Causal Exaggeration: Unconfounded but Inflated Causal Estimates

Vincent Bagilet *

November 28, 2023

Abstract

The credibility revolution in economics has made causal inference methods ubiquitous. Simultaneously, an increasing amount of evidence highlights that the literature strongly favors statistically significant results. I show that these two phenomena interact in a way that can substantially worsen the reliability of published estimates: while causal identification strategies alleviate bias caused by confounders, they reduce statistical power and can create another type of bias—exaggeration—when combined with selection on significance. This is consequential in fields such as environmental economics, as estimates turn into decision-making parameters for policy makers conducting cost-benefit analyses. I characterize this confounding-exaggeration trade-off theoretically and using realistic Monte Carlo simulations replicating prevailing identification strategies and document it in an example literature. I then discuss potential avenues to address this issue.

Link to the most recent version of the paper

*Columbia University, New York, USA. Email: vincent.bagilet@columbia.edu. 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, Clément De Chaisemartin, Jesse McDevitt-Irwin, 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, IPWSD, the Paris School of Economics and the Toulouse School of Economics.

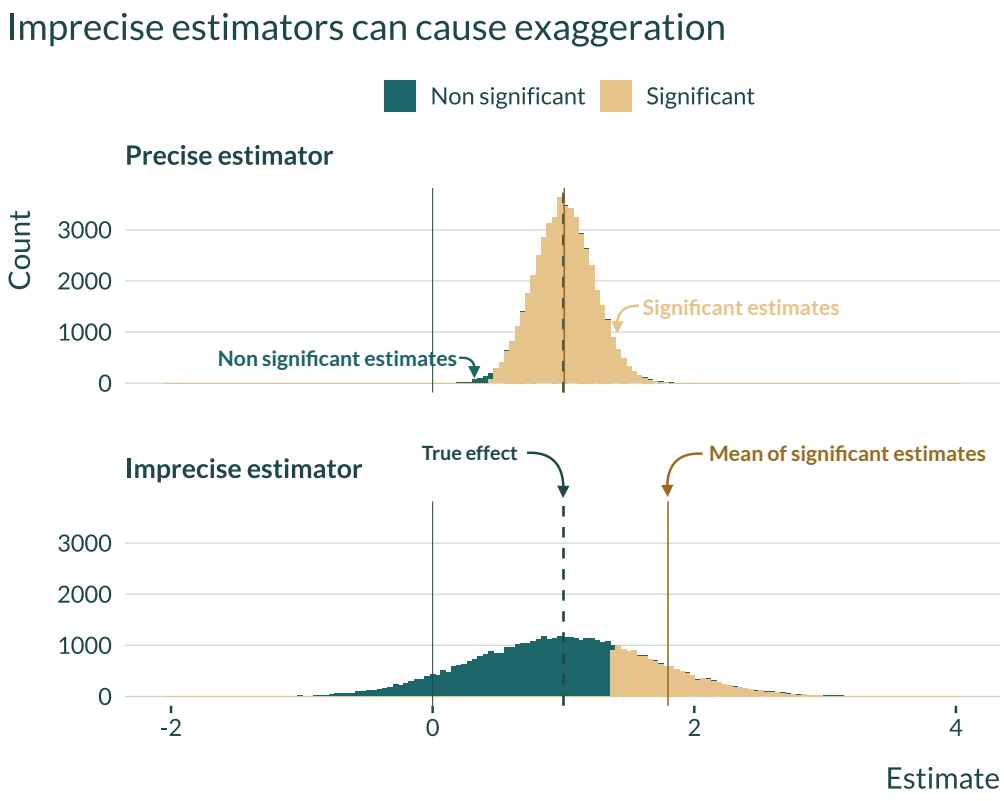
I Introduction

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

When statistical power is low, not only is the estimator imprecise but statistically significant estimates exaggerate the true effect size (Ioannidis 2008, Gelman and Carlin 2014, Lu et al. 2019, Zwet and Cator 2021). Only estimates at least 1.96 standard errors away from zero are statistically significant at the 5% level. In under-powered studies, these estimates make up a selected subsample 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 does not guarantee that it will be close to the true effect. An estimator $\hat{\beta}$ of the true effect β might be unbiased in the traditional sense of $\mathbb{E}[\hat{\beta}] = \beta$ but conditionally biased in the sense that $\mathbb{E}[\hat{\beta}|\text{ Significant}] \neq \beta$. For statistically significant estimates, the tension between statistical power and reducing confounding is thus a tension between reducing confounding and exaggerating the true effect size.

