Hahn, Jinyong

Article

Properties of least squares estimator in estimation of average treatment effects

SERIEs - Journal of the Spanish Economic Association

Provided in Cooperation with:
Spanish Economic Association

Suggested Citation: Hahn, Jinyong (2023) : Properties of least squares estimator in estimation of average treatment effects, SERIEs - Journal of the Spanish Economic Association, ISSN 1869-4195, Springer, Heidelberg, Vol. 14, Iss. 3/4, pp. 301-313, https://doi.org/10.1007/s13209-023-00279-x

This Version is available at:
https://hdl.handle.net/10419/286580

SERIEs (2023) 14:301–313
https://doi.org/10.1007/s13209-023-00279-x

ORIGINAL ARTICLE

Properties of least squares estimator in estimation of average treatment effects

Jinyong Hahn[1]

Received: 27 October 2022 / Accepted: 31 March 2023 / Published online: 29 April 2023

Abstract
Treatment effects are often estimated by the least squares estimator controlling for some covariates. This paper investigates its properties. When the propensity score is constant, it is a consistent estimator of the average treatment effects if it is viewed as a semiparametric partially linear regression estimator, but it is not necessarily more efficient than the simple difference-of-means estimator. If it is literally viewed as a least squares estimator with a finite number of controls, it is equal to the weighted average of conditional average treatment effects with potentially negative weights, although the negative weight issue does not exist under semiparametric interpretation. It is shown that the negative weight issue can be avoided by use of logit specification.

Keywords OLS · Negative weight · Efficiency

JEL Classification C14

1 Introduction

The semiparametric efficiency bound for the average treatment effects for the standard case (where the treatment is randomly assigned conditional on some covariates) is well understood. The efficient estimator there takes various forms,[1] but none takes the form of the least squares regression of the dependent variable on the binary treatment controlling for some linear function of covariates. Even then, the latter specification is commonly employed in the literature. It would therefore make sense to investigate the properties of the least squares estimator that controls for the covariates.

This paper makes contributions in this regard by examining the semiparametric efficiency properties of the least squares estimator. For this purpose, the least squares

[1] See, e.g., Hahn (1998), or Hirano et al. (2003).

✉ Jinyong Hahn
hahn@econ.ucla.edu

[1] UCLA, Los Angeles, USA

specification is given a semiparametric interpretation, where the researcher is assumed to have made a " nonparametric promise" to control for the covariates more and more flexibly as the sample size increases to infinity. Under such a promise, it can be shown that the least squares estimator can be interpreted to be an estimator of some weighted average of the treatment effects. When the propensity score is constant, it can be shown that the least squares consistently estimates the average treatment effects (ATE), as long as the nonparametric promise is kept. It is shown that even with such a nonparametric promise, the least squares estimator does not necessarily have an obvious advantage over the naive difference-of-means estimator from an efficiency perspective. Variants of this results in parametric frameworks have been known in the literature, but the current paper makes a contribution by deriving the results by explicitly adopting a nonparametric framework. The paper also discusses the interpretation of the least squares estimator when the nonparametric promise is *not* kept. It is well known that the least squares can be interpreted to be the weighted average of the treatment effects, and the literature has recently begun to pay attention to the fact that the weights can be negative. The negative weight is due to the implicit linear probability specification of the treatment indicator on the covariates. It is shown that the problem can be eliminated by using a logit specification. The interpretation is related to a recent discussion by Blandhol et al. (2022, Proposition 1).[2]

Throughout the paper, we adopt the assumption that the treatments are independent of potential outcomes given covariates. To be more specific, we consider the model where $(Y(0), Y(1))$ is independent of the binary treatment indicator D given the covariates X. We do *not* impose the constant treatment effects assumption where $Y(1) - Y(0)$ is a fixed constant. In this model, it is convenient to write

$$\begin{aligned} Y(0) &= \mu_0(X) + \sigma_0(X) u_0, \\ Y(1) &= \mu_1(X) + \sigma_1(X) u_1, \end{aligned}$$

where the us have mean equal to 0 and variance equal to 1 conditional on X. We can then write the observed outcome $Y = (1 - D) Y(0) + DY(1)$ as

$$Y = D\beta(X) + \alpha(X) + ((1 - D)\sigma_0(X) u_0 + D\sigma_1(X) u_1), \tag{1}$$

where $\alpha(X) \equiv \mu_0(X)$ and $\beta(X) \equiv \mu_1(X) - \mu_0(X)$. We assume that the propensity score $\pi(X) \equiv E[D|X]$ as well as $\alpha(X), \beta(X), \sigma_0(X), \sigma_1(X)$ are nonparametrically specified. We will also assume that $(Y_i(0), Y_i(1), X_i, D_i)$ $i = 1, 2, \ldots$ are independent and identically distributed (IID), and that the researcher observes $Y_i = (1 - D_i) Y_i(0) + D_i Y_i(1)$ as well as (X_i, D_i).

