

Decomposition and Interpretation of Treatment Effects in Settings with Delayed Outcomes*

Federico A. Bugni
Department of Economics
Northwestern University
federico.bugni@northwestern.edu

Ivan A. Canay
Department of Economics
Northwestern University
iacanay@northwestern.edu

Steve McBride
Head of Economy Science
Roblox

July 30, 2025

Abstract

This paper studies settings where the analyst is interested in identifying and estimating the average *direct* causal effect of a binary treatment on an outcome. We consider a setup in which the outcome realization does not get immediately realized after the treatment assignment, a feature that is ubiquitous in empirical settings. The period between the treatment and the realization of the outcome allows other observed actions to occur and affect the outcome. In this context, we study several regression-based estimands routinely used in empirical work to capture the average treatment effect and shed light on interpreting them in terms of ceteris paribus effects, indirect causal effects, and selection terms. We obtain three main and related takeaways under a common set of assumptions. First, the three most popular estimands do not generally satisfy what we call *strong sign preservation*, in the sense that these estimands may be negative even when the treatment positively affects the outcome conditional on any possible combination of other actions. Second, the most popular regression that includes the other actions as controls satisfies strong sign preservation *if and only if* these actions are mutually exclusive binary variables. Finally, we show that a linear regression that fully stratifies the other actions leads to estimands that satisfy strong sign preservation.

KEYWORDS: delayed outcomes, treatment effects, regression, total effects, direct effects, downstream engagement, sign preservation.

JEL classification codes: C12, C14.

*We thank Chuck Manski, Peng Ding, Lihua Lei, Filip Obradovic, and participants at various seminars and conferences for comments. We also thank Danil Fedchenko for excellent research assistance.

1 Introduction

We study settings where the analyst is interested in identifying and estimating an average *direct* causal effect of a binary treatment on an outcome, and the treatment status is determined in the context of a randomized controlled experiment or an observational study under conditional independence assumptions. We focus on settings where the outcome of interest does not get immediately realized after treatment assignment, a feature that is ubiquitous in empirical settings, including the study of long-run effects in economic history (Voigtländer and Voth, 2012; Angelucci et al., 2022), the analysis of long-term valuation in industry settings (Jain and Singh, 2002; Akhtari et al., 2021), and randomized experiments in economics and other social sciences (Beaman et al., 2013; Moderna, 2021). The delay in the realization of the outcomes creates a time window between the treatment assignment and the realization of the outcome that, in turn, opens up the possibility for other observed ex-post actions to take place before the outcome is finally realized; see Figure 1 for a graphical representation. In this context, we study the interpretation of several popular estimands that arise from running regressions of the outcome on the treatment and different ways of “controlling” for the other actions. We emphasize that our use of the term “regression” refers to a linear projection that is free of any modeling assumptions on potential outcomes or conditional means. Some of these estimands are not only popular in the economics literature, see, e.g., Fagereng et al. (2021); Heckman et al. (2013); Chernozhukov et al. (2021), but are also widely used across other sciences, like psychology, epidemiology, biostatistics, and political science, as shown by the large number of citations associated with the regression approach popularized by Baron and Kenny (1986). For each of these estimands, our results present a decomposition that facilitates their interpretation in terms of *ceteris paribus* effects of the treatment on the outcomes, indirect effects caused by the other actions, and selection terms; and provide a framework that allows us to clarify under what type of conditions the practice of “controlling” for the presence of other actions leads to estimands that admit the desired interpretation.

The main findings of this paper can be grouped into three sets of results. First, the standard practice of studying estimands that arise from a regression of an outcome on the treatment, with or without “controlling” for the other actions in such regressions, does not generally satisfy what we call *strong sign preservation*. Strong sign preservation, formally characterized in Definition 3.3, is satisfied when an estimand that intends to measure a *ceteris paribus* (or direct) causal effect of a treatment on an outcome is positive when the effect of the treatment on the outcome is positive conditional on *all* possible values of the other actions. Failure to satisfy strong sign preservation introduces a Simpson’s paradox-like sign reversal where the estimands may be negative even when the treatment positively affects the outcome for any possible combination of other actions. Second, the most popular estimand that linearly controls for the other actions in the regression, and that we label the long regression, does not generally provide benefits relative to the short regression that includes no controls whatsoever. More concretely, while neither the short nor the long regression satisfies strong sign preservation, the estimand associated with the long

regression admits a decomposition in terms of weighted averages of well-defined causal effects but where the weights could potentially be negative. This feature introduces yet another source that may separate the sign of the estimand from the sign of *ceteris paribus* causal effects. Notably, this feature also occurs when the regression includes interaction terms between the treatment and the other actions. Perhaps our most salient result is the one in Theorem 4.2, which shows that the long regression delivers easy-to-interpret results *if and only if* the other actions are all binary and mutually exclusive random variables. Finally, while non-parametric identification of the effects is straightforward under the stronger form of our assumptions and follows directly from a saturated regression, we also show that a linear regression that properly controls for other actions through complete stratification produces estimands that satisfy strong sign preservation. We term this the “strata fixed effects regression” due to its link to the practice of including strata fixed effects in randomized controlled trials with covariate adaptive randomization (see Bugni et al. (2018, 2019)).

The decompositions we derive for each of the five regression-based estimands can be interpreted as a partition into “direct” and “indirect” effects (and possibly “selection” effects depending on the assumptions), and so our results are linked to the vast literature on mediation analysis, see, e.g., Baron and Kenny (1986), Pearl (2001), Robins (2003), Kaufman et al. (2004), Kaufman (2009), Imai et al. (2010), VanderWeele et al. (2014), and Remark 2.1 for a discussion. See VanderWeele (2015) for a book-level treatment. However, as opposed to the literature on mediation that studies the type of assumptions that would identify the causal effects of the so-called *mediators*, which in our context would simply be the other actions taken before the outcome is realized, here our goal is not to identify these indirect effects but rather to gain a better understanding of how to properly interpret certain popular estimands of the effect of the treatment on the outcome without the aid of functional form assumptions.

Beyond the literature on mediation analysis, our paper also connects to several strands of research in econometrics and biostatistics. First, we are not the first to highlight the importance of distinguishing between “direct” (or “partial”) and “total” causal effects in econometrics. Early discussions include those in Manski (1997) and Heckman (2000); see also Remark 3.1. Second, our results relate to a large body of research that interprets regression-based estimands in various settings, often characterizing them as weighted averages of causal effects of interest—sometimes involving negative weights. Examples include the literature on two-way fixed effects in difference-in-differences settings (De Chaisemartin and d’Haultfoeuille, 2020; Goodman-Bacon, 2021; Callaway and Sant’Anna, 2021; Sun and Abraham, 2021; Borusyak et al., 2022); on local instrumental variables and marginal treatment effects (Canay et al., 2023); and on contamination bias when conditioning on covariates (Angrist, 1998; Goldsmith-Pinkham et al., 2022). Our paper is also connected to the literature on design-based analyses of regression estimands in factorial experiments (Zhao and Ding, 2022). These connections are primarily algebraic—stemming from shared properties of least squares—though the questions we address and the main takeaways of our analysis are distinct. Third, our results on the failure of strong sign preservation for popular estimands resonate with the “surrogate paradox” in the surrogacy literature, where a treatment may have a

positive effect on a surrogate, and the surrogate may be positively correlated with the outcome, yet the treatment effect on the outcome may be negative; see [Chen et al. \(2007\)](#) and [VanderWeele \(2013\)](#). Finally, the decomposition we derive for the short regression mirrors those developed in the literature on causal interactive effects—where units may be subject to multiple types of exposures (e.g., genetic and environmental)—as in [Robins and Greenland \(1992\)](#) and [VanderWeele and Tchetgen Tchetgen \(2014\)](#). Our focus on a broader range of regression estimands, however, sets our work apart.

The remainder of the paper is organized as follows. [Section 2](#) introduces the basic notation. [Section 3](#) defines the main concepts we use throughout the paper, including partial causal effects, direct causal effects, and strong sign preservation. [Section 4](#) introduces the five estimands we study and then presents the main results on how each of these estimands admits different decompositions into direct, indirect, and selection effects. Finally, [Section 5](#) concludes.

2 Setup and Notation

Consider a setting where Y denotes the observed outcome of interest, and the actions taken by individuals or units under study are divided into a “main” action of interest, denoted by D , and “other” actions, denoted by A :

$$(D, A) \in \mathcal{D} \times \mathcal{A} . \tag{1}$$

Let X represent other observed covariates, which include features beyond actions.

All actions are assumed to be discrete. The main action, D , is further assumed to be binary, i.e., $\mathcal{D} \equiv \{0, 1\}$. The other actions, A , form a K -dimensional vector taking values in $\mathcal{A} \equiv \{a = (a_1, \dots, a_K) : a_j \in \mathcal{A}_j \text{ for } j = 1, \dots, K\}$, where each \mathcal{A}_j is a finite set. For notational convenience, we assume $\mathcal{A}_j \subseteq \mathbb{N}$, $0 \in \mathcal{A}_j$, and that when $|\mathcal{A}_j| = 2$, we have $\mathcal{A}_j = \{0, 1\}$. These additional restrictions simplify the discussion and expressions but are not required for our results.

The setting we study in this paper is one with the following characteristics. First, the analyst controls the action of interest D via either a randomized controlled experiment or, alternatively, by an exogeneity assumption like selection on observables. We therefore alternatively call this action the “treatment”. Second, the outcome Y is not instantaneous and takes some time to be realized within the timeline of the experiment. In the period between treatment assignment and outcome realization, the other actions in A are chosen by the units—these may be modeled as conditionally exogenous in some applications, or left unmodeled altogether. [Figure 1](#) illustrates the setting. Below we describe some empirical applications in economics, social sciences, and industry that naturally fit into this setting.

The first class of applications that fit our framework is the literature that studies long-run outcomes in economics. There, interest typically lies in a treatment that happened several years in

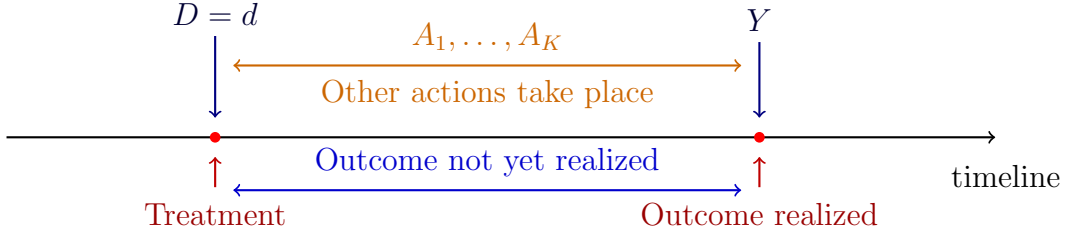


Figure 1: Timeline of actions. The first action, D , is assumed to be (conditionally) exogenous and is the main action of interest (so we refer to it as the treatment). The outcome is not instantaneous and may take a short or long period of time to get realized. In the meantime, units choose the value of the other actions A_1, \dots, A_K .

the past (oftentimes hundreds of years ago) on some outcome of interest in more recent times. For example, [Voigtländer and Voth \(2012\)](#) study the effect of the existence of Black Death pogroms in 1349 (D) on the level of anti-Semitism in Nazi Germany in 1920s (Y) at the city level. Other actions that are included in the analysis include city population in the 1920s (A), among other variables that are determined closer to the realization of the outcomes. Other examples include [Nunn \(2008\)](#) and [Angelucci et al. \(2022\)](#), among many others. In these applications, it is common to argue that the treatment is exogenous invoking selection on observables assumptions or, in some cases, relying on instrumental variables approaches. Our results are limited to settings exploiting conditional exogeneity.

The second class of applications arises in the analysis of long-term valuation in marketing and industry. In these settings, companies such as Airbnb, Uber, Google, and Microsoft operate platforms where users engage in a wide range of actions—such as making a booking, writing a review, subscribing to a service, or watching a video—and are interested in estimating the long-run value that a specific product or action generates for the firm. For instance, [Akhtari et al. \(2021\)](#) discusses how Airbnb measures the long-term value of user actions by tracking the revenue generated over a 365-day window. The example focuses on the effect of a guest making a booking (D) on long-term revenue (Y), while other intermediate actions—such as cancellations or writing reviews—are captured in the vector A . Although randomized experiments are often feasible for some actions, others cannot easily be randomized due to ethical, legal, or user experience concerns. In such cases, it is common to rely on conditional exogeneity (i.e., selection on observables) and focus on one action of interest at a time. Importantly, in these applications, the goal is often to identify a *direct* causal effect of the focal action (D) on the outcome (Y), holding other actions (A) fixed. This approach helps avoid the problem known as *double counting*—that is, attributing the same impact multiple times through correlated or causally related behaviors. For this reason, isolating the direct effect of D is not merely of academic interest, but a practical necessity in how firms estimate the marginal contribution of individual features or interventions. Related examples and discussions can be found in [Lewis et al. \(2011\)](#) and [Xu et al. \(2015\)](#), who discuss how correlated consumer behaviors—termed “activity bias”—can lead to substantial overestimation of advertising

effectiveness.

The third class of applications includes clinical trials and randomized experiments with a follow-up period between the treatment assignment and the realization of the outcome of interest. There, researchers typically fully control the assignment of the main treatment of interest but are unable to restrict behavior in the follow-up period. For example, the clinical trial run by [Moderna \(2021\)](#) to study the efficacy of the Moderna COVID-19 vaccine against SARS-CoV-2 infections. Participants in the study were randomized to Immediate Vaccination Group 1 (receiving the Moderna COVID-19 Vaccine on Day 1 and Day 29) or Standard of Care Group 2, with vaccination given at months 4 and 5. During the months following vaccinations, participants received visits that checked for infections and could include blood collection, nasal swabs, SARS-CoV-2 screening, COVID-19 symptom checks, and questionnaires. In this example, D would be an indicator of whether the participant received a vaccine, Y would be an indicator of whether the participant got infected within the 4 months of the study, and A would include other actions taken by the participants that could affect infection rates, like whether the participants wear masks in public, whether the participants avoid large gatherings, etc.¹ Another example is [Beaman et al. \(2013\)](#), who conducted a randomized field experiment that provides free fertilizer to female rice farmers in Mali to measure how they choose to use the fertilizer and the overall impact on profitability. The authors distributed the fertilizer in May 2020 and conducted two follow-up surveys, one in August 2020 and one in December 2020, right after the harvest. In this example, D would be an indicator of whether the farmer received free fertilizer, Y would be a measure of output like crop yield or just profits, and A would include all relevant complementary agricultural inputs, such as labor, herbicides, and water usage. Due to its simplicity, we use this farming example to illustrate the concepts we define in the next section.

Remark 2.1. What we call the other actions in [Figure 1](#) can be alternatively labeled as “mediator” variables since these are post-treatment variables that occur before the outcome is realized, see, e.g., [Baron and Kenny \(1986\)](#), [Pearl \(2001\)](#), [Robins \(2003\)](#), [Imai et al. \(2010\)](#), [VanderWeele et al. \(2014\)](#), and [VanderWeele \(2015\)](#), among many others. However, our work deviates from this literature in two important ways. First, while the literature on causal mediation analysis focuses on the identification of causal effects induced by mediators, our focus in this paper is to understand whether common estimands that are used to capture causal effects of main action D on the outcome Y admit clear interpretations through the lens of total and direct effects. Second, our decompositions in terms of direct and indirect effects are defined in terms of potential outcomes for all of the actions, including those that may be labeled as mediators, and this implies “indirect” effects in our context do not coincide with the definition of indirect effects in the mediation literature but rather with the so-called “controlled” effects discussed by [Pearl \(2001\)](#) and [Robins \(2003\)](#); see [Remark 3.2](#) for additional discussion on this distinction. It is worth noting, however, that several of our results have implications for the causal mediation literature, and we discuss these implications as we present

¹In this paper we abstract away from spillover effects between individuals, which could be relevant in the context of this and other examples we describe.

our main results. ■

Remark 2.2. The problems associated with the presence of the other “endogenous” actions A could arise in settings that do not require the presence of “delayed outcomes”. Throughout the paper, we use the term “endogenous” to refer to actions that may respond to the treatment, in the sense that their distribution may vary with the value of D . For example, the factorial experiments considered by [Zhao and Ding \(2022\)](#), the problems associated with “bad controls” discussed by [Angrist and Pischke \(2008\)](#), or the mediation framework described in [Remark 2.1](#), are all about endogenous actions that are not necessarily related to delayed outcomes. While we do not need to invoke delayed outcomes to introduce the type of interpretation challenges we discuss here, we choose to do so because it directly speaks to the examples that motivated this paper. ■

We denote potential outcomes by $Y(d, a)$ and their expectation by

$$\mu(d, a) \equiv E[Y(d, a)] . \tag{2}$$

We also introduce the concept of a pooled potential outcome to isolate the counterfactual outcome associated with the main action of interest (the treatment),

$$Y(d) = Y(d, A(d)) = \sum_{a \in \mathcal{A}} Y(d, a) I\{A(d) = a\} , \tag{3}$$

where $A(d)$ denotes potential outcomes for the actions A as a function of the treatment d . Finally, the observed outcome Y is related to potential outcomes by the relationship

$$Y = \sum_{(d,a) \in \mathcal{D} \times \mathcal{A}} Y(d, a) I\{(D, A) = (d, a)\} . \tag{4}$$

In the next section, we define the type of causal effects we are interested in this paper, as well as the assumptions that we invoke to interpret the decompositions we derive for a variety of estimands.

