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

December 15, 2022

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., $\log(1 + Y)$, $\text{arcsinh}(Y)$) 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 seemingly reasonable 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. Moreover, we show that any parameter of the form $\theta_g = E[g(Y(1), Y(0))]$ is necessarily scale dependent if it is point-identified and defined with zero-valued outcomes. 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, $E[\log(Y(1)) - \log(Y(0))]$. An appealing feature of the ATE in logs is that its units are easily interpretable. Since $\log(Y(1)/Y(0)) \approx \frac{Y(1)-Y(0)}{Y(0)}$ when $Y(1)/Y(0)$ is close to 1, the units of the ATE in logs correspond approximately with percentages. Moreover, like percentages, the units of the ATE in logs are *scale-invariant*, meaning that the ATE in logs is the same regardless of the seemingly arbitrary choice of whether the outcome is measured in, say, dollars, cents, or yuan.

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 in a similar way to logs, i.e. as (approximate) percentage effects. We reviewed papers published in the *American Economic Review* since 2018 to determine how they interpret the units of regressions using $\text{arcsinh}(Y)$ as an outcome variable, and found that 86% (12 of 14) interpret the coefficients as percents or elasticities.[1] One of the authors has also interpreted treatment effects for $\text{arcsinh}(Y)$ in this way (Chen, Glaeser and Wessel, 2022).

An underappreciated feature of commonly-used alternative transformations such as $\log(1+Y)$ and $\text{arcsinh}(Y)$ is that—in contrast to percentages or log points—they can be sensitive to the scaling of the outcome. In an application to the National Supported Work program, for example, Aihounton and Henningsen (2021) found that the estimated treatment effect on earnings using the inverse hyperbolic sine transformation varied between 31% and 2,451% depending on the units used for the outcome.

This raises the following questions. First, when will the ATEs for transformations such as $\log(1+Y)$ or $\text{arcsinh}(Y)$ depend on the scaling of the outcome, and how much will the scaling matter? Second, can we eliminate the sensitivity to the scaling of the outcome by using an alternative transformation? We answer each of these questions in turn.

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 $m : [0, \infty) \to \mathbb{R}$ such 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

[1] 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.

if treatment affects the probability that $Y = 0$—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. In particular, the ATE for $m(aY)$ diverges as $a \to \infty$ and converges to zero as $a \to 0$. Therefore, for such a function $m(\cdot)$, it is inappropriate to interpret the ATE in $m(\cdot)$ as a percentage, since a percentage is inherently unit-invariant.

The intuition for these results 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 transformations $m(Y)$ 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 non-zero value of Y has a very large impact on the ATE, and so the ATE places an extremely large weight on the extensive margin. If treatment changes the probability that $Y > 0$, 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.

Our second main result establishes that it is not possible to achieve scale-invariance by choosing a different transformation of the potential outcomes, at least if one requires point identification. Specifically, we consider average treatment effect parameters of the form $\theta_g = E[g(Y(1), Y(0))]$, where $g(y_1, y_0)$ is increasing in y_1. We show that when Y takes on only positive values, then θ_g is generically point-identified from the marginal distributions of the potential outcomes and invariant to the scaling of the outcome if and only if θ_g is an affine transformation of the ATE in logs (i.e. $g(y_1, y_0) = c\log(y_1/y_0) + d)$). Since $\log(0)$ is not well-defined, an immediate corollary of this result is that there is no parameter θ_g that is both point-identified and scale-invariant when g is well-defined for zero-valued outcomes. Hence, the scale dependence of the $\text{arcsinh}(Y)$ and $\log(1 + Y)$ transformations discussed above is not the result of having chosen the "wrong" transformation, but rather a fundamental feature of identifiable average treatment effects that are defined at zero.