Yet, exaggeration only arises under two conditions: 1) a publication bias favors statistically significant results and 2) statistical power is low. A large body of literature has shown that the economics literature selects results based on statistical significance (Rosenthal 1979, 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 exaggeration. Ioannidis et al. (2017) finds that the median statistical power in a wide range of areas of economics

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 $\mathcal{N}(1, 0.5)$.

is no more than 18%. Despite the widespread use of convincing causal identification strategies and usually large sample sizes, they show that nearly 80% of estimates are likely exaggerated by a factor of two. In environmental economics, using a more conservative approach, Ferraro and Shukla (2020) finds that 56% of estimates are exaggerated by a factor of two or more. In the present paper, I provide evidence of exaggeration in a subfield of this literature, that on the acute health effects of air pollution. I further expand this analysis in a companion paper (Bagilet 2023). The magnitude of exaggeration is thus considerable and in some situations could be on par with that of a bias caused by confounders, as illustrated in section II. It is therefore crucial to take exaggeration into account and to understand its drivers.

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, as could be detrimental policy implications.

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

All causal 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, even if the initial sample size was large. 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 changes in treatment status. Approaches that do not actually leverage natural experiments but aim to identify a causal effect by controlling for confounders also limit the variation used. Matching prunes units that cannot be matched and thus reduces the effective sample size. Adding controls or fixed effects to the model can increase the variance of the estimator and exaggeration if they absorb more of the variation in the treatment than in the outcome variable.

Since causal identification strategies can be interpreted as ways of controlling for confounders, this last point actually ties all the strategy-specific arguments together. 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. When these identification strategies absorb more of the variation in the treatment than in the outcome, they increase the variance 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 $\sigma^2_{\hat{\beta}_{\text{OVB}}}$ and $\sigma^2_{\hat{\beta}_{\text{CTRL}}}$, the variance of the estimators for β when omitting w (short regression) and controlling for it (long regression) are respectively:

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

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, $\sigma^2_{y^{\perp x}}$ and $\sigma^2_{y^{\perp x,w}}$ are the variances of the parts of y that are orthogonal to x and to x and w respectively, σ^2_x is the variance of x and $\sigma^2_{x^{\perp w}}$ is the variance of the part of x orthogonal to w. Thus,

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

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

In the remainder of the paper, I first document the magnitude of the trade-off. To do so, I focus on an example literature, that of the short term health effects of air pollution. I further expand this analysis in a companion paper (Bagilet 2023). In the present paper, I first document evidence of selection on significance and exaggeration in causal studies from this literature. In particular, I show that less precise studies report larger standardized effect sizes. I then focus on IV designs as they allow for a within-study comparison between causal and non-causal estimates, the Two-Stage Least-Squares (TSLS) and the corresponding "naive" Ordinary Least Squares (OLS) respectively. For many studies in this literature, the IV both reduces precision, limits statistical power and yields much larger estimates. The TSLS estimators are often drastically less precise than the OLS ones, with a median ratio of their respective standard errors of 3.8. This is suggestive of exaggeration: by reducing precision, the causal identification strategies likely induce exaggeration and produce

biased estimates. Only 2% of the IV designs would retrieve an effect size equal to that of the "naive" OLS—at the conventional 80% power threshold. The median exaggeration factor would be of 4.5. For an example study from this literature, I show that the bias of the IV could be 3.1 times larger than that of the OLS and that exaggeration likely explains 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 show that it increases as the strength of the instruments decreases, the number of exogenous shocks decreases or when controlling for a confounder absorbs more of the variation in the treatment than in the outcome.

