Causal Exaggeration: Unconfounded but Inflated Causal Estimates

Vincent Bagilet[*]

October 30, 2025

Abstract

The credibility revolution has made causal inference methods ubiquitous in economics, yet it has developed within a discipline that rewards statistically significant findings. I show that these two forces interact in ways that reduce the reliability of published estimates: while causal identification strategies alleviate bias from confounders, they reduce statistical power and can generate another type of bias—exaggeration—when combined with selection on significance. I characterize this confounding–exaggeration trade-off theoretically and via realistic Monte Carlo simulations, document its prevalence in the literature, and propose practical solutions, including a tool to identify the variation actually driving identification.

[*] ENS de Lyon, CERGIC, France. Email: vincent.bagilet@ens-lyon.fr. A previous version of this paper (*CEEP Working Paper Series*, 20) was co-authored with Léo Zabrocki-Hallak; I cannot thank him enough for his invaluable and far-reaching contributions to the project. I am very grateful to Jeffrey Shrader for his guidance and thank Sylvain Chabé-Ferret, Mathieu Couttenier, Clément De Chaisemartin, Jesse McDevitt-Irwin, Andrew Gelman, David McKenzie, José Luis Montiel Olea, Hélène Ollivier, Suresh Naidu, Claire Palandri, Julian Reif, Stephan Thies and Roberto Zuniga Valladares for helpful comments, as well as lab members at Columbia and seminars participants at Columbia, Ecole Normale Supérieure Lyon, EM Lyon, the Institute for Public Policy, IPWSD, the Paris School of Economics and the Toulouse School of Economics. I am very grateful to Alwyn Young for providing data from his analysis of the IV literature.

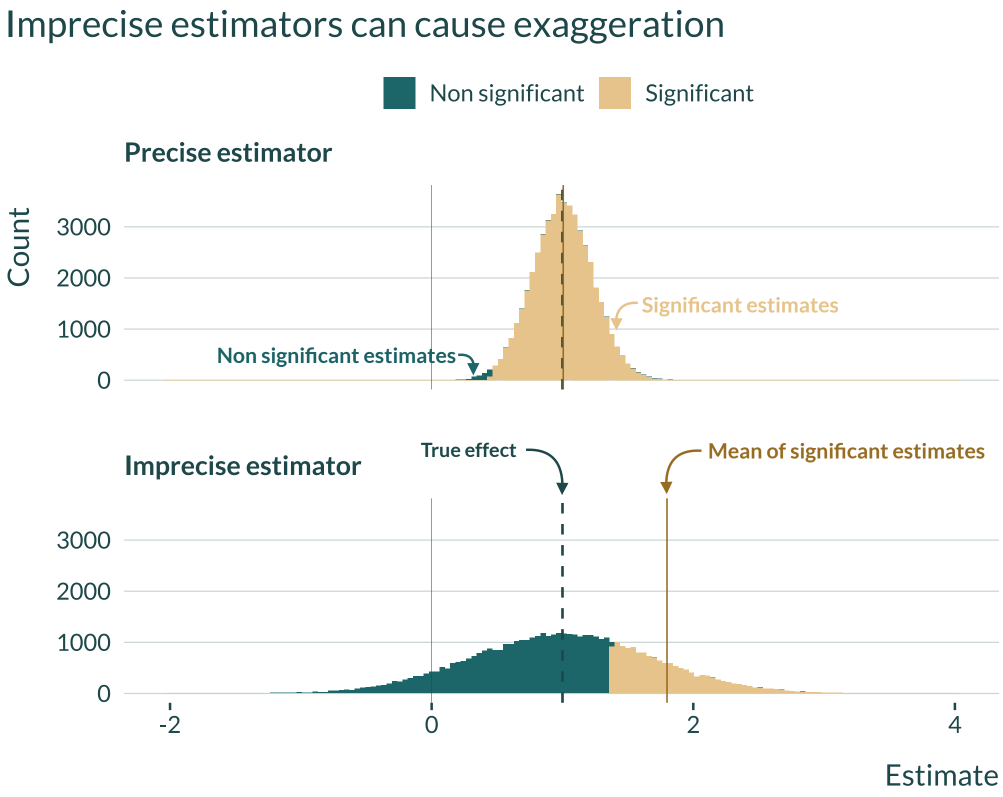
I Introduction

One of the main challenges of empirical economics is identifying causal effects. Identification 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 a portion 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. Limiting 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 (Gelman and Tuerlinckx 2000, Ioannidis 2008, Gelman and Carlin 2014). 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. They exaggerate the true effect and 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 guarantees that it will be far from 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. While identification strategies alleviate confounding bias, this paper argues that they can introduce another form of bias—exaggeration—by limitating the identifying variation.

Yet, exaggeration only arises under two conditions: 1) a publication bias favors a certain type of results and 2) statistical power is low. A large body of literature underlines that the economics literature selects results based on statistical significance (Rosenthal 1979, Brodeur et al. 2016, Andrews and Kasy 2019, Abadie 2020, Brodeur et al. 2020, for instance). Additional studies have highlighted its frequent and substantial lack of statistical power and resulting exaggeration (Ioannidis et al. 2017, Ferraro and Shukla 2020). Even in experimental economics, with a high level of control and an arguable absence of confounders, studies from top economics journals failed to replicate, the original estimates being on average inflated by a factor of at least 1.5 (Camerer et al. 2016). In the non-experimental economics literature, where statistical power is rarely a central consideration under current practices, several meta-analyses provide evidence of consequential

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 $\mathcal{N}(1, 0.5)$. Significance level: 5%

exaggeration. Ioannidis et al. (2017) estimates the median statistical power in a wide range of areas of economics to only 18%. Despite usually large sample sizes, they show that nearly 80% of estimates are likely exaggerated by a factor of two. As further illustrated in section II, the magnitude of exaggeration can be considerable and in some situations could be on par with that of a bias caused by confounders. Taking exaggeration into account and understanding its drivers is therefore crucial.

Accurate point estimates are instrumental as they often inform policy decisions through Cost-Benefit Analyses (CBA). For instance, environmental economics estimates enter the computation of the Social Cost of Carbon or routinely help policy makers decide of the implementation of regulations. Yet, the underlying effects can be relatively small and thus difficult to capture, making the studies subject to exaggeration. Estimates of the impact of environmental regulations on job losses constitute a good example (Gray et al. 2023). For instance, Walker (2011) documents the impact of the Clean Air Act amendments of 1990 on employment and finds an effect of -14.2% (s.e. 4.3). For similar policies, other studies find smaller effects, of the order of magnitude of -3% (Greenstone 2002, Gray et al. 2023). The design of Walker (2011) would not be precise enough to retrieve an effect size of this magnitude. If the true effect was in fact of this magnitude, the statistical power of the study would be 11% and significant estimates would exaggerate the true effect by a factor of 3.5 on average. In this example, bias caused by exaggeration would be substantial and could have detrimental policy implications.[1]

In the present paper, I argue that the use of causal identification strategies can cause exaggeration by limiting the variation used for identification. Reviewing the literature, 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. To limit the threat of confounding, causal inference methods discard variation and can therefore reduce statistical power. Combined with the publication filter based on significance this results in exaggeration bias. While causal identification strategies are essential to describe causal relationships, the present paper emphasizes that a perfectly convincing identification does not guarantee an absence of "bias" and that improving identification can actually pull estimates away from the true effect. The same strategies which remove bias caused by confounding actually introduce another type of bias.

All causal identification strategies discard variation in order to identify causal effects but the confounding-exaggeration trade-off is mediated through a distinctive channel for each of them. RD designs discard part of the variation by only considering observations within the bandwidth, decreasing the effective sample size and thus precision. An IV strategy only uses 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 treated observations. 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 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.[2] When these identification strategies absorb more of the variation in the treatment than in the outcome, they increase the

[1] Of course, in the particular setting of Walker (2011), the true effect might be larger than 3%. I am not claiming that the study is flawed but instead that its level of precision would produce inflated significant estimates if the true effect was of the order of magnitude of 3%.

[2] Fixed Effects (FEs) based identification strategies such as DiD control for the invariant, unobserved, and arguably endogenous part of the variation in the outcome. An IV approach essentially partials out the variation in x unexplained by the instruments. Fuzzy-RD and propensity score matching can 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.

variance of the resulting estimator and can 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 correlated with x and u, an error term. Under usual assumptions and using the Frisch-Waugh-Lovell theorem, we get that $\mathbb{V}_{\hat{\beta}_{\text{OVB}}}$ and $\mathbb{V}_{\hat{\beta}_{\text{CTRL}}}$, the variance of the estimators for β when omitting w (short regression) and controlling for it (long regression) are respectively:

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

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 or equivalently the variances of the parts of y that are orthogonal to x and to x and w respectively ($\sigma^2_{y^{\perp x}}$ and $\sigma^2_{y^{\perp x,w}}$), 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,

$$\mathbb{V}_{\hat{\beta}_{ovb}} < \mathbb{V}_{\hat{\beta}_{ctrl}} \quad \Leftrightarrow \quad \frac{\sigma^2_{y^{\perp x}}}{n\ \sigma^2_x} < \frac{\sigma^2_{y^{\perp x,w}}}{n\ \sigma^2_{x^{\perp w}}} \quad \Leftrightarrow \quad \frac{\sigma^2_{x^{\perp w}}}{\sigma^2_x} < \frac{\sigma^2_{y^{\perp x,w}}}{\sigma^2_{y^{\perp x}}}$$

Controlling for w increases 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 ($y^{\perp x}$), the variance of the estimator increases. Since exaggeration increases with estimator's variance, controlling for a confounder can increase exaggeration. Controls, including fixed effects, limit the identifying variation and may cause identification to rely on a small effective sample. Observations whose treatment status is largely explained by these controls contribute little, if at all, to identification. Under some circumstances discussed in sections III and IV, controlling can even produce an exaggeration bias larger than the OVB that would result from an absence of controls.

In the remainder of the paper, I first document the magnitude of the trade-off. To do so, I build on existing literature reviews (Brodeur et al. 2020, Young 2022, Bagilet 2023, Lal et al. 2024) to discuss evidence of exaggeration bias in a large set of causal studies mostly published in top journals. These analyses reveal heterogeneity across studies: while exaggeration might be limited in some analyses, it is likely substantial in many others. In the set of papers reviewed in Young (2022), I show that half of the IV designs would exaggerate a true effect size of the magnitude of the "naive" OLS estimate by a factor larger than 3.2—and by a factor larger than 2.0 for headline designs. I then directly compare confounding and exaggeration biases for an example study. I show that in this instance the bias of the IV could be on par or even larger than that of the OLS and that exaggeration may explain this difference.

Next, I 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 demonstrate 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 show that this "causal exaggeration" arises for realistic parameter values by further exploring its drivers in realistic settings in which there is no closed form formula for power and exaggeration available. Exaggeration being defined with respect to the true effect, a quantity which is never known in real world settings, I turn to simulations to be able to compute this quantity. Monte-Carlo simulations also allow varying the value of the parameter of interest *ceteris paribus*, something that would not be possible otherwise. To make these simulations concrete, I calibrate them to emulate existing studies from education, environmental, health, labor, and political economics. I find that causal exaggeration can be substantial in realistic settings where the variation remaining for identification is limited.

Finally, I discuss concrete avenues to address causal exaggeration when carrying out a non-experimental study[3]. A series of tools can be used to evaluate the potential magnitude of confounding and exaggeration issues separately. Sensitivity analyses help with the former while power calculations help with the latter. Considering the attention given to bias avoidance in the economics literature, I underline that making power central to non-experimental analyses, even after an effect has been found, would help limiting 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. 2022). Retrospective power calculations allow evaluating whether a study would have had enough power to confidently estimate a range of smaller but credible effect sizes (Gelman and Carlin 2014, Stommes et al. 2023). Focusing more specifically on the trade-off and its drivers, I consider and develop tools to identify the variation actually used for identification when using causal identification strategies. I introduce a `R` package, `ididvar`, providing a set of tools to easily identify the identifying variation in a regression and the set of observations actually contributing to the estimation of the effect of interest.

This paper contributes to four strands of the applied economics literature. First, the idea that causal identification estimators, while unbiased, may be imprecise is not new; it 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. Limited precision resulting from

[3] In experimental studies, there are essentially no confoundings. A solution to increase power and reduce exaggeration is generally to increase sample size, reduce noise by improving measurement or improving balance or to focus on larger potential effects.

the use of causal identification strategies not only makes it difficult to draw clear conclusions about the exact magnitude of effects, but, I argue, can also inherently lead to inflated published estimates, creating another form of "bias." In this sense, the bias–variance trade-off can in practice become a bias–bias trade-off.

Second, studies discussing the exaggeration of statistically significant estimates due to a lack of power usually do not investigate its determinants or focus on specific causal identification strategies separately (Ioannidis et al. 2017, Schell et al. 2018, Ferraro and Shukla 2020, Black et al. 2022, Stommes et al. 2023, Arel-Bundock et al. 2022). In a companion paper, I highlight tangible design parameters that can cause exaggeration in a wide range of designs (Bagilet 2023). In the present paper, I take a step back and propose an overarching mechanism inherent to causal identification strategies as a whole: although each strategy does so through different means, in essence they discard part of the variation, thereby increasing the risks of exaggeration.

Third, this paper makes a methodological and practical contribution by facilitating the computation of concrete measures of each observation's contribution to identification and, by extension, of the effective sample underlying the estimated effect. It extends and complement existing approaches (Angrist and Pischke 2009, Aronow and Samii 2016) and facilitates their implementation and interpretability in applied economics research. The accompanying `R` package enables users to compute, visualize, and analyze these measures in a straightforward and consistent way. It allows applied researchers to diagnose potential exaggeration risks ex post, without assumptions about true effect sizes. Beyond the scope of this paper, the package also provides a framework for identifying and communicating the set of observations contributing to identification and how this set changes across specifications, levels of fixed effects and set of control variables.

Finally, this study contributes to the literature on replicability in economics (Camerer et al. 2016, Ioannidis et al. 2017, Christensen and Miguel 2018, Kasy 2021). The trade-off presented in this paper suggests that the widespread reliance on convincing causal identification strategies in economics does not necessarily shield the field from potential replication threats. In addition, the tools discussed and introduced can be readily used in replication analyses or referee reports to transparently document statistical power and the sources of identifying variation and can help assess robustness across different specifications, subsamples, or sets of controls.

In the following section, I document evidence of causal exaggeration in the causal economics literature. In section III, 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 IV, I implement realistic Monte-Carlo simulations to illustrate the existence of the confounding-exaggeration trade-off. I discuss potential solutions to navigate this trade-off and introduce the `ididvar` package in section V and conclude in section VI.

II Causal Exaggeration in the Literature

The trade-off presented in this paper has meaningful implications only if causal identification strategies lead to substantial exaggeration, especially as compared to the amount of confounding bias they allow avoiding. This section aims to documents the existence of such exaggeration in the causal economics literature.

A first piece of evidence emerges from the observation that causal methods routinely yield larger estimates than association-based approaches. The data in Young (2022) reveals that among 30 reproducible papers published in journals from the *American Economic Association*[4], the median ratio of headline IV estimates over the corresponding OLS is 2.3, with 25% of estimates exceeding a ratio of 5.4. A comparable pattern appears in top political science journals (Lal et al. 2024). In the literature on the acute health effects of air pollution, which I explore in a companion paper, causal estimates are substantially larger than what would have been predicted by the standard epidemiology literature, estimates being regularly more than 10 times larger (Bagilet 2023). Even within a given setting and study, the median of the ratio of the obtained 2SLS to their corresponding "naive" OLS estimates is 3.8.