2 Interpretation of partially linear regression specification

We will consider the interpretation of the partially linear regression of Y on D using X as the control variable. To be more precise, we consider computing the estimate of

[2] See (Goldsmith-Pinkham et al. 2022) for related discussion for the case involving multiple treatments.

β_{LS} by fitting a semiparametric model

$$Y = D\beta_{LS} + g(X) + \varepsilon, \tag{2}$$

where $g(X)$ is nonparametrically specified. Let $\hat{\beta}_{LS}$ denote the estimated coefficient of D in such a regression.[3] We first examine the pseudo-parameter β_{LS} that $\hat{\beta}_{LS}$ estimates, i.e., its probability limit. It is straightforward to recognize that the pseudo-parameter is the population regression estimate of Y on $D - \pi(X)$, where $\pi(X)$ is the propensity score. We present an interpretation of β_{LS} as a weighted average of the $\beta(X)$, i.e., the average treatment effects (ATE) conditional on X.[4] Angrist (1998) derived such a representation for the case where X has a multinomial distribution, and the representation below is a nonparametric generalization when X has an arbitrary distribution:

Proposition 1

$$\beta_{LS} = \frac{E[\pi(X)(1-\pi(X))\beta(X)]}{E[\pi(X)(1-\pi(X))]}. \tag{3}$$

The interpretation (3) is from the semiparametric perspective, based on the "nonparametric promise" that the covariates will be controlled by richer and richer specification as a function of the sample size. Note that it is a weighted average of $\beta(X)$ weighted by $\pi(X)(1-\pi(X))/E[\pi(X)(1-\pi(X))]$. Because the $\pi(X)$ denotes the true conditional probability of D given X, the weight is always nonnegative. Therefore, the negative weight problem discussed in Blandhol et al. (2022, Section 4) does not exist under the nonparametric specification/interpretation of g in (2).

Suppose that a practitioner does not make or keep such a " nonparametric promise," and that he/she adopts a literally parametric approach where $g(X)$ is linear in X. We can show that the probability limit of $\hat{\beta}_{LS}$ is now equal to

$$\frac{E\left[\left(D - X'\gamma\right)Y\right]}{E\left[\left(D - X'\gamma\right)^2\right]} = \frac{E\left[\pi(X)\left(1 - X'\gamma\right)\beta(X)\right]}{E\left[\left(D - X'\gamma\right)^2\right]} + \frac{E\left[\left(\pi(X) - X'\gamma\right)\alpha(X)\right]}{E\left[\left(D - X'\gamma\right)^2\right]}, \tag{4}$$

where $X'\gamma$ is the linear projection of D on X. If the true $\alpha(X)$ is linear in X, we can show that the estimand (4) simplifies to

$$\frac{E\left[\left(D - X'\gamma\right)Y\right]}{E\left[\left(D - X'\gamma\right)^2\right]} = \frac{E\left[\pi(X)\left(1 - X'\gamma\right)\beta(X)\right]}{E\left[\left(D - X'\gamma\right)^2\right]}, \tag{5}$$

which implies that it is a weighted average of the conditional expectation $\beta(X)$ of the treatment effects given X. Because $X'\gamma$ can lie outside of the $(0, 1)$ range, it raises

[3] The parameter β_{LS} and the error term ε in equation (2) are defined by the partially linear regression applied to the true model (1). It can be shown that the ε does not satisfy $E[\varepsilon|D, X] = 0$, so (2) is not a partially linear regression " model," and the $\hat{\beta}_{LS}$ is a special case of the partially linear " projection" considered by Newey and Robins (2018).

[4] All proofs are collected in the appendix.

the possibility of negative weights. See, e.g., Blandhol et al. (2022, Section 4). This results from the partitioned regression interpretation of multiple regression, and the implicit linear probability specification there; the estimate of the coefficient of D in the regression of Y on D and X is equivalent to the estimate when Y is regressed on the residual when D is regressed on X, and because the regression of D on X uses the linear specification, the fitted value can exceed the (0,1) interval, thereby leading to the possibility that the residual can be negative.

