Causal Interpretation of Least Square Estimand under Model- and Design-Based Specification

Fangzhou Yu
Department of Economics, UNSW

November 18, 2024

Abstract

This paper studies the least square estimand of partially linear regressions with a binary or continuous treatment in the presence of heterogeneous treatment effects. We show that the least square estimand generally identify weighted average effects. However, the weights convey different causal interpretation under model-based specification (i.e. the regression equation leverages a correct model of potential outcomes) and design-based specification (i.e. the regression equation leverages a correct model of treatment assignment). Moreover, the design-based estimand is equivalent to overlap-weighted Inverse Propensity Weighting (IPW) under standard causal assumptions. We also show that the least square estimand is optimal in the class of weighted average effects, in the sense of achieving the nonparametric efficiency bound when the outcome is homoskedastic. We illustrate our results through an application of LaLonde (1986) study on the effect of NSW training program on earnings.

1 Introduction

Recently, a collection of influential papers have shown that ordinary least squares (OLS) and instrumental variable (IV) regressions may not identify a convex average of the treatment effects. Negative weights arise in the two-way fixed effects (TWFE) regressions of Difference-in-Differences (DiD) when treatment timing is staggered and treatment effects vary over time (e.g. De Chaisemartin & d'Haultfoeuille 2020, Goodman-Bacon 2021). In IV regression, negative weights are also observed when covariates are included (Small et al. 2017, Blandhol et al. 2022). Borusyak & Hull (2024) extend the discussion to the basic linear regression setting with a single treatment or instrument. They show that while design-based specifications (i.e. the regression equation leverages a correct model of treatment assignment) guarantee convex weights, negative weights generally exist when the specification is model-based (i.e. the regression equation leverages a correct model of potential outcomes) .

In this paper, we revisit the setting in Borusyak & Hull (2024) and extensively explore the least square estimand of linear and partially linear regressions. In contrast to the result in Borusyak & Hull (2024), we show that the concern of negative weights over model-based specifications is not as pronounced as previously suggested. For binary treatment with model-based specifications, the risk of negative weights is due to the projection of treatment on covariates exceeding unit value and does not generally exist in practice. Additionally, there exists common model-based specifications in empirical research which preclude negative weights. Motivated by this result, we further show that the concern of negative weights is eliminated when the treatment is continuous.

Building on the weighting scheme of least square estimands, we propose general causal interpretations under model- and design-based specifications. Specifically, for binary treatment, design-based specifications identify a weighted average treatment effect (WATE), where the weights are given by the conditional variance of the treatment. Moreover, under a stronger design-based specification, OLS is equivalent to overlap-weighted inverse propensity weighting (IPW) proposed by Li et al. (2018). On the other hand, model-based specifications identify a weighted average effect on the treated. For continuous treatment, model- and design-based specifications identify similar forms of weighted average derivative effects (WADE) with weights depending on the marginal distribution of the treatment given the covariates.

In the literature on causal interpretation of OLS and IV (e.g. Imbens & Angrist 1994, Angrist et al. 1996, Angrist 1998, Aronow & Samii 2016, Słoczyński 2022), the estimand is usually examined under binary treatment and assumptions that are often more restrictive than those typically used in empirical studies. For example, Angrist (1998) studies the case of saturated model; Słoczyński (2022) assumes that the potential outcomes are linear functions of the propensity score; Aronow & Samii (2016) consider a linear individual effect model for continuous treatment. We adopt weak assumptions on both model- and design-based specifications, as those considered by Goldsmith-Pinkham et al. (2022) and Borusyak & Hull (2024), and extend the identification results to continuous treatment. In particular, we show that the least square estimand with continuous treatment identifies a WADE, a causal parameter which extends the incremental treatment effect introduced by Rothenhäusler & Yu (2019).

The average derivative effect was originally proposed as the target parameter in index models (e.g. Powell et al. 1989, Härdle & Stoker 1989). Our identification result on WADE establishes a connection between index models and causal inference. Our primary focus is on the partially linear regression, a special case of index models, where we show that the treatment coefficient identifies a specific WADE. This result builds on the work of Powell et al. (1989), who derive the product-moment representation (Riesz Representation) of density-weighted average derivative of a general regression function. In the context of our study, the Riesz Representation is given by the least square estimand, and we trace back the weighting function. This general result incorporates binary treatment as a special case, offering a unified view of the weighting scheme in least square estimands.

We also show that the least square estimand is optimal within the class of WADE, in the sense of achieving the nonparametric efficiency bound when the outcome is homoskedastic. This finding extends the established efficiency bound of WATE with binary treatment, as shown by Crump et al. (2006) and Li et al. (2018), and encourages the use of partially linear regression for estimating treatment effects over methods targeting other parameters in the class of WADE.

The rest of the paper is organized as follows. Section 2 shows the simple result for OLS and IV with binary treatment. The general result for partially linear regression is represented in Section 3. Our findings are illustrated through an application of LaLonde (1986) study on the effect of NSW training program on earnings in Section 4. The proofs are collected in Appendix.

2 Motivating Example

2.1 OLS with Binary Treatment

In this section, we revisit the model- and design-based assumptions outlined in Borusyak & Hull (2024) and show our result in a simple setting of OLS. But instead of examining a general treatment, we focus on the case of binary treatments, which allows us to follow the potential outcome framework by Rubin (1974) and conveniently define the heterogeneous treatment effects. Formally, for each unit i in the sample, let $Y_i(0)$ denote the potential outcome for unit i under control and $Y_i(1)$ denote the potential outcome for unit i under treatment. Let $D_i = 1$ if unit i receives treatment and $D_i = 0$ otherwise. The observed outcome is

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0).$$

By rearranging the right-hand side, we have

$$Y_i = D_i(Y_i(1) - Y_i(0)) + Y_i(0) = \tau_i D_i + Y_i(0)$$

where $\tau_i = Y_i(1) - Y_i(0)$ is the individual treatment effect. In addition, we observe a vector of covariates X_i of each unit i. Consider OLS estimation of the following regression equation:

$$Y_i = \tau_{ols} D_i + X_i \beta + \epsilon_i. \tag{2.1}$$

Let $\tilde{D}_i$ be the residual of the population projection of D_i on X_i. By the Frisch-Waugh-Lovell theorem, OLS estimand can be written as

$$\tau_{ols} = \frac{\mathbb{E}[\tilde{D}_i Y_i]}{\mathbb{E}[\tilde{D}_i D_i]} = \frac{\mathbb{E}[\tau_i \tilde{D}_i D_i] + \mathbb{E}[\tilde{D}_i Y_i(0)]}{\mathbb{E}[\tilde{D}_i D_i]}. \tag{2.2}$$

Now we investigate the weighting scheme in τ_{ols} under model- and design-based specifications. Consider a model-based specification that leverages a linear model for $Y_i(0)$:

$$\mathbb{E}[Y_i(0)|D_i, X_i] = X_i\eta. \tag{2.3}$$

And a design-based specification that leverages a model for treatment assignment:

$$\mathbb{E}[D_i|Y_i(0), Y_i(1), X_i] = X_i\gamma. \tag{2.4}$$

The model-based specification assumes that the conditional mean of the untreated potential outcome under control is linear in covariates, given the treatment status. A linear model for the treated outcome is also implied by this assumption since $Y_i(1) = \tau_i + Y_i(0)$. This assumption incorporates linear models for the potential outcomes, which could be the reduced form obtained from a set of structural equations or a specification directly imposed on the outcome. An example is the parallel trend assumptions, where X_i includes time and unit dummies with i indexing a pair of time and unit. The design-based specification assumes that the conditional mean of treatment assignment is linear in covariates, given the potential outcomes. A pair of sufficient conditions is unconfoundedness coupled with linear propensity score of the treatment ($\mathbb{E}[D_i|X_i] = X_i\gamma$). An example is when D_i is the eligibility for a program which is determined by some features X_i of the individual, and τ_i is the intention-to-treat effect.

We explore the weighting scheme of OLS estimand under model- and design-based specifications. Under Equation 2.3,

$$\tau_{ols} = \frac{\mathbb{E}[\lambda_i\tau_i]}{\mathbb{E}[\lambda_i]}, \quad \text{where } \lambda_i = \tilde{D}_i D_i. \tag{2.5}$$

Under Equation 2.4,

$$\tau_{ols} = \frac{\mathbb{E}[\omega_i\tau_i]}{\mathbb{E}[\omega_i]}, \quad \text{where } \omega_i = Var(D_i|Y_i(0), Y_i(1), X_i). \tag{2.6}$$

Under either assumption, the OLS estimand can be interpreted as a weighted average of individual treatment effects. For Equation 2.5, first notice $\lambda_i = 0$ when $D_i = 0$, indicating that the individual effects of control units do not contribute to τ_{ols}. Secondly, λ_i cannot preclude negative weights, which appear when the projection of D_i on X_i exceeds unit value, resulting in $\tilde{D}_i < 0$ for treated units. Therefore, model-based specification identifies a weighted ATT (average effects on treated) with potentially negative weights. In contrast, ω_i is guaranteed to be convex as it is identified as the conditional variance of D_i under design-based specification.

Similar result to Equation 2.5 is also presented in Borusyak & Hull (2024) and they concern that "... (model-based) weighting scheme is not convex. This is because $\tilde{D}_i$ necessarily takes on both positive and negative values, since $\mathbb{E}[\tilde{D}_i] = 0$". However, we show that zero expectation of $\tilde{D}_i$ is not the cause of negative weights, but rather the fact that D_i is binary and the projection of D_i on X_i is not necessarily the propensity model of D_i under Equation 2.3.

While design-based specifications avoid negative weights and model-based specifications do not, design-based specifications are not necessarily more reliable in empirical practice. Under model-based specification, a sufficient and necessary condition for convex weights is having all fitted values of D_i less than 1, which is less restrictive than requiring a linear probability model for treatment, as in Equation 2.4. An example where model-based weights can be convex, yet design-based specification in Equation 2.4 is unlikely to hold, is when the proportion of treated units is small. In Section 4, we illustrate this argument in the Lalonde dataset, where the proportion of treated units is 1.1%.

The result in Equation 2.5 further motivates the investigation of least square estimand with more flexible specifications and continuous treatment. In particular, consider the least square estimand in a partially linear regression. $\tilde{D}_i$ in Equation 2.5 is replaced by the residuals from a projection of D_i onto an infinite-dimensional linear space. Although this space may not include the true treatment model, the projection is expected to provide a more accurate estimation of treatment probabilities, thereby mitigating concerns of negative weights. The concern should also be alleviated when the treatment is continuous, because unlike binary treatments, the fitted values of continuous treatments are typically not bounded. The formal discussion is provided in Section 3.

In the rest of this section, we discuss scenarios where convex weights are guaranteed in model-based OLS, as well as the connection between design-based OLS and IPW. Note that the potential threat of negative weights is likely to persist in stronger models of the potential outcomes than Equation 2.3, such as assuming conditional independence between potential outcomes and treatment. Because negative weights arise as the treatment is residualized by variables in the outcome model which may not predict the treatment well. However, extra assumptions on the covariates or conditional average effects can be helpful for obtaining convex weights.