What can explain that causal methods yield such large effects sizes, as compared to non-causal methods? They could arguably remove omitted variable bias, reduce attenuation bias caused by classical measurement error or target a different causal estimand. But as I argue in this section, exaggeration and imprecision could also explain part of this difference. For instance, published IV papers typically have low power and make most 2SLS estimates statistically indistinguishable from the corresponding OLS (Young 2022).

II. 1 Quantifying exaggeration

As discussed in the introduction, lack of power is widespread and exaggeration substantial in the economics literature; Ioannidis et al. (2017) estimates that nearly 80% of results published in a wide array of empirical economic literatures are likely exaggerated, typically by a factor of two and one-third by a factor of four or more.

To document exaggeration in the causal inference literature more specifically, and to compute exaggeration in general, one needs to hypothesize true effect sizes. I begin by examining broad literatures that in return require strong assumptions about true effect sizes, then progressively narrow my focus to more specific literatures that allow these assumptions to be relaxed. I first leverage data from Brodeur et al. (2020). This paper reviews of the universe of hypothesis tests

[4] The exact selection process is described in section 3 of Young (2022). The sample consists of papers using the keyword "instrument", published up through July 2016, that are reproducible, relying on open access data and Stata code, using linear methods and standard covariance estimates.

reported in papers published in the top-25 economics journals in 2015 and 2018 and using RCT, DID, RDD or IV. It shows that IV and to a lesser extent DID are particularly subject to publication bias—one of the two ingredients of exaggeration. Since the exaggeration ratio is the expected value of the absolute value of significant estimates over the true effect, it can only be calculated by hypothesizing true effect sizes. To circumvent this limitation, I first evaluate the proportion of studies in Brodeur et al. (2020) that would have a design reliable enough to retrieve an effect size equal to half of the obtained estimate. I also compute the exaggeration of significant estimates under this hypothesis. There is no *a priori* reason to believe that the magnitude of the true effect of a specific estimation would be equal to half its point estimate. However, since Ioannidis et al. (2017) finds a typical exaggeration of two in the economics literature, we may expect this assumption to be reasonable, on average. Regardless, I do not claim that the true effect is equal to this hypothesized value but rather wonder what would be the power and exaggeration under this justifiable assumption. This approach is also to some extent conservative: hypothesized effect sizes based on exaggerated estimates will be larger and will thus minimize exaggeration.

The exaggeration is computed by drawing a large number of times from a normal distribution centred on the hypothetical effect size and with a standard deviation equal to the standard error of the estimate found in the study and by then computing the average of the draws that are 1.96 standard errors away from 0. I further discuss how to compute such power calculations in section V. 2.2. Under this assumption regarding the effect size, the median power and median exaggeration ratio of significant estimates in the data from Brodeur et al. (2020) would be 37% and 1.6 respectively. Statistical power would be larger than the usual 80% threshold for only 22% of designs. 28% of the designs would, on average, exaggerate the true effect sizes by more than a factor of 2. While exaggeration is reassuringly limited in some settings, it can be substantial in many others. These results are relatively comparable across methods (IV, DID, RDD and RCT).

While the previous hypothesis regarding true effect sizes provides a broad overview of the literature, it may seem somewhat arbitrary. To reduce the burden of assumptions, I narrow the focus to IV designs, which enable a more systematic intra-study comparison between the 2SLS estimates and the corresponding "naive" OLS one. I do not claim that the OLS estimate accurately represents the true effect but it may be reasonable to expect the IV to have enough power to retrieve an effect of the magnitude of the OLS point estimate, if it was indeed the true effect. To explore this question, I leverage the data from Young (2022). IV designs that produced significant headline results would only have a median power of 25% to detect an effect size of the magnitude of the OLS estimate. Significant estimates from half of these designs would exaggerate the OLS estimate by a factor larger than 2 and a quarter would do so by a factor larger than 3.5. Only 17% of these designs would reach the conventional 80% power threshold. These figures further underline heterogeneity in vulnerability to exaggeration and power issues. While many analyses are likely immune to exaggeration, a substantial share of them would not be powered enough to detect effects of realistic magnitude leading significant estimates to considerably exaggerate these effects.[5] These results complement the analysis in Young (2022) that highlights the imprecision of published 2SLS estimates, making them statistically indistinguishable from the corresponding OLS estimate, despite the fact that 2SLS point estimates substantially differ in magnitude from their OLS counterpart and are even regularly of the opposite sign.

II. 2 Illustration of the trade-off

In order to limit hypotheses and to investigate causal exaggeration further, I focus on an example IV study investigating the impact of PM2.5 air pollution on mortality (He et al. 2020). It allows comparing the bias of the "naive" OLS to that of the IV by computing their distance to an estimate of the "true effect" they target. I define this "true effect" in two ways. First, I use the results of a meta-analysis of epidemiological studies (Shah et al. 2015). By pooling a number of studies carried out in various contexts, this meta-estimate might represent the average effect one may expect from such a study. However, the studies in this meta-analysis do not rely on canonical causal identification strategies used in economics and may be thought of as suffering from counfounding. I thus consider the result of Deryugina et al. (2019)—a precise causal study that may be less exposed to exaggeration—as an alternative estimate of the "true effect". This estimate may be context specific and the "true effect" in He et al. (2020) may deviate from these. The present discussion is conditional on this true underlying effect being close to these hypothesized true effect sizes.

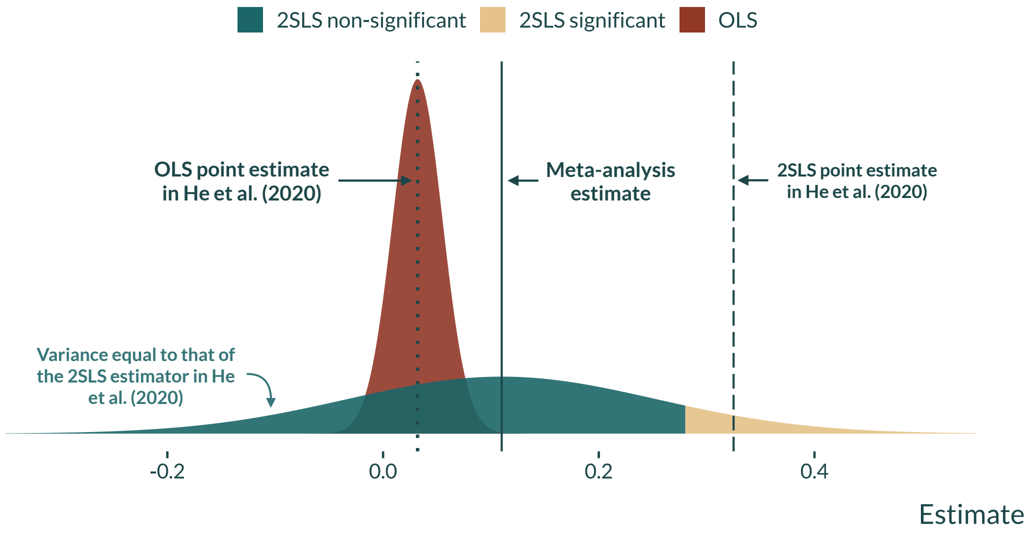
He et al. (2020) finds that a "$10 \mu g.m^{-3}$ increase in PM2.5 increases mortality by 3.25%" (s.e. 1.43%). Their corresponding OLS results suggest a 0.32% increase (s.e. 0.23%). For a similar increment in air pollution, Shah et al. (2015) and Deryugina et al. (2019) document a 1.1% and 1.8% increase in mortality respectively. The OLS estimate in He et al. (2020) is closer to the "true effect" based on Shah et al. (2015) than their 2SLS estimate. Provided that the three estimands are comparable, the bias of the IV is larger than that of the OLS. If the true effect was in fact closer to the one found by Deryugina et al. (2019), both biases would be roughly equal and the bias of the IV still substantial.

Exaggeration could explain this difference. Even if the 2SLS estimator effectively removes all conventional biases, the design in He et al. (2020) would still yield exaggerated statistically

[5] I find similar results in the literature on acute health effects of air pollution and in the political science literature, leveraging reviews implemented in Bagilet (2023) and Lal et al. (2024) respectively. The experimental literature displays less exaggeration on average but still substantial heterogeneity Camerer et al. (2016). These analyses are described on the project's website.

significant estimates. Figure 2 illustrates this point. Even if the 2SLS estimator is unbiased, i.e., centered on the meta-estimate found in Shah et al. (2015), the lack of precision of this design would lead significant estimates to substantially exaggerate the true effect, by a factor 3.2 on average. The 2SLS estimate found in He et al. (2020) could be one of these estimates.

Figure 2 – Illustration of the Confounding-Exaggeration Trade-off in He et al. (2020)



Notes: The distribution for the 2SLS estimator is centered on the true effect, represented by the solid line and defined as the meta-estimate found in Shah et al. (2015). Its variance is equal to the one of the 2SLS estimator in He et al. (2020). The distribution for the OLS estimator is centered on the OLS estimate found in He et al. (2020) and its variance equal to that of this same estimator. The dashed and dotted lines represent the 2SLS and OLS estimates found in He et al. (2020) respectively. I ignore exaggeration of the OLS for clarity but since the OLS is biased downward, inflating OLS estimates bring them closer to the true effect.

This example illustrates that in a published study where a causal identification strategy substantially reduces the precision of the estimator, the resulting statistically significant estimates may be further away from the true effect than the "naive" OLS estimate, even if the estimator is unbiased. Note that a comparable result holds if the true effect is equal to the one found in Deryugina et al. (2019). With this design, the average exaggeration would be 3.1 times larger than the OVB (or 1.3 time if the true effect is closer to the one found in Deryugina et al. (2019)).

III Mathematical derivation

In this section, I formally prove the existence of the confounding-exaggeration trade-off and describe its drivers in a simple setting. To do so, I first define an exaggeration ratio and show that it increases with the variance of normally distributed biased estimators. I then compute the asymptotic distributions of a series of estimators and study drivers of their variances, and ultimately of their exaggeration ratios. Finally, I show that, for any magnitude of OVB, exaggeration can be greater when using a causal inference method than the overall bias combining exaggeration and OVB in the naive regression.

III. 1 Properties of the exaggeration ratio

Following Gelman and Carlin (2014), we can define the exaggeration ratio E, as the expectation of the absolute value of significant estimates over the absolute value of the true effect. For an estimator $\hat{\beta}$ of a true effect β, with standard error σ and a two-sided hypothesis test of size α with threshold value z_α, let

$$E(\hat{\beta}, \sigma, \beta, z_\alpha) = \frac{\mathbb{E}\left[|\hat{\beta}| \;\middle|\beta, \sigma, |\hat{\beta}| > z_\alpha \sigma\right]}{|\beta|} \tag{1}$$

Lu et al. (2019) and Zwet and Cator (2021) showed that, for given test and true effect sizes, the exaggeration ratio increases with the variance of an unbiased normally distributed estimator. We can extend this proof to biased estimators and get that: [6]

Lemma 1. *For an estimator $\hat{\beta}_b \sim \mathcal{N}(\beta + b, \sigma^2)$ of a true effect of magnitude β and a fixed bias b of the same sign as and independent from the true effect,*

- — *E is a decreasing function of the Signal-to-Noise Ratio (SNR) $\frac{\beta}{\sigma}$, ie the relative precision of the estimator, and only depends on σ through this SNR.*
- — $\lim_{\sigma \to \infty} E(\hat{\beta}_b, \sigma, \beta, z_\alpha) = +\infty$.

Figure 1 provides a clear intuition for these results in the unbiased case. Note that here, we conservatively focus on cases in which the bias is in the same direction as the true effect so that exaggeration from causal inference methods and OVB do not cancel each other.

Based on lemma 1, one only needs to show asymptotic normality and study how the variances of these estimators evolve with these parameters to study how exaggeration evolves with the IV strength in an IV setting, the number of exogenous shocks in a reduced form and the correlation between the explanatory variable of interest and the omitted variable of interest. This relies on the assumption that the sample size is large enough so that the sample distribution of the estimator is well approximated by their asymptotic distribution.

[6] All the proofs of the lemma and theorems are reported in appendix A.

III. 2 Setting and data generating process

Consider a usual linear homoskedastic regression model with an omitted variable. For any individual $i \in \{1, ..., n\}$, we write:

$$y_i = \beta_0 + \beta_1 x_i + \delta w_i + u_i \tag{2}$$

where y is the outcome, x the explanatory variable, w an unobserved omitted variable, u an unobserved error term. $(\beta_0, \beta_1, \delta) \in \mathbb{R}^3$ are unknown parameters. β_1 is the parameter of interest.

Assume homogeneous treatment effects and homoskedasticity, along with the usual OLS assumptions (*i.i.d.* observations, finite second moments, positive-definiteness of $\mathbb{E}[\mathrm{x}_i \mathrm{x}_i']$—with $\mathrm{x}_i = (1, x_i)'$— and u_i conditional mean-zero and uncorrelated with x_i and w_i). Assume that w_i is unobserved, correlated with x_i and that $\delta \neq 0$. To simplify the derivations, I further assume that the unobserved variable is centered, *i.e.* $\mathbb{E}[w_i] = 0$. I also assume that the variance of the component of w_i that is orthogonal to x_i (denoted $w_i^{\perp x}$) does not vary with x_i, *i.e.*, $\text{Var}(w_i^{\perp x}|x_i) = \text{Var}(w_i^{\perp x})$. Consider the following data generating process for x_i:

$$x_i = \mu_x + \gamma w_i + \epsilon_i \tag{3}$$

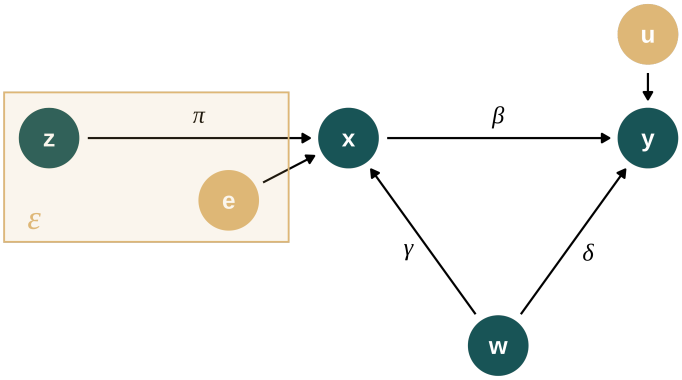
where $\gamma \in \mathbb{R}^*$ since x and w are correlated. Set $\rho_{xw} = \text{corr}(x, w) = \frac{\gamma \sigma_w}{\sigma_x}$. In the IV and reduced form sections, I further assume that there exists a valid instrumental variable z_i for x_i, *i.e.* that $\mu_x + \epsilon_i = \pi_0 + \pi_1 z_i + e_i$ where $(\pi_0, \pi_1) \in \mathbb{R}^2$ are unknown parameters. The existence or not of this valid instrument does not affect the results in the controlled and OVB cases. Since the instrument is valid, it satisfies exogeneity, *i.e.* $\mathbb{E}[\mathrm{z}_i u_i] = 0$ and $\mathbb{E}[z_i w_i] = 0$, relevance, *i.e.* $\text{rank}(\mathbb{E}[\mathrm{z}_i \mathrm{x}_i']) = 2$, and positive-definiteness of $\mathbb{E}[\mathrm{z}_i \mathrm{z}_i']$. The data generating process for x_i becomes:

$$x_i = \pi_0 + \pi_1 z_i + \gamma w_i + e_i \tag{4}$$

I assume that e_i is uncorrelated with z_i and w_i, *i.e.* $\mathbb{E}[z_i e_i] = 0$ and $\mathbb{E}[w_i e_i] = 0$. I also assume homoskedasticity for this term, such that $\mathbb{E}[e_i^2 | z_i, w_i] = \sigma_e^2$ is constant.

