AEA Papers and Proceedings 2024, 114: 597–600
https://doi.org/10.1257/pandp.20241046

TREATMENT EFFECTS: THEORY AND IMPLEMENTATION

Negative Weights Are No Concern in Design-Based Specifications[†]

By Kirill Borusyak and Peter Hull*

A recent and influential literature raises a concern with popular ordinary least squares (OLS) and instrumental variable (IV) specifications: they may fail to estimate convex averages of heterogeneous treatment effects, even when they succeed at avoiding omitted variable bias (OVB). A leading example is two-way fixed effects regressions, which address OVB by modeling untreated potential outcomes as linear in unit and time dummies (a "parallel trends" assumption). Several papers show how such specifications can suffer from what Small et al. (2017) call *sign reversals*: the regression estimand can be negative, despite all causal effects being positive, because of negative weights placed on some or many causal effects.[1] More flexible specifications have been proposed to address this problem (e.g., Wooldridge 2021; Borusyak, Jaravel, and Spiess 2023).

We show that conventional specifications avoid this concern when they are "design based"—that is, when they leverage a correct model of treatment (or instrument) assignment rather than a model of potential outcomes. Specifically, we consider OLS and IV regressions in which the controls are chosen to span the expected treatment or instrument value given the potential outcomes (a generalization of the propensity score, outside of binary treatments). Such specifications attach possibly negative *ex post* weights—which depend on treatment realizations—to causal effects. However, the estimands of these design-based specifications also have an average-effect representation with *ex ante* weights: the expectations of ex post weights over the assignment distribution. The ex ante weights are guaranteed to be convex in the OLS case, and this property extends to the IV case under a general first-stage monotonicity condition. Thus, negative ex post weights pose no problems in design-based specifications.

This analysis makes two contributions to a classic literature on convex weights with OLS and IV (e.g., Imbens and Angrist 1994; Angrist 1998; Angrist and Krueger 1999; Angrist, Graddy, and Imbens 2000). First, we jointly analyze ex post and ex ante weights, which are usually studied separately.[2] This clarifies the distinction—and its implications—between specifications justified by models for unobservables and specifications based on the assignment process of observed shocks.

Second, we prove the convexity of ex ante weights under a mean-independence condition that is weaker than the typical assumption of conditional ignorability. While ignorability may be no less plausible in settings with clear design, such as randomized trials, this difference highlights the key role of what Borusyak and Hull (2023) term the *expected instrument* to avoid both OVB and sign reversals with simple specifications.

Our results also relate to a recent literature on design-based causal inference with "formula" treatments and instruments—those

*Borusyak: University of California, Berkeley (email: k.borusyak@berkeley.edu); Hull: Brown University (email: peter_hull@brown.edu). We thank Dmitry Arkhangelsky, Peng Ding, Avi Feller, Xavier Jaravel, and Jesse Shapiro for helpful comments.

[†]Go to https://doi.org/10.1257/pandp.20241046 to visit the article page for additional materials and author disclosure statement(s).

[1]See, for example, de Chaisemartin and D'Haultfœuille (2020); Goodman-Bacon (2021); and Borusyak, Jaravel, and Spiess (2023). The no-sign-reversal property is what Blandhol et al. (2022) call "weakly causal" estimands.

[2]Notable exceptions include Arkhangelsky et al. (2023); Goldsmith-Pinkham, Hull, and Kolesár (2022); and de Chaisemartin and Lei (2023), who provide joint analyses in certain special cases.

constructed from a common set of exogenous shocks and nonrandom measures of exposure. Our mean-independence assumption builds on Borusyak, Hull, and Jaravel (2022), who establish convex ex ante weights with shift-share instruments; we show that their result holds more generally.[3]

I. Ex Post and Ex Ante Weights

We first show the results in a simple setting. Let y_i and x_i be an outcome and treatment observed in a sample of units i. Consider OLS estimation of

$$y_i = \beta x_i + w_i'\gamma + e_i, \tag{1}$$

where w_i is a low-dimensional vector of controls that includes a constant.

