A Response to Recent Critiques of Hainmueller, Mummolo and Xu (2019) on Estimating Conditional Relationships[*]

Jens Hainmueller   Jiehan Liu   Ziyi Liu   Jonathan Mummolo
(Stanford)   (Stanford)   (Berkeley)   (Princeton)

Yiqing Xu
(Stanford)

February 11, 2025

Abstract

Simonsohn (2024a) and Simonsohn (2024b) critique Hainmueller, Mummolo and Xu (2019, HMX), arguing that failing to model nonlinear relationships between the treatment and moderator leads to biased marginal effect estimates and uncontrolled Type-I error rates. While these critiques highlight the issue of under-modeling nonlinearity in applied research, they are fundamentally flawed in several key ways. First, the causal estimand for interaction effects and the necessary identifying assumptions are not clearly defined in these critiques. Once properly stated, the critiques no longer hold. Second, the kernel estimator HMX proposes recovers the true causal effects in the scenarios presented in these recent critiques, which compared effects to the wrong benchmark, producing misleading conclusions. Third, while Generalized Additive Models (GAM) can be a useful exploratory tool (as acknowledged in HMX), they are not designed to estimate marginal effects, and better alternatives exist, particularly in the presence of additional covariates. Our response aims to clarify these misconceptions and provide updated recommendations for researchers studying interaction effects through the estimation of conditional marginal effects.

Keywords: conditional relationships, interaction model, marginal effect, GAM, double/debiased machine learning

---

A recent blog post (Simonsohn, 2024a) and accompanying paper (Simonsohn, 2024b) critiques a paper published in *Political Analysis* titled "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice" (Hainmueller, Mummolo and Xu, 2019, hereafter HMX 2019). These critiques argue that the binning and kernel estimators proposed in HMX (2019) are inappropriate for estimating and testing conditional relationships using observational data.

We, the authors of HMX (2019), have appreciated these critiques and shared their initial response in private. However, since some important substantive points were not reflected in the blog post, we (HMX and two additional contributors) have decided to provide the following public response. We believe this will help clarify the issues raised and contribute to improving research practices.

Critiques of HMX (2019)

Let Y, D, X, and Z denote the outcome, treatment, moderator, and a set of additional covariates, respectively. Empirical researchers are often interested in how the effect of treatment D on outcome Y varies with a covariate X, which we call the moderator. There may exist other covariates Z, and we denote $V = (X, Z)$. While HMX (2019) also considers panel settings, for simplicity, we focus here on a cross-sectional setting where (Y_i, D_i, V_i) are i.i.d. from a super-population. Traditionally, researchers often estimate the following linear interaction model:

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 DX + \gamma Z + \epsilon, \tag{1}$$

where ϵ represents an idiosyncratic error, and interpret $\widehat{ME}(x) = \hat{\beta}_1 + \hat{\beta}_3 x$ as the conditional marginal effect of D on Y with respect to X (Brambor, Clark and Golder, 2006).

HMX (2019) highlights that Model (1) is often unrealistic because this functional form implies that the effect of D on Y changes linearly with X, and the overlap assumption

is frequently violated. To address these challenges, HMX (2019) recommends (i) visual diagnostics such as subgroup scatterplots, (ii) the **Generalized Additive Models (GAM) plot**, and (iii) the binning estimator, as diagnostic tools. Additionally, HMX (2019) proposes a semiparametric kernel estimator, implemented via local linear regressions, to relax the functional form assumption:

$$Y = f(X) + g(X)D + \gamma(X)Z + \epsilon. \tag{2}$$

Hence, the conditional marginal effects are estimated with a more flexible functional form which can vary freely across levels of X: $\widehat{ME}(x) = \hat{g}(x)$.

The primary concern in Simonsohn (2024a) and Simonsohn (2024b) is the misspecification bias that arises from nonlinear relationships between D, X, and Y, particularly when D and X are correlated. In these analyses—both in the paper and the blog—no additional covariates Z are included. As noted in Simonsohn (2024b), this bias increases the likelihood of detecting spurious interactions, leading to an elevated false positive rate. To account for nonlinearity, GAMs of the following form are recommended:

$$Y = f(D, X) + \epsilon.^1$$

Simonsohn (2024a) provides an intriguing simulated sample with the following data-generating process (DGP): $Y = D^2 - 0.5D + \epsilon$, where X and D are correlated: $Corr(X, D) = 0.5$. The author claims that the marginal effect of D on Y conditional on X is zero, i.e., that $ME(x) = \partial Y / \partial D = 0$, while the linear estimator, the binning estimator, and the kernel estimator all produce marginal effect estimators that are linearly and positively correlated with x (Figure 1). Based on the shared `R` code, we believe that the estimation actually

---

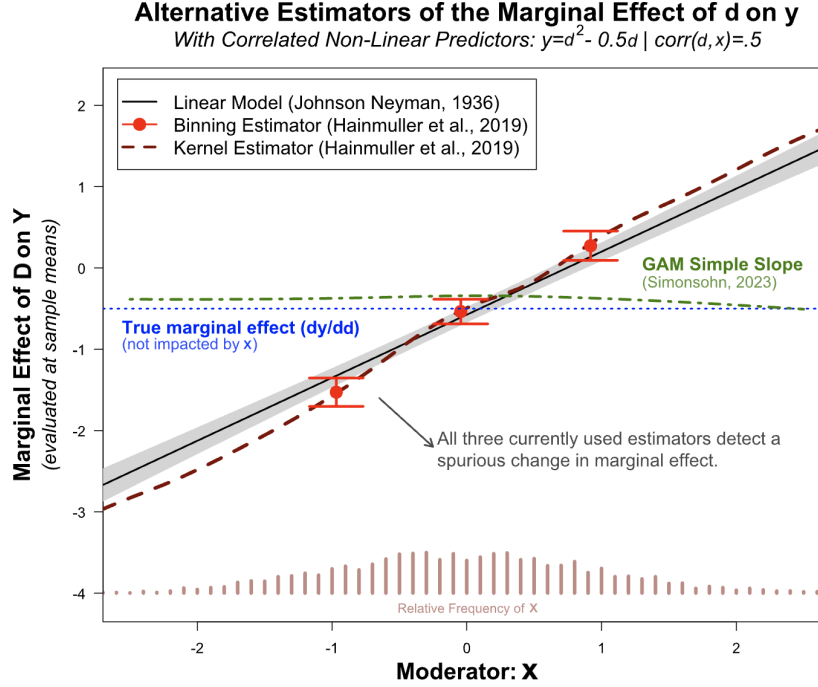[1]Specifically, Simonsohn (2024b) advocates for a model incorporating three smooth functions:

$$Y = f_1(D) + f_2(X) + f_3(D, X) + \epsilon.$$

Each function is estimated with a separate smoothing (penalty) parameter, and $f_3(D, X)$ is intended to capture the interaction effect. It is implemented using `gam(y ~ s(x1) + s(x2) + ti(x1, x2), data = foo)` with the `mgcv` package in `R`.

targets $\mathrm{E}\left[\frac{\partial Y}{\partial D}\big|\ D = 0, X\right]$.[2] However, this approach does *not* align with the conventional definition of marginal effects (see further discussion below).