3 Causal Treatment Effects

We start by discussing the type of counterfactual treatment effects that could interest the researcher in the canonical setting where D is binary. Viewing $Y(d, a)$ as a function of two types of actions immediately suggests that there could be total effects, direct effects, and indirect effects, all of which may or may not be of interest in the context of a concrete application. Understanding the variety of causal effects that one could describe, in turn, will help us provide representations and interpretations of commonly used target parameters, like the average treatment effect (ATE), in terms of these types of causal effects. We start with what is perhaps one of the most natural types of *ceteris paribus* effects in [Definition 3.1](#). ■

Definition 3.1 (Average Partial Causal Effect). An **average partial causal effect** of D on the outcome Y is any difference of the form $\mu(1, a) - \mu(0, a)$, where the value $a \in \mathcal{A}$ is kept constant.

Definition 3.1 defines an average partial causal effect of the main action as a mean comparison that keeps the value of the other actions unchanged in both states of the comparison (*ceteris paribus*). In the farming application of Beaman et al. (2013), it would capture the average causal effect of using fertilizer on the crop yield, while keeping other inputs, like labor, herbicides, and water usage, constant in the counterfactual comparison.

The *ceteris paribus* effect in Definition 3.1 could also be defined conditional on certain events or subpopulations. To account for this, we also consider the concept in Definition 3.1 conditional on some set Ω , i.e., $E[Y(1, a) - Y(0, a) \mid \Omega]$, where Ω is a function of (D, A, X) . For example, $\Omega = I\{D = 1\}$ would lead to an average partial causal effect on the treated and $\Omega = I\{X = x\}$ would lead to an average partial causal effect for units with covariates x .

The definition of a partial causal effect for the main action delivers a potentially different causal effect for each possible value of the other actions or, alternatively, provides a collection of partial causal effects indexed by $a \in \mathcal{A}$. While the goal could just be to identify such a collection of effects, in many settings it may be natural to aggregate this collection of partial effects in a way that summarizes the effect of the main action on the outcome of interest. The following definition defines a direct causal effect as any weighted average of partial causal effects.

Definition 3.2 (Average Direct Causal Effect). The **average direct causal effect** of D on the outcome Y is any convex combination of average partial causal effects of D on Y . That is,

$$\sum_{a \in \mathcal{A}} \omega(a) (\mu(1, a) - \mu(0, a)) , \tag{5}$$

where $\omega(a) \in [0, 1]$ for all $a \in \mathcal{A}$ and $\sum_{a \in \mathcal{A}} \omega(a) = 1$.

In the context of our farming example with A only capturing low and high water usage for simplicity, the parameter (5) combines the average causal effect of using fertilizer on the crop yield for units with high water usage, say $A = 1$, and units with low water usage, say $A = 0$. Average direct causal effects could also be defined conditional on a set Ω . The definition does not determine how the groups are weighted, but it requires that no group is assigned a negative weight. In this sense, any average direct causal effect satisfies *strong sign preservation*, as defined below.

Definition 3.3 (Strong Sign Preservation). Let Δ be an estimand that intends to capture the causal effect of the treatment D on the outcome Y . We say that Δ satisfies **strong sign preservation** if

$$\mu(1, a) - \mu(0, a) > 0 \text{ for all } a \in \mathcal{A} \text{ implies } \Delta > 0 .$$

In our farming example with A only capturing low and high water usage, strong sign preservation of a parameter Δ implies that whenever fertilizers improve the expected crop yield both for units

with high water usage and units with low water usage, Δ should be positive as well. As the name suggests, strong sign preservation does not allow for the possibility of what it is typically referred to as *sign reversal*, understood as a situation where $\Delta < 0$ when $\mu(1, a) - \mu(0, a) > 0$ for all $a \in \mathcal{A}$.

A premise of this paper is that researchers employ regression-based estimands to obtain an average direct causal effect, as defined in Definition 3.2. That is, the primary interest is not in identifying *indirect* effects—capturing the impact of D on Y through the actions A —nor in identifying the total effect, which allows the actions A to vary across treatment groups. While such alternative effects may be of interest in certain applications, our aim is to clarify when the estimands described in the next section can be interpreted in line with Definition 3.2.

Remark 3.1. While strong sign preservation may be perceived as a key requirement for parameters that intend to identify partial causal effects, it may not be a reasonable requirement in settings where the counter-factual question of interest involves total effects. The distinctions between “partial” and “total” causal effects have appeared in the literature in a variety of contexts, even beyond the mediation analysis literature discussed in Remark 2.1, where Pearl (2001) and Robins (2003) provide comprehensive treatments on these distinctions. For example, Heckman (2000) defines a causal effect as the partial derivative with respect to an isolated action. He acknowledges that assuming that an action can vary independently of all other actions is strong, yet he regards it as “...essential to the definition of a causal parameter”. Manski (1997, pages 1321 and 1323), in turn, provides two interpretations of potential outcomes (one that keeps other actions fixed and another one that lets the other actions change in response to the main action) and clarifies that the interpretation of treatment effects depends on how we think about potential outcomes. Here, we do not dwell on discussions about the relative merits of partial or total effects but rather seek to understand whether commonly used estimands in empirical work admit either of these (commonly sought after) interpretations under different assumptions. ■

Remark 3.2. Our definitions of average causal partial effects and average direct causal effects align with the concept of “controlled” effects in the mediation analysis literature; see Pearl (2001), Robins (2003), and Kaufman (2009), among others. These differ from the more commonly emphasized “natural” direct effects. In our notation, the average natural direct effect is given by $E[Y(1, A(d)) - Y(0, A(d))]$, where the action (or mediator) A is set to the level it would naturally take under the reference treatment $D = d$. The average natural indirect effect, in turn, is given by $E[Y(d, A(1)) - Y(d, A(0))]$ for $d \in \{0, 1\}$. While the total effect, $E[Y(1) - Y(0)]$, decomposes into natural direct and indirect effects without additional assumptions, the decomposition into controlled direct and indirect effects generally introduces selection terms unless further assumptions are imposed—as we show in the next section. On the other hand, identifying natural effects typically requires stronger assumptions than those needed for identifying controlled effects; see VanderWeele et al. (2014). ■

The decompositions we present in this paper are purely algebraic and, in their most general form, do not rely on any identification assumptions. However, to interpret the terms in these

decompositions as causal effects—or to simplify their expressions—it is useful to introduce assumptions that both facilitate interpretation and connect our framework to assumptions commonly used by practitioners and related work in the literature.

We begin with a baseline assumption that allows for the identification of a *total* effect of D on Y , i.e., $E[Y(1) - Y(0)]$. This assumption is standard in the literature on treatment effects in settings that do not explicitly account for the presence of the other actions A .

Assumption 3.1. For all $d \in \mathcal{D}$, the treatment D is conditionally independent of the potential outcome $Y(d)$ given covariates X , i.e.,

$$D \perp Y(d) \mid X .$$

Assumption 3.1 can be justified by the design of a randomized controlled experiment (e.g., [Beaman et al., 2013](#)) or by relying on a sufficiently rich set of covariates—as is often the case in industry applications—to make the conditional independence assumption credible (as in [Akhtari et al., 2021](#)). Three points are worth emphasizing. First, as previously noted, this assumption is not required to derive our decompositions, but rather to aid in their causal interpretation. Second, as noted by a referee, one could replace Assumption 3.1 with $D \perp (Y(d, a), A(d)) \mid X$ for all $(d, a) \in \mathcal{D} \times \mathcal{A}$. Although this yields equivalent interpretations, we favor Assumption 3.1 as it more closely aligns with applications where other actions are not explicitly modeled. Third, while Assumption 3.1 suffices to identify the average treatment effect of D on Y , it does not generally support a causal interpretation of estimands intended to capture *ceteris paribus* or “direct” effects of D when A is also present. For that purpose, we consider an alternative assumption that grants conditional independence of $Y(d, a)$ with respect to both D and A given X .

Assumption 3.2. For all $(d, a) \in \mathcal{D} \times \mathcal{A}$, the pair (D, A) is conditionally independent of the potential outcome $Y(d, a)$ given X , i.e.,

$$(D, A) \perp Y(d, a) \mid X .$$

Assumption 3.2 reframes the problem as one involving multiple (conditionally exogenous) treatments. Three comments are worth pointing out. First, this assumption is strong in the sense that it enables nonparametric identification of the full mean response function $\mu(d, a)$, as discussed in Section 4.5. In fact, it is equivalent to Assumptions (i) and (ii) in [VanderWeele et al. \(2014\)](#), which are sufficient for identifying controlled direct effects. Second, as noted by a referee, one could instead impose the stronger condition $(D, A(1), A(0)) \perp Y(d, a) \mid X$ for all (d, a) , which leads to the same interpretations. We favor Assumption 3.2 as it aligns better with our focus on controlled effects. Third, while it may hold in settings with a rich set of covariates, it may be questionable when only D is experimentally assigned and A reflects behavioral responses to D .

To illustrate, in the agricultural production example, if fertilizer (D) is randomly assigned but

farmers adjust water usage (A) in response to unobserved shocks that also affect yields—such as weather or soil quality—then Assumption 3.2 would likely fail. By contrast, in many industry applications, it is common to condition on high-dimensional covariates and assume that all actions, including A , are conditionally exogenous. In these settings, D is often designated as the action of interest not because it is fundamentally different from the other actions, but because of how it enters a business decision—for example, it may correspond to the action assigned to a particular team or product line. As such, the full vector (D, A) may be plausibly treated as jointly exogenous given a rich enough set of observed variables.

Importantly, while Assumption 3.2 permits identification of $\mu(d, a)$, we show in later sections that this is still not sufficient for interpreting common regression-based estimands as average direct causal effects of D on Y , as defined in Definition 3.2. This underscores the fragility of such interpretations even under relatively strong identification assumptions.

Remark 3.3. The results in this paper are expressed in terms of conditional expectations and would still hold if the conditional independence assumed in Assumptions 3.1 and 3.2 were weakened to conditional mean independence. We choose to retain the stronger form of these assumptions to facilitate comparison with the typical assumptions used in the literature, which are usually stated in terms of conditional statistical independence. ■

4 Decomposing Common Estimands

In this section, we analyze five natural and highly popular estimands intended to capture treatment effects of D on Y . For each of these estimands, we derive a decomposition in terms of parameters that can be labeled according to Definitions 3.1 and 3.2 and discuss under what assumptions they can be interpreted as intended. To keep our exposition as simple as possible and be able to zoom in on the type of concerns we intend to highlight, in what follows we abstract away from issues related to improper control of the covariates X . In other words, we ignore the role of the covariates in the type of regressions we consider. This could be interpreted as a situation where the covariates are discrete, and the regressions are viewed as within-cell regressions with cells given by $X = x$, or more generally, where the covariates have been already accounted for by other means, like clustering or via a partially linear model, among many possibilities. In particular, we do not consider the possibility that the analyst improperly controls for these covariates by simply including a linear term in X in the regressions, as this would create additional problems to those discussed here; see Goldsmith-Pinkham et al. (2022) for a detailed treatment on the consequences of not properly controlling for confounders. We also re-iterate that our use of the term “regression” refers to a linear projection that is free of any modeling assumptions on potential outcomes or conditional means.

The first such estimand is the usual difference in means, which we write here as the slope

coefficient Δ_{short} in a regression (projection) of Y on D and a constant term,

$$\textbf{Short regression: } Y = \beta + \Delta_{\text{short}}D + U , \quad (6)$$

where $E[UD] = 0$ by properties of projections and $E[U|D] = 0$ follows from D being binary. We call this the short regression.

The second estimand is the slope coefficient D in a linear regression of Y on D , a constant term, and the K actions A_1, \dots, A_K ,

$$\textbf{Long regression: } Y = \Delta_{\text{long}}D + \theta_0 + \sum_{j=1}^K \theta_j A_j + V , \quad (7)$$

where $E[VD] = E[VA_j] = 0$ by properties of projections. We call this the long regression.

The third estimand is the slope coefficient D in a linear regression of Y on D , a constant term, the K actions A_1, \dots, A_K , and their interactions with D ,

$$\textbf{Long regression with interactions: } Y = \Delta_{\text{inter}}D + \theta_0 + \sum_{j=1}^K \theta_j A_j + \sum_{j=1}^K \lambda_j A_j D + e , \quad (8)$$

where $E[eD] = E[eA_j] = E[eA_jD] = 0$ by properties of projections. We call this the long regression with interactions. Note that this is not a fully saturated regression in general, since the random variables A_j are allowed to take arbitrary values in $\mathbb{N} \cup \{0\}$.

The fourth estimand is the slope coefficient D in a regression of Y on D and a set of indicator functions for all the values that A takes,

$$\textbf{Strata fixed effects (SFE) regression: } Y = \Delta_{\text{sfe}}D + \sum_{a \in \mathcal{A}} \theta(a) I\{A = a\} + \nu , \quad (9)$$

where $E[\nu D] = E[\nu I\{A = a\}] = 0$ by properties of projections. Note that this is a regression of Y on D with “strata fixed effects”, where the event $\{A = a\}$ defines a stratum for each value of a . As a result, we call this the strata fixed effect (SFE) regression.

The last set of estimands are the slope coefficients $\Delta_{\text{sat}}(a)$, for $a \in \mathcal{A}$, in a saturated regression of Y on a set of indicator functions for all the values that A takes and their interactions with D ,

$$\textbf{Saturated (SAT) regression: } Y = \sum_{a \in \mathcal{A}} \gamma(a) I\{A = a\} + \sum_{a \in \mathcal{A}} \Delta_{\text{sat}}(a) I\{A = a\} D + \epsilon , \quad (10)$$

where $E[\epsilon D I\{A = a\}] = 0$ by properties of projections and $E[\epsilon | D, I\{A = a\}] = 0$ follows from D and $I\{A = a\}$ being binary for all $a \in \mathcal{A}$. We call this the saturated (SAT) regression.

Remark 4.1. The use of short, long, and long with interaction regressions in the social science literature is ubiquitous. When Glynn (2012) discusses the popularity of these regressions, he writes

that long regressions are so pervasive within the social science and empirical mediation literature that “examples are too numerous to cite.” Indeed, [Baron and Kenny \(1986\)](#), the paper that largely established the use of these and related regressions, has over 137,000 citations as of 2025. ■

4.1 Short regression

The short regression is algebraically very simple, so we build up toward the main result introducing the main concepts and notation along the way. The other regressions, on the contrary, have more opaque derivations, and so in those cases, we first present the formal results and then discuss their interpretation.

The slope coefficient Δ_{short} in (6) equals $\Delta_{\text{short}} = E[Y|D = 1] - E[Y|D = 0]$ by elementary arguments. If we define

$$\pi_d(a) \equiv P\{A = a|D = d\} , \quad (11)$$

and note that

$$E[Y|D = d] = \sum_{a \in \mathcal{A}} E[Y(d, a)|D = d, A = a]\pi_d(a) ,$$

we can decompose Δ_{short} into the following three terms,

$$\Delta_{\text{short}} = \Delta_{\text{dce}}^{\text{s}} + \Delta_{\text{ind}}^{\text{s}} + \Delta_{\text{sel}}^{\text{s}} \quad (12)$$

where

$$\Delta_{\text{dce}}^{\text{s}} \equiv \sum_{a \in \mathcal{A}} \pi_1(a) E[Y(1, a) - Y(0, a)|D = 1, A = a] \quad (13)$$

$$\Delta_{\text{ind}}^{\text{s}} \equiv \sum_{a \in \mathcal{A}} (\pi_1(a) - \pi_0(a)) (E[Y(0, a)|D = 0, A = a] - E[Y(0, 0)|D = 0, A = 0]) \quad (14)$$

$$\Delta_{\text{sel}}^{\text{s}} \equiv \sum_{a \in \mathcal{A}} \pi_1(a) (E[Y(0, a)|D = 1, A = a] - E[Y(0, a)|D = 0, A = a]) . \quad (15)$$

The three terms in the above decomposition for Δ_{short} have a clear interpretation and show that there are two channels of endogeneity introduced by the fact that the other actions, A , take place in-between the treatment assignment and the realization of the outcome of interest. The term in (13), $\Delta_{\text{dce}}^{\text{s}}$, captures an average direct causal effect on the treated, as in Definition 3.2. Note that this term conditions on $\Omega = I\{D = 1, A = a\}$ and so it is a conditional effect like those previously defined. The term in (14), $\Delta_{\text{ind}}^{\text{s}}$, admits a clean interpretation for each value $a \in \mathcal{A}$ under additional assumptions we introduce below. Without additional assumptions, this term is a type of “indirect” effect that contains the product of the difference in conditional probabilities, $\pi_1(a) - \pi_0(a)$, and a term that confounds the average partial causal effect of A moving from 0 to a on Y , with selection that arises from the distinct conditioning sets $\{D = 0, A = a\}$ and $\{D = 0, A = 0\}$. Finally, the term in (15), $\Delta_{\text{sel}}^{\text{s}}$, is a selection term that captures the fact that $Y(0, a)$ may not be independent

of (D, A) .

