Causal Exaggeration: Unconfounded but Inflated Causal Estimates

Vincent Bagilet [*]

Abstract

The credibility revolution in economics has made causal inference methods ubiquitous. At the same time, an increasing amount of evidence has highlighted that the literature strongly favors statistically significant results. I show that these two phenomena interact in a way that can substantially worsen the reliability of published estimates: even when causal identification strategies successfully reduce bias caused by confounders, they can decrease statistical power and create another type of bias, leading to exaggerated effect sizes. This exaggeration is consequential in environmental economics, as cost-benefit analyses turn estimates into decision-making parameters for policy makers. I characterize this trade-off using a formal mathematical derivation and realistic Monte Carlo simulations replicating prevailing identification strategies. I then discuss potential avenues to address it.

Preliminary introduction

[*] Columbia University, New York, USA. Email: vincent.bagilet@columbia.edu. A previous version of this paper was co-authored with Léo Zabrocki-Hallak; I cannot thank him enough for his invaluable contributions to the project. I am very grateful to Hélène Ollivier and Jeffrey Shrader for their guidance and thank Sylvain Chabé-Ferret, Jesse McDevitt-Irwin, David McKenzie, José Luis Montiel Olea, Suresh Naidu, Claire Palandri, Stephan Thies and Roberto Zuniga Valladares for helpful comments, as well as lab members at Columbia and seminars participants at Columbia, IPWSD, the Paris School of Economics and the Toulouse School of Economics.

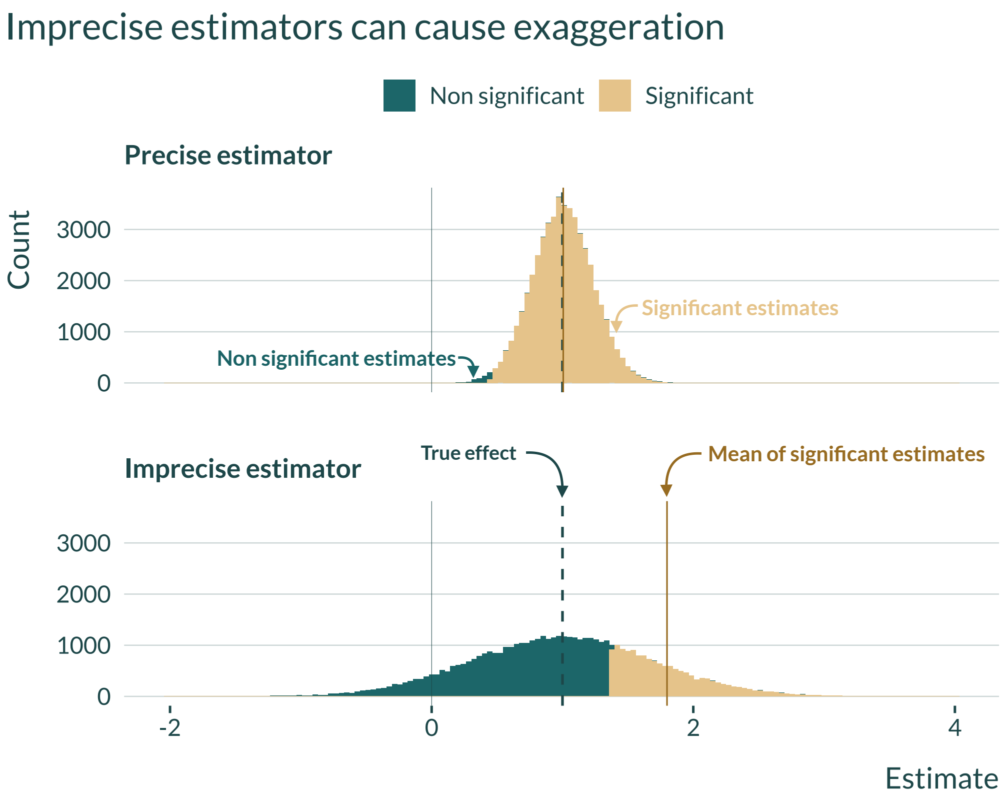
I Introduction

One of the main challenges of empirical economics is identifying causal effects. Identifications strategies such as Regression Discontinuity (RD), Instrumental Variables (IV), Difference-in-Differences (DID) and event studies help us achieve this goal. To do so, these strategies only use part of the variation in the data. They exploit the exogenous part of the variation in the treatment or decrease the sample size by only considering observations for which the as-if random assignment assumption is credible. This reduction in the variation used can decrease precision and thus statistical power—the probability of rejecting the null hypothesis when it is false, or put simply, the probability of obtaining a statistically significant estimate. There is therefore a tension between reducing confounding and statistical power.

