Log-like? Identified ATEs defined with zero-valued outcomes are (arbitrarily) scale-dependent*

Jiafeng Chen
Harvard Business School
Department of Economics, Harvard University

Jonathan Roth
Department of Economics, Brown University

January 2, 2023

Abstract

Researchers frequently estimate the average treatment effect (ATE) in logs, which has the desirable property that its units approximate percentages. When the outcome takes on zero values, researchers often use alternative transformations (e.g., $\log(1+Y)$, $\operatorname{arcsinh}(Y)$) that behave like $\log(Y)$ for large values of Y, and interpret the units as percentages. In this paper, we show that ATEs for transformations other than $\log(Y)$ cannot be interpreted as percentages, at least if one imposes the requirement that a percentage does not depend on the original scaling of the outcome (e.g. dollars versus cents). We first show that if $m(y)$ is a function that behaves like $\log(y)$ for large values of y and the treatment affects the probability that $Y = 0$, then the ATE for $m(Y)$ can be made arbitrarily large or small in magnitude by re-scaling the units of Y. More generally, we show that any parameter of the form $\theta_g = E[g(Y(1), Y(0))]$ that is well-defined with zero-valued outcomes cannot be both scale-invariant and point-identified. Thus, researchers must either consider parameters outside of this class, or give up on either scale-invariance or point-identification. We conclude by outlining a variety of options available to empirical researchers dealing with zero-valued outcomes, including (i) estimating ATEs for normalized outcomes, (ii) explicitly calibrating the value placed on the extensive versus intensive margins, or (iii) estimating separate effects for the intensive and extensive margins.

*We thank Isaiah Andrews, Kirill Borusyak, Edward Glaeser, Peter Hull, Erzo Luttmer, Giovanni Mellace, John Mullahy, David Ritzwoller, Brad Ross, Pedro Sant'Anna, Jesse Shapiro, and seminar participants at Boston University, Georgetown University, Southern Denmark University, Vanderbilt, and the SEA annual conference for helpful comments and suggestions. Bruno Lagomarsino provided superb research assistance.

1 Introduction

When the outcome of interest Y is strictly positive, researchers often estimate the average treatment effect (ATE) in logs, $E[\log(Y(1)) - \log(Y(0))]$, which has the appealing feature that its units approximate percentage changes in the outcome.[1] A practical challenge in many economic settings, however, is that the variable of interest may be equal to zero with positive probability, and thus the ATE in logs is not well-defined. When this is the case, it is common for researchers to estimate treatment effects for alternative transformations of the outcome such as $\log(1+Y)$ or $\text{arcsinh}(Y) = \log\left(\sqrt{1+Y^2}+Y\right)$, which behave similarly to $\log(Y)$ for large values of Y but are well-defined at zero. The treatment effects for these alternative transformations are typically interpreted like the ATE in logs, i.e. as (approximate) percentage effects. For example, we found that among papers papers published in the *American Economic Review* since 2018 that discuss the units of a regression using $\text{arcsinh}(Y)$ as the outcome, 86% (12 of 14) interpret the coefficients as percents or elasticities.[2] One of the authors has also interpreted treatment effects for $\text{arcsinh}(Y)$ in this way in prior work (Chen, Glaeser and Wessel, 2022).

The main point of this paper is that identified average treatment effects that are well-defined with zero-valued outcomes cannot be interpreted as percentages, at least if one imposes the seemingly reasonable requirement that a percentage effect does not depend on the baseline units in which the outcome is measured (e.g. dollars, cents, or yuan).

Our first main result shows that if $m(y)$ is a function that behaves like $\log(y)$ for large values of y, then the ATE for $m(Y)$ will be *arbitrarily sensitive* to the units of Y, in the sense that one can obtain an ATE of *any magnitude* by re-scaling the units. Specifically, we consider continuous, increasing functions $m : [0, \infty) \to \mathbb{R}$ that "look like" $\log(y)$ for large values of y, in the sense that

$$\lim_{y\to\infty} \frac{m(y)}{\log(y)} = 1. \tag{1}$$

The common $\log(1 + y)$ and $\text{arcsinh}(y)$ transformations satisfy this property. We then show that if treatment affects the extensive margin (i.e. $P(Y(1) = 0) \neq P(Y(0) = 0)$) then one can obtain any magnitude for the ATE of $m(Y)$ by rescaling the outcome by some positive factor a. It is therefore inappropriate to interpret the ATE for $m(Y)$ as a percentage, since a percentage is inherently a unit-invariant quantity.

The intuition for this result is that a "proportional" treatment effect is not well-defined for individuals whose outcome moves from zero to non-zero values. Any average treatment effect that is well-defined with zero-valued outcomes must therefore (implicitly) assign a value for a change along the extensive margin relative to proportional changes along the intensive margin. For transforma-

[1] In particular, $\log(Y(1)/Y(0)) \approx \frac{Y(1)-Y(0)}{Y(0)}$ when $Y(1)/Y(0)$ is close to 1.

[2] We found 20 papers overall using $\text{arcsinh}(Y)$ as an outcome variable, of which 14 interpret the units. Among the two out of the 14 that discuss the units but do not explicitly give a percentage interpretation, one exponentiates the coefficients and interprets the units in levels, and the other writes that "we interpret the inverse hyperbolic sine transformation as an approximation of the natural logarithm" but does not explicitly interpret the units as percentages.

tions $m(Y)$ that behave like $\log(Y)$ for large values of Y, the relative importance of the extensive margin is determined implicitly by the units of Y. Intuitively, if the units of the non-zero values of Y are made very large, then a change from a zero to a typical non-zero value of Y has a very large impact on $m(Y)$, and so the ATE places an extremely large weight on the extensive margin. If treatment has an extensive margin effect, then the ATE can thus be made large in magnitude by making the units of the non-zero values of Y very large. By contrast, if the units of Y are very small, then $m(Y) \approx m(0)$ for all values of Y, and thus the ATE will be small. By varying the units of the outcome, we can thus obtain any arbitrary magnitude for the ATE.

What alternative options are available in settings with zero-valued outcomes? Our second main result helps characterize the possibilities. We show that when there are zero-valued outcomes, there is no treatment effect parameter that satisfies all three of the following properties:

(a) The parameter is an average of a function of the potential outcomes, $\theta_g = E_P[g(Y(1), Y(0))]$, where the function g increases for higher values of $Y(1)$.

(b) The parameter is invariant to re-scaling of the units of the outcome (i.e. $g(y_1, y_0) = g(ay_1, ay_0)$).

(c) The parameter is point-identified from the marginal distributions of the potential outcomes.

This "trilemma" implies that any approach that accommodates zero-valued outcomes must necessarily jettison (at least) one of the three requirements above. Which requirement the researcher prefers to forgo will generally depend on their motivation for using a log-like transformation in the first place.

We conclude by highlighting a menu of approaches that may be attractive depending on the researcher's core motivation. We first consider the case where the researcher is interested in obtaining a causal parameter with an intuitive "percentage" interpretation. In this case, it may be natural to consider a normalized parameter outside of the class $E_P[g(Y(1), Y(0))]$. One prominent example is $\theta_{\text{ATE}\%} = \frac{E[Y(1)-Y(0)]}{E[Y(0)]}$, which is the ATE in *levels* expressed as a *percentage* of the control mean. Alternatively, the researcher might target the ATE for a normalized outcome that has a more intuitive interpretation. For example, if Y is employment in a region, and X is its population, then $\tilde{Y} = Y/X$ has the intuitive interpretation as the employment-to-population ratio, which is measured in percentages. We next consider a setting where the researcher's core motivation is to capture concave preferences over the outcome—such as when the researcher considers income gains to be more meaningful for individuals who are initially poor. In this case, it is natural to directly specify how much the researcher values a change along the extensive margin relative to the intensive margin—e.g., specifying that a change from 0 to 1 (in a particular unit) is valued the same as an x percentage point increase along the intensive margin. Finally, the researcher may be interested in separately understanding the effects of the treatment along both the intensive and extensive margins. In this case, the researcher may target separate parameters for the two margins—e.g., $E[\log(Y(1)) - \log(Y(0)) \mid Y(1) > 0, Y(0) > 0]$, the average (log) effect for individuals with positive outcomes under both treatments, captures the intensive margin. Separate effects for the intensive and extensive margins are not point-identified from the marginal distributions of the potential outcomes, but can be bounded using the method in Lee (2009), or point-identified with additional distributional assumptions (Zhang, Rubin and Mealli, 2009).

Related work. Previous work has illustrated in simulations or selected empirical applications that that results for transformations such as $\log(1+Y)$ or $\operatorname{arcsinh}(Y)$ may be sensitive to the units of the outcome (Aihounton and Henningsen, 2021; de Brauw and Herskowitz, 2021). We complement this work by providing theoretical results showing that scale-dependence is a necessary feature of *any* identified ATE that is well-defined with zero-valued outcomes, and that the dependence on units is arbitrarily bad for transformations that behave like $\log(Y)$ for large values of Y.

In concurrent work, Mullahy and Norton (2022) show that the marginal effects from linear regressions using $\log(1+Y)$ or $\operatorname{arcsinh}(Y)$ are sensitive to the scaling of the outcome. Specifically, they show that the marginal effects converge to those of either a levels regression or a (normalized) linear probability model, depending on whether the units are made small or large. Since the marginal effects of linear models correspond with ATEs under unconfoundedness, these results help provide intuition for our result that the ATE for any log-like transformation is *arbitrarily* sensitive to the scaling of the outcome. Our results show that sensitivity to scaling is an issue beyond the class of linear models using the $\log(1+Y)$ or $\operatorname{arcsinh}(Y)$ transformations considered in Mullahy and Norton (2022), however. In particular, we show that any identified ATE must be scale-dependent, and thus our results imply that any consistent estimator of an ATE must be scale-dependent, regardless of whether it is linear or specifically uses the $\operatorname{arcsinh}(Y)$ or $\log(1+Y)$ transformations.