Three points are worth highlighting. First, the decomposition above does not rely on either Assumption 3.1 or Assumption 3.2. While Assumption 3.1 (ignoring the X) ensures that

$$\Delta_{\text{short}} = E[Y|D = 1] - E[Y|D = 0] = E[Y(1) - Y(0)], \quad (16)$$

where $Y(d)$ denotes the pooled potential outcomes defined in (3), it is not sufficient to interpret Δ_{short} as an average direct causal effect or as a parameter satisfying strong sign preservation. In particular, the two endogeneity channels— Δ_{ind}^s and Δ_{sel}^s —in the decomposition of Δ_{short} may be positive or negative and, crucially, may cause Δ_{short} to have the opposite sign of Δ_{dce}^s . Second, these two endogeneity channels are conceptually distinct. The components in Δ_{ind}^s reflect indirect pathways and are generally difficult to eliminate, while the selection term Δ_{sel}^s becomes zero under Assumption 3.2. Finally, under a given set of assumptions, the decomposition we derive for Δ_{short} —as well as for the other estimands analyzed in subsequent sections—is unique up to the choice of conditioning set (in our case, $\Omega = \{D = d, A = a\}$) and the normalization (here, the indirect effect is defined relative to $\mu(0, 0)$). Once Assumption 3.2 is imposed, the conditioning event becomes irrelevant, and the only remaining degree of freedom lies in the choice of normalization for the indirect effect.

Under Assumption 3.2 the three terms entering the decomposition for Δ_{short} simplify in the following way: $\Delta_{\text{sel}}^s = 0$ and

$$\Delta_{\text{dce}}^s = \sum_{a \in \mathcal{A}} \pi_1(a)(\mu(1, a) - \mu(0, a)) \quad (17)$$

$$\Delta_{\text{ind}}^s = \sum_{a \in \mathcal{A}} (\pi_1(a) - \pi_0(a))(\mu(0, a) - \mu(0, 0)). \quad (18)$$

That is, the selection term Δ_{sel}^s is no longer present, and the direct effect Δ_{dce}^s and indirect effect Δ_{ind}^s are now a function of the unconditional expectations $\mu(d, a)$. Importantly, the term Δ_{ind}^s is still part of the decomposition since Assumption 3.2 does not restrict how A may affect outcomes, so that $\mu(0, a) - \mu(0, 0) \neq 0$, nor does it affect how the main action may affect the other ones, so that $\pi_1(a) - \pi_0(a) \neq 0$. Aside from removing the term capturing selection bias, Assumption 3.2 also delivers a clean interpretation of the indirect effects captured by Δ_{ind}^s . Each summand in Δ_{ind}^s contains the average partial causal effect of A moving from 0 to a on Y , $\mu(0, a) - \mu(0, 0)$, multiplied by the difference $\pi_1(a) - \pi_0(a)$, which admits a causal interpretation of an average direct causal effect of D on A under the additional assumption $A(d) \perp D$.

We can interpret the terms entering the decomposition of Δ_{short} in (12) in the context of the examples we introduced in Section 2. For example, in the farming example where Y is crop yield, D is an indicator of the use of fertilizer, and A is, for simplicity, an indicator of high water usage. In this setting, Δ_{dce}^s captures the average direct causal effect of using fertilizer on the crop yield, where the effect weights units with high and low water usage according to the respective probabilities of

these actions happening for the treated, $\pi_1(a)$. The term Δ_{ind}^s , in turn, captures a piece of the causal effect of water usage on crop yield that depends on the magnitude of differential water usage between the treated and the untreated. If water usage causally improves crop yield in the absence of fertilizer, and getting an exogenous fertilizer incentivizes units to increase their water usage, this term would be positive.

The following theorem summarizes our discussion above.

Theorem 4.1. Consider the short regression in (6) and assume $P\{D = d, A = a\} > 0$ for all $(d, a) \in \mathcal{D} \times \mathcal{A}$. Then, Δ_{short} can be decomposed as in (12)-(15). If Assumption 3.2 holds, then $\Delta_{\text{sel}}^s = 0$ and Δ_{dce}^s and Δ_{ind}^s simplify to the expressions in (17) and (18).

Remark 4.2. It is important to note that, even under the exogeneity condition in Assumption 3.2, the parameter Δ_{short} does not satisfy strong sign preservation as defined in Definition 3.3. Indeed, it is certainly possible that $\mu(1, a) - \mu(0, a) > 0$ for all $a \in \mathcal{A}$ and yet $\Delta_{\text{short}} < 0$ due to $\Delta_{\text{ind}}^s < -\Delta_{\text{dce}}^s < 0$. This phenomenon, which is reminiscent of the Simpson’s paradox, is present in similar ways in other causal inference settings; including two-way fixed effects as in [De Chaisemartin and d’Haultfoeuille \(2020\)](#) or the surrogate paradox as in [VanderWeele \(2013\)](#).² ■

Remark 4.3. Under Assumption 3.2 and $A(d) \perp D$, Δ_{short} is a linear combination of average partial causal effects, and it captures a “total” effect rather than a “partial” effect, as discussed in Remark 3.1. To understand this, notice that Δ_{ind} in (18) is the product of $\pi_1(a) - \pi_0(a)$ and $\mu(0, a) - \mu(0, 0)$ for each $a \in \mathcal{A}$. Both of these terms are partial effects, where $\pi_1(a) - \pi_0(a)$ is the average partial effect of D on A and $\mu(0, a) - \mu(0, 0)$ is the average partial effect of moving A from 0 to a on the outcome Y for units with $D = 0$. With this interpretation, Δ_{short} captures a total effect of D on Y that adds up the direct effect of D on Y , captured by Δ_{dce}^s , and the indirect effect that D has on Y via its effect on A and how A affects Y . This distinction between partial and total effects mimics the usual one associated with total and partial derivatives in mathematical analysis. Whether total or partial effects are most relevant will depend on the application at hand; see, for example, [Manski \(1997\)](#); [Heckman \(2000\)](#); [Imai et al. \(2010\)](#); [Glynn \(2012\)](#). That said, the premise of this paper is that regression-based estimands like Δ_{short} are often used with the goal of recovering an average direct causal effect, as defined in Definition 3.2. Our objective is therefore to clarify the conditions under which such interpretations are valid. ■

Assumption 3.1 suffices to interpret a difference-in-means as the total (average) treatment effect $E[Y(1) - Y(0)]$, but it is generally insufficient to remove selection effects—see Example A.1. Moreover, as emphasized in the empirical mediation literature, direct and indirect effects cannot be separately identified in randomized experiments without additional assumptions; see [Robins and Greenland \(1992\)](#). Similar recognition has emerged in applied work, particularly in development

²As pointed out by a referee, strong sign preservation would hold under additional assumptions that ensure $\text{sign}(\Delta_{\text{ind}}^s) = \text{sign}(\Delta_{\text{dce}}^s)$. Examples include monotonicity in $\pi_d(a)$ and $\mu(0, a)$, or linearity assumptions on $\mu(d, a)$, as we discuss in detail in the next section.

economics—for example, [Mel et al. \(2009\)](#), [Duflo et al. \(2011\)](#), and [Beaman et al. \(2013\)](#). For these reasons, we prioritize results derived under [Assumption 3.2](#), which enables clean interpretations of regression-based estimands in terms of direct and indirect causal effects.

4.2 Long regression

A seemingly natural, and certainly popular, way to mitigate the presence of indirect effects and obtain an estimand that satisfies strong sign preservation is to control for the other actions linearly as in (7); an approach we called the long regression. Our main result below shows that the slope coefficient Δ_{long} in (7) admits a decomposition similar to that derived by Δ_{short} , and thus includes a combination of direct effects and indirect effects. However, except in some special cases, the coefficients multiplying each average partial causal effect, as in [Definition 3.1](#), could be negative and so Δ_{long} may fail to satisfy strong sign preservation even in the absence of indirect effects. We formalize this below and provide a proof in [Appendix A.1](#).

Theorem 4.2. Let [Assumption 3.2](#) hold and assume that $P\{D = d, A = a\} > 0$ for all $(d, a) \in \mathcal{D} \times \mathcal{A}$ and that the covariance matrix of (D, A) is positive definite. Then, the coefficient Δ_{long} in (7) admits the decomposition

$$\Delta_{\text{long}} = \Delta_{\text{dce}}^1 + \Delta_{\text{ind}}^1, \quad (19)$$

where

$$\Delta_{\text{dce}}^1 \equiv \sum_{a \in \mathcal{A}} \omega_{\text{dce}}^1(a) (\mu(1, a) - \mu(0, a)) \quad (20)$$

$$\Delta_{\text{ind}}^1 \equiv \sum_{a \in \mathcal{A}} \omega_{\text{ind}}^1(a) (\mu(0, a) - \mu(0, 0)), \quad (21)$$

and $\{\omega_{\text{dce}}^1(a) : a \in \mathcal{A}\}$ and $\{\omega_{\text{ind}}^1(a) : a \in \mathcal{A}\}$ are as defined in [Theorem A.1](#) and satisfy $\sum_{a \in \mathcal{A}} \omega_{\text{dce}}^1(a) = 1$ and $\sum_{a \in \mathcal{A}} \omega_{\text{ind}}^1(a) = 0$. Furthermore, the following statements are equivalent:

- (a) A are mutually exclusive binary variables, i.e., $\mathcal{A}_j = \{0, 1\}$ for $j = 1, \dots, K$ and $A_j A_l = 0$ for all $j, l = 1, \dots, K$ with $j \neq l$.
- (b) For any distribution of (A, D) , $\omega_{\text{dce}}^1(a) \geq 0$ for all $a \in \mathcal{A}$.
- (c) For any distribution of (A, D) , $\omega_{\text{ind}}^1(a) = 0$ for all $a \in \mathcal{A}$.
- (d) For any distribution of (A, D) , Δ_{long} satisfies strong sign preservation.

[Theorem 4.2](#) shows that Δ_{long} can be decomposed into direct and indirect effects, but it leaves open the possibility that the coefficients entering each of these terms could, in general, be negative. An immediate implication is that, except in the special case where the actions in A are all mutually exclusive binary variables, which includes the case where A is a scalar binary variable as a special

case, the term Δ_{dce}^1 could be negative even if $\mu(1, a) - \mu(0, a) > 0$ for all $a \in \mathcal{A}$. This is because $\omega_{\text{dce}}(a)$ may be negative for some $a \in \mathcal{A}$. As a result, Δ_{long} generally does not satisfy strong sign preservation for the following two reasons. First, it may be possible that $\Delta_{\text{ind}}^1 < -\Delta_{\text{dce}}^1$, so that the indirect effect dominates the direct effect. This phenomenon is the same as the one we discussed for the short regression. Second, even in the absence of indirect effects, where $\Delta_{\text{ind}}^1 = 0$, the term Δ_{dce}^1 could be negative by itself even if $\mu(1, a) - \mu(0, a) > 0$ for all $a \in \mathcal{A}$, due to $\omega_{\text{dce}}(a) < 0$ for some $a \in \mathcal{A}$. This second possibility represents a stark distinction between the long regression estimand, Δ_{long} , and the short regression estimand, Δ_{short} , in the sense that Δ_{long} does not even measure a total causal effect of D on Y without additional assumptions, cf. Remark 4.3. It is also important to emphasize that this result does not depend on the distribution of Y given (A, D) . As a consequence, both Δ_{long} and Δ_{dce}^1 can be arbitrarily negative—even if $\mu(1, a) - \mu(0, a) > 0$ for all $a \in \mathcal{A}$ —when the actions are not mutually exclusive and binary.

We emphasize that the results in Theorem 4.2 are all *equivalent*. Obtaining an easy-to-interpret characterization of Δ_{long} in terms of partial effects when the actions are all binary and mutually exclusive should not come as particularly surprising given related results on the properties of this estimand in different but related settings (see, for example, Zhao and Ding, 2022; Goldsmith-Pinkham et al., 2022). To the best of our knowledge, however, a novel lesson of Theorem 4.2 is that this is precisely *the only way* to guarantee a proper weighted average representation for Δ_{long} for any distribution of (A, D) . This result could also be interpreted in the context of “bad controls” discussed in Angrist and Pischke (2008), as it clarifies the unique way in which seemingly bad controls could actually become “good controls”.

Remark 4.4. The role of Assumption 3.2 in our decompositions is purely to simplify the expressions and facilitate their interpretation. Removing this assumption leads to a decomposition of Δ_{long} that differs from Theorem 4.2 in three key ways. First, the term Δ_{dce}^1 becomes a linear combination of expectations that condition on $\Omega = \{D = 1, A = a\}$. Second, the interpretation of Δ_{ind}^1 becomes more convoluted, for the same reasons discussed earlier in the context of Δ_{ind}^s . Third, the decomposition includes an additional selection term that is conceptually identical to Δ_{sel}^s in the short regression case. Moreover, without Assumption 3.2, the equivalence between (a), (b), and (c) in Theorem 4.2 remains valid. However, in this case, the selection effect that appears in the decomposition breaks down the equivalence between (d) and the other items. For details on the decomposition of Δ_{long} without imposing Assumption 3.2, see Theorem A.1 in Appendix A.1. ■

The possibility of $\omega_{\text{dce}}^1(a)$ being negative for some $a \in \mathcal{A}$ does not depend on pathological data-generating processes. It can occur in relatively simple settings with reasonable distributions for (A, D) . This raises a significant concern for the use of linear-in- A regressions, as they can lead to results that are difficult to interpret and, in many cases, offer no improvement over the short regression in (6). Below we illustrate this situation with two canonical simple cases: one where A is a scalar random variable taking multiple values, and another one where $A = (A_1, A_2)$ with A_1 and A_2 being binary.

Consider first the case where $A = A_1$ is a scalar random variable taking values in $\mathcal{A}_1 = \{0, 1, 2, \dots, \bar{a}_1\}$. The regression in (7) simplifies to

$$Y = \Delta_{\text{long}}D + \theta_0 + \theta_1 A_1 + V .$$

Theorem A.1 provides general closed-form expressions for $\{\omega_{\text{dce}}^1(a) : a \in \mathcal{A}\}$ and $\{\omega_{\text{ind}}^1(a) : a \in \mathcal{A}\}$ that, when applied to this specific example, lead to

$$\omega_{\text{dce}}^1(a) \propto \left(\pi_1(a) - \frac{\text{Cov}(D, A_1)(a - E[A_1])}{\text{Var}(A_1)(1 - p)} \right) , \quad (22)$$

where $p = P\{D = 1\}$. From this expression, it follows that any distribution of (A, D) for which $\text{Cov}(D, A_1)(a - E[A_1]) > \text{Var}(A_1)(1 - p)\pi_1(a)$, would lead to negative weights. For example, consider the case where $p = 0.8$, $\bar{a}_1 = 3$, $\{A|D = 1\} \sim \text{Bi}(3, 0.8)$, and $\{A|D = 0\} \sim \text{Bi}(3, 0.2)$, where $\text{Bi}(n, \pi)$ denotes a Binomial distribution with n trials and success probability π . In this case, $\omega_{\text{dce}}^1(3) = -0.41 < 0$. Figure 2 plots the weights $\omega_{\text{dce}}^1(a)$ as a function of p and shows that $\omega_{\text{dce}}^1(3)$ is negative for any $p > 0.4$ in this example.