FIGURE 1. FIGURE ADAPTED FROM SIMONSOHN $(2024a)$
(WITH MODIFIED LABELS)

Alternative Estimators of the Marginal Effect of d on y
With Correlated Non-Linear Predictors: y=d²- 0.5d | corr(d,x)=.5

Note: In the original plot, the treatment and moderator were labeled as X and Z, respectively; we have modified them to D and X to maintain consistency with the notation used in this response.

Merits of the Critique. There is value in these critiques. First, they bring renewed attention to the issue of potential nonlinearity in estimating and testing conditional relationships. Although the kernel estimator can recover the intended estimand (as we will show below), it requires a large amount of data when nonlinear terms of D are present,

---

[2]The shared code is provided in the footnote of Figure 3 in Simonsohn $(2024a)$. The procedure begins by fitting a GAM: `gam(y ∼ s(D) + s(X) + ti(D, X), data = foo)`. Then, the predicted values of Y are computed over a range of X values while holding D fixed at 0 and 0.1. The "marginal effects" $\partial Y / \partial D$ are then estimated using a numerical approximation-specifically, by taking the difference between the predicted Y values at $D = 0.1$ and $D = 0$, and dividing by 0.1. A similar procedure for GAM simple slopes is advocated in Simonsohn $(2024b)$.

highlighting the advantages of modern causal inference methods. Second, these critiques present thought-provoking examples that challenge existing methods. Although they are fundamentally flawed, they demonstrate the importance of clearly defining estimands and often-unstated identifying assumptions.

Summary of our response. We address these critiques with three main points:

1. The most critical flaw in the critique is the lack of a clear definition of the estimand (and identifying assumptions). The critique focuses solely on prediction (using GAM) while simultaneously using strong causal language, such as "effect" or "marginal effect." We will show that once these elements are clearly defined, the critique no longer holds—in fact, the GAM output does not correspond to any marginal effect as the author claims.

2. Once the estimand, the conditional marginal effect (CME), and key assumptions are clearly stated, existing methods can effectively address the nonlinearity between D and X. These methods include the kernel estimator (with sufficient data and an appropriately small bandwidth), augmented inverse propensity-score weighting (AIPW), post-double-selection LASSO (PDS-LASSO), and double/debiased machine learning (DML) estimators. AIPW, PDS-LASSO, and DML are doubly robust, making them more resilient to model misspecification in applied settings.

3. The method advocated in Simonsohn (2024a) and Simonsohn (2024b), GAM, is primarily a predictive tool and is not well-suited for estimating the CME. While it can be adapted for causal inference, it relies heavily on researchers' prior knowledge of where nonlinearity may occur—a requirement that is often unrealistic in applied settings, especially when dealing with high-dimensional covariates Z. This limitation makes GAM susceptible to both regularization and misspecification biases. Although GAM can serve as a useful exploratory tool, it is not designed for estimating causal quantities such as the CME.

Properly Defining the Estimand Is Critical

First, we highlight the critical importance of clearly defining the causal estimand, which is often neglected in social science research (Lundberg, Johnson and Stewart, 2021). HMX (2019), targeting an audience familiar with the regression framework, does not formally adopt modern causal inference notation; we address this limitation in this response. We show that once the estimand, the CME, is properly specified, the critiques in Simonsohn (2024*a,b*) no longer hold. To illustrate this, we revisit the key simulated example from Simonsohn (2024*a*), demonstrating that its misleading conclusion stems from a misunderstanding of the estimand.

We begin by defining the relevant estimands and outlining the key identifying assumptions within the Neyman-Rubin potential outcomes framework (Neyman, 1923; Rubin, 1974).

Assumption 1. (SUTVA). Potential outcome of unit i can be written as:

$$Y_i(d_1, d_2, \cdots, d_n) = Y_i(d_i)$$

in which d_i denote the potential treatment value for unit i. Therefore, $Y_i = Y(D_i)$, which connects the observed outcome with potential outcomes.

Given SUTVA, we define the main estimand, the causal marginal effect (CME):

$$\theta(x) = \mathbb{E}\left[\left.\frac{\partial Y_i(d)}{\partial d}\right|X_i = x\right].$$

It is *causal* because it represents an aggregation of comparisons between potential outcomes for the same unit. When D_i is binary, it simplifies to

$$\theta(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x].$$

which is essentially conditional average treatment effect (CATE) at covariate value v, $\tau(v) = \mathbb{E}[Y_i(1) - Y_i(0) \mid V_i = v]$, *marginalized* over the distribution of additional covariates Z_i.

It is important to note that the CME does *not* imply that the CATE changes with X in a causal manner. The concept of *causal moderation* (also known as causal interaction), by contrast, refers to the causal effect of X on the treatment effect of D on Y (VanderWeele, 2009; Bansak, 2020). To define causal moderation, we need to extend the potential outcome framework to include x explicitly and define causal moderation at x as:

$$\delta(x) = \mathbb{E}\left[\frac{\partial^2 Y_i(d, x)}{\partial d \partial x}\bigg| X_i = x\right],$$

where the expectation is taken over D and Z. Identifying $\delta(x)$ requires much stronger assumptions, including the quasi-randomization of X, which are far more stringent than those needed for identifying the CATE or CME.

Another estimand, often conflated with the CME, is the *conditional average partial effect* (CAPE):

$$\rho(d, x, z) = \mathbb{E}\left[\frac{\partial Y_i(d)}{\partial d}\bigg| D_i = d, X_i = x, Z_i = z\right],$$

which represents the average partial effect of D on Y given specific values of D, X, and Z. When D is binary, it simplifies to $\mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = d, V_i = v], d = \{0, 1\}$, making it a special case of the CATE. This estimand differs from the conventional interpretation of marginal effects in that (i) the influence of Z is not marginalized; (ii) it is conditional on a specific d. Therefore, we rarely see this estimand used in empirical work. However, as we will see below, this may be the implicit estimand in Simonsohn (2024a).

