Research Article

Goeun Lee, Jin-young Choi, and Myoung-jae Lee*

Minimally capturing heterogeneous complier effect of endogenous treatment for any outcome variable

https://doi.org/10.1515/jci-2022-0036
received May 17, 2022; accepted May 26, 2023

Abstract: When a binary treatment D is possibly endogenous, a binary instrument δ is often used to identify the "effect on compliers." If covariates X affect both D and an outcome Y, X should be controlled to identify the "X-conditional complier effect." However, its nonparametric estimation leads to the well-known dimension problem. To avoid this problem while capturing the effect heterogeneity, we identify the complier effect heterogeneous with respect to only the one-dimensional "instrument score" $E(\delta|X)$ for non-randomized δ. This effect heterogeneity is minimal, in the sense that any other "balancing score" is finer than the instrument score. We establish two critical "reduced-form models" that are linear in D or δ, even though no parametric assumption is imposed. The models hold for any form of Y (continuous, binary, count, ...). The desired effect is then estimated using either single index model estimators or an instrumental variable estimator after applying a power approximation to the effect. Simulation and empirical studies are performed to illustrate the proposed approaches.

Keywords: endogenous treatment, complier effect, instrument score, propensity score, single index model, instrumental variable estimator

MSC 2020: 62G05, 62P10, 62P25

1 Introduction

A typical treatment effect analysis [1–6] involves a binary treatment D, an outcome Y, and covariates X. If the potential versions of Y for $D = 0, 1$ are denoted as (Y^0, Y^1), we can typically find $E(Y^1 - Y^0)$ or $E(Y^1 - Y^0|X)$ when D is exogenous. However, the identification of $E(Y^1 - Y^0)$ or $E(Y^1 - Y^0|X)$ is difficult when D is endogenous. Suppose a binary instrument variable (IV) δ is available for D, which meets some conditions including $\delta \perp\!\!\!\perp (Y^0, Y^1)|X$, with $\perp\!\!\!\perp$ for independence. Denoting the potential treatments of D for $\delta = 0, 1$ as (D^0, D^1), let the compliers (CP) be those with $(D^0 = 0, D^1 = 1)$ [7,8]. For endogenous D, we can identify $E(Y^1 - Y^0|\text{CP})$ or $E(Y^1 - Y^0|\text{CP}, X)$. Note that $D\{\equiv D^0 + (D^1 - D^0)\delta\} = \delta$ holds for CP.

The instrumental variable estimator (IVE) for Y on D with δ as the IV can estimate $E(Y^1 - Y^0|\text{CP}, X = x)$ for a discrete X using only the subpopulation $X = x$. However, for a high-dimensional X, this subsample approach runs into the well-known dimension problem as in Frölich [9] who estimated $E(Y^1 - Y^0|\text{CP}, X)$

*** Corresponding author: Myoung-jae Lee,** Department of Economics, Korea University, 145, Anam-ro, Seongbuk-gu, Seoul 02841, Republic of Korea, e-mail: myoungjae@korea.ac.kr
Goeun Lee: Jinhe Center for Economic Research, Xi'an Jiaotong University, 28, Xianning West Road, Xi'an, Shaanxi 710049, China, e-mail: goeunlee@xjtu.edu.cn
Jin-young Choi: Division of Economics, Hankuk University of Foreign Studies, 107, Imun-ro, Dongdaemun-gu, Seoul 02450, Republic of Korea, e-mail: choiecon@hufs.ac.kr

nonparametrically. To avoid this problem, Frölich [9] considered conditioning on "instrument score (IS)," but did not pursue it owing to an efficiency concern. Abadie [10] parametrized a local average response function (and the IS), and Tan [11] parametrized $E(Y|\delta, X)$ or $E(\delta|X)$. Ogburn et al. [12] estimated $E(Y^1 - Y^0|\text{CP}, \tilde{X})$ for some $\tilde{X} \in X$. Estimators based on a parametrization for Y, as in the work of Ogburn et al. [12], are valid only for certain types of Y (e.g., cardinal Y).

Treatment effect heterogeneity is an important problem that has been addressed by several researchers such as Imai and Ratkovic [13] and Künzel et al. [14]. Athey and Imbens [15] noted the following three problems: (i) estimating heterogeneous effects, (ii) finding an optimal policy allocating subjects to the treatment or control based on X (e.g., [16,17]), and (iii) low dimensionally representing the effect heterogeneity as in Athey and Imbens [18]. We address (iii) in this article by representing the effect heterogeneity in a unidimensional manner.

We find the X-heterogeneous effect to recommend D to those who would benefit from D. Consider a fat (Y) reducing drug D. We can search for heterogeneous effects for obese people; however, obesity has many dimensions. Hence, for simplicity, we often make recommendations based on only the body mass index (BMI), e.g., "take the drug if BMI ≥ 25." However, since $E(Y^1 - Y^0|\text{BMI})$ is an average, $E(Y^1 - Y^0|\text{BMI}, \text{age})$ may take considerably different values, leading to different recommendations depending on (BMI, age). We may further consider (BMI, age, gender), and so on. There exists no limit to augmenting the conditioning set for a maximal heterogeneity. Instead, we seek a minimal heterogeneity representation.

A potential candidate for the one-dimensional heterogeneity function of X is the propensity score (PS) $E(D|X)$ when D is exogenous. However, the PS is inappropriate if D is endogenous. Instead, if a binary IV δ is available for D, then the IS $\lambda_X \equiv E(\delta|X)$ can be used. Since IS is of no use when δ is randomized, we focus on *non-randomized IV* δ and $E(Y^1 - Y^0|\text{CP}, \lambda_X)$ in this article. In a related study, Choi et al. [19] addressed the randomized IV case by using the "control PS" $E(D|\delta = 0, X)$ as a dimension-reducing device, whereas we use $E(D|\text{CP}, X) = \lambda_X$; this equality is proven below.

The main motivation to condition on λ_X among all functions of X is that λ_X summarizes the information in X for the relationship between δ and (D^0, D^1, Y^0, Y^1) in the sense that

$$\delta \perp\!\!\!\perp (D^0, D^1, Y^0, Y^1)|X \Rightarrow \delta \perp\!\!\!\perp (D^0, D^1, Y^0, Y^1)|\lambda_X,$$

which is proven later. In other words, λ_X is a "sufficient statistic" for the parameter capturing the relationship between δ and (D^0, D^1, Y^0, Y^1). Taking δ as the underlying treatment, as in this study, is well rooted in the literature, particularly in the ratio form "Wald estimator."

Another attractive property of λ_X is that it is a "balancing score" (i.e., $\delta \perp\!\!\!\perp X|\lambda_X$). If ω_X is also a balancing score ($\delta \perp\!\!\!\perp X|\omega_X$), then ω_X should be finer than λ_X, as $\lambda_X = f(\omega_X)$ for a function f. In this sense, λ_X *minimally captures the effect heterogeneity*. If there is no problem accepting PS as a minimal dimension-reducing device because PS is the coarsest balancing score with $D \perp\!\!\!\perp X|\text{PS}$ [20] for exogenous D, then there should be no problem accepting IS with $\delta \perp\!\!\!\perp X|\text{IS}$ for endogenous D.

To see how informative λ_X is relative to X, note that $E(Y^1 - Y^0|\text{CP}, \lambda_X)$ being a non-constant function of λ_X indicates at least the presence of effect heterogeneity. Furthermore, knowing λ_X is as good as knowing X for CP in the sense $E(D|\text{CP}, X) = \text{IS}$:

$$E(D|\,\text{CP}\,, X) = E(\delta|\,\text{CP}\,, X) = E(\delta|X) \equiv \lambda_X \quad \text{under } \delta \perp\!\!\!\perp (D^0, D^1, Y^0, Y^1)|X.$$

For us, this is the primary advantage (along with the weak requisite assumptions) of the CP effect. The disadvantages are that it is specific to each IV, it is not a generally interesting effect, the "monotonicity condition" ($D^0 \le D^1$) can be violated in observational data [21], and the CPs are not the policy-relevant population unless the policy is the same as δ [22].

Let $1[A] \equiv 1$ if A holds and 0 otherwise. For parameters (β_δ, ζ_x) with $\beta_\delta > 0$ and an error term $U \perp\!\!\!\perp (\delta, X)$, suppose $D = 1[0 < Y^1 - Y^0 + \beta_\delta\delta]$ and $Y^1 - Y^0 = X'\zeta_x - U$: $D = 1$ if the "gain" $Y^1 - Y^0$ (plus $\beta_\delta\delta$) from the treatment is greater than 0, which is plausible. Since $D^0 = 1[0 < Y^1 - Y^0]$ and $D^1 = 1[0 < Y^1 - Y^0 + \beta_\delta]$, the CPs satisfy $X'\zeta_x < U < X'\zeta_x + \beta_\delta$. Then, $E(Y^1 - Y^0|\text{CP}, X)$ depends on X only through $X'\zeta_X$ because

$$E(X'\zeta_x - U|X'\zeta_x < U < X'\zeta_x + \beta_\delta, X) = X'\zeta_x - \int_{X'\zeta_x}^{X'\zeta_x+\beta_\delta} u\mathrm{d}F_u(u)/\{F_u(X'\zeta_x + \beta_\delta) - F_u(X'\zeta_x)\},$$

F_u is the distribution function of U. It would be difficult to extend the CP effect to non-CPs, as $E(Y^1 - Y^0|\mathrm{CP}, X)$ is nonlinear in $X'\zeta_x$, whereas $E(Y^1 - Y^0|X) = X'\zeta_x$.

In the aforementioned example, although λ_X captures the effect heterogeneity minimally, it can capture it "maximally" if $E(\cdot|\mathrm{CP}, \lambda_X) = E(\cdot|\mathrm{CP}, X'\zeta_x)$, because $E(Y^1 - Y^0|\mathrm{CP}, \lambda_X) = E\{E(Y^1 - Y^0|\mathrm{CP}, X)|\mathrm{CP}, \lambda_X\}$. This can happen if D is designed to benefit some subjects and δ is encouraging them to take D, as both D and δ would be based on the treatment gain $Y^1 - Y^0 = X'\zeta_x - U$.

To accomplish the goal of this research, we use two nonparametric reduced forms (RFs):

$$\begin{aligned} &Y = \mu_1(\lambda_X)D + \mu_0(\lambda_X) + U_1, \quad E(U_1|\delta, \lambda_X) = 0, \quad \mu_1(\lambda_X) \equiv E(Y^1 - Y^0|\,\mathrm{CP}\,, \lambda_X), \\ &\mu_0(\lambda_X) \equiv E\{(Y^1 - Y^0)D^0 + Y^0|\lambda_X\} - \mu_1(\lambda_X)E(D^0|\lambda_X), \end{aligned} \tag{1}$$

$$\begin{aligned} &Y = \mu_1(\lambda_X)P(\,\mathrm{CP}\,|\lambda_X)\delta + \mu_2(\lambda_X) + U_2, \quad E(U_2|\delta, \lambda_X) = 0, \quad P(\,\mathrm{CP}\,|\lambda_X) \equiv E(D^1 - D^0|\lambda_X), \\ &\mu_2(\lambda_X) \equiv E\{(Y^1 - Y^0)D^0 + Y^0|\lambda_X\}, \end{aligned} \tag{2}$$

where μ_1, μ_0, $P(\mathrm{CP}|\lambda_X)$, and μ_2 are unknown functions, and (U_1, U_2) are error terms. These RFs hold for any Y, as long as $Y^1 - Y^0$ makes sense. For example, for categorical Y, $Y^1 - Y^0$ does not make sense, but $Y_j^1 - Y_j^0$ does for the dummy variable Y_j for category j.

We can estimate $\lambda_X \equiv E(\delta|X)$ nonparametrically, but to make our approaches practical, we either adopt the single index assumption $\lambda_X = \Lambda(X'\alpha)$ for an unknown function $\Lambda(\cdot)$ and parameter α or specify $\lambda_X = \Phi(X'\theta)$ for the $N(0, 1)$ distribution function $\Phi(\cdot)$ and a parameter θ. Clearly, the latter strategy is more restrictive than the former. Considering these aspects, *we propose the following three estimators for the λ_X -heterogeneous CP effect*.

The first estimator is based on the single index assumption $\lambda_X = \Lambda(X'\alpha)$ and

$$\frac{E(Y|\lambda_X, \delta = 1) - E(Y|\lambda_X, \delta = 0)}{E(D|\lambda_X, \delta = 1) - E(D|\lambda_X, \delta = 0)} = \frac{\mu_1(\lambda_X)P(\,\mathrm{CP}\,|\lambda_X)}{P(\,\mathrm{CP}\,|\lambda_X)} = \mu_1(\lambda_X) \quad \{\text{from (2)}\}. \tag{3}$$

Hence, we can identify $\mu_1(\lambda_X) \equiv E(Y^1 - Y^0|\mathrm{CP}, \lambda_X)$ with the ratio on the left side. However, instead of conditioning on $\lambda_X = \Lambda(X'\alpha)$ as in (3), we condition on $X'\alpha$ because estimating $\Lambda(\cdot)$ is more challenging than estimating α.

In our second estimator, we assume that $\lambda_X = \Phi(X'\theta)$ to avoid estimating $\Lambda(\cdot)$. In this case, conditioning on $\lambda_X = \Phi(X'\theta)$ as in (3) than on $X'\theta$ is more advantageous because $\Phi(\cdot)$ is a known function well bounded by $[0, 1]$. Because of the assumption $\lambda_X = \Phi(X'\theta)$, the assumptions of the second estimator are more restrictive than those of the first. Both the first and second estimators converge in distribution more slowly than $\sqrt{N}$.

The two ratio estimators suffer from the "excessively small denominator" problem: A near-zero denominator can blow up the ratio. To avoid this, our third estimator applies a power approximation to $\mu_1(\lambda_X)$ in (1). Then, the IVE can estimate $\mu_1(\lambda_X)$ as the slope of D. Although power approximation is a nonparametric method, we regard the approximation to be exact to make our proposal more practical, and we adopt $\lambda_X = \Phi(X'\theta)$ as in the second estimator. Hence, the third estimator is the most restrictive among our three estimators, but it is $\sqrt{N}$-consistent. Because we need only the RF of λ_X and not the structural form (SF), i.e., because we only use the scalar λ_X (i.e., $X'\alpha$ or $X'\theta$) and not individual elements of α or θ, misspecification problems in $\lambda_X = \Phi(X'\theta)$ are not considerably worrisome.

If desired, $E(Y^1 - Y^0|\mathrm{CP})$ can be found from $\mu_1(\lambda_X)$. Note that the CPs are identified conditionally on λ_X under $\delta \perp\!\!\!\perp (D^0, D^1, Y^0, Y^1)|\lambda_X$ and the monotonicity $D^0 \le D^1$:

$$P(D = 1|\delta = 1, \lambda_X) - P(D = 1|\delta = 0, \lambda_X) = P(\,\mathrm{AT\ or\ CP}\,|\lambda_X) - P(\,\mathrm{AT}\,|\lambda_X) = P(\,\mathrm{CP}\,|\lambda_X),$$

where AT indicates "always takers" with $(D^0 = 1, D^1 = 1)$. Moreover, integrating out λ_X renders $P(\mathrm{CP})$. Then, denoting the distribution of $A|B$ as $F_{A|B}$, we can obtain

$$E(Y^1 - Y^0|\,\mathrm{CP}) = \int \mu_1(l)\mathrm{d}F_{\lambda_X|\,\mathrm{CP}}(l) = \int \mu_1(l)P(\,\mathrm{CP}\,|\lambda_X = l)\mathrm{d}F_{\lambda_X}(l)/P(\,\mathrm{CP}).$$

In the remainder of this article, Sections 2 and 3 explain the two ratio estimators and IVE, respectively; Sections 4 and 5 present simulation and empirical studies, respectively; and Section 6 concludes this article. Most proofs are presented in the appendix. As usual, we assume independent and identically distributed observations $(\delta_i, D_i, X_i, Y_i)$, $i = 1, \ldots, N$.

2 Ratio approaches

2.1 Identification

Since (2) motivates the ratio estimators, we first prove (2) in Theorem 1, and (1) is proven in the next section. As preliminaries, we present *our assumptions on IV*:

$$\begin{aligned} &\text{(i)}: \quad \delta \perp\!\!\!\perp (D^0, D^1, Y^0, Y^1)|\lambda_X \quad \text{for all } \lambda_X \quad \text{(IV exogeneity)}, \\ &\text{(ii)}: \quad D^0 \le D^1|\lambda_X \quad \text{for all } \lambda_X \quad \text{(Monotonicity)}, \\ &\text{(iii)}: \quad E(D^1 - D^0|\lambda_X) \neq 0 \quad \text{for all } \lambda_X \quad \text{(IV relevance)}. \end{aligned} \tag{4}$$

The "IV exclusion restriction" is implicit in the notation (Y^0, Y^1), because if δ directly affected Y, then the potential responses would be double-indexed by (δ, D).

The appendix proves the main heterogeneity-dimension reduction idea:

$$\delta \perp\!\!\!\perp (D^0, D^1, Y^0, Y^1)|X \Rightarrow \delta \perp\!\!\!\perp (D^0, D^1, Y^0, Y^1)|\lambda_X. \tag{5}$$

Although (4)(i) is enough for our estimators below, bear in mind that "$\lambda_X = E(D|\mathrm{CP}, X)$" noted in the preceding section to better motivate/interpret λ_X requires "$\delta \perp\!\!\!\perp (D^0, D^1, Y^0, Y^1)|X$" which is stronger than (4)(i) as (5) shows.