Other previous work has considered the interpretation of regression models using $\operatorname{arcsinh}(Y)$ or $\log(1+Y)$ from the perspective of structural equations models, as opposed to the potential outcomes model considered here. This literature has reached diverging conclusions: Bellemare and Wichman (2020) conclude that coefficients from $\operatorname{arcsinh}(Y)$ regressions have an interpretation as a semi-elasticity, while Cohn, Liu and Wardlaw (2022) conclude that these estimators are inconsistent and advocate for Poisson regression instead. In Appendix D, we show that these diverging conclusions stem from the fact that the structural equations considered in these papers implicitly impose different restrictions on the potential outcomes—some of which are incompatible with zero-valued outcomes—and consider different target causal estimands. This highlights the value of a potential outcomes framework such as ours, which makes transparent what causal estimands are identifiable, and what properties they can have.

1.1 Setup and notation

Let $D \in \{0,1\}$ be a binary treatment and let $Y \in \mathcal{Y}$ be a weakly positively-valued observed outcome, where $\mathcal{Y} \subseteq [0,\infty)$.[3] We assume that $Y = DY(1)+(1-D)Y(0)$, where $Y(1)$ and $Y(0)$ are respectively the potential outcomes under treatment and control. We suppose that in some population of interest, $(Y(1), Y(0)) \sim P$ for some (unknown) joint distribution P. We denote the marginal distribution of

[3] See Appendix C.1 for extensions of our results to settings with continuous treatments.

$Y(d)$ under P by $P_{Y(d)}$ for $d = 0, 1$. We assume that neither $P_{Y(0)}$ nor $P_{Y(1)}$ assigns probability one to zero.

2 Sensitivity to scaling for transformations that behave like $\log(Y)$

We first consider average treatment effects of the form $\theta = E_P[m(Y(1)) - m(Y(0))]$. We are interested in how θ changes if we change the units of Y by a factor of a. That is, how does

$$\theta(a) = E_P[m(aY(1)) - m(aY(0))]$$

depend on a? Setting $a = 100$, for example, might correspond with a change in units between dollars and cents.

We consider the case where $m(y)$ behaves like $\log(y)$ for large values of y, in the sense of (1). This property is satisfied by $\log(1 + y)$ and $\operatorname{arcsinh}(y)$, for example. Our first main result then shows that if the treatment affects the probability that $Y = 0$, then $|\theta(a)|$ can be made to take any desired value through the appropriate choice of a.

Proposition 1. *Suppose that:*

1. *(The function m is continuous and increasing) $m : [0, \infty) \to \mathbb{R}$ is a continuous, weakly increasing function.*
2. *(The function m behaves like* log *for large values) $m(y)/\log(y) \to 1$ as $y \to \infty$.*
3. *(Treatment affects the extensive margin) $P(Y(1) = 0) \neq P(Y(0)) = 0)$.*
4. *(Finite expectations) $E_P[|\log(Y(d))| \mid Y(d) > 0] < \infty$ for $d = 0, 1$.*

Then, for every $\theta^* \in (0, \infty)$, there exists an $a > 0$ such that $|\theta(a)| = \theta^*$. In particular, $\theta(a)$ is continuous with $\theta(a) \to 0$ as $a \to 0$ and $|\theta(a)| \to \infty$ as $a \to \infty$.

Proposition 1 casts serious doubt on the interpretation of ATEs for functions like $\log(1 + Y)$ or $\operatorname{arcsinh}(Y)$ as (approximate) percentage effects. While a percent (or log point) is entirely invariant to scaling, Proposition 1 shows that, in sharp contrast, the ATEs for these transformations are arbitrarily dependent on units.

Remark 1 (ATEs for $\log(c + Y)$)**.** In some settings, researchers consider the ATE for $m_c(Y) = \log(c+Y)$ and consider sensitivity to the parameter c. Observe that $\log(1+aY) = \log(a(1/a+Y)) = \log(a)+\log(1/a+Y)$, and thus the ATE for $\log(1+aY)$ is equal to the ATE for $\log(1/a+Y)$. Hence, varying the constant term for $\log(c + Y)$ is isomorphic to varying the scaling of the outcome for $m(y) = \log(1 + y)$. Proposition 1 thus implies that if treatment affects the probability that $Y = 0$, one can obtain any desired magnitude for the ATE using $\log(c + Y)$ via the choice of c.

2.1 Intuition for Proposition 1

The result in Proposition 1 intuitively derives from the fact that a "percentage" treatment effect is not well-defined for individuals who have $Y(0) = 0$ but $Y(1) > 0$, or vice versa. Any ATE that is well-defined with zero-valued outcomes must implicitly determine how much weight to place on changes along the extensive margin relative to proportional changes along the intensive margin.

When $m(Y)$ behaves like $\log(Y)$ for large values of Y, the importance of the extensive margin is implicitly determined by the units of Y. For intuition, suppose that we re-scale the outcomes so that the non-zero values of Y are very large. Then for an individual for whom treatment changes the outcome from zero to non-zero, the treatment effect will be very large, since $m(Y(1)) \gg m(Y(0)) = m(0)$. Extensive margin treatment effects thus have a large impact on the ATE when the units of Y are made large. By contrast, changing the units of Y does not change the importance of treatment effects along the intensive margin by much, since for $Y(1), Y(0) > 0$, we have that $m(Y(1)) - m(Y(0)) \approx \log(Y(1)/Y(0))$, which does not depend on the units of the outcome.

To see the roles of the extensive and intensive margins more formally, for simplicity consider the case where $P(Y(1) = 0, Y(1) > 0) = 0$, so that, for example, everyone who has positive income without receiving a training also has positive income when receiving the training.[4] Then, by the law of iterated expectations, we can write

$$
\begin{aligned}
E[m(aY(1)) - m(aY(0))] = {} & P(Y(1) > 0, Y(0) > 0) \underbrace{E[m(aY(1)) - m(aY(0)) \mid Y(1) > 0, Y(0) > 0]}_{\text{Intensive margin}} \\
& + P(Y(1) > 0, Y(0) = 0) \underbrace{E[m(aY(1)) - m(0) \mid Y(1) > 0, Y(0) = 0]}_{\text{Extensive margin}} .
\end{aligned}
$$

When a is large, $m(ay) \approx \log(ay)$ for non-zero values of y, and thus the intensive margin effect in the previous display is approximately equal to $E[\log(Y(1)) - \log(Y(0)) \mid Y(1) > 0, Y(0) > 0]$, the treatment effect in logs for individuals with positive outcomes under both treatment and control. This, of course, does not depend on the scaling of the outcome. However, the extensive margin effect grows with a, since $m(aY(1)) \approx \log(a) + \log(Y(1))$ is increasing in a while $m(0)$ does not change. Thus, as a grows large, the ATE for $m(aY)$ places more and more weight on the extensive margin effect of the treatment relative to the intensive margin. We can therefore make $|\theta(a)|$ arbitrarily large by sending $a \to \infty$. By contrast, if $a \approx 0$, then $m(aY(d)) \approx 0$ with very high probability, and thus the ATE for $m(aY)$ is approximately equal to 0.

2.2 Quantifying the sensitivity to scaling

Proposition 1 shows that any magnitude of $|\theta(a)|$ can be achieved via the appropriate choice of a. How much does $\theta(a)$ change for finite changes in the scaling a? The following Proposition gives an (approximate) answer to this question.

[4] A related argument goes through without this restriction, but now there are two extensive margins, one for people with $Y(1) > 0 = Y(0)$, and the other for those with $Y(0) > Y(1) = 0$.

Proposition 2. *Under the conditions of Proposition 1,*[5] *as* $a \to \infty$,

$$E_P[m(a \cdot Y(1)) - m(a \cdot Y(0))] = (P(Y(1) > 0) - P(Y(0) > 0)) \cdot \log(a) + o(\log(a)).$$

Proposition 2 shows that multiplying the units of Y by the factor a (for a large a) increases the ATE by about $\log(a)$ times the treatment effect on the probability of having a non-zero outcome. Thus, for example, if treatment increases the probability of a non-zero outcome by 10 percentage points, then changing the units from dollars to cents ($a = 100$) would be expected to change the estimated treatment effect by $0.1 \times \log(100) = 0.46$, or what is typically (although incorrectly) interpreted as a change in the treatment effect of around 46 percentage points. When the extensive margin is not very close to zero, changing the scaling by a factor of just 100 can thus generate a substantial impact on the ATE.

In empirical work, researchers sometimes check how the estimated ATE for $\log(c + Y)$ changes with c for a finite set of values c, which as noted in Remark 1 is equivalent to rescaling the outcome using $m(Y) = \log(1 + Y)$. Proposition 2 implies that checking whether the results for $\log(c + Y)$ are relatively insensitive to c is essentially equivalent to checking whether the extensive margin effect is small.

2.3 Additional remarks and extensions

Remark 2 (Extension to continuous treatments)**.** We focus on ATEs for binary treatments for expositional simplicity, although similar results apply with continuous treatments. In Appendix C.1, we show that when d is a continuous treatment, any treatment effect contrast that averages across possible values of d (i.e. a parameter of the form $\int \omega(d) E[m(aY(d)])$ is sensitive to scaling when there is an extensive margin effect.

Remark 3 (Extension to OLS estimands)**.** It is worth noting that the results in this section show that population ATEs for $m(Y)$ are sensitive to the units of Y. These results are about *estimands*, and thus any consistent *estimator* of the ATE for $m(Y)$ will be sensitive to scaling (at least asymptotically). Thus, our results apply to ordinary least squares (OLS) estimators when they have a causal interpretation, but also to non-linear estimators such as inverse-probability weighting or doubly-robust methods. Nevertheless, given the prominence of OLS in applied work, and the fact that OLS is sometimes used for non-causal estimands, in Appendix C.2 we provide a result specifically on the scale-sensitivity of the population regression coefficient for a random variable of the form $m(Y)$ on an arbitrary random variable X. Our results shows that the coefficients on X will be arbitrarily sensitive to scaling when the coefficients of a regression of $\mathbb{1}[Y > 0]$ on X are non-zero. Thus, the OLS estimand using functions that "look like" log on the left-hand side will be sensitive to scaling even when they do not have a causal interpretation.