Table 1 summarizes these estimands, each distinct yet closely related. CATE measures the average effect of a change in treatment D on the outcome Y, among units characterized by $(X_i = x, Z_i = z)$. This estimand is typically explored when researchers are interested in the heterogeneity of treatment effect. The CAPE is a special case of CATE when a specific treatment level $D_i = d$ is being conditioned on, which lacks practical justification in continuous treatment settings. CME measures how the effect of treatment D on outcome Y varies across units with characteristic X, independent of other characteristics Z or the

TABLE 1. SUMMARY OF ESTIMANDS

Estimand	Definition	Interpretation
Conditional average treatment effect (CATE)	$\mathbb{E}\left[\frac{\partial Y_i(d)}{\partial d}\Big| X_i = x, Z_i = z\right]$	The average effect of a treatment on a subgroup defined by $(X_i = x, Z_i = z)$, marginalized over different values of D
Conditional average partial effect (CAPE)	$\mathbb{E}\left[\frac{\partial Y_i(d)}{\partial d}\Big| D_i = d, X_i = x, Z_i = z\right]$	CATE conditional on a specific value of D (rarely used in continuous D settings)
Conditional marginal effect (CME)	$\mathbb{E}\left[\frac{\partial Y_i(d)}{\partial d}\Big| X_i = x\right]$	CATE marginalized over D and Z (not including X).
Causal moderation	$\mathbb{E}\left[\frac{\partial^2 Y_i(d,x)}{\partial d \partial x}\Big| X_i = x\right]$	Also known as causal interaction effect; marginalized over D and Z.

treatment level. In applied contexts, CME is frequently referred to as an interaction effect or conditional marginal effect, highlighting how the treatment effect varies across levels of X, the moderator, while disregarding other covariates. Lastly, causal moderation examines the second derivative with respect to both $D_i = d$ and $X_i = x$. Instead of asking how the effect of D on Y changes with the value of X *descriptively*, causal moderation examines how X *causally* impacts the effect of D on Y; therefore, its identification typically requires (quasi-) randomization of X.

Conventionally, researchers use linear interaction models to estimate the CME, which captures how the causal effect of D on Y varies with the value of X (marginalizing over the distribution of D and Z at that value of X, hence the term "marginal effects"). However, recent critiques of HMX (2019) lack clarity regarding the estimand, causing confusion. These critiques may potentially target either the CAPE (according to the shared `R` code) or causal moderation (according to the narrative which focuses on interaction effect). The former being an unconventional estimand and the latter requiring very strong identifying assumptions, such as (quasi-)randomization of both D and X.

The key simulated example. Now, we examine the key simulated example through the lens of the CME. Consistent with the notation in HMX (2019), D represents the treatment, and X denotes the moderator. Note that Simonsohn (2024a) uses X and Z to represent the treatment and moderator, respectively. From the simulation setup, we know the data-generating process (DGP) is as follows:

$$Y = D^2 - 0.5\,D + \varepsilon,$$

in which $\mathbb{E}[\varepsilon \mid D, X] = 0$ and

$$(D, X) \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right).$$

The normality assumption gives us closed-form expressions for $\mathbb{E}[D \mid X = x]$, a key component for deriving the true CME. In Appendix Section A.1, we show that the true CME is

$$\theta(x) = \mathbb{E}\left[\frac{\partial Y}{\partial D} \,\Big|\, X = x\right] = x - 0.5.$$

That is, if we hold $X = x$ fixed, the average slope of Y with respect to D is $(x - 0.5)$. Unlike what Simonsohn (2024a) implies, the true CME is *not* zero and increases monotonically with x. The intuition is that as X increases, the distribution of D also shifts upward. Because the partial effect of D on Y is increasing in D, the estimated marginal effect also increases, even though X does not directly enter the outcome equation.

In Appendix Section A.1, we prove that the linear estimator is asymptotically biased for the CME, whereas the kernel estimator proposed in HMX (2019) can accurately recover the CME in this simulated example. The intuition is that while the effect of D on Y is nonlinear in the full sample, within the neighborhood of $X = x_0$, $\theta(x_0)$ can be well approximated by a linear (Taylor) expansion. The kernel estimator, using local linear regression, effectively captures this approximation when there are abundant data near $X = x_0$. The binning esti-

mator will be biased but will have less bias than the linear estimator, as it is a (rectangular) kernel estimator with a large bandwidth (bin size).

Interestingly, these results are presented in Simonsohn (2024a) (see Figure 1); however, Simonsohn (2024a) misinterprets them as evidence that the kernel (and binning) estimators are inherently biased. This misinterpretation arises because he conflates the CME, $\theta(x) = \mathbb{E}\left[\frac{\partial Y_i(d)}{\partial d}\big| X_i = x\right]$, with the CAPE, $\rho(d, x) = \mathbb{E}\left[\frac{\partial Y_i(d)}{\partial d} \mid D_i = d, X_i = x\right]$. We replicate the results in Figure 2, where the red line represents the true CME. As expected, the left panel shows that the linear estimator is biased, whereas the right panel demonstrates that, far from being "false," the kernel estimator can recover the true CME fairly well in finite samples ($n = 5,000$), even when Y depends on D in a quadratic form and X and D are correlated. Specifically, the true CME are within the 95% uniform confidence intervals from the kernel estimator.

As for the binning estimator, HMX (2019) proposes it as a diagnostic tool and develops a test based on it for the null hypothesis that the CME does not change with X. Of course, this is a crude way to approximate the CME, as few researchers would actually assume that the CME is piecewise linear and only varies by terciles of X. Because the binning estimator uses data beyond a small neighborhood of the evaluation point x_0 when estimating the CME, it is likely biased if the relationship between D and Y is nonlinear at x_0. Nevertheless, as shown in Figure 1, the estimates produced by the binning estimator remain fairly close to the true CME.