The proof for (5) also establishes the "balancing score" property of IS: The distribution of X is the same across the $\delta = 0, 1$ groups once λ_X is conditioned on. Then, it follows from Theorem 2 of Rosenbaum and Rubin [20] that *λ_X is the coarsest balancing score*. In other words, a function, e.g., ω_X, of X is a balancing score (i.e., $\delta \perp\!\!\!\perp X|\omega_X$), iff $\lambda_X = f(\omega_X)$ for some function $f(\cdot)$. In this sense, $\mu_1(\lambda_X) \equiv E(Y^1 - Y^0|\mathrm{CP}, \lambda_X)$ minimally captures the CP effect heterogeneity, thereby avoiding the dimension problem in $E(Y^1 - Y^0|\mathrm{CP}, X)$.

Theorem 1. *Under* (4)(*i*), (4)(*ii*), *and the support-overlap condition* $0 < \lambda_X < 1$, (2) *holds for any form of Y as long as* $Y^1 - Y^0$ *makes sense*.

The qualifier "for all λ_X" in (4)(i) and (ii) is sufficient but not necessary, because if the conditions hold only for some values of λ_X, then Theorem 1 holds only for those values of λ_X for which the CP effect can still be identified.

2.2 Ratio estimator under single index assumption

Our single index assumption for dimension reduction with an unknown $\Lambda(\cdot)$ and α is

$$\lambda_X \equiv E(\delta|X) = \Lambda(X'\alpha). \tag{6}$$

To estimate $\Lambda(\cdot)$ and α, we minimize $\sum_i\{\delta_i - L(X_i'a)\}^2$ with respect to (wrt) $\{L(\cdot), a\}$ as Ichimura [23] did, which raises identification issues. First, the intercept for $X'\alpha$ is not identified, because $\Lambda(\alpha_0 + X'\alpha)$ can be written as $\Lambda_0(X'\alpha)$, where $\Lambda_0(\cdot) \equiv \Lambda(\alpha_0 + \cdot)$; i.e., we can identify both $\{\Lambda(\cdot), (\alpha_0, \alpha)\}$ and $\{\Lambda_0(\cdot), \alpha\}$ equally well. Second,

since $\Lambda(X'\alpha) = \Lambda(c \cdot X'\alpha/c)$ for any $c \neq 0$, instead of identifying $\{\Lambda(\cdot), \alpha\}$, we can identify $\{\Lambda_c(\cdot), \alpha/c\}$ equally well, where $\Lambda_c(\cdot) \equiv \Lambda(c\,\cdot)$. Clearly, the scale and sign of α are not identified.

One approach to overcome the identification problems is to *assume a continuous regressor with a non-zero slope*, e.g., the last regressor X_k. Then, divide $X'\alpha$ by $\alpha_k \neq 0$ to obtain the identified parameter $(\alpha_1/\alpha_k, \ldots, \alpha_{k-1}/\alpha_k, 1)$. If there is no such regressor, then $X'\alpha$ would be discrete, and we would not be able to trace the entire shape of $\Lambda(\cdot)$ with $X'\alpha$.

Since we further assume a strictly increasing $\Lambda(\cdot)$, we can identify the sign of α_k. We divide $X'\alpha$ by $|\alpha_k|$ and not by α_k to obtain $\{\alpha_1/|\alpha_k|, \ldots, \alpha_{k-1}/|\alpha_k|, sign(\alpha_k)\}$, where $sign(\alpha_k) = 1$ if $\alpha_k > 0$, 0 if $\alpha_k = 0$, and -1 if $\alpha_k < 0$. Then, we try both $(\alpha_1/|\alpha_k|, \ldots, \alpha_{k-1}/|\alpha_k|, 1)$ and $(\alpha_1/|\alpha_k|, \ldots, \alpha_{k-1}/|\alpha_k|, -1)$ to select the one that minimizes $\sum_i\{\delta_i - L(X_i'a)\}^2$.

Let $S_i \equiv X_i'\alpha$, and let G_j denote the group with $\delta = j$, $j = 0, 1$. Given kernel K, bandwidth h, and the sample size N_j for G_j, and recalling $\mu_1(\lambda_X) \equiv E(Y^1 - Y^0|\text{CP}, \lambda_X)$, our first ratio estimator for $\mu_1(s) = E(Y^1 - Y^0|\text{CP}, S = s)$ is

$$\begin{aligned}
\hat{\mu}_1(s) &\equiv \frac{\hat{b}(s)}{\hat{a}(s)} \quad \text{where } \hat{a}(s) \equiv \hat{a}_1(s) - \hat{a}_0(s), \quad \hat{b}(s) \equiv \hat{b}_1(s) - \hat{b}_0(s), \\
\hat{a}_j(s) &\equiv \frac{(N_j h)^{-1}\sum_{i\in G_j} K\{(S_i - s)/h\}D_i}{(N_j h)^{-1}\sum_{i\in G_j} K\{(S_i - s)/h\}} \equiv \frac{\hat{a}_{jd}(s)}{\hat{c}_j(s)}, \quad j = 0, 1, \\
\hat{b}_j(s) &\equiv \frac{(N_j h)^{-1}\sum_{i\in G_j} K\{(S_i - s)/h\}Y_i}{(N_j h)^{-1}\sum_{i\in G_j} K\{(S_i - s)/h\}} \equiv \frac{\hat{b}_{jy}(s)}{\hat{c}_j(s)}, \quad j = 0, 1.
\end{aligned} \tag{7}$$

The numerator of $\hat{a}_j(s)$ is defined as $\hat{a}_{jd}(s)$, where subscript d refers to D in $\hat{a}_j(s)$, and the numerator of $\hat{b}_j(s)$ is defined as $\hat{b}_{jy}(s)$ for the analogous reason. The common denominator of $\hat{a}_j(s)$ and $\hat{b}_j(s)$ is $\hat{c}_j(s) \to^p f_{Sj}(s)$, which is the density of $S|(\delta = j)$.

There are six estimators in $\hat{\mu}_1(s)$: $\{\hat{a}_{0d}(s), \hat{b}_{0y}(s), \hat{c}_0(s)\}$ and $\{\hat{a}_{1d}(s), \hat{b}_{1y}(s), \hat{c}_1(s)\}$. Computing $\hat{\mu}_1(s)$ is not involved, but estimating the asymptotic variance of $\hat{\mu}_1(s)$ is: it consists of six variances, and six covariances among $\{\hat{a}_{0d}(s), \hat{b}_{0y}(s), \hat{c}_0(s)\}$ and $\{\hat{a}_{1d}(s), \hat{b}_{1y}(s), \hat{c}_1(s)\}$. Regarding the selection of h, since only one-dimensional nonparametric estimators appear in $\hat{\mu}_1(s)$, it is preferable to select h in practice through "eye-balling" or cross-validation.

The probability limits of $\{\hat{a}_j(s), \hat{b}_j(s)\}$ and those of $\{\hat{a}(s), \hat{b}(s), \hat{\mu}_1(s)\}$ can be seen in

$$\begin{aligned}
&a_j(s) \equiv E(D|S = s, \delta = j), \quad b_j(s) \equiv E(Y|S = s, \delta = j), \\
\Rightarrow\ &\mu_1(s) = \frac{b(s)}{a(s)} \equiv \frac{b_1(s) - b_0(s)}{a_1(s) - a_0(s)} = \frac{E(Y|S = s, \delta = 1) - E(Y|S = s, \delta = 0)}{E(D|S = s, \delta = 1) - E(D|S = s, \delta = 0)}.
\end{aligned} \tag{8}$$

That α in $S = X'\alpha$ has to be estimated can be ignored, as α can be estimated $\sqrt{N}$-consistently. In other words, α is as good as known for $\hat{\mu}_1(s)$ that is $\sqrt{Nh}$-consistent. This is the reason why we condition on $S = X'\alpha$ instead of $\Lambda(X'\alpha)$. To make conditioning on $S = X'\alpha$ equivalent to that on $\Lambda(X'\alpha)$, we require $\Lambda(\cdot)$ to be strictly increasing. In the following, we often write "$S = s$" or "$P = p$" in conditioning sets simply as "s" or "p".

Theorem 2. $\sqrt{Nh}\{\hat{\mu}_1(s) - \mu_1(s)\} \to^d N\{0, \sum_{j=1}^6 (V_j + 2C_j)\}$, *where the* V_j*'s and* C_j*'s are in* (A6)–(A7) *of the appendix, with* $\kappa \equiv \int K(t)^2 \mathrm{d}t$ *and* $\pi_j \equiv \lim_{N\to\infty} N_j/N$, *and the requisite assumptions are as follows.* (*i*) (4) *holds for all* $\lambda_X \equiv \Lambda(X'\alpha)$ *and* $0 < \lambda_X < 1$; (*ii*) $\Lambda(\cdot)$ *is strictly increasing and differentiable*; (*iii*) *a continuous* X_k *with* $\alpha_k \neq 0$ *exists*; (*iv*) α *is interior to a compact parameter space* A_α; (*v*) $K(\cdot)$ *is symmetric about* 0 *and twice continuously differentiable with* $\int K(t)\mathrm{d}t = 1$, *the second derivative bounded, and* $K(t) = 0$ *for* $|t| > 1$; (*vi*) $Nh^8 \to 0$ *and* $Nh^{3+\nu}/(-\ln h) \to \infty$ *for an arbitrarily small* $\nu > 0$ *as* $N \to \infty$; (*vii*) $f_{Sj}(s) > 0$ *for* $j = 0, 1$ *and* $s = x'\alpha$, *with* x *on a compact set* $\check{X}$; *also,* $f_{Sj}(s)$, $E(Y|s, \delta = j)$, $E(D|s, \delta = j)$, $E(Y^2|s, \delta = j)$, *and* $E(YD|s, \delta = j)$ *are three times continuously differentiable wrt* $s = x'\alpha$, *with the third derivatives bounded uniformly on* A_α.

In Theorem 2, (i) is assumed to identify a; note that (4)(iii) implies $a(s) > 0$ in view of (3), because $E(D^1 - D^0|\lambda_X) = P(\text{CP}|\lambda_X)$. The strictly increasing $\Lambda(\cdot)$ part in (ii) and (iii) is aimed at identifying $\Lambda(\cdot)$, and the differentiability in (ii) pertains to Assumption 4.1 of Ichimura [23, p. 81]. Also, (iv) is a standard assumption for asymptotic normality. In (v), "$K(t) = 0$ *for* $|t| > 1$" is the same as Assumption 5.6 (4) of Ichimura [23, p. 88]. The conditions for h in (vi) pertain to the $\sqrt{N}$-consistency and asymptotic normality of $\hat{a}$ in Ichimura's Theorem 5.2 [23, p. 94] with his $m = \infty$ for binary D as the dependent variable. Part of the conditions in (vii) is used to satisfy Assumptions 5.3–5.5 of Ichimura [23, p. 87].

Although not mentioned in Theorem 2, Assumption 4.2 of Ichimura [23, p. 82] must be introduced if several components of X are deterministically determined by other components in X. We do not mention this assumption in our Theorem 2, as the incentive to use such components is weak in single index estimation. For instance, $\Lambda(X'a) = (X'a) + (X'a)^2$ means that all squared components of X are used in $\Lambda(X'a)$, and thus, these squared components do not need to be used as part of X.

Theorem 2 and $\hat{\mu}_1(s)$ may look daunting, but they can be simply summarized as follows. The first step is the single index estimation in which $\{L(\cdot), a\}$ is chosen to minimize $\sum_i\{\delta_i - L(X_i'a)\}^2$. We use the algorithm proposed by Hastie et al. [24, p. 391], which rapidly converges to a local minimum, if not the global minimum. The second step is to obtain the six kernel estimators constituting $\hat{\mu}_1(s)$. Hence, the only complication in our proposal is estimating the asymptotic variance of $\hat{\mu}_1(s)$. Our simulation study will demonstrate that the asymptotic variance formula works fairly well with $N = 500$, and very well with $N = 5{,}000$.

For our third estimator using (1) under $\lambda_X = \Phi(X'\theta)$, let $\hat{\theta}$ be the probit estimator and $\hat{\theta}_k$ be the slope of X_k. In comparing $\hat{\mu}_1(s)$ to the third estimator, we condition on $X'\hat{a}\cdot|\hat{\theta}_k|$ and not on $X'\hat{a}$ to ensure that the slope of X_k in $X'\hat{a}\cdot|\hat{\theta}_k|$ becomes $\hat{\theta}_k$ as in $X'\hat{\theta}$. To show this aspect, suppose $\theta_k > 0$. Then, $a_k = 1$, and thus, the slope of X_k in $X'\hat{a}\cdot|\hat{\theta}_k|$ is $\hat{\theta}_k$ as in $X'\hat{\theta}$. Now, suppose $\theta_k < 0$. Then, $a_k = -1$, so that the slope of X_k in $X'\hat{a}\cdot|\hat{\theta}_k|$ is $-|\hat{\theta}_k| = \hat{\theta}_k$ as in $X'\hat{\theta}$.

2.3 Ratio estimator under the probit IS

The requisite conditions for the preceding ratio estimator and its asymptotic variance are not simple. Our second ratio estimator is simpler in terms of the requisite conditions and asymptotic variance, although it requires imposing the probit assumption $\lambda_X = \Phi(X'\theta)$. Since θ can be estimated $\sqrt{N}$-consistently by the probit estimator $\hat{\theta}$, we treat θ as known for the second ratio estimator conditioning on $\Phi(X'\hat{\theta})$. Although the assumption $\lambda_X = \Phi(X'\theta)$ may appear restrictive, it is not so, because we need only the RF λ_X. For instance, suppose $\delta = 1[0 < \xi(X) + \sigma(X)\varepsilon]$, where $\xi(X)$ and $\sigma(X) \neq 0$ are unknown functions and $\varepsilon \sim N(0,1) \perp\!\!\!\perp X$. Then, $E(\delta|X) = \Phi\{\xi(X)/\sigma(X)\}$, and we estimate the linearized version $X'\theta \propto \xi(X)/\sigma(X)$ to use only $\Phi(X'\theta)$ and not the individual components of θ.

Under $\lambda_X = \Phi(X'\theta)$, the appendix proves that the ratio in (3) with $\lambda_X = p$ is

$$\frac{E(B|\lambda_X = p)}{E(A|\lambda_X = p)} \quad \text{where} \quad B \equiv Y(\delta - p) \text{ and } A \equiv D(\delta - p). \tag{9}$$

Then, with $P_i \equiv \lambda_{X_i}$, our estimator for (9) is

$$\tilde{\mu}_1(p) \equiv \left\{\frac{\sum_i K\{(P_i - p)/h\}B_i}{\sum_i K\{(P_i - p)/h\}}\right\} \Big/ \left\{\frac{\sum_i K\{(P_i - p)/h\}A_i}{\sum_i K\{(P_i - p)/h\}}\right\} = \frac{\sum_i K\{(P_i - p)/h\}B_i}{\sum_i K\{(P_i - p)/h\}A_i};$$

the notation $\tilde{\mu}_1(\cdot)$ is used for our second ratio estimator.

Theorem 3. $\sqrt{Nh}\{\tilde{\mu}_1(p) - \mu_1(p)\}$ *is asymptotically normal with variance*

$$\frac{\kappa}{f_\lambda(p)\{E(A|p)\}^4}[E(A^2|p)\{E(B|p)\}^2 + E(B^2|p)\{E(A|p)\}^2 - 2E(A|p)E(B|p)E(AB|p)]$$

under the following assumptions, where $\kappa \equiv \int K(t)^2 \mathrm{d}t$, *and* f_λ *is the density of* λ_X. (*i*) (4) *holds for all* $\lambda_X \equiv \Phi(X'\theta)$, *and* $0 < \lambda_X < 1$; (*ii*) *a continuous* X_k *with* $\theta_k \neq 0$ *exists*; (*iii*) θ *is interior to a compact parameter space* A_θ; (*iv*) $K(\cdot)$ *is symmetric about* 0 *and twice continuously differentiable with* $\int K(t)\mathrm{d}t = 1$; (v) $Nh^4 \to 0$ *and* $Nh \to \infty$ *as* $N \to \infty$; *and* (*vi*) $f_\lambda(\cdot)$, $E(A|\lambda_X)$, $E(A^2|\lambda_X)$, $E(B|\lambda_X)$, $E(B^2|\lambda_X)$, *and* $E(AB|\lambda_X)$ *are twice continuously differentiable with* $E(A|\lambda_X), f_\lambda(\lambda_X) > 0$ *for all* λ_X.

The aforementioned assumptions are weaker than those in Theorem 2 because single index estimation is not needed. Moreover, (ii) is not essential, because no continuous covariate means that nonparametric estimation is not needed. If D is exogenous such that $\delta = D$, then we can replace A with 1 in $\tilde{\mu}_1(p)$ to obtain the usual kernel estimator for $E(B|p)$. In this case, the asymptotic variance in Theorem 3 is simplified to $\kappa V(B|p)/f_\lambda(p)$. Note that although we use the same notation X for probit, X should be augmented by the constant 1 for probit.