Remark 4 (When most values are large)**.** Researchers often have the intuition that if most of the values of the outcome are "large", then ATEs for transformations like $\log(1 + Y)$ or $\text{arcsinh}(Y)$ will

[5] Continuity of m is not needed for this result.

approximate elasticities, since $m(Y) \approx \log(Y)$ for most values of Y. Indeed, in an influential paper, Bellemare and Wichman (2020) recommend that researchers using the $\text{arcsinh}(Y)$ transformation should transform the units of their outcome so that most of the non-zero values of Y are large. The results in this section suggest—perhaps somewhat counterintuitively—that if one rescales the outcome such that the non-zero values are all large, the behavior of the average treatment effect will be driven nearly entirely by the effect of the treatment on zero-valued outcomes and *not* on the distribution of outcomes conditional on these being positive. Moreover, the rescaling can be chosen to generate any magnitude for the ATE if the treatment affects the probability of a zero-valued outcome.

3 Sensitivity to scaling for other ATEs

Our results so far show that ATEs for transformations that are defined at zero and "look like" $\log(y)$ are arbitrarily sensitive to scaling. What other options are available when there are zero-valued outcomes? To help delineate alternative options, in this section we provide a result showing what desirable properties a parameter defined with zero-valued outcomes can have. Specifically, we establish a "trilemma": when there are zero-valued outcomes, there is no parameter that (a) takes the form $\theta_g = E_P[g(Y(1), Y(0))]$, (b) is scale-invariant, and (c) is point-identified. Any approach for settings with zero-valued potential outcomes must therefore abandon one of the properties (a)-(c); in Section 4 below we discuss several approaches that relax one (or more) of these requirements.

Before stating our formal result, we must first make precise what we mean by scale-invariance and point-identification. We say that g is scale-invariant if its value is the same under any re-scaling of the units of y by a positive constant a.

Definition 1. We say that the function g is *scale-invariant* if it is homogeneous of degree zero, i.e. $g(y_1, y_0) = g(ay_1, ay_0)$ for all $a, y_1, y_0 > 0$.

To define point-identification, recall that we assume that, in some population of interest, $(Y(1), Y(0)) \sim P$ for some (unknown) joint distribution P, and we denote the marginal distribution of $Y(d)$ under P by $P_{Y(d)}$ for $d = 0, 1$. We consider the setting where the marginal distributions of the potential outcomes $P_{Y(1)}, P_{Y(0)}$ are identified from the data, but not the joint distribution P, as in e.g. Fan, Guerre and Zhu (2017). In a randomized controlled trial, for example, the marginal distributions of the potential outcomes in the full population are identified from the distributions of $Y \mid D = d$, but the joint distribution is not identified since we never observe both $Y(1)$ and $Y(0)$ for any particular individual.[6] We will thus say θ_g is point-identified if it depends only on the marginal

[6] The marginal distributions of $P_{Y(d)}$ for the full population are likewise identified under conditional unconfoundedness and overlap assumptions in observable studies. In other common settings, the marginal distributions of the potential outcomes are identified for a subset of the population, e.g. for compliers in instrumental variables settings (Abadie, 2003) and for treated individuals in generalized DiD settings (Athey and Imbens, 2006; Roth and Sant'Anna, Forthcoming). Note that under the standard parallel trends assumption, only the marginal *means* of the potential outcomes distributions are identified, an even weaker notion of identification.

distributions of the potential outcomes, and not on the joint distribution, which cannot be learned directly from the data even in an experiment.

Definition 2 (Identification)**.** We say that θ_g is *point-identified from the marginals at P* if for every joint distribution Q with the same marginals as P (i.e. such that $Q_{Y(d)} = P_{Y(d)}$ for $d = 0, 1$), $E_P[g(Y(1), Y(0))] = E_Q[g(Y(1), Y(0))]$. For a class of distributions $\mathcal{P}$, we say that θ_g is *point identified over $\mathcal{P}$* if for every $P \in \mathcal{P}$, θ_g is point identified from the marginals distributions at P.

We will denote by $\mathcal{P}_+$ the set of distributions on $[0, \infty)^2$. Thus, θ_g is identified over $\mathcal{P}_+$ if it is always identified when Y takes on zero or weakly positive values. Our next result formalizes that when Y can take on zero values, it is not possible to have a parameter of the form $E_P[g(Y(1), Y(0))]$ that is both scale-invariant and point-identified over $\mathcal{P}_+$.

Proposition 3 (A trilemma)**.** *The following three properties cannot hold simultaneously:*

(a) $\theta_g = E_P[g(Y(1), Y(0))]$ *for a non-constant function* $g : [0, \infty)^2 \to \mathbb{R}$ *that is weakly increasing in its first argument.*

(b) The function g is scale-invariant.

(c) θ_g *is point-identified over* $\mathcal{P}_+$*.*

To establish the proof of Proposition 3, we prove that the only parameter satisfying properties (a) and (b) that is point-identified over distributions for which Y is *strictly* positively-valued is the ATE in logs (up to an affine transformation). Since $\log(0)$ is not well-defined, it follows that there are no parameters satisfying the three properties when one allows for zero-valued outcomes. Any parameter that is well-defined when there are zero-valued outcomes must therefore abandon at least one of properties (a)–(c).

4 Empirical approaches with zero-valued outcomes

Our theoretical results above imply that when there are zero-valued outcomes, the researcher cannot merely take a log-like transform of the outcome and interpret the units of the ATE as a percentage, since, unlike a percentage, the resulting ATE is dependent on the units. In this section, we highlight a variety of alternative estimands that are well-defined and easily interpreted when there are zero-valued outcomes. Of course, any approach must necessarily drop one of the requirements in the trilemma in Proposition 3, but the choice of which requirement to drop may depend on the researcher's motivation.

To inform our discussion of alternative estimands, it is useful to first consider several reasons why empirical researchers may contemplate log-transforming their data in the first place, rather than just reporting the ATE in levels:

(i) The researcher is interested in reporting an ATE with easily-interpretable units, such as "percents."

(ii) The researcher believes that there are decreasing returns to the outcome, and thus wants to place more weight on treatment effects for individuals with low initial outcomes. For instance, the researcher may perceive it to be more meaningful to raise income from $Y(0) = \$10{,}000$ to $Y(1) = \$20{,}000$ than from $Y(0) = \$100{,}000$ to $Y(1) = \$110{,}000$, yet both of these treatment effects contribute equally to the ATE in levels.

(iii) The researcher is interested in both the intensive and extensive margin effects of the treatment, and is using the ATE for a log-like transformation as an approximation to the proportional effect along the intensive margin.

We note that all of these motivations will typically be more salient in settings where the outcome Y spans several orders of magnitude (i.e. is right-skewed). The ATE in levels may be difficult to interpret when Y spans several orders of magnitude—e.g., a change in employment of 10,000 is more meaningful in a small town than in New York City—thus motivating attention to a parameter with more interpretable units. Likewise, diminishing marginal utility from Y will be more important when there is a large variance in Y across units. And understanding the intensive margin will typically be more relevant when there is substantial variation in the outcome among individuals with non-zero values.

These three motivations also suggest different ways of breaking out of the trilemma in Proposition 3. If the goal is to achieve a percentage interpretation, then one can either consider scale-invariant parameters outside of the class $E_P[g(Y(1), Y(0))]$, or consider the ATE in levels for a normalized parameter $\tilde{Y}$ that already has a percentage interpretation. Alternatively, if the goal is to capture concave social preferences over the income, then it is natural to specify how much we value the intensive margin relative to the extensive margin—thus abandoning scale-invariance. Finally, if the goal is to separately understand the intensive margin effect, the researcher can abandon partial identification (from the marginal distributions) and directly target the partially-identified parameter $E[\log Y(1) - \log Y(0) \mid Y(0) > 0, Y(0) > 0]$, the effect in logs for individuals with positive outcomes under both treatments. We address each of these cases in turn, with a summary of our discussion provided in Table 1.

4.1 When the goal is interpretable units

We first consider the case where the researcher's primary goal is to obtain a treatment effect parameter with easily-interpretable units, such as percentages.

Normalizing the ATE in levels. One possibility is to target the parameter

$$\theta_{\text{ATE\%}} = \frac{E[Y(1) - Y(0)]}{E[Y(0)]},$$

which is the ATE *in levels* expressed as a *percentage of the control mean.* A researcher studying a job training program, for example, might report the ATE in dollars, as a percentage of the average

Description	Estimand	Main property sacrificed?	Pros/Cons
Normalized ATE	$E[Y(1) - Y(0)]/E[Y(0)]$	$E[g(Y(1), Y(0))]$	*Pro:* Percent interpretation *Con:* Does not capture decreasing returns
Normalized outcome	$E[Y(1)/X - Y(0)/X]$	$E[g(Y(1), Y(0))]$	*Pro:* Per-unit-X interpretation *Con:* Need to find sensible X
Explicit tradeoff of intensive/extensive margins	ATE for $m(y) = \begin{cases} \log(y) & y > 0 \\ -x & y = 0 \end{cases}$	Scale-invariance	*Pro:* Explicit tradeoff of two margins *Con:* Need to choose x; Monotone only if support outside $(0, e^{-x})$
Intensive margin effect	$E\left[\log \frac{Y(1)}{Y(0)} \mid Y(1) > 0, Y(0) > 0\right]$	Point-identification	*Pro:* ATE in logs for the intensive margin *Con:* Partial identification

Table 1: Summary of alternative estimands

earnings for the control group. This parameter is point-identified and scale-invariant, and thus has an intuitive percentage interpretation. It is not an average of individual-level effects, however—i.e., it does not take the form $E_P[g(Y(1), Y(0))]$—and thus avoids the trilemma in Proposition 3. The parameter $\theta_{\text{ATE}\%}$ is also what Poisson regression (see Chapter 18.2 in Wooldridge, 2010) estimates under the potential outcomes model and an appropriate identifying assumption (e.g. unconfoundedness).[7]

We emphasize that $\theta_{\text{ATE}\%}$ is not an average of individual-level percentage effects, but rather the percentage change in the mean outcome between the treatment and control distributions.[8] Since $\theta_{\text{ATE}\%}$ is a re-scaling of the ATE in levels, it thus may be dominated by individuals with large values of the outcome when Y has a long right-tail. Whether this is desirable or not will depend on the context: if the researcher is studying an intervention designed to reduce healthcare spending, then it may be reasonable to study average spending per person, even if this is driven mainly by a small fraction of individuals with catostrophic medical spending. In this case, $\theta_{\text{ATE}\%}$ corresponds with the percentage cost savings from the program. On the other hand, a researcher studying an anti-poverty program meant to increase the earnings for poor individuals may not want the estimand to be dominated by a small number of individuals with very large earnings. If Y is highly skewed, this also may lead to large standard errors for the sample analog to $\theta_{\text{ATE}\%}$. We thus suspect that $\theta_{\text{ATE}\%}$ will typically be most attractive in settings where the researcher wants a parameter with a percentage interpretation, but the support of Y does not span several orders of magnitude. We next turn to alternative approaches that may be more appropriate in settings where Y is highly-skewed, such that interpreting the ATE in levels (expressed as a percentage) may be difficult.