Next, consider the case where $A = (A_1, A_2)$ with A_1 and A_2 both being binary variables, so that $\mathcal{A} = \{(0, 0), (1, 0), (0, 1), (1, 1)\}$. The regression in (7) simplifies to

$$Y = \Delta_{\text{long}}D + \theta_0 + \theta_1 A_1 + \theta_2 A_2 + V .$$

The closed-form expressions for $\{\omega_{\text{dce}}^1(a) : a \in \mathcal{A}\}$ and $\{\omega_{\text{ind}}^1(a) : a \in \mathcal{A}\}$ derived in Theorem A.1 also simplify to this case and lead to simple conditions for which

$$\omega_{\text{dce}}^1(1, 1) = -\omega_{\text{dce}}^1(1, 0) . \quad (23)$$

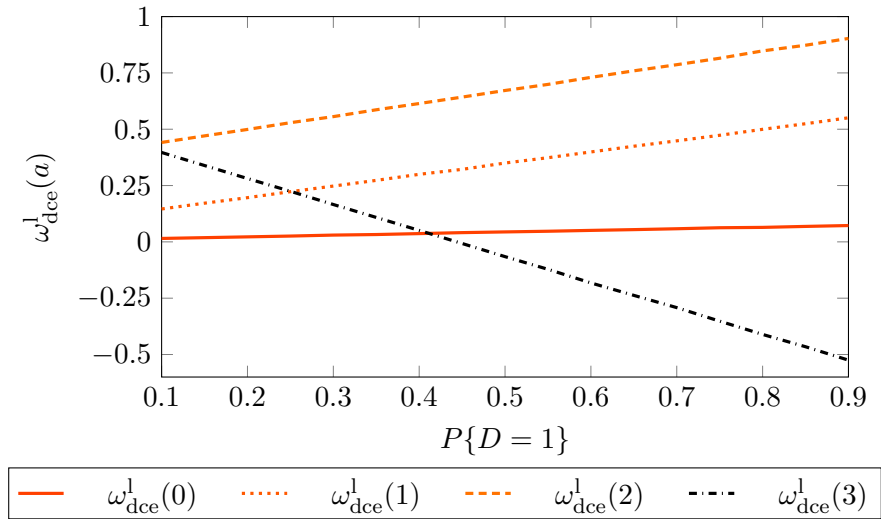


Figure 2: Weights $\omega_{\text{dce}}^1(a)$ as a function of p when $\{A|D = 1\} \sim \text{Bi}(3, 0.8)$, and $\{A|D = 0\} \sim \text{Bi}(3, 0.2)$

That is, whenever one of the average partial causal effects gets a positive weight, the other necessarily gets a negative one. As an illustrative example, when $\text{Cov}(A_1, A_2) = 0$,

$$P\{D = 1\} = P\{A_2 = 1\} = \frac{1}{2}, \quad P\{A_1 = 1 \mid D = 1\} = 2P\{A_1 = 1\}, \quad (24)$$

and

$$P\{A_1 = A_2 = 1 \mid D = 1\} = P\{A_1 = 1, A_2 = 0 \mid D = 1\} = \frac{1}{4}. \quad (25)$$

Using the expressions in Theorem A.1, we obtain $\omega_{\text{dce}}^1(1, 0) = -\omega_{\text{dce}}^1(1, 1) = -0.30$, which one more time illustrates that negative weights arise naturally in settings with non-pathological DGPs. In the proof of Theorem 4.2, we present even simpler counter-examples that also illustrate how the weights $\{\omega_{\text{ind}}^1(a) : a \in \mathcal{A}\}$ are generally non-zero and potentially negative, as well as how $\omega_{\text{dce}}^1(a)$ may be negative without necessarily satisfying (23).

Remark 4.5. It is important to note that, even under the stronger exogeneity condition in Assumption 3.2, the parameter Δ_{long} does not generally satisfy strong sign preservation as defined in Definition 3.3. Indeed, it is certainly possible that $\mu(1, a) - \mu(0, a) > 0$ for all $a \in \mathcal{A}$ and yet $\Delta_{\text{long}} < 0$ due to either $\Delta_{\text{ind}}^1 < -\Delta_{\text{dce}}^1 < 0$ or simply $\Delta_{\text{dce}}^1 < 0$ because of negative weights $\{\omega_{\text{dce}}^1(a) : a \in \mathcal{A}\}$. This second condition implies that *not even* Δ_{dce}^1 satisfies strong sign preservation and thus, in general, Δ_{long} does not offer much of a benefit relative to Δ_{short} , as Δ_{short} can at least be interpreted as a total effect, as discussed in Remark 4.3. ■

Remark 4.6. As discussed in Remark 4.1, the long regression in (7) is used extensively in the social sciences and the mediation literature. In economics, Heckman et al. (2013, Eq. (6)) consider a long regression in the context of a more restrictive model for potential outcomes that is linear and separable in (d, a) . More recently, Fagereng et al. (2021, Eq. (7)) use the same mediation model from Heckman et al. (2013), in combination with the long regression in (7), to disentangle the average causal effect on outcomes into direct and indirect effects. The “direct” causal interpretation assigned to the estimands in this last set of papers is correct under the modeling assumptions for potential outcomes, despite both applications involving actions A that are multidimensional and non-mutually exclusive. Our results, however, imply that the main conclusions from such an analysis delicately rely on a linear model for $\mu(d, a)$ and do not generally extend to nonlinear ones. ■

The results in Theorem 4.2 are, to the best of our knowledge, novel—though they are certainly related in spirit to a large literature in biostatistics and epidemiology concerned with mediators and indirect effects. Given the breadth of that literature, we highlight a few representative contributions rather than attempting an exhaustive review.³ While our results are thematically aligned

³Two illustrative examples that were suggested by a referee are Groenwold et al. (2021) and Wilcox et al. (2011). The first paper presents a decompositions of certain regression estimands similar in form to ours, but do so under the assumption that all variables are linearly related to one another—an assumption that, as we show below, has strong implications for interpretation. The second paper warns against including certain variables (such as gestational age) as controls in regression analyses, noting the potential for collider bias. Their critique is motivated by a specific empirical context and supported by simulation, but they do not offer a formal decomposition nor do they analyze the unadjusted case (analogous to our short regression).

with several ideas in this literature, we are not aware of prior work that establishes the same characterizations with the level of formality and generality presented here. A key contribution of our paper is to show that the case in which the components of A are binary and mutually exclusive is essentially the *only* one in which the long regression can be interpreted as capturing direct causal effects—without requiring additional assumptions such as linearity. This insight, which is formalized in Theorem 4.2, sharpens and unifies a number of informal arguments that appear across distinct literatures.

Perhaps the most relevant work for our purposes is Imai et al. (2010), who study the interpretation of the long regression—originally popularized by Baron and Kenny (1986)—under a linear potential outcome model and an assumption they call sequential ignorability. We restate that assumption below to facilitate comparison.

Assumption 4.1 (Sequential Ignorability). Let $A(d)$ denote the potential outcome for A and assume that: (i) $(Y(d', a), A(d)) \perp D | X = x$, and (ii) $Y(d', a) \perp A(d) | D = d, X = x$, both for $d, d' = 0, 1$ and all x .

The results in Imai et al. (2010) about the long regression in (7) invoke (a) sequential ignorability, (b) a scalar random variable A (though not necessarily binary), and (c) a linear model for $\mu(d, a)$ in (d, a) . Under these conditions (a)-(c), and ignoring the X for simplicity, Imai et al. (2010, Theorem 2) shows that Δ_{long} identifies $\bar{\zeta} = \bar{\zeta}(1) = \bar{\zeta}(0)$ where $\bar{\zeta}(d) \equiv E[Y(1, A(d))] - E[Y(0, A(d))] = \sum_{a \in \mathcal{A}} (\mu(1, a) - \mu(0, a)) \pi_d(a)$, and the equality follows from Assumption 4.1 implying $Y(d', a) \perp A(d)$; see Lemma A.2 in Appendix A.2. The linear model for $\mu(d, a)$ implies that the difference $\mu(1, a) - \mu(0, a)$ does not depend on the value of a , and so it is just a constant that we can denote by $\bar{\zeta}$ without loss of generality. The fact that $\bar{\zeta} = \bar{\zeta}(1) = \bar{\zeta}(0)$ then follows from $\sum_{a \in \mathcal{A}} \pi_d(a) = 1$ for $d \in \{0, 1\}$.

The additional assumptions (a)-(c) mentioned above have implications on the conclusions of Theorem 4.2, which does not invoke any of these assumptions. In particular, the linearity of $\mu(d, a)$ implies that Δ_{dce}^1 in (20) equals $\bar{\zeta} \sum_{a \in \mathcal{A}} \omega_{\text{dce}}^1(a) = \bar{\zeta}$ by the weights adding up to one according to Theorem 4.2. The same linearity assumption also implies that

$$\Delta_{\text{ind}}^1 \equiv \sum_{a \in \mathcal{A}} \omega_{\text{ind}}^1(a) (\mu(0, a) - \mu(0, 0)) \propto \sum_{a \in \mathcal{A}} \omega_{\text{ind}}^1(a) a = 0 ,$$

where the last equality follows from Theorem A.1 in the appendix. We conclude that Theorem 4.2 coincides with the results in Imai et al. (2010) in delivering Δ_{long} being equal to $\bar{\zeta}$ under the additional assumption that $\mu(d, a)$ is linear in (d, a) . This means that, while sequential ignorability is a stronger assumption than Assumption 3.2 (we prove this claim in Lemma A.1 in Appendix A.2), the main driving force of this result is the linear model for the potential outcomes or, equivalently, the linear model for $\mu(d, a)$.⁴ As we discussed in Remark 4.6, this linearity assumption has been

⁴Sequential ignorability without a linear model has also been used in the literature on causal mediation analysis to construct semi-parametrically efficient estimators; see Tchetgen Tchetgen and Shpitser (2012).

used in economic applications, e.g., Heckman et al. (2013); Fagereng et al. (2021), and while it imposes enough restrictions to provide a clean interpretation to the coefficient Δ_{long} , our results show that such a clean interpretation generally breaks down when $\mu(d, a)$ is not linear in (d, a) .

Remark 4.7. We note that while Theorem 4.2 is a result on how to properly interpret Δ_{long} in the context of a long regression, Imai et al. (2010, Theorem 2) is a result on the identification of *natural* indirect effects via the same type of regression and the additional conditions (a)-(c) above. Related to our discussion in Remark 3.2, the *natural* indirect effect does not coincide with our notion of indirect effect in Theorem 4.2. To see the difference, note that the *natural* indirect effect, defined as $\bar{\delta}(d) = E[Y(d, A(1)) - Y(d, A(0))]$, can be written as

$$\bar{\delta}(d) = \sum_{a \in \mathcal{A}} (\mu(d, a) - \mu(d, 0))(\pi_1(a) - \pi_0(a)) , \quad (26)$$

under sequential ignorability, and is distinct from Δ_{ind}^1 in (21) because $\omega_{\text{ind}}^1(a) \neq \pi_1(a) - \pi_0(a)$. Conceptually, the literature on mediation analysis defines an indirect effect as a target parameter and then determines conditions under which such indirect effects could be identified from the data. In contrast, we characterize the decomposition of estimands in terms of average direct causal effects, as defined in Definition 3.2, and then group the remaining terms as indirect or selection terms, depending on the case. We note, however, that our indirect effects coincide with those characterized by $\bar{\delta}(d)$ in the case of the short regression, i.e., Δ_{ind}^s in (18) equals $\bar{\delta}(0)$. ■

Remark 4.8. Under the conditions of Theorem 4.2 and assuming that $\mu(d, a)$ is linear, it follows that $\Delta_{\text{ind}}^1 = 0$, and so Δ_{long} captures the direct causal effect. In this case, Δ_{short} equals the direct effect plus the indirect effect, justifying the common practice of using $\Delta_{\text{short}} - \Delta_{\text{long}}$ to estimate the indirect effect. However, Theorems 4.1 and 4.2 show that this interpretation fails when $\mu(d, a)$ is nonlinear. Then, (i) the direct effects identified by the two regressions may differ, and (ii) Δ_{long} may itself include an indirect component if actions are not mutually exclusive and binary. As a result, $\Delta_{\text{short}} - \Delta_{\text{long}}$ no longer isolates the indirect effect. ■

Remark 4.9. Our findings conceptually align with prior critiques in the literature on mediation analysis regarding the limitations of interpreting regression coefficients as causal direct effects. In particular, Kaufman et al. (2004) provide an early warning against using covariate adjustment to infer biological mediation. Using a potential outcomes framework, they show that even in randomized experiments, adjusting for post-treatment variables can yield misleading inferences unless strong and often unrealistic assumptions hold—such as no unmeasured mediator-outcome confounding or no interaction at the unit level. Kaufman (2009) further advocates for grounding mediation analysis in formal causal estimands, such as controlled direct effects, rather than relying on statistical associations. These contributions reinforce our argument that regression-based estimands like Δ_{short} and Δ_{long} may not recover interpretable causal effects in the absence of strong identifying assumptions. ■

4.3 Long regression with interactions

A common variant of the long regression additionally includes interactions between the K actions, A_1, \dots, A_K , and the treatment D ; that is, the slope coefficient Δ_{inter} in (8). We refer to this as the long regression with interactions. Our result below shows that the slope coefficient Δ_{inter} in (8) admits a decomposition with the same shortcomings as the one we derived for Δ_{long} , including the possibility of Δ_{inter} not satisfying strong sign preservation, even in the absence of indirect effects. We formalize this below and provide a proof in Appendix A.1.

Theorem 4.3. Let Assumption 3.2 hold and assume that $P\{D = d, A = a\} > 0$ for all $(d, a) \in \mathcal{D} \times \mathcal{A}$ and that the covariance matrix of (D, A, AD) is positive definite. Then, the coefficient Δ_{inter} in (8) admits the decomposition

$$\Delta_{\text{inter}} = \Delta_{\text{dce}}^i + \Delta_{\text{ind}}^i, \quad (27)$$

where

$$\begin{aligned} \Delta_{\text{dce}}^i &\equiv \sum_{a \in \mathcal{A}} \omega_{\text{dce}}^i(a) (\mu(1, a) - \mu(0, a)) \\ \Delta_{\text{ind}}^i &\equiv \sum_{a \in \mathcal{A}} \omega_{\text{ind}}^i(a) (\mu(0, a) - \mu(0, 0)), \end{aligned}$$

and $\{\omega_{\text{dce}}^i(a) : a \in \mathcal{A}\}$ and $\{\omega_{\text{ind}}^i(a) : a \in \mathcal{A}\}$ are as defined in Theorem A.2 and satisfy $\sum_{a \in \mathcal{A}} \omega_{\text{dce}}^i(a) = 1$ and $\sum_{a \in \mathcal{A}} \omega_{\text{ind}}^i(a) = 0$. Furthermore, the following statements are equivalent:

- (a) A are mutually exclusive binary variables, i.e., $\mathcal{A}_j = \{0, 1\}$ for $j = 1, \dots, K$ and $A_j A_l = 0$ for all $j, l = 1, \dots, K$ with $j \neq l$.
- (b) For any distribution of (A, D) , $\omega_{\text{dce}}^i(a) \geq 0$ for all $a \in \mathcal{A}$.
- (c) For any distribution of (A, D) , $\omega_{\text{ind}}^i(a) = 0$ for all $a \in \mathcal{A}$.
- (d) For any distribution of (A, D) , Δ_{inter} satisfies strong sign preservation.

Theorem 4.3 is analogous to Theorem 4.2 and has very similar implications. Except in the special case where the actions in A are all mutually exclusive binary variables, which includes the case where A is a scalar binary variable as a special case, the term Δ_{dce}^i could be negative even if $\mu(1, a) - \mu(0, a) > 0$ for all $a \in \mathcal{A}$. This is because $\omega_{\text{dce}}^i(a)$ may be negative for some $a \in \mathcal{A}$.⁵ As a result, Δ_{inter} generally does not satisfy strong sign preservation for the same two reasons that Δ_{long} fails to do so. First, it is possible that $\Delta_{\text{ind}}^i < -\Delta_{\text{dce}}^i$, and second, even if $\Delta_{\text{ind}}^i = 0$, the term Δ_{dce}^i itself can be negative—despite $\mu(1, a) - \mu(0, a) > 0$ for all $a \in \mathcal{A}$ —due to negative weights. Again, this second failure mode distinguishes Δ_{long} and Δ_{inter} from Δ_{short} .

⁵The possibility of $\omega_{\text{dce}}^i(a)$ being negative for some $a \in \mathcal{A}$ can occur in simple settings with reasonable distributions for (A, D) ; see the proof of Theorem 4.3.