Special Case 1: Saturated X_i

If X_i is saturated, the projection of D_i on X_i takes values between $[0, 1]$ for all i and thus, the weights λ_i are guaranteed to be convex. Moreover, if D_i is conditionally mean-independent of the potential outcomes, saturated X_i also implies Equation 2.4. This coincides with the result in Angrist (1998) that, without considering the specification assumption, OLS with saturated covariates estimates a convex combination of heterogeneous treatment effects. An example of saturated X_i is the One-Way Fixed Effect model

$$Y_i = \beta_s + \tau_{owfe} D_i + \epsilon_i$$

where β_s are strata fixed effects. Let $n_{1s} = \sum_{i \in s} D_i$, $n_{0s} = \sum_{i \in s} 1 - D_i$ and $\bar{Y}_{ds}$ be the sample mean of Y_i in the treatment and control group for $d = 1, 0$ in stratum s. We can rewrite τ_{owfe} as

$$\tau_{owfe} = \frac{\mathbb{E}[\lambda_{owfe,s} \tau_s]}{\mathbb{E}[\lambda_{owfe,s}]}$$

where each unit i in stratum s has the same weight $\lambda_s = \frac{n_{1s} n_{0s}}{n_{1s} + n_{0s}}$ and $\tau_s = \bar{Y}_{1s} - \bar{Y}_{0s}$.

Special Case 2: Full Interaction Regression

Consider an extra assumption on the treatment effect $\mathbb{E}[\tau_i | D_i, X_i] = X_i \xi$. Then the concern of negative weights can be solved by running a full interaction regression:

$$Y_i = \tau_{ob} D_i + (X_i - \bar{X})\beta + D_i(X_i - \bar{X})\xi + \epsilon_i$$

where $\bar{X}$ is the sample mean of X_i and τ_{ob} is an estimand for ATE. This method is equivalent to separately estimating two OLS coefficients $\hat{\pi}_1$ using the treatment group data only, and $\hat{\pi}_0$ using the control group data only. The estimator $\hat{\tau}_{ob}$ can be recovered by computing $\hat{\tau}_{ob} = n^{-1} \sum_{i=1}^{n} (\hat{Y}_i(1) - \hat{Y}_i(0))$ where

$$\hat{Y}_i(d) = \begin{cases} Y_i, & D_i = d, \\ X_i \hat{\pi}_d, & D_i \neq d. \end{cases}$$

This procedure is also algebraically equivalent to the Oaxaca-Blinder method (Oaxaca 1973, Blinder 1973), and is subsequently studied in the context of experiment (Freedman 2008, Lin 2013, Guo & Basse 2023, Negi & Wooldridge 2021) and survey data (Wooldridge 2010, Kline 2011). In particular, Kline (2011) investigated the connection between Oaxaca-Blinder and IPW estimator for the ATT and found that the Oaxaca-Blinder is equivalent to IPW with a linear model for the conditional odds of being treated. The assumptions imposed by Kline (2011) are two linear models for $Y_i(1)$ and $Y_i(0)$ with unconfoundedness of treatment, similar to our combination of Equation 2.3 and conditional expectation of τ_i. It is also shown in the paper that Oaxaca-Blinder estimator is doubly robust (Robins et al. 1994), which implies convex weight of τ_{ob} under either model- or design-based specifications.

Special Case 3: Canonical Difference-in-Differences

Canonical DiD refers to the 2 × 2 DiD with the parameter of interest ATT estimated by TWFE or regression with group and period indicators. The parallel trend assumption is a model-based specification and models $Y_i(0)$ as a linear function of group and time fixed effects. This is the case where negative weights do not exist and all treated units are equally weighted such that the ATT is identified.

Consider the regression equation of TWFE

$$Y_{it} = \beta_i + \gamma_t + \tau_{twfe} D_{it} + \epsilon_{it} \tag{2.7}$$

where β_i and γ_t are group and time fixed effects, respectively.

Consider residualizing the treatment D_{it} with fixed effects, which is equivalent to the within transformation

$$\tilde{D}_{it} = D_{it} - \bar{D}_i - \bar{D}_t + \bar{D}$$

where $\bar{D}_i = T^{-1} \sum_t D_{it}$ is proportion of observations in the post-treatment period; $\bar{D}_t = N^{-1} \sum_i D_{it}$ is the proportion of units in the treatment group; and $\bar{D} = (NT)^{-1} \sum_{it} D_{it}$ is the proportion of observations that received treatment. Then negative values of $\tilde{D}_{it}$ arise for observations in the treatment group and pre-treatment period; or in the control group and post-treatment period. For example, in a balanced panel where each unit is observed in both pre and post treatment periods, and assume that p is the proportion of treatment group, the values for $\tilde{D}_{it}$ are listed in Table 1.

Table 1: The values of $\tilde{D}_{it}$ in balanced panel data

	Treated	Control
pre	$-(1-p)/2$	$p/2$
post	$(1-p)/2$	$-p/2$

Nonetheless, $D_{it} = 1$ for treatment group units in post period and otherwise zero, which implies that $\lambda_{it} = (1-p)/2$ for all observations that received treatment and otherwise zero. Thus, τ_{twfe} identifies the ATT and there is no concern of negative weights.

2.2 Connection Between OLS and IPW

Although stronger assumptions than Equation 2.4 are not necessary to guarantee convex weights, they may be able to simplify the weighting function. In this section, we explore the unconfoundedness assumption coupled with linear propensity model, and show that the OLS estimand is equivalent to overlap-weight IPW. Formally, we consider the following assumptions.

Assumption 2.1 (Linear Propensity)**.**

$$e(x) = \mathbb{E}[D_i | X_i = x] = X_i \delta$$

where $e(x)$ is the propensity score of treatment.

Assumption 2.2 (Unconfoundedness and Overlap)**.**

1. Unconfoundedness: $D_i \perp (Y_i(0), Y_i(1))|X_i$.
2. Overlap: $\exists\, \xi > 0,$ s.t. $\xi \leqslant e(x) \leqslant 1 - \xi$.

Assumption 2.1 is assumed in Aronow & Samii (2016), Słoczyński (2022) and Goldsmith-Pinkham et al. (2022). Assumption 2.2 is standard in the literature of IPW (e.g. Rubin 1974, Rosenbaum & Rubin 1983, Hahn 1998, Heckman et al. 1998).

Proposition 2.1 (Equivalence of OLS and Overlap-Weighted IPW Estimands)**.** *Under Assumption 2.1*

$$\tau_{ols} = \tau_{overlap}$$

where $\tau_{overlap}$ is the estimand of overlap-weighted IPW

$$\tau_{overlap} = \mathbb{E}\left[\frac{e(X_i)(1-e(X_i))}{\mathbb{E}[e(X_i)(1-e(X_i))]} \cdot \left(\frac{Y_i D_i}{e(X_i)} - \frac{Y_i(1-D_i)}{1-e(X_i)}\right)\right]. \tag{2.8}$$

Assumption 2.2 further yields the identification of overlap-weighted ATE:

$$\tau_{ols} = \mathbb{E}\left[\frac{e(X_i)(1-e(X_i))}{\mathbb{E}[e(X_i)(1-e(X_i))]} \cdot \tau_i\right].$$

The weight $e(X_i)(1-e(X_i))/\mathbb{E}[e(X_i)(1-e(X_i))]$ is referred to as overlap weight in the literature of IPW. It belongs to a class of generalized IPW estimators first proposed by Hirano et al. (2003), with its properties discussed in detail by Li et al. (2018). In particular, Li et al. (2018) show that the overlap weight is less sensitive to propensities close to 0 or 1, and leads to the minimum variance in the class of generalized IPW estimators under mild conditions.

Proposition 2.1 shows that the extra assumption required for establishing the connection between OLS and IPW is linear propensity score. Proposition 2.1 can be seen as a stronger version of Equation 2.6. Equation 2.4 is implied by the linear propensity coupled with unconfoundedness, and the design-based weight ω_i reduces to $Var(D_i|X_i) = e(X_i)(1 - e(X_i))$. The overlap assumption guarantees $Var(D_i|X_i) > 0$.

It is also shown that OLS is inferior to overlap-weighted IPW in terms of flexibility as the latter does not require linear propensity to identified causal effects. When estimating $\tau_{overlap}$, one can choose an estimator of $e(X_i)$, such as logit or probit, and plug in the sample analogy of $\tau_{overlap}$. While for τ_{ols}, $e(X_i)$ is implicitly estimated by OLS. This is no longer a constraint if we consider partially linear regression. On the other hand, OLS provides convenience when conducting hypothesis tests, as the asymptotic variance does not need to take the estimated $e(X_i)$ into account.

2.3 Extension to IV Regression

The analysis on binary treatment can be extended to the identification of local average treatment effect by IV regression. Suppose there is a binary instrument Z_i. Consider the IV version of the model-based specification:

$$\mathbb{E}[Y_i(0)|Z_i, X_i] = X_i\eta. \tag{2.9}$$

And design-based specification:

$$\mathbb{E}[Z_i|Y_i(0), Y_i(1), X_i] = X_i\gamma. \tag{2.10}$$

Following Small et al. (2017), we also assume

Assumption 2.3 (Stochastic First-Stage Monotonicity)**.**

$Pr(D_i = 1|Y_i(0), Y_i(1), Z_i = 1, X_i = x) \geqslant Pr(D_i = 1|Y_i(0), Y_i(1), Z_i = 0, X_i = x)$ for all x.

Assumption 2.3 is a relaxation of the traditional deterministic monotonicity assumption. Deterministic monotonicity assumes that the treatment level for each unit is monotonically increasing with the instrument level, while Assumption 2.3 only requires the monotonic increasing relationship conditioning on a set of covariates.

Let $\tilde{Z}_i$ be the residual of the population projection of Z_i on X_i. We require relevance between Z_i and D_i such that $\mathbb{E}[\tilde{Z}_i D_i] \neq 0$. Then, by the Frisch-Waugh-Lovell theorem, IV estimand can be written as

$$\tau_{iv} = \frac{\mathbb{E}[\tilde{Z}_i Y_i]}{\mathbb{E}[\tilde{Z}_i D_i]} = \frac{\mathbb{E}[\tau_i \tilde{Z}_i D_i] + \mathbb{E}[\tilde{Z}_i Y_i(0)]}{\mathbb{E}[\tilde{Z}_i D_i]}. \tag{2.11}$$

We obtain the weighting scheme for IV regression. Under Equation 2.9,

$$\tau_{iv} = \frac{\mathbb{E}[\lambda_i \tau_i]}{\mathbb{E}[\lambda_i]}, \quad \text{where } \lambda_i = \tilde{Z}_i D_i.$$

Under Equation 2.10,

$$\tau_{iv} = \frac{\mathbb{E}[\omega_i \tau_i]}{\mathbb{E}[\omega_i]}, \quad \text{where } \omega_i = Cov(\tilde{Z}_i, Pr(D_i = 1|Y_i(0), Y_i(1), Z_i, X_i)).$$

And ω_i is non-negative under Assumption 2.3.