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

Finally, I discuss concrete avenues to address this causal exaggeration when carrying out a non-experimental study [1]. First, 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. For instance, the sensitivity analysis tools developed in Cinelli and Hazlett (2020) enable to assess how strong confounders would have to be to change the estimate of the treatment effect beyond a given level we are interested in. Then, considering the attention given to bias avoidance in the economics literature, I underline that making power central to non-experimental analyses, even after an effect has been found, would help limit bias caused by exaggeration. Prospective power simulations help identify the design parameters affecting power and exaggeration by approximating the data generating process (Gelman 2020, Black et al. 2021). Retrospective power calculations allow to evaluate whether a study would have enough power to confidently estimate a range of smaller but credible effect sizes (Gelman and Carlin 2014, Stommes et al. 2021). Focusing more specifically on the trade-off and its drivers, I present tools to

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

visualize the variation actually used for identification when using causal identification strategies. The companion website describes in details how such analyses can be implemented. Finally, I briefly discuss potential solutions to mitigate this trade-off.

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

Second, studies discussing the exaggeration of statistically significant estimates due to a lack of power usually do not investigate its determinants and focus on specific causal identification methods separately (Ioannidis et al. 2017, Schell et al. 2018, Ferraro and Shukla 2020, Black et al. 2021, Stommes et al. 2021, Young 2021). In a companion paper, I highlight tangible design parameters that can cause exaggeration for a wide range of empirical 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, and that can explain these issues: although each strategy does so through different means, in essence they discard part of the variation, thereby increasing the risks of exaggeration.

Third, this study contributes to the literature on 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 use of convincing causal identification methods in economics may not shield the field from potential replication threats.

In the following section, I document evidence of causal exaggeration in an example literature, that of the short-term health effects of air pollution. 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 in section V and conclude in section VI.

II Causal Exaggeration in an Example Literature

The trade-off presented in this paper only has concrete implications if the use of causal identification strategies yields substantial exaggeration, especially as compared to the amount of bias caused by confounders these methods allow to avoid. I therefore document the magnitude of ex-

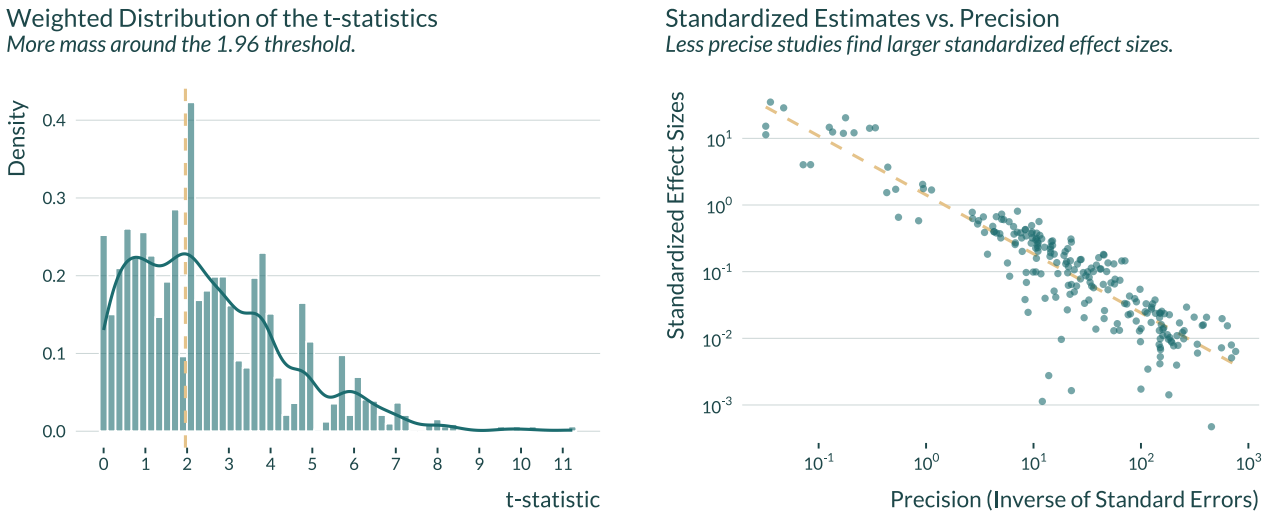
aggeration, comparing it to the bias of a standard regression model. To do so, I focus on the literature on the short term health effects of air pollution, one of the many literatures targeting small effects. The historical literature mostly relies on associations (Dominici and Zigler 2017, Bind 2019). Newly obtained results based on causal identification strategies confirm the adverse effects of air pollution in the short term (Schwartz et al. 2015; 2018, Deryugina et al. 2019). Yet, causal estimates are substantially larger than what would have been predicted by the standard epidemiology literature, some estimates being more than 10 times larger. Even within a given setting, my literature review shows that 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? Causal strategies could arguably remove omitted variable bias, reduce attenuation bias caused by classical measurement error in air pollution exposure or target a different causal estimand. But exaggeration could also explain part of this difference as studies on the short-term health effects of air pollution often display a low relative precision as a result of typically small effect sizes and relatively coarse data, at the city-day level (Peng et al. 2006, Peng and Dominici 2008).