One way to avoid this problem is to use logit specification instead of the linear probability specification. Suppose that we adopt a logit specification where $\Pr[D=1|X]$ is specified as $\Lambda\left(X'\delta\right)$, where $\Lambda(t)=e^t/\left(1+e^t\right)$. If we use $\Lambda\left(X'\hat{\delta}\right)$ as the fitted value (instead of $X'\gamma$ which would be used if the linear probability model is adopted) and regress Y on $D-\Lambda\left(X'\hat{\delta}\right)$, we would get the estimator

$$\frac{\sum_{i=1}^{n}\left(D-\Lambda\left(X'\hat{\delta}\right)\right)Y}{\sum_{i=1}^{n}\left(D-\Lambda\left(X'\hat{\delta}\right)\right)^2},$$

which converges in probability to

$$\frac{E\left[\left(D-\Lambda\left(X'\delta\right)\right)Y\right]}{E\left[\left(D-\Lambda\left(X'\delta\right)\right)^2\right]},$$

where δ denotes the probability limit of $\hat{\delta}$. Lee (2018) also consider using the nonlinear parametric specification for the propensity score, although he proposes a different estimator. Lee (2018) shows that his estimand can be interpreted to be the weighted average of $Y-E[Y|p(X)]$, where $p(X)$ denotes the probability limit of the parametric specification of $E[Y|X]$, but the current paper makes a contribution by providing an interpretation of the estimand as a weighted average of the conditional treatment effects $\beta(X)$.[5]

For this purpose, write

$$\begin{aligned}E\left[\left(D-\Lambda\left(X'\delta\right)\right)Y\right]=&E\left[\left(D-\Lambda\left(X'\delta\right)\right)D\beta(X)\right]+E\left[\left(D-\Lambda\left(X'\delta\right)\right)\alpha(X)\right]\\&+E\left[\left(D-\Lambda\left(X'\delta\right)\right)(1-D)\sigma_0(X)u_0\right]\\&+E\left[\left(D-\Lambda\left(X'\delta\right)\right)D\sigma_1(X)u_1\right].\end{aligned}$$

Because

$$E\left[\left(D-\Lambda\left(X'\delta\right)\right)(1-D)\sigma_0(X)u_0\middle|X\right]=E\left[\left(D-\Lambda\left(X'\delta\right)\right)(1-D)\sigma_0(X)\,\middle|\,X\right]E[u_0|X]$$

by conditional independence, we conclude that the third term above is equal to zero. The last term is also equal to zero by the same reasoning. We recall that the logit MLE

[5] Belloni et al. (2014) also propose to use nonlinear parametric specification of the propensity score, but they use the estimated propensity score as part of the efficient influence function of Hahn (1998), so the nonlinearity is not employed to overcome the negative weight problem.

implies the first-order condition (FOC) such that[6]

$$\sum_{i=1}^{n}\left(D-\Lambda\left(X'\hat{\delta}\right)\right)X=0. \tag{6}$$

It follows that $\sum_{i=1}^{n}\left(D-\Lambda\left(X'\hat{\delta}\right)\right)g\left(X\right)=0$ for any linear function $g\left(X\right)$ of X. In particular, it will be satisfied for $\alpha\left(X\right)$ if it were indeed linear, which implies that $E\left[\left(D-\Lambda\left(X'\delta\right)\right)\alpha\left(X\right)\right]=0$ as long as $\alpha\left(X\right)$ is linear in X. We would also have $E\left[\left(D-\Lambda\left(X'\delta\right)\right)\alpha\left(X\right)\right]=0$ if $\Lambda\left(X'\delta\right)$ is a correct specification of the true propensity score. Therefore, we can see that the pseudo-parameter that the new estimator estimates is

$$\begin{aligned}\frac{E\left[\left(D-\Lambda\left(X'\delta\right)\right)Y\right]}{E\left[\left(D-\Lambda\left(X'\delta\right)\right)^{2}\right]}&=\frac{E\left[\left(D-\Lambda\left(X'\delta\right)\right)D\beta\left(X\right)\right]}{E\left[\left(D-\Lambda\left(X'\delta\right)\right)^{2}\right]}\\&=\frac{E\left[\pi\left(X\right)\left(1-\Lambda\left(X'\delta\right)\right)\beta\left(X\right)\right]}{E\left[\left(D-\Lambda\left(X'\delta\right)\right)^{2}\right]},\end{aligned} \tag{7}$$

which will retain the " positive weight" feature, as long as $\alpha\left(X\right)$ is literally linear (or the propensity score indeed has a logit specification).[7]

Because $E\left[\left(D-\Lambda\left(X'\delta\right)\right)^{2}\right]=E\left[\pi\left(X\right)-2\pi\left(X\right)\Lambda\left(X'\delta\right)+\Lambda\left(X'\delta\right)^{2}\right]$, the "weight" in (7) does not necessarily add up to 1. This problem can be eliminated by considering instead an IV estimator

$$\frac{\sum_{i=1}^{n}\left(D-\Lambda\left(X'\hat{\delta}\right)\right)Y}{\sum_{i=1}^{n}\left(D-\Lambda\left(X'\hat{\delta}\right)\right)D},$$

which converges in probability to

$$\frac{E\left[\left(D-\Lambda\left(X'\delta\right)\right)Y\right]}{E\left[\left(D-\Lambda\left(X'\delta\right)\right)D\right]}=\frac{E\left[\pi\left(X\right)\left(1-\Lambda\left(X'\delta\right)\right)\beta\left(X\right)\right]}{E\left[\pi\left(X\right)\left(1-\Lambda\left(X'\delta\right)\right)\right]},$$

where the weights are nonnegative and add up to 1.