Under model-based specification, λ_i may be non-convex, similar to the problem with OLS. For units with $D_i = 0$, all $\lambda_i = 0$. While for units with $D_i = 1$, $\tilde{Z}_i$ can be negative when Z_i is either 0 or 1 and the projection of Z_i on X_i exceeds the unit interval. Under design-based specification and stochastic monotonicity, the weight ω_i is non-negative because both $\tilde{Z}_i$ and $Pr(D_i = 1|Y_i(0), Y_i(1), Z_i, X_i)$ are weakly increasing functions of Z_i. The result on design-based weight is consistent with Small et al. (2017), who show that nonparametric Wald estimators identify a convex average of individual treatment effects for the stochastic complier group under the assumption of unconfounded instrument.

3 General Problem

3.1 Preparation

We now introduce the general problem. Suppose we have n iid observations of random variables (Y_i, D_i, X_i) distributed according to unknown distribution P, where $Y_i \in \mathbb{R}$ is the outcome, treatment $D_i \in \mathbb{R}$ is continuous and $X_i \in \mathbb{R}^p$ is a p-dimensional vector of covariates. The potential outcome $Y_i(d) = Y_i$ under treatment $D_i = d$. Let $\mu(D_i, X_i) = \mathbb{E}[Y_i|D_i, X_i]$. Let $\mathcal{H}$ be a Hilbert space of functions $l : \mathbb{R}^{p+1} \mapsto \mathbb{R}$ equipped with inner-product $\langle l, h\rangle = \mathbb{E}[l(D_i, X_i)h(D_i, X_i)]$ and norm $\|l\| = \langle l, l\rangle^{1/2}$. Assume that $\mu \in \mathcal{H}$.

We consider a general class of parameter that is a weighted average of derivative effect (WADE) $\tau_{wade} = \mathbb{E}[w(D_i, X_i)\mu'(D_i X_i)]$ for some weight $w(d, x)$ and the superscript prime denotes the derivative with respect to D_i. By the Riesz Representation theorem, there exists a unique Riesz Representer $\alpha_w \in \mathcal{H}$ such that $\tau_{wade} = \mathbb{E}[\alpha_w(D_i, X_i)\mu(D_i, X_i)]$ and hence $\mathbb{E}[\alpha_w(D_i, X_i)Y_i]$. To provide empirical relevance and interpretability, we restrict the class of WADE such that α_w belongs to the set

$$\mathcal{A} = \{\alpha \in \mathcal{H} | \mathbb{E}[\alpha(D_i, X_i)D_i] = 1, \mathbb{E}[\alpha(D_i, X_i)|X_i] = 0\}.$$

We show in Appendix that these τ_{wade} are normalized in the sense that $\mathbb{E}[w(D_i, X_i)] = 1$ and least square estimands belong to this class.

The main motivation to consider WADE of continuous treatments is that a large number of empirical research involves evaluating the causal effect of a continuous treatment on the outcome. In the recent work by Rothenhäusler & Yu (2019), they propose the incremental effects of continuous treatment and show that the ADE $\mathbb{E}[\mu'(D_i, X_i)]$ identifies the incremental treatment effect. We highlight the connection by reframing their result here. Define the incremental effect as

$$\tau^{\nu} = \frac{\mathbb{E}[Y_i(d+\nu)] - \mathbb{E}[Y_i(d)]}{\nu}$$

where ν is a shift in the treatment level. However, directly evaluating τ^{ν} imposes issues in practice. First, it is unclear which shift should be considered when estimating τ^{ν}. Second, there may be large shifts which are unrealistic to some units in the population. In light of these concerns, Rothenhäusler & Yu (2019) propose a causal estimand $\tau^0 = \lim_{\nu \to 0} \tau^{\nu}$ and show that $\mathbb{E}[\mu'(D_i, X_i)]$ identifies τ^0 under a local version of unconfoundedness assumption and that the conditional density $f(d|x)$ is continuous and differentiable (See Proposition 1 in Rothenhäusler & Yu (2019) for details). We adapt their identification result by showing that $\mu'(d, x)$ identifies conditional incremental effect under either model- or design-based assumption in the proof of Theorem 3.1, and extend τ^0 to the class of WADE.

In the literature, another framework for exploring continuous treatment effect has been studied by Callaway et al. (2024) and Borusyak & Hull (2024). They restrict the treatment value to be $D_i \geqslant 0$ and consider either the average causal derivative $\partial\mathbb{E}[Y_i(d)]/\partial d$ or the direct derivative $\partial Y_i(d)/\partial d$. However, they provide limited discussion on the general identification of the derivatives, and the restricted support of D_i can be violated in empirical works.

The Riesz Representation of $\mathbb{E}[w(D_i, X_i)\mu'(D_i, X_i)]$ was originally considered for estimation of parameters in index models (e.g. Härdle & Stoker 1989, Powell et al. 1989). In particular, Powell et al. (1989) use integration by parts to derive the Riesz Representer of $\mathbb{E}[w(X_i)\mu'(X_i)]$ where $\mu(X_i) = \mathbb{E}[Y_i|X_i]$:

$$\alpha_w(x) = -\frac{\mathrm{d}w(x)}{\mathrm{d}x} - w(x) \cdot \frac{\mathrm{d}\log f(x)}{\mathrm{d}x} \tag{3.1}$$

where $f(x)$ is the density function of X_i. In the context of our research question, the Riesz Representation is given by the least square estimand with the form $\mathbb{E}[\alpha_w(D_i, X_i)Y_i]$. We reversely derive the associated weights as in $\mathbb{E}[w(D_i, X_i)\mu'(D_i, X_i)]$ and interpret the weights under model- and design-based assumptions. Note that in the setting of binary D_i, the analogous WATE is $\tau_{wate} = \mathbb{E}[w(D_i, X_i)(\mu(1, X_i) - \mu(0, X_i))]$. By assuming a linear form of $\mu(D_i, X_i)$, the Riesz Representation gives the results for OLS as in Section 2.1.

3.2 Least Square Estimand as WADE

Consider estimating the average derivative effect using the following partially linear regression:

$$Y_i = \tau_{ls} D_i + g(X_i) + \epsilon_i. \tag{3.2}$$

The least square estimand τ_{ls} and g are defined by

$$(\tau_{ls}, g) = \arg \min_{\tau \in \mathbb{R}, g^* \in \mathcal{G}} \mathbb{E}\left[(Y_i - \tau D_i - g^*(X_i))^2\right]$$

for some linear space of functions $\mathcal{G}$. $\mathcal{G}$ can be specified as a class of parametric or nonparametric functions. If $\mathcal{G} = \{x\beta : \beta \in \mathbb{R}^{\dim(X_i)}\}$, then the problem reduces to the case in Section 2.1.

Let $\tilde{D}_i$ denote the residuals from projecting D_i onto $\mathcal{G}$, formally, $\tilde{D}_i = D_i -$

$\arg\min_{g^* \in \mathcal{G}} \mathbb{E}[(D_i - g^*(X_i))^2]$. By the projection theorem,

$$\tau_{ls} = \frac{\mathbb{E}[\tilde{D}_i Y_i]}{\mathbb{E}[\tilde{D}_i D_i]}.$$

The above estimand is a Riesz Representation of the form $\mathbb{E}[\alpha_w(D_i, X_i) Y_i]$ with $\alpha_w(D_i, X_i) = \tilde{D}_i / \mathbb{E}[\tilde{D}_i D_i]$. It is straight forward to see that $\alpha_w \in \mathcal{A}$.

To interpret τ_{ls}, consider the following two assumptions:

Assumption 3.1 (Model-Based Specification)**.** Let $\mu^d(X_i) = \mathbb{E}[Y_i(d)|X_i]$.

$$\mathbb{E}[Y_i(d)|D_i, X_i] = \mu^d(X_i), \text{ and } \mu^d \in \mathcal{G} \text{ for all } d.$$

Assumption 3.2 (Design-Based Specification)**.** Let $q(X_i) = \mathbb{E}[D_i|X_i]$.

$$\mathbb{E}[D_i|Y_i(d), X_i] = q(X_i) \text{ for all } d, \text{ and } q \in \mathcal{G}.$$

Theorem 3.1 (Weighting Scheme in LS Estimand)**.** *Under Assumption 3.1,*

$$\tau_{ls} = \mathbb{E}\left[\lambda(D_i, X_i)\mu'(D_i, X_i)\right], \textit{where } \lambda(d, x) = -\frac{Cov(\tilde{D}_i, \mathbb{1}(D_i \leqslant d)|X_i = x)}{f(d|x)\mathbb{E}[\tilde{D}_i D_i]}$$

Under Assumption 3.2,

$$\tau_{ls} = \mathbb{E}\left[\omega(D_i, X_i)\mu'(D_i, X_i)\right], \textit{where } \omega(d, x) = -\frac{Cov(\tilde{D}_i, \mathbb{1}(D_i \leqslant d)|X_i = x)}{f(d|x)\mathbb{E}[Var(D_i|Y_i(\cdot), X_i)]}$$

Theorem 3.1 is an extension of the results in Section 2.1 to partially linear regression with continuous treatment. Since $\tilde{D}_i$ is increasing in D_i and $\mathbb{1}(D_i \leqslant d)$ is non-increasing in D_i, $-Cov(\tilde{D}_i, \mathbb{1}(D_i \leqslant d)|X_i = x) \geqslant$ is non-negative, implying that both λ and ω are convex weights[1]. Unlike the binary treatment case, the weights depends on the unknown

[1] The result for continuous treatment is also different from Borusyak & Hull (2024), who focus on a different causal parameter and show that negative weights exists in model-based specification. We note that in their proof of Proposition 1, they use the equality of $\mathbb{E}[\psi_i(d)\beta_i(d)|Y_i(\cdot), X_i] = \beta_i(d)\mathbb{E}[\psi_i(d)|Y_i(\cdot), X_i]$ with $\beta_i(d) = (\partial/\partial d)Y_i(d)$ and ψ indicating the model-based weight function. The equation does not hold in general. One counter example is by letting $Y_i(d) = dV_i + \epsilon_i$ where ϵ_i and V_i are random variables

distribution through the conditional density $f(d|x)$. Consider a stronger design-based specification where a parametric distribution of $f(d|x)$ is assumed, then the weights may be further simplified and provide a more interpretable form. For example, if $f(d|x)$ is a conditionally normal distribution, $\omega(d, x)$ reduces to $Var(D_i|X_i)/\mathbb{E}[Var(D_i|X_i)]$, which coincides with the design-based weights for binary treatment. In Section A.1, we also derive the weight function when $f(d|x)$ is Gamma and Beta distribution.

3.3 Least Square Estimand is Optimal

In the literature of WATE with binary treatment, Crump et al. (2006) and Li et al. (2018) show that the overlap weight Equation 2.8 optimizes the efficiency bound in the class of WATE, under the assumption of unconfoundedness. In this section, we extend this result to the WADE of continuous treatment. In particular, we show that the least square estimand of partially linear regression is optimal in the class of WADE when the outcome variable is homoskedastic.