When statistical power is low, not only is the estimator imprecise but statistically significant estimates exaggerate the true effect size (Ioannidis 2008, Gelman and Carlin 2014, Lu et al. 2019, Zwet and Cator 2021). Only estimates at least 1.96 standard errors away from zero are statistically significant at the 5% level. In under-powered studies, these estimates make up a selected sub-sample of all estimates, located in the tails of the distribution of all possible estimates. The average of these statistically significant estimates differs from the true effect, located at the center of the distribution if the estimator is unbiased. In addition, the less precise the estimator, the larger exaggeration is. Figure 1 illustrates the inflation of significant estimates caused by imprecision. When power is low, obtaining a statistically significant estimate from an unbiased estimator does not guarantee that it will be close to the true effect. An estimator $\hat{\beta}$ of the true effect β might be unbiased in the traditional sense of $\mathbb{E}[\hat{\beta}] = \beta$ but conditionally biased in the sense that $\mathbb{E}[\hat{\beta}| \text{ Significant}] \neq \beta$. For statistically significant estimates, the tension between statistical power and reducing confounding is thus a tension between reducing confounding and exaggerating the true effect size.

This consequence of low statistical power could be non-problematic absent two circumstances. First, a large literature underlined the existence of a publication bias favoring statistically significant results in economics (Rosenthal 1979, Andrews and Kasy 2019, Abadie 2020, Brodeur et al. 2020, for instance). Published estimates from under-powered studies thus form a biased sample of the distribution of estimates and can greatly exaggerate true effect sizes. Second, the economics literature, as others, suffers from a frequent and substantial lack of statistical power. The resulting exaggeration has been documented as participating in the current replication crisis affecting various fields such as economics, epidemiology, medicine or psychology (Button et al. 2013, Open Science Collaboration 2015, Camerer et al. 2016, Chang and Li 2022). Even in experimental economics, with a high level of control and an arguable absence of confounders, estimates published in top economic journals that have been replicated were on average inflated by a factor of at least

Figure 1 – Significance and distribution of two unbiased estimators with different variances



Notes: 100,000 draws from two normal distributions $\mathcal{N}(1, 0.05)$ and $\mathcal{N}(1, 0.5)$.

1.5 (Camerer et al. 2016). Quasi-experimental studies are likely even more exposed to this exaggeration issue as, in current practices, statistical power is not central to such analyses. Several meta-analyses provide clear evidence of consequential exaggeration in the non-experimental economics literature. Ioannidis et al. (2017) finds that the median statistical power in a wide range of areas of economics is no more than 18%. Despite the widespread use of convincing causal identification strategies and usually large sample sizes, they show that nearly 80% of estimates are likely exaggerated by a factor of two. In environmental economics, **?** finds that 56% of estimates are exaggerated by a factor of two or more. In a companion paper, I document evidence of substantial exaggeration in a subfield of this literature, that on the acute health effects of air pollution (**?**). The magnitude of exaggeration is thus considerable and could in some situations be on par with that of a bias caused by confounders. It is thus crucial to take exaggeration into account and to understand its drivers.

In this paper, I argue that the use of causal identification strategies contributes to exaggeration. Using a mathematical derivation and Monte Carlo simulations, I show that design choices in quasi-experimental studies can be seen as a trade-off between avoiding confounding and overestimating true effect sizes due to a resulting loss in power. To limit the threat of confounding, causal inference methods discard variation and therefore reduce statistical power. When combined with a statistical significance filter, this results in exaggeration bias. While causal identification strategies are essential to describe causal relationships, this paper emphasizes that a perfectly convincing identification does not guaranty an absence of "bias" and that improving identification can actually pull us away from the true effect. The same strategies which remove the bias caused by confounding factors actually introduce another type of bias.