The same logic of linear approximation extends to more complex settings, where both D and X enter the outcome equation. For example, consider the following DGP, which is used in Simonsohn (2024b): $Y = f(D, X) + \epsilon$. When $f(D, X)$ is a smooth function (i.e., twice continuously differentiable), the kernel estimator can still consistently estimate the CME in regions of X where data are sufficiently abundant. We provide a simple proof in the appendix. Although we do not have access to the data to definitively confirm, we believe

FIGURE 2. ESTIMATED CME ON THE SIMULATED EXAMPLE

(a) Linear and Binning Estimators   **(b) Kernel Estimator**
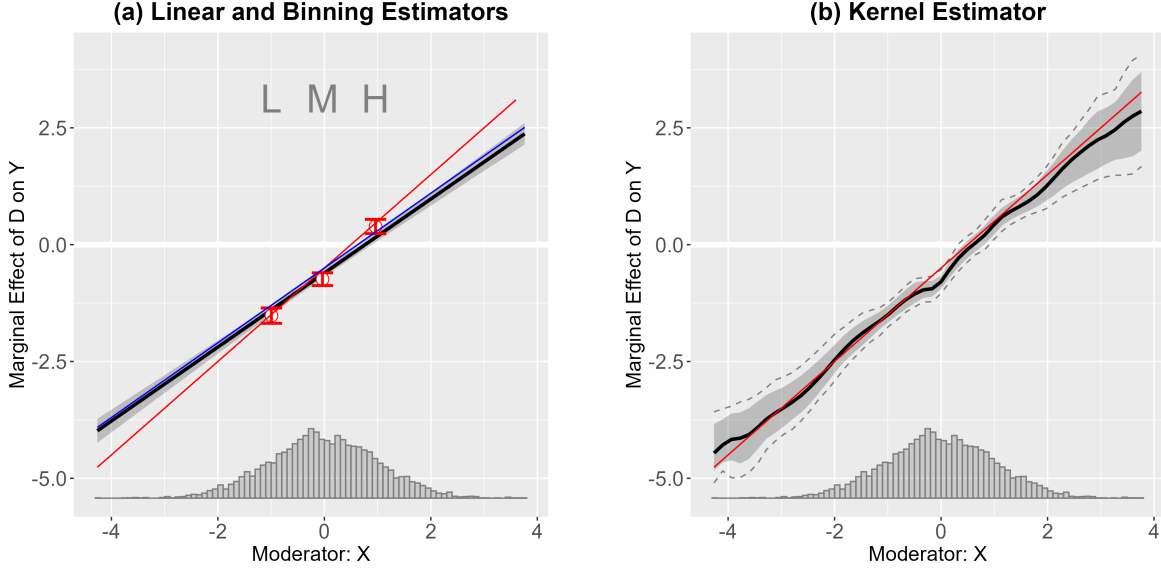
Note: The true CME is marked in red. The shaded area and dashed lines represent 95% pointwise confidence intervals and 95% uniform confidence intervals, respectively.

that in most, if not all, examples in Simonsohn (2024b), the kernel estimator can accurately recover the CME.

Summary   Contrary to the claim in Simonsohn (2024a) that "[A]ll three currently used estimators detect a spurious change in marginal effect," we demonstrate that the monotonically increasing marginal effect (or the CME) is not spurious, and the kernel estimator accurately recovers the true CME, and the binning estimator remains a useful diagnostic tool, even with more complex DGP. The confusion arises precisely because the causal estimand, CME, is not clearly defined in Simonsohn (2024a) and is potentially conflated with either the CAPE or causal moderation. The claim that the HMX estimators fail to recover the correct answer is incorrect because he used the wrong benchmark. The kernel estimator was designed to identify the CME; when evaluated against the correct estimand, it produces the correct result, whereas the GAM approach he proposed does not (see below).

Advancements in Methods for Estimating the CME

Simonsohn (2024a) advocates for using GAM to estimate potentially nonlinear conditional relationships. While GAM is an important and easy-to-use exploratory tool, we do not consider it an appealing estimation strategy for the CME in light of recent advancements in the causal inference literature. These advancements include (i) a shift in focus toward treatment assignment mechanisms (or "designs") and (ii) the development of robust estimation methods for targeted parameters with relaxed functional form assumptions, such as the introduction of double/debiased machine learning (DML) techniques (e.g., Van der Laan and Rose, 2011; Chernozhukov et al., 2017; Athey et al., 2019). See also Imbens and Xu (2024) for a recent survey of methods based on the unconfoundedness assumption.

We will explain the shortcomings of GAM in the next section. In this section, we briefly explain these advancements in the cross-sectional setting under the unconfoundedness and overlap assumptions.[3] We also illustrate the advantages of the proposed estimation strategies. First, we introduce two key identifying assumptions.

Assumption 2. (Unconfoundedness).

$$\{Y_i(d_i)\} \perp\!\!\!\perp D_i \mid V_i = v, \text{ for all } v \in \mathcal{V}.$$

Assumption 3. (Strict overlap)

$$f_{D|V}(d \mid v) > 0 \quad \text{for all } d \in \mathcal{D} \text{ and } v \in \mathcal{V}$$

where $\mathcal{D}$ is the support of the treatment D, $\mathcal{V}$ is the support of the covariates V, and $f_{D|V}(d \mid v)$ is the conditional probability density function of D given V. When D is binary, it simplifies to: For some positive η,

$$\eta \leq \mathbb{P}(W_i = 1 \mid V_i = v) \leq 1 - \eta, \quad \text{with probability 1.}$$

---

[3]More detailed discussions will be provided in *Estimating and Interpreting Conditional Marginal Effects: A Modern Approach* by Liu, Liu and Xu (N.d.) currently under contract with the Cambridge University Press.

It means that, given the covariate values $V_i = v$, the probability of receiving the treatment is strictly bounded away from 0 or 1.

Under Assumptions 1-3, the CME can be nonparametrically identified when D, X, and Z are discrete. However, when some of these variables are continuous or when Z is high-dimensional, additional dimensional reduction tools are required. These tools may include specifying an outcome model or a model for the propensity score to facilitate identification.

Assumption 4. (Partially linear model, PLM)

$$Y = g(V)D + f(V) + \epsilon$$

$$\mathbb{E}[D \mid V] = e(V)$$

in which $g(.)$ and $f(.)$ are flexible functions.

Note that when D is binary, $e(V)$ is the propensity score, and $\mathbb{E}[\epsilon \mid V] = 0$ when unconfoundedness hold.

DML and doubly robust estimators. Semenova and Chernozhukov (2021) show that under Assumptions 1-4 and regularity conditions—for example, if $g(.)$ and $f(.)$ can be consistently estimated with a sufficiently fast convergence rate—the DML estimators, which satisfy the so-called Neyman orthogonality condition, can recover the CME and achieve consistency asymptotically.

The primary strength of DML estimators lies in their robustness to model misspecification, as they rely on flexible machine learning algorithms to estimate nuisance functions, e.g., $f(V)$ and $e(V)$. This flexibility enables them to handle high-dimensional covariates and complex relationships between variables. Moreover, the use of Neyman orthogonality conditions ensures that small errors in estimating these nuisance functions do not propagate to the final estimate of the CME, thereby enhancing robustness.