3 Power approximation approach

3.1 Identification

We apply a power approximation to $\mu_1(\lambda_X) \equiv E(Y^1 - Y^0|\mathrm{CP}, \lambda_X)$ in (1) to obtain

$$E(Y^1 - Y^0|\,\mathrm{CP}, \lambda_X) = M'\beta, \quad M \equiv (1, \lambda_X, \ldots, \lambda_X^J)', \quad \beta \equiv (\beta_0, \beta_1, \ldots, \beta_J)' \Rightarrow Y = DM'\beta + \mu_0(\lambda_X) + U_1. \tag{10}$$

This yields a moment condition $E(\mathrm{IV} \times \mathrm{error}) = 0$ with the IV $(\delta - \lambda_X)M$:

$$0 = E[(\delta - \lambda_X)M \cdot\{\mu_0(\lambda_X) + U_1\}] = E\{(\delta - \lambda_X)M \cdot(Y - DM'\beta)\}.$$

With $E^{-1}(\cdot)$ denoting $\{E(\cdot)\}^{-1}$, we solve $E\{(\delta - \lambda_X)M \cdot(Y - DM'\beta)\} = 0$ for β:

$$\beta = E^{-1}\{(\delta - \lambda_X)DMM'\}E\{(\delta - \lambda_X)MY\}, \quad E(Y^1 - Y^0|\,\mathrm{CP}, \lambda_X) = M'\beta. \tag{11}$$

The main identification finding in this section is Theorem 4.

Theorem 4. *Under* (4)(*i*) *and* (*ii*), (1) *holds for any* Y *as long as* $Y^1 - Y^0$ *makes sense. If* (4)(*iii*), $0 < \lambda_X < 1$, *and* $\mu_1(\lambda_X) \equiv E(Y^1 - Y^0|\mathrm{CP}, \lambda_X) = \sum_{j=0}^J \beta_j \lambda_X^j = M'\beta$ *hold additionally, then* $\beta = (\beta_0, \beta_1, \ldots, \beta_J)'$ *is identified.*

The proof for Theorem 4 in the appendix reveals

$$E\{(\delta - \lambda_X)DMM'\} = E\{E(D^1 - D^0|\lambda_X)(1 - \lambda_X)\lambda_X MM'\}. \tag{12}$$

Hence, even if $E(D^1 - D^0|\lambda_X) \neq 0$ only for some values of λ_X in (4)(iii), Theorem 4 holds as long as $E\{(\delta - \lambda_X)DMM'\}$ is invertible.

3.2 Power approximation estimator

Before estimation, we introduce a modification to $Y = \mu_1(\lambda_X)D + \mu_0(\lambda_X) + U_1$ in (1) in case λ_X is misspecified. When we use $\delta - \lambda_X$ as the IV, we can ignore $\mu_0(\lambda_X)$ in the composite error $\mu_0(\lambda_X) + U_1$, because the IV $\delta - \lambda_X$ is orthogonal to $\mu_0(\lambda_X)$. However, λ_X may be misspecified in practice, which can result in a bias because the misspecified $\delta - \lambda_X$ may be correlated with $\mu_0(\lambda_X)$. Hence, it is better to remove $\mu_0(\lambda_X)$ from the error. To this end, we use $Y - E(Y|\lambda_X)$ as the outcome variable as follows (see Chernozhukov et al. [25] and Lee [26,27] for closely related ideas).

Take $E(\cdot|\lambda_X)$ on (1) to obtain $E(Y|\lambda_X) = \mu_1(\lambda_X)E(D|\lambda_X) + \mu_0(\lambda_X)$. Subtract this expression from (1) to remove $\mu_0(\lambda_X)$ and obtain

$$Y - E(Y|\lambda_X) = \mu_1(\lambda_X)D - \mu_1(\lambda_X)E(D|\lambda_X) + U_1. \tag{13}$$

We use (13) instead of (1) to implement our IVE, where $-\mu_1(\lambda_X)E(D|\lambda_X) + U_1$ is the composite error. Accounting for $E(Y|\lambda_X)$ explicitly in (13) tends to improve the finite sample performance of the following IVE by decreasing the error term standard deviation (SD) and making the IVE relatively insensitive to misspecifications of λ_X.

We present two versions of the power approximation estimator: one conditioned on $X'\theta$ and the other conditioned on $\Phi(X'\theta)$. The former estimator is simpler, but the latter estimator is likely to perform better when $X'\theta$ is large, thereby causing the power functions of $X'\theta$ to become even larger to generate outliers. In contrast, power functions of $\Phi(X'\theta)$ become smaller as $X'\theta$ becomes large, and thus, no outlier problem arises.

First, we condition on $X'\theta$ and apply a power approximation to $E(Y|X'\theta)$ and $\mu_1(X'\theta)$: For some γ parameters and error term U,

$$Y - W'\gamma = DW'\beta - \mu_1(X'\theta)E(D|X'\theta) + U, \quad \text{where } W = \{1, X'\theta, \ldots, (X'\theta)^J\}' \quad \text{and} \quad \gamma \equiv (\gamma_0, \ldots, \gamma_J)'.$$

Using the same J-order power approximation for both $W'\gamma$ and $W'\beta$ is not necessary, because $W'\gamma$ in $Y - W'\gamma$ can be replaced by $\bar{Y}$ (with $J = 0$) or even zero. If desired, we can also obtain $E(Y^1 - Y^0|\text{CP}, \lambda_X = p)$ from $W'\beta$: Replacing $X'\theta$ in W with $\Phi^{-1}(p)$ yields

$$E(Y^1 - Y^0|\,\text{CP}, \lambda_X = p) = [1, \Phi^{-1}(p), \ldots, \{\Phi^{-1}(p)\}^J]\beta. \tag{14}$$

Let $\hat{\theta}$ be the probit estimator of δ on X. Replace $X'\theta$ with $X'\hat{\theta}$, and then replace $E(Y|X'\theta)$ with the ordinary least-squares estimator (OLS) predicted value $\hat{W}'\hat{\gamma}$, where

$$\hat{\gamma} \text{ is the OLS of } Y \text{ on } \hat{W} \equiv W(\hat{\theta}) \equiv \{1, X'\hat{\theta}, \ldots, (X'\hat{\theta})^J\}'.$$

Then, we obtain the IVE of $Y - \hat{W}'\hat{\gamma}$ on $D\hat{W}$ with $\hat{\varepsilon}\hat{W} \equiv \{\delta - \Phi(X'\hat{\theta})\}\hat{W}$ as the IV:

$$\hat{\beta} \equiv \left(\sum_i \hat{\varepsilon}_i D_i \hat{W}_i \hat{W}_i'\right)^{-1} \sum_i \hat{\varepsilon}_i \hat{W}_i (Y_i - \hat{W}_i'\hat{\gamma}), \quad \hat{\varepsilon}_i \equiv \delta_i - \Phi(X_i'\hat{\theta}). \tag{15}$$

The appendix proves the next two theorems.

Theorem 5. *The IVE in* (15) *is asymptotically normal with variance* Ω_1:

$$\begin{aligned}
&\sqrt{N}(\hat{\beta} - \beta) \to^d N(0, \Omega_1), \quad \Omega_1 \equiv E^{-1}(\varepsilon DWW')E(\eta_1\eta_1')E^{-1}(\varepsilon DWW'),\\
&\eta_1 \equiv V\varepsilon W - E\{D\nabla W'\beta\varepsilon WX' + V\phi(X'\theta)WX' - V\varepsilon\nabla WX'\}\eta_{\hat{\theta}}, \quad \varepsilon \equiv \delta - \Phi(X'\theta),\\
&V \equiv Y - W'\gamma - DW'\beta, \quad \nabla W \equiv \{0, 1, 2(X'\theta), \ldots, J(X'\theta)^{J-1}\}',
\end{aligned}$$

and $\eta_{\hat{\theta}}$ *is the influence function for* $\sqrt{N}(\hat{\theta} - \theta)$. Ω_1 *can be consistently estimated with*

$$\begin{aligned}
&\hat{\Omega}_1 \equiv \left(\frac{1}{N}\sum_i \hat{\varepsilon}_i D_i \hat{W}_i \hat{W}_i'\right)^{-1} \frac{1}{N}\sum_i \hat{\eta}_{1i}\hat{\eta}_{1i}' \left(\frac{1}{N}\sum_i \hat{\varepsilon}_i D_i \hat{W}_i \hat{W}_i'\right)^{-1}, \quad \hat{\varepsilon}_i \equiv \delta_i - \Phi(X_i'\hat{\theta}),\\
&\hat{\eta}_{1i} \equiv \hat{V}_i\hat{\varepsilon}_i\hat{W}_i - \frac{1}{N}\sum_k \{D_k\nabla\hat{W}_k'\hat{\beta}\hat{\varepsilon}_k\hat{W}_kX_k' + \hat{V}_k\phi(X_k'\hat{\theta})\hat{W}_kX_k' - \hat{V}_k\hat{\varepsilon}_k\nabla\hat{W}_kX_k'\}\hat{\eta}_{\hat{\theta}i},\\
&\hat{V}_i \equiv Y_i - \hat{W}_i'\hat{\gamma} - \hat{W}_i'D_i'\hat{\beta}, \quad \nabla\hat{W}_i \equiv \{0, 1, 2(X_i'\hat{\theta}), \ldots, J(X_i'\hat{\theta})^{J-1}\}',\\
&\hat{\eta}_{\hat{\theta}i} \equiv \left(\frac{1}{N}\sum_k \hat{s}_k\hat{s}_k'\right)^{-1}\hat{s}_i, \quad \textit{where } \hat{s}_i \equiv \frac{\{\delta_i - \Phi(X_i'\hat{\theta})\}\phi(X_i'\hat{\theta})}{\Phi(X_i'\hat{\theta})\{1 - \Phi(X_i'\hat{\theta})\}}X_i.
\end{aligned}$$

Now, we condition on λ_X and apply power approximations to $E\{Y|\Phi(X'\theta)\}$ and $\mu_1\{\Phi(X'\theta)\}$:

$$Y - M'\gamma = DM'\beta - \mu_1\{\Phi(X'\theta)\}E\{D|\Phi(X'\theta)\} + U, \quad M = \{1, \Phi(X'\theta), \ldots, \Phi(X'\theta)^J\}'.$$

Replace $E\{Y|\Phi(X'\theta)\}$ with the OLS-predicted value $\hat{M}'\tilde{\gamma}$, where

$$\tilde{\gamma} \text{ is the OLS of } Y \text{ on } \hat{M} \equiv M(\hat{\theta}) \equiv \{1, \Phi(X'\hat{\theta}), \ldots, \Phi(X'\hat{\theta})^J\}'.$$

Obtain the IVE of $Y - \hat{M}'\tilde{\gamma}$ on $D\hat{M}$ with $\hat{\varepsilon}\hat{M} \equiv \{\delta - \Phi(X'\hat{\theta})\}\hat{M}$ as the IV:

$$\tilde{\beta} \equiv \left(\sum_i \hat{\varepsilon}_i D_i \hat{M}_i \hat{M}_i'\right)^{-1} \sum_i \hat{\varepsilon}_i \hat{M}_i (Y_i - \hat{M}_i'\tilde{\gamma}). \tag{16}$$

Theorem 6. *The IVE in* (16) *is asymptotically normal with variance* Ω_2:

$$\sqrt{N}(\tilde{\beta} - \beta) \to^d N(0, \Omega_2), \quad \Omega_2 \equiv E^{-1}(\varepsilon DMM')E(\eta_2\eta_2')E^{-1}(\varepsilon DMM'),$$
$$\eta_2 \equiv \Gamma\varepsilon M - E\{D\nabla M'\beta\varepsilon MX' + \Gamma\phi(X'\theta)MX' - \Gamma\varepsilon\nabla MX'\}\eta_{\hat{\theta}},$$
$$\Gamma \equiv Y - M'\gamma - DM'\beta, \quad \nabla M \equiv \{0, \phi(X_i'\theta), 2\phi(X'\theta), \ldots, J\phi(X'\theta)^{J-1}\}',$$
$$\hat{\Omega}_2 \equiv \left(\frac{1}{N}\sum_i \hat{\varepsilon}_i D_i \hat{M}_i \hat{M}_i'\right)^{-1} \frac{1}{N}\sum_i \hat{\eta}_{2i}\hat{\eta}_{2i}' \left(\frac{1}{N}\sum_i \hat{\varepsilon}_i D_i \hat{M}_i \hat{M}_i'\right)^{-1} \to^p \Omega_2,$$
$$\hat{\eta}_{2i} \equiv \tilde{\Gamma}_i\hat{\varepsilon}_i\hat{M}_i - \frac{1}{N}\sum_k \{D_k\nabla\hat{M}_k'\tilde{\beta}\hat{\varepsilon}_k\hat{M}_kX_k' + \tilde{\Gamma}_k\phi(X_k'\hat{\theta})\hat{M}_kX_k' - \tilde{\Gamma}_k\hat{\varepsilon}_k\nabla\hat{M}_kX_k'\}\hat{\eta}_{\hat{\theta}i},$$
$$\tilde{\Gamma}_i \equiv Y_i - \hat{M}_i'\tilde{\gamma} - \hat{M}_i'D_i'\tilde{\beta}, \quad \nabla\hat{M}_i \equiv \{0, \phi(X_i'\hat{\theta}), 2\phi(X_i'\hat{\theta}), \ldots, J\phi(X_i'\hat{\theta})^{J-1}\}'.$$

4 Simulation study

With N = 500, 5,000, and 10,000 simulation repetitions, our simulation setting is

$$\delta = 1[0 < \pi_1 + \pi_2X_2 + \pi_3X_3 + \xi], \quad X_2 \text{ discrete uniform on } \{-0.5, -0.25, 0, 0.25, 0.5\},$$
$$X_3 \sim U[-0.5, 0.5], \quad \xi \sim N(0,1) \bigsqcup (X_2, X_3), \quad \pi_1 = 0, \quad \pi_2 = \pi_3 = 1,$$
$$D = 1[0 < \tau_1 + \tau_2X_2 + \tau_3X_3 + \tau_\delta\delta + \psi], \quad \psi \sim N(0,1) \bigsqcup (\xi, X_2, X_3),$$
$$\tau_1 = 0, \quad \tau_2 = \tau_3 = \tau_\delta = 1, \quad Y^{0*} = 1 + X_2 + X_3 + U, \quad U = e + \psi,$$
$$e \sim N(0,1) \bigsqcup (\xi, \psi, X_2, X_3), \quad Y^{1*} = Y^{0*} + \rho_1(X_2 + X_3) + \rho_2(X_2 + X_3)^2, \quad \rho_1 = 1,$$
$$\rho_2 = 0, -0.5 \text{ (linear, quadratic effect)}, \; Y = Y^0 + (Y^1 - Y^0)D, \quad \text{where}$$
$$Y^d = Y^{d*} \text{ for continuous } Y, \text{ and } Y^d = 1[0 < Y^{d*}] \text{ for binary } Y, \quad d = 0, 1.$$

Since $\pi_2 = \pi_3 = 1$ in the δ model, $X'\pi = X_2 + X_3$ has the range $[-1, 1]$ owing to the range of (X_2, X_3). The error term U in Y^0 is correlated with the error term ψ in D with $\text{Cor}(U, \psi) \simeq 0.71$ to ensure that D is endogenous. We use four simulation designs:

Design 1: Y is continuous, and $\rho_2 = 0$ (effect is $X'\pi$);
Design 2: Y is continuous, and $\rho_2 = -0.5$ (effect is $X'\pi - 0.5(X'\pi)^2$);
Design 3: Y is binary, and $\rho_2 = 0$ (effect is non-linear in $X'\pi$);
Design 4: Y is binary, and $\rho_2 = -0.5$ (effect is non-linear in $X'\pi - 0.5(X'\pi)^2$).

For the first ratio estimator $\hat{\mu}_1(s)$, we can set the slope of X_3 as -1 or 1, but we set the value only to 1 because the sign of the slope of X_3 can be estimated at a rate faster than $N^{-1/2}$ as only one of the two values is selected. The lack of consideration of -1 decreases the simulation time by half. With $\tau_\delta = 1$ in the D model, the CP effect is not zero, but the simulation program crashes when the denominator of $\hat{\mu}_1$ and $\tilde{\mu}_1$ approaches zero. In this case, the simulation run is abandoned, and the data are redrawn. For $K(\cdot)$ of $\hat{\mu}_1$ and $\tilde{\mu}_1$, we use the simple $N(0,1)$ kernel. The bounded quartic kernel is used as well, but then dropped, as the kernel choice does not result in a significant difference. The bandwidth h is chosen initially by cross-validation for a number of times and then fixed throughout the simulation repetitions, as doing cross-validation at each simulation run is time-consuming.

Let $X = (X_2, X_3)'$ and $X_{+1} \equiv (1, X_2, X_3)'$ such that $E(\delta|X) = \Phi(X_{+1}'\theta)$, although we use sometimes the expression $E(\delta|X) = \Phi(X'\theta)$ for simplicity. The following tables present

(1) : $\hat{\mu}_1(s)$ at $s = -0.5,\ 0,$ and 0.5 for $s = x'\hat{\alpha}$ (evaluation points);

(2) : $\tilde{\mu}_1(p)$ at $p = 0.31 = \Phi(-0.5), 0.5 = \Phi(0),$ and $0.69 = \Phi(0.5)$ for $p = \Phi(x'_{+1}\hat{\theta})$;

(3) : $\hat{w}'_J\hat{\beta} \equiv \sum_{j=0}^{J} \hat{\beta}_j (x'_{+1}\hat{\theta})^j$ for $J = 1, 2$ at $x'_{+1}\hat{\theta} = -0.5, 0,$ and 0.5;

(4) : $\hat{m}'_J\tilde{\beta} \equiv \sum_{j=0}^{J} \tilde{\beta}_j \Phi(x'_{+1}\hat{\theta})^j$ for $J = 1, 2$ at $\Phi(x'_{+1}\hat{\theta}) = 0.31, 0.5,$ and 0.69.

Overall, six estimators are compared at the three evaluation points.