We rely on the influence function to characterize the efficiency bound of the WADE. Given an efficient estimator $\hat{\tau}_{wade}$ of WADE τ_{wade}, the asymptotic variance of the estimator is given by the variance of the corresponding influence function. Formally,

$$\sqrt{n}(\hat{\tau}_{wade} - \tau_{wade}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(Y_i, D_i, X_i) + o_p(1)$$

where $\phi(Y_i, D_i, X_i)$ is the influence function of τ_{wade}. And thus,

$$\sqrt{n}(\hat{\tau}_{wade} - \tau_{wade}) \xrightarrow{d} N(0, V), \text{ where } V = Var(\phi(Y_i, D_i, X_i)).$$

According to Newey & Stoker (1993), the influence function of τ_{wade} can be constructed by $\alpha_w(y - \mu) + w \cdot \mu' - \tau_{wade}$, which gives

$$\phi(Y_i, D_i, X_i) = \alpha_w(D_i, X_i)(Y_i - \mu(D_i, X_i)) + w(D_i, X_i)\mu'(D_i, X_i) - \tau_{wade}. \tag{3.3}$$

By minimizing $Var(\phi(Y_i, D_i, X_i))$ with respect to $w(D_i, X_i)$ subject to the constraint $\alpha_w \in$

independent of X_i. Then $\mathbb{E}[\beta_i(d)|Y_i(\cdot), X_i] = \mathbb{E}[V_i|dV_i + \epsilon_i] \neq \beta_i(d)$.

$\mathcal{A}$, we show that τ_{ls} is the optimal WADE when the outcome is homoskedastic.

Theorem 3.2 (Least Square Estimand is Optimal WADE)**.** *Suppose that Y_i is homoskedastic, i.e. $Var(Y_i|D_i, X_i) = \sigma^2$ is a constant and $D_i \perp Y_i(d)|X_i$ for all d. Then the optimally efficient WADE is τ_{ls}.*

The efficiency property shown in Theorem 3.2 encourages the use of partially linear regression for estimating the treatment effect over other methods targeting different parameters in the class of WADE. In the proof in Appendix, we also show that Theorem 3.2 incorporates Theorem 5.4 in Crump et al. (2006) and Theorem 2 in Li et al. (2018) when the treatment is binary.

4 Application

4.1 Estimation

Under the assumption of linear model- and design-based specifications, OLS estimation of the WATE is straightforward. When the specifications are known but nonlinear, one can modify $g(x)$ in Equation 3.2 to fit the specification. For example, consider a design-based specification where the propensity score of D_i is known to be a logit model of X_i, then $g(x)$ can be replaced by $e(x)$. And τ_{ls} can be estimated by running OLS of Y_i on D_i and the logistic estimate $\hat{e}(X_i)$ with standard error calculated by the sample analog of $\sqrt{Var(\phi(Y_i, D_i, X_i))}$.

When the specifications are unknown, estimation with partially linear regression is still an active research topic. Here, we provide some discussions on the estimation method. The influence function in Equation 3.3 can be used to construct efficient estimator for WADE. By plugging in the $\alpha_w(d, x)$ and $w(d, x)$ of τ_{ls}, we have the influence function

$$\phi_{ls}(Y_i, D_i, X_i) = \frac{D_i - e(X_i)}{\mathbb{E}[(D_i - e(X_i))^2]}(Y_i - \rho(X_i) - \tau_{ls}(D_i - e(X_i))) \tag{4.1}$$

where $e(x) = \mathbb{E}[D_i|X_i = x]$ and $\rho(x) = \mathbb{E}[Y_i|X_i = x]$. The efficient estimator of τ_{ls} can be obtained by setting the sample mean of ϕ_{ls} to zero and solve for τ_{ls}:

$$\hat{\tau}_{ls} = \frac{\sum_{i=1}^{n}(D_i - \hat{e}(X_i))(Y_i - \hat{\rho}(X_i))}{\sum_{i=1}^{n}(D_i - \hat{e}(X_i))^2} \tag{4.2}$$

where $\hat{e}(x)$ and $\hat{\rho}(x)$ are nonparametric estimators of the nuisance parameters $e(x)$ and $\rho(x)$, respectively. This is so-called the estimating equation estimator. This form of $\hat{\tau}_{ls}$ has been explored in the literature of partially linear model with an emphasis on the convergence rate of the nuisance estimators. In particular, Robins et al. (1994) imposes Donsker condition on the quantities $(D_i - \hat{e}(X_i))(Y_i - \hat{\rho}(X_i))$ and $(D_i - \hat{e}(X_i))^2$, and shows the consistency and asymptotic normality of $\hat{\tau}_{ls}$. However, the Donsker condition is not guaranteed to hold when $\hat{e}(x)$ and $\hat{\rho}(x)$ are flexible machine learning estimators. Chernozhukov et al. (2018) provide the method of cross-fitting to avoid the Donsker condition, which requires $\hat{e}(x)$ and $\hat{\rho}(x)$ to be estimated from a sample independent of the one used to construct $\hat{\tau}_{ls}$. Here, we provide a set of conditions such that $\hat{\tau}_{ls}$ is a regular asymptotically linear estimator (RAL) by unifying the conditions in Robins et al. (1994) and Chernozhukov et al. (2018).

Proposition 4.1. *Assume that the following conditions hold:*

a. $\|e - \hat{e}\| = o_P(n^{-a/4})$ *and* $\|\rho - \hat{\rho}\| = o_P(n^{-b/4})$ *where* $a \geqslant 1,\ b \geqslant 0$ *and* $a + b \geqslant 2$.

b. $n^{-1}\sum_{i=1}^{n}(D_i - \hat{e}(X_i)) > 0$ *and* $\mathbb{E}[D_i - e(X_i)] > 0$.

c. There exists a constant $0 < C < \infty$ *such that* $Var(D_i|X_i) < C$ *and* $Var(Y_i|X_i) < C$.

And assume that at least one of d or e holds:

d. (Donsker condition) The quantities $(D_i - \hat{e}(X_i))(Y_i - \hat{\rho}(X_i))$ *and* $(D_i - \hat{e}(X_i))^2$ *fall within a P-Donsker class with probability approaching* 1.

e. (Cross-fitting) The sample used to estimate $\hat{e}(x)$ *and* $\hat{\rho}(x)$ *is independent of the sample used to construct* $\hat{\tau}_{ls}$.

Then, $\hat{\tau}_{ls}$ *is a RAL of* τ_{ls} *with influence function* ϕ_{ls}. *Hence,* $\sqrt{n}(\hat{\tau}_{ls} - \tau_{ls}) \xrightarrow{d} N(0, V)$ *and* $V = Var(\phi_{ls}(Y_i, D_i, X_i))$.

Note that Assumption a implies the robustness of design-based specification from the perspective of estimation. In particular, $\hat{\rho}$ can converge slowly as long as $\hat{e}$ converges sufficiently fast with a rate of at least $n^{1/4}$. We next illustrate the model- and design-based weights with various estimators in an application.

4.2 The Effect of NSW Job Training Program

To illustrate our results, we revisit the NSW-CPS data constructed by LaLonde (1986) regarding the effect of NSW training program on participants' earnings. This dataset has been analyzed in several influential papers using various methods including propensity score methods (Dehejia & Wahba 1999, Smith & Todd 2005) and OLS estimation (Angrist & Pischke 2009, Słoczyński 2022). The treatment variable D_i is a binary indicator of whether the individual participated in the program and our focus is on the identification and estimation of the WATE. The outcome variable Y_i is the individual's earnings in 1978. We consider four sets of covariates, same as those in Table 3.3.3 in Angrist & Pischke (2009).

Various estimators are presented in Table 2. The least square estimators targets the WATE presented in Section 2.1 and $\hat{\tau}_{ipw}$ is the Hajek-IPW estimator of the ATE with propensity score $e(X_i)$ estimated by logistic regression of D_i on X_i. A direct comparison to $\hat{\tau}_{ipw}$ is $\hat{\tau}_{ls,logit}$, which is the OLS estimator of Y_i on D_i and logistic $\hat{e}(X_i)$. $\hat{\tau}_{ls,cf}$ is the estimator in Equation 4.2 with both nuisance parameters estimated by random forest with 5-fold cross-fitting following the algorithm by Chernozhukov et al. (2018). $\hat{\tau}_{ls,donsker}$ follows the same estimation procedure but without cross-fitting[2].

The first row in Table 2 replicates Table 3.3.3 in Angrist & Pischke (2009). A comparison between the least square estimators and $\hat{\tau}_{ipw}$ reveals that the least square estimand can significantly differ from the ATE, with notably smaller standard errors[3], which aligns with its efficiency property. Since the Donsker condition is unlikely to hold on random forest estimator, $\hat{\tau}_{ls,donsker}$ may be biased, as indicated by the difference between $\hat{\tau}_{ls,donsker}$ and other least square estimators in column 2 and 4 of Table 2.

The weights in $\hat{\tau}_{ols}$ and $\hat{\tau}_{ls,cf}$ from column 4 Table 2 are presented in Figure 1 and Figure 2,

[2] The estimation of $\hat{\tau}_{ls,cf}$ and $\hat{\tau}_{ls,donsker}$ are implemented in R using `DoubleML` package, with $\hat{\rho}$ estimated by random forest learner `regr.ranger` and $\hat{e}$ by `classif.ranger`.

[3] The Hajek-IPW estimator is also an efficient estimator (Hirano & Imbens 2001), so the difference in the standard errors between $\hat{\tau}_{ls}$ and $\hat{\tau}_{ipw}$ is likely not attributable to the efficiency of the estimators but rather to different estimands.

respectively, under model- and design-based specifications. We use box plots to display the distribution of the weights (the first row) and scatter plots to illustrate the relationship with outcome values. The first column in Figure 1 and Figure 2 presents the estimated normalized weight $\lambda_i/\mathbb{E}[\lambda_i]$ under model-based specifications Equation 2.3 and Assumption 3.1. For the second column, since the design-based weight ω_i under Equation 2.4 and Assumption 3.2 is infeasible to estimate directly, we adopt the stronger design-based assumption of unconfoundedness and estimate the weight $\omega_i = Var(D_i|X_i)$.

In the sample, only 1.1% of individuals are treated, resulting in just 1.1% of the model-based weights being non-zero. And after normalization by $\mathbb{E}[\lambda_i]$, these non-zero weights can reach extreme values up to 100. Consequently, the OLS estimand under model-based specification identifies an ATT-type causal effect, with large weights assigned to treated units. Note that there are no negative model-based weights in this sample, as the small proportion of treated units drives the linear prediction line of D_i to fall below the horizontal line of one, ensuring $\tilde{D}_i > 0$ for all treated units. On the other hand, the design-based weights are centered at a small value with a narrower range compared to model-based weights, but negative values do arise, which reflects a violation of the linearity assumption in Equation 2.4. This is an example that model-based specifications may be more reliable than design-based specifications as discussed in Section 2.1. When the linearity assumption is relaxed using partially linear regression with the propensity score estimated by random forest, the negative values of design-based weights disappear, as shown in the right panel of Figure 2.