To interpret the estimate of β, we suppose that the outcome is generated by a causal model with linear but heterogeneous effects β_i:

$$y_i = x_i\beta_i + \varepsilon_i.$$

Here, ε_i is an untreated potential outcome—that is, the outcome that unit i would see when x_i is set to zero.[4]

Suppose that appropriate asymptotics apply, such that OLS consistently estimates:

$$\beta = \frac{E\left[\tilde{x}_i y_i\right]}{E\left[\tilde{x}_i^2\right]} = \frac{E\left[\tilde{x}_i x_i \beta_i\right] + E\left[\tilde{x}_i \varepsilon_i\right]}{E\left[\tilde{x}_i^2\right]},$$

where $\tilde{x}_i$ denotes the residuals from the population projection of x_i on w_i.

Now consider two assumptions, either of which might motivate specification (1):

ASSUMPTION 1: $E\left[\varepsilon_i \mid x_i, w_i\right] = w_i'\gamma$.

ASSUMPTION 2: $E\left[x_i \mid \varepsilon_i, \beta_i, w_i\right] = w_i'\lambda$.

Assumption 1 models untreated potential outcomes as being linear in the controls, given the treatment. An example is the parallel trends assumption, where i indexes unit-period pairs in a panel and w_i includes unit and time dummies.[5] In constrast, Assumption 2 specifies treatment as conditionally mean-independent of potential outcomes, with a linear expected treatment $E\left[x_i \mid w_i\right]$ (e.g., the propensity score, for binary x_i). An example is an experiment where dosage x_i is randomly assigned but with different probabilities depending on strata captured by a set of dummies w_i.[6]

Under either assumption, OLS estimates from (1) avoid OVB: $E\left[\tilde{x}_i \varepsilon_i\right] = 0$.[7] Hence,

$$\beta = E\left[\psi_i \beta_i\right] / E\left[\psi_i\right], \quad \psi_i = \tilde{x}_i x_i,$$

using the fact that $E\left[\tilde{x}_i x_i\right] = E\left[\tilde{x}_i^2\right]$. This shows that the regression estimand can be interpreted as a weighted average of heterogeneous effects β_i, with weights ψ_i. We term these *ex post* weights, as they are functions of (i.e., determined after) the treatment.

Except in special cases, the ex post weighting scheme is not convex. This is because $\tilde{x}_i$ necessarily takes on both positive and negative values, since $E\left[\tilde{x}_i\right] = 0$. Thus, $\tilde{x}_i x_i$ will typically also take on positive and negative values.[8] Suppose, for example, that x_i is strictly positive. Units with low x_i (and thus $\tilde{x}_i < 0$) serve as the effective control group, and receive negative weights, but they also contribute $x_i\beta_i$ to y_i. This can lead to sign reversals under Assumption 1.

This intuition is misleading for design-based specifications justified by Assumption 2, however. In an experiment, it is random which units are in the effective control group: each unit can be assigned a low x_i, with ex post weight $\psi_i < 0$, but can as well be assigned a high x_i with $\psi_i > 0$. On average, prior to treatment assignment, all units in a strata expect the same weight. As it turns out, these expected (or *ex ante*) weights are always nonnegative under Assumption 2—avoiding sign reversals.

[3] Other settings in this literature include network spillovers (Borusyak and Hull 2023) and simulated IVs for policy eligibility (Borusyak and Hull 2021).

[4] Effect linearity is without loss for binary x_i; we consider a general causal model in Section II.

[5] The assumption of low-dimensional w_i is violated with unit fixed effects in short panels, but the negative ex post weight issue extends directly to that case.

[6] Here, Assumption 1 also holds since w_i is saturated.

[7] $E\left[\tilde{x}_i \varepsilon_i\right] = E\left[\tilde{x}_i E\left[\varepsilon_i \mid x_i, w_i\right]\right] = E\left[\tilde{x}_i w_i'\gamma\right] = 0$ under Assumption 1 since $E\left[\tilde{x}_i w_i\right] = 0$. Under Assumption 2, $E\left[\tilde{x}_i \varepsilon_i\right] = E\left[E\left[\tilde{x}_i \mid w_i, \varepsilon_i\right]\varepsilon_i\right] = 0$ since $E\left[\tilde{x}_i \mid w_i, \varepsilon_i\right] = 0$.