On a related note, (Blandhol et al. 2022, Proposition 1) discussed the interpretation of 2SLS without any " nonparametric promise" under the assumption that $E\left[Y\left(0\right)|X\right]$ and $E\left[Y\left(1\right)|X\right]$ are linear in X. It was shown that the pseudo-parameter estimated by 2SLS can be decomposed into two terms. The first term is a weighted average of the conditional treatment effects among compliers, and the weight $\omega\left(\mathrm{CP},X\right)$ can be negative. The second term is a weighted average of the conditional treatment effects

[6] The counterpart of (6) is not satisfied in probit specification.

[7] Belloni et al. (2014) advocated the use of double machine learning in the partially linear regression context. In their framework, the counterpart of D was not restricted to be a binary variable, so the concern about the negative weight was not an issue. The preceding discussion suggests that one may consider double machine learning, where some least squares type machine learning is adopted for the Y on X submodel, but some logit type machine learning is adopted for the D on X submodel.

among always takers, and the weight ω (AT, X) may not be equal to zero. Their analysis is based on the equivalence that the 2SLS is equal to the IV estimator using the residual $\tilde{Z}$ from the regression of binary instrument Z on X, which implicitly adopts a linear probability model specification. If the nonparametric promise is made and kept, this is not an issue, but without such a promise, it would lead to the phenomenon that the fitted value from the regression of Z on X can exceed 1, which leads to the possibility of negative weight ω (CP, X). This issue can be resolved by using a logit specification of Z on X, and replacing $\tilde{Z}$ in their equation (1) by $Z - \Lambda\left(X'\delta\right)$.[8] As for the nonzero ω (AT, X), it can be attributed to two reasons (1) the nonparametric promise is not kept; or (2) the implicit linear probability specification is incorrect.

3 Efficiency of semiparametric regression adjustments in randomized experiments

In this section, we examine the semiparametric efficiency properties of $\hat{\beta}_{LS}$ in the semiparametric specification (2). In particular, we consider the case where the propensity score $\pi(\cdot)$ is constant, and try to understand the efficiency properties of the regression adjustments from a semiparametric perspective. Freedman (2008a, b) considered comparison of parametric version of $\hat{\beta}_{LS}$ with the difference-in-means estimator. More precisely, he considered the case where $g(X)$ in (2) is specified as a linear function of X, and concluded that efficiency ranking is impossible adopting an asymptotic framework where the finite population with size n changes as a function of n.[9] Lin (2013) adopted an identical framework, and investigated how efficiency improvement is possible,[10] which is confirmed by Negi and Wooldridge (2021) under an IID framework. Negi and Wooldridge (2021) also use the linear parametric specification of $g(X)$, so the result in this section can be understood to be a fully semiparametric generalization of the earlier results in a familiar IID asymptotic framework.

In order to understand the efficiency properties of $\hat{\beta}_{LS}$, it is useful to derive its asymptotic variance. Below is a characterization of the asymptotic properties of $\hat{\beta}_{LS}$, obtained under the nonparametric specification of $\alpha(X)$, $\beta(X)$, $\sigma_0(X)$, $\sigma_1(X)$, and $\pi(X)$.

Proposition 2 *The asymptotic variance of* $\hat{\beta}_{LS}$ *is*

$$\frac{E\left[\pi(X)^2(1-\pi(X))\sigma_0^2(X)\right]}{\left(E\left[\pi(X)(1-\pi(X))\right]\right)^2} + \frac{E\left[(1-\pi(X))^2\pi(X)\sigma_1^2(X)\right]}{\left(E\left[\pi(X)(1-\pi(X))\right]\right)^2} + \frac{E\left[(D-\pi(X))^4(\beta(X)-\beta_{LS})^2\right]}{\left(E\left[\pi(X)(1-\pi(X))\right]\right)^2}.$$

[8] Because X and $Z - \Lambda\left(X'\delta\right)$ are orthogonal due to the first order condition of the logit MLE as in (6), their proof of Proposition 2 goes through without any modification. The only change is that the negative ω (CP, X) phenomenon disappears.

[9] Therefore, in his framework, the n is the size of population.

[10] He does so by using a parametric variant of the semiparametric efficient estimator developed in Hahn (1998).