Taken together, our results imply that when Y can take on zero values, researchers should not take a "log-like" transformation of the outcome that is defined at zero and interpret the units as percentages—at least if one requires the seemingly reasonable notion that a percentage does not depend on the units of the outcome. We conclude by highlighting three alternative approaches for settings where researchers currently use log-like transformations. First, researchers might estimate an ATE for a normalized outcome—e.g. the treatment effect on earnings expressed as a percentage of pre-treatment earnings. Second, researchers might explicitly calibrate how much they value the intensive versus extensive margins by directly specifying that a change from 0 to 1 (in a particular unit) is valued as a certain percentage change in the outcome for those with $Y > 0$. Finally, the researcher might report separate estimates for the treatment effect along the intensive versus extensive margins, e.g. using bounding approaches such as in Lee (2009).

1.1 Setup and notation

We consider a setting with a binary treatment $D \in \{0, 1\}$ and a weakly positively-valued observed outcome $Y \in \mathcal{Y}$, where $\mathcal{Y} \subseteq [0, \infty)$. 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 $Y(d)$ 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. *(m is increasing and continuous) $m : [0, \infty) \to \mathbb{R}$ is a continuous, weakly increasing function.*
2. *(m behaves like* log *for large values) $m(y)/\log(y) \to 1$ as $y \to \infty$.*
3. *(Treatment affects 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 contrast, the ATEs for these transformations are arbitrarily dependent on units.

2.1 Intuition for Proposition 1

We first provide some intuition for why $|\theta(a)|$ grows large as $a \to \infty$. Heuristically, if we re-scale the outcome by a large factor a, then $m(aY)$ becomes very large for values of $Y > 0$, but does not change at all when $Y = 0$. Thus, the average value of $m(aY(d))$ becomes much larger for the treatment arm with a higher fraction of non-zero values of Y. More formally, consider $y > 0$. Then when a is large, $m(ay) \approx \log(ay) = \log(a) + \log(y)$, and thus $m(ay)$ increases at the rate $\log(a)$ when $y > 0$. On the other hand, if $y = 0$, then $m(ay) = m(0)$ and does not increase with a. It follows that $E[m(aY(d))] = P(Y(d) > 0) \cdot \log(a) \cdot (1 + o(1))$. Hence, when we re-scale Y by a factor a, the ATE increases by approximately $(P(Y(1) > 0) - P(Y(0) > 0)) \cdot \log(a)$, which diverges as a is made large. On the other hand, if we re-scale the outcome so that most of the non-zero values are small, i.e. set $a \approx 0$, then since m is a continuous function, $m(aY) \approx m(0)$, and thus the treatment effect is small.

A second related intuition for the result in Proposition 1 is that when $m(Y)$ "looks like" $\log(Y)$ for large values of Y, the units of Y implicitly determine how much the ATE weights the intensive and extensive margins. To see this, for simplicity consider the case where $P(Y(1) = 0, Y(0) > 0) = 0$, so that, e.g., everyone who has positive income without receiving a training also has positive income when receiving the training.[2] Then, by the law of iterated expectations, we can write

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

If $m(ay) \approx \log(ay)$ for non-zero values of y, then 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 $m(aY(1))$ 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.

This highlights a fundamental challenge with attempting to estimate an average "percentage" effect when there are zero outcomes: the percentage change in earnings is inherently not well-defined when an individual's outcome moves from zero under control to non-zero under treatment (or vice versa). Any average treatment effect that is well-defined with zero outcomes must therefore implicitly determine how much weight to place on the extensive margin relative to percentage changes along the intensive margin. The argument above shows that for transformations that look like $\log(Y)$ for large values of Y, the weights placed on the intensive versus extensive margins are implicitly pinned down by the scale of Y.

[2] 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 $Y(0) > Y(1) = 0$.

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. One might wonder, however, how much $\theta(a)$ changes for finite changes in the scaling a. The following Proposition gives an (approximate) answer to this question.