For each entry in each following table, four numbers appear at a given evaluation point: the (i) absolute bias (|Bias|); (ii) SD; (iii) averaged SD (across 10,000 repetitions) based on the asymptotic variance to be compared with (ii); and (iv) proportion of the 95% point-wise confidence intervals (CI) capturing the true value. We do not present the root mean-squared error (RMSE) to save space: in most cases, the absolute bias is much smaller than the SD, and thus, the RMSE is similar to the SD. The entries with the subscript "avg" indicate the simple averages across the three evaluation points, which are used as summary measures.

To make $s = x'\hat{\alpha}$ comparable to $p = \Phi(x'_{+1}\hat{\theta})$, we set $s = x'\hat{\alpha} \cdot \hat{\theta}_3$ and not $x'\hat{\alpha}$, so that the slope of x_3 in $x'\hat{\alpha} \cdot \hat{\theta}_3$ becomes $\hat{\theta}_3$ as in $x'_{+1}\hat{\theta}$. For $\hat{\mu}_1(s)$ and $\tilde{\mu}_1(p)$, we abandon the simulation run when the denominator is smaller than 0.01. In 10,000 repetitions with $N = 500$, about 1.67% of the runs are abandoned for $\tilde{\mu}_1(p)$, but no runs are abandoned for $\hat{\mu}_1(s)$.

Table 1 shows the results for continuous Y with $N = 500$, where the true effect is linear or quadratic in the single index $X'\pi$. The ratio estimators $\hat{\mu}_1$ and $\tilde{\mu}_1$ perform well for both linear and quadratic effects. $\hat{\mu}_1$ tends to be more biased than $\tilde{\mu}_1$ but has a smaller SD. Since the bias magnitude is considerably smaller, the difference in SD dominates that in the bias, and thus, $\hat{\mu}_1$ performs better than $\tilde{\mu}_1$. The highest performing estimators are $\hat{w}'_1\hat{\beta}$ and $\hat{m}'_1\tilde{\beta}$. The biases of $\hat{w}'_2\hat{\beta}$ and $\hat{m}'_2\tilde{\beta}$ are not large; however, their SDs are very high due to multi-collinearity problems among the regressors. Moreover, the corresponding asymptotic variance estimates grossly exaggerate the actual SDs. In terms of SD, $\hat{m}'_2\tilde{\beta}$ does better than $\hat{w}'_2\hat{\beta}$. The SDs of most of $\hat{\mu}_1$, $\tilde{\mu}_1$, $\hat{w}'_1\hat{\beta}$, and $\hat{m}'_1\tilde{\beta}$ match closely with the averaged asymptotic SDs, which demonstrates the correctness of their asymptotic variances. The CI coverage proportion is too small for $\hat{\mu}_1$ and $\tilde{\mu}_1$ and too large for $\hat{w}'_2\hat{\beta}$ and $\hat{m}'_2\tilde{\beta}$. Overall, the ranking in Table 1 can be summarized as follows, with ">" meaning "better than":

Table 1: Continuous Y: |Bias|, SD, Avg.Asy.SD, and CI coverage proportion ($N = 500$)

Effect	Linear	Quadratic	Effect	Linear	Quadratic
$\hat{\mu}_{1(0.5)}$	0.056 0.698 0.488 0.83	0.074 0.588 0.447 0.81	$\tilde{\mu}_{1(0.31)}$	0.018 0.735 0.730 0.89	0.006 0.781 0.765 0.89
$\hat{\mu}_{1(0)}$	0.044 0.364 0.369 0.96	0.067 0.370 0.368 0.96	$\tilde{\mu}_{1(0.5)}$	0.121 0.899 0.966 0.94	0.111 0.897 0.940 0.94
$\hat{\mu}_{1(0.5)}$	0.114 1.14 0.782 0.85	0.155 1.11 0.771 0.87	$\tilde{\mu}_{1(0.69)}$	0.019 1.93 2.88 0.84	0.077 1.86 2.76 0.84
$\hat{\mu}_{1\text{Avg}}$	0.071 0.73 0.55 0.88	0.099 0.69 0.53 0.88	$\tilde{\mu}_{1\text{Avg}}$	0.053 1.2 1.5 0.89	0.065 1.2 1.5 0.89
$\hat{w}'_1\hat{\beta}_{(0.5)}$	0.041 0.486 0.503 0.93	0.015 0.486 0.496 0.93	$\hat{w}'_2\hat{\beta}_{(0.5)}$	0.143 1.70 5.69 0.96	0.180 1.53 5.37 0.96
$\hat{w}'_1\hat{\beta}_{(0)}$	0.034 0.345 0.353 0.96	0.118 0.353 0.353 0.96	$\hat{w}'_2\hat{\beta}_{(0)}$	0.034 1.16 5.50 0.98	0.011 1.28 6.60 0.98
$\hat{w}'_1\hat{\beta}_{(0.5)}$	0.027 0.670 0.685 0.98	0.029 0.660 0.680 0.98	$\hat{w}'_2\hat{\beta}_{(0.5)}$	0.019 3.08 17.0 0.99	0.008 3.49 18.6 1.0
$\hat{w}'_1\hat{\beta}_{\text{Avg}}$	0.034 0.50 0.51 0.97	0.054 0.50 0.51 0.95	$\hat{w}'_2\hat{\beta}_{\text{Avg}}$	0.065 2.0 9.4 0.98	0.066 2.1 10 0.98
$\hat{m}'_1\tilde{\beta}_{(0.31)}$	0.053 0.488 0.502 0.95	0.028 0.490 0.497 0.93	$\hat{m}'_2\tilde{\beta}_{(0.31)}$	0.113 0.694 0.847 0.96	0.133 0.755 0.911 0.96
$\hat{m}'_1\tilde{\beta}_{(0.5)}$	0.032 0.346 0.354 0.96	0.116 0.354 0.353 0.96	$\hat{m}'_2\tilde{\beta}_{(0.5)}$	0.030 0.537 0.746 0.98	0.007 0.545 0.764 0.98
$\hat{m}'_1\tilde{\beta}_{(0.69)}$	0.010 0.676 0.689 0.98	0.045 0.668 0.682 0.97	$\hat{m}'_2\tilde{\beta}_{(0.69)}$	0.023 1.45 2.46 1.0	0.092 1.49 2.46 1.0
$\hat{m}'_1\tilde{\beta}_{\text{Avg}}$	0.032 0.50 0.52 0.97	0.063 0.50 0.51 0.95	$\hat{m}'_2\tilde{\beta}_{\text{Avg}}$	0.055 0.90 1.4 0.98	0.077 0.93 1.4 0.98

Avg.Asy.SD: average across 10,000 reps of the asymptotic SD formula in theorems; $\hat{\mu}_1$ and $\tilde{\mu}_1$: ratio estimates; $\hat{w}'_J\hat{\beta}$ & $\hat{m}'_J\tilde{\beta}$: power-approximation estimates with $J = 1, 2$; Avg: simple average across three evaluation points.

$$\hat{m}_1'\tilde{\beta} \approx \hat{w}_1'\hat{\beta} > \hat{\mu}_1 > \hat{m}_2'\tilde{\beta} > \tilde{\mu}_1 > \hat{w}_2'\hat{\beta}. \tag{17}$$

When N increases to 5,000 in Table 2, all estimators become stable, and the averaged asymptotic SDs are closely matched with the corresponding simulation SDs. $\hat{\mu}_1$ outperforms $\tilde{\mu}_1$ in terms of both the bias and SD. $\hat{w}_1'\hat{\beta}$ and $\hat{m}_1'\tilde{\beta}$ perform the best when the true effect is linear, but exhibit substantially large biases when the true effect is quadratic. The performances of $\hat{w}_2'\hat{\beta}$ and $\hat{m}_2'\tilde{\beta}$ are satisfactory, even though they were the lowest performing estimators in Table 1, with $N = 500$; $\hat{w}_2'\hat{\beta}$ and $\hat{m}_2'\tilde{\beta}$ exhibit the minimum bias under quadratic effects. $\hat{\mu}_1$ is comparable to $\hat{w}_2'\hat{\beta}$ and $\hat{m}_2'\tilde{\beta}$, and $\tilde{\mu}_1$ has the largest SDs. The CI coverage proportion is close to 95%, except when the bias is large for $\hat{w}_1'\hat{\beta}$ and $\hat{m}_1'\tilde{\beta}$ under quadratic effects. Overall, the ranking in Table 2 is

$$\hat{m}_2'\tilde{\beta} > \hat{\mu}_1 \approx \hat{w}_2'\hat{\beta} > \hat{m}_1'\tilde{\beta} \approx \hat{w}_1'\hat{\beta} > \tilde{\mu}_1. \tag{18}$$

The findings in Table 3 with binary Y and $N = 500$ are similar to those in Table 1 except for the CI coverage that is considerably lower for $\hat{\mu}_1$. The ranking (17) still holds for Table 3. When N increases to 5,000 in Table 4, the asymptotic variance formulas work well, and the large biases of $\hat{m}_1'\tilde{\beta}$ and $\hat{w}_1'\hat{\beta}$ (slightly larger than the biases in Table 3) and resulting low CI coverage are notable. Overall, the ranking in Table 4 can be expressed as

$$\hat{m}_2'\tilde{\beta} \approx \hat{w}_2'\hat{\beta} > \hat{\mu}_1 > \hat{m}_1'\tilde{\beta} \approx \hat{w}_1'\hat{\beta} > \tilde{\mu}_1. \tag{19}$$

Thus far, we examined the CI coverage at each evaluation point separately. For joint coverage across m evaluation points, the 95% joint coverage requires that the CI at each point has a higher confidence level. Solving $(1 - \alpha')^m = 1 - \alpha$ for α' with $\alpha = 0.05$ gives about $\alpha' = 0.05/m$, allowing the confidence band (CB) across the m points to capture the true effect curve in 95% of the trials. Figure 1 shows 50 CBs randomly selected from our 10, 000 simulation runs when Y is continuous, the effect is linear, and the sample size is 5,000, using $m = 20$ equally spaced evaluation points over $p \in [0.25, 0.75]$. For each estimator, only 1 ~ 3 CBs do not capture the true line, resulting in a joint coverage of 94 ~ 98%.

Overall, there exist trade-offs among the bias, SD, CI coverage, ease in implementation, and closeness of the asymptotic variance formula to the actual variance. We note that $\hat{\mu}_1$ performs reasonably well overall, whereas $\tilde{\mu}_1$ exhibits a low performance. However, considering the trade-offs, we recommend the use of $(\hat{w}_1'\hat{\beta}, \hat{m}_1'\tilde{\beta})$ for small samples and $(\hat{w}_2'\hat{\beta}, \hat{m}_2'\tilde{\beta})$ for large samples, which are particularly easy to implement with only probit and OLS.

Table 2: Continuous Y: |Bias|, SD, Avg.Asy.SD, and CI coverage proportion ($N = 5{,}000$)

Effect	Linear	Quadratic	Effect	Linear	Quadratic
$\hat{\mu}_{1(0.5)}$	0.007 0.165 0.152 0.91	0.019 0.165 0.154 0.90	$\tilde{\mu}_{1(0.31)}$	0.031 0.231 0.233 0.94	0.033 0.230 0.234 0.92
$\hat{\mu}_{1(0)}$	0.015 0.126 0.129 0.96	0.030 0.127 0.130 0.95	$\tilde{\mu}_{1(0.5)}$	0.019 0.246 0.257 0.95	0.011 0.244 0.259 0.95
$\hat{\mu}_{1(0.5)}$	0.026 0.265 0.244 0.91	0.037 0.259 0.240 0.91	$\tilde{\mu}_{1(0.69)}$	0.027 0.621 0.682 0.91	0.054 0.631 0.689 0.92
$\hat{\mu}_{1\text{Avg}}$	0.016 0.19 0.18 0.93	0.029 0.18 0.18 0.92	$\tilde{\mu}_{1\text{Avg}}$	0.026 0.37 0.39 0.93	0.032 0.37 0.39 0.93
$\hat{w}_1'\hat{\beta}_{(0.5)}$	0.007 0.142 0.142 0.95	0.013 0.138 0.139 0.85	$\hat{w}_2'\hat{\beta}_{(0.5)}$	0.011 0.145 0.147 0.95	0.010 0.145 0.149 0.95
$\hat{w}_1'\hat{\beta}_{(0)}$	0.003 0.101 0.101 0.95	0.081 0.101 0.102 0.86	$\hat{w}_2'\hat{\beta}_{(0)}$	0.005 0.141 0.150 0.95	0.003 0.132 0.142 0.95
$\hat{w}_1'\hat{\beta}_{(0.5)}$	0.001 0.201 0.199 0.96	0.075 0.197 0.198 0.83	$\hat{w}_2'\hat{\beta}_{(0.5)}$	0.018 0.361 0.362 0.97	0.001 0.312 0.329 0.98
$\hat{w}_1'\hat{\beta}_{\text{Avg}}$	0.004 0.15 0.15 0.95	0.056 0.15 0.15 0.85	$\hat{w}_2'\hat{\beta}_{\text{Avg}}$	0.011 0.22 0.22 0.96	0.005 0.20 0.21 0.96
$\hat{m}_1'\tilde{\beta}_{(0.31)}$	0.018 0.142 0.142 0.95	0.002 0.138 0.140 0.84	$\hat{m}_2'\tilde{\beta}_{(0.31)}$	0.022 0.141 0.142 0.96	0.033 0.145 0.146 0.95
$\hat{m}_1'\tilde{\beta}_{(0.5)}$	0.003 0.101 0.101 0.95	0.081 0.101 0.102 0.85	$\hat{m}_2'\tilde{\beta}_{(0.5)}$	0.001 0.129 0.133 0.95	0.005 0.128 0.134 0.96
$\hat{m}_1'\tilde{\beta}_{(0.69)}$	0.012 0.201 0.199 0.95	0.085 0.196 0.197 0.83	$\hat{m}_2'\tilde{\beta}_{(0.69)}$	0.012 0.283 0.281 0.97	0.009 0.278 0.276 0.98
$\hat{m}_1'\tilde{\beta}_{\text{Avg}}$	0.011 0.15 0.15 0.95	0.056 0.15 0.15 0.84	$\hat{m}_2'\tilde{\beta}_{\text{Avg}}$	0.012 0.18 0.19 0.96	0.015 0.18 0.19 0.96

Table 3: Binary Y: |Bias|, SD, Avg.Asy.SD, and CI coverage proportion (N = 500)

Effect	Linear	Quadratic	Effect	Linear	Quadratic
$\hat{\mu}_{1(0.5)}$	0.029 0.373 0.274 0.83	0.020 0.359 0.269 0.82	$\tilde{\mu}_{1(0.31)}$	0.051 0.578 0.627 0.92	0.019 0.552 0.602 0.92
$\hat{\mu}_{1(0)}$	0.024 0.149 0.145 0.95	0.034 0.149 0.147 0.94	$\tilde{\mu}_{1(0.5)}$	0.133 0.564 0.626 0.97	0.158 0.589 0.649 0.97
$\hat{\mu}_{1(0.5)}$	0.026 0.214 0.176 0.68	0.027 0.215 0.179 0.71	$\tilde{\mu}_{1(0.69)}$	0.098 0.869 1.37 0.86	0.108 0.856 1.40 0.86
$\hat{\mu}_{1\text{Avg}}$	0.026 0.25 0.20 0.82	0.027 0.24 0.20 0.82	$\tilde{\mu}_{1\text{Avg}}$	0.094 0.67 0.87 0.92	0.095 0.67 0.89 0.91
$\hat{w}_1'\hat{\beta}_{(0.5)}$	0.013 0.225 0.227 0.93	0.028 0.226 0.226 0.90	$\hat{w}_2'\hat{\beta}_{(0.5)}$	0.091 1.50 3.96 0.96	0.052 0.621 2.54 0.94
$\hat{w}_1'\hat{\beta}_{(0)}$	0.059 0.121 0.124 0.94	0.083 0.123 0.126 0.91	$\hat{w}_2'\hat{\beta}_{(0)}$	0.001 0.481 2.63 0.97	0.013 0.450 2.98 0.97
$\hat{w}_1'\hat{\beta}_{(0.5)}$	0.030 0.184 0.195 0.98	0.049 0.182 0.198 0.97	$\hat{w}_2'\hat{\beta}_{(0.5)}$	0.007 1.17 8.37 0.99	0.027 1.11 8.90 0.99
$\hat{w}_1'\hat{\beta}_{\text{Avg}}$	0.034 0.18 0.18 0.95	0.053 0.18 0.18 0.93	$\hat{w}_2'\hat{\beta}_{\text{Avg}}$	0.033 1.1 5.0 0.97	0.031 0.73 4.8 0.97
$\hat{m}_1'\tilde{\beta}_{(0.31)}$	0.008 0.228 0.229 0.93	0.023 0.229 0.229 0.90	$\hat{m}_2'\tilde{\beta}_{(0.31)}$	0.047 0.377 0.382 0.96	0.038 0.350 0.369 0.95
$\hat{m}_1'\tilde{\beta}_{(0.5)}$	0.058 0.121 0.123 0.94	0.082 0.123 0.126 0.91	$\hat{m}_2'\tilde{\beta}_{(0.5)}$	0.004 0.192 0.250 0.98	0.010 0.190 0.242 0.98
$\hat{m}_1'\tilde{\beta}_{(0.69)}$	0.036 0.185 0.194 0.98	0.056 0.182 0.196 0.97	$\hat{m}_2'\tilde{\beta}_{(0.69)}$	0.005 0.314 0.542 0.99	0.003 0.300 0.526 0.99
$\hat{m}_1'\tilde{\beta}_{\text{Avg}}$	0.034 0.18 0.18 0.95	0.054 0.18 0.18 0.92	$\hat{m}_2'\tilde{\beta}_{\text{Avg}}$	0.019 0.30 0.39 0.97	0.017 0.28 0.38 0.97