So far, we allowed for the possibility that $\pi(X)$ is not a constant, which meant that β_{LS} may not be the ATE. Now, let us assume that $\pi(X) = \pi$. If so, we have $\beta_{LS} = E[\beta(X)] = \beta_{ATE}$ by (3). Specializing asymptotic variance formula in Proposition 2 for this situation, we find that the asymptotic variance of β_{LS} now simplifies to

$$E\left[\frac{\sigma_0^2(X)}{1-\pi} + \frac{\sigma_1^2(X)}{\pi} + \frac{E\left[(D-\pi)^4\right]}{(\pi(1-\pi))^2}(\beta(X) - \beta_{ATE})^2\right]. \tag{8}$$

Relative to the efficiency bound for the ATE in Hahn (1998)

$$E\left[\frac{\sigma_0^2(X)}{1-\pi} + \frac{\sigma_1^2(X)}{\pi} + (\beta(X) - \beta_{ATE})^2\right] \tag{9}$$

specialized for the case where $\pi(X)$ is constant, it can be shown that (8) is greater than or equal to (9) in general. In other words, the $\hat{\beta}_{LS}$ is not semiparametrically efficient, even under the special assumption of constant propensity score.[11]

We now consider the difference-in-means estimator for the same case where $\pi(X)$ is constant and equal to π. It is trivial to show that its asymptotic variance[12] is equal to

$$\frac{\text{Var}[Y(0)]}{1-\pi} + \frac{\text{Var}[Y(1)]}{\pi} = \frac{\sigma_0^2(X)}{1-\pi} + \frac{\sigma_1^2(X)}{\pi} + \frac{\text{Var}[\mu_0(X)]}{1-\pi} + \frac{\text{Var}[\mu_1(X)]}{\pi}. \tag{10}$$

Again making the same comparison with the efficiency bound (9), we can show that (10) is greater than or equal to (9) in general, so the difference-in-means estimator is not semiparametrically efficient either.[13]

Efficiency ranking between the partially linear projection estimator and the difference-in-means estimator boils down to the comparison of (8) and (10). Because both estimators do not achieve the semiparametric efficiency bound even under the case where the propensity score is constant, we can make an educated guess that efficiency ranking between these two estimators is impossible in general. In the appendix, we present two numerical examples to confirm this educated guess, confirming the Freedman (2008a, b) critique in the nonparametric framework.

On the other hand, we can find a reasonably interesting situation where the partially linear regression specification does lead to efficiency:

Proposition 3 *Suppose that* $\pi(X) = \pi = 1/2$. *Then,* (8) *is equal to* (9)

We can see that the asymptotic variance of the partially linear regression achieves the efficiency bound when $\pi = 1/2$. Therefore, if the treatment is assigned by flipping a fair coin, it would be sensible to adopt the partially linear regression model specification. The special role of $\pi = 1/2$ was discussed by Freedman (2008b) and Lin (2013), as well as Negi and Wooldridge (2021) for the case where $g(X)$ is given

[11] See the proof of Proposition 4 in the appendix for more detailed derivation.

[12] A proper subset of calculations here can be found in Hahn (1998).

[13] See the proof of Proposition 5 in the appendix for more detailed derivation.

a linear parametric specification. As noted above, Freedman (2008b) and Lin (2013) adopt an asymptotic framework characterized by a sequence of finite populations, which makes it different from Negi and Wooldridge (2021). Therefore, Proposition 3 can also be understood to be a fully semiparametric generalization of the earlier results in an IID asymptotic framework.

Data availability Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose.

Appendix

A Proofs

Proof of Proposition 1 Using

$$\begin{aligned}(D-\pi(X))\,Y = {} & (D-\pi(X))\,D\beta(X) + (D-\pi(X))\,\alpha(X) \\ & + (D-\pi(X))\left((1-D)\,\sigma_0(X)\,u_0 + D\sigma_1(X)\,u_1\right)\end{aligned}$$

and the facts that[14]

$$\begin{aligned}E\left[(D-\pi(X))\,D\beta(X)|\,X\right] &= E\left[(D-\pi(X))\,D|\,X\right]\beta(X) \\ &= \pi(X)\,(1-\pi(X))\,\beta(X)\,, \\ E\left[(D-\pi(X))\,\alpha(X)|\,X\right] &= \alpha(X)\,E\left[D-\pi(X)|\,X\right] = 0, \\ E\left[(D-\pi(X))\,\sigma_0(X)\,u_0|\,X\right] &= \sigma_0(X)\,E\left[(D-\pi(X))\,u_0|\,X\right] \\ &= \sigma_0(X)\,E\left[D-\pi(X)|\,X\right]E\left[u_0|\,X\right] = 0, \\ E\left[(D-\pi(X))\,D\sigma_0(X)\,u_0|\,X\right] &= \sigma_0(X)\,E\left[(D-\pi(X))\,Du_0|\,X\right] \\ &= \sigma_0(X)\,E\left[(D-\pi(X))\,D|\,X\right]E\left[u_0|\,X\right] = 0, \\ E\left[(D-\pi(X))\,D\sigma_1(X)\,u_1|\,X\right] &= \sigma_1(X)\,E\left[(D-\pi(X))\,Du_1|\,X\right] \\ &= \sigma_1(X)\,E\left[(D-\pi(X))\,D|\,X\right]E\left[u_1|\,X\right] = 0,\end{aligned}$$