Overall, this DGP is close to the usual textbook one but with an additional omitted variable. The Directed Acyclic Graph (DAG) in figure 3 represents the data generating process.

Figure 3 – DAG of the data generating process



Notes: for clarity the error terms are represented in this graph, in beige. Model parameters are noted as edge labels.

III. 3 Asymptotic distributions of the estimators

I now derive the asymptotic distributions of the various estimators. For each model, the goal is to show asymptotic normality and to study the evolution of the sampling distribution variances with the value of the parameter of interest, *i.e.*, a measure of the correlation between x and w (γ) in the controlled case, of the IV strength (π_1) in the IV case and of the number of exogenous shocks (σ_z^2 when z is a dummy) in the reduced form case. I assume that the sampling distributions are well approximated by the asymptotic distributions. In order for the variation of one factor not to impact other factors of interest, I consider the variances of the variables ($\sigma_y^2, \sigma_x^2, \sigma_w^2$ and σ_z^2) as fixed but adjust for the variances of the error terms (σ_u^2 and σ_ϵ^2) when varying the values of one of the parameters (γ, δ and π_1). This corresponds to thinking in terms of shares of the variance of x and y explained by "defined" variables (*i.e.*, observed variables and w) *versus* by residuals. Finally note that comparison between cases with and without OVB for different parameter values is only relevant if varying the parameter of interest does not affect the OVB. I thus make comparative statics analyses at bias fixed, *i.e.*, as shown below, for $\gamma\delta = \kappa = cst$.

III. 3.1 Naive regression (OVB)

First, let us study the benchmark against which we are going to compare our causal approaches. Consider the "naive" regression of y on x (with w omitted).

Lemma 2. *Based on the data generating process described in section III. 2, for $\hat{\beta}_{\text{OVB}}$ the OLS estimate of β_1 in the regression of y on x, $\hat{\beta}_{\text{OVB}} \xrightarrow{d} \mathcal{N}(\beta_1 + b_{\text{OVB}},\ \sigma^2_{\text{OVB}})$, with*

$$b_{\text{OVB}} = \frac{\delta\gamma\sigma_w^2}{\sigma_x^2} \qquad \textit{and} \qquad \sigma^2_{\text{OVB}} = \frac{\sigma_u^2 + \delta^2\sigma_w^2(1-\rho_{xw}^2)}{n\ \sigma_x^2}$$

The intuition for the formula of the asymptotic variance has been discussed in the introduction: $\sigma_u^2 + \delta^2\sigma_w(1-\rho_{xw}^2)$ is the part of the variance in y that is not explained by x ($\sigma_{y^{\perp x}}^2$).

Varying the parameter of interest, ρ_{xw}, will change the bias and σ_u^2. Since σ_x^2 and σ_w^2 are fixed, reasoning at $b_{\text{OVB}} = cst$ is equivalent to considering that $\gamma\delta = \kappa = const$. Then, noting that $\forall i, u_i = y_i - \beta_0 - \beta_1 x_i - \delta w_i$ and computing its variance, we can rewrite the variance of the estimator as a function of fixed variances and one or less varying parameter:

$$\sigma_{\text{OVB}}^2 = \frac{\sigma_y^2 - \beta_1^2\sigma_x^2 - 2\beta_1\kappa\sigma_w^2 - \kappa^2\frac{\sigma_w^4}{\sigma_x^2}}{n\ \sigma_x^2}$$

This expression underlines that, for a given bias, σ_{OVB}^2 does not vary with γ, or equivalently δ, the parameters of interest. Applying lemma 1 proves that E_{OVB} does not either.

III. 3.2 Controlled regression

Next, let us turn to the "ideal" case in which no variable is omitted, *i.e.* we control for the omitted variable w and thus partial out confounders. The model considered accurately represents the DGP. This corresponds to the usual OLS setting with a constant and two regressors that are uncorrelated with the error: y regressed on x and w.

Lemma 3. *Based on the data generating process mentioned previously, for $\hat{\beta}_{\text{CTRL}}$ the OLS estimator of β_1 in the regression of y on x and w, $\hat{\beta}_{\text{CTRL}} \xrightarrow{d} \mathcal{N}(\beta_1,\ \sigma_{\text{CTRL}}^2)$, with*

$$\sigma_{\text{CTRL}}^2 = \frac{\sigma_u^2}{n\ \sigma_x^2(1-\rho_{xw}^2)}$$

$\sigma_x^2(1-\rho_{xw}^2)$ is the part of the variance of x that is not explained by w ($\sigma_{x^{\perp w}}^2$) and σ_u^2 the part of the variance of y that is not explained by x nor w ($\sigma_{y^{\perp x,w}}^2$); here too we retrieved a result described in introduction. For a given bias, we then rewrite σ_{CTRL}^2 as a function of fixed variances and one varying parameter, γ:

$$\sigma_{\text{CTRL}}^2 = \frac{\sigma_y^2 - \beta_1^2\sigma_x^2 - \frac{\kappa^2}{\gamma^2}\sigma_w^2 - 2\beta_1\kappa\sigma_w^2}{n\ (\sigma_x^2 - \gamma^2\sigma_w^2)}$$

Since the numerator and denominator respectively increase and decrease with $|\gamma|$, σ_{CTRL}^2 increases with $|\gamma|$. For a given bias, the more w is correlated with x (and thus roughly the less it is with y since $\delta\gamma = const$), the larger the variance of the estimator. In addition, we can note that, for a given bias, the variance of the estimator can be arbitrarily large since $\lim_{\gamma^2 \to \frac{\sigma_x^2}{\sigma_w^2}} \sigma_{\text{CTRL}}^2 = +\infty$.

III. 3.3 Instrumental Variables

In the previous section, we considered a case in which we removed variation that included unwanted endogenous variation. We now turn to the IV, a converse situation where we select variation we want, exogenous variation. We estimate the IV model in which we regress y on $\mathrm{x}_i = (1, x_i)'$ instrumented by $\mathrm{z}_i = (1, z_i)'$. We are thus in a just-identified case and $\hat{\boldsymbol{\beta}}_{2\text{SLS}} = \hat{\boldsymbol{\beta}}_{\text{IV}}$.

Lemma 4. *Based on the data generating process mentioned above, for $\hat{\beta}_{\text{IV}}$ the IV estimator of β_1 in the regression of y on x instrumented by z, $\hat{\beta}_{\text{IV}} \xrightarrow{d} \mathcal{N}(\beta_1,\ \sigma^2_{\text{IV}})$, with*

$$\sigma^2_{\text{IV}} = \frac{\sigma_u^2 + \delta^2\sigma_w^2}{n\ \sigma_x^2\rho_{xz}^2}$$

The numerator is $\sigma^2_{y\perp\hat{x}}$, the part of the variance in y that is not explained by $\hat{x}$, the predicted value of x in the first stage and the denominator is $\sigma^2_{\hat{x}}$. For a given bias, noting that $\rho_{xz} = \text{corr}(x, z) = \pi_1\frac{\sigma_z}{\sigma_x}$ and replacing σ_u^2, we can rewrite σ^2_{IV} as a function of fixed variances and one varying parameter, π_1:

$$\sigma^2_{\text{IV}} = \frac{\sigma_y^2 - \beta_1^2\sigma_x^2 - 2\beta_1\kappa\sigma_w^2}{n\ \pi_1^2\sigma_z^2}$$

Clearly, the smaller π_1, the larger σ^2_{IV}. In addition, $\lim_{\pi_1\to 0}\sigma^2_{\text{IV}} = +\infty$.

III. 3.4 Reduced form

Let us now assume that we want to directly estimate the effect of the instrument on the outcome of interest. Plugging equation 4 into equation 2 yields:

$$y_i = (\beta_0 + \beta_1\pi_0) + (\beta_1\pi_1)z_i + ((\delta + \beta_1\gamma)w_i + u_i + \beta_1 e_i)$$

Note that if we directly regress the outcome on the instrument, the resulting estimand will be different from that of the other models. To make them comparable, we could set π_1 to 1 so that an increase of 1 in the instrument causes an increase of β_1 in y. Regardless of whether we make this assumption or not, regressing y on z corresponds to the usual univariate, unbiased case and directly gives the following result:

Lemma 5. *Based on the data generating process mentioned previously, for $\hat{\beta}_{\text{RED}}$, the OLS estimator of the reduced form regression of y on z, $\hat{\beta}_{\text{RED}} \xrightarrow{d} \mathcal{N}(\beta_1,\ \sigma^2_{\text{RED}})$, with*

$$\sigma^2_{\text{RED}} = \frac{\sigma_y^2 - \beta_1^2\pi_1^2\sigma_z^2}{n\ \sigma_z^2}$$

The numerator is the part of the variance of y that is not explained by z ($\sigma^2_{y^{\perp z}}$). In addition, it is clear that the smaller σ_z^2, the larger σ^2_{RED} and $\lim_{\sigma_z \to 0} \sigma^2_{\text{RED}} = +\infty$.

In the binary case, $\sigma_z^2 = p_1(1-p_1)$ with p_1 the proportion of treated observations, *i.e.*, the proportion of 1 in z. When most observations have the same treatment status, *i.e.*, p_1 close to 0 or 1, σ_z^2 tends to zero and σ^2_{RED} shoots up. There is not enough variation in the treatment status to precisely identify the effect of interest.

III. 4 Exaggeration ratios

Combining the results from lemma 2 through 5 regarding the asymptotic distribution of the various estimators with lemma 1 stating that exaggeration increases with the variance of a normally distributed estimator yields:

Theorem 1. *For the data generating process described in section III. 2, the exaggeration ratio of the controlled, IV and reduced form estimators, respectively $E_{\text{CTRL}}, E_{\text{IV}}$ and E_{RED}, are such that:*

— *E_{CTRL} increases with the correlation between the omitted variable and the explanatory variable of interest (i.e. $|\gamma|$ or $|\rho_{xw}|$), for a given bias,*

— *E_{IV} decreases with the strength of the IV (i.e. with $|\pi_1|$ or $|\rho_{xz}|$),*

— *E_{RED} increases when the number of exogenous shocks decreases in the binary case*

Also using the same lemma and the limit properties of the variances described in section III. 2, and since, at fixed bias, E_{OVB} does not vary with the parameters of interest, we get:

Theorem 2. *For the data generating process described in section III. 2, $\forall\ b_{\text{OVB}}$,*

— $\exists\ \gamma$ *s.t.* $E_{\text{CTRL}} > E_{\text{OVB}}$

— $\exists\ \pi_1$ *s.t.* $E_{\text{IV}} > E_{\text{OVB}}$

— $\exists\ \sigma_z$ *s.t.* $E_{\text{RED}} > E_{\text{OVB}}$

For some parameter values and a given underlying true effect, statistically significant estimates can be larger on average when using a convincing causal identification strategy that eliminates the omitted variable bias than when embracing the bias and running a naive biased regression.

IV Simulations

To study the drivers of exaggeration in concrete settings, I build simulations that reproduce real-world examples from economics of education for RDD, political economy for IV, health economics for exogenous shocks and environmental and labor economics for control and fixed effects approaches. I split the simulations by identification strategy. Real-world settings enable clearly grasping the relationships between the different variables, setting realistic parameter values, based on existing studies and showing that exaggeration can arise for parameter values consistent with existing studies. To underline that causal exaggeration is not bound to specific studies, I do not reproduce a particular study but instead calibrate my simulations to emulate a typical non-experimental study from each literature. To check the representativity of my simulations, I compare their Signal-to-Noise Ratios (SNR) to the estimate/standard error ratios in studies from the corresponding literature. As underlined by lemma 1, the SNR is a sufficient statistic for the exaggeration ratio; a SNR consistent with the literature will ensure the representativity of the simulations with respect to the feature of interest, exaggeration. I also make many conservative simulation choices that ease the recovery of the effect of interest and limit estimation challenges. 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, thus taking identification as given. Extensively documented code for each simulation procedure is available on the project's website.

IV. 1 Regression Discontinuity Design

Intuition. A RDD relies on the assumption that for values close to the threshold, treatment assignment is quasi-random. It focuses on observations within a certain bandwidth around this threshold and discards observations further away. The effective sample used for causal identification is thus smaller than the total sample. A smaller bandwidth and effective sample size reduce precision and can create exaggeration. Here, the confounding-exaggeration trade-off is mediated by the size of the bandwidth.

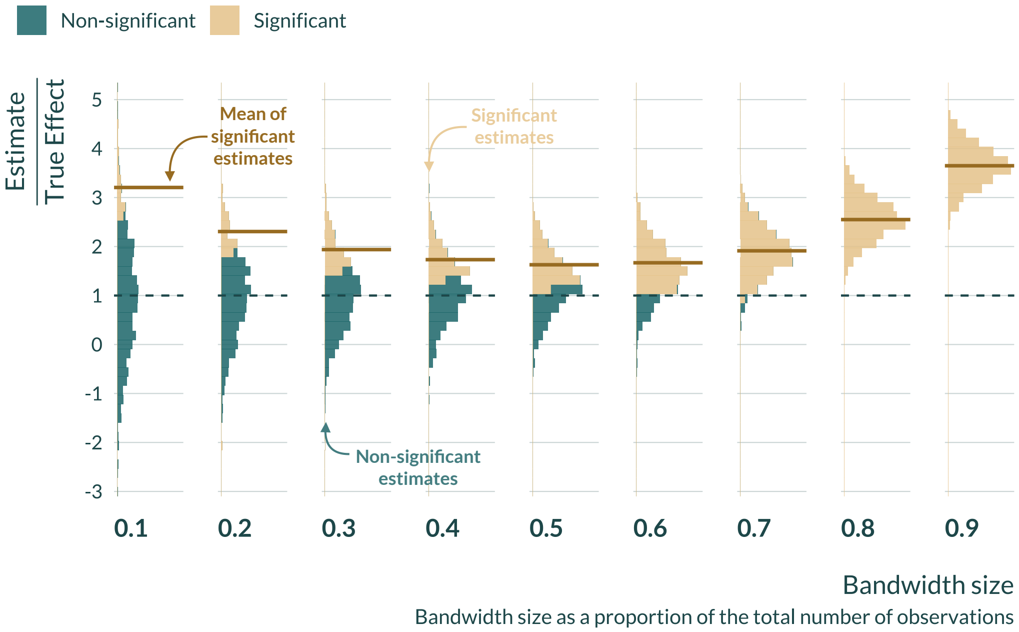
Case-study and simulation procedure. To illustrate this trade-off, I consider a standard application of the sharp RD design from economics of education in which students are assigned to additional lessons based on the score they obtained on a standardized test. Thistlethwaite and Campbell (1960) introduced the concept of RDD using a similar type of quasi-experiment. Students with test scores below a given threshold receive the treatment while those above do not. Since students far above and far below the threshold may differ along unobserved characteristics such as ability, a RDD estimates the effect of the treatment by comparing outcomes of students whose initial test scores are immediately below and above this threshold.