Table 2: The Effects of the NSW Training Program on Earnings

Estimator		(1)	(2)	(3)	(4)
Parametric	$\hat{\tau}_{ols}$	-3437	-77	663	794
		(612)	(596)	(610)	(619)
	$\hat{\tau}_{ls,logit}$	-3590	-38	817	902
		(619)	(603)	(615)	(616)
	$\hat{\tau}_{ipw}$	-7807	-9380	-6067	-3741
		(1691)	(931)	(1358)	(3072)
Nonparametric	$\hat{\tau}_{ls,cf}$	-3446	31	1019	927
		(639)	(616)	(692)	(672)
	$\hat{\tau}_{ls,donsker}$	-3429	1353	907	1577
		(630)	(599)	(634)	(574)
Demographics		$\surd$		$\surd$	$\surd$
Earnings in 1974					$\surd$
Earnings in 1975			$\surd$	$\surd$	$\surd$
$P(D_i = 1)$		1.1%	1.1%	1.1%	1.1%
N		16177	16177	16177	16177

The demographic controls are age, age squared, education, race, marriage status and an indicator for whether an individual has a degree. The standard errors are in parentheses. For $\hat{\tau}_{ols}$, the standard errors are the Huber-White robust errors; $\hat{\tau}_{ipw}$ is the Hajek-IPW estimator that targets the ATE; for other least square estimators, the standard errors are calculated by the sample variance of the influence function in Equation 4.1.

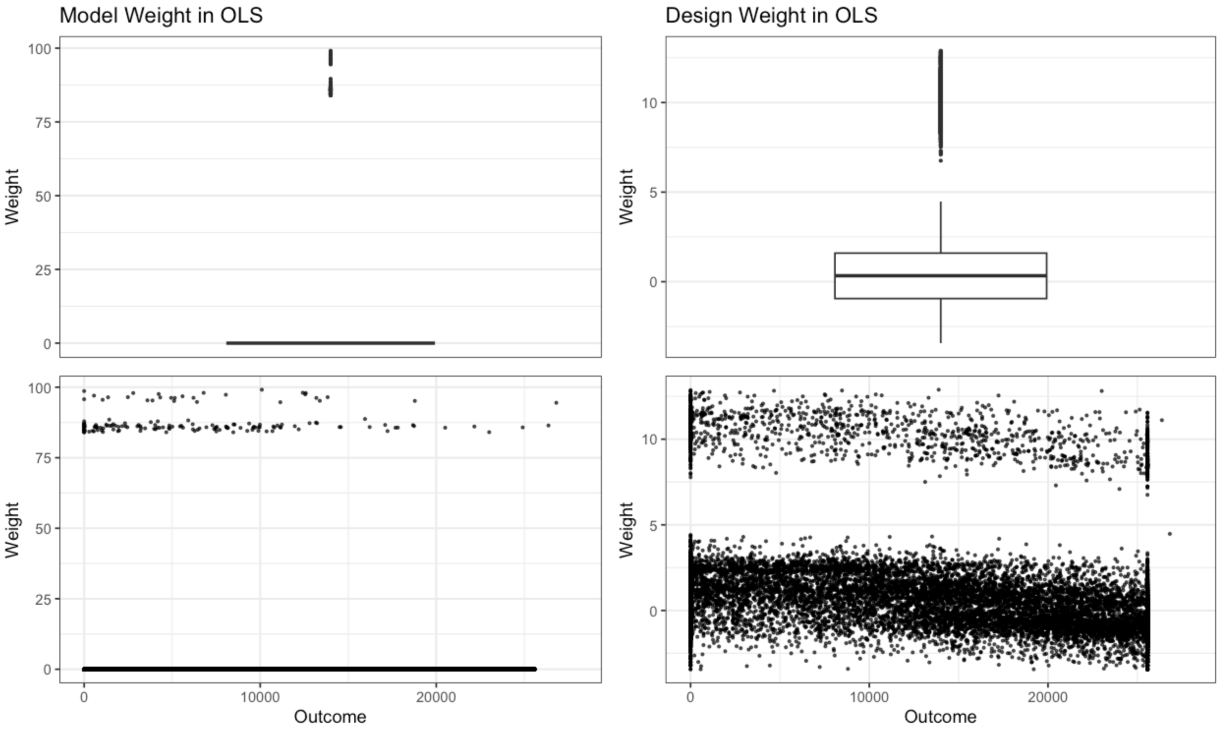

Figure 1: The Weights in OLS of Column 4 Table 2

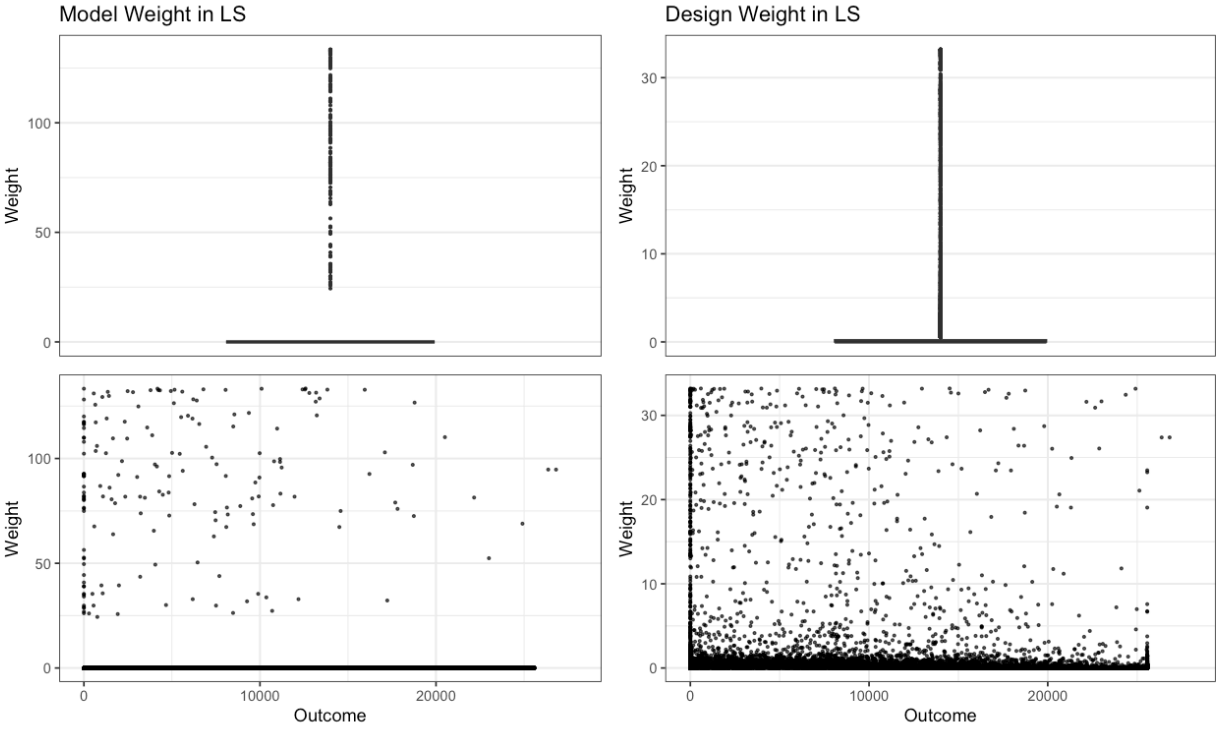

Figure 2: The Weights in Partially Linear Regression of Column 4 Table 2

5 Conclusion

This paper provides an extensive examination of the least squares estimand in the context of causal inference, offering both theoretical and practical guidance for researchers employing regression-based methods for estimating causal effects. Our study shows that the least squares estimand generally identifies weighted average effect and its interpretation depends on whether the specification is model-based or design-based. Under model-based specification with binary treatment, least square estimand identifies a weighted ATT and there is a potential risk of negative weights. However, these issues do not arise with continuous treatment. On the other hand, design-based specifications consistently guarantee convex weights for binary and continuous treatment.

Our results are illustrated using an application of the Lalonde dataset. The analysis highlights how different specification assumptions and estimation methods influence the weighting scheme of the estimands. Notably, violation of design-based specification is also a threat to convex weights in OLS estimation. In practice, we suggest empirical researchers to run OLS of treatment on covariates and check the distribution of residuals as diagnostics of convex weights. The partially linear regression is also recommended to avoid concerns of negative weights.

Finally, it is important to note that this study concentrates on a single treatment variable. In cases with multiple treatments or one treatment with multiple doses, negative weights arise under design-based specifications (Goldsmith-Pinkham et al. 2022). Nonetheless, in the special case of DiD with staggered treatments, negative weights can be addressed by a more flexible model-based specification (Sun & Abraham 2021).

References

Angrist (1998), 'Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants', *Econometrica* pp. 249–288.

Angrist, J. D., Imbens, G. W. & Rubin, D. B. (1996), 'Identification of causal effects using instrumental variables', *Journal of the American statistical Association* **91**(434), 444–455.

Angrist & Pischke, J.-S. (2009), *Mostly harmless econometrics: An empiricist's companion*, Princeton university press.

Aronow, P. M. & Samii, C. (2016), 'Does regression produce representative estimates of causal effects?', *American Journal of Political Science* **60**(1), 250–267.

Blandhol, C., Bonney, J., Mogstad, M. & Torgovitsky, A. (2022), When is tsls actually late?, Technical report, National Bureau of Economic Research.

Blinder, A. S. (1973), 'Wage discrimination: reduced form and structural estimates', *Journal of Human resources* pp. 436–455.

Borusyak, K. & Hull, P. (2024), Negative weights are no concern in design-based specifications, *in* 'AEA Papers and Proceedings', Vol. 114, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, pp. 597–600.

Callaway, B., Goodman-Bacon, A. & Sant'Anna, P. H. (2024), Difference-in-differences with a continuous treatment, Technical report, National Bureau of Economic Research.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018), 'Double/debiased machine learning for treatment and structural parameters'.

Crump, R. K., Hotz, V. J., Imbens, G. & Mitnik, O. (2006), 'Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand'.

De Chaisemartin, C. & d'Haultfoeuille, X. (2020), 'Two-way fixed effects estimators with heterogeneous treatment effects', *American Economic Review* **110**(9), 2964–2996.

Dehejia, R. H. & Wahba, S. (1999), 'Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs', *Journal of the American statistical Association* **94**(448), 1053–1062.

Freedman, D. A. (2008), 'On regression adjustments to experimental data', *Advances in Applied Mathematics* **40**(2), 180–193.

Goldsmith-Pinkham, P., Hull, P. & Kolesár, M. (2022), Contamination bias in linear regressions, Technical report, National Bureau of Economic Research.

Goodman-Bacon, A. (2021), 'Difference-in-differences with variation in treatment timing', *Journal of Econometrics* **225**(2), 254–277.

Guo, K. & Basse, G. (2023), 'The generalized oaxaca-blinder estimator', *Journal of the American Statistical Association* **118**(541), 524–536.

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

Härdle, W. & Stoker, T. M. (1989), 'Investigating smooth multiple regression by the method of average derivatives', *Journal of the American statistical Association* **84**(408), 986–995.

Heckman, J. J., Ichimura, H. & Todd, P. (1998), 'Matching as an econometric evaluation estimator', *The review of economic studies* **65**(2), 261–294.

Hirano, K. & Imbens, G. W. (2001), 'Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization', *Health Services and Outcomes research methodology* **2**, 259–278.