[7]With a randomly assigned D, for example, Poisson regression using the pseudo-likelihood $Y \mid D \sim \text{Pois}(e^{\alpha+\beta D})$ estimates the population coefficient $e^{\beta} = E[Y(1)]/E[Y(0)] = \theta_{\text{ATE}\%} + 1$.

[8]This is loosely parallel to how the τ^{th} quantile treatment effect gives the change in the τ^{th} quantile of the potential outcomes distributions, but is not the τ^{th} percentile of the treatment effects without strong assumptions.

Normalizing the outcome. A second, related approach to obtaining a treatment effect with more intuitive units is to estimate the ATE for a transformed outcome of the form $\tilde{Y} = Y/X$, where Y is the original outcome and X is some pre-determined characteristic. For example, consider a setting where Y is employment in a particular area. The treatment effect in levels for Y may be difficult to interpret, since a change in employment of 1,000 means something very different in New York City versus a small rural town. However, if X is the area's population, then $\tilde{Y}$ is the employment-to-population ratio, which may be more comparable across places, and is already in percentage (i.e. per capita) units. If Y and X are typically of the same order of magnitude, then this normalization will also help to reduce the skew in $\tilde{Y}$. We note that the ATE for $\tilde{Y}$ can be viewed as a scale-invariant, point-identified parameter of the form $\theta = E_P[g(Y(1), Y(0), X)]$, and thus escapes the trilemma in Proposition 3 by avoiding property (a).[9] The viability of this approach, of course, depends on having a variable X such that the normalized outcome $\tilde{Y}$ is of economic interest. We suspect that in many contexts, reasonable options will be available, including pre-treatment observations of the outcome (assuming these are positive), or the *predicted* control outcome given some observable characteristics ($X = E[Y(0) \mid W]$, for observable characteristics W).

4.2 When the goal is to capture decreasing returns

We next consider the case where the motivation for using a log-like transformation of the outcome is to capture some form of decreasing marginal utility over the outcome. For example, when Y is strictly positively valued, the ATE in logs corresponds with the change in utility from implementing the treatment for a utilitarian social planner with log utility over the outcome, $U = E[\log(Y)]$. Intuitively, this social welfare function captures the fact that the planner values a percentage point change in the outcome equally for all individuals, regardless of their initial level of the outcome. We note that the derivative of $\log(Y)$ goes to 0 as $Y \to \infty$, and thus log-utility will typically not be dominated by the right-tail of Y.

Of course, log utility makes sense only when there is only an intensive margin, and is not well-defined when there is an extensive margin. Thus, a planner with log utility over the intensive margin must specify how much they value the intensive versus extensive margins. Recall from Section 2.1 that when using transformations like $\log(1 + y)$ or $\text{arcsinh}(y)$, the scaling of the outcome implicitly determines the weights placed on the intensive versus extensive margins.

Instead of implicitly weighting the margins via the scaling of Y, one can directly specify how much they value each margin of treatment. If one has log utility over changes on the intensive margin, then a natural approach is to specify how much one values a change in earnings from 0 to 1 relative to a percentage change in earnings for non-zero earnings. If, for example, one values the extensive margin effect of moving from 0 to 1 the same as an x percent increase in earnings, then one might consider setting $m(y) = \log(y)$ for $y > 0$ and $m(0) = -x$. The ATE for this transformation can be interpreted as an approximate percentage (log point) effect, where an increase from 0 to 1

[9] It is scale-invariant in the sense that $g(y_1, y_0, x) = g(ay_1, ay_0, ax)$.

is valued at x log points.[10]

We emphasize that for a fixed value of x, this approach necessarily depends on the scaling of the outcome (thus avoiding the trilemma in Proposition 3). However, this may not be so concerning since the appropriate choice of x also depends on the units of the outcome—e.g. saying a change from 0 to 1 is worth x percent means something very different if 1 corresponds with one dollar versus a million dollars. In other words, ATEs for transformations such as $\text{arcsinh}(Y)$ may be difficult to interpret because the scaling of the outcome implicitly determines the relative importance of the intensive and extensive margins; this approach avoids that difficulty by *explicitly* taking a stand on the tradeoff between these two margins. Nevertheless, a challenge with this approach is that researchers may have differing opinions over the appropriate choice of x.

4.3 When the goal is to understand intensive and extensive margins

Finally, we consider the case where the researcher is interested in understanding the intensive and extensive margin effects, separately. A common question in the literature on job training programs (Card, Kluve and Weber, 2010), for instance, is whether a training raises participants' earnings by helping them find a job—which would be expected only to have an extensive-margin effect—or by increasing human capital, which would be expected to also affect the intesive margin. In settings like this, it is natural to target separate parameters for the intensive and extensive margins. For example, the parameter

$$\theta_{\text{Intensive}} = E[\log(Y(1)) - \log(Y(0)) \mid Y(1) > 0, Y(0) > 0]$$

captures the ATE in logs for those who would have a positive outcome regardless of their treatment status. The parameter $\theta_{\text{Intensive}}$ is scale-invariant but is not point-identified from the marginal distributions of the potential outcomes (thus avoiding the trilemma in Proposition 3), and therefore cannot be consistently estimated without further assumptions.[11] However, Lee (2009) popularized a method for obtaining bounds on $\theta_{\text{Intensive}}$ under the monotonicity assumption that, for example, everyone with positive earnings without receiving a training would also have positive earnings when receiving the training. See, also, Zhang and Rubin (2003) for related results, including bounds without the monotonicity assumption. Estimates of $\theta_{\text{Intensive}}$ can be reported alongside measures of the extensive margin effect, such as the change in the probability of having a non-zero outcome, $P(Y(1) > 0) - P(Y(0) > 0)$, and the average (log) outcome for individuals for whom treatment affects the extensive margin, $E[\log(Y(1)) \mid Y(1) > 0, Y(0) = 0]$, which can be bounded similarly

[10] Note that this transformation will generally only be sensible if the support of Y excludes $(0, e^{-x})$, since otherwise the function $m(y)$ is not monotone in y. It is common, however, to have a lower-bound on non-zero values; e.g., a firm cannot have between 0 and 1 employees.

[11] $\theta_{\text{Intensive}}$ also does not take the form $E_P[g(Y(1), Y(0))]$, although it can be written as

$$\frac{E_P\left[\mathbb{1}[Y(1) > 0, Y(0) > 0] \log(Y(1)/Y(0))\right]}{E_P[\mathbb{1}[Y(1) > 0, Y(0) > 0]]},$$

where both the numerator and denominator take this form.

to $\theta_{\text{Intensive}}$.[12] One can potentially tighten the bounds (or restore point identification) by imposing additional assumptions on the joint distribution of the potential outcomes—e.g. by assuming that the outcome distribution under treatment stochastically dominates that under control (Zhang et al., 2009).

We note that the parameter $\theta_{\text{Intensive}}$ is generally distinct from the "intensive margin" marginal effects implied by two-part models (2PMs), which were recommended for scenarios with zero-valued outcomes by Mullahy and Norton (2022). In Appendix E, we consider the causal interpretation of the marginal effects of 2PMs, building on the discussion in Angrist (2001). Our decomposition shows that the marginal effects from 2PMs yield the sum of a causal parameter similar to $\theta_{\text{Intensive}}$ as well as a "selection term" comparing potential outcomes for individuals for whom treatment only has an intensive margin effect to those with an extensive margin effect. It thus may be difficult to ascribe a causal interpretation to the marginal effects of 2PMs without assumptions about this selection.

5 Conclusion

It is common in empirical work to estimate ATEs for transformations such as $\log(1+Y)$ or $\text{arcsinh}(Y)$ which are well-defined at zero and behave like $\log(Y)$ for large values of Y. We show that the ATEs for such transformations cannot be interpreted as percentages, since they depend arbitrarily on the units of the outcome. Further, we show that any parameter of the form $\theta_g = E_P[g(Y(1), Y(0))]$ must be scale-dependent if it is identified and well-defined at zero. We discuss several alternative approaches, including estimating ATEs for normalized outcomes, explicitly calibrating the value placed on the intensive versus extensive margins, and separately estimating effects for the intensive and extensive margins.

References

Abadie, Alberto, "Semiparametric instrumental variable estimation of treatment response models," *Journal of Econometrics*, April 2003, *113* (2), 231–263.

Aczél, J., *Lectures on Functional Equations and Their Applications*, Academic Press, January 1966. Google-Books-ID: n7vckU_1tY4C.

Aihounton, Ghislain B D and Arne Henningsen, "Units of measurement and the inverse hyperbolic sine transformation," *The Econometrics Journal*, June 2021, *24* (2), 334–351.

[12]Under the monotonicity assumption imposed by Lee (2009),

$$E[\log(Y(1)) \mid Y(1) > 0] = (1-\theta)E[\log(Y(1)) \mid Y(1) > 0, Y(0) > 0] + \theta E[\log(Y(1)) \mid Y(1) > 0, Y(0) = 0],$$

where $\theta = P(Y(0) > 0)/P(Y(1) > 0)$. Since $E[\log(Y(1)) \mid Y(1) > 0]$ and θ are identified (from the marginal distributions of the potential outcomes), bounds on $E[\log(Y(1)) \mid Y(1) > 0, Y(0) > 0]$ as derived in Lee (2009) immediately imply bounds on $E[\log(Y(1)) \mid Y(1) > 0, Y(0) = 0]$.