The simulation framework is as follows. If a student i has an initial scores $Qual_i$ below a cutoff C, they must take additional lessons, making the allocation of the treatment T sharp: $T_i = \mathbb{I}[Qual_i < C]$. Both qualification and final test scores are affected by students' unobserved ability w in a non-linear (cubic) way. A high or low ability has a strong positive impact on test scores while an average one does not strongly impact test scores. The qualifying test score of student i is thus: $Qual_i = \mu_q + \gamma f(w_i) + \epsilon_i$, where f a non-linear function (here cubic) and $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ random noise. Their final test score is: $Final_i = \beta_0 + \beta_1 T_i + \eta Qual_i + \delta f(w_i) + u_i$, where β_1 is the causal parameter of interest. In order that ability impacts qualifying and final scores similarly, I set $\delta = \gamma(1 - \eta)$.

These simulations are built and calibrated to emulate a realistic study from this literature. I derive parameters of the distribution of grades from statistics from the Department of Education, treatment effect size is based on a meta-analysis of RCTs in economics of education (Kraft 2020). Sample and bandwidth sizes are consistent with an existing study leveraging an RDD to explore a similar question (Jacob and Lefgren 2004). I define ability such that it creates a large bias for large bandwidths and a limited bias for small bandwidths. Further details on calibration are available on the project's website. Given these parameters values, I generate 1000 datasets with 60,000 observations. For each dataset, I estimate the treatment effect by regressing the final score on the treatment status and the qualifying score for different bandwidth sizes. The SNRs obtained in the simulations are aligned with SNRs observed in the literature, suggesting a realistic calibration.

Results. Figure 4 displays the results of these simulations. For large bandwidth sizes, the distribution of estimates is far away from the true effect; there is omitted variable bias. As the bandwidth size decreases, the identification strategy gets rid of the OVB and the distribution of the estimates centres on the true effet. At the same time, as the bandwidth and the effective sample size decrease, the distribution widens and significant estimates start representing a smaller and smaller subset of the distribution, located in the tails of the distribution, creating exaggeration. While the average of all estimates gets close to the true effect as bandwidth size and thus OVB decrease, in this setting the average of statistically significant estimates never gets close to the true effect. For large bandwidths, the omitted variable biases the effect while for small bandwidths, the small effective sample size creates exaggeration issues. The optimal bandwidth literature describes a similar trade-off but with different consequences (Imbens and Kalyanaraman 2012). They consider a bias-precision trade-off, I consider an omitted variable bias-exaggeration bias trade-off. Here, the parameter mediating the trade-off can directly be adjusted in a continuous way by the researchers and the more we reduce one of these two biases, the more we increase the other.

Figure 4 – Evolution of the Bias with Bandwidth Size in Regression Discontinuity Design, conditional on significance.



Notes: 1000 simulations. Significativity level: 5%, N = 60,000. The brown lines represent the average of significant estimates. The bandwidth size is expressed as the proportion of the total number of observations in the entire sample. Details on the simulation are available at this link.

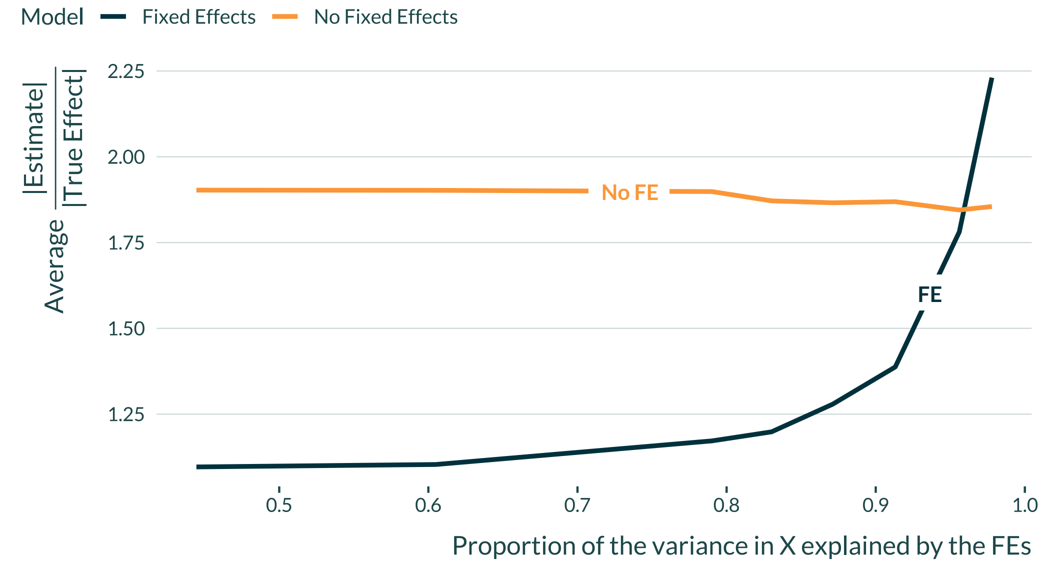
IV. 2 Fixed Effects and Controlling for Confounders

Intuition. To identify a causal effect and avoid the risk of confounding, an "ideal" approach would be to partial these confounders out by directly controlling for them. However, as discussed in the introduction and section III, controlling for an additional variable may increase the variance of the estimator if it absorbs more variation in the explanatory variable of interest than in the outcome variable. The same reasoning applies to a cornerstone causal identification approach: Fixed Effects (FEs). If including FEs partials out more of the variation in x than in y, it will increase the variance of the estimator and can lead to exaggeration.

Case-study and simulation procedure. To highlight this trade-off, I consider the case of studies on the impact of temperature on worker productivity. A typical approach in this literature consists in estimating the link between different temperature bins and productivity, focusing in particular on high temperature bins (Lai et al. 2023, for a review of the literature). Usual approaches to explore this question typically rely on High Dimensional Fixed Effects models, including time and location fixed effects to adjust for invariant characteristics such as seasonal demand or location differences. For readability, in my simulations I only consider time fixed effects and abstract from location ones. I assume that the true data generating process for temperature at time t in period τ is $Temp_{t\tau} = \mu_{Temp} + \gamma\lambda_\tau + \epsilon_t$ where μ_{Temp} is an intercept representing the average temperature and λ_τ are period fixed-effects with mean 0 and variance 1, such that their variance can be modified *via* γ. This particular structure of fixed effects allows controlling for the intensity of the fixed effects via the parameter γ, determining the proportion of the variation in temperature that comes from period-to-period (month) variations in average temperature. The productivity of worker i is defined as $Prod_{it\tau} = \beta_0 + \eta_i + \delta\lambda_\tau + \beta_1 \cdot \mathbb{1}\{Temp_{t\tau} \in (T_L, T_H]\} + u_{it}$ where η_i are individual fixed effects and $(T_L, T_H]$ the temperature bin of interest. The effects for other bins are not simulated and set to zero. For the omitted variable bias to remain constant when γ varies, I fix $\kappa = \gamma\delta$. I also adjust for the variances of ϵ and η to keep the variance of x and y constant.

I calibrate my simulations to replicate the distribution of the variables and interactions between variables from a typical study from the literature of interest (Stevens 2017, Somanathan et al. 2021, Lai et al. 2023, LoPalo 2023, among others). A more detailed description of calibration and modeling choices is available on the project's website. For each value of γ, or equivalently each proportion of the variation in temperature that is period-specific considered, I generate 1000 datasets with 49,000 observations and estimate two models, one including and the other omitting period fixed effects. The SNRs obtained are in line with many of those observed in the literature.

Figure 5 – Evolution of the Bias with the Proportion of the Variance in X explained by the Fixed Effects, for Statistically Significant Estimates.



Notes: The blue line indicates the average bias for estimates from the model including FEs that are statistically significant at the 5%. The orange line represents the bias of statistically significant estimates from the model without the fixed effects. In this simulation, N = 50,000. Details on the simulation and calibration are available at this link.

Results. Figure 5 displays the results of these simulations. When the (period) fixed effects explain a large part of the variation in x (temperature), the estimator is imprecise and leads to exaggeration, to the point that it becomes larger than the OVB simulated. While this phenomenon arises when fixed effects explain a relatively important share of the variance in temperature, it remains non-negligible for month-to-month variations in temperature routinely observed in actual situations. In addition, this setting presents large effect sizes and relatively large samples; other settings with smaller effect and sample sizes likely exhibit substantial exaggeration for smaller shares of variation captured by fixed effects. The inclusion of additional fixed effects might also further increase exaggeration.

IV. 3 Instrumental Variables Strategy

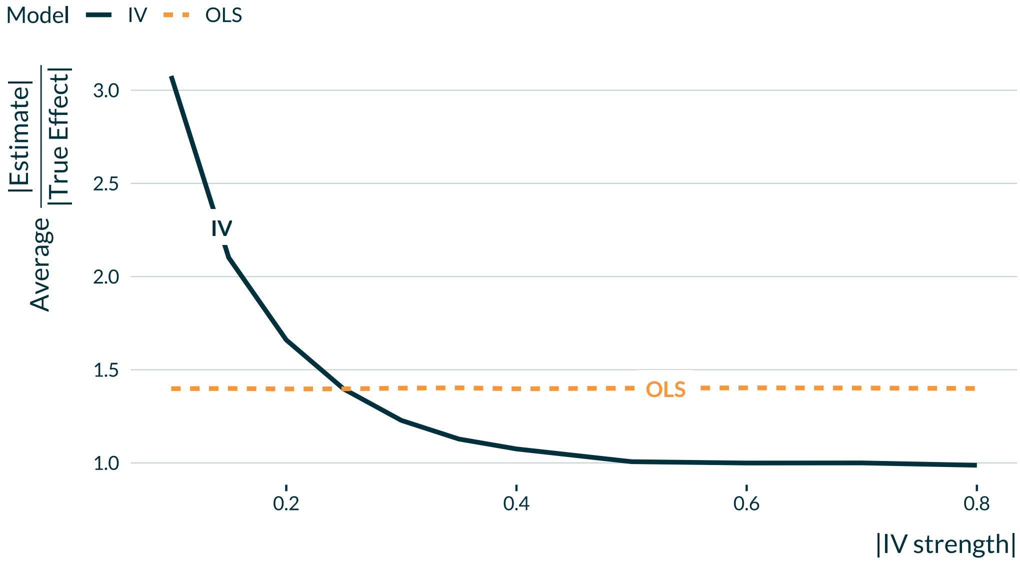
Intuition. Instrumental variables strategies overcome the issue of unobserved confounding by only considering the exogenous variation in the treatment, *i.e.* the variation that is explained by the instrument. Even when this exogenous fraction of the variation is limited, the instrument can successfully eliminate confounding on average. However, in such cases, the IV estimator will be imprecise and statistical power low. In the case of the IV, the confounding-exaggeration trade-off is mediated by the strength of the instrument considered. The weaker the instrument, the more inflated statistically significant estimates will be.

Case-study and simulation procedure. To illustrate this trade-off and its drivers, I consider the example of the impact of voter turnout on election results. To avoid the threat of confounding in this setting, existing studies take advantage of exogenous factors such as rainfall that affect voter turnout.[7] I reproduce such setting and assume that the true data generating process for the republican vote share is such that in location i, $Share_i = \beta_0 + \beta_1 Turnout_i + \delta w_i + u_i$, where w is an unobserved variable and $u \sim \mathcal{N}(0, \sigma_u^2)$ some random noise. The causal parameter of interest is β_1. In addition, turnout is affected by the amount of rain: $Turnout_i = \pi_0 + \pi_1 Rain_i + \gamma w_i + e_i$, where $Rain_i$ is the amount of rain in location i on the day of the election and e some random noise drawn form $\mathcal{N}(0, \sigma_e^2)$. I refer to π_1 as the strength of the instrumental variable.

To make the simulations realistic, I calibrate them on existing studies. I derive sample size, distribution parameters and effect sizes from a set of studies using similar variables (Gomez et al. 2007, Fujiwara et al. 2016, Cooperman 2017). Details on the calibration choices are available on the project's website. For each value of IV strength considered, I create 1000 datasets of 30,000 observations. I run both a naive OLS and a 2SLS model to estimate the impact of voter turnout on republican vote share. SNR obtained are aligned with SNR observed in this literature.

[7] I abstract from potential exclusion restriction violations of this instrument and simulate it as exogenous.

Figure 6 – Evolution of the Bias of Statistically Significant Estimates Against Strength of the Instrument in the IV Case.



Notes: The blue line indicates the average bias for IV estimates that are statistically significant at the 5%. The orange line represents the bias of statistically significant OLS estimates at the 5% level. The strength of the instrumental variable is expressed as the value of the linear parameter linking rainfall to turnout. In these simulations, N = 30,000. Details on the simulation are available at this link.

Results. Figure 6 displays, for different IV strengths, the average of statistically significant estimates scaled by the true effect size for both the IV and the naive regression model. When the instrument is strong, the IV will recover the true effect, contrarily to the naive regression model. Yet, when the IV strength decreases, the exaggeration of statistically significant estimates skyrockets. Even if the intensity of the omitted variable bias is large, for limited IV strengths, the exaggeration ratio can become larger than the omitted variable bias. When the only available instrument is weak, using the naive regression model would, on average, produce statistically significant estimates that are closer to the true effect size than the IV. Of interest for applied research, a large F-statistic does not necessarily attenuate this problem. This result complements limitations around the use of first-stage F-statistics with non-iid errors (Young 2022, Lal et al. 2024).

IV. 4 Exogenous shocks

Intuition. Taking advantage of exogenous variation in the treatment status caused by exogenous shocks or events can also enable avoiding confounding. In many settings, while the number of observations may be large, the number of events, their duration or the proportion of individuals affected might be limited. As a consequence, the number of (un)treated observations can be small and the variation available to identify the treatment limited. As extensively discussed in the randomized controlled trial literature, statistical power is maximized when the proportion of treated observations is equal to the proportion of untreated ones and drops when one of these proportions gets close to 0. In studies using discrete exogenous shocks, a confounding-exaggeration trade-off is thus mediated by the number of treated observations.

Case-study and simulation procedure. To illustrate this trade-off, I simulate a study of the impact of air pollution reduction on newborn weight of babies. To avoid confounding, one can exploit exogenous shocks to air pollution such as plant closures, creation of a low emission zone or of an urban toll (Currie et al. 2015, Lavaine and Neidell 2017). I simulate a typical analysis, at the zip code and monthly levels and focus on the example of toxic plant closures. I consider that the average birth weight in zip code z at time period t, bw_{zt}, depends on a zip code fixed effect ζ_z, a time fixed effect τ_t, and the treatment status T_{zt}, equal to one if a plant is closed in this period and 0 otherwise. The average birth weight in zip code z at time t is defined as follows: $bw_{zt} = \beta_0 + \beta_1 T_{zt} + \zeta_z + \tau_t + u_{zt}$. To further simplify the identification of the effect, I assume a non-staggered treatment allocation and constant and homogenous effects. I only vary the proportion of zip codes affected by toxic plant closings, keeping the length of the closures fixed.