[14] I used conditional independence when I wrote $E\left[(D-\pi(X))\,u_0|\,X\right] = E\left[D-\pi(X)|\,X\right]E\left[u_0|\,X\right]$ below.

along with the law of iterated expectations, we can see that the estimand is

$$\beta_{LS} = \frac{E\left[\left(D - \pi\left(X\right)\right)Y\right]}{E\left[\left(D - \pi\left(X\right)\right)^2\right]} = \frac{E\left[\pi\left(X\right)\left(1 - \pi\left(X\right)\right)\beta\left(X\right)\right]}{E\left[\pi\left(X\right)\left(1 - \pi\left(X\right)\right)\right]}.$$

□

Proof of (4) and (5) We have

$$\begin{aligned} E\left[\left(D - X'\gamma\right)Y\right] = {} & E\left[\left(D - X'\gamma\right)D\beta\left(X\right)\right] + E\left[\left(D - X'\gamma\right)\alpha\left(X\right)\right] \\ & + E\left[\left(D - X'\gamma\right)\left(1 - D\right)\sigma_0\left(X\right)u_0\right] \\ & + E\left[\left(D - X'\gamma\right)D\sigma_1\left(X\right)u_1\right]. \end{aligned}$$

If $\alpha\left(X\right)$ is linear in X, we have

$$E\left[\left(D - X'\gamma\right)\alpha\left(X\right)\right] = 0.$$

Furthermore, we have

$$\begin{aligned} E\left[\left(D - X'\gamma\right)\left(1 - D\right)\sigma_0\left(X\right)u_0 \middle| X\right] &= E\left[\left(D - X'\gamma\right)\left(1 - D\right)\sigma_0\left(X\right) \middle| X\right] E\left[u_0 \middle| X\right] = 0, \\ E\left[\left(D - X'\gamma\right)D\sigma_1\left(X\right)u_1 \middle| X\right] &= E\left[\left(D - X'\gamma\right)D\sigma_1\left(X\right) \middle| X\right] E\left[u_1 \middle| X\right] = 0 \end{aligned}$$

by conditional independence. Therefore, we have

$$E\left[\left(D - X'\gamma\right)Y\right] = E\left[\left(D - X'\gamma\right)D\beta\left(X\right)\right] = E\left[\pi\left(X\right)\left(1 - X'\gamma\right)\beta\left(X\right)\right],$$

from which (5) follows.

If $\alpha\left(X\right)$ is not linear in X, we have

$$\begin{aligned} & E\left[\left(D - X'\gamma\right)\alpha\left(X\right)\right] = E\left[\left(D - \pi\left(X\right)\right)\alpha\left(X\right)\right] + E\left[\left(\pi\left(X\right) - X'\gamma\right)\alpha\left(X\right)\right] \\ & = E\left[\left(\pi\left(X\right) - X'\gamma\right)\alpha\left(X\right)\right], \end{aligned}$$

which is not equal to zero in general. Therefore, the interpretation in (5) is incorrect in general without the linearity of $\alpha\left(X\right)$, and we have

$$\frac{E\left[\left(D - X'\gamma\right)Y\right]}{E\left[\left(D - X'\gamma\right)^2\right]} = \frac{E\left[\pi\left(X\right)\left(1 - X'\gamma\right)\beta\left(X\right)\right]}{E\left[\left(D - X'\gamma\right)^2\right]} + \frac{E\left[\left(\pi\left(X\right) - X'\gamma\right)\alpha\left(X\right)\right]}{E\left[\left(D - X'\gamma\right)^2\right]}$$

in general. □

Proof of Proposition 2 Using a result in Newey and Robins (2018), we find that the influence function for $\hat{\beta}_{LS}$ is equal to $\left(E\left[\left(D - \pi\left(X\right)\right)^2\right]\right)^{-1}$ times $\left(D - \pi\left(X\right)\right)\varepsilon$, where[15]

[15] We note that the ε in (11) does not satisfy $E\left[\varepsilon \middle| D, X\right] = 0$, so it does not satisfy the basic assumption of (Robinson (1988), equation (1)), and therefore, the $\hat{\beta}_{LS}$ is not an estimator for a partially linear regression

$$\begin{aligned}\varepsilon &\equiv Y - D\beta_{LS} - E\left[Y - D\beta_{LS}|\, X\right] \\ &= (D - \pi(X))(\beta(X) - \beta_{LS}) + ((1-D)\sigma_0(X)u_0 + D\sigma_1(X)u_1). \end{aligned} \tag{11}$$