Proposition 2. *Under the conditions of Proposition 1,*[3] *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, changing the units from dollars to cents ($a = 100$) is expected to increase the ATE by $\log(100) = 4.6$ times the treatment effect on having a non-zero outcome. If treatment increases the probability of a non-zero outcome by 10 percentage points, then changing the units from dollars to cents would be expected to change the estimated treatment effect by 0.46, or what is typically (although incorrectly) interpreted as a change in the effect of 46 percentage points.

2.3 Additional remarks and extensions

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

Remark 2 (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. Our results thus imply 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.

Remark 3 (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

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

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 4 (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, 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 size will be sensitive to scaling even when they do not have a causal interpretation.

3 Sensitivity to scaling for other ATEs

Our results so far show that the ATEs for transformations that are defined at zero and "look like" $\log(y)$ are arbitrarily sensitive to scaling. Our next result shows that *some* sensitivity to scaling is a necessary feature of identified parameters that take the more general form $\theta_g = E_P[g(Y(1), Y(0))]$. Specifically, we show that the only parameter of this form that is scale-invariant and point-identified when Y is strictly positive is the ATE in logs (up to an affine transformation), and thus there is no such scale-invariant parameter that is well-defined at zero.

Recall that 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 $Y(d)$ 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. The marginal distributions of $P_{Y(d)}$ 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).[4]

We first make precise that point identification of $\theta_g = E_P[g(Y(1), Y(0))]$ means θ_g does not depend on the joint distribution of $(Y(1), Y(0))$. The joint distribution of the potential outcomes

[4]Under the standard parallel trends assumption, only the marginal *means* of the potential outcomes distributions are identified, an even weaker notion of identification.

is never identified from the data.

Definition 1 (Identification)**.** 1. 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 $d = 0, 1$), we have that $E_P[g(Y(1), Y(0))] = E_Q[g(Y(1), Y(0))]$.

2. We say that θ_g is *generically point identified* over $\mathcal{P}$ if for every $P \in \mathcal{P}$, θ_g is point identified from the marginals distributions at P.

Definition 2. We say that the function $g : (0, \infty)^2 \to \mathbb{R}$ is *scale-invariant* if it is homogeneous of degree zero, i.e. $g(y_1, y_0) = g(ay_1, ay_0)$ for all $a, y_1, y_0 > 0$.

We now characterize exactly when θ_g is generically point-identified and scale invariant when $Y > 0$. Specifically, let $\mathcal{P}_{++}$ denote the class of distributions with support contained in a compact subset of the positive reals. (We focus on compact subsets to avoid complications related to $E[\log(Y(d))]$ being infinite.) Our next result shows that if g is scale-invariant and increasing in y_1, then θ_g is identified from the marginals if and only if θ_g is an affine transformation of the ATE in logs.

Proposition 3. *Suppose* $g : (0, \infty)^2 \to \mathbb{R}$ *is weakly increasing in* y_1 *and scale-invariant. Then* θ_g *is generically 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* $d \in \mathbb{R}$.

An immediate consequence of Proposition 3 is that if g is well-defined for zero-valued outcomes, then either (i) θ_g is not generically identified over $\mathcal{P}_{++}$, or (ii) g must not be scale-invariant.

Corollary 1. *Suppose* $g : [0, \infty)^2 \to \mathbb{R}$ *is weakly increasing in its first argument. If* g *is not a constant function, then either (i)* θ_g *is not generically point-identified over* $\mathcal{P}_{++}$*, or (ii)* g *is not scale-invariant.*

Note that if $g(y_1, y_0) = m(y_1) - m(y_0)$, so that θ_g is the ATE for $m(Y)$ as in the previous section, then θ_g is generically identified from the marginals. It follows that if $m(0)$ is well-defined, then θ_g must necessarily be sensitive to scaling. Hence, we see that the sensitivity to scaling of commonly-used transformation such as $\log(1 + y)$ or $\text{arcsinh}(y)$ is not a specific property of the ATE for these transformations, but rather a generic feature of *any* scale-invariant average treatment effect that is well-defined with zero-valued outcomes.