Table 4: Binary Y: |Bias|, SD, Avg.Asy.SD, and CI coverage proportion (N = 5,000)

Effect	Linear	Quadratic	Effect	Linear	Quadratic
$\hat{\mu}_{1(0.5)}$	0.002 0.096 0.092 0.90	0.005 0.099 0.094 0.90	$\tilde{\mu}_{1(0.31)}$	0.004 0.170 0.190 0.96	0.001 0.176 0.193 0.96
$\hat{\mu}_{1(0)}$	0.009 0.053 0.053 0.94	0.011 0.053 0.053 0.93	$\tilde{\mu}_{1(0.5)}$	0.010 0.152 0.168 0.97	0.012 0.152 0.169 0.97
$\hat{\mu}_{1(0.5)}$	0.008 0.061 0.058 0.89	0.003 0.062 0.060 0.89	$\tilde{\mu}_{1(0.69)}$	0.056 0.313 0.351 0.95	0.051 0.310 0.351 0.95
$\hat{\mu}_{1\text{Avg}}$	0.006 0.070 0.068 0.91	0.006 0.072 0.069 0.91	$\tilde{\mu}_{1\text{Avg}}$	0.023 0.21 0.24 0.96	0.022 0.21 0.24 0.96
$\hat{w}_1'\hat{\beta}_{(0.5)}$	0.025 0.066 0.066 0.85	0.035 0.066 0.066 0.68	$\hat{w}_2'\hat{\beta}_{(0.5)}$	0.014 0.072 0.072 0.95	0.019 0.071 0.073 0.92
$\hat{w}_1'\hat{\beta}_{(0)}$	0.046 0.036 0.036 0.71	0.068 0.037 0.037 0.45	$\hat{w}_2'\hat{\beta}_{(0)}$	0.009 0.048 0.049 0.95	0.014 0.048 0.048 0.93
$\hat{w}_1'\hat{\beta}_{(0.5)}$	0.044 0.058 0.057 0.61	0.073 0.059 0.059 0.32	$\hat{w}_2'\hat{\beta}_{(0.5)}$	0.008 0.059 0.061 0.98	0.015 0.055 0.057 0.98
$\hat{w}_1'\hat{\beta}_{\text{Avg}}$	0.038 0.054 0.053 0.72	0.058 0.054 0.054 0.48	$\hat{w}_2'\hat{\beta}_{\text{Avg}}$	0.010 0.060 0.061 0.96	0.016 0.058 0.060 0.94
$\hat{m}_1'\tilde{\beta}_{(0.31)}$	0.021 0.067 0.067 0.84	0.031 0.066 0.067 0.66	$\hat{m}_2'\tilde{\beta}_{(0.31)}$	0.004 0.074 0.073 0.95	0.006 0.073 0.074 0.93
$\hat{m}_1'\tilde{\beta}_{(0.5)}$	0.046 0.036 0.036 0.70	0.068 0.037 0.037 0.44	$\hat{m}_2'\tilde{\beta}_{(0.5)}$	0.005 0.048 0.049 0.95	0.007 0.049 0.050 0.95
$\hat{m}_1'\tilde{\beta}_{(0.69)}$	0.047 0.057 0.056 0.59	0.077 0.057 0.058 0.30	$\hat{m}_2'\tilde{\beta}_{(0.69)}$	0.003 0.056 0.057 0.98	0.010 0.052 0.058 0.98
$\hat{m}_1'\tilde{\beta}_{\text{Avg}}$	0.038 0.053 0.053 0.71	0.058 0.053 0.054 0.47	$\hat{m}_2'\tilde{\beta}_{\text{Avg}}$	0.004 0.059 0.060 0.96	0.008 0.058 0.061 0.95

5 Empirical analysis

Our empirical analysis is for the effects of 401(k) retirement programs D on savings Y. Many studies have investigated whether contributions to tax-deferred retirement plans increase savings or simply crowd out other types of savings [10,28–32]. Since D is correlated with unobserved individual preferences for savings, Abadie [10] used the eligibility δ for 401(k) programs as an IV for D to overcome the endogeneity problem.

Because the eligibility for the programs is exogenously set, δ is unlikely to be correlated with the preferences for savings once we control for X such as the income. The IV exclusion restriction is plausible for δ, as δ is likely to affect savings only through the income. Since only the eligible persons can apply for a 401(k) account, monotonicity holds trivially, and the IV relevance condition is verified by the OLS of D on (δ, X).

We use the same data as those used by Poterba et al. [29] and Abadie [10], derived from the Survey of Income and Program Participation for 1991. The observation unit is a household, and the sample is restricted to households with at least one member employed. Table 5 presents the sample mean (SD) of the variables, where

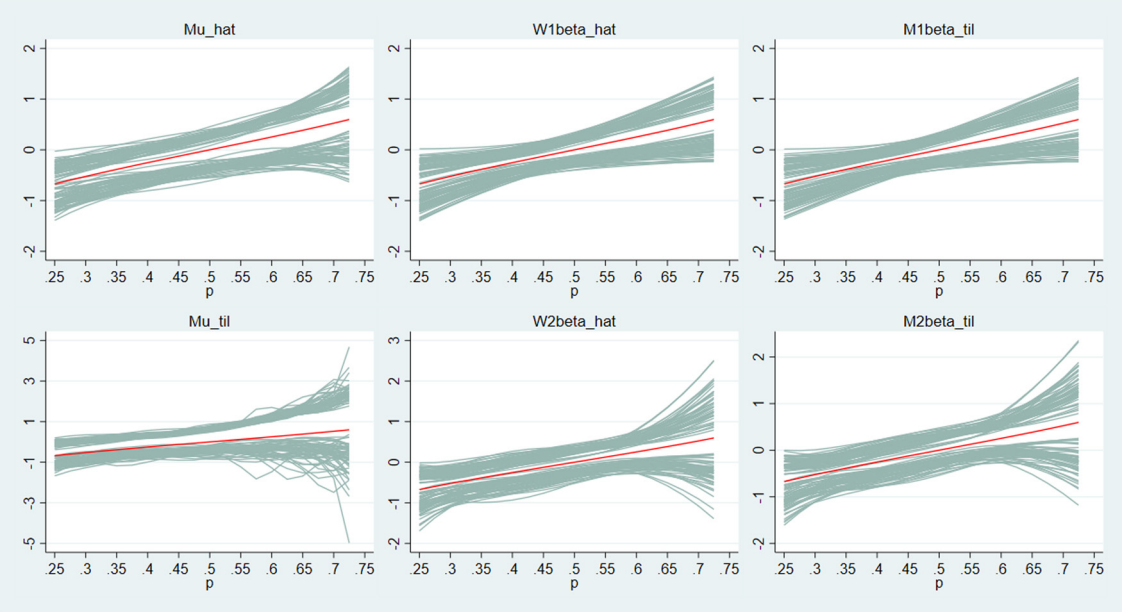

Figure 1: True effect curve and 50 CBs for continuous Y with linear effect and $N = 5{,}000$.

Y is the net financial assets in \$1,000. X consists of the household income in \$1,000 (Inc), age, marital status (Mar), and household size (Hsize). N_d is the group size for $D = d$. 39% of the households are eligible for the 401(k) programs ($\delta = 1$), and 72% = 100 × (0.28/0.39) of them have $D = 1$.

Before we examine the effect estimates in Table 6, we explain the steps implemented for facilitating the comparison of the estimators. Recall that $\tilde{\mu}_1$ and $\hat{M}_J'\tilde{\beta}$ are conditioned on λ_X, whereas $\hat{\mu}_1$ and $\hat{W}_J'\hat{\beta}$ are conditioned on the linear indices. Although (14) is used to make $\hat{W}_J'\hat{\beta}$ comparable to $\tilde{\mu}_1$ and $\hat{M}_J'\tilde{\beta}$, the same step cannot be implemented for $\hat{\mu}_1$ because the single index for $\hat{\mu}_1$ does not include an intercept. The only way to make $\hat{\mu}_1$ comparable to $(\tilde{\mu}_1, \hat{W}_J'\hat{\beta}, \hat{M}_J'\tilde{\beta})$ is to condition $\hat{\mu}_1$ on $\hat{\Lambda}(X'\hat{\alpha})$, instead of $X'\hat{\alpha}$. However, since $\hat{\Lambda}(X'\hat{\alpha})$ is $\sqrt{Nh}$-consistent, this modification requires the derivation of its asymptotic distribution anew, which is prohibitively complicated. Hence, we use bootstrapping (300 bootstrap repetitions) to do inference for $\hat{\mu}_1$ conditioned on $\hat{\Lambda}(X'\hat{\alpha})$. If a comparison with the other estimators is not required, conditioning $\hat{\mu}_1$ on $X'\hat{\alpha}$ is adequate.

For the estimation, first, both age and age^2 are used as regressors in the probit of δ on X for $\hat{M}_J'\tilde{\beta}$. Notably, the use of age^2 introduces certain difficulties for $\hat{\mu}_1$ because using a functionally dependent regressor in $\hat{\mu}_1$ requires a complex identification assumption, as mentioned in relation to Theorem 2. Second, for $K(\cdot)$ of $\hat{\mu}_1$ and $\tilde{\mu}_1$, we use the simple $N(0,1)$ kernel. The bounded quartic kernel is not used because the estimators crash if there are no observations around several evaluation points. According to our simulation study, the choice of the two kernels does not considerably affect the results. Third, the bandwidth h is set as $h_0 \in [1,2]$ in $h = h_0 SD(\hat{\lambda}_X)N^{-1/5}$ through "eye-balling"; the bandwidth for $\hat{\Lambda}(\cdot)$ is the rule-of-thumb bandwidth h_Λ =

Table 5: Mean (SD) of variables ($N = 9{,}275$; $N_1 = 2{,}562$, $N_0 = 6{,}713$)

	Pooled	$D = 1$	$D = 0$		**Pooled**	$D = 1$	$D = 0$
Y	19.1 (64)	38.5 (79.3)	11.7 (055.3)	Inc	39.3 (24.1)	49.8 (26.8)	35.2 (21.6)
				Age	41.1 (10.3)	41.5 (9.65)	41.9 (10.0)
D	0.28	1	0	Mar	0.63	0.69	0.60
δ	0.39	1	0.16	Hsize	2.89 (1.53)	2.92 (1.47)	2.87 (1.55)

Y: Net financial assets in \$1,000; Inc: family income; Mar: marital status; Hsize: family size.

Table 6: Complier effects for financial assets in \$1,000 (N = 9,275)

$p = \lambda_X$:	0.2 (SE)	0.3 (SE)	0.4 (SE)	0.5 (SE)	0.6 (SE)
$\hat{\mu}_1$	4.37 (4.16)	4.56 (4.94)	9.11 (6.61)	*15.9 (6.60)	+21.7 (15.0)
$\tilde{\mu}_1$	5.77 (4.12)	+4.69 (2.85)	**16.9 (3.04)	**29.4 (5.52)	*33.2 (13.1)
$\hat{W}_1'\hat{\beta}$	6.38 (11.3)	+8.24 (4.80)	**9.82 (1.70)	+11.3 (6.42)	12.8 (11.6)
$\hat{W}_2'\hat{\beta}$	7.69 (30.7)	*5.64 (2.30)	+14.7 (8.86)	**21.3 (6.70)	*26.1 (10.6)
$\hat{M}_1'\tilde{\beta}$	5.76 (8.51)	*7.69 (3.92)	**9.62 (1.67)	*11.6 (5.88)	13.5 (10.5)
$\hat{M}_2'\tilde{\beta}$	5.11 (18.7)	**5.96 (1.90)	*14.7 (6.44)	**21.0 (4.40)	*25.0 (10.9)

**, *, +: 1%, 5%, 10% significance levels; $\hat{\mu}_1$ & $\tilde{\mu}_1$: 1st & 2nd ratio estimators; $\hat{W}_J'\hat{\beta}$ & $\hat{M}_J'\tilde{\beta}$: power-approximation estimators conditioned on $X'\hat{\theta}$ & $\Phi(X'\hat{\theta})$.

$SD\{\Lambda(X'\hat{\alpha})\}N^{-1/5}$. Fourth, for $\hat{W}_J'\hat{\beta}$ and $\hat{M}_J'\tilde{\beta}$, we set $J = 1, 2$, and thus, six estimators are compared, as in our simulation study. The comparison is performed over $\lambda_X \in [0.2, 0.6]$, which contains most $\hat{\lambda}_X$ values.

For $\hat{\Lambda}(X'\hat{\alpha})$ in $\hat{\mu}_1$, Inc, Age, and Mar are the significant variables, whereas Inc, Age, and Hsize are the significant variables for $\Phi(X'\hat{\theta})$, with the standard error (SE) in $(\cdot)$.

	Inc	Age	Age^2	Mar	Hsize
$\hat{\alpha}$ in $\hat{\Lambda}(X'\hat{\alpha})$	** 0.046 (0.011)	** 0.055 (0.018)	*−0.002 (0.001)	**−0.20 (0.055)	−0.034 (0.028)
$\hat{\theta}$ in $\Phi(X'\hat{\theta})$	** 0.014 (0.001)	** 0.039 (0.005)	**−0.001 (0.0001)	0.019 (0.037)	**−0.034 (0.011)

In the results of $(\hat{\alpha}, \hat{\theta})$, the intercept is omitted, and ** and * denote the 1% and 5% significance levels, respectively. To make $\hat{\alpha}$ and $\hat{\theta}$ comparable, we normalize $\hat{\alpha}$ with $|\hat{\alpha}_{\text{hsize}}|$ and then multiply $\hat{\alpha}$ by $|\hat{\theta}_{\text{hsize}}|$, as explained at the end of Section 2.2. Income and age appear to be the two main variables driving the variation in IS.

In Table 6, all estimators show increasing effects of D across $p \in [0.2, 0.6]$, which become significant for $p \geq 0.3$, except for $\hat{\mu}_1$. Recalling $\bar{\delta} \simeq 0.4$, the effect on CPs at $p = 0.4$ is 9–17, and the effect based on $\hat{W}_2'\hat{\beta}$ and $\hat{M}_2'\tilde{\beta}$ is 14.7 (\$14,700). Recall that $\hat{W}_2'\hat{\beta}$ and $\hat{M}_2'\tilde{\beta}$ achieve the highest performance in our simulation study with large samples.

Table 6 shows no significant difference between $\hat{\mu}_1$ and $\tilde{\mu}_1$ at $p = 0.2, 0.3$; however, the estimates become much different for $p \geq 0.4$. For a given J, $\hat{W}_J'\hat{\beta}$ and $\hat{M}_J'\tilde{\beta}$ are similar, but $(\hat{W}_1'\hat{\beta}, \hat{M}_1'\tilde{\beta})$ differ much from $(\hat{W}_2'\hat{\beta}, \hat{M}_2'\tilde{\beta})$. With $J = 1$, the effect increases gradually from about 6 at $p = 0.2$ to about 13 at $p = 0.6$, but it becomes insignificant as p increases further. With $J = 2$, the effect increases dramatically from $-8 \sim -5$ to $25 \sim 26$ and is significant even at $p = 0.6$. As p increases, $\hat{W}_2'\hat{\beta}$ and $\hat{M}_2'\tilde{\beta}$ deviate from $\tilde{\mu}_1$ that uses the same IS.

Figure 2 shows $\hat{\mu}_1, \tilde{\mu}_1, \hat{W}_2'\hat{\beta}$, and $\hat{M}_2'\tilde{\beta}$ over 90% of the support points of IS. We omit $\hat{W}_1'\hat{\beta}$ and $\hat{M}_1'\tilde{\beta}$, as $J = 2$ is preferable over $J = 1$ in large samples. In Figure 2, $\hat{\mu}_1$, $\hat{W}_2'\hat{\beta}$, and $\hat{M}_2'\tilde{\beta}$ show more or less monotonically increasing effects as p increases, whereas $\tilde{\mu}_1$ is quadratic: $\tilde{\mu}_1$ increases up to 36 at $p = 0.58$ and then decreases sharply to 6 at $p = 0.67$. However, this decline is implausible, which might be due to the "boundary problem" in kernel estimators. Overall, recalling the BMI example in the introduction, we can state that *the* 401(*k*) *plan effect on the savings is positive for those with IS greater than approximately* 0.3.

It is puzzling why $\tilde{\mu}_1$ yields considerably different results from $\hat{M}_2'\tilde{\beta}$, even though both estimators use $\lambda_X = \Phi(X'\theta)$. For this, note that kernel estimation is a local nonparametric method, whereas power approximation is global. Hence, the former approach is superior when the focus is on an evaluation point, whereas the latter approach is superior when the focus is on the shape of a curve. In the former, observations far from the evaluation point are irrelevant, whereas all observations are relevant for all evaluation points in the latter.

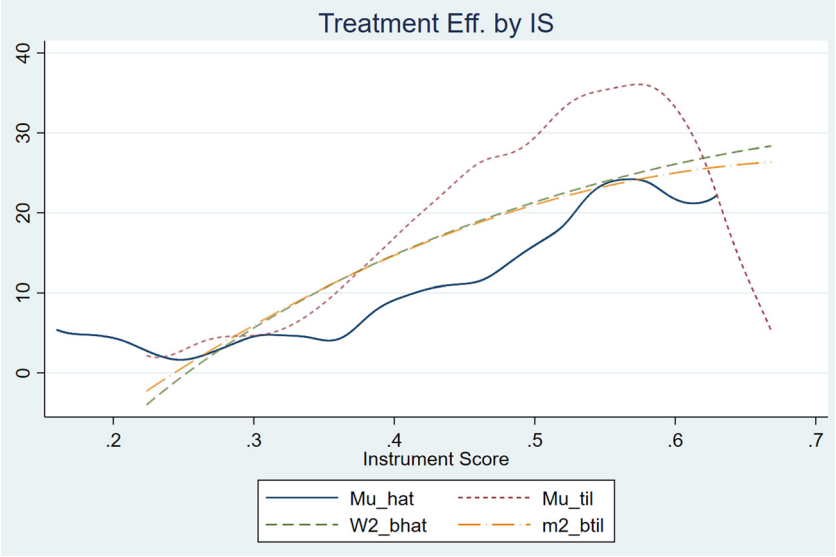

Figure 2: $E(Y^1 - Y^0 |\ \text{CP}, \lambda_x = p)$ versus p.