The PLM is general. It is straightforward to see that Model (1), the traditional parametric linear interaction model, and Model (2), the semiparametric kernel estimator proposed in HMX (2019), are both special cases of the PLM. However, both models are limited by their restrictions on functional form. While the kernel estimator significantly relaxes the functional form assumptions of Model (1), it cannot accommodate complex, nonlinear relationships between Z and Y or between Z and D. In other words, jointly modeling a complex response surface of Y given D and multiple Z using the kernel estimator is either computationally infeasible or leads to significant regularization bias, affecting the estimation of target parameters in finite samples. The starting point of HMX (2019) is the assumption that the relationship between Z and Y is correctly specified, which allow us to focus on the nonlinearity of the CME and overlap issues. The PLM relaxes this assumption.

The main limitations of DML estimators are twofold: (i) they are computationally intensive due to the need for cross-fitting and fine-tuning machine learning algorithms; and (ii) they require large sample sizes for reliable performance, as their effectiveness depends on accurately estimating nuisance functions. When data are scarce, DML estimators may not perform well. In such cases, doubly robust estimators based on parametric or semiparametric models, such as augmented inverse propensity score weighting (AIPW) (Robins and Ritov, 1997) and post-double-selection LASSO (PDS-LASSO) (Belloni, Chernozhukov and Hansen, 2014), offer reliable alternatives. Blackwell and Olson (2022) provide an implementation of PDS-LASSO for estimating interaction effects, focusing on missing interactions between X and Z as well as D and Z. The so-called *fully interacted model* is also implemented in the `interflex` package. Building on Simonsohn's critique, we can extend the model by incorporating greater flexibility in modeling potential nonlinearity D and V (including X).

In Figure 3, we illustrate the performance of various estimators using a simulated example where the treatment is binary. The DGP is designed so that the moderator X and two covariates, Z_1 and Z_2, enter the treatment assignment and outcome equations in complex

FIGURE 3. ESTIMATING THE CME WITH VARIOUS ESTIMATORS
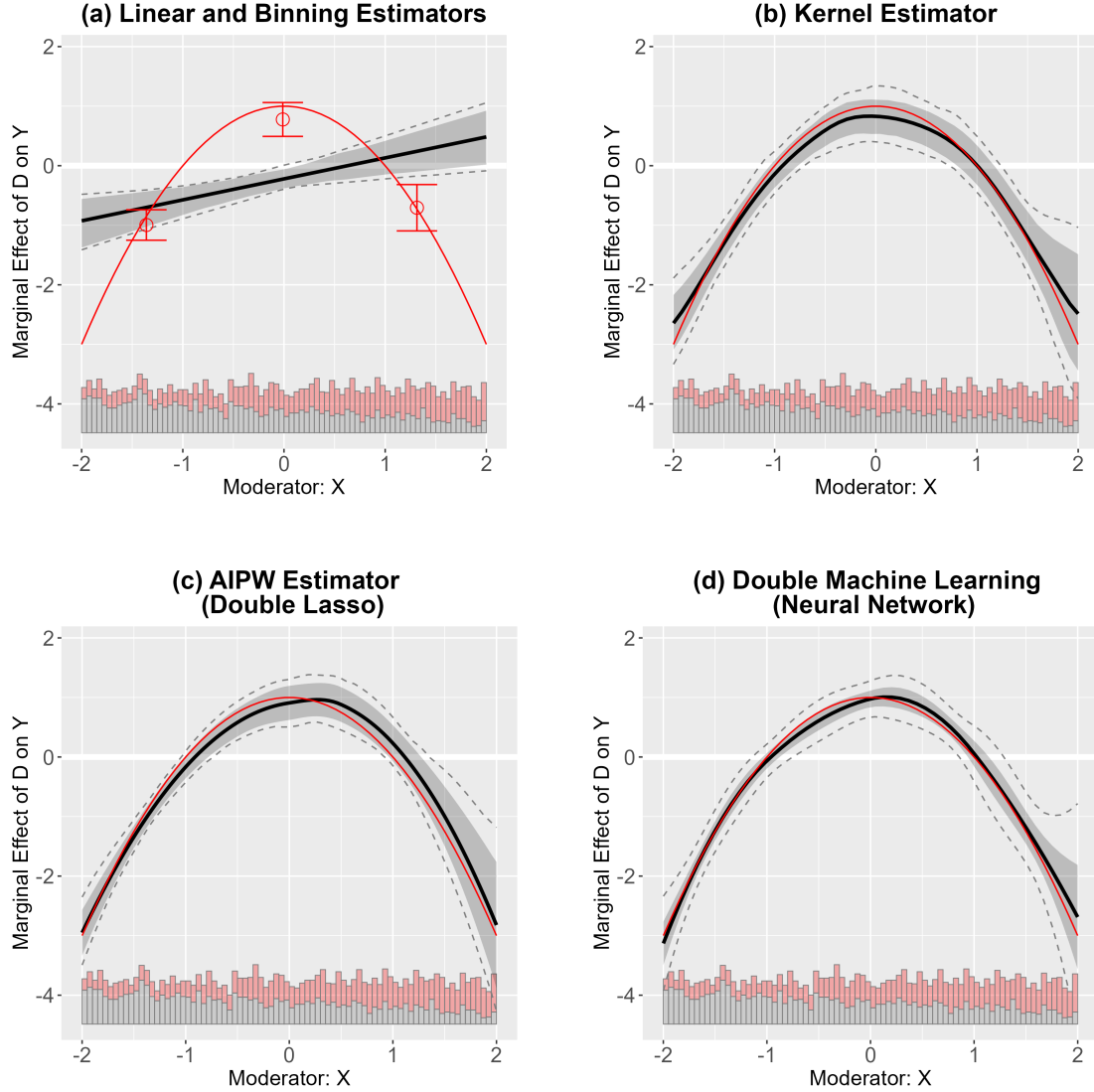
Note: In each subfigure, the red solid line and the black solid line represent the true and estimated CME, respectively. The shaded gray areas and the gray dashed lines depict the point-wise and uniform confidence intervals, respectively. At the bottom of each subplot, a histogram displays the distribution of treated (red) and control (gray) units across varying values of the moderator X. The sample size if 5,000. The moderator X and additional covariates Z_1 and Z_2 enter the outcome model and X and Z_1 enter the propensity score. The exact DGP can be found in Appendix Section A.2.

ways. The figure displays results for the linear and binning estimators (top-left), the kernel estimator (top-right), the AIPW-LASSO estimator (bottom-left), and the DML-NeuralNet

estimator (bottom-right). Figure 3 shows that the kernel, AIPW-LASSO, and DML estimators recover the CME reliably, while the binning estimator remains a useful diagnostic tool. Notably, the AIPW-LASSO estimator appears to be the most efficient in this setting. We do not include GAM because it is highly infeasible, and we detail the reasons in the next section.

