounding conceptually straightforward, it does not render the problem statistically trivial, yielding estimators with substantial sampling variance relative to the magnitude of the ATE.

Direct calculation shows that for a putative propensity function $w : \mathcal{X} \mapsto (0, 1)$, the corresponding IPW estimator may be written as

$$\bar{\tau}^w_{\text{ipw}} = \sum_x \left(\frac{N_x}{N} \right) \bar{\tau}^{w,x}_{\text{ipw}} \tag{2}$$

where

$$\bar{\tau}^{w,x}_{\text{ipw}} = \left(\frac{\hat{p}(x)}{w(x)} \bar{Y}_{x,Z=1} - \frac{1 - \hat{p}(x)}{1 - w(x)} \bar{Y}_{x,Z=0} \right) \tag{3}$$

and $\hat{p}(x) = \left(N_{x,Z=1} / N_x \right)$ is the proportion of treated units in each stratum.

Taking $w(x) = p(x)$ is the "true propensity score" case, while letting $w(x) = \hat{p}(x)$ is the "estimated propensity score" case. In the former case, the treated and untreated stratum averages are weighted by $\hat{p}(x)/p(x)$ and $\hat{q}(x)/q(x) = (1 - \hat{p}(x))/(1 - p(x))$, respectively; in the latter case the weights are identically one. This difference in weights leads to the following analogue of the results of Hahn (1998), Hirano, Imbens and Ridder (2003), and Robins, Mark and Newey (1992):

Proposition 1. *The variance difference between the true propensity score IPW and the estimated propensity score IPW may be expressed as*

$$\begin{aligned}\sum_{x\in\mathcal{X}} \mathbb{V}(\bar{\tau}^{p,x}_{ipw}) - \mathbb{V}(\bar{\tau}^{\hat{p},x}_{ipw}) &= \sum_{x\in\mathcal{X}} \left(\frac{\mu_1(x)}{p(x)} + \frac{\mu_0(x)}{1-p(x)}\right)^2 \frac{n_x p(x)(1-p(x))}{(n_x+2)^2} \\ &+ \sum_{x\in\mathcal{X}} \sigma^2_{x,1}\left(\frac{p(x)n_x+1}{(n_x+2)^2 p(x)^2} - \frac{1-(1-p(x))^{(n_x+1)}}{(n_x+1)p(x)}\right) \\ &+ \sum_{x\in\mathcal{X}} \sigma^2_{x,0}\left(\frac{(1-p(x))n_x+1}{(n_x+2)^2(1-p(x))^2} - \frac{1-p(x)^{(n_x+1)}}{(n_x+1)(1-p(x))}\right)\end{aligned} \tag{4}$$

where $n_x = N_x - 2$.

We arrive at this expression by deriving it for an individual stratum, suppressing the dependence on x for notational convenience. By the Law of Total Variance

$$\begin{aligned}\mathbb{V}\left(\frac{\hat{p}}{w}\bar{Y}_1 - \frac{1-\hat{p}}{1-w}\bar{Y}_0\right) &= \mathbb{V}\left(\mathbb{E}\left(\frac{\hat{p}}{w}\bar{Y}_1 - \frac{1-\hat{p}}{1-w}\bar{Y}_0 \mid Z\right)\right) + \mathbb{E}\left(\mathbb{V}\left(\frac{\hat{p}}{w}\bar{Y}_1 - \frac{1-\hat{p}}{1-w}\bar{Y}_0 \mid Z\right)\right) \\ &= \mathbb{V}\left(\frac{\hat{p}}{w}\mu_1 - \frac{1-\hat{p}}{1-w}\mu_0\right) + \mathbb{E}\left(\left(\frac{\hat{p}}{w}\right)^2 \frac{\sigma^2_1}{N_1} + \left(\frac{1-\hat{p}}{1-w}\right)^2 \frac{\sigma^2_0}{N_0}\right).\end{aligned} \tag{5}$$

The random variable in these expressions is $\hat{p} = N_1/N$ and $\hat{q} = 1 - \hat{p} = N_0/N$, respectively. We handle each term separately.