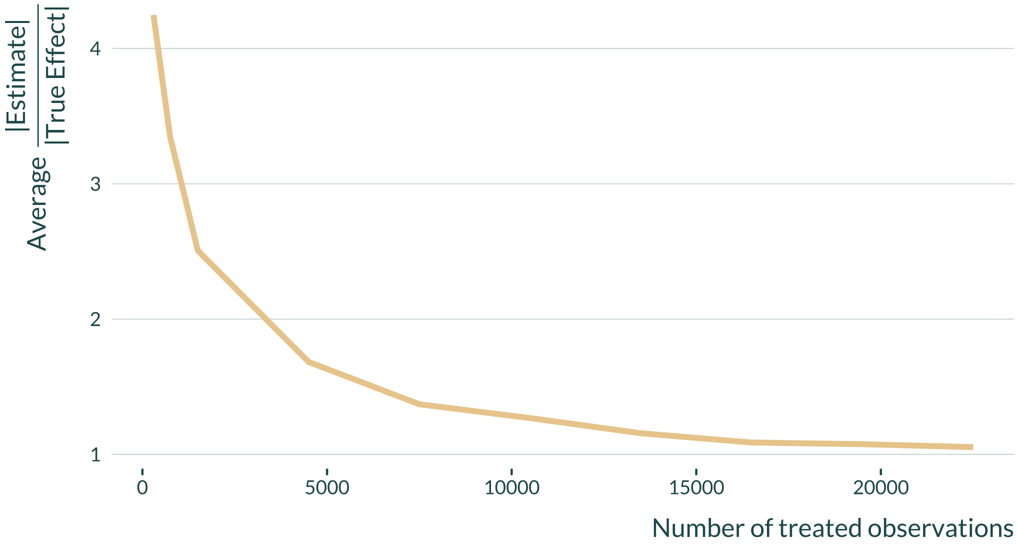
I derive parameters values from studies from the literature (Currie et al. 2015, Lavaine and Neidell 2017). The simulations mimic the design of such studies, considering a similar number of observations (150,000), distributions of variables, treatment allocation procedure and treatment effect sizes. I generate datasets with an increasing number of treated observations, varying the proportion of treated units and estimate the correct two-way fixed effects model. The simulation procedure is described in details on the project's website.

Results. Figure 7 displays the results of these simulations. As expected, exaggeration is larger for a smaller number of treated observations, keeping every thing else constant. Even though the actual sample size is extremely large in this example, if the number of treated observations is small, as it can be the case in this literature, exaggeration can be substantial. A very large number of observations does not necessarily shield us from exaggeration.

V Navigating the trade-off

In the previous sections, I showed using causal identification strategies induces a trade-off between avoiding confounding and exaggerating true effects. How can we, as applied researchers using observational data, arbitrate it? Since key pieces of information such as the true effect and the effect of omitted variables are inherently unknown, we cannot directly compute the biases

Figure 7 – Evolution of Bias With the Number of Treated Observations, for Statistically Significant Estimates, in the Exogenous Shocks Case



Notes: Significance level: 5%. In this simulation, N = 150,000. 1000 iterations for each number of treated observations considered. Details on the simulation are available at this link.

caused by confounders and exaggeration. In this section, I examine how we can, however, get a sense of threats from both sides of the trade-off and probe its main driver, the variation used for identification. I then discuss how changing attitudes towards statistical significance and replicating studies could limit the exaggeration issue.

V. 1 Gauging omitted variable bias

On one side of the trade-off lies the widely discussed bias caused by confounders. Although it is in essence impossible to measure, tools such as sensitivity analyses are available to gauge its magnitude (Rosenbaum 2002, Middleton et al. 2016, Oster 2019, Cinelli and Hazlett 2020). For instance, the method developed in Cinelli and Hazlett (2020) enables assessing how strong confounders would have to be to change the estimate of the treatment effect beyond a given level we are interested in. It offers bounds for the strength of the association between the treatment and potential omitted variables by weighting it against the measured association between the treatment and observed covariates. A typical conclusion from such an analysis would be: "omitted variables would have to explain as much residual variance of the outcome and the treatment as the observed covariate x (age for instance) to bring down the estimate to a value of β_l". The authors also implement graphical tools to facilitate this comparison. I suggest using quantitative bias analyses to assess how restrictive a causal approach must be to limit unobserved confounding to acceptable levels. Extremely strict controls and fixed effects may leave little scope for confounders but can generate substantial exaggeration. In such situations, the potential influence of confounders should be weighed carefully: given the risks of exaggeration—and concerns about representativity—causal estimates should not be treated as definitive benchmarks for assessing naive estimates, nor should a discrepancy between the two be taken as sufficient evidence that the naive estimate is incorrect.

V. 2 Evaluating risks of low power and exaggeration

On the other side of the trade-off lies the exaggeration emerging when statistical power is low. As OVB, exaggeration and statistical power are in essence impossible to measure as their computation depends on the true effect which is always unknown. Yet, power calculations can help assess them by making hypothesizes on the magnitude of the true effect. In randomized controlled trials, such computations are not only an established practice but a requirement (Duflo et al. 2007, McConnell and Vera-Hernández 2015, Athey and Imbens 2016). They are, however, rarely reported in non-experimental studies. Yet, taking publication bias and the threat of exaggeration into account highlights the necessity of running power calculations in non-experimental studies as well. A low power or a relatively large variance not only makes it more difficult to detect an effect or to draw clear conclusions about its magnitude when detected but it can also create a bias. To avoid this bias, I advocate to make power more central to non-experimental analyses. Currently, in causal inference textbooks, very few pages are devoted to statistical power in non-experimental studies (Angrist and Pischke 2009; 2014, Imbens and Rubin 2015, Cunningham 2021). To the best of my knowledge, only two textbooks discuss the matter in depth (Shadish et al. 2002, Huntington-Klein 2021). Results from power and exaggeration calculations would not only be highly informative but would also be very easy to report in the robustness section of articles.

V. 2.1 Prospective power calculations

To evaluate the statistical power of a study, the risk of exaggeration and identifying the factors driving it, one can simulate the design of the study before implementing it (Hill 2011, Gelman 2020, Black et al. 2022). Simulating a data generating process from scratch requires thinking about the distribution of the variables, about their relationships and can also help underline the variation used for identification. I implemented such Monte Carlo simulations in Section IV. The replication material and `R` code I provide can be used as an example to implement simulations for most causal identification strategies, based on data generated from scratch. In situations where the relationships among covariates are too complex to emulate, one can also start from an existing dataset and add a fake and known treatment effect to the data. I implemented such real-data simulations in a companion paper and describe their implementation in its replication material

(Bagilet 2023).

V. 2.2 Retrospective power calculations

Running post-analysis power calculations can also help getting a sense of the statistical power associated with a research design, as seen in section II. Such *retrospective* calculations allow evaluating whether the design of the study would produce accurate and uninflated statistically significant estimates if the true effect was in fact smaller than the observed estimate (Gelman and Carlin 2014, Ioannidis et al. 2017, Stommes et al. 2023).

I illustrate how a retrospective analysis works by taking the example of Card (1993) on the relationship between human capital and income. The IV strategy in this paper finds that an additional year of education, instrumented by the distance of growing up near a four-year college, causes a 13.2% average increase in wage, with an associated standard error of 5.5%. Is there a risk of exaggeration with this design? Since, as noted by the author himself, the estimate is very imprecise we could expect so. Imagine the existing literature suggests that such effects are likely to be close to a 10% increase in wage. We may wonder if the design in Card (1993) would allow detecting such an effect. We can thus easily compute the statistical power and exaggeration of Card's design under the hypothesis of a true effect size of 10%.[8] The average statistically significant estimate at the 5% level would be roughly of 15%, therefore overestimating the true effect by a factor of 1.5. Statistical power would only be 44%. Conditional on a 10% true effect size being a reasonable assumption, this design would be under-powered and exaggeration substantial.

The usefulness of any retrospective power analysis lies on the assumption made regarding the true effect size. To identify a range of plausible effect sizes one can rely on results from meta-analyses or from existing studies that have a credible design (*e.g.*, a large randomized controlled trial). When such meta-analyses are available, one can use Bayesian shrinkage to adjust statistically significant estimates in light of the distribution of estimates obtained in prior studies (Zwet and Gelman 2021, Zwet and Cator 2021, Zwet et al. 2021). When such information is not available, power calculations can be ran for a range of smaller but credible effect sizes that can be for instance derived from theoretical findings. It is also possible to evaluate whether the design of our study would be able to detect smaller effects than the point estimate obtained, while keeping in mind that there might be shortcomings to this approach as the obtained estimate or previously published ones may themselves be exaggerated (Gelman and Carlin 2014).

[8] Timm et al. (2019) and Linden (2019) offer `R` and `Stata` packages that enable easily running these calculations through an extremely short command: `retrodesign(10, 5.5)`. As discussed previously, the exaggeration is computed by drawing a large number of times from a normal distribution centred on the hypothetical effect size of 10% and with a standard deviation equal to the standard error of the estimate found in the study (here 5.5%) and by then computing the average of the absolute value of the draws that are 1.96 standard errors away from 0.

V. 2.3 Precision matters even after obtaining a significant estimate

In non-experimental studies, estimator variance often plays a critical role because a large variance may prevent rejecting the null hypothesis when it is false. As a result, variance is typically a central concern until a statistically significant estimate is obtained. However, exaggeration highlights that variance remains important for the reliability of the point estimate—and for reasons beyond mere uncertainty—even after statistical significance has been achieved.

Obtaining a statistically significant estimate from an imprecise estimator should not necessarily be interpreted as a sign of "success" in getting significance despite a wide confidence interval. Rather, it may signal that this estimate liens in the tails of the distribution and thus inaccurately represents the true effect. This observation reframes the conventional bias-variance trade-off as a bias-bias trade-off: a larger variance can lead to a larger bias, even in (conditional) expectation. This paper therefore calls for careful attention to the implications of design and modeling choices on estimator variance, even when large variance has not prevented statistical significance.

V. 3 Identifying the identifying variation

V. 3.1 A metric of identifying observations

The approaches outlined above for assessing exaggeration and confoundings are built on assumptions regarding the true effect size. To circumvents these assumptions, one can instead directly focus on the driver of the bias-variance trade-off: the variation used for identification. When the identifying variation is too limited, statistical power is low and exaggeration is large. Pinpointing the specific variation on which estimation relies can provide a direct way to gauge the risk of exaggeration. This section introduces tools to identify that variation.

To identify the identifying variation in an applied study, one usually begins with a thought exercise: where does the variation leveraged for identification come from? This reflection is integral to the implementation of any applied economics study. While it is typically undertaken primarily for identification and internal validity purposes, this paper also highlights its importance for exaggeration reasons, *i.e.* after a significant estimate has been obtained. Although this exercise may be straightforward in simple cases—such as one-way fixed effects—it becomes more challenging in complex designs involving multiple or interacting fixed effects, especially for readers less familiar with the setting.

Given this difficulty of intuitively identifying the source of variation in complex models, it can be useful to develop measures that quantify the contribution of each observation or group of observations to the estimation of the treatment effect. A range of existing tools from the statistics literature, such as leverage and Cook's distance, measure the influence of individual observations on regression parameters. These measures, however, assess influence on the *entire* parameter vector and are not directly suited for applied economics where interest is typically confined to a single parameter: the coefficient of the treatment variable. To get to a more suited measure, I propose to first apply the Frisch-Waugh-Lovell theorem and then compute leverage for the regression of the residualized outcome on the residualized treatment. The residuals are obtained from regressions on the full set of controls, including fixed effects and other identification-related controls such as control functions. This produces observation-specific weights describing the extent to which each observation contributes to the estimation of the treatment effect.

These weights are equivalent—up to a normalization to one—to the multiple regression weights introduced in Aronow and Samii (2016) and previously discussed in Angrist and Pischke (2009). The estimate of the treatment coefficient in a simple linear regression can be interpreted as a weighted average of individual treatment effects. The weight ω_i of individual i is the squared difference between its treatment status T_i and the value of this treatment status as predicted by the other covariates X: $\omega_i = (T_i - \mathbb{E}[T_i|X_i])^2$. At the group level, these weights connect to well-established insights, particularly in the widely used case of fixed effects. A group's weight, *ie* the sum of the individual weights within that group, corresponds to the within-group variance of the conditional treatment status. Since including fixed effects changes the type of variation used for identification from pooled to within-group variation, groups with little within variation in treatment status only marginally contribute to identification. These weights also have crucial implications at the observation level: individual observations whose treatment status is largely explained by covariates contribute little, if at all, to estimation. Variation in the explanatory variable of interest is absorbed by the controls. In that sense, the weights not only highlight the role of groups but also the uneven contribution of individual observations within them. More broadly, in settings without a natural grouping structure, observation-level weights help assess how much each unit actually contributes to identification. Aronow and Samii (2016) uses these weights to argue that, when treatment effects are heterogeneous, some observations may be disproportionately represented in the estimated average effect, raising concerns about external validity. Even in the absence of heterogeneity, however, relying on observations whose treatment status is well explained by covariates can yield a very small *effective* sample size and, in turn, exaggeration. Whereas Aronow and Samii (2016) emphasizes the representativity of the effective sample for external validity reasons, the present paper focuses on its size, for fear of exaggeration.

V. 3.2 A tool to identify the identifying variation

Although the computation of these weights is relatively straightforward, it involves several steps, and their subsequent analysis and comparison are more challenging. To facilitate this process, I develop the `ididvar` package for `R`.[9] The package provides a unified tool to compute identifying-variation weights and related contribution metrics, along with a series of visualization tools. By lowering the cost of identifying which observations drive estimation, the package aims helping assess potential risks of exaggeration when identifying variation is limited. Unlike power calculations, these metrics can be computed without assumptions about the true effect size.

Most functions are high-level and easy to use, requiring only the regression output and the variable of interest. A first set of functions computes and visualizes weights, including their distribution across space, groups, or time. This enables users to detect low-weight observations or clusters that contribute little to identification. Because the distribution of weights is often heterogenous, building meaningful visualization can be challenging. To address this, I propose a discretized and logged scale that compares each weight to the average weight, $1/n$, where n is the number of observations or groups. This representation clarifies variation and ensures comparability across specifications and studies.