Proof sketch. The proof proceeds in two steps (see Appendix B for the formal proof). The first step is to show that if θ_g is generically identified, then g must be additively separable, $g(y_1, y_0) = q(y_1) + r(y_0)$. This is achieved by showing that if g is not additive, then there exists a set of points on which g is strictly supermodular (or submodular). It follows from arguments as in Fan et al. (2017) that when P is a distribution over these points, the identified set for θ_g is determined by the Frechet-Hoeffding bounds, which yield an interval, and thus θ_g is not point-identified from the marginals. The second part of the proof shows that if g is increasing in y_1, additively separable, and scale-invariant, then it must be equal to $c \log(y_1/y_0) + d$. To see why this is the case, observe that if $g(y_1, y_0)$ is scale-invariant, then $g(y_1, y_0) = g(y_1/y_0, 1) =: h(y_1/y_0)$. From this and the additive separability of g, some basic algebra yields that

$$\tilde{h}(ab) = \tilde{h}(a) + \tilde{h}(b) \text{ for all } a, b > 0, \tag{2}$$

where $\tilde{h}(y) := h(y) - h(1)$. Equation (2) can be recognized as Cauchy's logarithmic functional equation, for which the only monotonic solution takes the form $c \log(y)$ (Aczél, 1966). □

Remark 5 (Implications for average proportional effects)**.** Although our focus is on settings where the outcome may be zero-valued, Proposition 3 has interesting implications for the case where $Y > 0$. Suppose that $Y > 0$ and we are interested in the average proportional effect of the treatment (or semi-elasticity),

$$\theta_{\text{prop}} = E\left[\frac{Y(1) - Y(0)}{Y(0)}\right].$$

This parameter is a convenient summary measure of the proportional effect of the treatment, and in some context is motivated by economic theory. For example, the Baily-Chetty formula for the optimal level of unemployment insurance involves a parameter of the form θ_{prop} where Y is consumption and D is unemployment (Baily, 1978; Chetty, 2006). Note that θ_{prop} corresponds with θ_g for $g(y_1, y_0) = (y_1 - y_0)/y_0$, which is a scale-invariant function that is not an affine transformation of $\log(y_1/y_0)$. Thus, Proposition 3 implies that θ_{prop} is not generically point-identified. It turns out, however, that θ_{prop} can be partially identified, with the upper and lower bounds given by the Frechet-Hoeffding inequality; we are working on a companion paper that derives these bounds as well as methods for estimation and inference.

3.1 Alternative version with smooth g

Proposition 3 shows that if g is scale-invariant and well-defined with zero-valued outcomes, then θ_g is not point-identified at *some* distribution $P \in \mathcal{P}_{++}$. We then might hope to find a scale-invariant treatment effect parameter that is point-identified for "most" distributions P. The following result shows that this is not possible either, at least in the case where g is smooth: if g is twice continuously differentiable and scale-invariant, then identification of θ_g fails for *all* P satisfying some basic regularity conditions.

Proposition 4. *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 $c \geqslant 0$.*

4 Recommendations

Our results above show that ATEs for common transformations such as $\log(1 + Y)$ and $\operatorname{arcsinh}(Y)$ can not be interpreted as percentage effects, given that their magnitudes depend arbitrarily on the units of the outcome. Indeed, any parameter of the form $E_P[g(Y(1), Y(0))]$ that is identified and well-defined at zero will depend on scale. Researchers using such transformations thus cannot merely interpret the units as a percentage, but must pay careful attention to the units. In this section, we outline a variety of alternative options that are available in settings where these transformations are often used.

As a guiding principle, we note that Proposition 3 informally states that only two of the following three properties are attainable at the same time when there are zero outcomes, and thus at least one of them must be relinquished:

(a) The parameter of interest is an average effect of the form $\theta_g = E[g(Y(1), Y(0))]$.