When $w = N_1/N$ (itself a random variable), as in the estimated IPW, then the first term becomes 0 (the variance of a constant). Meanwhile, when the weighting function is fixed at $w = p$ (and $1 - w = 1 - p = q$) we find that

$$\begin{aligned}\mathbb{V}\left(\frac{\hat{p}}{p}\mu_1 - \frac{1-\hat{p}}{1-p}\mu_0\right) &= \frac{\mu_1^2}{p^2N^2}\mathbb{V}(N_1) + \frac{\mu_0^2}{q^2N^2}\mathbb{V}(N_0) - 2\frac{\mu_1\mu_0}{pqN^2}\mathrm{Cov}\,(N_1, N_0) \\ &= \left(\frac{\mu_1^2}{p^2N^2} + \frac{\mu_0^2}{q^2N^2} + 2\frac{\mu_1\mu_0}{pqN^2}\right)npq, \\ &= \left(\frac{\mu_1}{p} + \frac{\mu_0}{q}\right)^2 \frac{npq}{(n+2)^2},\end{aligned} \tag{6}$$

for $n = N - 2$ (because our observations are constrained to have at least one treated and untreated unit) and using the fact that $N_1 = N - N_0$ and therefore $\mathrm{Cov}\,(N_1, N_0) = -\mathbb{V}(N_1) = -\mathbb{V}(N_0)$.

On to the second term, we consider

$$\mathbb{E}\left(\frac{\sigma^2_1}{N_1} + \frac{\sigma^2_0}{N_0}\right)$$

and

$$\mathbb{E}\left(\left(\frac{\hat{p}}{p}\right)^2 \frac{\sigma^2_1}{N_1} + \left(\frac{1-\hat{p}}{1-p}\right)^2 \frac{\sigma^2_0}{N_0}\right)$$

in turn. We apply, in each case, a result of Chao and Strawderman (1972) about negative moments of a positive random variable A; that is, expectations of the form $\mathbb{E}\,(1/(c+A))$ for c a known constant.

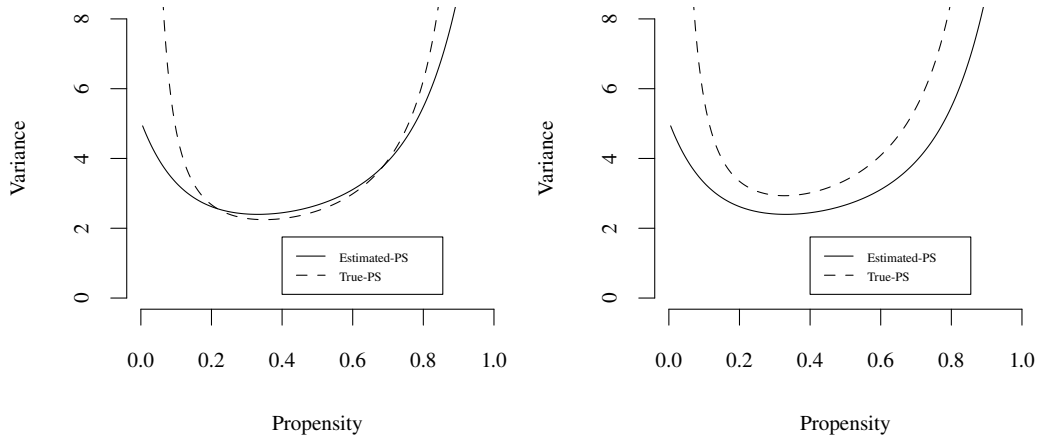

Figure 1. A comparison of the finite sample variances when $N_x = 17$, with $\sigma_1^2 = 4$ and $\sigma_0^2 = 16$. Left panel: $\mu_1 = \mu_0 = 0$. Right panel: $\mu_1 = 1$ and $\mu_0 = 3$.