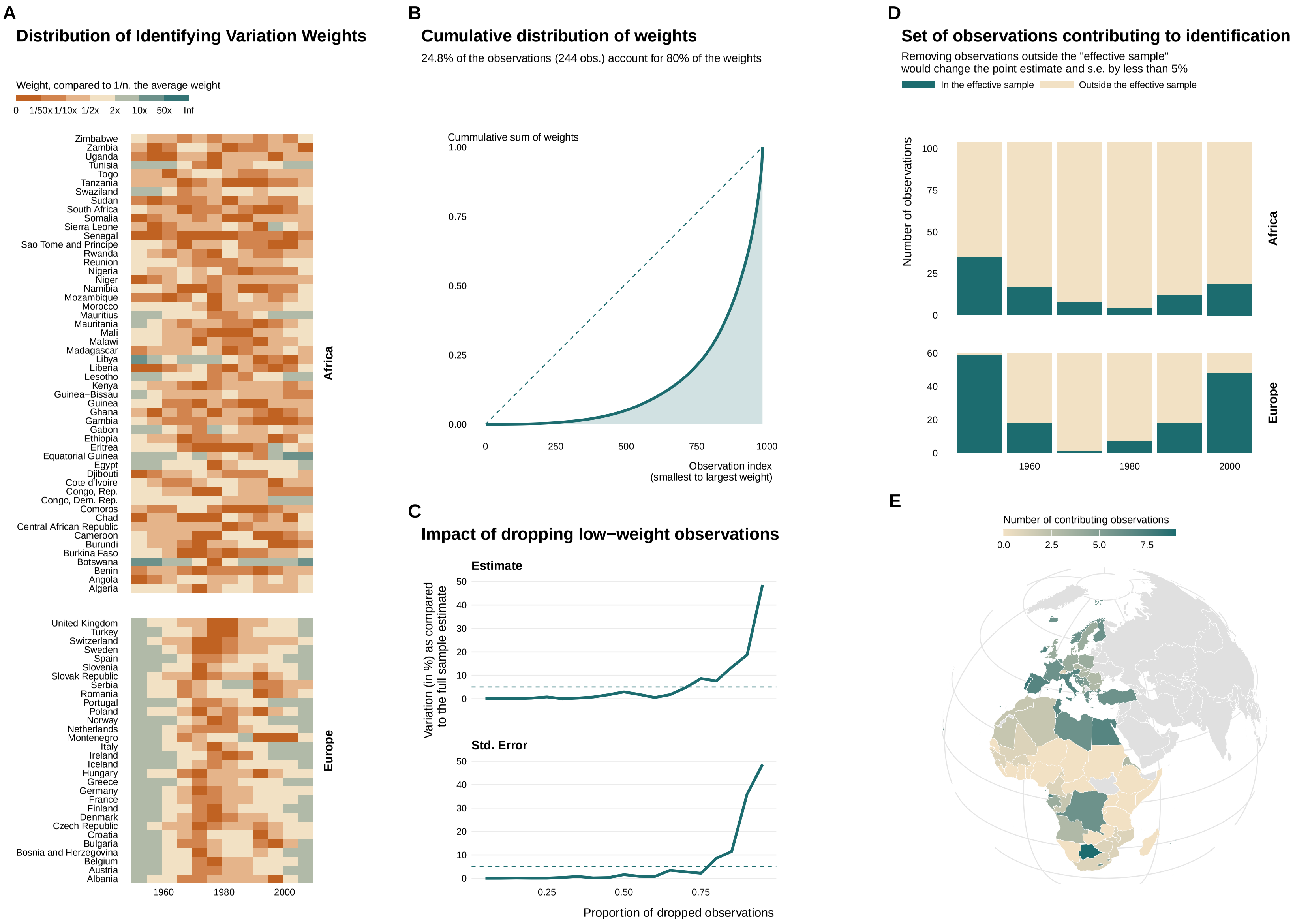
To illustrate the package's features, I consider a simple example using a country-level panel (Bryan et al. 2017, restricted to Africa and Europe for clarity) and estimate the within-country relationship between GDP per capita and life expectancy: $\log(PerCapitaGDP_{ct}) = \alpha_c + \beta LifeExp_{ct} + u_{ct}$. Figure 8 presents visualizations produced by the package. Panel A displays the identifying-variation weights for each observation,[10] while Panel B shows their cumulative distribution. The distribution is highly heterogeneous: many observations contribute little to identification, particularly those from mid-periods and European countries.

Weights alone, however, do not indicate which observations actually contribute to identification. A second set of functions identifies the *effective sample*: the smallest subset of observations that cannot be removed without altering the point estimate or standard error of the parameter of interest by more than a chosen margin, e.g., 5%. The procedure iteratively re-estimates the model, removing low-weight observations following an approach similar to influence-function methods. Reporting the effective sample size helps interpret results in light of the sample that actually drives identification, much as one would interpret the nominal sample size. It does not, however, imply that these observations should be excluded from the main estimation.

[9] This package can be extended to accommodate more specific settings. Readers are encouraged to suggest improvements through GitHub issues or pull requests.

[10] This graph is essentially generated in one line of code: `idid_viz_weights(reg, ”l_gdpPercap”, year, country)` where `reg` is the output of the regression of interest, `l_gdpPercap` the variable of interest and `var_x` and `var_y` specify the variables to be displayed n the x and y-axes respectively

Figure 8 – Illustration of the types of analyses ididvar allows



Notes: Details on the computation are available at this link.

In the illustrative example, Panel C of Figure 8 shows how point estimates and standard errors evolve as low-weight observations are sequentially dropped. 70% of the nominal sample can be dropped without affecting estimates and standard errors that differ by more than 5%. The effective sample is thus about three times smaller than the nominal sample. Results should be interpreted accordingly. Panels D and E characterize this effective sample: many observations from Africa and mid-periods do not contribute to identification, implying that the effective sample is not representative of the full dataset.

The package is flexible and can be applied to a wide range of identification strategies and estimation methods. It enables users to compare specifications and explore how the inclusion of different sets of fixed effects or controls affects the effective sample. More broadly, by making the structure and amount of identifying variation explicit, the package provides a concrete way to assess where a study lies along the confounding-exaggeration trade-off. In doing so, it aims to help users evaluate whether the credibility gained from additional controls comes at the cost of a loss of identifying variation and statistical power.

V. 4 Attitude Towards Statistical Significance and Replication

Exaggeration only arises in the presence of publication bias. As shown in the simulations, if estimates were not filtered by their statistical significance nor any other characteristics, even underpowered studies would on average recover the true effect, as long as the estimator is unbiased. The exaggeration issue could therefore be addressed by tackling publication bias. [11]

To identify broader pathways to eliminate this filtering of significant results, it is first helpful to discuss the processes that lead to statistically significant results when power and thus the probability of obtaining a significant estimate is low. In such situations, they can be obtained either by "chance" or as an outcome of the garden of forking paths (Simmons et al. 2011, Gelman and Loken 2013, Kasy 2021). Decision forks appear at various stages along the path of research, for instance in data preparation, regarding the inclusion of a given control variable in the model or observation in the sample or later, regarding whether to carry on with a research that yields non-significant results. Due to the structural flaw that favors significance, the path followed may be more likely to lead to a statistically significant result. These choices are most often not the result of bad researcher practices but instead a product of a structure that portrays significant results as an end goal of research.

[11] Note that even if this publication bias is often understood as selection of statistical significance, it can also extend to any situation in which a certain type of results, large or surprising for instance, are favored. The present paper focuses on the former type of selection owing to its primacy and the fact that its existence and the mechanisms causing it are well-documented.

The issue being structural, system level changes in scientific practices could also alleviate exaggeration and the trade-off described in this paper. First, many researchers advocate abandoning statistical significance as a measure of a study's quality (McShane et al. 2019). To be effective, this change should be paired with an effort to replicate studies (Christensen and Miguel 2018). Replications, even of low powered studies, would eventually enable building the actual distribution of the causal estimand of interest. Meta-analyzes would then reduce the uncertainty around the true value of the causal estimand by pooling estimates (Hernán 2022). Finally, the inflation of statistically significant estimates could be limited by interpreting confidence intervals and not point estimates and thus considering these intervals as compatibility intervals (Shadish et al. 2002, Amrhein et al. 2019, Romer 2020). The width of such intervals gives a range of effect sizes compatible with the data. These intervals will be wide in under-powered studies signalling that point estimates should not be taken at face value, even if statistically significant.

VI Conclusion

The economic literature suffers from an extensive lack of statistical power (Ioannidis et al. 2017) and strongly favors statistically significant findings (Rosenthal 1979, Andrews and Kasy 2019, Abadie 2020, Brodeur et al. 2020, for instance). In such situations, significant estimates from underpowered studies always exaggerate true effect sizes, even when the estimators are "unbiased" in the usual sense of $\mathbb{E}[\hat{\beta}] = \beta$ (Gelman and Tuerlinckx 2000, Ioannidis 2008, Gelman and Carlin 2014). It is therefore not surprising that many estimates published in economics have been found to be considerably exaggerated (Ioannidis et al. 2017), despite the extensive use of convincing causal inference methods. Yet, the determinants of these exaggeration and power problems have remained largely understudied. I argue that exaggeration is exacerbated by the foundational component of causal inference: the fact that it only leverages subsets of the variation for identification. Although causal methods enable avoiding confounding, they also reduce statistical power and thus increase the risk of exaggeration. The same aspect that makes these methods credible can create another type of bias. Systematically reporting statistical power and examining the variation actually used for identification could help avoid falling into this exaggeration trap.

References

Abadie, A. (2020), 'Statistical Nonsignificance in Empirical Economics', American Economic Review: Insights **2**(2), 193–208. 2, 33

Amrhein, V., Trafimow, D. and Greenland, S. (2019), 'Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication', The American Statistician **73**(sup1), 262–270. 33

Andrews, I. and Kasy, M. (2019), 'Identification of and Correction for Publication Bias', American Economic Review **109**(8), 2766–2794. 2, 33

Angrist, J. D. and Pischke, J.-S. (2009), Mostly Harmless Econometrics: An Empiricist's Companion, 1 edition edn, Princeton University Press, Princeton. 7, 26, 29

Angrist, J. D. and Pischke, J.-S. (2014), Mastering 'Metrics: The Path from Cause to Effect, Princeton University Press. 26

Arel-Bundock, V., Briggs, R. C., Doucouliagos, H., Mendoza Aviña, M. and Stanley, T. D. (2022), Quantitative Political Science Research is Greatly Underpowered, Working Paper 6, I4R Discussion Paper Series. 7

Aronow, P. M. and Samii, C. (2016), 'Does Regression Produce Representative Estimates of Causal Effects?', American Journal of Political Science **60**(1), 250–267. 7, 29

Athey, S. and Imbens, G. (2016), 'The Econometrics of Randomized Experiments', arXiv:1607.00698 [econ, stat] . 26

Bagilet, V. (2023), 'Accurate Estimation of Small Effects: Illustration Through Air Pollution and Health'. 5, 7, 8, 10, 27

Black, B., Hollingsworth, A., Nunes, L. and Simon, K. (2022), 'Simulated power analyses for observational studies: An application to the Affordable Care Act Medicaid expansion', Journal of Public Economics **213**, 104713. 6, 7, 26

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, 5, 8, 9, 33

Brodeur, A., Lé, M., Sangnier, M. and Zylberberg, Y. (2016), 'Star Wars: The Empirics Strike Back', American Economic Journal: Applied Economics **8**(1), 1–32. 2

Bryan, J., Kim, A. Y. and MacDonald, A. (2017), 'Jennybc/gapminder: V0.3.0'. 30

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, 7, 10

Card, D. (1993), Using Geographic Variation in College Proximity to Estimate the Return to Schooling, Working Paper 4483, National Bureau of Economic Research. 27

Christensen, G. and Miguel, E. (2018), 'Transparency, Reproducibility, and the Credibility of Economics Research', Journal of Economic Literature **56**(3), 920–980. 7, 33

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. 25

Cooperman, A. D. (2017), 'Randomization Inference with Rainfall Data: Using Historical Weather Patterns for Variance Estimation', Political Analysis **25**(3), 277–288. 22

Cunningham, S. (2021), Causal Inference: The Mixtape, Yale University Press. 26

Currie, J., Davis, L., Greenstone, M. and Walker, R. (2015), 'Environmental Health Risks and Housing Values: Evidence from 1,600 Toxic Plant Openings and Closings', American Economic Review **105**(2), 678–709. 24

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

Dehejia, R. H. and Wahba, S. (1999), 'Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs', Journal of the American Statistical Association **94**(448), 1053–1062. 45

Deryugina, T., Heutel, G., Miller, N. H., Molitor, D. and Reif, J. (2019), 'The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction', American Economic Review **109**(12), 4178–4219. 10, 11

Duflo, E., Glennerster, R. and Kremer, M. (2007), Using Randomization in Development Economics Research: A Toolkit, in T. P. Schultz and J. A. Strauss, eds, 'Handbook of Development Economics', Vol. 4, Elsevier, pp. 3895–3962. 26

Ferraro, P. J. and Shukla, P. (2020), 'Feature—Is a Replicability Crisis on the Horizon for Environmental and Resource Economics?', Review of Environmental Economics and Policy **14**(2), 339–351. 2, 7

Fujiwara, T., Meng, K. and Vogl, T. (2016), 'Habit Formation in Voting: Evidence from Rainy Elections', American Economic Journal: Applied Economics **8**(4), 160–188. 22

Gelman, A. (2020), Regression and Other Stories, Cambridge University Press, Cambridge New York, NY Port Melbourne, VIC New Delhi Singapore. 6, 26

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, 12, 27, 33

Gelman, A. and Loken, E. (2013), 'The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time'. 32

Gelman, A. and Tuerlinckx, F. (2000), 'Type S error rates for classical and Bayesian single and multiple comparison procedures', Computational Statistics **15**(3), 373–390. 2, 33

Gomez, B. T., Hansford, T. G. and Krause, G. A. (2007), 'The Republicans Should Pray for Rain: Weather, Turnout, and Voting in U.S. Presidential Elections', The Journal of Politics **69**(3), 649–663. 22

Gray, W. B., Shadbegian, R. and Wolverton, A. (2023), 'Environmental Regulation and Labor Demand: What Does the Evidence Tell Us?', Annual Review of Resource Economics **15**(1), 177–197. 3

Greenstone, M. (2002), 'The Impacts of Environmental Regulations on Industrial Activity: Evidence from the 1970 and 1977 Clean Air Act Amendments and the Census of Manufactures', Journal of Political Economy **110**(6), 1175–1219. 3

He, G., Liu, T. and Zhou, M. (2020), 'Straw burning, PM2.5, and death: Evidence from China', Journal of Development Economics **145**, 102468. 10, 11

Hernán, M. A. (2022), 'Causal analyses of existing databases: No power calculations required', Journal of Clinical Epidemiology **144**, 203–205. 33

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

Hill, J. L. (2011), 'Bayesian Nonparametric Modeling for Causal Inference', Journal of Computational and Graphical Statistics **20**(1), 217–240. 26

Huntington-Klein, N. (2021), The Effect: An Introduction to Research Design and Causality, 1 edn, Chapman and Hall/CRC, Boca Raton. 26

Imbens, G. and Kalyanaraman, K. (2012), 'Optimal Bandwidth Choice for the Regression Discontinuity Estimator', The Review of Economic Studies **79**(3), 933–959. 6, 19

Imbens, G. W. and Rubin, D. B. (2015), Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction, Cambridge University Press, Cambridge. 26

Ioannidis, J. P. A. (2008), 'Why Most Discovered True Associations Are Inflated', Epidemiology **19**(5), 640–648. 2, 33

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. 2, 3, 7, 8, 9, 27, 33

Jacob, B. A. and Lefgren, L. (2004), 'Remedial Education and Student Achievement: A Regression-Discontinuity Analysis', The Review of Economics and Statistics **86**(1), 226–244. 19

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. 7, 32

Kraft, M. A. (2020), 'Interpreting Effect Sizes of Education Interventions', Educational Researcher **49**(4), 241–253. 19