(b) The parameter of interest is scale-invariant.

(c) The parameter of interest is generically point-identified.

Section 4.1 introduces families of scale-invariant estimands that do not satisfy (a). Section 4.2 proposes to explicitly state the trade-off between the extensive and the intensive margin, under some fixed scaling, thereby jettisoning (b). Lastly, Section 4.3 proposes Lee (2009)-bounds for the intensive margin effect $E[\log Y(1) - \log Y(0) \mid Y(1), Y(0) > 0]$,[5] which is not generically point-identified.

4.1 Normalized treatment effects

One alternative in settings where researchers are interested in a treatment effect with a percentage interpretation is to express the ATE in levels as a percentage of the control mean,

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

This parameter is scale-invariant and has an intuitive percentage interpretation, and thus may be attractive in many contexts.[6] This parameter is also what Poisson regression estimates: Poisson regression (see Chapter 18.2 in Wooldridge, 2010)—using the pseudo-likelihood $Y \mid D \sim \text{Pois}(e^{\alpha+\beta D})$ for a randomly assigned treatment D—estimates the population coefficient $e^\beta = E[Y(1)]/E[Y(0)] = \theta_{\text{ATE\%}} + 1$. However, this parameter may difficult to interpret in contexts where Y spans several order of magnitudes. For example, the econometrician may perceive a change in income from \$5,000 to \$6,000 very differently from a change in income from \$100,000 to \$101,000, yet both those changes are treatment effects in levels of \$1,000 and thus contribute equally to $\theta_{\text{ATE\%}}$. Statistically, $\theta_{\text{ATE\%}}$ may also be difficult to estimate in finite samples if Y has a long-right tail.

In settings where the $\theta_{\text{ATE\%}}$ may be hard to interpret (or estimate), the researcher may instead consider estimating an ATE for a normalized variable $\tilde{Y}$ chosen so that the treatment effects for

[5] The quantity is proportional to $E[(\log Y(1) - \log Y(0))\mathbb{1}(Y(1) > 0, Y(0) > 0)]$, which takes the form of θ_g and is scale-invariant.

[6] Note that $\theta_{\text{ATE\%}}$ is not of the form $E[g(Y(1), Y(0))]$, and thus is not subject to the impossibility results in Proposition 3 and Corollary 1.

$\tilde{Y}$ are more comparable across individuals (and have less skew). Consider, for example, a setting where the researcher has access to post-treatment earnings Y as well as pre-treatment earnings Y_{pre}. Suppose that $Y_{\text{pre}} > 0$, as for example would occur if having prior work experience is a pre-requisite for participating in a training program.[7] Then the researcher might consider estimating the ATE for the normalized outcome $\tilde{Y} = Y/Y_{\text{pre}}$, i.e.

$$\theta_{\tilde{Y}} = E\left[\frac{Y(1)}{Y_{\text{pre}}} - \frac{Y(0)}{Y_{\text{pre}}}\right].$$

This parameter has the intuitive interpretation as the average treatment effect on earnings expressed as a percentage of pre-treatment earnings.

Along similar lines, the researcher might consider normalizing the outcome by the *expected* earnings given observable covariates. That is, define $\tilde{Y} = Y/E[Y(0) \mid X]$, i.e. an individual's outcome as a percentage of the average control earnings for people with the same observable characteristics X. If, say Y represents earnings and X includes education and pre-treatment earnings, then the ATE for $\theta_{\tilde{Y}}$ has the interpretation as the average change in earnings as a percentage of the control outcome for people with the same education and previous earnings.

We suspect that in many economic contexts, the outcome can be appropriately normalized so that the treatment effects for $\tilde{Y}$ are interpretable across individuals, and the transformed outcome is not too skewed. Nevertheless, this approach may not be applicable if no pre-treatment information or covariates are available, or if the distribution of the normalized outcome remains highly skewed. We therefore consider other alternatives next.