Hirano, K., Imbens, G. W. & Ridder, G. (2003), 'Efficient estimation of average treatment effects using the estimated propensity score', *Econometrica* **71**(4), 1161–1189.

Imbens, G. W. & Angrist, J. D. (1994), 'Identification and estimation of local average treatment effects', *Econometrica* **62**(2), 467–475.

Kennedy, E. H., Balakrishnan, S. & G'sell, M. (2020), 'Sharp instruments for classifying compliers and generalizing causal effects'.

Kline, P. (2011), 'Oaxaca-blinder as a reweighting estimator', *American Economic Review* **101**(3), 532–537.

LaLonde, R. J. (1986), 'Evaluating the econometric evaluations of training programs with experimental data', *The American economic review* pp. 604–620.

Li, F., Morgan, K. L. & Zaslavsky, A. M. (2018), 'Balancing covariates via propensity score weighting', *Journal of the American Statistical Association* **113**(521), 390–400.

Lin, W. (2013), 'Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique'.

Negi, A. & Wooldridge, J. M. (2021), 'Revisiting regression adjustment in experiments with heterogeneous treatment effects', *Econometric Reviews* **40**(5), 504–534.

Newey, W. K. & Stoker, T. M. (1993), 'Efficiency of weighted average derivative estimators and index models', *Econometrica: Journal of the Econometric Society* pp. 1199–1223.

Oaxaca, R. (1973), 'Male-female wage differentials in urban labor markets', *International economic review* pp. 693–709.

Powell, J. L., Stock, J. H. & Stoker, T. M. (1989), 'Semiparametric estimation of index coefficients', *Econometrica: Journal of the Econometric Society* pp. 1403–1430.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994), 'Estimation of regression coefficients when some regressors are not always observed', *Journal of the American statistical Association* **89**(427), 846–866.

Rosenbaum, P. R. & Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* **70**(1), 41–55.

Rothenhäusler, D. & Yu, B. (2019), 'Incremental causal effects', *arXiv preprint arXiv:1907.13258* .

Rubin, D. B. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies.', *Journal of educational Psychology* **66**(5), 688.

Słoczyński, T. (2022), 'Interpreting ols estimands when treatment effects are heterogeneous: Smaller groups get larger weights', *The review of economics and statistics* **104**(3), 501–509.

Small, D. S., Tan, Z., Ramsahai, R. R., Lorch, S. A. & Brookhart, M. A. (2017), 'Instrumental variable estimation with a stochastic monotonicity assumption'.

Smith, J. A. & Todd, P. E. (2005), 'Does matching overcome lalonde's critique of nonexperimental estimators?', *Journal of econometrics* **125**(1-2), 305–353.

Sun, L. & Abraham, S. (2021), 'Estimating dynamic treatment effects in event studies with heterogeneous treatment effects', *Journal of Econometrics* **225**(2), 175–199.

Van der Vaart, A. (1998), 'Functional delta method', *Asymptotic Statistics* pp. 291–303.

Wooldridge, J. M. (2010), *Econometric analysis of cross section and panel data*, MIT press.

A Appendix

A.1 Proofs

Proof of results in Section 2. We give a general proof for IV where OLS is a special case with $D_i = Z_i$. We start from Equation 2.11 and consider Equation 2.9. First, we have

$$\mathbb{E}[\tilde{Z}_i Y_i(0)] = \mathbb{E}[\mathbb{E}[\tilde{Z}_i Y_i(0)|Z_i, X_i]] = \mathbb{E}[\tilde{Z}_i \mathbb{E}[Y_i(0)|Z_i, X_i]] = \mathbb{E}[\tilde{Z}_i X_i \eta] = 0$$

where the second and fourth equality hold by construction of $\tilde{D}_i$ and the third equality hold under Equation 2.9. Then we have

$$\tau_{iv} = \frac{\mathbb{E}[\tau_i \tilde{Z}_i D_i]}{\mathbb{E}[\tilde{Z}_i D_i]} = \frac{\mathbb{E}[\lambda_i \tau_i]}{\mathbb{E}[\lambda_i]}$$

Under Equation 2.10,

$$\mathbb{E}[\tilde{Z}_i Y_i(0)] = \mathbb{E}[Y_i(0)\mathbb{E}[\tilde{Z}_i|Y_i(0), Y_i(1), X_i]] = \mathbb{E}[Y_i(0)\mathbb{E}[Z_i - X_i\gamma|Y_i(0), Y_i(1), X_i]] = 0$$

and

$$\mathbb{E}[\tau_i \tilde{Z}_i D_i] = \mathbb{E}[\mathbb{E}[\tilde{Z}_i D_i|Y_i(0), Y_i(1), Z_i, X_i]\tau_i]$$

Thus,

$$\tau_{iv} = \frac{\mathbb{E}[\omega_i \tau_i]}{\mathbb{E}[\omega_i]}, \quad \omega_i = \mathbb{E}[\tilde{Z}_i D_i|Y_i(0), Y_i(1), Z_i, X_i]$$

The weight ω_i is non-negative under Assumption 2.3, because

$$\mathbb{E}[\tilde{Z}_i D_i|Y_i(0), Y_i(1), Z_i, X_i] = Cov(\tilde{Z}_i, Pr(D_i = 1|Y_i(0), Y_i(1), Z_i, X_i))$$

where $\tilde{Z}_i$ and $Pr(D_i = 1|Y_i(0), Y_i(1), Z_i, X_i)$ are both weakly increasing in Z_i. Note that when $D_i = Z_i$ in the case of OLS, Assumption 2.3 is trivially satisfied and ω_i reduces to $Var(D_i|Y_i(0), Y_i(1), X_i)$ □

Proof of Proposition 2.1. By the Frisch-Waugh-Lovell theorem, we have

$$\tau_{ols} = \frac{\mathbb{E}[Y_i \tilde{D}_i]}{\mathbb{E}[\tilde{D}_i^2]}$$

where $\tilde{D}_i = D_i - e(X_i)$ where $e(X_i) = X_i\delta$ under Assumption 2.1. Then we have

$$\begin{aligned}
\tau_{ols} &= \frac{\mathbb{E}[Y_i(D_i - e(X_i))]}{\mathbb{E}[(D_i - e(X_i))^2]} \\
&= \frac{\mathbb{E}[Y_i(D_i - e(X_i))]}{\mathbb{E}[e(X_i)(1 - e(X_i))]} \\
&= \mathbb{E}\left[\frac{e(X_i)(1 - e(X_i))}{\mathbb{E}[e(X_i)(1 - e(X_i))]} \frac{Y_i(D_i - e(X_i))}{e(X_i)(1 - e(X_i))}\right] \\
&= \mathbb{E}\left[\frac{e(X_i)(1 - e(X_i))}{\mathbb{E}[e(X_i)(1 - e(X_i))]} \left(\frac{Y_i D_i}{e(X_i)} - \frac{Y_i(1 - D_i)}{1 - e(X_i)}\right)\right] \\
&= \tau_{overlap}
\end{aligned}$$

The first equation follows from the linear propensity score assumption, the second equation follows from the fact that D_i is generated from a Bernoulli distribution with probability $e(X_i)$ with conditional variance $e(X_i)(1 - e(X_i))$.

Under Assumption 2.2, $\tau_{overlap}$ identifies the overlap weighted ATE

$$\tau_{overlap} = \mathbb{E}\left[\frac{e(X_i)(1 - e(X_i))}{\mathbb{E}[e(X_i)(1 - e(X_i))]}(Y_i(1) - Y_i(0))\right]$$

which follows from standard proof of IPW. □

Proof of Theorem 3.1. The proof consists of 3 parts. First, we show that $\mu'(D_i, X_i)$ identifies incremental treatment effect under Assumption 3.1 and Assumption 3.2. Second, we provide a general weighting function for the class of WADE. Third, we interpret the weights of τ_{ls} under Assumption 3.1 and Assumption 3.2.

Part 1

Suppose $Y(d)$ is continuously differentiable with derivative $Y_i'(d)$. There exists a random variable $d' \in [d, d+\nu]$ such that

$$\frac{Y_i(d+\nu) - Y_i(d)}{\nu} = Y_i'(d').$$

By dominated convergence,

$$\lim_{\nu \to 0} \frac{\mathbb{E}[Y_i(d+\nu)|D_i = d, X_i] - \mathbb{E}[Y_i(d)|D_i = d, X_i]}{\nu} = \mathbb{E}[Y'(d)|D_i = d, X_i].$$

Under either Assumption 3.1 or Assumption 3.2, since $Y(d)$ and D_i are mean-independent, we have

$$\begin{aligned} \mathbb{E}[Y_i(d+\nu) - Y_i(d)|D_i = d, X_i] &= \mathbb{E}[Y_i(d+\nu)|D_i = d, X_i] - \mathbb{E}[Y_i(d)|D_i = d, X_i] \\ &= \mathbb{E}[Y_i(d+\nu)|D_i = d+\nu, X_i] - \mathbb{E}[Y_i(d)|D_i = d, X_i]. \end{aligned}$$

Dividing both sides by ν and taking the limit,

$$\mathbb{E}[Y_i'(d)|D_i = d, X_i] = (\partial/\partial d)\mathbb{E}[Y_i(d)|D_i = d, X_i].$$

As $Y_i = Y_i(d)$ given $D_i = d$, then

$$\mathbb{E}[Y_i'(d)|D_i = d, X_i] = (\partial/\partial d)\mathbb{E}[Y_i|D_i = d, X_i]$$

where the left-hand-side is the incremental effect as shown above, the right-hand-side is $\mu'(D_i, X_i)$. This completes the proof that $\mu'(D_i, X_i)$ identifies the conditional incremental effect.

Part 2

We extend Equation 3.1 to the context of WADE. Suppose (I) $w(d,x)f(d|x)$ is differentiable with respect to d, and (II) $w(\underline{d}, x)f(\underline{d}|x) = w(\bar{d}, x)f(\bar{d}|x) = 0$ where $\underline{d}$ and $\bar{d}$ are the lower and upper bound of support of D_i. Then for any $\mu \in \mathcal{H}$, we have

$$\begin{aligned}\mathbb{E}[w(D_i, X_i)\mu'(D_i, X_i)|X_i = x] &= \int_{\underline{d}}^{\bar{d}} w(d,x)\mu'(d,x)f(d|x)\mathrm{d}d \\ &= w(\underline{d},x)f(\underline{d}|x) - w(\bar{d},x)f(\bar{d}|x) - \int_{\underline{d}}^{\bar{d}} \mu(d,x)\frac{\partial w(d,x)f(d|x)}{\partial d}\mathrm{d}d \\ &= -\int_{\underline{d}}^{\bar{d}} \mu(d,x)\frac{\partial w(d,x)f(d|x)}{\partial d}\mathrm{d}d \\ &= \mathbb{E}[\alpha_w(D_i,X_i)\mu(D_i,X_i)|X_i = x]\end{aligned}$$

where we let

$$\alpha_w(d,x) = -\frac{\partial w(d,x)f(d|x)}{\partial d} = -\frac{\partial w(d,x)}{\partial d} - w(d,x)\cdot\frac{\partial \log f(d|x)}{\partial d}.$$

The extension to boundaries of $\pm\infty$ is straightforward by considering the limit of $w(d,x)f(d|x)$.