Specifically, for $A \sim \text{Bin}(n, p)$ and $c = 1$, they show that

$$\mathbb{E}\left(\frac{1}{(1+A)}\right) = \frac{1-q^{(n+1)}}{(n+1)p}.$$

We apply this result to the estimated propensity IPW by writing $N_1 = 1 + A$ where A has parameters $n = N - 2$ and p. The number of observed controls, N_0 has a similar representation, but with parameter $q = 1 - p$. Therefore

$$\mathbb{E}\left(\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}\right) = \sigma_1^2 \frac{1-q^{(n+1)}}{(n+1)p} + \sigma_0^2 \frac{1-p^{(n+1)}}{(n+1)q}. \tag{7}$$

The true propensity IPW is more straightforward, but uses the same representation of the treated (respectively, control) counts as $1 + A$ for a binomial random variable with parameters $n = N - 2$ and p (respectively, $q = 1 - p$):

$$\begin{aligned}\mathbb{E}\left(\left(\frac{\hat{p}}{p}\right)^2 \frac{\sigma_1^2}{N_1} + \left(\frac{1-\hat{p}}{1-p}\right)^2 \frac{\sigma_0^2}{N_0}\right) &= \frac{\sigma_1^2}{N^2p^2}\mathbb{E}(N_1) + \frac{\sigma_1^2}{N^2(1-p)^2}\mathbb{E}(N_0) \\ &= \sigma_1^2 \frac{(pn+1)}{(n+2)^2p^2} + \sigma_0^2 \frac{(qn+1)}{(n+2)^2q^2}.\end{aligned} \tag{8}$$

Gathering like terms and summing over strata, indexed by x, demonstrates the proposition.

Figure 1 illustrates these variances as a function of p when $\sigma_1^2 = 4$ and $\sigma_0^2 = 16$. In the left panel we observe that when the data generating process has zero mean for both potential outcomes $\mu_1 = \mu_0 = 0$, the true propensity estimator can have lower variance for a certain range of p; the asymmetry in the plot is due to differing potential outcome variances, $\sigma_1^2 \neq \sigma_0^2$. The right panel shows that if either potential outcome surface is sufficiently far from zero, the quadratic term dominates and the estimated propensity IPW has lower variance than the true propensity IPW for all values of p.

3. Known propensity IPW as variable selection

Observe that for two strata, which we will denote x_a and x_b, the true-propensity IPW effectively collapses cells whenever $p(x_a) = p(x_b)$:

$$\begin{aligned}
\frac{N_a}{N}\bar{\tau}^{p,a} + \frac{N_b}{N}\bar{\tau}^{p,b} &= \frac{N_a}{N}\left(\frac{\hat{p}_a}{p}\bar{Y}_{a,1} - \frac{1-\hat{p}_a}{1-p}\bar{Y}_{a,0}\right) + \frac{N_b}{N}\left(\frac{\hat{p}_b}{p}\bar{Y}_{b,1} - \frac{1-\hat{p}_b}{1-p}\bar{Y}_{b,0}\right) \\
&= \frac{N_a}{N}\left(\frac{N_{a,1}/N_a}{p}\frac{S_{a,1}}{N_{a,1}} - \frac{N_{a,0}/N_a}{p}\frac{S_{a,0}}{N_{a,0}}\right) + \\
&\quad \frac{N_b}{N}\left(\frac{N_{a,1}/N_a}{p}\frac{S_{a,1}}{N_{a,1}} - \frac{N_{a,0}/N_a}{p}\frac{S_{a,0}}{N_{a,0}}\right) \\
&= \frac{1}{N}\left(\frac{S_{a,1}+S_{b,1}}{p} - \frac{S_{a,0}+S_{b,0}}{1-p}\right) \\
&= \left(\frac{\hat{p}}{p}\bar{Y}_1 - \frac{1-\hat{p}}{1-p}\bar{Y}_0\right)
\end{aligned} \tag{9}$$

where, for notational ease, we write $N = N_a + N_b$ and $p = p(x_a) = p(x_b)$. This demonstrates that, for discrete control variables, IPW estimators are fundamentally stratification estimators.