While Theorem 4.3 focuses on the properties of the estimand Δ_{inter} in settings with interactions terms, it is most often the case that the analyst would rather focus on the estimand $\Delta_{\text{inter}} + \sum_{j=1}^K \lambda_j E[A_j]$ (or, simply, $\Delta_{\text{inter}} + \sum_{j=1}^K \lambda_j a_j$ for given values $a_j, j = 1, \dots, K$). In Lemma A.4 in Appendix A.2, we show that

$$\Delta_{\text{inter}} + \sum_{j=1}^K \lambda_j E[A_j] = \sum_{a \in \mathcal{A}} \omega_{\text{dce}}^{\text{i}\star}(a) (E[Y(1, a) - Y(0, a)]) + \omega_{\text{ind}}^{\text{i}\star}(a) (E[Y(0, a) - Y(0, 0)]) ,$$

where the “weights” $\{(\omega_{\text{dce}}^{\text{i}\star}(a), \omega_{\text{ind}}^{\text{i}\star}(a)) : a \in \mathcal{A}\}$ may be negative in general, and thus leading to an estimand with similar properties to those of Δ_{inter} . The one special case where the estimand $\Delta_{\text{inter}} + \sum_{j=1}^K \lambda_j E[A_j]$ identifies the intended causal effects is when $\mu(d, a)$ is assumed to take the functional form $\mu(d, a) = \kappa_0 + \kappa_1 d + \sum_{j=1}^K \kappa_{2,a} a_j + d \sum_{j=1}^K \kappa_{3,j} a_j$, which is equivalent to assuming that the conditional mean of the observed outcome, Y , is correctly specified in the interaction regression in (8). Lemma A.5 in Appendix A.2 shows that $\Delta_{\text{inter}} + \sum_{j=1}^K \lambda_j a_j = \mu(1, a) - \mu(0, a)$ in this case, delivering an average partial causal effect given $a = (a_1, \dots, a_K)$, as defined in Definition 3.1. It follows from these results that a clean interpretation of Δ_{inter} or $\Delta_{\text{inter}} + \sum_{j=1}^K \lambda_j a_j$ in terms of the definitions introduced in Section 3 essentially depends on a correctly specified parametric model for potential outcomes and does not generally apply to nonlinear models, similarly to our results about the long regression in Section 4.2.

Remark 4.10. The observations made in Remark 4.4 for Δ_{long} also extend to Δ_{inter} . In particular, one can derive a decomposition of Δ_{inter} without imposing Assumption 3.2. Dropping this assumption leads to a decomposition that differs from Theorem 4.3 in three key ways. First, the term $\Delta_{\text{dce}}^{\text{i}}$ becomes a linear combination of expectations conditional on $D = 1, A = a$. Second, the interpretation of $\Delta_{\text{ind}}^{\text{i}}$ becomes convoluted, for the same reasons discussed in the context of $\Delta_{\text{ind}}^{\text{s}}$ and $\Delta_{\text{ind}}^{\text{l}}$. Finally, the decomposition includes an additional selection term that is conceptually identical to those appearing in $\Delta_{\text{ind}}^{\text{s}}$ and $\Delta_{\text{ind}}^{\text{l}}$. As with the long regression case, the equivalence between (a), (b), and (c) in Theorem 4.3 remains valid even without Assumption 3.2. However, in the absence of this assumption, Δ_{inter} includes the selection term, which breaks the equivalence between (d) and the other items. For details on the decomposition of Δ_{long} without Assumption 3.2, see Theorem A.2 in Appendix A.1. ■

Remark 4.11. Similar to the long regression in (7) that we discussed in Remarks 4.1 and 4.6, the interaction regression is used extensively in the mediation literature. In the context of mediation analysis, this variant has been popularized and advocated by Judd and Kenny (1981); Kraemer et al. (2002, 2008). However, the main goal in that particular setting has been to test for the existence of mediation effects using the estimated coefficients in (8), see Kraemer et al. (2008) for details on the proposed test and Imai et al. (2010) for a result that shows that, under Assumption 4.1, such a test does not provide evidence in favor or against the parameter $\bar{\delta}(d)$ in (26) being zero. Our results, on the other hand, imply that even in settings where mediation effects are a nuisance and the main goal is to interpret the coefficients directly related to the treatment D , the main

conclusions depend on the distribution of (A, D) . ■

4.4 Strata fixed effects (SFE) regression

Theorems 4.2 and 4.3 show that adding other actions linearly in the regression is generally useful only when the actions are binary and mutually exclusive. If we could make the actions binary and mutually exclusive, we would obtain an estimand free from indirect effects and with strong sign preservation. This is possible by considering the slope coefficient Δ_{sfe} in (9), where the regression controls for all possible values of A , i.e., $\{I\{A = a\} : a \in \mathcal{A}\}$. We call this a strata-fixed effects regression, due to its connection to the use of strata fixed effects in randomized controlled trials with covariate adaptive randomization; see Bugni et al. (2018, 2019). Our result below shows that Δ_{sfe} in (9) has a decomposition free from indirect effects (see Appendix A.1 for a proof).

Theorem 4.4. Let Assumption 3.2 hold and assume that $P\{D = d, A = a\} > 0$ for all $(d, a) \in \mathcal{D} \times \mathcal{A}$. Let $\pi_d(a)$ be defined as in (11). Then

$$\Delta_{\text{sfe}} = \sum_{a \in \mathcal{A}} \omega_{\text{sfe}}(a) (\mu(1, a) - \mu(0, a)) ,$$

where the weights $\{\omega_{\text{sfe}}(a) : a \in \mathcal{A}\}$ are given by

$$\omega_{\text{sfe}}(a) \equiv \frac{\pi_1(a)\pi_0(a)/p_a}{\sum_{a' \in \mathcal{A}} \pi_1(a')\pi_0(a')/p_{a'}} \quad (28)$$

for $p_a \equiv P\{A = a\}$, and satisfy $\sum_{a \in \mathcal{A}} \omega_{\text{sfe}}(a) = 1$ and $\omega_{\text{sfe}}(a) \geq 0$. Furthermore, Δ_{sfe} satisfies strong sign preservation.

Theorem 4.4 shows that Δ_{sfe} identifies an average direct causal effect as in Definition 3.2. Importantly, it does not contain indirect effects and, as a result, Δ_{sfe} satisfies strong sign preservation as in Definition 3.3. The weight $\omega_{\text{sfe}}(a)$ admits a simple representation and depends only on the conditional probabilities that the action a happens for the treated and control group, $\pi_1(a)$ and $\pi_0(a)$. These weights are generally different than the weights associated with the direct effect in the short regression, $\Delta_{\text{dce}}^{\text{s}}$, which are simply $\pi_1(a)$, unless D and A are independent. We emphasize that the result in Theorem 4.4 does not require the other actions, A , to be singled-valued or mutually exclusive.

The estimand Δ_{sfe} has been studied in other settings. For example, Angrist (1998) studies regressions of Y on D under Assumption 3.1 and considers a linear regression that saturates on the covariates, X , to obtain an expression that parallels the one in Theorem 4.4; see Angrist (1998, footnote 11). Indeed, Assumption 3.2 implies that $Y(d, a) \perp D|A$ which makes the connection to Angrist (1998) immediate. In this sense, our result illustrates another instance where stratifying confounding variables leads to easy-to-interpret results (though, it is important to notice that

this result relies on the fact that the treatment variable D in our case is binary, c.f. [Goldsmith-Pinkham et al. \(2022\)](#)). As another example, [Bugni et al. \(2018, 2019\)](#) study the properties of this estimand and present results on how to properly compute standard errors in randomized controlled experiments with covariate adaptive randomization. These papers, however, do not represent Δ_{sfe} as a weighted average of causal effects since the strata are not viewed as counter-factual features in such experiments.

4.5 Saturated (SAT) regression

We now turn our attention to the last set of estimands we study in this paper; the slope coefficient Δ_{sat} in (10). As we have stated in the introduction, under Assumption 3.2, it follows that $\mu(d, a)$ is immediately identified from $E[Y|D = d, A = a]$ for any $d \in \{0, 1\}$ and $a \in \mathcal{A}$ and so identification of any contrast of means $\mu(d, a)$ is straightforward. From this, it immediately follows that the same result could be achieved by running a saturated regression, as in (10), that we re-write here for readability,

$$Y = \sum_{a \in \mathcal{A}} \gamma(a) I\{A = a\} + \sum_{a \in \mathcal{A}} \Delta_{\text{sat}}(a) I\{A = a\} D + \epsilon .$$

Under Assumption 3.2, standard results on saturated regressions imply that $\Delta_{\text{sat}}(a) = \mu(1, a) - \mu(0, a)$ for all $a \in \mathcal{A}$, and so $\Delta_{\text{sat}}(a)$ captures an average causal partial effect of D on Y for each value of the other actions, $a \in \mathcal{A}$, aligned with Definition 3.1. For completeness, we state and prove this result formally in Theorem A.4 in the appendix. The same result also shows that replacing Assumption 3.2 with Assumption 3.1 leads to a decomposition of $\Delta_{\text{sat}}(a)$ that includes a selection term, as it was also the case for the other estimands we considered. Finally, we note that this approach is equivalent to a multi-way ANOVA, since the mean for the outcome variable depends exclusively on categorical variables.

5 Concluding Remarks

In this paper we analyze settings where the researcher seeks to estimate the average *direct* causal effect of a binary treatment D on an outcome Y , in situations where the outcome is realized with a delay and other endogenous actions A may occur in the interim. A central premise of our analysis is that researchers often use regression-based estimands with the intent of isolating a direct causal effect of D on Y , not a total effect that also includes indirect pathways through A .

We focus on several popular estimands arising from regressions that control for A in various ways—linearly, with interactions, or not at all—and decompose them to assess whether they admit a causal interpretation. Our decompositions are algebraic and do not rely on identification assumptions, but we examine the additional conditions under which causal interpretations are possible. The results are “negative” in spirit: they show that, even under strong exogeneity assumptions,

these commonly used regression-based estimands generally do not recover the average direct causal effect of interest unless we impose restrictive assumptions (e.g., linearity in the potential outcomes model). We highlight that only in the special case where the actions are binary and mutually exclusive does the long or interacted regression estimand admit a causal interpretation of a direct effect without further assumptions—and our result shows that this is an *if and only if* condition.

The goal of our exercise is not to propose new estimators or to discuss nonparametric identification of direct and indirect effects, but rather to clarify the conditions under which standard approaches yield meaningful causal parameters. We hope this work helps practitioners recognize the limitations of common estimands and adopt a more deliberate approach when interpreting them as capturing direct effects.

References

- AKHTARI, M., CHEN, J., LEMIONET, A., NGUYEN, D., OBEID, H. and ZHU, Y. (2021). How Airbnb measures future value to standardize tradeoffs. *Medium.com*. URL <https://medium.com/airbnb-engineering/how-airbnb-measures-future-value-to-standardize-tradeoffs-3aa99a941ba5>.
- ANGELUCCI, C., MERAGLIA, S. and VOIGTLÄNDER, N. (2022). How merchant towns shaped parliaments. *American Economic Review*, **112** 3441–3487.
- ANGRIST, J. D. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, **66** 249–288.
- ANGRIST, J. D. and PISCHKE, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press.
- BARON, R. M. and KENNY, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, **51** 1173.
- BEAMAN, L., KARLAN, D., THUYSBAERT, B. and UDRY, C. (2013). Profitability of Fertilizer: Experimental Evidence from Female Rice Farmers in Mali. *American Economic Review*, **103** 381–386.
- BORUSYAK, K., JARAVEL, X. and SPIESS, J. (2022). Revisiting event study designs: Robust and efficient estimation. *Available at SSRN 2826228*.
- BUGNI, F. A., CANAY, I. A. and SHAIKH, A. M. (2018). Inference under covariate adaptive randomization. *Journal of the American Statistical Association (Theory & Methods)*, **113** 1741–1768.

- BUGNI, F. A., CANAY, I. A. and SHAIKH, A. M. (2019). Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics*, **10** 1741–1768.
- CALLAWAY, B. and SANT’ANNA, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of econometrics*, **225** 200–230.
- CANAY, I. A., MOGSTAD, M. and MOUNTJOY, J. (2023). On the use of outcome tests for detecting bias in decision making. Tech. rep., National Bureau of Economic Research.
- CHEN, H., GENG, Z. and JIA, J. (2007). Criteria for surrogate end points. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **69** 919–932.
- CHERNOZHUKOV, V., KASAHARA, H. and SCHRIMPF, P. (2021). Causal impact of masks, policies, behavior on early covid-19 pandemic in the us. *Journal of Econometrics*, **220** 23–62.
- DE CHAISEMARTIN, C. and D’HAULTFOEUILLE, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, **110** 2964–96.
- DUFLO, E., KREMER, M. and ROBINSON, J. (2011). Nudging Farmers to Use Fertilizer: Theory and Experimental Evidence from Kenya. *American Economic Review*, **101** 2350–2390.
- FAGERENG, A., MOGSTAD, M. and RØNNING, M. (2021). Why do wealthy parents have wealthy children? *Journal of Political Economy*, **129** 703–756.
- GLYNN, A. N. (2012). The product and difference fallacies for indirect effects. *American Journal of Political Science*, **56** 257–269.
- GOLDSMITH-PINKHAM, P., HULL, P. and KOLESÁR, M. (2022). Contamination bias in linear regressions. [2106.05024](#).
- GOODMAN-BACON, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, **225** 254–277.
- GROENWOLD, R. H. H., PALMER, T. M. and TILLING, K. (2021). Some recommendations for causal inference in observational studies of effects of prenatal exposures on birth and perinatal outcomes. *Paediatric and Perinatal Epidemiology*, **35** 308–310.
- HECKMAN, J., PINTO, R. and SAVELYEV, P. (2013). Understanding the mechanisms through which an influential early childhood program boosted adult outcomes. *American Economic Review*, **103** 2052–86.
- HECKMAN, J. J. (2000). Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective*. *Quarterly Journal of Economics*, **115** 45–97.
- IMAI, K., KEELE, L. and YAMAMOTO, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical science*, **25** 51–71.

- JAIN, D. and SINGH, S. S. (2002). Customer lifetime value research in marketing: A review and future directions. *Journal of interactive marketing*, **16** 34–46.
- JUDD, C. M. and KENNY, D. A. (1981). Process analysis: Estimating mediation in treatment evaluations. *Evaluation review*, **5** 602–619.
- KAUFMAN, J. S. (2009). Commentary: Gilding the black box. *International Journal of Epidemiology*, **38** 845–847.
- KAUFMAN, J. S., MACLEHOSE, R. F. and KAUFMAN, S. (2004). A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation. *Epidemiologic Perspectives & Innovations*, **1** 4. Open Access.
- KRAEMER, H., KIERNAN, M., ESSEX, M. and KUPFER, D. J. (2008). How and why criteria defining moderators and mediators differ between the baron & kenny and macarthur approaches. *Health Psychology*, **27** S101.
- KRAEMER, H. C., WILSON, G. T., FAIRBURN, C. G. and AGRAS, W. S. (2002). Mediators and moderators of treatment effects in randomized clinical trials. *Archives of general psychiatry*, **59** 877–883.
- LEWIS, R. A., RAO, J. M. and REILEY, D. H. (2011). Here, there, and everywhere: Correlated online behaviors can lead to overestimates of the effects of advertising. In *Proceedings of the 20th ACM International World Wide Web Conference*. ACM, 157–166.
- MANSKI, C. F. (1997). Monotone Treatment Response. *Econometrica*, **65** 1311.
- MEL, S. D., MCKENZIE, D. and WOODRUFF, C. (2009). Returns to Capital in Microenterprises: Evidence from a Field Experiment. *The Quarterly Journal of Economics*, **124** 423–423.
- MODERNA (2021). A Study of SARS CoV-2 Infection and Potential Transmission in Individuals Immunized With Moderna COVID-19 Vaccine (CoVPN 3006). *ClinicalTrials.gov Identifier: NCT04811664*. URL <https://clinicaltrials.gov/ct2/show/study/NCT04811664>.
- NUNN, N. (2008). The long-term effects of africa’s slave trades. *The Quarterly Journal of Economics*, **123** 139–176.
- PEARL, J. (2001). Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. UAI’01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 411–420.
- ROBINS, J. M. (2003). Semantics of causal dag models and the identification of direct and indirect effects. In *Highly Structured Stochastic Systems* (N. L. H. P. J. Green and S. Richardson, eds.). Oxford University Press, 70–81.

- ROBINS, J. M. and GREENLAND, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 143–155.
- SUN, L. and ABRAHAM, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, **225** 175–199.
- TCHETGEN TCHETGEN, E. J. and SHPITSER, I. (2012). Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of statistics*, **40** 1816.
- VANDERWEELE, T. J. (2013). Surrogate measures and consistent surrogates. *Biometrics*, **69** 561–565.
- VANDERWEELE, T. J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, New York.
- VANDERWEELE, T. J. and TCHETGEN TCHETGEN, E. J. (2014). Attributing effects to interactions. *Epidemiology (Cambridge, Mass.)*, **25** 711.
- VANDERWEELE, T. J., VANSTEELANDT, S. and ROBINS, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator–outcome confounder. *Epidemiology*, **25** 300–306.
- VOIGTLÄNDER, N. and VOTH, H.-J. (2012). Persecution perpetuated: the medieval origins of anti-semitic violence in nazi germany. *The Quarterly Journal of Economics*, **127** 1339–1392.
- WILCOX, A. J., WEINBERG, C. R. and BASSO, O. (2011). On the pitfalls of adjusting for gestational age at birth. *International Journal of Epidemiology*, **40** 1234–1241.
- XU, Y., CHEN, N., FERNANDEZ, A., SINNO, O. and BHASIN, A. (2015). From infrastructure to culture: A/B testing challenges in large scale social networks. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2227–2236.
- ZHAO, A. and DING, P. (2022). Regression-based causal inference with factorial experiments: estimands, model specifications and design-based properties. *Biometrika*, **109** 799–815.