4.2 Directly valuing the intensive versus extensive margin

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—i.e., for individuals with positive earnings, one values a percentage point change in earnings the same regardless of the initial level—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, under some fixed unit, 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 is valued at x log points.[8]

We emphasize that for a fixed value of x, this approach does necessarily depend on the scaling of the outcome; however, the appropriate choice of x *should* depend on the units of the outcome—e.g.

[7] It is common for UI agencies to run training programs, and UI recipients by definition were previously employed.

[8] 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.

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. This approach thus makes explicit the tradeoff between the extensive margin and the intensive margin effect for the particular choice of units.

4.3 Separating the intensive and extensive margins

Finally, the researcher might consider separately estimating effects along 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 average effect on earnings for those who would have positive earnings regardless of their treatment status. Although $\theta_{\text{intensive}}$ is not point-identified from the marginal distributions of the potential outcomes, Lee (2009) popularized a method for bounding it 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 to $\theta_{\text{intensive}}$.[9] We note that 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, Rubin and Mealli, 2009).

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.

[9] 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 $E[\log(Y(1)) \mid Y(1) > 0, Y(0) = 0]$.

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.

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

Baily, Neil, "Some Aspects of Optimal Unemployment Insurance," *Journal of Public Economics*, 1978, pp. 379–402.

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.

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.

Chetty, Raj, "A General Formula for the Optimal Level of Social Insurance," *Journal of Public Economics*, November 2006, *90*, 1879–1901.

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.

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. *(m is increasing and continuous) $m : [0, \infty) \to \mathbb{R}$ is a continuous, weakly increasing function.*
2. *(m behaves like $\log$ for large values) $m(y)/\log(y) \to 1$ as $y \to \infty$.*
3. *(Treatment affects 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 $a - 1 < a_n < a + 1$ 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. □

Proposition 2. *Under the conditions of Proposition 1,*[10] *as $a \to \infty$,*

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

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

$$\frac{1}{\log a_n} E_P[m(a_n Y(1)) - m(a_n Y(0))] \to P(Y(1) = 0) - P(Y(0) = 0). \quad (3)$$

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

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

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

Proposition 3. *Suppose $g : (0, \infty)^2 \to \mathbb{R}$ is weakly increasing in y_1 and scale-invariant. Then θ_g is generically 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 generic point-identification 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

$$E_{P_1}(g(Y(1), Y(0))) = \frac{1}{2}\left(g(y_1, y_0) + g(y_1 + a, y_0 + b)\right)$$

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

and thus θ_g is not generically identified. Hence, if θ_g is generically identified, 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 $a,b>0$,

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

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 , Theorem 2 in Section 2.1.2.[11] 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. $\square$

Corollary 1. *Suppose $g:[0,\infty)^2\to\mathbb{R}$ is weakly increasing in its first argument. If g is not a constant function, then either (i) θ_g is not generically point-identified over $\mathcal{P}_{++}$, or (ii) g is not scale-invariant.*

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

Proof. 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 result is then immediate from Proposition 3. $\square$

Proposition 4. *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 $A = 1$,[12] Theorem 3.1 in 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.

[12]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 $A = 1$, $g(y_1^*, y_0^* + \epsilon_0) \leqslant g(Y(1), Y(0)) \leqslant g(y_1^* + \epsilon_1, y_0^*)$.

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 3. $\square$

C Extension to non-binary treatments and OLS estimands

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.[13]

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.

Proposition 5. *Suppose that:*

1. *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 $f_n(d) \to$

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

$\omega(d)E[m(aY(d))]$ pointwise. We also showed in the proof to Proposition 1 that for a_n sufficiently close to a,

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

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

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

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

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

Letting $f_n(d) = \omega(d)\frac{E[m(a_nY(d))]}{\log(a_n)}$, 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

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

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 4, 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. *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. $\square$

References for the Appendix

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

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