Angrist, Joshua D, "Estimation of Limited Dependent Variable Models With Dummy Endogenous Regressors," *Journal of Business & Economic Statistics*, January 2001, *19* (1), 2–28. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/07350010152472571.

Athey, Susan and Guido Imbens, "Identification and Inference in Nonlinear Difference-in-Differences Models," *Econometrica*, 2006, *74* (2), 431–497.

Bellemare, Marc F. and Casey J. Wichman, "Elasticities and the Inverse Hyperbolic Sine Transformation," *Oxford Bulletin of Economics and Statistics*, 2020, *82* (1), 50–61. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/obes.12325.

Bellégo, Christophe, David Benatia, and Louis Pape, "Dealing with Logs and Zeros in Regression Models," March 2022. arXiv:2203.11820 [econ, stat].

Belotti, Federico, Partha Deb, Willard G. Manning, and Edward C. Norton, "Twopm: Two-Part Models," *The Stata Journal*, April 2015, *15* (1), 3–20. Publisher: SAGE Publications.

Card, David, Jochen Kluve, and Andrea Weber, "Active Labour Market Policy Evaluations: A Meta-Analysis," *The Economic Journal*, November 2010, *120* (548), F452–F477.

Chen, Jiafeng, Edward Glaeser, and David Wessel, "JUE Insight: The (non-)effect of opportunity zones on housing prices," *Journal of Urban Economics*, 2022, p. 103451.

Cohn, Jonathan B., Zack Liu, and Malcolm I. Wardlaw, "Count (and count-like) data in finance," *Journal of Financial Economics*, November 2022, *146* (2), 529–551.

de Brauw, Alan and Sylvan Herskowitz, "Income Variability, Evolving Diets, and Elasticity Estimation of Demand for Processed Foods in Nigeria," *American Journal of Agricultural Economics*, 2021, *103* (4), 1294–1313. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajae.12139.

Fan, Yanqin, Emmanuel Guerre, and Dongming Zhu, "Partial identification of functionals of the joint distribution of "potential outcomes"," *Journal of Econometrics*, March 2017, *197* (1), 42–59.

Lee, David S., "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, July 2009, *76* (3), 1071–1102.

Mullahy, John and Edward C. Norton, "Why Transform Y? A Critical Assessment of Dependent-Variable Transformations in Regression Models for Skewed and Sometimes-Zero Outcomes," December 2022.

Roth, Jonathan and Pedro Sant'Anna, "When Is Parallel Trends Sensitive to Functional Form," *Econometrica*, Forthcoming.

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

Zhang, Junni L. and Donald B. Rubin, "Estimation of Causal Effects via Principal Stratification When Some Outcomes are Truncated by "Death"," *Journal of Educational and Behavioral Statistics*, December 2003, *28* (4), 353–368. Publisher: American Educational Research Association.

_ , _ , and Fabrizia Mealli, "Likelihood-Based Analysis of Causal Effects of Job-Training Programs Using Principal Stratification," *Journal of the American Statistical Association*, March 2009, *104* (485), 166–176.

A Proofs for Section 2 (Sensitivity to scaling for transformations that behave like $\log(Y)$)

Proposition 1. *Suppose that:*

1. *(The function m is continuous and increasing) $m : [0, \infty) \to \mathbb{R}$ is a continuous, weakly increasing function.*

2. *(The function m behaves like $\log$ for large values) $m(y)/\log(y) \to 1$ as $y \to \infty$.*

3. *(Treatment affects the extensive margin) $P(Y(1) = 0) \neq P(Y(0)) = 0)$.*

4. *(Finite expectations) $E_P[|\log(Y(d))| \mid Y(d) > 0] < \infty$ for $d = 0, 1$.*

Then, for every $\theta^* \in (0, \infty)$, there exists an $a > 0$ such that $|\theta(a)| = \theta^*$. In particular, $\theta(a)$ is continuous with $\theta(a) \to 0$ as $a \to 0$ and $|\theta(a)| \to \infty$ as $a \to \infty$.

Proof. Note that $\theta(0) = E_P[m(0)] - E_P[m(0)] = 0$. Additionally, Proposition 2 implies that $|\theta(a)| \to \infty$ as $a \to \infty$. To establish the proof, it thus suffices to show that $\theta(a)$ is continuous on $[0, \infty)$. The desired result is then immediate from the intermediate value theorem.

To establish continuity, fix some $a \in [0, \infty)$ and consider a sequence $a_n \to a$. Without loss of generality, assume $a_n < a + 1$ for all n. Let $m_{a_n}(y) = m(a_n y)$. Since m is continuous, $m_{a_n}(y) \to m_a(y)$ pointwise. Since $m(y)/\log(y) \to 1$ as $y \to \infty$, there exists $\overline{y}$ such that $m(y) < 2\log(y)$ for all $y \geqslant \overline{y}$. From the monotonicity of m, it follows that

$$m(0) \leqslant m(y) \leqslant \mathbb{1}[y \leqslant \overline{y}]m(\overline{y}) + \mathbb{1}[y > \overline{y}]2\log(y)$$

and hence

$$\begin{aligned} m(0) \leqslant m_{a_n}(y) &\leqslant \mathbb{1}[a_n y \leqslant \overline{y}]m(\overline{y}) + \mathbb{1}[a_n y > \overline{y}]2\log(a_n y) \\ &\leqslant |m(\overline{y})| + 2 \cdot \mathbb{1}[y > 0] \cdot (|\log(a+1)| + |\log(y)|) =: \overline{m}(y). \end{aligned}$$

for all n. Hence, we have that $|m_{a_n}(y)| \leqslant |m(0)| + |\overline{m}(y)|$ for all n, and the bounding function is integrable for $Y(d)$ for $d = 0, 1$ by the fourth assumption of the proposition. It follows from the dominated convergence theorem that $E_P[m_{a_n}(Y(d))] \to E_P[m_a(Y(d))]$ for $d = 0, 1$, and thus $\theta(a_n) \to \theta(a)$, as we wished to show. $\square$

Proposition 2. *Under the conditions of Proposition 1,*[13] *as $a \to \infty$,*

$$E_P[m(a \cdot Y(1)) - m(a \cdot Y(0))] = (P(Y(1) > 0) - P(Y(0) > 0)) \cdot \log(a) + o(\log(a)).$$

[13] Continuity of m is not needed for this result.

Proof. Fix a sequence $a_n \to \infty$, and without loss of generality, assume $a_n > e$. We will show that

$$\frac{1}{\log a_n} E_P[m(a_n Y(1)) - m(a_n Y(0))] \to P(Y(1) = 0) - P(Y(0) = 0). \tag{2}$$

Define $f_n(y) = m(a_n y)/\log(a_n)$. Note that $f_n(y) \to \mathbb{1}[y > 0]$ pointwise, since $f_n(0) = m(0)/\log(a_n) \to 0$, while for $y > 0$,

$$f_n(y) = \frac{m(a_n y)}{\log(a_n)} = \frac{m(a_n y)}{\log(a_n y)} \frac{\log(a_n) + \log(y)}{\log(a_n)} \to 1,$$

where we use the fact that $m(y)/\log(y) \to 1$ as $y \to \infty$ by assumption. We showed in the proof to Proposition 1 that

$$|m(y)| \leqslant \kappa + 2 \cdot \mathbb{1}[y > 0] \cdot |\log(y)|$$

where κ is a constant not depending on y. It follows that

$$|f_n(y)| = \frac{|m(a_n y)|}{\log(a_n)} \leqslant \kappa + 2 \cdot \mathbb{1}[y > 0] \cdot (1 + |\log(y)|).$$

Further, since $E_P[|\log(Y(d))| \mid Y(d) > 0]$ is finite by assumption, the upper bound is integrable for $y = Y(d)$ for $d = 0, 1$. It follows from the dominated convergence theorem that

$$E_P[f_n(Y(d))] = E_P\left[\frac{m(a_n Y(d))}{\log(a_n)}\right] \to E_P[\mathbb{1}[Y(d) > 0]] = P(Y(d) > 0).$$

Equation (2) follows immediately from the continuous mapping theorem, which completes the proof. □

B Proofs for Section 3 (Sensitivity to scaling for other ATEs)

Proposition 3 (A trilemma)**.** *The following three properties cannot hold simultaneously:*

(a) $\theta_g = E_P[g(Y(1), Y(0))]$ *for a non-constant function* $g : [0, \infty)^2 \to \mathbb{R}$ *that is weakly increasing in its first argument.*

(b) The function g *is scale-invariant.*

(c) θ_g *is point-identified over* $\mathcal{P}_+$*.*

Proof. To establish the proof of Proposition 3, we first prove the following result, which shows that the only scale-invariant parameter of the form $E_P[g(Y(1), Y(0))]$ that is identified over distributions on the positive reals is the ATE in logs (up to an affine transformation).

Proposition 4. *Let* $\mathcal{P}_{++}$ *denote the set of distributions over compact subsets of* $(0, \infty)^2$*. Suppose* $g : (0, \infty)^2 \to \mathbb{R}$ *is weakly increasing in* y_1 *and scale-invariant. Then* θ_g *is point-identified over* $\mathcal{P}_{++}$ *if and only if* $g = c \cdot (\log(y_1) - \log(y_0)) + d$*, for constants* $c \geqslant 0$ *and* $d \in \mathbb{R}$*.*

Proof. We first show that point-identification over $\mathcal{P}_{++}$ implies that $g(\cdot,\cdot)$ must be additively separable. We do so by considering the points $\{y_0, y_0+b\} \times \{y_1, y_1+a\}$ on a rectangular grid. If $g(\cdot,\cdot)$ is not additively separable, then its expectation with respect to distributions supported on the rectangular grid depends on the correlation.