This aspect leads to adverse results for the kernel method when it does not work well at a chosen evaluation point. For example, if the monotonicity condition is violated at the point, then the kernel estimate would be unsatisfactory. In contrast, because the power approximation method is global, the effect estimates at other points can help mitigate the poor local estimate. Owing to this property, as well as the more significant estimates in Table 6, $\hat{M}_2'\tilde{\beta}$ and $\hat{W}_2'\hat{\beta}$ are the preferred estimators in this empirical analysis.

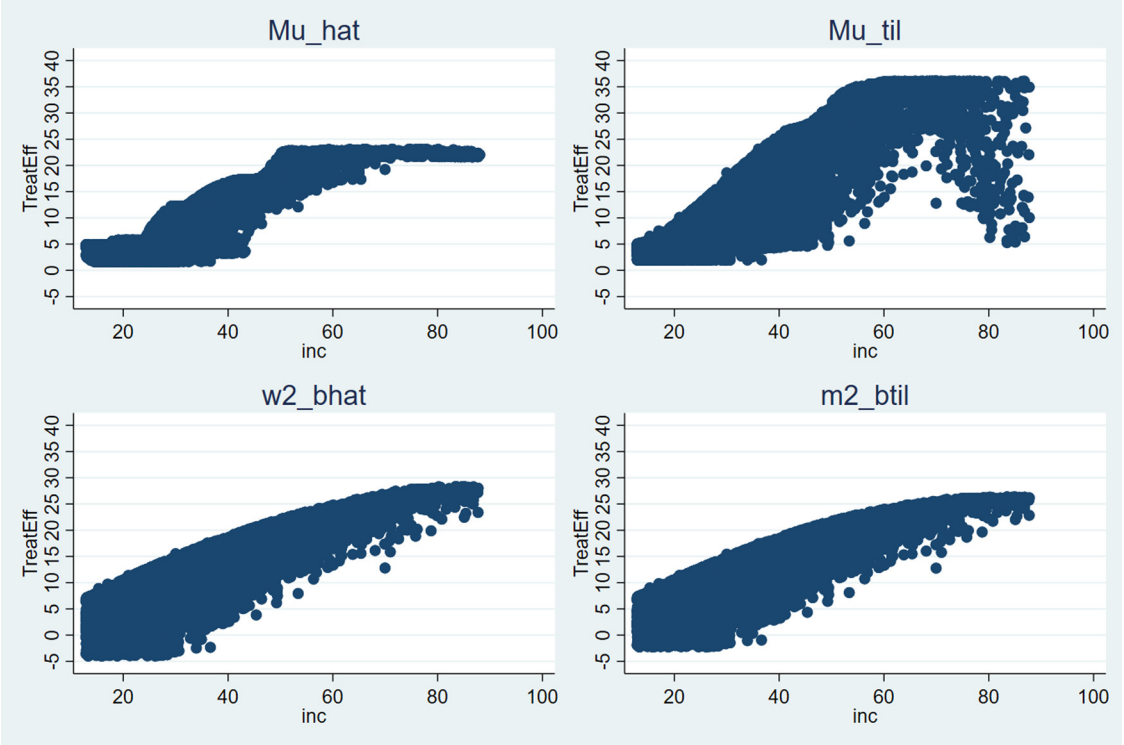

Figure 3: $E(Y^1 - Y^0 |$ CP, income).

The benefits of the global nonparametric approach versus the local approach can be further highlighted. The inverted matrix for $\hat{M}_2'\tilde{\beta}$ is essentially $E\{E(D^1 - D^0|\lambda_X)(1 - \lambda_X)\lambda_X MM'\}$ as (12) shows, whereas the denominator of $\tilde{\mu}_1$ is essentially $E(D^1 - D^0|\lambda_X) = P(\text{CP}|\lambda_X)$ as (3) shows. Consequently, if $P(\text{CP}|\lambda_X)$ approaches zero at some λ_X, $\tilde{\mu}_1$ would suffer, whereas $\hat{M}_2'\tilde{\beta}$ would suffer less because it uses the averaged version of $P(\text{CP}|\lambda_X)$.

To examine how informative $E(Y^1 - Y^0|\text{CP}, \lambda_X)$ is, Figure 3 shows the plots of $E(Y^1 - Y^0|\text{CP}, \text{Inc}_i)$ versus Inc_i for all $i = 1, \ldots, N$. The first row is for $\hat{\mu}_1$ and $\tilde{\mu}_1$, and the second row is for $\hat{W}_2'\hat{\beta}$ and $\hat{M}_2'\tilde{\beta}$. Figure 3, which is based on the income, reveals increasing patterns similar to those shown in Figure 2. This trend suggests that the λ_X-heterogeneous effect in Figure 2 is mostly driven by the income. We also tried other covariates, but could not find any informative pattern. In practice, our estimators can be tried initially, and if effect heterogeneity is found, then a more extensive analysis with X can be performed.

The effect of a covariate can be found using the graph for $W'\hat{\beta}$ (Figure 2). For example, when X_2 increases by one unit at $X'\hat{\theta} = s$ (i.e., at $W'\hat{\beta} = (1, s, \ldots, s^J)\hat{\beta}$), $W'\hat{\beta}$ increases by $\hat{\theta}_2$ times $\{0, 1, 2(X'\hat{\theta}), \ldots, J(X'\hat{\theta})^{J-1}\}\hat{\beta}$, and this effect can be found in Figure 2 by moving to the right from $(1, s, \ldots, s^J)\hat{\beta}$ by $\hat{\theta}_2 \cdot \{0, 1, 2(X'\hat{\theta}), \ldots, J(X'\hat{\theta})^{J-1}\}\hat{\beta}$ and then comparing the values of the graph at the two positions.

Specifically, when the income increases by one unit ($1,000) at the mean of $\bar{X}'\hat{\theta} \approx -0.287$, the IS $\Phi(\bar{X}'\hat{\theta})$ hardly changes from 0.39, as the change is only from 0.387 to 0.392. Then, because $\hat{\theta}_{inc} = 0.014$, $W_2'\hat{\beta}$ increases by $0.014 \times \{0, 1, -0.574\}\hat{\beta} \simeq 0.53$. Hence, the effect of the increase in the income of $1,000 on Y is an increase of $530 for the CPs with the IS of approximately 0.39. This value is considerably smaller than the "naive income effect" observed in Table 5: the mean group difference of Y divided by the mean group difference of the income is $(38.5 - 11.7)/(49.8 - 35.2) = 1.84$.

6 Conclusion

For a binary treatment D, an outcome Y, and covariates X, denoting the potential responses as (Y^0, Y^1) for $D = 0, 1$, the treatment effect heterogeneity (i.e., $E(Y^1 - Y^0|X)$ not being a constant) is a rule rather than an exception. However, when X is high-dimensional, the nonparametric estimation of $E(Y^1 - Y^0|X)$ runs into the well-known dimension problem. In the PS literature, $E(Y^1 - Y^0|PS)$ has been estimated instead to overcome the dimension problem under the D-exogeneity "$D \perp\!\!\!\perp (Y^0, Y^1)|X$."

When D is endogenous/confounded, however, PS matching and other estimators requiring D-exogeneity cannot be used, and at least a binary instrument (e.g., δ) is needed to overcome the problem. In this article, defining the potential treatments (D^0, D^1) for $\delta = 0, 1$ and "CP" as those with $(D^0 = 0, D^1 = 1)$, we showed that the role played by PS for exogenous D can be played by the "IS" $\lambda_X \equiv E(\delta|X)$ for endogenous D with a non-randomized δ. The IS becomes PS for exogenous D because $\delta = D$.

The dimension reduction achieved by PS for exogenous D cannot be realized by an arbitrary function of X. Similarly, the dimension reduction achieved by IS for endogenous D cannot be realized by an arbitrary function of X. By identifying and estimating $E(Y^1 - Y^0|\text{CP}, \lambda_X)$ conditioned only on the scalar λ_X, we capture the effect heterogeneity while avoiding the dimension problem. The heterogeneity captured by λ_X is minimal, in the sense that λ_X is the "coarsest balancing score" ($\delta \perp\!\!\!\perp X|\lambda_X$). *Since the endogenous D becomes exogenous for CPs because $D = \delta$, and since it is likely that $D = 1$ when $Y^1 - Y^0$ is positive, the IS with $D = \delta$ can capture the effect heterogeneity well.*

We proposed three estimators for $E(Y^1 - Y^0|\text{CP}, \lambda_X)$, motivated by two critical RF equations that are linear in either D or δ, even though no explicit linearity assumption is imposed. The equations and *our three estimators hold for any form of Y (continuous, binary, count, ...),* as long as $Y^1 - Y^0$ makes sense. The three estimators require progressively more restrictive assumptions to enhance their applicability.

The first estimator is a kernel nonparametric estimator based on a single index estimator for λ_X that enables the use of an unknown link function. The second estimator is the same as the first estimator, except for the assumption of the probit link for λ_X. The third estimator is an IVE formulated after approximating $E(Y^1 - Y^0|\text{CP}, \lambda_X)$ with a power function of λ_X. Since we take the power approximation to be exact, the third

estimator is $\sqrt{N}$-consistent, whereas the first two estimators converge at a slower rate. Among the three estimators, we recommend the third estimator because it is easy to implement with only OLS and probit, numerically stable, and not subject to the "excessively small denominator problem" inherent in the first two ratio estimators.

We presented an empirical illustration for the effects of 401(k) retirement programs (D) on savings (Y) with the eligibility δ for the programs as the IV. Our main finding is that the households with the IS greater than approximately 0.3 would increase their savings due to D, whereas those with the IS smaller than 0.3 would not. This kind of finding could be useful when D is a drug whose administration is self-selected by individuals (and is thus endogenous). If there exists an education/encouragement δ based on X for the possible benefits of the drug, we can find an analogous cutoff to prepare an easy-to-follow guideline that the drug would benefit those with the IS greater than the cutoff.

Acknowledgment: The authors are grateful to two anonymous reviewers for their helpful comments.

Funding information: Jin-young Choi's research has been supported by the Hankuk University of Foreign Studies Research Fund of 2023. Myoung-jae Lee's research has been supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2020R1A2C1A01007786).

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: The authors state no conflict of interest.

Ethical approval: The conducted research is not related to either human or animals use.

Data availability statement: The dataset analyzed during this study is from "Introductory Econometrics: A Modern Approach, 7e" by Jeffrey M. Wooldridge. The dataset named "401ksubs" is available in the textbook e-learning resource repository, https://www.cengage.com/cgi-wadsworth/course_products_wp.pl?fid=M20b&product_isbn_issn=9781111531041.

References

[1] Rosenbaum PR. Observational studies. 2nd ed. Springer; 2002. doi: https://doi.org/10.1007/978-1-4757-3692-2.
[2] Lee MJ. Micro-econometrics for policy, program, and treatment effects. Oxford University Press; 2005. doi: https://doi.org/10.1093/0199267693.001.0001.
[3] Lee MJ. Matching, regression discontinuity, difference in differences, and beyond. Oxford University Press; 2016. doi: https://doi.org/10.1093/acprof:oso/9780190258733.001.0001.
[4] Pearl J. Causality. 2nd ed. Cambridge University Press; 2009. doi: https://doi.org/10.1017/CBO9780511803161.
[5] Imbens GW, Rubin DB. Causal inference for statistics, social, and biomedical sciences: an introduction. Cambridge University Press; 2015. doi: https://doi.org/10.1017/CBO9781139025751.
[6] Abadie A, Cattaneo MD. Econometric methods for program evaluation. Annu Rev Econ. 2018;10:465–503. doi: https://doi.org/10.1146/annurev-economics-080217-053402.
[7] Imbens GW, Angrist J. Identification and estimation of local average treatment effects. Econometrica. 1994;62(2):467–76. doi: https://doi.org/10.2307/2951620.
[8] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Amer Stat Assoc. 1996;91(434):444–55. doi: https://doi.org/10.2307/2291629.
[9] Frölich M. Nonparametric IV estimation of local average treatment effects with covariates. J Econom. 2007;139(1):35–75. doi: https://doi.org/10.1016/j.jeconom.2006.06.004.
[10] Abadie A. Semiparametric instrumental variable estimation of treatment response models. J Econom. 2003;113(2):231–63. doi: https://doi.org/10.1016/S0304-4076(02)00201-4.

[11] Tan Z. Regression and weighting methods for causal inference using instrumental variables. J Amer Stat Assoc. 2006;101(476):1607–18. doi: https://doi.org/10.1198/016214505000001366.
[12] Ogburn EL, Rotnitzky A, Robins JM. Doubly robust estimation of the local average treatment effect curve. J R Stat Soc (Ser B). 2015;77(2):373–96. doi: https://doi.org/10.1111/rssb.12078.
[13] Imai K, Ratkovic M. Estimating treatment effect heterogeneity in randomized program evaluation. Ann Appl Stat. 2013;7(1):443–70. doi: https://doi.org/10.1214/12-AOAS593.
[14] Künzel SR, Sekhon JS, Bickel PJ, Yu B. Metalearners for estimating heterogeneous treatment effects using machine learning. Proc Nat Acad Sci. 2019;116(10):4156–65. doi: https://doi.org/10.1073/pnas.1804597116.
[15] Athey S, Imbens GW. Machine learning methods that economists should know about. Ann Rev Econom. 2019;11:685–725. doi: https://doi.org/10.1146/annurev-economics-080217-053433.
[16] Hirano K, Porter JR. Asymptotics for statistical treatment rules. Econometrica. 2009;77(5):1683–701. doi: https://doi.org/10.3982/ECTA6630.
[17] Dudik M, Erhan D, Langford J, Li L. Doubly robust policy evaluation and optimization. Stat Sci. 2014;29(4):485–511. doi: https://doi.org/10.1214/14-STS500.
[18] Athey S, Imbens GW. Recursive partitioning for heterogeneous causal effects. Proc Nat Acad Sci. 2016;113(27):7353–60. doi: https://doi.org/10.1073/pnas.1510489113.
[19] Choi JY, Lee G, Lee MJ. Endogenous treatment effect for any response conditional on control propensity score. Stat Probability Lett. 2023;196:109747. doi: https://doi.org/10.1016/j.spl.2022.109747.
[20] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55. doi: https://doi.org/10.1093/biomet/70.1.41.
[21] Swanson SA, Hernán MA. Think globally, act globally: an epidemiologist's perspective on instrumental variable estimation. Stat Sci. 2014;29(3):371–4. doi: https://doi.org/10.1214/14-STS491.
[22] Mogstad M, Torgovitsky A. Identification and extrapolation of causal effects with instrumental variables. Ann Rev Econ. 2018;10:577–613. doi: https://doi.org/10.1146/annurev-economics-101617-041813.
[23] Ichimura H. Semiparametric least squares (SLS) and weighted SLS estimation of single index models. J Econom. 1993;58(1–2):71–120. doi: https://doi.org/10.1016/0304-4076(93)90114-K.
[24] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. 2nd ed. Springer; 2009. doi: https://doi.org/10.1007/978-0-387-84858-7.
[25] Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al. Double/debiased machine learning for treatment and structural parameters. Econom J. 2018;21(1):C1–68. doi: https://doi.org/10.1111/ectj.12097.
[26] Lee MJ. Simple least squares estimator for treatment effects using propensity score residuals. Biometrika. 2018;105(1):149–64. doi: https://doi.org/10.1093/biomet/asx062.
[27] Lee MJ. Instrument residual estimator for any response variable with endogenous binary treatment. J R Stat Soc (Ser B). 2021;83(3):612–35. doi: https://doi.org/10.1111/rssb.12442.
[28] Gale WG, Scholz JK. IRAs and household saving. Amer Econ Rev. 1994;84(5):1233–60.
[29] Poterba JM, Venti SF, Wise DA. Do 401(k) contributions crowd out other personal saving?. J Public Econ. 1995;58(1):1–32. doi: https://doi.org/10.1016/0047-2727(94)01462-W.
[30] Madrian BC, Shea DF. The Power of suggestion: inertia in 401(k) participation and savings behavior. Quarter J Econ. 2001;116(4):1149–87. doi: https://doi.org/10.1162/003355301753265543.
[31] Benjamin DJ. Does 401(k) eligibility increase saving? Evidence from propensity score subclassification. J Public Econ. 2003;87(5–6):1259–90. doi: https://doi.org/10.1016/S0047-2727(01)00167-0.
[32] Chetty R, Friedman JN, Leth-Petersen S, Nielsen TH, Olsen T. Active vs. passive decisions and crowd-out in retirement savings accounts: evidence from Denmark. Quarter J Econ. 2014;129(3):1141–219. doi: https://doi.org/10.1093/qje/qju013.