[8] One special case with no negative ex post weights is when x_i is binary and w_i is saturated.

Formally, under Assumption 2, we have $\beta = E[\phi_i\beta_i]/E[\phi_i]$ for ex ante weights:[9]

$$\phi_i = E[\psi_i \mid w_i, \beta_i] = \text{var}(x_i \mid w_i, \beta_i) \geq 0.$$

Comparing Assumption 2 to possible alternatives shows that the key to convex weighting is design-based specification of the expected treatment. On one hand, stronger models of unobservables may not suffice: for example, even if Assumption 1 is enriched to include a model of causal effects, $E[\beta_i \mid x_i, w_i] = w_i'\delta$, sign reversals remain possible.[10] On the other hand, stronger design assumptions like conditional unconfoundedness, $x_i \perp\!\!\!\perp (\varepsilon_i, \beta_i) \mid w_i$ (maintaining linearity, $E[x_i \mid w_i] = w_i'\lambda$), are unnecessary.[11]

II. General Result

We now consider a general causal model with potential outcomes $y_i(x)$, where $y_i = y_i(x_i)$. We suppose that x_i is continuously distributed and $x_i \geq 0$; analogous results apply in the discrete case.[12] Causal effects in this model are written $\beta_i(x) = (\partial/\partial x)y_i(x)$.

Consider IV estimation of equation (1) with x_i instrumented by some z_i. OLS estimation corresponds to the case of $z_i = x_i$. We replace Assumptions 1 and 2 with their natural generalizations:

ASSUMPTION 1′: $E[y_i(0) \mid z_i, w_i] = w_i'\gamma$.

ASSUMPTION 2′: $E[z_i \mid y_i(\cdot), w_i] = w_i'\lambda$.

We also introduce a stochastic first-stage monotonicity condition, which is trivially satisfied in the OLS case:

ASSUMPTION 3: *Almost surely over* $(y_i(\cdot), w_i)$, $\Pr(x_i \geq x \mid z_i = z, y_i(\cdot), w_i)$ *is weakly increasing in* z *for all* x.

We assume that the IV estimator of (1) consistently estimates $\beta = E[\tilde{z}_i y_i]/E[\tilde{z}_i x_i]$, where $\tilde{z}_i$ denotes the residuals from the population projection of z_i on w_i and $E[\tilde{z}_i x_i] \neq 0$. Online Appendix B then proves the general result.

PROPOSITION 1: *Under either Assumption 1′ or Assumption 2′,*

$$\beta = E\Big[\int \psi_i(x)\beta_i(x)dx\Big]/E\Big[\int \psi_i(x)dx\Big],$$

with ex post weights $\psi_i(x) = \tilde{z}_i \cdot \mathbf{1}\{x_i \geq x\}$ *that may be negative. Assumption 2′, however, further yields*

$$\beta = E\Big[\int \phi_i(x)\beta_i(x)dx\Big]/E\Big[\int \phi_i(x)dx\Big],$$

with ex ante weights $\phi_i(x) = E[\psi_i(x) \mid y_i(\cdot), w_i] = \text{cov}(\tilde{z}_i, \mathbf{1}\{x_i \geq x\} \mid y_i(\cdot), w_i)$ *that are nonnegative under Assumption 3.*

Proposition 1 extends classic results on convex weighting in two ways. First, like Assumption 2, Assumption 2′ only imposes mean-independence of the instrument from potential outcomes given the controls. Second, as with the condition in Small et al. (2017), Assumption 3 is weaker than conventional first-stage monotonicity. In particular, it does not require the first-stage relationship between z_i and x_i to be causal.

Online Appendix C shows how Proposition 1 applies to specifications with formula instruments, nesting existing results on their interpretation with heterogeneous effects.

III. Conclusion

We have shown that design-based OLS and IV specifications avoid the recent sign-reversal concern. Four caveats to this result are worth highlighting. First, even with specifications based on outcome models with negative ex post weights, sign reversals need not occur as effect heterogeneity can be limited or idiosyncratic (see, e.g., de Chaisemartin and D'Haultfœuille 2020). Second, avoiding sign reversals may not be enough to bring the estimand close to the (unweighted) average treatment effect or other

[9] We prove this in online Appendix A. Note that the convex-average representation is nonunique: β can also be written as averaging β_i with weights $E[\tilde{x}_i^2 \mid \beta_i]$ or $\tilde{x}_i^2$.