Thus, consider an estimated-propensity IPW, but restricted to the unique values of $p(\mathcal{X})$ rather than all the unique levels of $\mathcal{X}$. That is, knowledge of the true propensity function may be used to define the strata without insisting that the finite-sample deviations in treatment assignment be ignored. Is such a "hybrid" IPW ever preferred over the estimated PS approach? Direct variance calculation reveals that, yes, the true propensity function can improve estimation in some cases.

Here we demonstrate these calculations for two strata. The hybrid, collapsed strata, estimator has variance $\mathbb{E}(\sigma_1^2/N_1 + \sigma_0^2/N_0)$, while the conventional, uncollapsed, estimator has variance

$$\frac{N_a^2}{N^2}\mathbb{E}\left(\frac{\sigma_{a,1}^2}{N_{a,1}} + \frac{\sigma_{a,0}^2}{N_{a,0}}\right) + \frac{N_b^2}{N^2}\mathbb{E}\left(\frac{\sigma_{b,1}^2}{N_{b,1}} + \frac{\sigma_{b,0}^2}{N_{b,0}}\right).$$

In general, which of these is smaller depends on the data generating process, but the collapsed estimator can certainly obtain lower variance if it omits extraneous controls. That is, consider the case where the means and variances for both potential outcomes are constant across the strata so that $\sigma_{a+b,z}^2 = \sigma_{a,z}^2 = \sigma_{b,z}^2$ for $z \in \{0,1\}$ and further suppose that $N_a = N_b$ for simplicity. In that case, it can be shown that the hybrid (collapsing, or variable selecting) estimator has lower variance by showing that

$$\frac{\sigma_z^2}{4}\mathbb{E}\left(\frac{1}{N_{a,z}} + \frac{1}{N_{b,z}}\right) = \frac{\sigma_z^2}{2}\mathbb{E}\left(\frac{1}{N_{a,z}}\right) \geq \sigma_z^2\mathbb{E}\left(\frac{1}{N_{a,z}+N_{b,z}}\right).$$

We first re-express this inequality so that the formulas from Chao and Strawderman (1972) may again be applied, this time for both $c = 1$ and $c = 2$. Let $N_a = N_b$. Because we require that each stratum has at least one treatment-control contrast, we have that $N_{a,1} = 1 + A$, where A is a binomial random variable with parameter $n = N_a - 2$. Similarly $N_{b,1} = 1 + B$ for B an independent random variable with the same distribution as A, because $p(x_a) = p(x_b)$ in the present example. Additionally, this implies that $N_{a+b,1} = 2 + A + B$ for the same random variables, and we note that $A + B$ is a binomial random variable with parameter $N_a + N_b - 4 = 2n$. Therefore, showing that the collapsed estimator has lower

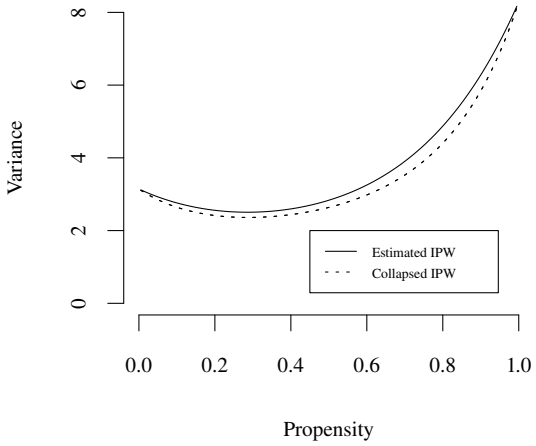

Figure 2. A comparison of the finite sample variances when $N_x = 17$, with $\sigma_1^2 = 4$ and $\sigma_0^2 = 16$ for the collapsed and "non-collapsed" estimators.

variance amounts to showing that

$$\mathbb{E}\left(\frac{1}{2(1+A)}\right) - \mathbb{E}\left(\frac{1}{2+A+B}\right) = \frac{1-q^{(n+1)}}{2(n+1)p} - \left(\frac{1}{(2n+1)p} - \frac{\left(1-q^{(2n+2)}\right)}{(2n+1)(2n+2)p^2}\right)$$

is nonnegative for all $p \in (0,1)$, where $q = 1-p$. We verify the nonnegativity of this difference in Appendix 1; Figure 2 illustrates the inequality.