All causal inference methods discard variation in order to identify causal effects but the confounding-exaggeration trade-off is mediated through a distinctive channel for each of them. In RD designs, even when the initial sample size is large, we discard part of the variation by only considering observations within the bandwidth, decreasing the effective sample size and thus precision. In an IV setting, we only use the subset of the variation in the treatment that is explained by the instrument. In studies leveraging exogenous shocks, the variation used to identify an effect sometimes only comes from a limited number of changes in treatment status. Approaches that do not actually leverage natural experiments but aim to identify a causal effect by controlling for confounders also limit the variation used. Matching prunes units that cannot be matched and thus reduces the effective sample size. Adding controls or fixed effects to the model can increase the variance of the estimator and exaggeration if they absorb more of the variation in the treatment than in the outcome variable.

Since causal identification strategies can be interpreted as ways of controlling for confounders, this last point actually ties all the strategy-specific arguments together. We implement Fixed Effects (FEs) based identification strategies such as DiD to control for the invariant, unobserved, and arguably endogenous part of the variation in the outcome. In the Control Function (CF) approach to IV, we control for the variation in x unexplained by the instruments. Fuzzy-RD and propensity score matching can also be thought of as control function approaches, of the forcing variable and propensity score respectively. In addition, excluding observations that are outside the bandwidth or unmatched is equivalent to controlling for observation-level fixed effects for these observations. When these methods for controlling for confounders absorb more of the variation in the treatment than in the outcome, they will increase the variance and cause exaggeration. Considering a simple linear homoskedastic model gives the intuition for this trade-off between exaggeration and omitted variable bias (OVB) for control approaches. Let $y_i = \alpha + \beta x_i + \delta w_i + u_i$, $\forall i \in \{1, .., n\}$, with x the variable of interest, w a potentially unobserved variable and u an error term. Under usual assumptions and using the Frisch-Waugh-Lovell theorem, we get that σ^2_{OVB} and σ^2_{CTRL}, the variance of the estimators for β when omitting w (short regression) and controlling for it (long regression) are respectively:

$$\sigma^2_{\text{OVB}} = \frac{\sigma^2_{u_{\text{OVB}}}}{n\ \sigma^2_x} = \frac{\sigma^2_{y^{\perp x}}}{n\ \sigma^2_x} \qquad \text{and} \qquad \sigma^2_{\text{CTRL}} = \frac{\sigma^2_{u_{\text{CTRL}}}}{n\ \sigma^2_{x^{\perp w}}} = \frac{\sigma^2_{y^{\perp x, w}}}{n\ \sigma^2_{x^{\perp w}}}$$

where $\sigma^2_{u_{\text{OVB}}}$ and $\sigma^2_{u_{\text{CTRL}}}$ are the variances of the residuals in the regression of y on x and of y on x and w respectively, $\sigma^2_{y^{\perp x}}$ and $\sigma^2_{y^{\perp x, w}}$ are the variances of the parts of y that are orthogonal to x and to x and w respectively, σ^2_x is the variance of x and $\sigma^2_{x^{\perp w}}$ is the variance of the part of x orthogonal to w. Thus,

$$\sigma^2_{\text{CTRL}} > \sigma^2_{\text{OVB}} \quad \Leftrightarrow \quad \frac{\sigma^2_{y^{\perp x, w}}}{n\ \sigma^2_{x^{\perp w}}} > \frac{\sigma^2_{y^{\perp x}}}{n\ \sigma^2_x} \quad \Leftrightarrow \quad \frac{\sigma^2_{y^{\perp x, w}}}{\sigma^2_{y^{\perp x}}} > \frac{\sigma^2_{x^{\perp w}}}{\sigma^2_x}$$

Controlling for w will increase the variance of the estimator if the fraction of the variance unexplained by w is greater for $y^{\perp x}$ than for x. Put differently, if controlling absorbs more of the variation in x than in the residual part of y ($y^{\perp x}$), it will increase the variance of the estimator. As briefly discussed above, since exaggeration increases with the variance of the estimator, controlling may increase exaggeration. I develop a formal proof showing that exaggeration can be larger when controlling, even when accounting for bias, in section **??**.

