Negative Weights are No Concern in Design-Based Specifications*

Kirill Borusyak† Peter Hull‡

December 30, 2023

Abstract

Recent work shows that popular partially-linear regression specifications can put negative weights on some treatment effects, potentially producing incorrectly-signed estimands. We counter by showing that negative weights are no problem in design-based specifications, in which low-dimensional controls span the conditional expectation of the treatment. Specifically, the estimands of such specifications are convex averages of causal effects with "ex-ante" weights that average the potentially negative "ex-post" weights across possible treatment realizations. This result extends to design-based instrumental variable estimands under a first-stage monotonicity condition, and applies to "formula" treatments and instruments such as shift-share instruments.

*We thank Dmitry Arkhangelsky, Peng Ding, Avi Feller, Xavier Jaravel, and Jesse Shapiro for comments.
†University of California, Berkeley. Email: k.borusyak@berkeley.edu
‡Brown University. Email: peter_hull@brown.edu

1 Introduction

A recent and influential literature raises a concern with popular ordinary least squares (OLS) and instrumental variable (IV) specifications: that 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 assuming untreated potential outcomes are linear in unit and time dummies (the popular "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 individual effects.[1] More flexible specifications have been proposed to tackle this problem (e.g. Wooldridge, 2021; Borusyak, Jaravel and Spiess, 2023).

We show that conventional specifications avoid this concern when they are "design-based": i.e., 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 the "expected instrument" (Borusyak and Hull, 2023) to avoid both OVB and sign reversals with simple specifications.

Our results also relate to a recent literature on design-based causal inference with "for-

[1] See, e.g., 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 are Arkhangelsky et al. (2023) and Goldsmith-Pinkham, Hull and Kolesár (2022), who provide joint analyses in certain special cases.

mula" treatments and instruments—those constructed from a common set of exogenous shocks and non-random 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 their result holds more generally.[3]

2 *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. We 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 resulting estimate of β, we suppose 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: i.e., the outcome that unit i would see when x_i is set to zero. Effect linearity is without loss for binary x_i; we consider a more general potential outcome model in Section 3.

Suppose appropriate asymptotics apply, such that OLS consistently estimates:

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

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. $E[\varepsilon_i \mid x_i, w_i] = w_i'\gamma$.

Assumption 2. $E[x_i \mid \varepsilon_i, \beta_i, w_i] = w_i'\lambda$.

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.[4] In constrast, Assumption 2 specifies treatment as conditionally mean-independent of potential outcomes, with a linear expected treatment $E[x_i \mid w_i]$ (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.[5]

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

[4] While our assumption that w_i is low-dimensional is violated with unit fixed effects in short panels, the negative ex-post weight issue discussed below extends directly to that case.

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

Under either assumption, OLS estimates from (1) avoid OVB: $E[\tilde{x}_i\varepsilon_i] = 0$.[6] Hence:

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

using the fact that $E[\tilde{x}_i x_i] = E[\tilde{x}_i^2]$. This shows 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 $E[\tilde{x}_i] = 0$. Thus, $\tilde{x}_i x_i$ will typically also take on positive and negative values.[7] Suppose, e.g., 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 could 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 non-negative under Assumption 2—avoiding sign reversals.

Formally, under Assumption 2 we have $\beta = E[\phi_i\beta_i]/E[\phi_i]$ for ex-ante weights:[8]

$$\phi_i = E[\psi_i \mid w_i, \beta_i] = Var(x_i \mid w_i, \beta_i) \geq 0.$$

Comparing Assumption 2 to different alternatives reveals that the key to convex weighting is the design-based specification of the expected treatment. On one hand, stronger models of unobservables may not suffice: e.g., 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.[9] On the other hand, stronger design assumptions like conditional unconfoundedness, $x_i \perp\!\!\!\perp (\varepsilon_i, \beta_i) \mid w_i$ (coupled with the linearity assumption $E[x_i \mid w_i] = w_i'\lambda$), turn out to be unnecessary.[10]

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

[7] One special case with no negative ex-post weights is when x_i are binary w_i is saturated.

[8] This follows since $E[\psi_i\beta_i] = E[E[\psi_i \mid w_i, \beta_i]\beta_i]$ and $E[\psi_i \mid w_i, \beta_i] = E[\tilde{x}_i^2 \mid w_i, \beta] + E[\tilde{x}_i \mid w_i, \beta_i]w_i'\lambda = Var(x_i \mid w_i, \beta_i)$, and similarly for $E[\psi_i]$. Note the convex-average representation is non-unique: β can also be written as averaging β_i with weights $E[\tilde{x}_i^2 \mid \beta_i]$ or $\tilde{x}_i^2$.