A Appendix

A.1 Proofs

Proof of Theorem 4.1. This proof follows from derivations in Section 4.1 and basic algebraic manipulations.

■

Theorem A.1. Consider the long regression in (6) and assume that $P\{D = d, A = a\} > 0$ for all $(d, a) \in \mathcal{D} \times \mathcal{A}$. Assume that the variance-covariance matrix of (A, D) is positive definite and let $M = \text{Cov}(D, A) \text{Var}(A)^{-1}$. Then,

$$\begin{aligned} \Delta_{\text{long}} &= \sum_{a \in \mathcal{A}} \omega_{\text{dce}}^1(a) E[Y(1, a) - Y(0, a) | D = 1, A = a] \\ &+ \sum_{a \in \mathcal{A}} \omega_{\text{ind}}^1(a) (E[Y(0, a) | D = 0, A = a] - E[Y(0, 0) | D = 0, A = 0]) \\ &+ \sum_{a \in \mathcal{A}} \omega_{\text{dce}}^1(a) (E[Y(0, a) | D = 1, A = a] - E[Y(0, a) | D = 0, A = a]), \end{aligned} \quad (\text{A-1})$$

where

$$\begin{aligned} \omega_{\text{dce}}^1(a) &\equiv \frac{\pi_1(a) [\text{Var}(D) - P\{D = 1\} \sum_{j=1}^K M_j (a_j - E[A_j])]}{\text{Var}(D) - \text{Cov}(D, A) \text{Var}(A)^{-1} \text{Cov}(A, D)} \\ \omega_{\text{ind}}^1(a) &\equiv \frac{\text{Var}(D) [\pi_1(a) - \pi_0(a)] - P\{A = a\} \sum_{j=1}^K M_j (a_j - E[A_j])}{\text{Var}(D) - \text{Cov}(D, A) \text{Var}(A)^{-1} \text{Cov}(A, D)}, \end{aligned} \quad (\text{A-2})$$

and $\pi_d(a)$ is defined in (11). Furthermore, $\sum_{a \in \mathcal{A}} \omega_{\text{dce}}^1(a) = 1$, $\sum_{a \in \mathcal{A}} a \omega_{\text{ind}}^1(a) = \mathbf{0}$, and $\sum_{a \in \mathcal{A}} \omega_{\text{ind}}^1(a) = 0$.

Proof. Let $\theta \equiv (\theta_j : j = 1, \dots, K)$. By properties of projections,

$$E[(1, D, A)'](Y - (\Delta_{\text{long}} D + \theta_0 + \theta' A)) = \mathbf{0}. \quad (\text{A-3})$$

Profiling θ_0 leads to

$$\text{Cov}(D, Y) = \text{Var}(D) \Delta_{\text{long}} + \text{Cov}(A, D)' \theta \quad (\text{A-4})$$

$$\text{Cov}(A, Y) = \text{Cov}(A, D) \Delta_{\text{long}} + \text{Var}(A) \theta. \quad (\text{A-5})$$

Since $\text{Cov}(D, A)$ is positive definite, $\text{Var}(A)$ is positive definite. Then, (A-5) implies that $\theta = \text{Var}(A)^{-1} (\text{Cov}(A, Y) - \text{Cov}(A, D) \Delta_{\text{long}})$. If we plug this into (A-4), we get

$$(\text{Var}(D) - \text{Cov}(D, A) \text{Var}(A)^{-1} \text{Cov}(A, D)) \Delta_{\text{long}} = \text{Cov}(D, Y) - M \text{Cov}(A, Y). \quad (\text{A-6})$$

Since $\text{Cov}(D, A)$ is positive definite, $\text{Cov}(D, A) \text{Var}(A)^{-1} \text{Cov}(A, D) > 0$, and so (A-6) implies that

$$\Delta_{\text{long}} = \frac{\text{Var}(D) \Delta_{\text{short}} - \sum_{j=1}^K M_j \text{Cov}(A_j, Y)}{\text{Var}(D) - \text{Cov}(D, A) \text{Var}(A)^{-1} \text{Cov}(A, D)}, \quad (\text{A-7})$$

where we used that $\text{Var}(D) \Delta_{\text{short}} = \text{Cov}(D, Y)$. For any $j = 1, \dots, K$, some algebra shows that

$$\begin{aligned} \text{Cov}(A_j, Y) &= \sum_{a \in \mathcal{A}} E[Y(1, a) - Y(0, a) | D = 1, A = a] (a_j - E[A_j]) \pi_1(a) E[D] \\ &+ \sum_{a \in \mathcal{A}} (E[Y(0, a) | D = 0, A = a] - E[Y(0, 0) | D = 0, A = 0]) (a_j - E[A_j]) P\{A = a\} \\ &+ \sum_{a \in \mathcal{A}} (E[Y(0, a) | D = 1, A = a] - E[Y(0, a) | D = 0, A = a]) (a_j - E[A_j]) \pi_1(a) E[D]. \end{aligned} \quad (\text{A-8})$$

Then, (A-1) follows from combining (12), (A-7), and (A-8).

To show $\sum_{a \in \mathcal{A}} \omega_{\text{dce}}^1(a) = 1$, consider the following derivation.

$$\begin{aligned} \sum_{a \in \mathcal{A}} \omega_{\text{dce}}^1(a) &\stackrel{(1)}{=} \frac{\text{Var}(D) - P\{D = 1\} \sum_{j=1}^K M_j (E[A_j|D = 1] - E[A_j])}{\text{Var}(D) - \text{Cov}(D, A) \text{Var}(A)^{-1} \text{Cov}(A, D)} \\ &\stackrel{(2)}{=} \frac{\text{Var}(D) - M \text{Cov}(A, D)}{\text{Var}(D) - \text{Cov}(D, A) \text{Var}(A)^{-1} \text{Cov}(A, D)} \stackrel{(3)}{=} 1, \end{aligned}$$

where (1) holds by $\sum_{a \in \mathcal{A}} \pi_1(a) = 1$ and $\sum_{a \in \mathcal{A}} \pi_1(a) a_j = E[A_j|D = 1]$, and (2) holds by $P\{D = 1\}(E[A_j|D = 1] - E[A_j]) = \text{Cov}(A_j, D)$, and (3) holds by definition of M .

We show $\sum_{a \in \mathcal{A}} \omega_{\text{ind}}^1(a) = 0$ by the following derivation applied to its numerator:

$$\text{Var}(D) \sum_{a \in \mathcal{A}} [\pi_1(a) - \pi_0(a)] - \sum_{a \in \mathcal{A}} P\{A = a\} \sum_{j=1}^K M_j (a_j - E[A_j]) = 0,$$

where the equality holds by $\sum_{a \in \mathcal{A}} \pi_1(a) = \sum_{a \in \mathcal{A}} \pi_0(a) = \sum_{a \in \mathcal{A}} P\{A = a\} = 1$ and $\sum_{a \in \mathcal{A}} P\{A = a\} a_j = E[A_j]$.

Finally, we show $\sum_{a \in \mathcal{A}} a_u \omega_{\text{ind}}^1(a) = 0$ for any $u = 1, \dots, K$. Once again, we focus on the following derivation applied to its numerator:

$$\begin{aligned} \sum_{a \in \mathcal{A}} a_u \text{Var}(D) [\pi_1(a) - \pi_0(a)] - \sum_{a \in \mathcal{A}} a_u P\{A = a\} \sum_{j=1}^K M_j (a_j - E[A_j]) \\ \stackrel{(1)}{=} \text{Var}(D) [E[A_u|D = 1] - E[A_u|D = 0]] - M \text{Cov}(A_u, A) \\ \stackrel{(2)}{=} \text{Cov}(D, A_u) - \text{Cov}(D, A) \text{Var}(A)^{-1} \text{Cov}(A, A_u) \stackrel{(3)}{=} 0, \end{aligned}$$

where (1) holds by $\sum_{a \in \mathcal{A}} \pi_d(a) a_j = E[A_j|D = d]$ for $d = 0, 1$, and $\sum_{a \in \mathcal{A}} a_u P\{A = a\} (a_j - E[A_j]) = \text{Cov}(A_u, A_j)$, (2) holds by $\text{Var}(D) [E[A_u|D = 1] - E[A_u|D = 0]] = \text{Cov}(D, A_u)$ and the definition of M , and (3) holds by the fact that $\text{Var}(A)^{-1} \text{Cov}(A, A_u)$ equals a column vector with zeros except for a one in the u th position. ■

Proof of Theorem 4.2. The first part follows from Theorem A.1, which also yields $\sum_{a \in \mathcal{A}} \omega_{\text{dce}}^1(a) = 1$ and $\sum_{a \in \mathcal{A}} \omega_{\text{ind}}^1(a) = 0$. To complete the proof, we now show the equivalence between (a), (b), (c), and (d).

First, we show that (a) implies (b), (c), and (d). To this end, assume (a) holds. Then, the long regression in (7) is equivalent to an SFE regression in (9). To see why, note that (a) implies that A is a K -dimensional vector that is either equal to 0 or a canonical vector (i.e., a vector with a 1 in only one of its coordinates and zeroes otherwise). If we then let $\theta(a) = \theta_0$ for $a = 0$ and $\theta(a) = \theta_j$ for a being the canonical vector with j th coordinate equal to one, we get

$$\theta_0 + \theta' A = \sum_{a \in \mathcal{A}} \theta(a) I\{A = a\}.$$

Therefore, $\Delta_{\text{long}} = \Delta_{\text{sfe}}$ and Theorem A.3 imply (b) (with $\omega_{\text{dce}}^1(a) = \omega_{\text{sfe}}(a)$), (c), and (d).

Second, we show that (b), (c), or (d) implies (a) or, equivalently, the negation of (a) implies the negation of (b), the negation of (c), and the negation of (d).

Start by considering the case when $K = 1$. Then, (a) fails when $\mathcal{A} \neq \{0, 1\}$. For example, consider the case where $\mathcal{A} = \{0, 1, 2\}$ with $\{A|D = 1\} \sim \text{Bi}(2, 0.3)$, $\{A|D = 0\} \sim \text{Bi}(2, 0.9)$, and $P\{D = 1\} = 0.5$, where $\text{Bi}(n, p)$ denotes a Binomial distribution with n trials and probability p . With this distribution of (A, D) , the weights in (A-2) become $\omega_{\text{dce}}^1 \approx [-0.1, 0.76, 0.34]$ and $\omega_{\text{ind}}^1 \approx [-0.14, 0.28, -0.14]$, and so (b) and (c) fail. To verify that (d) fails, consider the case with $\mu(0, a) = 0$ for all $a \in \mathcal{A}$, $\mu(1, 0) = 12$, and $\mu(1, 1) = \mu(1, 2) = 1$. We then have $\mu(1, a) - \mu(0, a) > 0$ for all $a \in \mathcal{A}$, yet $\Delta_{\text{long}} = \Delta_{\text{dce}}^1 = -0.1 < 0$, and so (d) fails.

Next, consider the case when $K = 2$. In this case, (a) can fail when (i) $\mathcal{A}_j \neq \{0, 1\}$ for some $j = 1, 2$ or (ii) $\mathcal{A}_j = \{0, 1\}$ for all $j = 1, 2$ but $A_1 A_2 \neq 0$. For (i), consider $\{A_1|D = 0\} \sim \text{Bi}(2, 0.9)$, $\{A_1|D = 1\} \sim \text{Bi}(2, 0.3)$, $P\{D = 1\} = 0.5$, $A_2 \perp \{D, A_1\}$, and $\text{Var}(A_2) > 0$. The fact that $A_2 \perp \{D, A_1\}$ and $\text{Var}(A_2) > 0$ implies that A_2 drops out of the expressions in (A-2), and the example becomes identical to the one considered when $K = 1$, where (b), (c), and (d) fail. For (ii), let $\text{Ber}(p)$ denote a Bernoulli distribution with parameter p and consider $\{A_j|D = 0\} \sim \text{Ber}(0.1)$ and $\{A_j|D = 1\} \sim \text{Ber}(0.7)$ for $j = 1, 2$, with $P\{D = 1\} = 0.5$, so that $\mathcal{A}_j = \{0, 1\}$ for $j = 1, 2$ and $P\{A_1 A_2 = 0\} \approx 0.45$. With this distribution of (A, D) , the weights in (A-2) become $\omega_{\text{dce}}^1 \approx [0.34, 0.38, 0.48, -0.10]$ and $\omega_{\text{ind}}^1 \approx [-0.14, 0.14, 0.14, -0.14]$, and so (b) and (c) fail. To verify that (d) fails, consider the case with $\mu(0, a) = 0$ for all $a \in \mathcal{A}$, $\mu(1, (0, 0)) = \mu(1, (1, 0)) = \mu(1, (0, 1)) = 1$, and $\mu(1, (1, 1)) = 13$. We then have $\mu(1, a) - \mu(0, a) > 0$ for all $a \in \mathcal{A}$, yet $\Delta_{\text{long}} = \Delta_{\text{dce}}^1 = -0.1 < 0$, and so (d) fails.

Finally, consider the case $K > 2$. Then, (a) can fail when (i) $\mathcal{A}_j \neq \{0, 1\}$ for some $j = 1, \dots, K$ or (ii) $\mathcal{A}_j = \{0, 1\}$ for all $j = 1, \dots, K$ but $A_j A_l \neq 0$ for some $j, l = 1, \dots, K$ with $j \neq l$. In either case, we can repeat the examples used for $K = 2$ by adding coordinates $j = 3, \dots, K$ with $\{A_j : j > 2\} \perp \{D, \{A_j : j \leq 2\}\}$, and $\text{Var}(A_j) > 0$ for $j > 2$. By construction, $\{A_j : j > 2\}$ drops out of the expressions in (A-2), and the examples considered with $K = 2$ imply the failure of (b), (c), and (d). ■

Theorem A.2. Consider the long regression with interactions in (8) and assume that $P\{D = d, A = a\} > 0$ for all $(d, a) \in \mathcal{D} \times \mathcal{A}$. Assume that the variance-covariance matrix of (A, D, AD) , denoted by Σ_{inter} , is positive definite and let $M = \text{Cov}(D, W) \text{Var}(W)^{-1}$ with $W \equiv (A', A'D)'$. Then,

$$\begin{aligned} \Delta_{\text{inter}} &= \sum_{a \in \mathcal{A}} \omega_{\text{dce}}^i(a) E[Y(1, a) - Y(0, a) | D = 1, A = a] \\ &\quad + \sum_{a \in \mathcal{A}} \omega_{\text{ind}}^i(a) (E[Y(0, a) | D = 0, A = a] - E[Y(0, 0) | D = 0, A = 0]) \\ &\quad + \sum_{a \in \mathcal{A}} \omega_{\text{dce}}^i(a) (E[Y(0, a) | D = 1, A = a] - E[Y(0, a) | D = 0, A = a]) , \end{aligned} \quad (\text{A-9})$$

where

$$\begin{aligned} \omega_{\text{dce}}^i(a) &\equiv \frac{\pi_1(a) \left[\sigma_D^2 - p \sum_{j=1}^K M_j (a_j - E[A_j]) - p \sum_{j=1}^K M_{j+K} (a_j - pE[A_j|D = 1]) \right]}{\sigma_D^2 - M \text{Cov}(W, D)} \\ \omega_{\text{ind}}^i(a) &\equiv \frac{\sigma_D^2 (\pi_1(a) - \pi_0(a)) - \sum_{j=1}^K M_j p_a (a_j - E[A_j]) - p \sum_{j=1}^K M_{j+K} (\pi_1(a) a_j - p_a E[A_j|D = 1])}{\sigma_D^2 - M \text{Cov}(W, D)} , \end{aligned} \quad (\text{A-10})$$

$p = P\{D = 1\}$, $p_a = P\{A = a\}$, $\sigma_D^2 = \text{Var}(D)$, and $\pi_d(a)$ is defined in (11). Furthermore, $\sum_{a \in \mathcal{A}} \omega_{\text{dce}}^i(a) = 1$,

$\sum_{a \in \mathcal{A}} a \omega_{\text{ind}}^i(a) = \mathbf{0}$, and $\sum_{a \in \mathcal{A}} \omega_{\text{ind}}^i(a) = 0$.