GAM is Not an Appealing Alternative

GAM has significant drawbacks compared to the kernel estimator, as well DML and other doubly robust estimators specifically designed for estimating the CME. The key issue is that, as a method for estimating the response surface, i.e., the conditional expecation of the *outcome*, GAM does not explicitly model the CME. In contrast, kernel estimators based on local linear regression approximate derivatives more directly. In practice, Simonsohn (2024a) proposes the following procedure: (i) fit a GAM to model $Y = f(X, D) + \epsilon$; (ii) estimate "marginal effects" at $D = d_0$ using the difference in predicted values from the estimated model $\hat{f}$:

$$\hat{\rho}(d_0, x) = \mathbb{E}\left[\frac{\partial Y_i(d)}{\partial d} \mid D_i = d_0, X_i = x\right] \approx \frac{\hat{f}(x, d_0 + 0.1) - \hat{f}(x, d_0)}{0.1}.$$

This approach estimates CAPE instead of CME, as it relies on finite differences rather than derivatives. Moreover, the CAPE estimate depends on the choice of d_0, which introduces researcher discretion.

Second, as a spline-based semiparametric estimator, GAM requires users to specify particular nonlinear relationships to include in the model, making it prone to model misspecification. When additional covariates are present, manually selecting which nonlinear relationships to focus on becomes impractical. As demonstrated by Hainmueller and Hazlett (2014), GAM is not a reliable conditioning strategy when multiple covariates (even a small number) need to be modeled, such as in the case of $f(V)$. In other words, jointly modeling a complex

response surface of Y given D and multiple V is computationally infeasible (typically, `mgcv` can handle at most two variables, including D).

Third, unlike DML estimators, GAM does not satisfy the Neyman orthogonality condition, which prevents small errors in estimating nuisance functions from propagating to the final estimates. As a result, it suffers from regularization bias due to its reliance on penalized splines, particularly in high-dimensional, data-rich settings. Similarly, because GAM does not model the treatment assignment mechanism, it lacks the doubly robust property of methods like AIPW-LASSO or PDS-LASSO, making it a less appealing choice in small-to-medium sample size scenarios.

Below, we demonstrate the first problem, which we consider the most detrimental. We simulate a dataset with a continuous treatment D, a continuous moderator X, and no additional covariates. The exact DGP is provided in Appendix Section A.2. The true CME is $\theta(x) = x - x^2$. However, the standard implementation of GAM using the `mgcv` package in R—as in both Simonsohn (2024*a*,*b*)—does not estimate the CME but instead produces the average conditional partial effect, $\rho(d, x)$. As shown in Figure 4, $\rho(d, x)$ varies drastically depending on d, the level of D being conditioned on, which can lead to significant confusion in applied research. This issue highlights the importance of clearly defining the estimand from the outset to avoid misinterpretation.

Conclusion and Recommendations

Recent critiques of HMX (2019) reignite important discussions about robustly estimating conditional causal relationships. However, these critiques have two critical flaws. First, the lack of a well-defined estimand creates confusion about the theoretical target, leading to erroneous conclusions (such as the claim that the kernel estimator cannot recover the key quantity of interest). Second, the main recommendation—using GAM—fall short in estimating the CME or addressing key empirical challenges, particularly when additional

FIGURE 4. ESTIMATED CME: GAM VS KERNEL
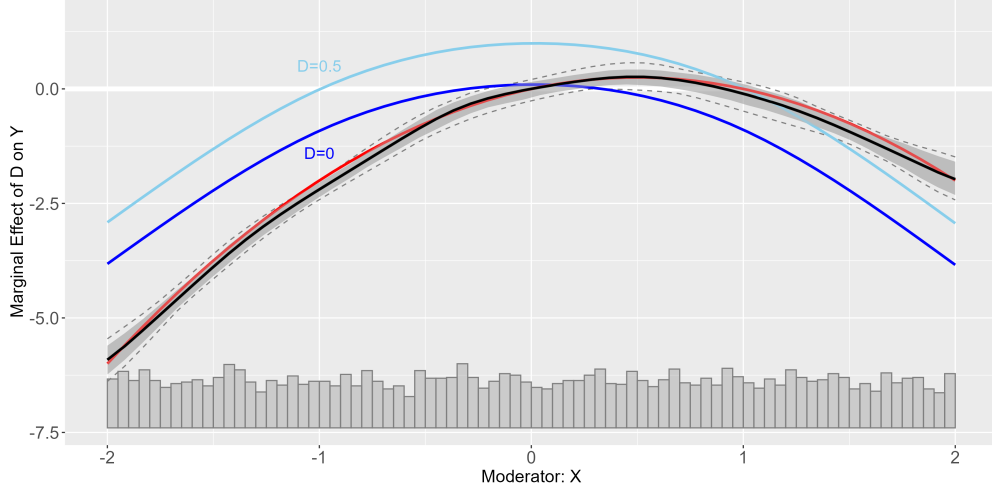
Note: The red solid line and the black solid line represent the true CME, $\theta(x) = x - x^2$, and estimated CME using the kernel method, respectively. The shaded gray areas and the gray dashed lines depict the point-wise and uniform confidence intervals, respectively. The histogram at the bottom displays the distribution of units across different values of X. In addition, the blue line depicts the estimated average conditional partial effect using the GAM estimator conditional on $D = 0$, with the true value $\rho(0, x) = -x^2$, while the light blue line shows the estimated average conditional partial effect using the GAM estimator conditional on $D = 0.5$, with the true value $\rho(0.5, x) = 1 - x^2$.

covariates are present. Building on HMX (2019) and subsequent works, we propose the following recommendations:

- Clearly state the quantity of interest and the key identifying and modeling assumptions before proceeding with the analysis.

- Assess data quality by checking for missingness and outliers, and evaluate whether the overlap assumption is satisfied; improve overlap by trimming the data if necessary.

- Use GAM, the binning estimator, and the kernel estimator as exploratory and diagnostic tools to understand potential nonlinear relationships.

- Adopt flexible modeling strategies, such as the kernel estimator, DML techniques, and other doubly robust estimators to handle potential nonlinearity and high-dimensional covariates.

Setting	Suitable Estimators
Experimental	Kernel estimator
Observational w/ a small-to-medium n (e.g., $n \leq 5,000$)	AIPW-LASSO, PDS-LASSO
Observational w/ a large n (e.g., $n > 5,000$)	DML estimators

Note: The doubly robust and DML estimators will reduce to the kernel estimator when no additional covariates exist or confound the relationship between D and Y. In addition, the kernel estimator remains a valid the tool when the covariates enter the outcome model linearly. The DML estimators have already been incorporated into the `interflex` package. In the coming days, we will update the package to include AIPW-LASSO and PDS-LASSO as well.