Formally, suppose that there there exist positive values $y_1, y_0, a, b > 0$ such that

$$g(y_1, y_0) + g(y_1 + a, y_0 + b) \neq g(y_1 + a, y_0) + g(y_1, y_0 + b).$$

Now, consider the marginal distributions $P_{Y(d)}$ such that $P(Y(1) = y_1) = \frac{1}{2} = P(Y(1) = y_1 + a)$ and $P(Y(0) = y_0) = \frac{1}{2} = P(Y(0) = y_0 + b)$. Let P_1 and P_2 denote the joint distributions corresponding with these marginals and perfect positive and negative correlation of the potential outcomes, respectively. Then we have that

$$\begin{aligned} E_{P_1}(g(Y(1), Y(0))) &= \frac{1}{2}\left(g(y_1, y_0) + g(y_1 + a, y_0 + b)\right) \\ &\neq \frac{1}{2}\left(g(y_1 + a, y_0) + g(y_1, y_0 + b)\right) \\ &= E_{P_2}(g(Y(1), Y(0))), \end{aligned}$$

and thus θ_g is not point-identified from the marginals of P_1. Hence, if θ_g is identified over $\mathcal{P}_{++}$, then it must be that

$$g(y_1, y_0) + g(y_1 + a, y_0 + b) = g(y_1 + a, y_0) + g(y_1, y_0 + b) \text{ for all } y_1, y_0, a, b,$$

and hence

$$g(y_1 + a, y_0) - g(y_1, y_0) = g(y_1 + a, y_0 + b) - g(y_1, y_0 + b) \text{ for all } y_1, y_0, a, b.$$

It follows that we can write $g(y_1, y_0) = r(y_1) + q(\frac{1}{y_0})$, where $r(y_1) = g(y_1, 1) - g(1, 1)$ and $q(\frac{1}{y_0}) = g(1, y_0)$.

Second, we show that homogeneity of degree zero, combined with monotonicity, implies that g must be a difference in logarithms. Observe that

$$g(y_1, y_0) = g\left(\frac{y_1}{y_0}, \frac{y_0}{y_0}\right) = g\left(\frac{y_1}{y_0}, 1\right) =: h\left(\frac{y_1}{y_0}\right),$$

where h is an increasing function. Since g is scale-invariant, we thus, have that for any $a, b > 0$,

$$\begin{aligned} g(1,1) &= h(1) = r(1) + q(1) \\ g(a,1) &= h(a) = r(a) + q(1) \\ g\left(1, \frac{1}{b}\right) &= h(b) = r(1) + q(b) \end{aligned}$$

$$g\left(a, \frac{1}{b}\right) = h(ab) = r(a) + q(b)$$

and hence $h(ab) = h(a)+h(b)-h(1)$. It follows that $\tilde{h}(x) = h(x)-h(1)$ is an increasing function such that $\tilde{h}(ab) = \tilde{h}(a) + \tilde{h}(b)$ for all $a, b \in \mathbb{R}$, i.e. an increasing function satisfying Cauchy's logarithmic function equation: $\phi(ab) = \phi(a) + \phi(b)$ for all positive reals a, b. Recall that if a function is increasing, then it has countably many discontinuity points, and thus is continuous somewhere. It is a well-known result in functional equations that the only solutions to Cauchy's logarithmic equation are of the form $\phi(t) = c\log(t)$, if we require that these solutions are continuous at some point; see Aczél (1966), Theorem 2 in Section 2.1.2.[14] Since we require monotonicity, the constant $c \geqslant 0$. Thus, $g(y_1, y_0) = h(y_1/y_0) = \tilde{h}(y_1/y_0) + \tilde{h}(1) = c\log(y_1) - c\log(y_0) + \tilde{h}(1)$. Letting $d = \tilde{h}(1)$ completes the proof of Proposition 4. □

Note that if $g : [0, \infty)^2 \to \mathbb{R}$ is increasing in y_1, then it cannot be equal to $c\log(y_1/y_0) + d$ for $c > 0$ everywhere on $(0, \infty)^2$, since this would imply that $\lim_{y_1 \to 0} g(y_1, 1) = -\infty < g(0, 1)$. The proof of Proposition 3 is then immediate from Proposition 4, which shows that that if properties (a) and (b) are satisfied, and θ_g is point-identified over $\mathcal{P}_{++} \subset \mathcal{P}_+$, then $g = c\log(y_1/y_0) + d$ on $(0, \infty)^2$. □

C Extensions

C.1 Extension to continuous treatments

Although we focus on binary treatment in the main text for simplicity, similar issues arise with continuously distributed D. Suppose now that D can take a continuum of values on some set $\mathcal{D} \subseteq \mathbb{R}$, and let $Y(d)$ denote the potential outcome at the dose d. Consider the parameter

$$\theta(a) = \int_{\mathcal{D}} \omega(d) E[m(aY(d))],$$

which is a weighted sum of the average values of $m(aY(d))$ across different values of d with weights $\omega(d)$. For example, in an RCT with a continuous treatment, a regression of $m(aY)$ on D yields a parameter of the form $\theta(a)$ where, by the Frisch-Waugh-Lovell theorem, the weights are proportional to $(d - E[D])p(d)$ and integrate to 0.[15]

We now show that $\theta(a)$ can be made to have arbitrary magnitude via the choice of a when there is an extensive margin effect. In particular, by an extensive margin effect we mean that $\int \omega(d)P(Y(d) > 0) \neq 0$, i.e. when there is an average effect on the probability of a zero outcome, using the same weights $\omega(d)$ that are used for $\theta(a)$. When $\theta(a)$ is the regression of $m(aY)$ on D in an RCT, for example, $\int \omega(d)P(Y(d) > 0) \neq 0$ if the regression of $\mathbb{1}[Y > 0]$ on D yields a non-zero coefficient.

[14] Correspondingly, non-trivial solutions to Cauchy's logarithmic equations are highly ill-behaved.

[15] Here, $p(d)$ denotes the density of D at d over the randomization distribution.

Proposition 5. *Suppose that:*

1. *The function m satisfies parts 1 and 2 of Proposition 1.*
2. *(Extensive margin effect) $\int_{\mathcal{D}} \omega(d) P(Y(d) > 0) \neq 0$.*
3. *(Bounded expectations) For all d, $E[|\log(Y(d))| \mid Y(d) > 0] < \infty$*
4. *(Regularity for weights) The weights $\omega(d)$ satisfy $\int_{\mathcal{D}} \omega(d) = 0$, $\int_{\mathcal{D}} |\omega(d)| < \infty$ and $\int_{\mathcal{D}} |\omega(d)| \cdot E[|\log(Y(d))| \mid Y(d) > 0] < \infty$.*

Then for every $\theta^* \in (0, \infty)$, there exists $a > 0$ such $\theta(a) = \theta^*$. In particular, $\theta(a)$ is continuous and $\theta(a) \to 0$ as $a \to 0$ and $|\theta(a)| \to \infty$ as $a \to \infty$.

Proof. Note that $\theta(0) = \int \omega(d) m(0) = 0$. It thus suffices to show that $\theta(a)$ is continuous for $a \in [0, \infty)$ and that $|\theta(a)| \to \infty$ as $a \to \infty$. The result then follows from the intermediate value theorem.

We first show continuity. Fix $a \in [0, \infty)$ and a sequence $a_n \to a$. Let $f_n(d) = \omega(d) E[m(a_n Y(d))]$. We showed in the proof to Proposition 1 that $E[m(a_n Y(d))] \to E[m(aY(d))]$, and thus $f_n(d) \to \omega(d) E[m(aY(d))]$ pointwise. We also showed in the proof to Proposition 1 that for a_n sufficiently close to a,

$$|m(a_n Y)| \leqslant \kappa + 2 \cdot \mathbb{1}[y > 0] \cdot |\log(y)|,$$

for a constant κ not depending on n. It follows that

$$|f_n(d)| \leqslant |\omega(d)| \cdot |\kappa| + 2|\omega(d)| \cdot E[|\log(Y(d))| \mid Y(d) > 0],$$

and the upper bound is integrable by part 4 of the Proposition. Hence, by the dominated convergence theorem, we have that $\theta(a_n) = \int_{\mathcal{D}} f_n(d) \to \int_{\mathcal{D}} \omega(d) E[m(aY(d)] = \theta(a)$, as needed.

To show that $|\theta(a)| \to \infty$ as $a \to \infty$, we will show that

$$\frac{\theta(a)}{\log(a)} \to \int_{\mathcal{D}} \omega(d) P[Y(d) > 0]$$

as $a \to \infty$. Consider $a_n \to \infty$, and suppose without loss of generality that $a_n > e$. Observe that

$$\frac{\theta(a_n)}{\log(a_n)} = \int_{\mathcal{D}} \omega(d) \frac{E[m(a_n Y(d))]}{\log(a_n)}.$$

We showed in the proof to Proposition 2 that for each d,

$$\frac{E[m(a_n Y(d))]}{\log(a_n)} \to P(Y(d) > 0).$$

Letting $f_n(d) = \omega(d) \dfrac{E[m(a_n Y(d))]}{\log(a_n)}$, we thus have that $f_n(d) \to \omega(d) P(Y(d) > 0)$ pointwise.

Moreover, we showed in the proof to Proposition 1 that

$$|m(y)| \leqslant \kappa + 2 \cdot \mathbb{1}[y > 0] \cdot |\log(y)|$$

where κ is a constant not depending on y. It follows that

$$\frac{|m(a_n y)|}{\log(a_n)} \leqslant \kappa + 2 \cdot \mathbb{1}[y > 0] \cdot (1 + |\log(y)|)$$

and thus that

$$|f_n(d)| \leqslant |\omega(d)| \cdot (\kappa + 2 + 2E[|\log(Y(d)| \mid Y(d) > 0])$$

where the upper bound is integrable by the fourth part of the proposition. The result then follows from dominated convergence.

□

C.2 Extension to OLS estimands

As noted in Remark 3, our results imply that any consistent estimator of the ATE for an outcome of the form $m(aY)$ will be (asymptotically) sensitive to scaling when there is an extensive margin effect. Our results thus cover the OLS estimand when it is consistent (e.g. in an RCT or under unconfoundedness). Given the prominence of OLS in applied work—and the fact that it is sometimes used for non-causal analyses—we now provide a direct result on the sensitivity to scaling of the OLS estimator of an outcome of the form $m(aY)$ on an arbitrary random variable X.