In the remainder of the paper, I first derive a formal proof of the existence of the trade-off for prevailing causal identification strategies. Specifically, I show that the bias caused by exaggeration can be larger than the one caused by confounders. I also analyze the drivers of exaggeration and show that it increases as the strength of the instruments decreases, the number of exogenous shocks decreases or when controlling for a confounder absorbs more of the variation in the treatment than in the outcome.

Then, I illustrate the existence of this "causal exaggeration" in realistic settings using examples drawn from environmental, education, labor, health and political economics. The exaggeration of statistically significant estimates can be defined as the ratio of the estimated effect over the true effect, which is never known in a real world setting. I order to know the true effect and be able to compute this quantity, I turn to simulations. In addition, Monte-Carlo simulations allow to vary the value of the parameter of interest *ceteris paribus*. An actual setting would for instance only allow to observe one strength for a given instrument. Since these simulations have an illustrative purpose only, I intentionally focus on settings in which statistical power can be low. All other simulation assumptions are chosen to make it as easy as possible to recover the effect of interest. I consider simple linear models with constant and homogenous treatment effects, *i.i.d.* observations and homoskedastic errors. All the models are correctly specified and accurately represent the data generating process, except for the omitted variable.

Finally, I discuss concrete avenues to address this causal exaggeration when carrying out a non-experimental study[1]. First, I advocate for the use of tools to evaluate the potential magnitude and risk of both confounding and exaggeration issues separately. Sensitivity analyses help with the former while power calculations help with the latter. For instance, the sensitivity analysis tools developed in Cinelli and Hazlett (2020) enable to assess how strong confounders would have to be to change the estimate of the treatment effect beyond a given level we are interested in. Then, considering the attention given to bias avoidance in the economics literature, I advocate to make power central to non-experimental analyses, even after an effect has been found, in order to limit bias caused by exaggeration. Prospective power simulations help identify the design parameters affecting power and exaggeration by approximating the data generating process (Gelman 2020, Black et al. 2021). Retrospective power calculations allow to evaluate whether a study would have enough power to confidently estimate a range of smaller but credible effect sizes (Gelman and Carlin 2014, Stommes et al. 2021). Focusing more specifically on the trade-off and its drivers, I present tools to visualize the variation actually used for identification when using causal identification strategies. The companion website describes in details how such analyses can be implemented. Finally, I briefly discuss potential solutions to mitigate this trade-off.

This paper contributes to three strands of the applied economics literature. First, the idea that causal identification estimators, while unbiased, may be imprecise is not new; this is part of the well-known bias-variance trade-off (Imbens and Kalyanaraman 2012, Deaton and Cartwright 2018, Hernán and Robins 2020, Ravallion 2020). I approach this literature from a different angle: through the prism of statistical power and publication bias. Not only the limited precision resulting from the use of causal identification methods could make it difficult to draw clear conclusions regarding the exact magnitude of the effect but I argue that it might also inherently lead to inflated published effect sizes, creating another "bias". The bias-variance trade-off can in fact be a bias-bias trade-off.

Second, recent studies discussing the exaggeration of statistically significant estimates due to low power focused on specific causal identification methods separately and usually do not investigate the determinants of this exaggeration (Schell et al. 2018, Black et al. 2021, Stommes et al. 2021, Young 2021). In a companion paper, I consider a wide range of empirical designs and highlight tangible design parameters that can cause exaggeration (**?**). In the present paper, I take a step back and propose an overarching mechanism, inherent to causal identification strategies as a whole, and that can explain these issues: although each strategy does so through different means, in essence they discard part of the variation, thereby increasing the risks of exaggeration.

Third, this study contributes to the literature on reproducibility in economics (Camerer et al. 2016, Ioannidis et al. 2017, Christensen and Miguel 2018, Kasy 2021). The trade-off presented in this paper may contribute to explaining the replication failures observed in empirical economics,

[1] In experimental studies, a solution to increase power is generally to increase sample size, reduce noise by improving measurement or improving balance or focus on larger potential effects.

despite the widespread use of convincing causal identification methods.

In the following section, I study the drivers of exaggeration and formally show in a simple setting that the use of causal identification strategies can exacerbate it. In section **??**, I implement realistic Monte-Carlo simulations to illustrate the existence of the confounding-exaggeration trade-off. I provide a series of recommendations to navigate this trade-off in section **??** and conclude in section **??**.