In the present section, I review both the standard epidemiology and the causal inference literature. I built an algorithm based on regular expressions to retrieve estimates and confidence intervals from abstracts of the former literature and perform a manual review for the latter. Details on their implementation are available here. The final corpora are composed of 2155 estimates from 668 articles and 537 estimates from 36 articles respectively. I use them to show evidence of publication bias and exaggeration in this literature. I then quantify the later for the whole literature, focusing on differences between causal and standard approaches. Finally, for an example study, I compare the exaggeration of the causal approach to the confounding bias of the "naive" regression.

II. 1 Evidence of publication bias in this literature

Exaggeration only arises in the presence of publication bias. The left panel of Figure 2 reveals its presence in causal studies from this literature. Following the approach used in Brodeur et al. (2016; 2020), I show that there is an excess mass in the t-statistics distribution at the 5% statistical significance threshold. The right panel of Figure 2 produces further evidence of this favoring of significant estimates but also points to a consequence of this publication bias: published estimates from imprecise studies might be exaggerated. In this plot we observe that less precise studies display larger standardized effect sizes. If published estimates captured true effects, their standardized effect size should be independent of the precision of the study. This figure constitutes suggestive evidence of selection on significance and exaggeration in this literature.

Figure 2 – Suggestive Evidence of Publication Bias and Exaggeration in the Causal Inference Literature on Acute Health Effects of Air Pollution.



Notes: The sample in the left panel includes all 537 estimates reported in articles from the causal literature, including "naive" OLS estimates and placebo tests. Following Brodeur et al. (2020), the weights are equal to the inverse of the number of tests displayed in the same table multiplied by the inverse of the number of tables in the article. In the right panel we exclude the "naive" OLS estimates and placebo tests. Both axes are on a log10 scale. Limiting the sample to economics journal leaves the figures essentially unchanged (see supplemental material). Distinguishing between top 5 and other journals shows that even if there standardized effect sizes are typically smaller in top 5 journals, the same inverse relationship can be observed.

II. 2 Quantifying exaggeration

The exaggeration of an estimator corresponds to the expected value of the absolute value of significant estimates. As such, it depends on the true magnitude of the estimand of interest; this true effect determines the value the estimator is centered on. Exaggeration can therefore only be calculated by hypothesizing true effect sizes. Considering the wide variety of treatments and outcomes in the literature on the acute health effects of air pollution, there is a multitude of estimands and "true effects". To circumvent this limitation and since Ioannidis et al. (2017) and Ferraro and Shukla (2020) find a typical exaggeration of two in the economics literature, I first evaluate the proportion of studies that would have a design reliable enough to retrieve an effect size equal to half of the obtained estimate.

If the true effect size was equal to half of the obtained estimate, 58% of the standard epidemiology studies would have a power below the conventional 80% target. The median exaggeration factor would be 1.3. These figures however hide a lot of heterogeneity across studies. For one quarter of studies, the exaggeration would be larger than 1.9. This hypothesis on the true effect size, despite enabling to get an overview of this heterogenous literature, suffers from an important

limitation. It is particularly conservative: hypothesized effect sizes based on exaggerated estimates will be too large and will thus minimize exaggeration. For a more homogenous subset of this literature, I thus make more informed guesses about potential true effect sizes, using results from a meta-analysis. Shah et al. (2015) gathered 94 studies on the effects of several air pollutants on mortality and emergency admissions for stroke. I find that 63% of the studies in Shah et al. (2015) have a statistical power below 80% and that the median exaggeration ratio is 1.6.