We have

$$\begin{aligned}(D - \pi(X))\varepsilon = (D - \pi(X))^2(\beta(X) - \beta_{LS}) + (D - \pi(X))(1-D)\sigma_0(X)u_0 \\ + (D - \pi(X))D\sigma_1(X)u_1.\end{aligned}$$

The three terms on the right have mean zero, and their cross products have also mean zero. Moreover, we have

$$\begin{aligned}E\left[(D - \pi(X))^2 \middle| X\right] &= \pi(X)(1 - \pi(X)), \\ E\left[(D - \pi(X))^2(1-D)^2\sigma_0^2(X) \middle| X\right] &= \pi(X)^2(1 - \pi(X))\sigma_0^2(X), \\ E\left[(D - \pi(X))^2 D\sigma_1^2(X) \middle| X\right] &= (1 - \pi(X))^2\pi(X)\sigma_1^2(X),\end{aligned}$$

which means that the asymptotic variance is equal to from which the conclusion follows. □

Proposition 4 *Suppose that the propensity score is constant. The $\hat{\beta}_{LS}$ is not semiparametrically efficient*

Proof We show that (8) is greater than or equal to (9). We have that

$$(8) - (9) = E\left[\frac{E\left[(D-\pi)^4\right]}{(\pi(1-\pi))^2}(\beta(X) - \beta_{ATE})^2\right] - E\left[(\beta(X) - \beta_{ATE})^2\right],$$

but because $E\left[(D - \pi(X))^2\right] = \pi(1-\pi)$, we see that

$$\frac{E\left[(D-\pi)^4\right]}{(\pi(1-\pi))^2} \geq 1$$

by Cauchy-Schwartz, so the asymptotic variance (8) from the partially linear regression is larger than the asymptotic variance bound (9) in general. □

Proposition 5 *Suppose that the propensity score is constant. The difference-in-means estimator is not semiparametrically efficient*

Footnote 15 continued
" model." It is a special case of the partially linear " projection" considered by Newey and Robins (2018). Because the resultant estimators are identical, we continue to call $\hat{\beta}_{LS}$ an estimator for the partially linear regression model, reflecting the familiarity of the terminology. In other words, we may adopt a practical point of view, and interpret that the underlying " model" is $Y = D\beta_{LS} + g(X) + \varepsilon$, where ε is defined in (11) and $g(X) = \pi(X)(\beta(X) - \beta_{LS}) + \alpha(X)$.

Proof We show that (10) is greater than or equal to (9). When $\pi(X)$ is constant and equal to π, we have

$$
\begin{aligned}
&(10)-(9)\\
&=\frac{\operatorname{Var}[\mu_0(X)]}{1-\pi}+\frac{\operatorname{Var}[\mu_1(X)]}{\pi}-E\left[(\beta(X)-\beta_{ATE})^2\right]\\
&=\frac{\operatorname{Var}[\mu_0(X)]}{1-\pi}+\frac{\operatorname{Var}[\mu_1(X)]}{\pi}-\operatorname{Var}[\mu_1(X)-\mu_0(X)]\\
&=\frac{\operatorname{Var}[\mu_0(X)]}{1-\pi}+\frac{\operatorname{Var}[\mu_1(X)]}{\pi}-\operatorname{Var}[\mu_0(X)]-\operatorname{Var}[\mu_1(X)]+2\operatorname{Cov}(\mu_0(X),\mu_1(X))\\
&=\frac{p}{1-p}\operatorname{Var}[\mu_0(X)]+\frac{1-p}{p}\operatorname{Var}[\mu_1(X)]+2\operatorname{Cov}(\mu_0(X),\mu_1(X))\\
&=\operatorname{Var}\left[\sqrt{\frac{p}{1-p}}\mu_0(X)+\sqrt{\frac{1-p}{p}}\mu_1(X)\right]\geq 0.
\end{aligned}
$$

□

Proof of Proposition 3 If $\pi=1/2$, we see that

$$
\frac{E\left[(D-\pi)^4\right]}{(\pi(1-\pi))^2}=\frac{\left(1-\frac{1}{2}\right)^4\frac{1}{2}+\left(0-\frac{1}{2}\right)^4\left(1-\frac{1}{2}\right)}{\left(\frac{1}{2}\left(1-\frac{1}{2}\right)\right)^2}=1
$$

so (8) becomes equal to

$$
E\left[\frac{\sigma_0^2(X)}{1-\pi}+\frac{\sigma_1^2(X)}{\pi}+(\beta(X)-\beta_{ATE})^2\right],
$$

i.e., the efficiency bound (9). □

B Numerical comparison for Sect. 3

From (8) and (10), we can see that efficiency ranking between the two estimators boils down to comparison between

$$
\frac{E\left[(D-\pi)^4\right]}{(\pi(1-\pi))^2}E\left[(\beta(X)-\beta_{ATE})^2\right]=\frac{E\left[(D-\pi)^4\right]}{(\pi(1-\pi))^2}\operatorname{Var}[\mu_0(X)-\mu_1(X)]
$$

and

$$
\frac{\operatorname{Var}[\mu_0(X)]}{1-p}+\frac{\operatorname{Var}[\mu_1(X)]}{p}.
$$