The weighting function can be obtained by solving the above equation as an ordinary differential equations for $w(d,x)$.

$$w(d,x) = \frac{F(d|x)(1-F(d|X))}{f(d|x)}(\mathbb{E}[\alpha_w(D_i,X_i)|D_i > d, X_i = x] - \mathbb{E}[\alpha_w(D_i,X_i)|D_i \leqslant d, X_i = x]).$$

This weighting function satisfies (I) by construction, and satisfies (II) because $F(d|x)(1 - F(d|x))$ is zero at the boundaries of the support of D_i.

Note that

$$\mathbb{E}[\alpha_w(D_i,X_i)|X_i] = \mathbb{E}[\alpha_w(D_i,X_i)|D_i \leqslant d, X_i](1-F(d|X_i)) + \mathbb{E}[\alpha_w(D_i,X_i)|D_i > d, X_i]F(d|X_i)$$

Since $\alpha_w \in \mathcal{A}$, $\mathbb{E}[\alpha_w(D_i, X_i)|X_i] = 0$. Therefore, the weighting function reduces to

$$\begin{aligned} w(d, z) &= -\frac{\mathbb{E}[\alpha_w(D_i, X_i)|D_i \leqslant d, X_i = x]F(d|x)}{f(d|x)} \\ &= \frac{-1}{f(d|x)} \int_{\underline{d}}^{d} \alpha_w(d^*, x) f(d^*|x) \mathrm{d}d^* \\ &= \frac{-1}{f(d|x)} \int_{\underline{d}}^{\bar{d}} \alpha_w(d^*, x) \mathbb{1}(d^* \leqslant d) f(d^*|x) \mathrm{d}d^*. \end{aligned}$$

Lastly, we show that the weights $w(d, x)$ are normalized.

$$\begin{aligned} \mathbb{E}[w(D_i, X_i)|X_i = x] &= \int_{\underline{d}}^{\bar{d}} w(d, x) f(d|x) \mathrm{d}d \\ &= -\int_{\underline{d}}^{\bar{d}} \int_{\underline{d}}^{\bar{d}} \alpha_w(d^*, x) \mathbb{1}(d^* \leqslant d) f(d^*|x) \mathrm{d}d^* \mathrm{d}d \\ &= \int_{\underline{d}}^{\bar{d}} \alpha_w(d^*, x) \int_{\underline{d}}^{\bar{d}} -\mathbb{1}(d^* \leqslant d) \mathrm{d}d f(d^*|x) \mathrm{d}d^* \\ &= \int_{\underline{d}}^{\bar{d}} \alpha_w(d^*, x)(d^* - \bar{d}) f(d^*|x) \mathrm{d}d^* \\ &= \mathbb{E}[\alpha_w(D_i, X_i)D_i|X_i = x] - \bar{d}\mathbb{E}[\alpha_w(D_i, X_i)|X_i = x] \\ &= \mathbb{E}[\alpha_w(D_i, X_i)D_i|X_i = x] \\ &= 1. \end{aligned}$$

Part 3

By plugging in $\alpha_w(D_i, X_i) = \tilde{D}_i / \mathbb{E}[\tilde{D}_i D_i]$, we have

$$\lambda(d, x) = w(d, x) = -\frac{Cov(\tilde{D}_i, \mathbb{1}(D_i \leqslant d)|X_i = x)}{f(d|x)\mathbb{E}[\tilde{D}_i D_i]}$$

Under Assumption 3.2, $\mathbb{E}[\tilde{D}_i D_i] = \mathbb{E}[Var(D_i|Y_i(d), X_i)]$. This concludes the proof. □

Proof of Design-based Weights with Parametric Distributions. Consider a function $\tilde{f}(d|x) = \omega(d, x) f(d|x)$ where $\omega(d, x)$ is design-based weight.Since the weight is convex, it follows that $\tilde{f}(d|x)$ is also a density function. Under design-based specifcation, the Riesz

Representer for least square estimand is

$$\alpha(d,x) = \frac{d - \mathbb{E}[D_i|X_i = x]}{\mathbb{E}[Var(D_i|X_i)]}.$$

In this proof, it is convenient to focus on the Riesz Representer

$$\tilde{\alpha}(d,x) = \frac{d - \mathbb{E}[D_i|X_i = x]}{Var(D_i|X_i = x)}$$

which is proportional to α by a factor $Var(D_i|X_i = x)/\mathbb{E}[Var(D_i|X_i)]$. Let $\tilde{f}(d|x) = \tilde{\omega}(d,x)f(d|x)$ where $\tilde{\omega}(d,x)$ is the design-based weight associated with $\tilde{\alpha}(d|x)$. It is straightforward to show the least square weight $\omega(d,x) = \frac{Var(D_i|X_i=x)}{\mathbb{E}[Var(D_i|X_i)]}\tilde{\omega}(d,x)$.

According to the previous proof, we have

$$\tilde{\omega}(d,x) = \frac{-1}{f(d|x)} \int_{\underline{d}}^{d} \alpha_w(d^*,x) f(d^*|x) \mathrm{d}d^* = \frac{\tilde{f}(d|x)}{f(d|x)}.$$

Therefore, by plugging in $\tilde{\alpha}$, there is relationship

$$\tilde{f}(d|x) = \int_{\underline{d}}^{d} \frac{m(x) - d^*}{v(x)} f(d^*|x) \mathrm{d}d^*$$

where $m(x) = \mathbb{E}[D_i|X_i = x]$ and $v(x) = Var(D_i|X_i = x)$ are parametric mean and variance of D_i given $X_i = x$.

To derive the weight function with parametric $f(d|x)$, it suffices to find the function $\tilde{f}(d|x)$ such that its derivative with respect to d is $\frac{m(x)-d}{v(x)} f(d|x)$. Now, we provide 3 examples of parametric distributions.

Normal distribution: Let $f(d|m,v)$ denote normal density function with mean $m = m(x)$ and variance $v = v(x)$. Note that $\frac{m(x)-d}{v(x)} f(d|x)$ is exactly the derivative of normal density. Therefore, $\tilde{f}(d|x) = f(d|x)$ and $\tilde{\omega}(d,x) = 1$. The lease square weight is $\omega(d,x) = \frac{Var(D_i|X_i=x)}{\mathbb{E}[Var(D_i|X_i)]}$, which coincides with the design-based weights for binary treatment.

Gamma distribution: Let $f(d|a,b)$ denote gamma density function with shape parameter

$a = a(x)$ and rate parameter $b = b(x)$,

$$f(d|a,b) = \frac{b^a}{\Gamma(a)} d^{a-1} e^{-bd}.$$

Note that

$$f'(d|a+1,b) = \frac{m-d}{v} f(d|a,b).$$

Therefore,

$$\tilde{\omega}(d,x) = \frac{f(d|a+1,b)}{f(d|a,b)} = \frac{d}{m}$$

and the least square weight is

$$\omega(d,x) = \frac{Var(D_i|X_i = x)}{\mathbb{E}[Var(D_i|X_i)]} \frac{d}{m}$$

Beta distribution: Let $f(d|a,b)$ denote beta density function with shape parameters $a = a(x)$ and $b = b(x)$,

$$f(d|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} d^{a-1} (1-d)^{b-1}.$$

Note that

$$f'(d|a+1,b+1) = \frac{m-d}{v} f(d|a,b).$$

Therefore,

$$\tilde{\omega}(d,x) = \frac{f(d|a+1,b+1)}{f(d|a,b)} = \frac{d(1-d)}{m(1-m)-v}$$

and the least square weight is

$$\omega(d,x) = \frac{Var(D_i|X_i = x)}{\mathbb{E}[Var(D_i|X_i)]} \frac{d(1-d)}{m(1-m)-v}$$

□

Proof of Theorem 3.2. We start from $\sqrt{n}(\hat{\tau}_{wade} - \tau_{wade}) \overset{d}{\to} N(0,V)$ where $V = Var(\phi(Y_i, D_i, X_i))$. Note that $\mathbb{E}[w(D_i,X_i)\mu'(D_i,X_i)] = \tau_{wade}$, the variance reduces to

$$V = \mathbb{E}[\alpha_w^2(D_i,X_i)(Y_i - \mu(D_i,X_i))^2] = \mathbb{E}[\alpha_w^2(D_i,X_i)\sigma^2(D_i,X_i)]$$

where $\sigma^2(D_i, X_i) = Var(Y_i|D_i, X_i)$.

Let $\gamma \in H$ be a function with conditional mean $\bar{\gamma}(x) = \mathbb{E}[\gamma(D_i, X_i)|X_i = x]$. The function $\alpha(d, x) = \gamma(d, x) - \bar{\gamma}(x)$ satisfies the condition $\mathbb{E}[\alpha(D_i, X_i)|X_i] = 0$, and hence, $\alpha \in R$ if and only if $\mathbb{E}[\alpha(D_i, X_i)D_i] = \mathbb{E}[\alpha(D_i, X_i)(D_i - e(X_i))] = 1$. We therefore consider the Lagrangian

$$L(\gamma, \lambda) = \mathbb{E}[(\gamma(D_i, X_i) - \bar{\gamma}(X_i))^2\sigma^2(D_i, X_i) - 2\lambda((\gamma(D_i, X_i) - \bar{\gamma}(X_i))(D_i - e(X_i))) - 1]$$

where λ is the Lagrange multiplier. Consider replacing $\gamma(D_i, X_i)$ with $\gamma(D_i, X_i) + \delta\eta(D_i, X_i)$ with δ being a positive constant. Thus, we obtain

$$\frac{L(\gamma + \delta\eta, \lambda) - L(\gamma, \lambda)}{\delta} = 2\mathbb{E}[(\eta(D_i, X_i) - \bar{\eta}(X_i))((\gamma(D_i, X_i) - \bar{\gamma}(X_i))\sigma^2(D_i, X_i) - \lambda(D_i - e(X_i)))] + O(\delta).$$

where $\bar{\eta}(X_i) = \mathbb{E}[\eta(D_i, X_i)|X_i]$. Note that

$$\mathbb{E}[\bar{\eta}(X_i)((\gamma(D_i, X_i) - \bar{\gamma}(X_i))\sigma^2(D_i, X_i) - \lambda(D_i - e(X_i)))] = \mathbb{E}[\eta(D_i, X_i)Cov(\gamma(D_i, X_i), \sigma^2(D_i, X_i)|X_i)]$$

and let

$$l(d, x, \gamma, \lambda) = (\gamma(d, x) - \bar{\gamma}(x))\sigma^2(d, x) - \lambda(d - e(x)) - Cov(\gamma(D_i, X_i), \sigma^2(D_i, X_i)|X_i = x).$$