[9] 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 (c.f. Imbens and Wooldridge, 2009, Ch. 5.3).

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

3 General Result

We now consider a general causal model with potential outcomes $y_i(x)$, where $y_i = y_i(x_i)$. We suppose x_i is continuously distributed and $x_i \geq 0$; analogous results apply in the discrete case.[11] Causal effects in this model are written $\beta_i(x) = \frac{\partial}{\partial x} y_i(x)$.

We 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 assumption, 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 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 $E[\tilde{z}_i x_i] \neq 0$. Appendix A then proves the general result:

Proposition 1. *Under either Assumption 1′ or Assumption 2′:*

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

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:*

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

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

Proposition 1 extends classic results on convex weighting in design-based OLS and IV specifications in two ways. First, like Assumption 2, Assumption 2′ only imposes mean-independence of the instrument from potential outcomes given the controls. Full independence is not needed to avoid sign reversals. Second, as with the stochastic monotonicity condition of 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.

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

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

4 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 to other 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; Arkhangelsky et al., 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): 249–288.

Angrist, Joshua D., and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." In *Handbook of Labor Economics*. Vol. 3, 1277–1366. Elsevier.

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

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

Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky. 2022. "When is TSLS actually late?" National Bureau of Economic Research Working Paper 29709.

Borusyak, Kirill, and Peter Hull. 2021. "Efficient estimation with non-random exposure to exogenous shocks." Mimeo.

Borusyak, Kirill, and Peter Hull. 2023. "Non-Random Exposure to Exogenous Shocks." *Econometrica*, 91(6): 2155–2185.

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: 2107.13737.

de Chaisemartin, Clément, and Xavier D'Haultfœuille. 2020. "Two-way fixed effects estimators with heterogeneous treatment effects." *American Economic Review*, 110(9): 2964–2996.

Freedman, David A. 2008. "On regression adjustments to experimental data." *Advances in Applied Mathematics*, 40(2): 180–193.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár. 2022. "Contamination bias in linear regressions." National Bureau of Economic Research Working Paper 30108.

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics*, 225(2): 254–277.

Imbens, Guido W, and Jeffrey M Wooldridge. 2009. "Recent developments in the econometrics of program evaluation." *Journal of Economic Literature*, 47(1): 5–86.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica*, 62(2): 467–475.

Roth, Jonathan, and Pedro HC 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).

Wooldridge, Jeffrey M. 2021. "Two-way fixed effects, the two-way mundlak regression, and difference-in-differences estimators." SSRN, 3906345.

A Proof of Proposition 1

We assume that all the relevant moments exist and the conditions for Fubini's theorem hold. Consider first the numerator of $\beta = E[\tilde{z}_i y_i]/E[\tilde{z}_i x_i]$. Following the same steps as in footnote 6, $E[\tilde{z}_i y_i(0)] = 0$ under either Assumption $1'$ or $2'$. Thus:

$$\begin{aligned} E[\tilde{z}_i y_i] &= E[\tilde{z}_i y_i(0)] + E\left[\tilde{z}_i \int_0^{x_i} \frac{\partial}{\partial x} y_i(x) dx\right] \\ &= E\left[\int_0^{\infty} \psi_i(x)\beta_i(x) dx\right] \end{aligned}$$

for $\psi_i(x) \equiv \tilde{z}_i \mathbf{1}[x_i \geq x]$. Similarly, for the denominator, $E[\tilde{z}_i x_i] = E\left[\int_0^{\infty} \psi_i(x) dx\right]$. The ex-post $\psi_i(x)$ weights can clearly be negative, since $E[\tilde{z}_i] = 0$.

Under Assumption $2'$:

$$E\left[\int_0^\infty \psi_i(x)\beta_i(x)dx\right] = E\left[\int_0^\infty E\left[\psi_i(x)\beta_i(x) \mid y_i(\cdot), w_i\right] dx\right]$$
$$= E\left[\int_0^\infty \phi_i(x)\beta_i(x)dx\right]$$

for $\phi_i(x) \equiv E[\psi_i(x) \mid y_i(\cdot), w_i] = Cov(\tilde{z}_i, \mathbf{1}[x_i \geq x] \mid y_i(\cdot), w_i)$. Moreover, under Assumption 3, the ex-ante $\phi_i(x)$ weights are non-negative:

$$Cov(\tilde{z}_i, \mathbf{1}[x_i \geq x] \mid y_i(\cdot), w_i) = Cov(\tilde{z}_i, Pr(x_i \geq x \mid z_i, y_i(\cdot), w_i) \mid y_i(\cdot), w_i) \geq 0,$$

since both $\tilde{z}_i$ and $Pr(x_i \geq x \mid z_i, y_i(\cdot), w_i)$ are non-decreasing in z_i, conditional on $(y_i(\cdot), w_i)$.