First, we consider the case where $p = 1/3$ with $\mu_0(X)$ and $\mu_1(X)$ independent of each other. In this case,

$$\frac{E\left[(D-\pi)^4\right]}{(\pi(1-\pi))^2} = \frac{\left(1-\frac{1}{3}\right)^4 \frac{1}{3} + \left(0-\frac{1}{3}\right)^4 \left(1-\frac{1}{3}\right)}{\left(\frac{1}{3}\left(1-\frac{1}{3}\right)\right)^2} = \frac{3}{2}.$$

If $\mu_0(X)$ and $\mu_1(X)$ are independent, we have $\mathrm{Var}[\mu_0(X) - \mu_1(X)] = \mathrm{Var}[\mu_0(X)] + \mathrm{Var}[\mu_1(X)]$, so

$$(8) = E\left[\frac{\sigma_0^2(X)}{1-\pi} + \frac{\sigma_1^2(X)}{\pi}\right] + E\left[\frac{3}{2}\mathrm{Var}[\mu_0(X)] + \frac{3}{2}\mathrm{Var}[\mu_1(X)]\right],$$
$$(10) = E\left[\frac{\sigma_0^2(X)}{1-\pi} + \frac{\sigma_1^2(X)}{\pi}\right] + E\left[3\,\mathrm{Var}[\mu_0(X)] + \frac{3}{2}\mathrm{Var}[\mu_1(X)]\right].$$

It follows that the asymptotic variance of the partially linear regression is smaller in this case.

Second, we consider the case where $p = 0.1$. In this case,

$$\frac{E\left[(D-\pi)^4\right]}{(\pi(1-\pi))^2} = \frac{(1-0.1)^4(0.1) + (0-0.1)^4(1-0.1)}{((0.1)(1-0.1))^2} = 8.1111$$

so the comparison boils down to

$$(8) = E\left[\frac{\sigma_0^2(X)}{1-\pi} + \frac{\sigma_1^2(X)}{\pi}\right] + E\left[8.1111\,\mathrm{Var}[\mu_0(X) - \mu_1(X)]\right],$$
$$(10) = E\left[\frac{\sigma_0^2(X)}{1-\pi} + \frac{\sigma_1^2(X)}{\pi}\right] + E\left[\frac{10}{9}\mathrm{Var}[\mu_0(X)] + 10\,\mathrm{Var}[\mu_1(X)]\right].$$

If $\mu_1(X) = -\mu_0(X)$, we can see that

$$8.1111\,\mathrm{Var}[\mu_0(X) - \mu_1(X)] = 8.1111\,\mathrm{Var}[2\mu_0(X)] = 32.444\,\mathrm{Var}[\mu_0(X)],$$
$$\left(\frac{10}{9} + 10\right)\mathrm{Var}[\mu_0(X)] = 11.111\,\mathrm{Var}[\mu_0(X)],$$

so the asymptotic variance of the partially linear regression is larger in this case.

References

Belloni A, Chernozhukov V, Hansen C (2014) Inference on treatment effects after selection among high-dimensional controls. Rev Econ Stud 81:608–650

Blandhol C, Bonney J, Mogstad M, Torgovitsky A (2022) When is TSLS Actually LATE? NBER Working Paper No. 29709

Freedman DA (2008) On regression adjustments in experiments with several treatments. Ann Appl Stat 2:176–196
Freedman DA (2008) On regression adjustments to experimental data. Adv Appl Math 40:180–193
Goldsmith-Pinkham P, Hull P, Kolesár M (2022) Contamination Bias in Linear Regressions, unpublished working paper
Hahn J (1998) On the role of propensity score in efficient semiparametric estimation of average treatment effects. Econometrica 66:315–332
Hirano K, Imbens GW, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71:1161–1189
Lee M (2018) Simple least squares estimator for treatment effects using propensity score residuals. Biometrika 105:149–164
Lin W (2013) Agnostic notes on regression adjustments to experimental data: reexamining freedman's critique. Ann Appl Stat 7:295–318
Negi A, Wooldridge JM (2021) Revisiting regression adjustment in experiments with heterogeneous treatment effects. Econom Rev 40:504–534
Newey WK, Robins JM (2018) Cross-fitting and fast remainder rates for semiparametric estimation, unpublished working paper
Robinson PM (1988) Root-N-consistent semiparametric regression. Econometrica 56:931–954