Note that, although the difference between the estimator variances appears small, it can be made arbitrarily large by considering additional extraneous stratification; these equations only compare the variance increase due to a single extraneous stratum.

While the above calculation shows that a variable-selected IPW *can* improve upon an estimated-propensity IPW, the result is obtained by considering response-irrelevant strata. In particular, unnecessary propensity strata *could* improve treatment effect estimation if they facilitated more precise estimation of potential outcome means. That is, "estimating the propensity scores" using variables known to be irrelevant for treatment assignment can nonetheless increase statistical precision; in such cases the IPW estimator is acting as a direct regression adjustment in disguise. Conversely, a known propensity function does not discern *instruments* — variables that only impact the outcome via their causal effect on treatment assignment. But the above variance comparisons equally as well apply to comparing between two distinct valid propensity functions, one with instruments and one without. By definition, the instrumental strata have identical potential outcome means and variances and so are "noise" variables in the relevant sense.

4. Discussion

This paper has considered stratification estimators in the case of discrete control variables with finite support. Discrete covariates are both common in practice (indeed, more common than continuous covariates) and pedagogically illuminating, and therefore worthy of careful study. However, in the main, previous literature has eschewed this approach. For instance, we read in the textbook of Imbens and Rubin (2015) (Section 12.2.2):

> If...we view the covariates as having a discrete distribution with finite support, the implication of unconfoundedness is simply that one should stratify by the values of the covariates. In that case there will be, with high probability, in sufficiently large samples, both treated and control units with the exact same values of the covariates. In this way we can immediately remove all biases arising from differences between covariates, and many adjustment methods will give similar, or even identical, answers.
>
> However[...]this case rarely occurs in practice. In many applications it is not feasible to stratify fully on all covariates, because too many strata would have only a single unit.
>
> The differences between various adjustment methods arise precisely in such settings where it is not feasible to stratify on all values of the covariates, and mathematically these differences are most easily analyzed in settings with random samples from large populations using effectively continuous distributions for the covariates...[Therefore] for the purpose of discussing various frequentist approaches to estimation and inference under unconfoundedness...it is helpful to view the covariates as having been randomly drawn from an approximately continuous distribution.

The two main premises of this quote are sensible: a) confounding — and, more specifically, *de*confounding — is relatively easy to understand in the case of discrete covariates with finite support, and b) complete stratification is infeasible in many applications. But the conclusion — that the stylized setting of continuous covariates is therefore better suited to studying statistical methods for causal inference — does not necessarily follow. Indeed, this paper pursues a different mathematical assumption — that each strata has at least one treated-control contrast — and find that, even in that case, interesting bias variance trade-offs emerge. Moreover, these trade-offs can be studied directly for finite samples, without resorting to asymptotic arguments, which may be untrustworthy guides to a method's operating characteristics in practice. As a case in point, Hahn (2004) concludes that foreknowledge of which variables are instruments is asymptotically irrelevant for regression estimators of average treatment effects; the calculations of Section 3 show that foreknowledge of instruments is certainly relevant to the finite sample behavior of stratification estimators. Practitioners have, of course, long known this.

Regarding the true-propensity IPW estimator, the finite sample variance calculation highlights that it is a defective estimator that has variance that increases unboundedly as the data are shifted away from the origin. But the lesson should not be that known propensity scores are intrinsically irrelevant, but rather that the value of a known propensity function lies in the fact that its range defines a sufficient set of strata for deconfounding; the specific values that the propensity function takes on those strata need not be utilized.

Unfortunately, while a known propensity function defines a sufficient set of control strata, this stratification is neither minimal nor optimal. Specifically, a propensity function may include instrumental strata that are not necessary for deconfounding and it may *not* include prognostic strata that would improve treatment effect estimation despite being unnecessary for deconfounding. We note, however, that these limitations of a known propensity function are entirely distinct from the fundamental flaw of the true-propensity IPW, which simply uses the available information in an unwise way.