Appendix

A Proofs

A.1 Proof for (5) and balancing score property of IS

Proof. Under $\delta \perp\!\!\!\perp (D^0, D^1, Y^0, Y^1)|X$, observe

$$\begin{aligned} E(\delta|D^0, D^1, Y^0, Y^1, \lambda_X) &= E\{E(\delta|D^0, D^1, Y^0, Y^1, X) \mid D^0, D^1, Y^0, Y^1, \lambda_X\} \\ &= E\{E(\delta|X) \mid D^0, D^1, Y^0, Y^1, \lambda_X\} = E(\lambda_X|D^0, D^1, Y^0, Y^1, \lambda_X) = \lambda_X = E(\delta|\lambda_X). \end{aligned}$$

The first and last expressions prove that δ is mean-independent of (D^0, D^1, Y^0, Y^1) given λ_X, but since δ is binary, the mean-independence is the same as (4)(i) to prove (5).

IS $\lambda_X \equiv E(\delta|X)$ is a balancing score because, for any fixed t,

$$E(\delta 1[X \le t]|\lambda_X) = E\{E(\delta 1[X \le t]|X) \mid \lambda_X\} = E\{E(\delta|X)1[X \le t] \mid \lambda_X\} = \lambda_X P(X \le t|\lambda_X) = E(\delta|\lambda_X)\cdot P(X \le t|\lambda_X),$$

take $E(\cdot|\lambda_X)$ on $\lambda_X \equiv E(\delta|X)$ to see $\lambda_X = E(\delta|\lambda_X)$. Dividing the first and last expressions with $E(\delta|\lambda_X)$, for which $0 < \lambda_X < 1$ is assumed, gives $P(X \le t|\delta = 1, \lambda_X) = P(X \le t|\lambda_X)$. Since the class of sets $\{X \le t, t\in(\text{real space})\}$ is a "probability determining class," the distribution of X is the same across the $\delta = 0, 1$ groups given λ_X. □

A.2 Proof for Theorem 1 regarding (2)

Proof. With $D = (D^1 - D^0)\delta + D^0$, due to (4)(i) and (ii), we have

$$\begin{aligned} E(D|\delta, \lambda_X) &= E(D^1 - D^0|\delta, \lambda_X)\delta + E(D^0|\delta, \lambda_X) \\ &= E(D^1 - D^0|\lambda_X)\delta + E(D^0|\lambda_X) \\ &= P(D^1 = 1, D^0 = 0|\lambda_X)\delta + E(D^0|\lambda_X) \\ &= P(\text{CP}|\lambda_X)\delta + E(D^0|\lambda_X). \end{aligned} \tag{A1}$$

The first and last expressions reveal that $E(D^0|\lambda_X)$ is the λ_X-conditional intercept, and $P(\text{CP}|\lambda_X)$ is the λ_X-conditional slope of δ to show the effect of δ on D given λ_X.

Take $E(\cdot|\delta, \lambda_X)$ on $Y = (Y^1 - Y^0)\{(D^1 - D^0)\delta + D^0\} + Y^0$: due to (4)(i) and (ii),

$$\begin{aligned} E(Y|\delta, \lambda_X) &= E[(Y^1 - Y^0)\{(D^1 - D^0)\delta + D^0\}|\delta, \lambda_X] + E(Y^0|\delta, \lambda_X) \\ &= E\{(Y^1 - Y^0)(D^1 - D^0)|\delta, \lambda_X\}\delta + E\{(Y^1 - Y^0)D^0 + Y^0|\delta, \lambda_X\} \\ &= E\{(Y^1 - Y^0)(D^1 - D^0)|\lambda_X\}\delta + E\{(Y^1 - Y^0)D^0 + Y^0|\lambda_X\} \\ &= E(Y^1 - Y^0|\text{CP}, \lambda_X)P(\text{CP}|\lambda_X)\delta + E\{(Y^1 - Y^0)D^0 + Y^0|\lambda_X\} \\ &= \mu_1(\lambda_X)P(\text{CP}|\lambda_X)\delta + E\{(Y^1 - Y^0)D^0 + Y^0|\lambda_X\}. \end{aligned} \tag{A2}$$

Define $U_2 \equiv Y - E(Y|\delta, \lambda_X) \Leftrightarrow E(Y|\delta, \lambda_X) = Y - U_2 \Rightarrow E(U_2|\delta, \lambda_X) = 0$ to rewrite the first and last expressions of (A2) as (2). □

A.3 Proof for Theorem 2 (first ratio estimator)

Proof. Under the assumptions in Theorem 2, $\hat{\alpha}$ is $\sqrt{N}$-consistent [23, Theorem 5.2]. Hence, using $\hat{\alpha}$ in $\hat{\mu}_1(s)$ is as good as using α, and thus, the following deals with the asymptotic distribution for $\hat{\mu}_1(s)$ with α known.

(1) Linearization

The linearization of $\hat{\mu}_1(s)$ to be used is

$$\begin{aligned}\frac{\hat{b}(s)}{\hat{a}(s)} - \frac{b(s)}{a(s)} &= -\frac{b(s)}{a(s)^2}\{\hat{a}(s) - a(s)\} + \frac{1}{a(s)}\{\hat{b}(s) - b(s)\} + o_p\left(\frac{1}{\sqrt{Nh}}\right) \\ &= -\frac{b(s)}{a(s)^2}[\hat{a}_1(s) - a_1(s) - \{\hat{a}_0(s) - a_0(s)\}] + \frac{1}{a(s)}[\{\hat{b}_1(s) - b_1(s)\} - \{\hat{b}_0(s) - b_0(s)\}] \\ &\quad + o_p\left(\frac{1}{\sqrt{Nh}}\right);\end{aligned} \tag{A3}$$

suppress $o_p\{(Nh)^{-1/2}\}$ henceforth. Apply this linearization also to $\hat{a}_j(s)$ and $\hat{b}_j(s)$:

$$\hat{a}_j(s) - a_j(s) = \frac{\hat{a}_{jd}(s)}{\hat{c}_j(s)} - \frac{a_{jd}(s)}{c_j(s)} = -\frac{a_{jd}(s)}{c_j(s)^2}\{\hat{c}_j(s) - c_j(s)\} + \frac{1}{c_j(s)}\{\hat{a}_{jd}(s) - a_{jd}(s)\},$$
$$\hat{b}_j(s) - b_j(s) = \frac{\hat{b}_{jy}(s)}{\hat{c}_j(s)} - \frac{b_{jy}(s)}{c_j(s)} = -\frac{b_{jy}(s)}{c_j(s)^2}\{\hat{c}_j(s) - c_j(s)\} + \frac{1}{c_j(s)}\{\hat{b}_{jy}(s) - b_{jy}(s)\},$$

where

$$\hat{a}_{jd}(s) \equiv \frac{1}{N_j h}\sum_{i\in G_j} K\{(S_i - s)/h\}D_i, \quad a_{jd}(s) \equiv E(D|s, \delta = j)f_{Sj}(s),$$
$$\hat{b}_{jy}(s) \equiv \frac{1}{N_j h}\sum_{i\in G_j} K\{(S_i - s)/h\}Y_i, \quad b_{jy}(s) \equiv E(Y|s, \delta = j)f_{Sj}(s),$$
$$\hat{c}_j(s) \equiv \frac{1}{N_j h}\sum_{i\in G_j} K\{(S_i - s)/h\}, \quad c_j(s) \equiv f_{Sj}(s).$$

Substitute the linearizations for $\hat{a}_j(s) - a_j(s)$ and $\hat{b}_j(s) - b_j(s)$ into (A3) to obtain

$$\begin{aligned}\frac{\hat{b}(s)}{\hat{a}(s)} - \frac{b(s)}{a(s)} &= -\frac{b(s)}{a(s)^2}\left[-\frac{a_{1d}(s)}{c_1(s)^2}\{\hat{c}_1(s) - c_1(s)\} + \frac{1}{c_1(s)}\{\hat{a}_{1d}(s) - a_{1d}(s)\} + \frac{a_{0d}(s)}{c_0(s)^2}\{\hat{c}_0(s) - c_0(s)\}\right. \\ &\quad \left. - \frac{1}{c_0(s)}\{\hat{a}_{0d}(s) - a_{0d}(s)\}\right] + \frac{1}{a(s)}\left[-\frac{b_{1y}(s)}{c_1(s)^2}\{\hat{c}_1(s) - c_1(s)\} + \frac{1}{c_1(s)}\{\hat{b}_{1y}(s) - b_{1y}(s)\}\right. \\ &\quad \left. + \frac{b_{0y}(s)}{c_0(s)^2}\{\hat{c}_0(s) - c_0(s)\} - \frac{1}{c_0(s)}\{\hat{b}_{0y}(s) - b_{0y}(s)\}\right].\end{aligned}$$

Rewrite the preceding display by collecting terms, and then multiply by $\sqrt{Nh}$:

$$\begin{aligned}&\sqrt{Nh}\left\{\frac{\hat{b}(s)}{\hat{a}(s)} - \frac{b(s)}{a(s)}\right\} \\ &= -\frac{b(s)}{a(s)^2}\frac{1}{c_1(s)}\frac{\sqrt{N_1 h}\{\hat{a}_{1d}(s) - a_{1d}(s)\}}{\sqrt{\pi_1}} + \frac{b(s)}{a(s)^2}\frac{1}{c_0(s)}\frac{\sqrt{N_0 h}\{\hat{a}_{0d}(s) - a_{0d}(s)\}}{\sqrt{\pi_0}} \\ &\quad + \frac{1}{a(s)}\frac{1}{c_1(s)}\frac{\sqrt{N_1 h}\{\hat{b}_{1y}(s) - b_{1y}(s)\}}{\sqrt{\pi_1}} - \frac{1}{a(s)}\frac{1}{c_0(s)}\frac{\sqrt{N_0 h}\{\hat{b}_{0y}(s) - b_{0y}(s)\}}{\sqrt{\pi_0}} \\ &\quad + \left\{\frac{b(s)}{a(s)^2}\frac{a_{1d}(s)}{c_1(s)^2} - \frac{1}{a(s)}\frac{b_{1y}(s)}{c_1(s)^2}\right\}\frac{\sqrt{N_1 h}\{\hat{c}_1(s) - c_1(s)\}}{\sqrt{\pi_1}} \\ &\quad - \left\{\frac{b(s)}{a(s)^2}\frac{a_{0d}(s)}{c_0(s)^2} - \frac{1}{a(s)}\frac{b_{0y}(s)}{c_0(s)^2}\right\}\frac{\sqrt{N_0 h}\{\hat{c}_0(s) - c_0(s)\}}{\sqrt{\pi_0}}.\end{aligned} \tag{A4}$$

The right side has six terms, which gives six asymptotic variances. Also, the three terms sharing the same subscript 1 are correlated with each other to give three covariances, and the three terms sharing the same subscript 0 also give three covariances. Hence, the asymptotic variance of $\sqrt{Nh}\{\hat{\mu}_1(s) - \mu_1(s)\}$ consists of 12 terms. We present some preliminaries next, and then turn to the 12 terms.

(2) Preliminaries with $\kappa \equiv \int K(t)^2 \mathrm{d}t$

With $\int K(t)t\mathrm{d}t = 0$ and the twice continuous differentiability of $E(Y|s,\delta)$ and $f_{Sj}(s)$,

$$E\left\{\frac{1}{h}K\left(\frac{S-s}{h}\right)Y|\delta = j\right\} = \int \frac{1}{h}K\left(\frac{t-s}{h}\right)E(Y|t,\delta = j)f_{Sj}(t)\mathrm{d}t$$
$$= \int K(v)E(Y|s+hv,\delta = j)f_{Sj}(s+hv)\mathrm{d}v = E(Y|s,\delta = j)f_{Sj}(s) + O(h^2);$$
$$E\left\{\frac{1}{h}K\left(\frac{S-s}{h}\right)^2 Y^2|\delta = j\right\} = \int \frac{1}{h}K\left(\frac{t-s}{h}\right)^2 E(Y^2|t,\delta = j)f_{Sj}(t)\mathrm{d}t$$
$$= \int K(v)^2 E(Y^2|s+hv,\delta = j)f_{Sj}(s+hv)dv = E(Y^2|s,\delta = j)f_{Sj}(s)\kappa + O(h^2).$$

Analogous expressions hold when Y is replaced by D or YD. Observe ($E^2(\cdot) \equiv \{E(\cdot)\}^2$)

$$\frac{1}{N_j}\sum_{i\in G_j}\frac{1}{h}K\left(\frac{S_i-s}{h}\right)Y_i - E\left\{\frac{1}{h}K\left(\frac{S_i-s}{h}\right)Y|\delta = j\right\} = \frac{1}{N_j}\sum_{i\in G_j}\left[\frac{1}{h}K\left(\frac{S_i-s}{h}\right)Y_i - E\left\{\frac{1}{h}K\left(\frac{S_i-s}{h}\right)Y|\delta = j\right\}\right];$$

$$E\left(\left[\frac{1}{h}K\left(\frac{S-s}{h}\right)Y - E\left\{\frac{1}{h}K\left(\frac{S-s}{h}\right)Y|\delta = j\right\}\right]^2|\delta = j\right)$$
$$= E\left[\frac{1}{h^2}K\left(\frac{S-s}{h}\right)^2 Y^2|\delta = j\right] - E^2\left\{\frac{1}{h}K\left(\frac{S-s}{h}\right)Y|\delta = j\right\}$$
$$= h^{-1}E(Y^2|s,\delta = j)f_{Sj}(s)\cdot\kappa + O(h) - \{E(Y|s,\delta = j)f_{Sj}(s) + O(h^2)\}^2.$$

Hence, invoking the Lindeberg CLT for triangular arrays,

$$\sqrt{N_j h}\left[\sum_{i\in G_j}\frac{1}{h}K\left(\frac{S_i-s}{h}\right)Y_i - E\left\{\frac{1}{h}K\left(\frac{S_i-s}{h}\right)Y|\delta = j\right\}\right] \to^d N\{0, E(Y^2|s,\delta = j)f_{Sj}(s)\kappa\}.$$

An analogous result holds with Y replaced by D, and the asymptotic covariance between the two normalized sums with Y and D is

$$E\left\{\frac{1}{h}K\left(\frac{S-s}{h}\right)^2 YD|\delta = j\right\} = E(YD|s,\delta = j)f_{Sj}(s)\kappa + O(h^2). \tag{A5}$$

(3) Variances

Because of

$$\sqrt{N_j h}\{\hat{a}_{jd}(s) - a_{jd}(s)\} \to^d N\{0, E(D|s,\delta = j)f_{Sj}(s)\kappa\},$$

the variance of the first and second terms in (A4) is, respectively,

$$\begin{aligned}\frac{b(s)^2}{a(s)^4}\frac{1}{c_1(s)^2\pi_1}E(D|s,\delta = 1)f_{S1}(s)\kappa &= \frac{b(s)^2}{a(s)^4}\frac{1}{f_{S1}(s)\pi_1}E(D|s,\delta = 1)\kappa \equiv V_1,\\ \frac{b(s)^2}{a(s)^4}\frac{1}{c_0(s)^2\pi_0}E(D|s,\delta = 0)f_{S0}(s)\kappa &= \frac{b(s)^2}{a(s)^4}\frac{1}{f_{S0}(s)\pi_0}E(D|s,\delta = 0)\kappa \equiv V_2.\end{aligned} \tag{A6}$$

Because of

$$\sqrt{N_j h}\{\hat{b}_{jy}(s) - b_{jy}(s)\} \to^d N\{0, E(Y^2|s,\delta = j)f_{Sj}(s)\kappa\},$$

the variance of the third and fourth terms in (A4) is, respectively,

$$\frac{1}{a(s)^2}\frac{1}{c_1(s)^2\pi_1}E(Y^2|s,\delta = 1)f_{S1}(s)\kappa = \frac{1}{a(s)^2}\frac{1}{f_{S1}(s)\pi_1}E(Y^2|s,\delta = 1)\kappa \equiv V_3,$$
$$\frac{1}{a(s)^2}\frac{1}{c_0(s)^2\pi_0}E(Y^2|s,\delta = 0)f_{S0}(s)\kappa = \frac{1}{a(s)^2}\frac{1}{f_{S0}(s)\pi_0}E(Y^2|s,\delta = 0)\kappa \equiv V_4.$$

Because of $\sqrt{N_j h}\{\hat{c}_j(s) - c_j(s)\} \longrightarrow^d N\{0, f_{Sj}(s)\kappa\}$, the variance of the fifth and sixth terms is, respectively,

$$\left\{\frac{b(s)}{a(s)^2}\frac{a_{1d}(s)}{c_1(s)^2} - \frac{1}{a(s)}\frac{b_{1y}(s)}{c_1(s)^2}\right\}^2 \frac{1}{\pi_1} f_{S1}(s)\kappa \equiv V_5, \quad \left\{\frac{b(s)}{a(s)^2}\frac{a_{0d}(s)}{c_0(s)^2} - \frac{1}{a(s)}\frac{b_{0y}(s)}{c_0(s)^2}\right\}^2 \frac{1}{\pi_0} f_{S0}(s)\kappa \equiv V_6.$$

(4) Covariances

For the covariance between the first and third terms, we need the expected value of the product between $\sqrt{N_1 h}\{\hat{a}_{1d}(s) - a_{1d}(s)\}$ and $\sqrt{N_1 h}\{\hat{b}_{1y}(s) - b_{1y}(s)\}$, where only the overlapping N_1 terms are non-zero whose expected value is $E(YD|s, \delta = 1) f_{S1}(s)\kappa$ as (A5) shows. Hence, the desired covariance is

$$-\frac{b(s)}{a(s)^2}\frac{1}{c_1(s)} \cdot \frac{1}{a(s)}\frac{1}{c_1(s)\pi_1} E(YD|s, \delta = 1) f_{S1}(s)\kappa = -\frac{b(s)}{a(s)^3}\frac{1}{f_{S1}(s)\pi_1} E(YD|s, \delta = 1)\kappa \equiv C_1.$$