Lai, W., Qiu, Y., Tang, Q., Xi, C. and Zhang, P. (2023), 'The Effects of Temperature on Labor Productivity', Annual Review of Resource Economics **15**(Volume 15, 2023), 213–232. 20, 21

Lal, A., Lockhart, M., Xu, Y. and Zu, Z. (2024), 'How Much Should We Trust Instrumental Variable Estimates in Political Science? Practical Advice Based on 67 Replicated Studies', Political Analysis pp. 1–20. 5, 8, 10, 23

Lavaine, E. and Neidell, M. (2017), 'Energy Production and Health Externalities: Evidence from Oil Refinery Strikes in France', Journal of the Association of Environmental and Resource Economists **4**(2), 447–477. 24

Linden, A. (2019), 'RETRODESIGN: Stata module to compute type-S (Sign) and type-M (Magnitude) errors', Boston College Department of Economics. 27

LoPalo, M. (2023), 'Temperature, Worker Productivity, and Adaptation: Evidence from Survey Data Production', American Economic Journal: Applied Economics **15**(1), 192–229. 21

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. 12, 40

McConnell, B. and Vera-Hernández, M. (2015), Going beyond simple sample size calculations: A practitioner's guide, Working Paper W15/17, IFS Working Papers. 26

McShane, B. B., Gal, D., Gelman, A., Robert, C. and Tackett, J. L. (2019), 'Abandon Statistical Significance', The American Statistician **73**(sup1), 235–245. 33

Middleton, J. A., Scott, M. A., Diakow, R. and Hill, J. L. (2016), 'Bias Amplification and Bias Unmasking', Political Analysis **24**(3), 307–323. 25

Oster, E. (2019), 'Unobservable Selection and Coefficient Stability: Theory and Evidence', Journal of Business & Economic Statistics **37**(2), 187–204. 25

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

Romer, D. (2020), 'In Praise of Confidence Intervals', AEA Papers and Proceedings **110**, 55–60. 33

Rosenbaum, P. R. (2002), Observational Studies, Springer Series in Statistics, Springer New York, New York, NY. 25

Rosenthal, R. (1979), 'The file drawer problem and tolerance for null results', Psychological Bulletin **86**(3), 638–641. 2, 33

Rubin, D. B. (2001), 'Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation', Health Services and Outcomes Research Methodology **2**(3), 169–188. 46

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. 7

Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002), Experimental and Quasi-experimental Designs for Generalized Causal Inference, Houghton Mifflin. 26, 33

Shah, A. S. V., Lee, K. K., McAllister, D. A., Hunter, A., Nair, H., Whiteley, W., Langrish, J. P., Newby, D. E. and Mills, N. L. (2015), 'Short term exposure to air pollution and stroke: Systematic review and meta-analysis', BMJ **350**, h1295. 10, 11

Simmons, J. P., Nelson, L. D. and Simonsohn, U. (2011), 'False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant', Psychological Science **22**(11), 1359–1366. 32

Somanathan, E., Somanathan, R., Sudarshan, A. and Tewari, M. (2021), 'The Impact of Temperature on Productivity and Labor Supply: Evidence from Indian Manufacturing', Journal of Political Economy **129**(6), 1797–1827. 21

Stevens, A. (2017), 'Temperature, Wages, and Agricultural Labor Productivity'. 21

Stommes, D., Aronow, P. M. and Sävje, F. (2023), 'On the reliability of published findings using the regression discontinuity design in political science', Research & Politics **10**(2). 6, 7, 27

Thistlethwaite, D. L. and Campbell, D. T. (1960), 'Regression-discontinuity analysis: An alternative to the ex post facto experiment', Journal of Educational Psychology **51**(6), 309–317. 18

Timm, A., Gelman, A. and Carlin, J. (2019), 'Retrodesign: Tools for Type S (Sign) and Type M (Magnitude) Errors'. 27

Walker, W. R. (2011), 'Environmental Regulation and Labor Reallocation: Evidence from the Clean Air Act', The American Economic Review **101**(3), 442–447. 3, 4

Young, A. (2022), 'Consistency without Inference: Instrumental Variables in Practical Application', European Economic Review **147**, 104112. 5, 8, 9, 10, 23

Zwet, E. and Gelman, A. (2021), 'A Proposal for Informative Default Priors Scaled by the Standard Error of Estimates', The American Statistician pp. 1–9. 27

Zwet, E., Schwab, S. and Senn, S. (2021), 'The statistical properties of RCTs and a proposal for shrinkage', Statistics in Medicine **40**(27), 6107–6117. 27

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. 12, 27, 40

A Mathematical proofs

I. 1 Variation of the exaggeration ratio (lemma 1)

Proof. Lu et al. (2019) and Zwet and Cator (2021) showed this in the case of $b = 0$. To extend it to the biased case, consider $E_b = \frac{\mathbb{E}\left[|\hat{\beta}_b| \, \middle| \beta_1, \sigma, |\hat{\beta}_b| > z_\alpha \sigma\right]}{|\beta_1|}$ the exaggeration ratio of interest. Note that, since $\hat{\beta}_b$ is an unbiased estimator of $\beta_1 + b$, $\tilde{E}_b = \frac{\mathbb{E}\left[|\hat{\beta}_b| \, \middle| \beta_1, \sigma, |\hat{\beta}_b| > z_\alpha \sigma\right]}{|\beta_{1+b}|}$ has the properties described in the lemma. Now, considering that $E_b = \left|\frac{\beta_1 + b}{\beta_1}\right| \tilde{E}_b$ proves the properties when β_1 and b have the same sign. □

I. 2 Asymptotic distribution of $\hat{\beta}_{\text{OVB}}$ (lemma 2)

For readability, let us introduce the usual vector notation such that for instance $y = (y_1, ..., y_n)'$ and set $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ and $\mathrm{x}_i = (1, x_i)'$. I also use capital letters to denote matrices (for instance $X = (\mathrm{x}_1', ..., \mathrm{x}_n')'$).

Proof. Since, we do not observe w, we consider the projection of y on X only:

$$y = X\boldsymbol{\beta}_{\text{OVB}} + u_{\text{OVB}} \tag{5}$$

where by definition of the projection, $\mathbb{E}[X' u_{\text{OVB}}] = 0$.

We first compute the bias of the estimator. From equation 5 we get:

$$\begin{aligned}
& X'y = X'X\boldsymbol{\beta}_{\text{OVB}} + X'u_{\text{OVB}} \\
\Rightarrow \quad & \mathbb{E}[X'y] = \underbrace{\mathbb{E}[X'X]}_{\text{pos. def.}} \boldsymbol{\beta}_{\text{OVB}} + \underbrace{\mathbb{E}[X'u_{\text{OVB}}]}_{0} \\
\Leftrightarrow \quad & \boldsymbol{\beta}_{\text{OVB}} = \mathbb{E}[X'X]^{-1}\mathbb{E}[X'(X\boldsymbol{\beta} + \delta w + u)] \quad \text{cf eq. } 2 \\
\Leftrightarrow \quad & \boldsymbol{\beta}_{\text{OVB}} = \boldsymbol{\beta} + \mathbb{E}[X'X]^{-1}\mathbb{E}[X'w]\delta
\end{aligned} \tag{6}$$

We then compute the asymptotic distribution. We can write:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{OVB}} - \boldsymbol{\beta}_{\text{OVB}}) = \left(\frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_i \mathrm{x}_i'\right)^{-1} \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \mathrm{x}_i u_{\text{OVB},i}\right)$$

Applying the Weak Law of Large Numbers (WLLN), the Central Limit Theorem (CLT) and Slutsky's theorem yields:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{OVB}} - \boldsymbol{\beta}_{\text{OVB}}) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}[\mathrm{x}_i\mathrm{x}_i']^{-1}\mathbb{E}[\mathrm{x}_i\mathrm{x}_i' u^2_{\text{OVB},i}]\mathbb{E}[\mathrm{x}_i\mathrm{x}_i']^{-1}\right) \tag{7}$$

We are interested in the second component of $\hat{\boldsymbol{\beta}}_{\text{OVB}}$. To retrieve it we need to compute $\mathbb{E}[\mathrm{x}_i\mathrm{x}_i']^{-1}$, $\mathbb{E}[\mathrm{x}_i w_i]$ and $\mathbb{E}[\mathrm{x}_i\mathrm{x}_i' u^2_{\text{OVB},i}]$.

$$\mathbb{E}[\mathrm{x}_i\mathrm{x}_i'] = \mathbb{E}\begin{bmatrix} 1 & x_i \\ x_i & x_i^2 \end{bmatrix} = \begin{bmatrix} 1 & \mu_x \\ \mu_x & \sigma_x^2 + \mu_x^2 \end{bmatrix} \quad \Rightarrow \quad \mathbb{E}[\mathrm{x}_i\mathrm{x}_i']^{-1} = \frac{1}{\sigma_x^2}\begin{bmatrix} \sigma_x^2 + \mu_x^2 & -\mu_x \\ -\mu_x & 1 \end{bmatrix}$$

$$\mathbb{E}[\mathrm{x}_i w_i] = \mathbb{E}\begin{bmatrix} w_i \\ x_i w_i \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbb{E}[x_i]\underbrace{\mathbb{E}[w_i]}_{0} + \text{cov}(x_i, w_i) \end{bmatrix} = \begin{bmatrix} 0 \\ \gamma\underbrace{\text{var}(w_i)}_{\sigma_w^2} + \underbrace{\text{cov}(\epsilon_i, w_i)}_{0} \end{bmatrix} = \begin{bmatrix} 0 \\ \gamma\sigma_w^2 \end{bmatrix}$$

$$\Rightarrow \quad \mathbb{E}[\mathrm{x}_i\mathrm{x}_i']^{-1}\mathbb{E}[\mathrm{x}_i w_i] = \frac{\gamma\sigma_w^2}{\sigma_x^2}\begin{bmatrix} -\mu_x \\ 1 \end{bmatrix} \tag{8}$$

Note that $\mathbb{E}[\mathrm{x}_i\mathrm{x}_i' u^2_{\text{OVB},i}] \overset{\text{LIE}}{=} \mathbb{E}[\mathrm{x}_i\mathrm{x}_i'\mathbb{E}[u^2_{\text{OVB},i}|\mathrm{x}_i]]$. We thus first compute $\mathbb{E}[u^2_{\text{OVB},i}|\mathrm{x}_i]$, noting that:

$$\begin{aligned}
u_{\text{OVB},i} &= y_i - \mathrm{x}_i'\boldsymbol{\beta}_{\text{OVB}} \\
&= \delta w_i + u_i + \mathrm{x}_i'(\boldsymbol{\beta} - \boldsymbol{\beta}_{\text{OVB}}) \\
&= \delta w_i + u_i - \underbrace{\mathrm{x}_i'\mathbb{E}[\mathrm{x}_i\mathrm{x}_i']^{-1}\mathbb{E}[\mathrm{x}_i w_i]}_{\text{projection of } w_i \text{ on } x_i}\,\delta \\
&= u_i + \delta\underbrace{\left(w_i - \frac{\gamma\sigma_w^2}{\sigma_x^2}(x_i - \mu_x)\right)}_{\text{part of } w_i \text{ orthogonal to } x_i} \\
&= u_i + \delta w_i^{\perp} \qquad\qquad \text{where } w_i^{\perp} = w_i - \frac{\gamma\sigma_w^2}{\sigma_x^2}(x_i - \mu_x)
\end{aligned}$$

And thus,

$$
\begin{aligned}
\mathbb{E}[u^2_{\text{OVB},i}|\mathrm{x}_i] &= \mathbb{E}[(u_i + \delta w_i^{\perp})^2|x_i] \\
&= \mathbb{E}[u_i^2|x_i] + 2\delta\mathbb{E}[u_i w_i^{\perp}|x_i] + \delta^2\mathbb{E}[(w_i^{\perp})^2|x_i] \\
&= \sigma_u^2 + 2\delta\left(\mathbb{E}[u_i w_i|x_i] - \frac{\gamma\sigma_w^2}{\sigma_x^2}(x_i - \mu_x)\underbrace{\mathbb{E}[u_i|x_i]}_{0}\right) + \delta^2\mathbb{E}[(w_i^{\perp})^2|x_i] \\
&\overset{\text{LIE}}{=} \sigma_u^2 + 2\delta\mathbb{E}[w_i\underbrace{\mathbb{E}[u_i|x_i, w_i]}_{0}|x_i] + \delta^2\mathbb{E}[(w_i^{\perp})^2|x_i] \\
&= \sigma_u^2 + \delta^2\mathbb{E}[(w_i^{\perp})^2|x_i]
\end{aligned}
$$

Notice that, by the law of total variance, $\mathbb{E}[(w_i^{\perp})^2|x_i] = \mathrm{Var}(w_i^{\perp}|x_i) + \mathbb{E}[w_i^{\perp}|x_i]^2$. Now, since $w_i^{\perp}$ is the component of w_i that is orthogonal to x_i and by the projection interpretation of the conditional variance, $\mathbb{E}[w_i^{\perp}|x_i] = 0$. And thus, since by assumption $\mathrm{Var}(w_i^{\perp}|x_i) = \mathrm{Var}(w_i^{\perp})$,

$$
\begin{aligned}
\mathbb{E}[(w_i^{\perp})^2|x_i] &= \mathrm{Var}(w_i^{\perp}|x_i) \\
&= \mathrm{Var}(w_i^{\perp}) \\
&= \mathbb{E}[(w_i^{\perp})^2] - \mathbb{E}[w_i^{\perp}]^2 \\
&= \mathbb{E}\left[\left(w_i - \frac{\gamma\sigma_w^2}{\sigma_x^2}(x_i - \mu_x)\right)^2\right] - \left(\underbrace{\mathbb{E}[w_i]}_{0} + \frac{\gamma\sigma_w^2}{\sigma_x^2}\underbrace{\mathbb{E}[x_i - \mu_x]}_{0}\right)^2 \\
&= \underbrace{\mathbb{E}[w_i^2]}_{\sigma_w^2} - 2\frac{\gamma\sigma_w^2}{\sigma_x^2}\left(\underbrace{\mathbb{E}[x_i w_i]}_{\gamma\sigma_w^2} - \mu_x\underbrace{\mathbb{E}[w_i]}_{0}\right) + \frac{\gamma^2\sigma_w^4}{\sigma_x^4}\underbrace{\mathbb{E}[(x_i - \mu_x)^2]}_{\sigma_x^2} \\
&= \sigma_w^2\left(1 - \frac{\gamma^2\sigma_w^2}{\sigma_x^2}\right)
\end{aligned}
$$

Note that this variance is well defined (positive) only if $\sigma_x^2 \geq \gamma^2\sigma_w^2$. Under this condition,

$$
\mathbb{E}[u^2_{\text{OVB},i}|\mathrm{x}_i] = \sigma_u^2 + \delta^2\sigma_w^2\left(1 - \frac{\gamma^2\sigma_w^2}{\sigma_x^2}\right) \tag{9}
$$

Thus, under our set of assumptions, $\mathbb{E}[u^2_{\text{OVB},i}|\mathrm{x}_i]$ does not depend on x_i and $\mathbb{E}[u^2_{\text{OVB},i}|\mathrm{x}_i] = \mathbb{E}[u^2_{\text{OVB},i}]$. We denote this quantity $\sigma^2_{u_{\text{OVB}}}$.