References

Abadie, A. (2020), 'Statistical Nonsignificance in Empirical Economics', American Economic Review: Insights **2**(2), 193–208. 2

Andrews, I. and Kasy, M. (2019), 'Identification of and Correction for Publication Bias', American Economic Review **109**(8), 2766–2794. 2

Black, B. S., Hollingsworth, A., Nunes, L. and Simon, K. I. (2021), Simulated Power Analyses for Observational Studies: An Application to the Affordable Care Act Medicaid Expansion, SSRN Scholarly Paper ID 3368187, Social Science Research Network, Rochester, NY. 6

Brodeur, A., Cook, N. and Heyes, A. (2020), 'Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics', American Economic Review **110**(11), 3634–3660. 2

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J. and Munafò, M. R. (2013), 'Power failure: Why small sample size undermines the reliability of neuroscience', Nature Reviews Neuroscience **14**(5), 365–376. 2

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M. and Wu, H. (2016), 'Evaluating replicability of laboratory experiments in economics', Science **351**(6280), 1433–1436. 2, 3, 6

Chang, A. C. and Li, P. (2022), 'Is Economics Research Replicable? Sixty Published Papers From Thirteen Journals Say "Often Not"', Critical Finance Review **11**. 2

Christensen, G. and Miguel, E. (2018), 'Transparency, Reproducibility, and the Credibility of Economics Research', Journal of Economic Literature **56**(3), 920–980. 6

Cinelli, C. and Hazlett, C. (2020), 'Making sense of sensitivity: Extending omitted variable bias', Journal of the Royal Statistical Society: Series B (Statistical Methodology) **82**(1), 39–67. 6

Deaton, A. and Cartwright, N. (2018), 'Understanding and misunderstanding randomized controlled trials', Social Science & Medicine **210**, 2–21. 6

Gelman, A. (2020), Regression and Other Stories, Cambridge University Press, Cambridge New York, NY Port Melbourne, VIC New Delhi Singapore. 6

Gelman, A. and Carlin, J. (2014), 'Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors', Perspectives on Psychological Science **9**(6), 641–651. 2, 6

Hernán, M. A. and Robins, J. M. (2020), Causal Inference: What If, boca raton: chapman & hall/crc edn. 6

Imbens, G. and Kalyanaraman, K. (2012), 'Optimal Bandwidth Choice for the Regression Discontinuity Estimator', The Review of Economic Studies **79**(3), 933–959. 6

Ioannidis, J. P. A. (2008), 'Why Most Discovered True Associations Are Inflated', Epidemiology **19**(5), 640–648. 2

Ioannidis, J. P. A., Stanley, T. D. and Doucouliagos, H. (2017), 'The Power of Bias in Economics Research', The Economic Journal **127**(605), F236–F265. 3, 6

Kasy, M. (2021), 'Of Forking Paths and Tied Hands: Selective Publication of Findings, and What Economists Should Do about It', Journal of Economic Perspectives **35**(3), 175–192. 6

Lu, J., Qiu, Y. and Deng, A. (2019), 'A note on Type S/M errors in hypothesis testing', British Journal of Mathematical and Statistical Psychology **72**(1), 1–17. 2

Open Science Collaboration (2015), 'Estimating the reproducibility of psychological science', Science **349**(6251), aac4716. 2

Ravallion, M. (2020), Should the Randomistas (Continue to) Rule?, Working Paper 27554, National Bureau of Economic Research. 6

Rosenthal, R. (1979), 'The file drawer problem and tolerance for null results', Psychological Bulletin **86**(3), 638–641. 2

Schell, T. L., Griffin, B. A. and Morral, A. R. (2018), Evaluating Methods to Estimate the Effect of State Laws on Firearm Deaths: A Simulation Study, Technical report, RAND Corporation. 6

Stommes, D., Aronow, P. M. and Sävje, F. (2021), 'On the reliability of published findings using the regression discontinuity design in political science', arXiv:2109.14526 [stat] . 6

Young, A. (2021), 'Leverage, Heteroskedasticity and Instrumental Variables in Practical Application', p. 43. 6

Zwet, E. W. and Cator, E. A. (2021), 'The significance filter, the winner's curse and the need to shrink', Statistica Neerlandica **75**(4), 437–452. 2