Specifically, suppose that $(X, Y) \sim Q$, for $Y \in [0, \infty)$ and $X \in \mathbb{R}^J$, where the first element of X is a constant. Consider the OLS estimand

$$\beta(a) = E_Q[XX']^{-1} E_Q[Xm(aY)].$$

We assume that $E_Q[XX']$ is full-rank so that $\beta(a)$ is well-defined. Letting $\beta_j(a) = e_j' \beta(a)$ be the j^{th} element of $\beta(a)$, we will show that $\beta_j(a)$ can be made to have arbitrary magnitude via the choice of a if $\gamma_j \neq 0$, where

$$\gamma = E_Q[XX']^{-1} E_Q[X \mathbb{1}[Y > 0]]$$

is the coefficient from a regression of $\mathbb{1}[Y > 0]$ on X.

Proposition 6. *Suppose that*

1. *The function m satisfies parts 1 and 2 of Proposition 1.*
2. *(Finite expectations) Suppose that $E_Q[\|X\|] < \infty$ and $E_Q[\|X \log(Y)\| \mid Y > 0] < \infty$.*
3. *For some $j \in \{2, ..., J\}$, $\gamma_j \neq 0$.*

Then for every $\beta_j \in (0, \infty)$, there exists $a > 0$ such that $|\beta_j(a)| = \beta_j$. In particular $\beta_j(a)$ is continuous with $\beta_j(a) \to 0$ as $a \to 0$ and $|\beta_j(a)| \to \infty$ as $a \to \infty$. Moreover, $\beta_j(a)/\log(a) \to \gamma_j$ as

$a \to \infty$.

Proof. Note that $\beta(0) = E_Q[XX']^{-1}E[Xm(0)]$, is the coefficient from a regression of a constant outcome $m(0)$ on X, and thus $\beta_1(0) = m(0)$ while $\beta_k(0) = 0$ for $k \geqslant 2$. Thus $\beta_j(0) = 0$. To complete the proof, we will show that $|\beta_j(a)| \to \infty$ as $a \to \infty$ and that $\beta_j(a)$ is continuous for $a \in [0, \infty)$. The result then follows from the intermediate value theorem.

For ease of notation, let $\nu' = e_j' E_Q[XX']^{-1}$, so that $\beta_j(a) = E_Q[\nu' X m(aY)]$.

We first show that $\beta_j(a) \to \infty$ as a diverges. Consider a sequence $a_n \to \infty$, and assume without loss of generality that $a_n > e$. Let $f_n(x, y) = \nu' x \cdot m(a_n y)/\log(a_n)$. Observe that $f_n(x, y) \to \nu' x \cdot \mathbb{1}[y > 0]$ pointwise, since $f_n(x, 0) = \nu' x \cdot m(0)/\log(a_n) \to 0$, while for $y > 0$,

$$f_n(x, y) = \nu' x \cdot \frac{m(a_n y)}{\log(a_n)} = \nu' x \cdot \frac{m(a_n y)}{\log(a_n y)} \frac{\log(a_n) + \log(y)}{\log(a_n)} \to \nu' x,$$

where we use the fact that $m(y)/\log(y) \to 1$ as $y \to \infty$. We showed in the proof to Proposition 2 that

$$\frac{|m(a_n y)|}{\log(a_n)} \leqslant \kappa + 2 \cdot \mathbb{1}[y > 0] \cdot (1 + |\log(y)|),$$

which implies that

$$|f_n(x, y)| \leqslant |\nu' x \cdot (\kappa + 2 \cdot \mathbb{1}[y > 0] \cdot (1 + |\log(y)|))| =: \overline{f}(x, y).$$

Moreover, part 2 of the proposition implies that $\overline{f}(X, Y)$ is integrable. From the dominated convergence theorem, it follows that

$$\frac{\beta_j(a_n)}{\log(a_n)} = E_Q[f_n(X, Y)] \to E_Q[\nu' X \mathbb{1}[Y > 0]] = \gamma_j \neq 0.$$

Hence, we see that $\beta_j(a_n) = \gamma_j \log(a_n) + o(\log(a_n))$, and thus $|\beta_j(a_n)| \to \infty$.

To complete the proof, we show continuity of $\beta_j(a)$. Fix $a \in [0, \infty)$, and consider a sequence $a_n \to a$. Assume without loss of generality that $a_n < a + 1$ for all n. Let $f_n(x, y) = \nu' x \cdot m(a_n y)$. From the continuity of m, we have that $f_n(x, y) \to \nu' x \cdot m(ay)$ pointwise. We showed in the proof to Proposition 1 that there exists some κ (not depending on n) such that

$$|m(a_n y)| \leqslant \kappa + 2\mathbb{1}[y > 0]|\log(y)|.$$

Hence,

$$|f_n(x, y)| \leqslant |\nu' x \cdot (\kappa + 2\mathbb{1}[y > 0]|\log(y)|)|.$$

Moreover, the bounding function is integrable over the distribution of (X, Y) by part 2 of the proposition. Applying the dominated convergence theorem again, we obtain that

$$\beta_j(a_n) = E_Q[f_n(X, Y)] \to E_Q[\nu' X \cdot m(aY)] = \beta_j(a),$$

as needed. $\square$

C.3 Extension to smooth g

Proposition 4 showed that if θ_g is point-identified, scale-invariant, and not the ATE in logs (or an affine transformation thereof), then there exists *some* distribution $P \in \mathcal{P}_{++}$ such that θ_g is not point-identified. The following result shows that if, in addition, g is smooth, then θ_g is not point-identified for *all* distributions P satisfying some basic regularity conditions.

Proposition 7. *Suppose that $P_{Y(d)}$ is absolutely continuous with respect to Lebesgue measure on $(0, \infty)$ with positive density everywhere, and that $E_P[\log(Y(d))]$ is well-defined for $d = 0, 1$. Suppose that $g : (0, \infty)^2 \to \mathbb{R}$ is increasing in y_1, scale-invariant, and twice continuously differentiable. Then θ_g is point identified from the marginals if and only if $g(y_1, y_0) = c(\log(y_1) - \log(y_0)) + d$ for $c \geqslant 0$.*

Proof. We first show that if g is point-identified from the marginals, then the cross derivative

$$\frac{\partial^2 g(y_1, y_0)}{\partial y_1 y_0} = 0$$

for all $y_1, y_0 > 0$. Towards contradiction, suppose there there exists (y_1^*, y_0^*) such that $\frac{\partial^2 g(y_1^*, y_0^*)}{\partial y_1 \partial y_0} \neq 0$. Without loss of generality, assume that

$$\frac{\partial^2 g(y_1^*, y_0^*)}{\partial y_1 \partial y_0} > 0.$$

By continuity, there exists a neighborhood of (y_1^*, y_0^*) on which $\frac{\partial^2 g(y_1, y_0)}{\partial y_1 \partial y_0} > 0$. Since $P_{Y(d)}$ is absolutely continuous with respect to Lebesgue measure, it follows that we can choose $\epsilon_1, \epsilon_0 > 0$ such that $P(Y(d) \in [y_d^*, y_d^* + \epsilon_d)) = p > 0$ and $\frac{\partial^2 g(y_1, y_0)}{\partial y_1 \partial y_0} > 0$ on the rectangle $[y_1^*, y_1^* + \epsilon_1) \times [y_1^*, y_1^* + \epsilon_1)$. For notational convenience, denote the interval $[y_d^*, y_d^* + \epsilon_d]$ by $\mathcal{I}_d$. Consider the joint distribution P_1 such that the marginals of P_1 match $P_{Y(d)}$ and $P_1(Y(1) \in \mathcal{I}_1, Y(0) \in I_2) = p$ (which implies that $P_1(Y(1) \in \mathcal{I}_1, Y(0) \notin \mathcal{I}_0) = 0 = P_1(Y(1) \notin \mathcal{I}_1, Y(0) \in \mathcal{I}_0)$), and such that $Y(1), Y(0)$ have perfect rank correlation conditional on the event that $(Y(1), Y(0)) \in \mathcal{I}_1 \times \mathcal{I}_0$ and conditional on the event that $(Y(1), Y(0)) \notin \mathcal{I}_1 \times \mathcal{I}_0$. Define P_2 analogously except assume that P_2 has perfect negative rank correlation conditional on $(Y(1), Y(0)) \in \mathcal{I}_1 \times \mathcal{I}_0$. For convenience, let $A = 1$ denote the event that $(Y(1), Y(0)) \in \mathcal{I}_1 \times \mathcal{I}_0$. By construction, we then have that

$$E_{P_1}[g(Y(1), Y(0))] - E_{P_2}[g(Y(1), Y(0))] = p \left(E_{P_1}[g(Y(1), Y(0)) | A = 1] - E_{P_2}[g(Y(1), Y(0)) | A = 1] \right).$$

However, g has strictly positive second derivative conditional on $A = 1$, and thus is strictly supermodular conditional on $A = 1$. Note that $E_{P_1}[g(Y(1), Y(0)) | A = 1]$ and $E_{P_2}[g(Y(1), Y(0)) | A = 1]$ correspond with the Frechet-Hoeffding upper and lower bounds for $E[g(Y(1), Y(0)) | A = 1]$ conditional on the marginals for $Y(d) | A = 1$. Since the distribution of $Y(d) | Y(d) \in \mathcal{I}_d$ is continuous by construction, and $g(Y(1), Y(0))$ is bounded conditional on $A = 1$,[16] Theorem 3.1 in Fan et al. (2017) implies the upper and lower bounds do not coincide, and thus $E_{P_1}[g(Y(1), Y(0))] - E_{P_2}[g(Y(1), Y(0))] \neq 0$, which is a contradiction.

We have thus shown that $\frac{\partial^2}{\partial y_1 \partial y_0} g(y_1, y_0) = 0$ for all $y_1, y_0 > 0$, and hence we can write $g(y_1, y_0) = r(y_1) + q(y_0)$. The remainder of the proof is then the same as for Proposition 4. $\square$

D Connection to structural equations models

Previous work has a considered a variety of estimators for settings with zero-valued outcomes beginning instead with a structural equations model. These papers have reached different results, with some concluding that regressions with $\text{arcsinh}(Y)$ have the interpreation of an elasticity, and others showing that they are inconsistent and advocating for other methods (e.g. Poisson regression) instead. In this section, we interpret the results in those papers from the perspective of the potential outcomes model, and show that these diverging conclusions stem from different implicit assumptions about the potential outcomes, as well as a focus on different causal estimands.