We can now compute the variance of the estimator $\hat{\boldsymbol{\beta}}_{\text{OVB}}$, noting that $\mathbb{E}[\mathrm{x}_i\mathrm{x}_i' u^2_{\text{OVB},i}] = \mathbb{E}[\mathrm{x}_i\mathrm{x}_i'\mathbb{E}[u^2_{\text{OVB},i}|\mathrm{x}_i]] = \mathbb{E}[\mathrm{x}_i\mathrm{x}_i'\sigma^2_{u_{\text{OVB}}}] = \sigma^2_{u_{\text{OVB}}}\mathbb{E}[\mathrm{x}_i\mathrm{x}_i']$. And thus $\mathbb{E}[\mathrm{x}_i\mathrm{x}_i']^{-1}\mathbb{E}[\mathrm{x}_i\mathrm{x}_i' u^2_{\text{OVB},i}]\mathbb{E}[\mathrm{x}_i\mathrm{x}_i']^{-1} = \sigma^2_{u_{\text{OVB}}}\mathbb{E}[\mathrm{x}_i\mathrm{x}_i']$.

Plugin this and equation 8 into equation 7, we get, for $\hat{\beta}_{\text{OVB}}$, the second component of $\hat{\boldsymbol{\beta}}_{\text{OVB}}$:

$$\hat{\beta}_{\text{OVB}} \xrightarrow{d} \mathcal{N}\left(\beta_1 + \frac{\delta\gamma\sigma_w^2}{\sigma_x^2},\ \frac{\sigma_u^2 + \delta^2\sigma_w^2\left(1 - \frac{\gamma^2\sigma_w^2}{\sigma_x^2}\right)}{n\ \sigma_x^2}\right)$$

Then, noting that $\rho_{xw} = \text{corr}(x, w) = \frac{\text{cov}(\mu_x + \gamma w + \epsilon, w)}{\sigma_x\sigma_w} = \frac{\gamma\sigma_w}{\sigma_x}$, we have:

$$\sigma^2_{\text{OVB}} = \text{avar}\left(\hat{\beta}_{\text{OVB}}\right) = \frac{\sigma_u^2 + \delta^2\sigma_w^2\left(1 - \rho_{xw}^2\right)}{n\ \sigma_x^2}$$

□

I. 3 Asymptotic distribution of $\hat{\beta}_{\text{CTRL}}$ (lemma 3)

Proof. The proof is the well know proof of the asymptotic distribution of the OLS. I simply compute $\mathbb{E}[x_{w,i}x'_{w,i}]^{-1}$ to retrieve the variance of the parameter of interest $\boldsymbol{\beta}_{\text{CTRL}}$. We know that we have:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{CTRL}} - \boldsymbol{\beta}_{\text{CTRL}}) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}[\mathrm{x}_{w,i}\mathrm{x}'_{w,i}]^{-1}\sigma_u^2\right)$$

We are interested in the second component of $\hat{\boldsymbol{\beta}}_{\text{CTRL}}$. To retrieve it we need to compute $\mathbb{E}[\mathrm{x}_{w,i}\mathrm{x}'_{w,i}]^{-1}$.

$$\mathbb{E}[\mathrm{x}_{w,i}\mathrm{x}'_{w,i}] = \mathbb{E}\begin{bmatrix} 1 & x_i & w \\ x_i & x_i^2 & x_i w_i \\ w_i & x_i w_i & w_i^2 \end{bmatrix} = \begin{bmatrix} 1 & \mu_x & 0 \\ \mu_x & \sigma_x^2 + \mu_x^2 & \gamma\sigma_w^2 \\ 0 & \gamma\sigma_w^2 & \sigma_w^2 \end{bmatrix}$$

Note that we have $\mathbb{E}[x_i w_i] = \mathbb{E}[x_i]\underbrace{\mathbb{E}[w_i]}_{0} + \text{cov}(x_i, w_i) = \gamma\underbrace{\text{var}(w_i)}_{\sigma_w^2} + \underbrace{\text{cov}(\epsilon_i, w_i)}_{0} = \gamma\sigma_w^2$.

Now, $\mathbb{E}[\mathrm{x}_{w,i}\mathrm{x}'_{w,i}]^{-1} = \dfrac{1}{\det(\mathbb{E}[\mathrm{x}_{w,i}\mathrm{x}'_{w,i}])}\ {}^t\mathrm{C}$ with C the comatrix of $\mathbb{E}[\mathrm{x}_{w,i}\mathrm{x}'_{w,i}]$. We have:

$$\det(\mathbb{E}[\mathrm{x}_{w,i}\mathrm{x}'_{w,i}]) = (\sigma_x^2 + \mu_x^2)\sigma_w^2 - \sigma_w^2\mu_x^2 - \gamma^2\sigma_w^4 = \sigma_w^2(\sigma_x^2 - \gamma^2\sigma_w^2)$$

and the "central" component of C, σ_w^2. Thus the central component of interest of $\mathbb{E}[\mathrm{x}_{w,i}\mathrm{x}'_{w,i}]^{-1}$ is $\frac{1}{\sigma_x^2 - \gamma^2\sigma_w^2}$. Therefore, for $\hat{\beta}_{\text{CTRL}}$, the second component of $\hat{\boldsymbol{\beta}}_{\text{CTRL}}$, we have:

$$\hat{\beta}_{\text{CTRL}} \xrightarrow{d} \mathcal{N}\left(\beta_1,\ \frac{\sigma_u^2}{n\ (\sigma_x^2 - \gamma^2\sigma_w^2)}\right) \tag{10}$$

Then, noting that $\rho_{xw} = \operatorname{corr}(x, w) = \frac{\operatorname{cov}(\mu_x+\gamma w+\epsilon, w)}{\sigma_x \sigma_w} = \frac{\gamma \sigma_w}{\sigma_x}$, we have:

$$\sigma^2_{\text{CTRL}} = \frac{\sigma_u^2}{n\ \sigma_x^2(1-\rho_{xw}^2)}$$

□

I. 4 Asymptotic distribution of $\hat{\beta}_{\mathbf{IV}}$ (lemma 4)

Proof. Since $u_{\text{IV}} = u_{\text{OVB}} = \delta w + u$, we have $\sigma^2_{u_{\text{IV}}} = \sigma_u^2 + \delta^2 \sigma_w^2$. Thus, the usual asymptotic distribution of the IV gives:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{IV}} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}\left(0, (\sigma_u^2 + \delta^2\sigma_w^2)\mathbb{E}[\mathrm{z}_i \mathrm{x}_i']^{-1}\mathbb{E}[\mathrm{z}_i \mathrm{z}_i']\left(\mathbb{E}[\mathrm{z}_i \mathrm{x}_i']^{-1}\right)'\right)$$

We are interested in the second component of $\hat{\boldsymbol{\beta}}_{\text{IV}}$. To retrieve it we need to compute $\mathbb{E}[\mathrm{z}_i \mathrm{z}_i]$, $\mathbb{E}[\mathrm{x}_i \mathrm{z}_i']^{-1}$ and its transpose.

$$\mathbb{E}[\mathrm{z}_i \mathrm{z}_i'] = \mathbb{E}\begin{bmatrix} 1 & z_i \\ z_i & z_i^2 \end{bmatrix} = \begin{bmatrix} 1 & \mu_z \\ \mu_z & \sigma_z^2 + \mu_z^2 \end{bmatrix}$$

$$\mathbb{E}[\mathrm{z}_i \mathrm{x}_i'] = \begin{bmatrix} 1 & \mathbb{E}[x_i] \\ \mathbb{E}[z_i] & \mathbb{E}[z_i x_i] \end{bmatrix} = \begin{bmatrix} 1 & \pi_0 + \pi_1 \mathbb{E}[z_i] + \gamma \underbrace{\mathbb{E}[w_i]}_{0} + \underbrace{\mathbb{E}[e_i]}_{0} \\ \mu_z & \pi_0 \mathbb{E}[z_i] + \pi_1 \mathbb{E}[z_i^2] + \gamma \underbrace{\mathbb{E}[z_i w_i]}_{0} + \underbrace{\mathbb{E}[z_i e_i]}_{0} \end{bmatrix} = \begin{bmatrix} 1 & \pi_0 + \pi_1 \mu_z \\ \mu_z & \pi_0 \mu_z + \pi_1(\sigma_z^2 + \mu_z^2) \end{bmatrix}$$

$$\Rightarrow \quad \mathbb{E}[\mathrm{z}_i \mathrm{x}_i']^{-1} = \frac{1}{\pi_1 \sigma_z^2}\begin{bmatrix} \pi_0 \mu_z + \pi_1(\sigma_z^2 + \mu_z^2) & -\pi_0 - \pi_1 \mu_z \\ -\mu_z & 1 \end{bmatrix}$$

Thus,

$$\mathbb{E}[\mathrm{z}_i \mathrm{x}_i']^{-1}\mathbb{E}[\mathrm{z}_i \mathrm{z}_i']\left(\mathbb{E}[\mathrm{z}_i \mathrm{x}_i']^{-1}\right)' = \frac{1}{\pi_1 \sigma_z^2}\begin{bmatrix} 2\pi_0 \mu_z + \pi_1(\sigma_z^2 + \mu_z^2) + \frac{\pi_0^2}{\pi_1} & -\mu_z - \frac{\pi_0}{\pi_1} \\ -\mu_z - \frac{\pi_0}{\pi_1} & \frac{1}{\pi_1} \end{bmatrix}$$

And so, for $\hat{\beta}_{\text{IV}}$, the second component of $\hat{\boldsymbol{\beta}}_{\text{IV}}$, we have:

$$\sqrt{n}\left(\hat{\beta}_{\text{IV}} - \beta_1\right) \xrightarrow{d} \mathcal{N}\left(0,\ \frac{\sigma_u^2 + \delta^2 \sigma_w^2}{n\ \pi_1^2 \sigma_z^2}\right) \tag{11}$$

Now, since $\rho_{xz} = \text{corr}(x_i, z_i) = \frac{\text{cov}(\pi_0+\pi_1 z_i+\gamma w_i+e_i,z_i)}{\sigma_x \sigma_z} = \pi_1 \frac{\sigma_z}{\sigma_x}$,

$$\sqrt{n}\left(\hat{\beta}_{\text{IV}} - \beta_1\right) \xrightarrow{d} \mathcal{N}\left(0,\ \frac{\sigma_u^2 + \delta^2 \sigma_w^2}{\sigma_x^2 \rho_{xz}^2}\right)$$

□

I. 5 Asymptotic distribution of $\hat{\beta}_{\text{RED}}$ (lemma 5)

Proof. The proof is straightforward: this is the usual univariate, unbiased case, with and error term equal to $(\delta + \beta_1\gamma)w_i + u_i + \beta_1 e_i$. Since w, u and ϵ_{RED} uncorrelated, its variance is $(\delta + \beta_1\gamma)^2\sigma_w^2 + \sigma_u^2 + \beta_1^2\sigma_e^2$. □

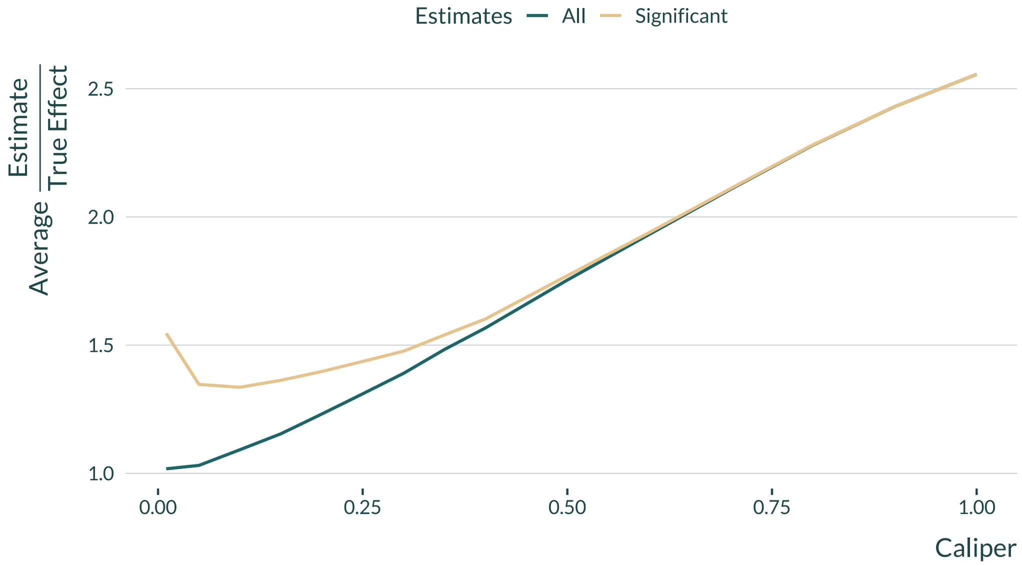
B Matching simulations

Intuition. Another approach to retrieve a causal effect in a situation of selection on observables is to use matching. This method defines "counterfactuals" for treated units by picking comparable units in the untreated pool. In the case of propensity score matching, treated units are matched to units that would have a similar predicted probability of taking the treatment, *i.e.* couple of units with a difference in propensity score lower than a critical value called the caliper. The smaller the caliper, the more comparable units have to be to be matched and therefore the lower the risk of confounding. Yet, with a stringent caliper, some units may not find a match and be pruned, decreasing the effective sample size. This can lead to a loss in statistical power and produce statistically significant estimates that are inflated. In the case of matching, the confounding-exaggeration trade-off is therefore mediated by the value of the caliper.

Case-study and simulation procedure. I illustrate this issue by simulating a labor training program where the treatment is not randomly allocated (Dehejia and Wahba 1999). Individuals self-select into the training program and may therefore have different characteristics from individuals who do not choose to enroll. To emulate this, I assume that the distribution of the propensity scores differ for treated and control groups: they are drawn from $\mathcal{N}(\mu_T, \sigma_T)$ and $\mathcal{N}(\mu_C, \sigma_C)$ respectively. This can be analogous to considering that matching is done based on the value of a unique covariate. Based on how these propensity scores are created, I define the potential monthly income of each individual i, under the treatment or not.

Based on this simulation framework, I generate 1000 datasets for each propensity score matching procedure with caliper values ranging from 0 to 1. Parameter values of the simulation are set to make them realistic and can be found here. Once units are matched, I simply regress the observed revenue on the treatment indicator.

Figure 9 – Evolution of Bias with the Caliper in Propensity Score Matching, Conditional on Statistical Significance.



Notes: The green line indicates the average bias for all estimates, regardless of their statistical significance. The beige line represents the inflation of statistically significant estimates at the 5% level. The caliper is expressed in standard deviation of the propensity score distribution. Details on the simulation are available at this link.

Results. Figure 9 indicates that the average bias of estimates, regardless of their statistical significance, decreases with the value of the caliper as units become more comparable. For large caliper values, units are not comparable enough and confoundings bias the effect. For small caliper values, they become comparable but the sample size becomes too small to allow for a precise estimation of the treatment effect and exaggeration arises. Statistically significant estimates never get close of the true effect. This imprecision, and thus exaggeration, results from the fact that the matching procedure does not use information on outcomes that would reduce the residual variance of the model but rather focuses on reducing bias arising from covariates imbalance (Rubin 2001).