The causal literature displays an even lower power: if the true effect size of each study was equal to half of the obtained estimate, the median power would be 33% and the median exaggeration ratio would be 1.7. Only 11% of studies would have a power greater than 80%. One quarter of the studies would, on average, exaggerate the true effect sizes by a factor greater than 2. Focusing on IV designs allows to make a less arbitrary intra-study comparison, comparing the 2SLS estimates to the "naive" OLS one. If conventional bias is limited and the true effect is equal to the naive estimate, the median power of the IVs would only be 8.4% and median exaggeration would reach 4.5. Only 2.0% of the IV designs would have a power of 80% to detect an effect size of the magnitude of the OLS estimate. Significant 2SLS estimates would be greatly biased. This can be explained by the fact that the IV drastically reduces precision; the median standard error of the 2SLS estimators is 3.8 times larger than the one of the corresponding OLS. This is suggestive of causal exaggeration: a large share of the IVs reduce precision and yield likely exaggerated estimates. [2]

II. 3 Illustration of the trade-off

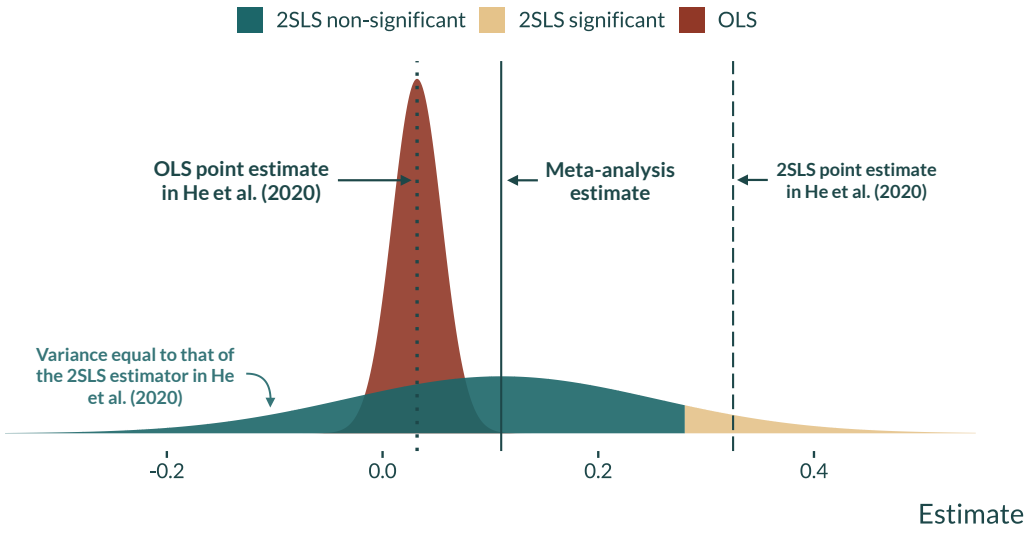
I then investigate causal exaggeration further, exploring whether the use of causal methods can actually induces bias through exaggeration. For an example study, He et al. (2020), I compare the bias of the "naive" OLS to that of the IV by computing their distance to an estimate of the "true effect" they 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 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.

2. Note that, as hinted by figure 2, some causal studies are more precise and unlikely to suffer from severe exaggeration issues.

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 statistical significant estimates. Figure 3 illustrates this point.

Figure 3 – 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.

The distribution of the 2SLS estimator represented in figure 3 assumes that the estimator is unbiased and thus centered on the meta-estimate found in Shah et al. (2015). The variance of this distribution corresponds to the variance of the 2SLS estimator found in He et al. (2020). Due to the lack of precision of this design, the statistically significant estimates are located in the tail of the distribution and 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. The distribution of the OLS estimator is the one obtained by the authors. I ignore exaggeration of the OLS for clarity but since the OLS is biased downward, inflating it would yield an estimate 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 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. This leads me to computing the asymptotic distributions of a series of estimators in order to prove their normality 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 deviation 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: [3]

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,*

3. All the proofs of the lemma and theorems are in appendix A.

— *E is a decreasing function of the Signal-to-Noise Ratio (SNR),* $\frac{\beta}{\sigma}$*, 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 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 III. 1, 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, we can show asymptotic normality and study how the variances of these estimators evolve with these parameters. 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.