Before discussing specific papers, we first note that, broadly speaking, structural equation models can be viewed as constraining the joint distribution of potential outcomes. Observe that, for any pair of potential outcomes $(Y(1), Y(0))$, we can represent them as $(Y(1, U), Y(0, U))$ for some function $Y(d, u)$ and individual-level unobservable (or "structural error") U. The potential outcomes framework we work with in this paper does not impose any functional form assumptions on $Y(d, u)$. Structural equation models, on the other hand, tend to specify explicit functional forms for $Y(d, u)$. In what follows, we consider the implicit restrictions placed on the potential outcomes as well as the target estimand in work related work that starts with a structural equations model.

D.1 Bellemare and Wichman (2020)

Bellemare and Wichman (2020) consider OLS regressions of the form[17]

$$\text{arcsinh}(Y) = \beta_0 + D\beta_1 + U. \tag{3}$$

Note that when D is binary and randomly assigned, $D \perp (Y(1), Y(0))$, then from the perspective of the potential outcomes model, the population coefficient β_1 is that ATE for $\text{arcsinh}(Y)$. Bellemare and Wichman (2020) instead consider the interpretation of β_1 when (3) is treated as structural. From the perspective of the potential outcomes model, this amounts to imposing that the potential outcomes $Y(d) := Y(d, U)$ take the form

$$\text{arcsinh}(Y(d, U)) = \beta_0 + d\beta_1 + U, \tag{4}$$

[16] Specifically, g is increasing in y_1 and HOD 0, which implies that $g(y_1, y_0) = h(y_1/y_0)$ for an increasing function h. Hence, conditional on $A = 1$, $g(y_1^*, y_0^* + \epsilon_0) \leqslant g(Y(1), Y(0)) \leqslant g(y_1^* + \epsilon_1, y_0^*)$.

[17] They also consider specifications with additional covariates on the right-hand side, although we abstract away from this for expsositional simplicity.

where the individual-level random variable U takes the same value for all values of d. Under (4), we have that

$$\beta_1 = \operatorname{arcsinh}(Y(1,U)) - \operatorname{arcsinh}(Y(0,U)).$$

Since $\operatorname{arcsinh}(y) \approx \log(2y)$ for y large, it follows that $\beta_1 \approx \log(Y(1,U)/Y(0,U))$ when $Y(1,U)$ and $Y(0,U)$ are large. Thus, Bellemare and Wichman (2020) argue that β_1 approximates the semi-elasiticity of the outcome with respect to d when the outcome is large. They likewise provide similar results for the elasticity of $Y(d,U)$ with respect to treatment when treatment is continuous. Their results thus imply that the ATE for $\operatorname{arcsinh}(Y)$ has a sensible interpretation as a (semi-)elasticity when the model for the potential outcomes given in (4) holds.

It is worth emphasizing, however, that (4) will generally be incompatible with the data when both $Y(1)$ and $Y(0)$ have point-mass at zero, and $\beta_1 \neq 0$. Specifically, note that (4) implies that

$$\operatorname{arcsinh}(Y(1,U)) - \operatorname{arcsinh}(Y(0,U)) = \beta_1.$$

If $\beta_1 > 0$, for example, this implies that $\operatorname{arcsinh}(Y(1,U)) > \operatorname{arcsinh}(Y(0,U))$, and hence $Y(1,U) > Y(0,U)$, since the $\operatorname{arcsinh}(y)$ function is strictly increasing for $y \geqslant 0$. However, if $Y(1,U) = 0$, this then implies that $Y(0,U) < 0$, which is a contradiction. Thus, the model in (4) is incompatible with $P(Y(1) = 0) > 0$ if $\beta_1 > 0$. By similar logic, the model is also incompatible with $P(Y(0) = 0) > 0$ if $\beta_1 < 0$. In settings where there is point-mass at zero, the model that Bellemare and Wichman (2020) show gives β_1 an interpretation as a semi-elasticity will therefore typically be rejected by the data. It is also worth noting that even if there are no zeros in the data, the model in (4) will generally be sensitive to functional form, in the sense that if (4) holds for Y measured in dollars, it will generally not hold when Y is measured in cents. The validity of the interpretation of β_1 as an elasticity thus depends on having chosen the "correct" scaling of the outcome such that (4) holds.

D.2 Cohn et al. (2022)

Cohn et al. (2022) consider structural equations of the form

$$Y = \exp(D\beta)U. \tag{5}$$

When $E[U \mid D] = 1$, they show that Poisson regression is consistent for β, whereas regressions of $\log(1+Y)$ or $\log(Y)$ on D may be inconsistent for β. Although Cohn et al. (2022) do not consider a potential outcomes interpretation of β, we can give β a causal interpreation if we treat (5) as structural, i.e. impose that the potential outcomes take the form

$$Y(d,U) = \exp(d\beta)U, \tag{6}$$

where U is an individual level shock common to all d, and $E[U] = 1$. Under (6), it follows that $\exp(\beta) = E[Y(1)]/E[Y(0)]$, which corresponds with the estimand considered in Section 4.1.[18]

[18] Bellégo, Benatia and Pape (2022) also consider (5), but consider the more general class of identifying restrictions

We note, however, that if one were instead to impose (5) with the assumption that $E[\log(U)|D] = 0$, then the regression of $\log(Y)$ on D would be consistent for β, whereas Poisson regression would generally be inconsistent for β. Indeed, under the potential outcomes model in (6) with the assumption that $E[\log(U)] = 0$, we have that $\beta = E[\log(Y(1)) - \log(Y(0))]$, the ATE in logs.[19]

This discussion highlights that whether or not an estimator is consistent depends on the specification of the *target parameter* (a.k.a. estimand). Our results help to illuminate what parameters can be consistently estimated by some estimators by enumerating the properties that identified causal estimands can (or cannot) have.

E Connection to two-part models

One approach recommended for settings with weakly-positive outcomes is to estimate a two-part model (Mullahy and Norton, 2022). In this section, we briefly review two-part models, and show that the marginal effects implied by these models do not correspond with ATEs for the intensive margin without further restrictions on the potential outcomes.

The idea of a two-part model is to separately model the conditional distribution $Y \mid D$ using (a) a first model for the probability that Y is positive given D, $P(Y > 0 \mid D)$ (b) a second model for the conditional expectation of Y given that it is positive, $E[Y \mid D, Y > 0]$. Common specifications include logit or probit for part (a), and a linear regression of Y on D for part b); see, e.g., Belotti, Deb, Manning and Norton (2015). After obtaining estimates of the two-part model, it is common to evaluate the marginal effects of D on both parts, i.e. the implied values of

$$\begin{aligned}
\tau_a &= P(Y > 0 \mid D = 1) - P(Y > 0 \mid D = 0) \\
\tau_b &= E[Y \mid Y > 0, D = 1] - E[Y \mid Y > 0, D = 0].
\end{aligned}$$

We now consider how the parameters of the two-part model relate to causal effects in the potential outcomes model. Suppose, for simplicity, that the two-part model is well-specified, so that it correctly models $P(Y > 0 \mid D)$ and $E[Y \mid Y > 0, D]$. Suppose further that D is randomly assigned, $D \perp\!\!\!\perp Y(1), Y(0)$. In this case, we have that

$$\begin{aligned}
\tau_a &= P(Y(1) > 0) - P(Y(0) > 0) \\
\tau_b &= E[Y(1) \mid Y(1) > 0] - E[Y(0) \mid Y(0) > 0]
\end{aligned}$$

From the previous display, we see that the marginal effect on the first margin, τ_a, has a causal interpretation: it is the treatment's effect on the probability that the outcome is positive.

The interpretation of the marginal effect on the second margin, τ_b, is more complicated however.

of the form $E[D \log(U + \delta)] = 0$, where δ is a tuning parameter. The appropriate choice of estimator then depends on δ, with Poisson regression and log regressions the limiting cases as $\delta \to \infty$ and $\delta \to 0$, respectively. We note, however, that Bellégo et al. (2022) impose that $E[U] = 1$, and thus the causal interpretation of β in the potential outcomes model in (6) is the same as in Cohn et al. (2022) regardless of the value of δ.

[19] Note that the assumption that $E[\log(U)] = 0$ implicitly implies that $U > 0$, and thus $Y > 0$.

For simplicity, suppose are willing to impose the "monotonicity" assumption discussed in Section 4, $P(Y(1)=0, Y(0)>0)=0$, so that anyone with a zero outcome under treatment also has a zero outcome under control. Then, letting $\alpha = P(Y(0)=0 \mid Y(1)>0)$, we can write τ_b as

$$\begin{aligned}
\tau_b &= (1-\alpha)E[Y(1) \mid Y(1)>0, Y(0)>0] + \alpha E[Y(1) \mid Y(1)>0, Y(0)=0] - E[Y(0) \mid Y(1)>0, Y(0)>0] \\
&= \underbrace{E[Y(1)-Y(0) \mid Y(1)>0, Y(0)>0]}_{\text{Intensive margin effect}} + \alpha \underbrace{\left(E[Y(1) \mid Y(1)>0, Y(0)=0] - E[Y(1) \mid Y(1)>0, Y(0)>0]\right)}_{\text{Selection term}},
\end{aligned}$$

where the first equality uses iterated expectations, and the second re-arranges terms.

The previous display shows that τ_b is the sum of two terms. The first is the ATE for individuals who would have a positive outcome regardless of treatment status (similar to $\theta_{\text{Intensive}}$ in Section 4, except using Y instead of $\log(Y)$). The second term is not a causal effect, but rather represents a selection term: it is proportional to the difference in the average value of $Y(1)$ for people who would have positive outcomes only under treatment versus people who would have positive outcomes regardless of treatment status. In many economic contexts, we may expect this selection effect to be negative. For example, we may suspect that individuals who would only get a job if they receive a particular training have lower ability, and hence lower values of $Y(1)$, than individuals who would have a job regardless of training status. The marginal effect τ_b thus only has an interpretation as an ATE along the intensive margin if either (a) there is no extensive margin effect ($\alpha = 0$) or (b) we are willing to assume that the selection term is zero. Angrist (2001) provided a similar decomposition (without imposing monotonicity), concluding that the two-part model "seems ill suited for causal inference," at least without further restrictions on the potential outcomes.