Proof. Let $\theta = (\theta_j : j = 1, \dots, K)$, $\lambda = (\lambda_j : j = 1, \dots, K)$, and $\alpha = (\theta', \lambda')'$. By properties of projections,

$$E[(1, D, A', DA')'(Y - (\Delta_{\text{inter}}D + \theta_0 + \alpha'W))] = \mathbf{0}. \quad (\text{A-11})$$

Profiling θ_0 leads to,

$$\text{Cov}(D, Y) = \text{Var}(D)\Delta_{\text{inter}} + \text{Cov}(W, D)'\alpha \quad (\text{A-12})$$

$$\text{Cov}(W, Y) = \text{Cov}(W, D)\Delta_{\text{inter}} + \text{Var}(W)\alpha. \quad (\text{A-13})$$

Since Σ_{inter} is positive definite, $\text{Var}(W)$ is positive definite. Then, (A-13) implies that $\alpha = \text{Var}(W)^{-1}(\text{Cov}(W, Y) - \text{Cov}(W, D)\Delta_{\text{inter}})$. If we plug this into (A-12), we get

$$(\text{Var}(D) - M \text{Cov}(W, D))\Delta_{\text{inter}} = \text{Cov}(D, Y) - M \text{Cov}(W, Y). \quad (\text{A-14})$$

Since Σ_{inter} is positive definite, $\text{Var}(D) - \text{Cov}(W, D)'\text{Var}(W)^{-1}\text{Cov}(W, D) > 0$ and so (A-14) implies that

$$\Delta_{\text{inter}} = \frac{\text{Cov}(D, Y) - \sum_{j=1}^K M_j \text{Cov}(A_j, Y) - \sum_{j=1}^K M_{j+K} \text{Cov}(DA_j, Y)}{\text{Var}(D) - M \text{Cov}(W, D)}, \quad (\text{A-15})$$

where we used that $\text{Var}(D)\Delta_{\text{short}} = \text{Cov}(D, Y)$. For any $j = 1, \dots, K$, some algebra shows that

$$\begin{aligned} \text{Cov}(A_j, Y) &= \sum_{a \in \mathcal{A}} E[Y(1, a) - Y(0, a) | D = 1, A = a](a_j - E[A_j])\pi_1(a)p \\ &+ \sum_{a \in \mathcal{A}} (E[Y(0, a) | D = 0, A = a] - E[Y(0, 0) | D = 0, A = 0])(a_j - E[A_j])p_a \\ &+ \sum_{a \in \mathcal{A}} (E[Y(0, a) | D = 1, A = a] - E[Y(0, a) | D = 0, A = a])(a_j - E[A_j])\pi_1(a)p, \end{aligned} \quad (\text{A-16})$$

and

$$\begin{aligned} \text{Cov}(DA_j, Y) &= p \sum_{a \in \mathcal{A}} E[Y(1, a) - Y(0, a) | D = 1, A = a]\pi_1(a)(a_j - pE[A_j | D = 1]) \\ &+ p \sum_{a \in \mathcal{A}} (E[Y(0, a) | D = 1, A = a] - E[Y(0, a) | D = 0, A = a])\pi_1(a)(a_j - pE[A_j | D = 1]) \\ &+ p \sum_{a \in \mathcal{A}} (E[Y(0, a) | D = 0, A = a] - E[Y(0, 0) | D = 0, A = 0])(\pi_1(a)a_j - p_a E[A_j | D = 1]). \end{aligned} \quad (\text{A-17})$$

By plugging in (12), (A-16), and (A-17) into (A-15), (A-9) follows.

To show $\sum_{a \in \mathcal{A}} \omega_{\text{dce}}^i(a) = 1$, consider the following derivation.

$$\begin{aligned} \sum_{a \in \mathcal{A}} \omega_{\text{dce}}^i(a) &\stackrel{(1)}{=} \frac{\sigma_D^2 - p \sum_{j=1}^K M_j (E[A_j | D = 1] - E[A_j]) - \sum_{j=1}^K M_{j+K} \sigma_D^2 E[A_j | D = 1]}{\sigma_D^2 - M \text{Cov}(W, D)} \\ &\stackrel{(2)}{=} \frac{\sigma_D^2 - \sum_{j=1}^K M_j \text{Cov}(D, A_j) - \sum_{j=1}^K M_{j+K} \text{Cov}(D, DA_j)}{\sigma_D^2 - M \text{Cov}(W, D)} \stackrel{(3)}{=} 1, \end{aligned}$$

where (1) holds by $\sum_{a \in \mathcal{A}} \pi_1(a) = 1$, $\sum_{a \in \mathcal{A}} \pi_1(a)a_j = E[A_j|D = 1]$, and $\sigma_D^2 = p(1 - p)$, (2) holds by $p(E[A_j|D = 1] - E[A_j]) = \text{Cov}(D, A_j)$ and $\sigma_D^2 E[A_j|D = 1] = \text{Cov}(DA_j, D)$, and (3) holds by definition of M .

We show $\sum_{a \in \mathcal{A}} \omega_{\text{ind}}^i(a) = 0$ by the following derivation applied to its numerator:

$$\sigma_D^2 \sum_{a \in \mathcal{A}} (\pi_1(a) - \pi_0(a)) - \sum_{j=1}^K M_j \sum_{a \in \mathcal{A}} p_a (a_j - E[A_j]) - p \sum_{j=1}^K M_{j+K} \sum_{a \in \mathcal{A}} (\pi_1(a)a_j - p_a E[A_j|D = 1]) = 0,$$

where the equality holds by $\sum_{a \in \mathcal{A}} \pi_1(a) = \sum_{a \in \mathcal{A}} \pi_0(a) = \sum_{a \in \mathcal{A}} p_a = 1$, $\sum_{a \in \mathcal{A}} p_a a_j = E[A_j]$, and $\sum_{a \in \mathcal{A}} \pi_1(a)a_j = E[A_j|D = 1]$.

Finally, we show $\sum_{a \in \mathcal{A}} a_u \omega_{\text{ind}}^i(a) = 0$ for any $u = 1, \dots, K$. Once again, we focus on the following derivation applied to its numerator:

$$\begin{aligned} & \sum_{a \in \mathcal{A}} a_u \sigma_D^2 (\pi_1(a) - \pi_0(a)) - \sum_{j=1}^K M_j \sum_{a \in \mathcal{A}} a_u p_a (a_j - E[A_j]) - p \sum_{j=1}^K M_{j+K} \sum_{a \in \mathcal{A}} a_u (\pi_1(a)a_j - p_a E[A_j|D = 1]) \\ & \stackrel{(1)}{=} \sigma_D^2 [E[A_u|D = 1] - E[A_u|D = 0]] \\ & - \sum_{j=1}^K M_j \text{Cov}(A_u, A_j) - \sum_{j=1}^K M_{j+K} p (E[A_u A_j|D = 1] - E[A_u]E[A_j|D = 1]) \\ & \stackrel{(2)}{=} \text{Cov}(D, A_u) - \sum_{j=1}^K M_j \text{Cov}(A_j, A_u) - \sum_{j=1}^K M_{j+K} \text{Cov}(DA_j, A_u) \\ & \stackrel{(3)}{=} \text{Cov}(D, A_u) - \text{Cov}(D, W) \text{Var}(W)^{-1} \text{cov}(W, A_u) \stackrel{(4)}{=} 0, \end{aligned}$$

where (1) holds by $\sum_{a \in \mathcal{A}} \pi_d(a)a_j = E[A_j|D = d]$ for $d = 0, 1$, $\sum_{a \in \mathcal{A}} a_u P\{A = a\}(a_j - E[A_j]) = \text{Cov}(A_u, A_j)$, and $\sum_{a \in \mathcal{A}} a_u p_a = E[A_u]$, (2) holds by $\text{Var}(D)[E[A_u|D = 1] - E[A_u|D = 0]] = \text{Cov}(D, A_u)$ and $p(E[A_u A_j|D = 1] - E[A_u]E[A_j|D = 1]) = \text{Cov}(DA_j, A_u)$, (3) holds by the definition of M , and (4) holds by the fact that $\text{Var}(W)^{-1} \text{Cov}(W, A_u)$ equals a column vector with zeros except for a one in the u th position. ■

Proof of Theorem 4.3. The first part follows from Theorem A.2, which also yields that $\sum_{a \in \mathcal{A}} \omega_{\text{dce}}^i(a) = 1$ and $\sum_{a \in \mathcal{A}} \omega_{\text{ind}}^i(a) = 0$. To complete the proof, we now show the equivalence between (a), (b), (c), and (d).

First, we show that (a) implies (b), (c), and (d). To this end, assume (a) holds. Then, the long with interactions regression in (8) is equivalent to an SAT regression in (9). To see why, note that (a) implies that $\mathcal{A} = \{\mathbf{0}_{K \times 1}, \{e_j : j = 1, \dots, K\}\}$, where $e_j \in \mathbb{R}^{K \times 1}$ has a one in the j 'th coordinate and zero otherwise. By defining $A_0 = 1 - \sum_{j=1}^K A_j$, $\gamma(a) = \theta_0$ and $\Delta_{\text{sat}}(a) = \Delta_{\text{inter}}$ for $a = 0$, and $\gamma(a) = \theta_0 + \theta_j$ and $\Delta_{\text{sat}}(a) = \Delta_{\text{inter}} + \lambda_j$ for $a = e_j$ with $j = 1, \dots, K$, we get

$$\Delta_{\text{inter}} D + \theta_0 + \theta' A + \lambda' A D = \sum_{a \in \mathcal{A}} \gamma(a) I\{A = a\} + \sum_{a \in \mathcal{A}} \Delta_{\text{sat}}(a) I\{A = a\} D.$$

Therefore, $\Delta_{\text{inter}} = \Delta_{\text{sat}}(0)$ and Theorem A.4 imply (b) (with $\omega_{\text{dce}}^i(0) = 1$ and $\omega_{\text{dce}}^i(e_j) = 0$ for $j = 1, \dots, K$), (c), and (d).

To conclude, we now show that (b), (c), or (d) implies (a) or, equivalently, the negation of (a) implies the negation of (b), the negation of (c), and the negation of (d).

First, consider the case when $K = 1$. Then, (a) fails when $\mathcal{A} \neq \{0, 1\}$. For example, if $\{A|D = 0\} \sim \text{Bi}(2, 0.3)$, $\{A|D = 1\} \sim \text{Bi}(2, 0.9)$, and $P\{D = 1\} = 0.5$, and so $\mathcal{A} = \{0, 1, 2\}$. By evaluating this information on (A-10), we get $\omega_{\text{dce}}^i \approx [0.19, 1.62, -0.81]$ and $\omega_{\text{ind}}^i \approx [-0.72, 1.44, -0.72]$, i.e., (b) and (c) fail. To verify that (d) fails, consider the case with $\mu(0, a) = 0$ for all $a \in \mathcal{A}$, $\mu(1, 0) = \mu(1, 1) = 1$, and $\mu(1, 2) = 3$. We then have $\mu(1, a) - \mu(0, a) > 0$ for all $a \in \mathcal{A}$, yet $\Delta_{\text{inter}} = \Delta_{\text{dce}}^i = -0.62 < 0$, and so (d) fails.

Second, consider the case when $K = 2$. Then, (a) can fail when (i) $\mathcal{A}_j \neq \{0, 1\}$ for some $j = 1, 2$ or (ii) $\mathcal{A}_j = \{0, 1\}$ for all $j = 1, 2$ but $A_1 A_2 \neq 0$. For (i), consider $\{A_1|D = 0\} \sim \text{Bi}(2, 0.3)$, $\{A_1|D = 1\} \sim \text{Bi}(2, 0.9)$, $P\{D = 1\} = 0.5$, $A_2 \perp \{D, A_1\}$, and $\text{Var}(A_2) > 0$. The fact that $A_2 \perp \{D, A_1\}$ and $\text{Var}(A_2) > 0$ implies that A_2 drops out of the expressions in (A-10), and the example becomes identical to the one considered when $K = 1$ and, thus, (b), (c), and (d) fail. For (ii), consider $\{A_j|D = 0\} \sim \text{Be}(0.3)$ and $\{A_j|D = 1\} \sim \text{Be}(0.9)$ for $j = 1, 2$, and $P(D = 1) = 0.5$, and so $\mathcal{A}_j = \{0, 1\}$ for $j = 1, 2$ and $P(A_1 A_2 = 0) \approx 0.25$. By evaluating this information on (A-10), we get $\omega_{\text{dce}}^i \approx [0.19, 0.81, 0.81, -0.81]$ and $\omega_{\text{ind}}^i \approx [-0.72, 0.72, 0.72, -0.72]$, i.e., (b) and (c) fail. To verify that (d) fails, consider the case with $\mu(0, a) = 0$ for all $a \in \mathcal{A}$, $\mu(1, (0, 0)) = \mu(1, (1, 0)) = \mu(1, (0, 1)) = 1$, and $\mu(1, (1, 1)) = 3$. We then have $\mu(1, a) - \mu(0, a) > 0$ for all $a \in \mathcal{A}$, yet $\Delta_{\text{inter}} = \Delta_{\text{dce}}^i = -0.62 < 0$, and so (d) fails.

Finally, consider $K > 2$. Then, (a) can fail when (i) $\mathcal{A}_j \neq \{0, 1\}$ for some $j = 1, \dots, K$ or (ii) $\mathcal{A}_j = \{0, 1\}$ for all $j = 1, \dots, K$ but $A_j A_l \neq 0$ for some $j, l = 1, \dots, K$ with $j \neq l$. In either case, we can repeat the examples used for $K = 2$ by adding coordinates $j = 3, \dots, K$ with $\{A_j : j > 2\} \perp \{D, \{A_j : j \leq 2\}\}$, and $\text{Var}(A_j) > 0$ for $j > 2$. By construction, $\{A_j : j > 2\}$ drops out of the expressions in (A-10), and the examples considered with $K = 2$ imply the failure of (b), (c), and (d). ■

Theorem A.3. Consider the SFE regression in (9), and assume that $P\{D = d, A = a\} > 0$ for all $(d, a) \in \mathcal{D} \times \mathcal{A}$. Then,

$$\Delta_{\text{sfe}} = \Delta_{\text{dce}}^f + \Delta_{\text{sel}}^f, \quad (\text{A-18})$$

where

$$\omega_{\text{sfe}}(a) \equiv \frac{P\{D = 0|A = a\}P\{D = 1|A = a\}P\{A = a\}}{\sum_{\tilde{a} \in \mathcal{A}} P\{D = 1|A = \tilde{a}\}P\{D = 0|A = \tilde{a}\}P\{A = \tilde{a}\}} \quad \text{for all } a \in \mathcal{A} \quad (\text{A-19})$$

$$\Delta_{\text{dce}}^f \equiv \sum_{a \in \mathcal{A}} \omega_{\text{sfe}}(a) E[Y(1, a) - Y(0, a)|D = 1, A = a] \quad (\text{A-20})$$

$$\Delta_{\text{sel}}^f \equiv \sum_{a \in \mathcal{A}} \omega_{\text{sfe}}(a) (E[Y(0, a)|D = 1, A = a] - E[Y(0, a)|D = 0, A = a]). \quad (\text{A-21})$$

Furthermore, note that $\sum_{a \in \mathcal{A}} \omega_{\text{sfe}}(a) = 1$ and $\omega_{\text{sfe}}(a) \geq 0$.