Q.E.D.

B Formula Treatments and Instruments

Borusyak and Hull (2023), henceforth BH, study formula treatments and instruments of the form $z_i = f_i(g, s)$ for known $f_i(\cdot)$, observed shocks $g = (g_1, \ldots, g_K)$ that potentially vary at a different "level" k, and other observed data s of arbitrary structure. They assume the shocks are exogenous in the sense of $g \perp\!\!\!\perp y(\cdot) \mid v$, where $y(\cdot)$ is the set of potential outcomes and v is some set of observed variables that includes s. They further assume the shock "assignment process"—i.e., the conditional distribution of g given v—is known or consistently estimable. Borusyak, Hull and Jaravel (2022), henceforth BHJ, study the class of shift-share formulas $z_i = \sum_k s_{ik} g_k$ under the identifying assumption closer to Assumption $2'$: that $E[g_k \mid y(\cdot), v] = q_k'\xi$ for some observed q_k collected in v along with $s = \{s_{ik}\}$. This assumption weakens the full independence condition in BH while also only specifying the mean of shocks, rather than the full assignment process.

BH show that OVB is avoided in their setting when the controls in w_i are functions of v and linearly span $\mu_i = E[f_i(g, s) \mid v]$: the expected instrument over draws of the shocks. Similarly, BHJ show that OVB is avoided with shift-share z_i when $\sum_k s_{ik} q_k$ is controlled for, which follows because w_i spans $\mu_i = E[\sum_k s_{ik} E[g_k \mid y(\cdot), v] \mid v] = \sum_k s_{ik} q_k'\xi$. While both BH and BHJ focus on constant-effects models, where OVB is the only concern, they also discuss identification of convex-weighted averages of heterogeneous effects.

We show Assumption $2'$ is satisfied in both frameworks. Suppose w_i is a function of v and linearly spans μ_i: i.e., $\mu_i = w_i'\lambda$. Then, under either the BH or BHJ assumptions:

$$\begin{aligned} E[z_i \mid y_i(\cdot), w_i] &= E[E[f_i(g, s) \mid v, y_i(\cdot)] \mid y_i(\cdot), w_i] \\ &= E[E[f_i(g, s) \mid v] \mid y_i(\cdot), w_i] \\ &= E[\mu_i \mid y_i(\cdot), w_i] \\ &= w_i'\lambda. \end{aligned}$$

We also show how convexity of ex-ante weights can follow with formula instruments in a way similar (but not identical) to imposing Assumption 3. Specifically, we verify $\phi_i(x) \geq 0$ for a shift-share case where treatment is generated by the general causal model $x_i = h(g, \eta_i)$ where h is monotone in g. Suppose the instrument $z_i = \sum_k s_{ik} g_k$ is constructed from the correct shocks g_k but the shares $s_{ik} \geq 0$ do not correctly capture the dependence of x_i on g, which can be nonlinear or linear with different shares. Thus, the first-stage is non-causal. Assume that, conditional on $(y_i(\cdot), \eta_i, s_i, w_i)$ for $s_i = \{s_{ik}\}$, the g_k are drawn independently and, without loss, with zero mean (such that $\mu_i = 0$, making controls unnecessary). Then, for any x and almost-surely:

$$Cov\left(\sum_k s_{ik} g_k,\ \mathbf{1}\left[h(g, \eta_i) \geq x\right] \mid y_i(\cdot), \eta_i, s_i, w_i\right) \geq 0$$

by Lemma 2 of BH, since both $\sum_k s_{ik} g_k$ and $\mathbf{1}\left[h(g, \eta_i) \geq x\right]$ are non-decreasing functions of the mutually independent g_k, conditionally on $(y_i(\cdot), \eta_i, s_i, w_i)$. Hence:

$$\phi_i(x) = E\left[Cov\left(\sum_k s_{ik} g_k, \mathbf{1}\left[h(g, \eta_i) \geq x\right] \mid y_i(\cdot), \eta_i, s_i, w_i\right) \mid y_i(\cdot), w_i\right] \geq 0.$$

We finally note that both BHJ and BH do not require random or exchangeable samples, allowing each unit to have its own data-generating process. Our results apply in that case if moments are replaced with their full-sample versions: i.e., replacing $E[\tilde{z}_i y_i]$ with $\frac{1}{N} \sum_{i=1}^{N} E[\tilde{z}_i y_i]$ where N is the number of observations.