[10] However, under that additional assumption, an alternative specification that includes the interaction $x_i \times w_i$ is not subject to the sign-reversal problem and in fact identifies the average effect $E[\beta_i]$ under an overlap condition (cf. Imbens and Wooldridge 2009, 28).

[11] A benefit of unconfoundedness is that the ex ante weights reduce to $\phi_i = \text{var}(x_i \mid w_i)$ (as in Angrist and Krueger 1999) and are thus identified. This allows for a reweighted specification identifying $E[\beta_i]$, again assuming overlap ($\phi_i > 0$).

[12] The lower bound of x_i is normalized to zero without loss. Under regularity conditions, the results extend to the trivial lower bound of $-\infty$.

policy-relevant averages. Third, negative ex ante weights do generally arise in design-based specifications involving multiple treatments, including multiple bins of the same treatment (Goldsmith-Pinkham, Hull, and Kolesár 2022). Fourth, our ex ante weight characterization may not apply to design-based specifications with high-dimensional fixed effects or other controls (Freedman 2008; de Chaisemartin and Lei 2023). On the other hand, in settings with a clear design, fixed-effect estimation may be unattractive for other reasons (see, e.g., Roth and Sant'Anna 2023).

REFERENCES

Angrist, Joshua D. 1998. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants." *Econometrica* 66 (2): 249–88.

Angrist, Joshua D., Kathryn Graddy, and Guido W. Imbens. 2000. "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish." *Review of Economic Studies* 67 (3): 499–527.

Angrist, Joshua D., and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." In *Handbook of Labor Economics*, Vol. 3A, edited by Orley C. Ashenfelter and David Card, 1277–366. Amsterdam: Elsevier.

Arkhangelsky, Dmitry, Guido W. Imbens, Lihua Lei, and Xiaoman Luo. 2023. "Design-Robust Two-Way-Fixed-Effects Regression for Panel Data." arXiv: 2107.13737v2.

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky. 2022. "When Is TSLS Actually LATE?" NBER Working Paper 29709.

Borusyak, Kirill, and Peter Hull. 2021. "Efficient Estimation with Non-Random Exposure to Exogenous Shocks." Unpublished.

Borusyak, Kirill, and Peter Hull. 2023. "Nonrandom Exposure to Exogenous Shocks." *Econometrica* 91 (6): 2155–85.

Borusyak, Kirill, Peter Hull, and Xavier Jaravel. 2022. "Quasi-Experimental Shift-Share Research Designs." *Review of Economic Studies* 89 (1): 181–213.

Borusyak, Kirill, Xavier Jaravel, and Jann Spiess. 2023. "Revisiting Event Study Designs: Robust and Efficient Estimation." arXiv: 2108.12419v4.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2020. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." *American Economic Review* 110 (9): 2964–96.

de Chaisemartin, Clément, and Ziteng Lei. 2023. "More Robust Estimators for Instrumental-Variable Panel Designs, with an Application to the Effect of Imports from China on US Employment." arXiv: 2103.06437.

Freedman, David A. 2008. "On Regression Adjustments to Experimental Data." *Advances in Applied Mathematics* 40 (2): 180–93.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár. 2022. "Contamination Bias in Linear Regressions." NBER Working Paper 30108.

Goodman-Bacon, Andrew. 2021. "Difference-in-Differences with Variation in Treatment Timing." *Journal of Econometrics* 225 (2): 254–77.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62 (2): 467–75.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47 (1): 5–86.

Roth, Jonathan, and Pedro H. C. Sant'Anna. 2023. "Efficient Estimation for Staggered Rollout Designs." *Journal of Political Economy Microeconomics* 1 (4): 669–709.

Small, Dylan S., Zhiqiang Tan, Roland R. Ramsahai, Scott A. Lorch, and M. Alan Brookhart. 2017. "Instrumental Variable Estimation with a Stochastic Monotonicity Assumption." *Statistical Science* 32 (4): 561–79.

Wooldridge, Jeffrey M. 2021. "Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators." Unpublished.