To write

$$\frac{L(\gamma + \delta\eta, \lambda) - L(\gamma, \lambda)}{\delta} = 2\mathbb{E}[\eta(D_i, X_i)l(D_i, X_i, \gamma, \lambda)] + O(\delta).$$

we require $l(d, x, \gamma, \lambda) = 0$, which gives

$$\alpha(D_i, X_i)\sigma^2(D_i, X_i) - \mathbb{E}[\alpha(D_i, X_i)\sigma^2(D_i, X_i)|X_i = x] = \lambda(D_i, X_i)(D_i - e(X_i))$$

by replacing$\alpha(d, x)$ with $\gamma(d, x) - \bar{\gamma}(x)$. The above equation holds by

$$\alpha(d, x) = \frac{\lambda(d - \tilde{e}(x))}{\sigma^2(d, x)}.$$

The constrain $\mathbb{E}[\alpha(D_i, X_i)|X_i] = 0$ gives

$$\tilde{e}(x) = \frac{\mathbb{E}[D_i/\sigma^2(D_i, X_i)|X_i = x]}{\mathbb{E}[1/\sigma^2(D_i, X_i)|X_i = x]}$$

and $\mathbb{E}[\alpha(D_i, X_i)D_i] = 1$ gives

$$\lambda = \mathbb{E}\left[\frac{D_i - \tilde{e}(X_i)D_i}{\sigma^2(D_i, X_i)}\right]^{-1}.$$

Therefore, the optimal WADE is

$$\frac{\mathbb{E}[Y_i(D_i - \tilde{e}(X_i))/\sigma^2(D_i, X_i)]}{\mathbb{E}[D_i(D_i - \tilde{e}(X_i))/\sigma^2(D_i, X_i)]}.$$

If Y_i is homoskedastic with $\sigma^2(D_i, X_i) = \sigma^2$, then the optimal WADE reduces to

$$\tau_{ls} = \frac{\mathbb{E}[\tilde{D}_i Y_i]}{\mathbb{E}[\tilde{D}_i D_i]}$$

which is the least square estimand with $\tilde{D}_i = D_i - e(X_i)$ under unconfoundedness assumption.

Theorem 3.2 also incorporates the binary treatment case. Note that for $D_i \in \{0, 1\}$,

$$\tilde{e}(x) = \frac{e(x)\sigma^2(0, x)}{e(x)\sigma^2(0, x) + (1 - e(x))\sigma^2(1, x)}$$

and hence, the optimal WATE is

$$\mathbb{E}\left[\frac{\tilde{w}(x)}{\mathbb{E}[\tilde{w}(X_i)]}\frac{(D_i - e(x))Y_i}{e(x)(1 - e(x))}\right], \text{ where } \tilde{w}(x) = \left(\frac{\sigma^2(1, x)}{e(x)} + \frac{\sigma^2(0, x)}{1 - e(x)}\right)^{-1}$$

which is Theorem 5.4 in Crump et al. (2006). Moreover, if Y_i is homoskedastic, the optimal WATE further reduces to $\tau_{overlap}$, which is Theorem 2 in Li et al. (2018). □

Proof of Proposition 4.1. We employ a common empirical process notation. Let $W_i = (Y_i, D_i, X_i)$, P and P_n be linear operators such that for some function $g(W_i)$, $P[g(W_i)] =$

$\mathbb{E}[g(W_i)]$ and $P_n[g(W_i)] = n^{-1}\sum_{i=1}^n g(W_i)$. Further define

$$\begin{aligned}
\gamma(P) &= \mathbb{E}[(Y_i - \rho(X_i)(D_i - e(X_i))] = \gamma \\
\gamma(\hat{P}) &= \mathbb{E}[(Y_i - \hat{\rho}(X_i)(D_i - \hat{e}(X_i))] \\
\eta(P) &= \mathbb{E}[(D_i - e(X_i))^2] = \eta \\
\eta(\hat{P}) &= \mathbb{E}[(D_i - \hat{e}(X_i))^2] \\
\phi_\gamma(W_i; P) &= (Y_i - \rho(X_i)(D_i - e(X_i)) - \gamma(P) \\
\phi_\gamma(W_i; \hat{P}) &= (Y_i - \hat{\rho}(X_i)(D_i - \hat{e}(X_i)) - \gamma(\hat{P}) \\
\phi_\eta(W_i; P) &= (D_i - e(X_i))^2 - \eta(P) \\
\phi_\eta(W_i; \hat{P}) &= (D_i - \hat{e}(X_i))^2 - \eta(\hat{P}) \\
\hat{\gamma} &= \gamma(\hat{P}) + P_n[\phi_\gamma(W_i; \hat{P})] \\
\hat{\eta} &= \eta(\hat{P}) + P_n[\phi_\eta(W_i; \hat{P})].
\end{aligned}$$

Our proving strategy is to consider $\hat{\tau}_{ls} = \hat{\gamma}/\hat{\eta}$ with the influence function $\phi_{ls}(w; P) = (\phi_\gamma(w; P) - \tau_{ls}\phi_\eta(w; P))/\eta(P)$. We show that under the assumptions in Proposition 4.1, $\hat{\gamma}$ and $\hat{\eta}$ are RAL in the sense that

$$\begin{aligned}
\hat{\gamma} - \gamma &= \frac{1}{n}\sum_{i=1}^n \phi_\gamma(W_i; P) + o_P(n^{-1/2}) \\
\hat{\eta} - \eta &= \frac{1}{n}\sum_{i=1}^n \phi_\eta(W_i; P) + o_P(n^{-1/2}).
\end{aligned}$$

Then it follows from the Slutsky's theorem that

$$\sqrt{n}(\hat{\tau}_{ls} - \tau_{ls}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \phi_{ls}(W_i; P) + o_P(1)$$

which is the desired result.

Consider $\hat{\gamma}$ first. By definition, we have

$$\begin{aligned}
\hat{\gamma} - \gamma &= \gamma(\hat{P}) - \gamma(P) + P_n[\phi_\gamma(W_i;\hat{P})] \\
&= (P_n - P)\phi_\gamma(W_i;\hat{P}) + R_n \\
&= (P_n - P)\phi_\gamma(W_i;P) + (P_n - P)[\phi_\gamma(W_i;\hat{P}) - \phi_\gamma(W_i;P)] + R_n \\
&= (P_n - P)\phi_\gamma(W_i;P) + E_n + R_n
\end{aligned}$$

where

$$\begin{aligned}
E_n &= (P_n - P)[\phi_\gamma(W_i;\hat{P}) - \phi_\gamma(W_i;P)] \\
R_n &= \gamma(\hat{P}) - \gamma(P) + P[\hat{\phi}_\gamma(W_i;\hat{P})].
\end{aligned}$$

Now the task is to show that the empirical process term $E_n = o_P(n^{-1/2})$ and the remainder term $R_n = o_P(n^{-1/2})$.

The empirical process term

$$\begin{aligned}
E_n &= (P_n - P)[\phi_\gamma(W_i;\hat{P}) - \phi_\gamma(W_i;P)] \\
&= (P_n - P)[\gamma(\hat{P}) - \gamma(P)] \\
&+ (P_n - P)[(Y_i - \rho(X_i))(e(X_i) - \hat{e}(X_i)] \\
&+ (P_n - P)[(D_i - e(X_i))(\rho(X_i) - \hat{\rho}(X_i))] \\
&+ (P_n - P)[(\rho(X_i) - \hat{\rho}(X_i))(e(X_i) - \hat{e}(X_i))].
\end{aligned}$$

where the first term $(P_n - P)[\gamma(\hat{P}) - \gamma(P)] = (\gamma(\hat{P}) - \gamma(P))(P_n - P)[1] = 0$.

Note that, under Assumption a and c, we have the following conditions:

$$\begin{aligned}
\mathbb{E}[(Y_i - \rho(X_i))^2(e(X_i) - \hat{e}(X_i))^2] = \mathbb{E}[Var(Y_i|X_i)(e(X_i) - \hat{e}(X_i))^2] = o_P(1) \\
\mathbb{E}[(D_i - e(X_i))^2(\rho(X_i) - \hat{\rho}(X_i))^2] = \mathbb{E}[Var(D_i|X_i)(\rho(X_i) - \hat{\rho}(X_i))^2] = o_P(1).
\end{aligned}$$

Lastly,

$$\mathbb{E}[(\hat{\rho}(X_i) - \rho(X_i))^2(\hat{e}(X_i) - e(X_i))^2] \leqslant \mathbb{E}[(\hat{\rho}(X_i) - \rho(X_i))^4]^{1/2}\mathbb{E}[(\hat{e}(X_i) - e(X_i))^4]^{1/2} = o_P(1)$$

which follows from Holder's inequality and Assumption a.

Under Donsker condition and Lemma 19.24 of Van der Vaart (1998), or under cross-fitting condition and Lemma 2 of Kennedy et al. (2020), the above conditions imply that all remaining terms in E_n are $o_P(n^{-1/2})$, which means $E_n = o_P(n^{-1/2})$.

The remainder term

Evaluating the remainder term gives

$$\begin{aligned}
R_n &= \mathbb{E}[\gamma(\widehat{P}) - \gamma(P) + \widehat{\phi}_\gamma(W_i; \widehat{P})] \\
&= \mathbb{E}[(Y_i - \widehat{\rho}(X_i))(D_i - \widehat{e}(X_i)) - (Y_i - \rho(X_i))(D_i - e(X_i))] \\
&= \mathbb{E}[(\rho(X_i) - \widehat{\rho}(X_i))(e(X_i) - \widehat{e}(X_i))].
\end{aligned}$$

Therefore,

$$R_n^2 \leqslant \mathbb{E}[(\rho(X_i) - \widehat{\rho}(X_i))^2]\mathbb{E}[(e(X_i) - \widehat{e}(X_i))^2] = o_P(n^{-1}).$$

by Cauchy-Schwarz inequality and Assumption a, which implies that R_n is $o_P(n^{-1/2})$.

Therefore, we have shown that $\widehat{\gamma} - \gamma = \frac{1}{n}\sum_{i=1}^{n} \phi_\gamma(W_i; P) + o_P(n^{-1/2})$. The proof for $\widehat{\eta}$ is similar by considering the special case of $Y_i = D_i$. This concludes the proof.

Note 1: Assumption a is used to ensure that the remainder term R_n in $\widehat{\gamma}$ and $\widehat{\eta}$ is $o_P(n^{-1/2})$, and this is why the convergence rate of $\widehat{\rho}$ can be small as long as $\widehat{e}$ converges at a sufficiently fast rate.

Note 2: The cross-fitting condition requires two independent samples to construct $\widehat{\tau}_{ls}$ and estimate $\widehat{\gamma}$ and $\widehat{\eta}$. A more efficient algorithm proposed by Chernozhukov et al. (2018) is to split the sample into K folds, use the kth fold for constructing $\widehat{\tau}_{ls}$ and the remaining $K-1$ folds for estimating $\widehat{\gamma}$ and $\widehat{\eta}$. The final estimate is the average of the K estimates. Using this algorithm, the empirical process term and the remainder term becomes

$$\sum_{k=1}^{K} \frac{n_k}{n}(R_{nk} + E_{nk})$$

where n_k is the number of observations in the kth fold, and R_{nk} and E_{nk} evaluated at $P_{nk}(\cdot) = \frac{1}{n_k}\sum_{i=1}^{n_k}(\cdot)$. The proof remains similar. □