Proof. By properties of projections,

$$E[YD] = \Delta_{\text{sfe}} E[D] + \sum_{a \in \mathcal{A}} \theta(a) E[I\{A = a\}D] \quad (\text{A-22})$$

$$E[YI\{A = a\}] = \Delta_{\text{sfe}} E[DI\{A = a\}] + \theta(a) P\{A = a\} \quad \text{for all } a \in \mathcal{A}. \quad (\text{A-23})$$

By $P\{A = a\} > 0$ for all $a \in \mathcal{A}$, (A-23) implies that

$$\theta(a) = E[Y|A = a] - \Delta_{\text{sfe}} E[D|A = a] \quad \text{for all } a \in \mathcal{A}. \quad (\text{A-24})$$

Then, (A-22), (A-24), and some algebra imply that

$$\begin{aligned} & E[Y|D = 1] - \sum_{a \in \mathcal{A}} E[Y|A = a]P\{A = a|D = 1\} \\ &= \Delta_{\text{sfe}} \sum_{a \in \mathcal{A}} \frac{P\{D = 1|A = a\}P\{D = 0|A = a\}P\{A = a\}}{P\{D = 1\}}, \end{aligned} \quad (\text{A-25})$$

Under $P\{A = a\} > 0$ and $P\{D = 1|A = a\} \in (0, 1)$ for all $a \in \mathcal{A}$, (A-25) implies that

$$\begin{aligned} \Delta_{\text{sfe}} &= \frac{P\{D = 1\}E[Y|D = 1] - \sum_{a \in \mathcal{A}} E[Y|A = a]P\{A = a, D = 1\}}{\sum_{a \in \mathcal{A}} P\{D = 1|A = a\}P\{D = 0|A = a\}P\{A = a\}} \\ &= \sum_{a \in \mathcal{A}} \omega_{\text{sfe}}(a) (E[Y|A = a, D = 1] - E[Y|A = a, D = 0]). \end{aligned} \quad (\text{A-26})$$

By doing algebra on (A-26), (A-18) follows. Finally, verifying $\sum_{a \in \mathcal{A}} \omega_{\text{sfe}}(a) = 1$ and $\omega_{\text{sfe}}(a) \geq 0$ is straightforward given the definition in (A-19). ■

Proof of Theorem 4.4. This result follows immediately from Theorem A.3. ■

Theorem A.4. Consider the SAT regression in (10) and assume that $P\{D = d, A = a\} > 0$ for all $(d, a) \in \mathcal{D} \times \mathcal{A}$. Then, for all $a \in \mathcal{A}$,

$$\Delta_{\text{sat}}(a) = \Delta_{\text{dce}}^{\text{t}}(a) + \Delta_{\text{sel}}^{\text{t}}(a), \quad (\text{A-27})$$

where

$$\Delta_{\text{dce}}^{\text{t}}(a) \equiv E[Y(1, a) - Y(0, a)|D = 1, A = a] \quad (\text{A-28})$$

$$\Delta_{\text{sel}}^{\text{t}}(a) \equiv E[Y(0, a)|D = 1, A = a] - E[Y(0, a)|D = 0, A = a]. \quad (\text{A-29})$$

Furthermore, under Assumption 3.2, $\Delta_{\text{sel}}^{\text{t}}(a) = 0$ and

$$\Delta_{\text{sat}}(a) = \Delta_{\text{dce}}^{\text{t}}(a) = \mu(1, a) - \mu(0, a). \quad (\text{A-30})$$

Proof. Fix $a \in \mathcal{A}$ arbitrarily throughout this proof. By projection,

$$\begin{aligned} E[YI\{A = a\}] &= \gamma(a)P\{A = a\} + \Delta_{\text{sat}}(a)E[DI\{A = a\}] \\ E[YDI\{A = a\}] &= (\gamma(a) + \Delta_{\text{sat}}(a))E[DI\{A = a\}]. \end{aligned} \quad (\text{A-31})$$

By $P\{A = a\} > 0$, (A-31) implies that

$$\gamma(a) = E[Y|A = a] - \Delta_{\text{sat}}(a)P\{D = 1|A = a\} \quad (\text{A-32})$$

$$E[YD|A = a] = (\gamma(a) + \Delta_{\text{sat}}(a))P\{D = 1|A = a\}. \quad (\text{A-33})$$

By plugging in (A-32) into (A-33), we get

$$E[YD|A = a] - E[Y|A = a]P\{D = 1|A = a\} = \Delta_{\text{sat}}(a)P\{D = 1|A = a\}P\{D = 0|A = a\}. \quad (\text{A-34})$$

By (A-34) and $P\{D = 1|A = a\} \in (0, 1)$, we get that

$$\Delta_{\text{sat}}(a) = E[Y(1, a)|D = 1, A = a] - E[Y(0, a)|D = 0, A = a]. \quad (\text{A-35})$$

The desired result follows from adding and subtracting $E[Y(0, a)|D = 1, A = 1]$ to (A-35). ■

A.2 Auxiliary Lemmas

Lemma A.1. The following statements are true.

- (a) Assumption 4.1 implies Assumption 3.2.
- (b) Assumption 3.2 does not imply Assumption 4.1.
- (c) Assumption 4.1 implies that $Y(\tilde{d}, a) \perp A(d) | X$ for $(\tilde{d}, d, a) \in \mathcal{D} \times \mathcal{D} \times \mathcal{A}$.

Proof. Part (a). For any $(\tilde{d}, \tilde{a}, y, a, d)$, we have

$$\begin{aligned} P\{Y(\tilde{d}, \tilde{a}) \leq y, A(d) = a, D = d|X\} &\stackrel{(1)}{=} P\{Y(\tilde{d}, \tilde{a}) \leq y|X\} P\{A(d) = a|X\} P\{D = d|X\} \\ &\stackrel{(2)}{=} P\{Y(\tilde{d}, \tilde{a}) \leq y|X\} P\{A(d) = a, D = d|X\} \\ &\stackrel{(3)}{=} P\{Y(\tilde{d}, \tilde{a}) \leq y|X\} P\{A = a, D = d|X\}, \end{aligned} \quad (\text{A-36})$$

where (1) holds by Assumption 4.1, (2) holds by Assumption 4.1(i), and (3) holds by $A(D) = A$. Since $(\tilde{d}, \tilde{a}, y, a, d)$ is arbitrary, (A-36) implies Assumption 3.2.

Part (b). Consider the following example. Assume $X \perp (D, (A(d) : d \in \mathcal{D})', (Y(\tilde{d}, a) : (\tilde{d}, a) \in \mathcal{D} \times \mathcal{A})')'$, $Y(d, a) = 0$ for all (d, a) , $(A(1), A(0)) = (D, D)$, and $D \sim \text{Be}(0.5)$. Since $Y(d, a) = 0$, it is independent of $(D, A(D)) = (D, D)$. Thus, Assumption 3.2 holds. By $Y(d, a) = 0$ for all (d, a) and also $A(d) = D$, we have $Y(\tilde{d}, a) \perp A(d)|D$, so Assumption 4.1(ii) holds. However, $(Y(\tilde{d}, a), A(d)) = (0, D) \not\perp D$, and so Assumption 4.1(i) fails.

Part (c). For any $(\tilde{d}, \tilde{a}, y, a, d)$, we have

$$\begin{aligned} P\{Y(\tilde{d}, \tilde{a}) \leq y, A(d) = a|X\} &\stackrel{(1)}{=} P\{Y(\tilde{d}, \tilde{a}) \leq y, A(d) = a|X, D\} \\ &\stackrel{(2)}{=} P\{Y(\tilde{d}, \tilde{a}) \leq y|X, D\} P\{A(d) = a|X, D\} \\ &\stackrel{(3)}{=} P\{Y(\tilde{d}, \tilde{a}) \leq y|X\} P\{A(d) = a|X\}, \end{aligned} \quad (\text{A-37})$$

where (1) and (3) hold by Assumption 4.1(i), and (2) holds by Assumption 4.1(ii). Since $(\tilde{d}, \tilde{a}, y, a, d)$ is arbitrary, (A-37) implies Assumption 3.2. ■

Lemma A.2. Assume the conditions in Theorem 4.2, and that

$$\mu(d, a) = \kappa_0 + \kappa_1 d + \kappa_2' a \quad \text{for all } (d, a) \in \{0, 1\} \times \mathcal{A} \quad (\text{A-38})$$

for some constants $\kappa_0, \kappa_1, \kappa_2$. First, the coefficients in (7) satisfy $\Delta_{\text{long}} = \kappa_1$, $\theta_0 = \kappa_0$, and $\theta_1 = \kappa_2$. Second, the terms in the decomposition in (19) are $\Delta_{\text{dec}}^1 = \kappa_1$ and $\Delta_{\text{ind}}^1 = 0$.

Proof. Assumption 3.2 implies that $E(Y|D = d, A = a) = \mu(d, a)$ which, combined with (A-38), implies that the conditional expectation of Y is linear in $(1, a, d)$. From here, the first result holds because the linear regression consistently estimates the parameters of a linear conditional expectation. The second part follows immediately from combining (A-38) with $\sum_{a \in \mathcal{A}} a \omega_{\text{ind}}^1(a) = \mathbf{0}$ and $\sum_{a \in \mathcal{A}} \omega_{\text{dce}}^1(a) = 1$ (both shown in Theorem A.1). ■

Lemma A.3. The examples used in the proofs of Theorem 4.2 and 4.3 can be completed to satisfy Assumption 4.1.

Proof. For brevity, we focus on the example in the proof of Theorem 4.2 when $K = 1$. A similar argument can be made for all other examples.

Recall that the example in the proof of Theorem 4.2 when $K = 1$ is as follows: $\{A|D = 0\} \sim \text{Bi}(2, 0.3)$, $\{A|D = 1\} \sim \text{Bi}(2, 0.9)$, and $P\{D = 1\} = 0.5$, and so $\mathcal{A} = \{0, 1, 2\}$. The example is silent about X or $\{Y(d, a) : (d, a) \in \mathcal{D} \times \mathcal{A}\}$, and so it is unclear whether Assumption 4.1 holds or not. We now provide one way to complete the specification of the example in a manner compatible with Assumption 4.1.

Assume that $X \perp (\{Y(d, a) : (d, a) \in \mathcal{D} \times \mathcal{A}\}, D, \{A(\tilde{d}) : \tilde{d} \in \mathcal{D}\})$, $\{Y(d, a) : (d, a) \in \mathcal{D} \times \mathcal{A}\}$ non-stochastic and equal to $\{\mu(d, a) : (d, a) \in \mathcal{D} \times \mathcal{A}\}$, $A(0) \sim \text{Bi}(2, 0.3)$, $A(1) \sim \text{Bi}(2, 0.9)$, $D \sim \text{Be}(0.5)$, and $\{A(1), A(0), D\}$ are independent random variables. These conditions imply that $A(0) \stackrel{d}{=} \{A(0)|D = 0\} = \{A|D = 0\} \sim \text{Bi}(2, 0.3)$, $A(1) \stackrel{d}{=} \{A(1)|D = 1\} = \{A|D = 1\} \sim \text{Bi}(2, 0.9)$, and $P\{D = 1\} = 0.5$, as required by the example. Next, we show that the completed example satisfies Assumption 4.1. First, we have that Assumption 4.1(i) holds from the fact that X is independent of the rest of the problem, $\{Y(d, a) : (d, a) \in \mathcal{D} \times \mathcal{A}\}$ is non-stochastic, and $A(d) \perp D$. Second, we have that Assumption 4.1(ii) follows from the fact that X is independent of the rest of the problem and $\{Y(d, a) : (d, a) \in \mathcal{D} \times \mathcal{A}\}$ is non-stochastic. ■

Lemma A.4. Consider the setup in Theorem 4.3 and that A is scalar. Then, the coefficients in (8) satisfy the following decomposition:

$$\Delta_{\text{inter}} + E[A]\lambda = \sum_{a \in \mathcal{A}} \omega_{\text{dce}}^{i*}(a)(E[Y(1, a) - Y(0, a)]) + \omega_{\text{ind}}^{i*}(a)(E[Y(0, a) - Y(0, 0)]) , \quad (\text{A-39})$$

where

$$\begin{aligned} \Delta &= \text{Var}(AD) \text{Var}(A) - (\text{Cov}(DA, A))^2 \\ \Psi &= 1 + \frac{E[A]}{\Delta} (\text{Cov}(A, DA) \text{Cov}(A, D) - \text{Var}(A) \text{Cov}(DA, D)) \\ \omega_{\text{dce}}^{i*}(a) &= \Psi \omega_{\text{dce}}^i(a) + \frac{E[A]}{\Delta} (\text{Var}(A)p\pi_1(a)(a - pE[A|D = 1]) - \text{Cov}(A, DA)(a - E[A])\pi_1(a)p) \\ \omega_{\text{ind}}^{i*}(a) &= \Psi \omega_{\text{ind}}^i(a) + \frac{E[A]}{\Delta} (\text{Var}(A)p(\pi_1(a)a - p_a E[A|D = 1]) - \text{Cov}(A, DA)(a - E[A])p_a) . \end{aligned} \quad (\text{A-40})$$

Moreover, $\sum_{a \in \mathcal{A}} \omega_{\text{dce}}^{i*}(a) = 1$ and $\sum_{a \in \mathcal{A}} \omega_{\text{ind}}^{i*}(a) = 0$. Furthermore, it is possible to have $\omega_{\text{dce}}^{i*}(a) < 0$ and $\omega_{\text{ind}}^{i*}(a) \neq 0$ for some $a \in \mathcal{A}$.

Proof. By properties of projection,

$$(\text{Var}(W))^{-1}(\text{Cov}(W, Y) - \text{Cov}(W, D)\Delta_{\text{inter}}) = \alpha = (\theta', \lambda')' .$$

We can use the fact that A is scalar to obtain an explicit formula for $(\text{Var}(W))^{-1}$. With this expression in hand, we get

$$\Delta_{\text{inter}} + E[A]\lambda = \Psi\Delta_{\text{inter}} + \frac{E[A]}{\Delta}(\text{Var}(A)\text{Cov}(DA, Y) - \text{Cov}(A, DA)\text{Cov}(A, Y)) , \quad (\text{A-41})$$

By plugging in the expressions for (A-9), (A-16), (A-17) on the right-hand side of (A-41), imposing Assumption 3.2, we obtain (A-39) and (A-40).

By the definition of $\{(\omega_{\text{dce}}^{\text{i}*}(a), \omega_{\text{ind}}^{\text{i}*}(a)) : a \in \mathcal{A}\}$ in (A-40) and repeating arguments used in the proof of Theorem A.2, it is immediate to show that $\sum_{a \in \mathcal{A}} \omega_{\text{dce}}^{\text{i}*}(a) = 1$ and $\sum_{a \in \mathcal{A}} \omega_{\text{ind}}^{\text{i}*}(a) = 0$.

To conclude, it suffices to find an example in which $\omega_{\text{dce}}^{\text{i}*}(a) < 0$ and $\omega_{\text{ind}}^{\text{i}*}(a) \neq 0$ for some $a \in \mathcal{A}$. To this end, consider an example with $\{A|D=0\} \sim \text{Bi}(2, 0.9)$, $\{A|D=1\} \sim \text{Bi}(2, 0.1)$, and $P\{D=1\} = 0.3$, and so $\mathcal{A} = \{0, 1, 2\}$. By evaluating this information on (A-40), we get $\omega_{\text{dce}} \approx [-0.2, 1.08, 0.12]$ and $\omega_{\text{ind}} \approx [-0.26, 0.52, -0.26]$, i.e., (b) and (c) fail. ■

Lemma A.5. Assume the conditions in Theorem 4.3, and that

$$\mu(d, a) = \kappa_0 + \kappa_1 d + \kappa_2 a + \kappa_3' ad \quad \text{for all } (d, a) \in \{0, 1\} \times \mathcal{A} \quad (\text{A-42})$$

for some constants $\kappa_0, \kappa_1, \kappa_2, \kappa_3$. Then, the coefficient in (8) satisfies $\Delta_{\text{inter}} = \kappa_1$, $\theta_0 = \kappa_0$, $\theta = \kappa_2$, and $\lambda = \kappa_3$. Furthermore, the decomposition in (27) are $\Delta_{\text{dce}}^{\text{i}} = \kappa_1$ and $\Delta_{\text{ind}}^{\text{i}} = 0$.

Proof. Assumption 3.2 implies that $E(Y|D=d, A=a) = \mu(d, a)$ which, combined with (A-42), implies that the conditional expectation of Y is linear in $(1, a, d, ad)$. From here, the first result follows from the fact that the linear regression consistently estimates the parameters of a linear conditional expectation. The second part follows from combining $\sum_{a \in \mathcal{A}} a\omega_{\text{ind}}^{\text{i}}(a) = \mathbf{0}$ (shown in Theorem A.2) and (A-42). ■

Example A.1. Theorem 4.1 shows that Assumption 3.2 implies that $\Delta_{\text{sel}}^{\text{s}} = 0$. We now provide an example that demonstrates Assumption 3.1 is insufficient to guarantee $\Delta_{\text{sel}}^{\text{s}} \neq 0$.

Consider the case where A is binary and D is independent of $(A(0), A(1), Y(0, 0), Y(1, 0), Y(0, 1), Y(1, 1))$. Assume in addition that $A(0), A(1)$ and D are all independent Bernoulli random variables with $p = 0.5$. Note that $A \perp D$ so that $\pi_1(1) = \pi_0(1)$. Next, specify the potential outcomes as $Y(0, 0) = Y(1, 0) = 0$, $Y(1, 1) = 1$, and $Y(0, 1) = I\{A(1) > A(0)\}$.

This example satisfies Assumption 3.1, as $Y(d) = Y(d, A(d)) \perp D$ for $d = 0, 1$. Using (13)–(15) and $\Delta_{\text{short}} = E[Y|D=1] - E[Y|D=0]$ yield:

$$\Delta_{\text{short}} = 0.5 , \quad \Delta_{\text{dce}}^{\text{s}} = 0.25 , \quad \Delta_{\text{ind}}^{\text{s}} = 0 , \quad \Delta_{\text{sel}}^{\text{s}} = 0.25 .$$

Concretely, while $\Delta_{\text{ind}}^{\text{s}} = 0$ follows directly from $\pi_1(1) = \pi_0(1)$, the example illustrates that $\Delta_{\text{sel}}^{\text{s}} \neq 0$ under Assumption 3.1. Finally, this example does not contradict the second part of Theorem 4.1, since Assumption 3.2 does not hold.