References

CHAO, M.-T. and STRAWDERMAN, W. (1972). Negative moments of positive random variables. *Journal of the American Statistical Association* **67** 429–431.

GRAHAM, B. S. (2011). Efficiency bounds for missing data models with semiparametric restrictions. *Econometrica* **79** 437–452.

HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 315–331.

HAHN, J. (2004). Functional restriction and efficiency in causal inference. *The Review of Economics and Statistics* **86** 73–76.

HERREN, A. and HAHN, P. R. (2023). On true versus estimated propensity scores for treatment effect estimation with discrete controls. *arXiv preprint arXiv:2305.11163*.

HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189.

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

LUNCEFORD, J. K. and DAVIDIAN, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine* **23** 2937–2960.

ROBINS, J. M., MARK, S. D. and NEWEY, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* 479–495.

ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.

Appendix 1

Here we give a proof that omitting noise variables from an IPW adjustment reduces variance, as discussed in Section 3. Specifically, we will demonstrate that

$$\frac{1-q^{(n+1)}}{2(n+1)p} - \left(\frac{1}{(2n+1)p} - \frac{\left(1-q^{(2n+2)}\right)}{(2n+1)(2n+2)p^2} \right) \geq 0.$$

To begin, observe that

$$\frac{1-q^{(n+1)}}{2(n+1)p} - \left(\frac{1}{(2n+1)p} - \frac{\left(1-q^{(2n+2)}\right)}{(2n+1)(2n+2)p^2} \right) = \frac{1-q^{(n+1)}}{2(n+1)p} - \frac{1}{(2n+1)p} + \frac{\left(1-q^{(2(n+1))}\right)}{(2n+1)(2(n+1))p^2}$$

$$= \frac{(1-q^{(n+1)})(2n+1)p - 2(n+1)p + (1-q^{(2(n+1))})}{2(n+1)(2n+1)p^2}$$

and that $2(n+1)(2n+1)p^2 \geq 0$ for all $p \in [0,1]$ and $n \geq 1$. Thus we aim to show that

$$(1-q^{(n+1)})(2n+1)(1-q) - 2(n+1)(1-q) + (1-q^{(2(n+1))}) \geq 0,$$

which simplifies to

$$q(1-q^{(n+1)})(q^n+1) - (2n+2)(1-q)q^{(n+1)} \geq 0,$$

the lefthand side of which is 0 at both $q = 0$ and $q = 1$. Because $q \geq 0$, dividing by q further reduces our task to showing that

$$(1 - q^{(n+1)})(q^n + 1) - (2n + 2)(1 - q)q^n \geq 0.$$

Here, the lefthand side is 0 at $p = 0$ (i.e. $q = 1$) and 1 at $p = 1$ (i.e. $q = 0$); therefore, we must show that this expression is nondecreasing in p, which we do by evaluating its derivative:

$$\begin{aligned}\frac{d}{dp}\left\{(1 - q^{(n+1)})(q^n + 1) - (2n + 2)(1 - q)q^n\right\} &= -\frac{d}{dq}\left\{(1 - q^{(n+1)})(q^n + 1) - (2n + 2)(1 - q)q^n\right\} \\ &= -(2n + 1)q^{n-1}\left[n(q - 1) - q(q^n - 1)\right].\end{aligned}$$

Following a similar strategy to before, we factor out $(2n + 1)q^{n-1}$, which is nonnegative in p, and aim to show that $-[n(q - 1) - q(q^n - 1)]$ is nonnegative for all $p \in [0, 1]$. As before, observe that $-[n(q - 1) - q(q^n - 1)] = 0$ when $p = 0$ and n when $p = 1$, meaning we must now show that *this* expression is nondecreasing in p. Again taking the derivative with respect to p, we find that

$$\begin{aligned}\frac{d}{dp} - [n(q - 1) - q(q^n - 1)] &= \frac{d}{dq}\left[n(q - 1) - q(q^n - 1)\right] \\ &= (n + 1)(1 - q^n) = (n + 1)(1 - (1 - p)^n),\end{aligned}$$

which is nonnegative for all $p \in [0, 1]$.