- Interpret results carefully. Use causal language with caution. For example, CME describes the causal effect of D along the moderator X, but it does not represent the causal effect of X itself.

Finally, we thank Professor Simonsohn again for the constructive criticism that inspired this response.

References

Athey, Susan, Julie Tibshirani, Stefan Wager et al. 2019. "Generalized Random Forests." *The Annals of Statistics* 47(2):1148–1178.

Bansak, Kirk. 2020. "Estimating Causal Moderation Effects with Randomized Treatments and Non-Randomized Moderators." *Journal of the Royal Statistical Society Series A: Statistics in Society* 184:65–86.

Belloni, Alexandre, Victor Chernozhukov and Christian Hansen. 2014. "Inference on Treatment Effects after Selection among High-dimensional Controls." *Review of Economic Studies* 81(2):608–650.

Blackwell, Matthew and Michael P Olson. 2022. "Reducing Model Misspecification and Bias in the Estimation of Interactions." *Political Analysis* 30(4):495–514.

Brambor, Thomas, William Roberts Clark and Matt Golder. 2006. "Understanding Interaction Mdels: Improving Empirical Analyses." *Political analysis* 14(1):63–82.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen and Whitney Newey. 2017. "Double/Debiased/Neyman Machine Learning of Treatment Effects." *American Economic Review, Papers and Proceedings* 107(5):261–65.

Hainmueller, Jens and Chad Hazlett. 2014. "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach." *Political Analysis* 22(2):143–168.

Hainmueller, Jens, Jonathan Mummolo and Yiqing Xu. 2019. "How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice." *Political Analysis* 27(2):163–192.

Imbens, Guido and Yiqing Xu. 2024. "LaLonde (1986) after Nearly Four Decades: Lessons Learned." *arXiv preprint arXiv:2406.00827* .

Liu, Jiehan, Ziyi Liu and Yiqing Xu. N.d. "Estimating and Interpreting Conditional Marginal Effects." under contract, Cambrdige University Press.

Lundberg, Ian, Rebecca Johnson and Brandon M Stewart. 2021. "What is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory." *American Sociological Review* 86(3):532–565.

Neyman, J. 1923. "On the Application of Probability Theory to Agricultural Experiments (with discussion)." *Statistical Science* 5:465–472.

Robins, James M and Yaacov Ritov. 1997. "Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semi-parametric Models." *Statistics in Medicine* 16.

Rubin, D. B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66:688–701.

Semenova, Vira and Victor Chernozhukov. 2021. "Debiased Machine Learning of Conditional Average Treatment Effects and Other Causal Functions." *The Econometrics Journal* 24(2):264–289.

Simonsohn, Uri. 2024*a*. "Dear Political Scientists: Don't Bin, GAM Instead." Data Colada blog.
  URL: *https://datacolada.org/121*

Simonsohn, Uri. 2024*b*. "Interacting with Curves: How to Validly Test and Probe interactions in the Real (Nonlinear) World." *Advances in Methods and Practices in Psychological Science* 7(1):25152459231207787.

Van der Laan, Mark J and Sherri Rose. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data.* Springer Science & Business Media.

VanderWeele, Tyler J. 2009. "On the Distinction Between Interaction and Effect Modification." *Epidemiology* 20:863–871.

A. Appendix

A.1. Dissecting the Key Simulated Example

We begin by deriving the true CME. We then show that the linear interaction model is biased, while the kernel estimator proposed in HMX (2019) can recover the CME when data are abundant.

The true CME.   The partial derivative of Y with respect to D is

$$\frac{\partial Y}{\partial D} \; = \; 2D \; - \; 0.5.$$

To obtain the conditional marginal effect at $X = x$, we compute

$$\mathbb{E}\!\left[\frac{\partial Y}{\partial D} \; \middle| \; X = x\right] \; = \; \mathbb{E}[\, 2D - 0.5 \mid X = x\,].$$

Under the bivariate normal assumption $\mathrm{Corr}(D, X) = 0.5$, it follows that

$$\mathbb{E}[D \mid X = x] = \mathbb{E}[D] + \frac{\mathrm{Cov}(D, X)}{\mathrm{Var}(X)}(x - \mathbb{E}[X]) = 0.5x.$$

Therefore, the true CME is

$$\theta(x) \; = \; \mathbb{E}[\, 2D - 0.5 \mid X = x\,] \; = \; 2\big(0.5\,x\big) - 0.5 \; = \; x - 0.5.$$

The linear interaction model is biased for CME.   We now consider the misspecified linear–interaction model:

$$Y \; = \; \beta_0 \; + \; \beta_d\, D \; + \; \beta_x\, X \; + \; \beta_{dx}\, (D \times X) \; + \; \text{error}.$$

Under our normality assumptions, the 4×4 matrix $M = \mathbb{E}\big[\mathbf{Z}\,\mathbf{Z}^{\top}\big]$, where $\mathbf{Z} = (1 \; D \; X \; D\,X)^{T}$, together with $\mathbf{b} = \mathbb{E}[\mathbf{Z}\,Y]$, fully determines the population OLS solution:

$$\beta \; = \; \big(\beta_0,\; \beta_d,\; \beta_x,\; \beta_{dx}\big)^{\top} \; = \; M^{-1}\,\mathbf{b}.$$

Carrying out the matrix calculations, we find

$$\beta_0 = 0.6, \quad \beta_d = -0.5, \quad \beta_x = 0, \quad \beta_{dx} = 0.8.$$

Thus, the fitted regression implies the conditional marginal effect

$$\widehat{\mathrm{CME}}(X) \;=\; \frac{\partial \hat{Y}}{\partial D} \;=\; \beta_d \;+\; \beta_{dx} X \;=\; -0.5 \;+\; 0.8\,X.$$

Note that this differs from the true CME, $X - 0.5$.