For the covariance between the first and fifth terms, we need the expected value of the product between $\sqrt{N_1 h}\{\hat{a}_{1d}(s) - a_{1d}(s)\}$ and $\sqrt{N_1 h}\{\hat{c}_1(s) - c_1(s)\}$, where only the overlapping N_1 terms are non-zero whose expected value is $E(D|s, \delta = 1) f_{S1}(s)\kappa$. Hence, the desired covariance is

$$-\frac{b(s)}{a(s)^2}\left\{\frac{b(s)}{a(s)^2}\frac{a_{1d}(s)}{c_1(s)^2} - \frac{1}{a(s)}\frac{b_{1y}(s)}{c_1(s)^2}\right\}\frac{1}{\pi_1} E(D|s, \delta = 1)\kappa \equiv C_2.$$

Analogously, for the covariance between the third and fifth terms, we need the expected value of the product between $\sqrt{N_1 h}\{\hat{b}_{1y}(s) - b_{1y}(s)\}$ and $\sqrt{N_1 h}\{\hat{c}_1(s) - c_1(s)\}$, where only the overlapping N_1 terms are non-zero whose expected value is $E(Y|s, \delta = 1) f_{S1}(s)\kappa$. Hence, the desired covariance is

$$\frac{1}{a(s)} \cdot \left\{\frac{b(s)}{a(s)^2}\frac{a_{1d}(s)}{c_1(s)^2} - \frac{1}{a(s)}\frac{b_{1y}(s)}{c_1(s)^2}\right\}\frac{1}{\pi_1} E(Y|s, \delta = 1)\kappa \equiv C_3.$$

As for the three covariance terms involving the three terms with the subscript 0, the terms analogous to C_1, C_2, and C_3 are, respectively,

$$\begin{aligned} -\frac{b(s)}{a(s)^3}\frac{1}{f_{S0}(s)\pi_0} E(YD|s, \delta = 0)\kappa &\equiv C_4, \\ -\frac{b(s)}{a(s)^2} \cdot \left\{\frac{b(s)}{a(s)^2}\frac{a_{0d}(s)}{c_0(s)^2} - \frac{1}{a(s)}\frac{b_{0y}(s)}{c_0(s)^2}\right\}\frac{1}{\pi_0} E(D|s, \delta = 0)\kappa &\equiv C_5, \\ \frac{1}{a(s)} \cdot \left\{\frac{b(s)}{a(s)^2}\frac{a_{0d}(s)}{c_0(s)^2} - \frac{1}{a(s)}\frac{b_{0y}(s)}{c_0(s)^2}\right\}\frac{1}{\pi_0} E(Y|s, \delta = 0)\kappa &\equiv C_6. \end{aligned} \tag{A7}$$

□

A.4 Proof for (9)

Proof. Take $E(\cdot|\lambda_X)$ on $\lambda_X \equiv E(\delta|X)$ to see $\lambda_X = E(\delta|\lambda_X)$, which implies $p = E(\delta|\lambda_X = p)$. Now, rewrite the ratio in (3) at $\lambda_X = p$ as:

$$\begin{aligned} &\frac{E(Y\delta|\lambda_X = p)/E(\delta|\lambda_X = p) - [E\{Y(1-\delta)|\lambda_X = p\}/\{1 - E(\delta|\lambda_X = p)\}]}{E(D\delta|\lambda_X = p)/E(\delta|\lambda_X = p) - [E\{D(1-\delta)|\lambda_X = p\}/\{1 - E(\delta|\lambda_X = p)\}]} \\ &= \frac{E[Y(\delta/p) - Y\{(1-\delta)/(1-p)\} \quad |\lambda_X = p]}{E[D(\delta/p) - D\{(1-\delta)/(1-p)\} \quad |\lambda_X = p]} \\ &= \frac{E[\{Y\delta(1-p) - Y(1-\delta)p\}/\{p(1-p)\}|\lambda_X = p]}{E[\{D\delta(1-p) - D(1-\delta)p\}/\{p(1-p)\}|\lambda_X = p]} \\ &= \frac{E\{Y\delta(1-p) - Y(1-\delta)p|\lambda_X = p\}}{E\{D\delta(1-p) - D(1-\delta)p \quad |\lambda_X = p\}} = \frac{E\{Y(\delta - p)|\lambda_X = p\}}{E\{D(\delta - p)|\lambda_X = p\}}. \end{aligned}$$

□

A.5 Proof for Theorem 3 (second ratio estimator)

Proof. Define $\hat{a}(p)$ and $\hat{b}(p)$ so that $\tilde{\mu}_1(p) = \hat{b}(p)/\hat{a}(p)$:

$$\hat{a}(p) \equiv \frac{1}{Nh}\sum_i K\left(\frac{P_i - p}{h}\right)A_i \quad \text{and} \quad \hat{b}(p) \equiv \frac{1}{Nh}\sum_i K\left(\frac{P_i - p}{h}\right)B_i.$$

Also, define their estimands: $a(p) \equiv E(A|p)f_\lambda(p)$ and $b(p) \equiv E(B|p)f_\lambda(p)$. The linearization analogous to (A3) holds to give

$$\sqrt{Nh}\left\{\frac{\hat{b}(p)}{\hat{a}(p)} - \frac{b(p)}{a(p)}\right\} = -\frac{b(p)}{a(p)^2}\sqrt{Nh}\{\hat{a}(p) - a(p)\} + \frac{1}{a(p)}\sqrt{Nh}\{\hat{b}(p) - b(p)\} + o_p(1).$$

The asymptotic variance from the two terms on the right side is

$$\begin{aligned}
&\left\{\frac{b(p)^2}{a(p)^4}E(A^2|p) + \frac{1}{a(p)^2}E(B^2|p) - 2\frac{b(p)}{a(p)^3}E(AB|p)\right\}\cdot \kappa f_\lambda(p) \\
&= \left\{\frac{E^2(B|p)f_\lambda(p)^2}{E^4(A|p)f_\lambda(p)^4}E(A^2|p) + \frac{1}{E^2(A|p)f_\lambda(p)^2}E(B^2|p) - 2\frac{E(B|p)f_\lambda(p)}{E^3(A|p)f_\lambda(p)^3}E(AB|p)\right\}\kappa f_\lambda(p) \\
&= \left\{\frac{E^2(B|p)}{E^4(A|p)f_\lambda(p)}E(A^2|p) + \frac{1}{E^2(A|p)f_\lambda(p)}E(B^2|p) - 2\frac{E(B|p)}{E^3(A|p)f_\lambda(p)}E(AB|p)\right\}\kappa \\
&= \frac{\kappa}{f_\lambda(p)E^4(A|p)}\{E(A^2|p)E^2(B|p) + E(B^2|p)E^2(A|p) - 2E(A|p)E(B|p)E(AB|p)\}.
\end{aligned}$$

□

A.6 Proof for Theorem 4

Proof. With $\mu_0(\lambda_X)$ defined in (1), rewrite the $E(Y|\delta,\lambda_X)$ equation in (A2) as:

$$\begin{aligned}
E(Y|\delta,\lambda_X) &= \mu_1(\lambda_X)P(\text{CP}|\lambda_X)\delta + \mu_0(\lambda_X) + \mu_1(\lambda_X)E(D^0|\lambda_X) \\
&= \mu_1(\lambda_X)D + \mu_0(\lambda_X) + \mu_1(\lambda_X)\{P(\text{CP}|\lambda_X)\delta + E(D^0|\lambda_X) - D\}.
\end{aligned} \tag{A8}$$

Define $\zeta \equiv Y - E(Y|\delta,\lambda_X) \Leftrightarrow E(Y|\delta,\lambda_X) = Y - \zeta \Rightarrow E(\zeta|\delta,\lambda_X) = 0$ to have

$$\begin{aligned}
Y &= \mu_1(\lambda_X)D + \mu_0(\lambda_X) + \mu_1(\lambda_X)\{P(\text{CP}|\lambda_X)\delta + E(D^0|\lambda_X) - D\} + \zeta \\
&= \mu_1(\lambda_X)D + \mu_0(\lambda_X) + U_1, \quad U_1 \equiv \mu_1(\lambda_X)\{P(\text{CP}|\lambda_X)\delta + E(D^0|\lambda_X) - D\} + \zeta.
\end{aligned}$$

$E(U_1|\delta,\lambda_X) = 0$, as $E(\zeta|\delta,\lambda_X) = 0$ and $E(D|\delta,\lambda_X) = P(\text{CP}|\lambda_X)\delta + E(D^0|\lambda_X)$ in (A1).

Define $(D^0 = 1, D^1 = 1)$ as "always takers," and $(D^0 = 0, D^1 = 0)$ as "never takers"; (4)(ii) rules out "defiers" $(D^0 = 1, D^1 = 0)$. With $D = (1-\delta)D^0 + D^1\delta$, we have

$$\begin{aligned}
\text{Cov}(\delta, D|\lambda_X) &= E(\delta D|\lambda_X) - \lambda_X E(D|\lambda_X) \\
&= E(\delta D^1|\lambda_X) - \lambda_X E\{(1-\delta)D^0 + D^1\delta|\lambda_X\} \\
&= E(\delta D^1|\lambda_X)(1-\lambda_X) - E\{(1-\delta)D^0|\lambda_X\}\lambda_X \\
&= P(D^1 = 1|\delta = 1,\lambda_X)\lambda_X(1-\lambda_X) - P(D^0 = 1|\delta = 0,\lambda_X)(1-\lambda_X)\lambda_X \\
&= \{P(\text{always taker, complier}|\lambda_X) - P(\text{always taker, defier}|\lambda_X)\}(1-\lambda_X)\lambda_X \\
&= P(\text{complier}|\lambda_X)(1-\lambda_X)\lambda_X = E(D^1 - D^0|\lambda_X)(1-\lambda_X)\lambda_X > 0 \quad \text{(defiers ruled out).}
\end{aligned}$$

Using the first and last expressions gives

$$E\{(\delta - \lambda_X)DMM'\} = E\{\text{Cov}(\delta, D|\lambda_X)MM'\} = E\{E(D^1 - D^0|\lambda_X)(1-\lambda_X)\lambda_X MM'\}.$$

□

A.7 Proof for Theorem 5

Proof. Let θ be the true value in the parameter space A_θ whose generic element is a. Define

$$W(\theta) \equiv \{1, (X'\theta), (X'\theta)^2, \ldots, (X'\theta)^J\}' \text{ so that } \frac{\partial W(\theta)}{\partial a'} = \nabla W \cdot X'.$$

The IV for $DW(\hat{\theta})$ is $\hat{\varepsilon}W(\hat{\theta})$, and the IVE $\hat{\beta}$ satisfies

$$\frac{1}{\sqrt{N}}\sum_i m(\hat{\theta}, \hat{\beta}, \hat{\gamma}) = 0, \quad m(a, b, g) = \{Y - W(a)'g - DW(a)'b\}\{\delta - \Phi(X'a)\}W(a),$$

where a is for θ, b is for β, and g is for γ. Taylor-expand the moment condition (times $\sqrt{N}$) around β to obtain, for some $\hat{\beta}^* \in (\hat{\beta}, \beta)$,

$$0 = \frac{1}{\sqrt{N}}\sum_i m(\hat{\theta}, \beta, \hat{\gamma}) + \frac{1}{N}\sum_i \frac{\partial m(\hat{\theta}, \hat{\beta}^*, \hat{\gamma})}{\partial b'}\sqrt{N}(\hat{\beta} - \beta). \tag{A9}$$

Solve this for $\sqrt{N}(\hat{\beta} - \beta)$ and then further Taylor-expand $m(\hat{\theta}, \beta, \hat{\gamma})$ around $m(\theta, \beta, \gamma)$:

$$\begin{aligned}\sqrt{N}(\hat{\beta} - \beta) = -E^{-1}\left\{\frac{\partial m(\theta, \beta, \gamma)}{\partial b'}\right\}\Bigg[\frac{1}{\sqrt{N}}\sum_i m(\theta, \beta, \gamma) + E\{\frac{\partial m(\theta, \beta, \gamma)}{\partial a'}\}\sqrt{N}(\hat{\theta} - \theta) \\ + E\{\frac{\partial m(\theta, \beta, \gamma)}{\partial g'}\}\sqrt{N}(\hat{\gamma} - \gamma)\Bigg] + o_p(1) \to^d N(0, \Omega_1),\end{aligned} \tag{A10}$$

where

$$\begin{aligned}\Omega_1 &= E^{-1}\left\{\frac{\partial m(\theta, \beta, \gamma)}{\partial b'}\right\} \cdot E(\eta_1\eta_1') \cdot E^{-1}\{\frac{\partial m(\theta, \beta, \gamma)}{\partial b}\}, \\ \eta_1 &\equiv m(\theta, \beta, \gamma) + E\left\{\frac{\partial m(\theta, \beta, \gamma)}{\partial a'}\right\}\eta_{\hat{\theta}} + E\left\{\frac{\partial m(\theta, \beta, \gamma)}{\partial g'}\right\}\eta_{\hat{\gamma}},\end{aligned}$$

and $(\eta_{\hat{\theta}}, \eta_{\hat{\gamma}})$ are the influence functions for $(\hat{\theta}, \hat{\gamma})$.

As $m(a, b, g) = \{Y - W(a)'g - DW(a)'b\}\{\delta - \Phi(X'a)\}W(a)$ and $\partial W(\theta)/\partial a' = \nabla W X'$,

$$\begin{aligned}\frac{\partial m(\theta, \beta, \gamma)}{\partial a'} &= -(\nabla W'\gamma + D\nabla W'\beta)\varepsilon WX' - V\phi(X'\theta)WX' + V\varepsilon\nabla WX', \\ \frac{\partial m(\theta, \beta, \gamma)}{\partial b'} &= -\varepsilon DWW', \quad \frac{\partial m(\theta, \beta, \gamma)}{\partial g'} = -\varepsilon WW'.\end{aligned}$$

$E\{\partial m(\theta, \beta, \gamma)/\partial g'\} = 0$ due to $E(\varepsilon|X) = E(\delta - \lambda_X|X) = 0$. As for $E\{\partial m(\theta, \beta, \gamma)/\partial a'\}$, we have $E(\nabla W'\gamma\varepsilon WX') = 0$. Hence, the asymptotic variance is

$$E^{-1}(\varepsilon DWW')E(\eta_1\eta_1')E^{-1}(\varepsilon DWW'), \quad \eta_1 \equiv V\varepsilon W - E\{D\nabla W'\beta\varepsilon WX' + V\phi(X'\theta)WX' - V\varepsilon\nabla WX'\}\eta_{\hat{\theta}}. \square$$

A.8 Proof for Theorem 6

Proof. The proof for Theorem 6 is almost the same as that for Theorem 5. Define

$$M(\theta) \equiv \{1, \Phi(X'\theta), \Phi(X'\theta)^2, \ldots, \Phi(X'\theta)^J\}' \text{ so that } \frac{\partial M(\theta)}{\partial a'} = \nabla M \cdot X'.$$

The IV for $D\hat{M}$ is $\hat{\varepsilon}\hat{M}$, and the IVE $\tilde{\beta}$ satisfies

$$\frac{1}{\sqrt{N}}\sum_i m(\hat{\theta},\tilde{\beta},\tilde{\gamma}) = 0, \quad m(a,b,g) = \{Y - M(a)'g - DM(a)'b\}\{\delta - \Phi(X'a)\}M(a).$$

(A9) and (A10) hold with $(\hat{\beta},\hat{\gamma})$ and (Ω_1,η_1) replaced by $(\tilde{\beta},\tilde{\gamma})$ and

$$\Omega_2 = E^{-1}\left\{\frac{\partial m(\theta,\beta,\gamma)}{\partial b'}\right\}\cdot E(\eta_2\eta_2')\cdot E^{-1}\left\{\frac{\partial m(\theta,\beta,\gamma)}{\partial b}\right\},$$
$$\eta_2 \equiv m(\theta,\beta,\gamma) + E\left\{\frac{\partial m(\theta,\beta,\gamma)}{\partial a'}\right\}\eta_{\hat{\theta}} + E\left\{\frac{\partial m(\theta,\beta,\gamma)}{\partial g'}\right\}\eta_{\tilde{\gamma}},$$

and $\eta_{\tilde{\gamma}}$ is the influence function for $\tilde{\gamma}$.

With $m(a,b,g) = \{Y - M(a)'g - DM(a)'b\}\{\delta - \Phi(X'a)\}M(a)$,

$$\frac{\partial m(\theta,\beta,\gamma)}{\partial a'} = -(\nabla M'\gamma + D\nabla M'\beta)\varepsilon MX' - \Gamma\phi(X'\theta)MX' + \Gamma\varepsilon\nabla MX',$$
$$\frac{\partial m(\theta,\beta,\gamma)}{\partial b'} = -\varepsilon DMM', \quad \frac{\partial m(\theta,\beta,\gamma)}{\partial g'} = -\varepsilon MM'.$$

$E\{\partial m(\theta,\beta,\gamma)/\partial g'\} = 0$ and $E(\nabla M'\gamma\varepsilon MX') = 0$, and the asymptotic variance is

$$E^{-1}(\varepsilon DMM')E(\eta_2\eta_2')E^{-1}(\varepsilon DMM'), \quad \eta_2 \equiv \Gamma\varepsilon M - E\{D\nabla M'\beta\varepsilon MX' + \Gamma\phi(X'\theta)MX' - \Gamma\varepsilon\nabla MX'\}\eta_{\hat{\theta}}.\square$$