The kernel estimator can recover the CME. Although the true DGP features a D^2 term, the kernel estimator can recover the true CME, that is, $\theta(x) = x - 0.5$ (in regions where data are abundant). To understand why, consider the following local-linear setup. The kernel estimator assumes:

$$Y \;=\; f(X) \;+\; g(X)\,D \;+\; \epsilon, \quad \mathbb{E}[\epsilon \mid X, D] = 0,$$

where f and g are unknown functions of X that we approximate *locally* around $X = x_0$ using linear functions. Specifically, for $|X - x_0| < r$ (a small neighborhood), write

$$f(X)_{|X-x_0|<r} \;\approx\; \mu(x_0) + \eta(x_0)\,(X - x_0), \quad g(X)_{|X-x_0|<r} \;\approx\; \alpha(x_0) + \beta(x_0)\,(X - x_0).$$

We then estimate the local coefficients $\{\mu(x_0),\, \eta(x_0),\, \alpha(x_0),\, \beta(x_0)\}$ by solving the following weighted least-squares problem:

$$\left(\hat{\mu}(x_0),\, \hat{\alpha}(x_0),\, \hat{\eta}(x_0),\, \hat{\beta}(x_0)\right) \;=\; \arg\min_{\tilde{\mu},\, \tilde{\alpha},\, \tilde{\eta},\, \tilde{\beta}} L\left(\tilde{\mu},\, \tilde{\alpha},\, \tilde{\eta},\, \tilde{\beta}\right),$$

$$L \;=\; \sum_{i=1}^{N}\left[Y_i \;-\; \tilde{\mu} \;-\; \tilde{\alpha}\,D_i \;-\; \tilde{\eta}\,(X_i - x_0) \;-\; \tilde{\beta}\left[D_i\,(X_i - x_0)\right]\right]^2 K\!\left(\frac{X_i - x_0}{h(x_0)}\right),$$

where $K(\cdot)$ is a *kernel weight* function, and $h(x_0)$ is the *bandwidth* controlling how quickly weights decay as X_i moves away from x_0.

For simplicity, consider the uniform kernel $K(z) = 1$, if $|z| < 1$ and 0 otherwise. As $h(x_0) \to 0$, this kernel enforces $|X_i - x_0| < h(x_0)$ to be very small, so effectively only points with $X_i \approx x_0$ remain in the sum. In that infinitesimal neighborhood, $(X_i - x_0) \approx 0$, the variation in $(X - x_0)$ is negligible, and thus the local-linear approximations for f and g

reduce to a local OLS regression of Y on $\{1, D\}$. Using the fact that (D, X) is bivariate normal with $\mathrm{Cov}(D, X) = 0.5$, we can show that, in the neighborhood of x_0, D's coefficient $\hat{b}_1(x_0)$ can be calculated as $\widehat{\mathrm{Cov}}(D, Y)$ divided by $\widehat{\mathrm{Var}}(D)$ (conditional on $X \approx x_0$):

$$\hat{b}_1(x_0) \approx \frac{\mathrm{Cov}(D, Y \mid X \approx x_0)}{\mathrm{Var}(D \mid X \approx x_0)} \approx x_0 - 0.5.$$

Thus, $\hat{\alpha}(x_0)$, which equals to $\hat{b}_1(x_0)$, has approach the *true* conditional marginal effect $\theta(x_0) = x_0 - 0.5$.

Extension to more complex settings. Consider the following DGP, which is used in Simonsohn (2024b): $Y = s(D, X) + \epsilon$. Assume $s(D, X)$ is a smooth function (i.e., twice continuously differentiable). In such cases, the kernel estimator can still consistently estimate the CME in regions of X where data are sufficiently abundant. We provide a simple proof in the appendix.

To see this, expand $s(D, X)$ via a first-order Taylor series around $X = x_0$:

$$s(D, X) \approx s(D, x_0) + \left.\frac{\partial s}{\partial X}\right|_{(D, x_0)} \cdot (X - x_0).$$

The CME at $X = x_0$ therefore is:

$$\theta(x_0) = \mathbb{E}\left[\frac{\partial Y}{\partial D} \,\middle|\, X = x_0\right] = \mathbb{E}\left[\frac{\partial s(D, X)}{\partial D} \,\middle|\, X = x_0\right].$$

The kernel estimator is implemented via local linear regressions weighted by a kernel $K\left(\frac{X - x_0}{h}\right)$, where h is the bandwidth. The regression model proposed in Hainmueller, Mummolo and Xu (2019) is:

$$Y = \beta_0 + \beta_1 D + \beta_2 (X - x_0) + \beta_3 D \cdot (X - x_0) + \epsilon.$$

which approximates $s(D, X)$ near $X = x_0$ as follows:

$$s(D, X) \approx \underbrace{\beta_0 + \beta_1 D}_{\text{Intercept and slope in } D} + \underbrace{\beta_2 (X - x_0) + \beta_3 D \cdot (X - x_0)}_{\text{Adjustment for } X \text{ near } x_0}.$$

At $X = x_0$, the interaction term $D \cdot (X - x_0)$ vanishes, leaving:

$$\left. \frac{\partial Y}{\partial D} \right|_{X=x_0} = \beta_1.$$

Thus, β_1 directly estimates $\theta(x_0) = \mathbb{E}\left[\frac{\partial s(D,X)}{\partial D} \mid X = x_0 \right]$.

Consistent estimation of the CME requires the following conditions: (i) as $n \to \infty$, the bandwidth $h \to 0$, ensuring that the neighborhood around x_0 shrinks; and (ii) the product $n \cdot h \to \infty$, guaranteeing that there are sufficient observations near x_0 to achieve reliable estimation. Under these conditions, the local linear regression minimizes the weighted least-squares error. By the properties of local polynomial regression:

$$\hat{\beta}_1 \xrightarrow{p} \left. \frac{\partial s(D, X)}{\partial D} \right|_{X=x_0}.$$

Thus, $\hat{\beta}_1$ is a consistent estimator of $\theta(x_0)$.

A.2. DGP for the Simulated Examples

DGP for the simulated example in Figure 3. The true DGP is defined as follows: the moderator X is uniformly distributed on $[-2, 2]$, the covariates Z_1 and Z_2 are independent standard normal variables. The treatment assignment is given by:

$$\mathbb{P}(D \mid V) = \frac{e^{(0.5X + 0.5Z_1)}}{1 + e^{(0.5X + 0.5Z_1)}}.$$

The outcome Y is generated according to

$$Y = 1 + X^2 + D - X^2 D + e^{Z_1 + 0.5X} Z_2 + Z_1^2 - 2 \cdot \mathbb{I}(Z_2 > 0)Z_2 + 3\sin(Z_1 + Z_2) + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, 1)$.

DGP for the simulated example in Figure 4. The true DGP is defined as follows: the moderator X is uniformly distributed on $[-2, 2]$. The treatment variable D is generated by

$$D = 0.5X + \epsilon_D,$$

where $\epsilon_D \sim \mathcal{N}(0, 1)$. The outcome Y is generated according to

$$Y = 1 + 1.5X + D^2 - DX^2 + \epsilon_Y,$$

where $\epsilon_Y \sim \mathcal{N}(0, 1)$. The true CME is

$$\mathbb{E}\left(\frac{\partial Y}{\partial D} \Big| X = x\right) = x - x^2.$$

.