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.*, $\mathrm{Var}(w_i^{\perp x} | x_i) = \mathrm{Var}(w_i^{\perp x})$. 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 $\rho_{xw} = \mathrm{corr}(x, w) = \frac{\gamma \sigma_w}{\sigma_x}$. 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

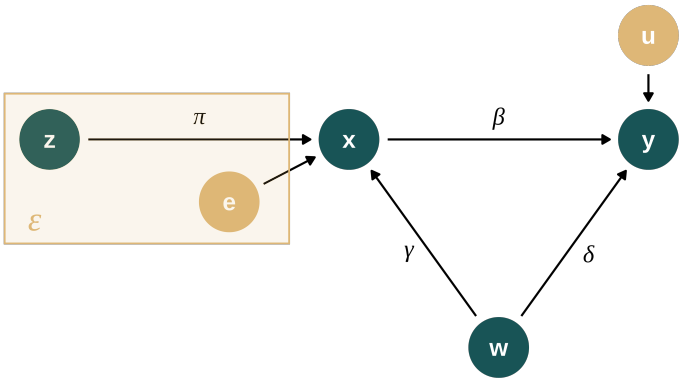
valid instrument does not affect the results in the controlled and OVB cases. Since the instrument is valid, it satisfies exogeneity, *ie* $\mathbb{E}[z_i u_i] = 0$ and $\mathbb{E}[z_i w_i] = 0$, relevance, *ie* $\text{rank}(\mathbb{E}[z_i x_i']) = 2$, and positive-definiteness of $\mathbb{E}[z_i z_i']$. 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, *ie* $\mathbb{E}[z_i e_i] = 0$ 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 4 represents the data generating process.

Figure 4 – 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*

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

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 ($\sigma^2_{y^{\perp x}}$).

Note that, 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:

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

This expression underlines that, for a given bias, σ^2_{OVB} does not vary with γ, or equivalently δ, the parameters of interest. Applying lemma III. 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 $\hat{\beta}_{CTRL}$ the OLS estimator of β_1 in the regression of y on x and w, $\hat{\beta}_{CTRL} \xrightarrow{d} \mathcal{N}(\beta_1,\ \sigma^2_{CTRL})$, with*

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

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

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

Since the numerator and denominator respectively increase and decrease with $|\gamma|$, σ^2_{CTRL} 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 $\lim_{\gamma^2 \to \frac{\sigma_x^2}{\sigma_w^2}} \sigma^2_{\text{CTRL}} = +\infty$.

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 $\hat{\boldsymbol{\beta}}_{\text{2SLS}} = \hat{\boldsymbol{\beta}}_{\text{IV}}$.

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*

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

Note that 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 $\hat{\beta}_{RED}$, the OLS estimator of the reduced form regression of y on z, $\hat{\beta}_{RED} \xrightarrow{d} \mathcal{N}(\beta_1,\ \sigma^2_{RED})$, with*

$$\sigma^2_{RED} = \frac{\sigma^2_y - \beta_1^2 \pi_1^2 \sigma^2_z}{n\ \sigma^2_z}$$

Note that 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^2_z, the larger σ^2_{RED}. In addition, $\lim_{\sigma_z \to 0} \sigma^2_{\text{RED}} = +\infty$.

In the binary case, $\sigma^2_z = p_1(1-p_1)$ 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^2_z 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 III. 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_{CTRL}, E_{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, $\forall\ b_{OVB}$,*

- *$\exists\ \gamma$ s.t. $E_{CTRL} > E_{OVB}$*
- *$\exists\ \pi_1$ s.t. $E_{IV} > E_{OVB}$*

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

For some parameter values, 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

For clarity, I split the simulations by identification strategy and build simulations that reproduce real-world examples from economics of education for RDD, labor economics for matching, political economy for IV, health economics for exogenous shocks and environmental economics for control and fixed effects approaches. Real-world settings enable to clearly grasp the relationships between the different variables and to set realistic parameter values. Since all these simulations have an illustrative purpose only, I intentionally focus on settings in which statistical power can be low. All the models are correctly specified and accurately represent the data generating process, except for the omitted variable bias (OVB).

For each identification strategy, I start by laying out how the method enables to retrieve a causal effect. It naturally points to the key parameter through which the confounding/exaggeration trade-off is mediated. I then briefly describe the example setting considered and the simulation assumptions. I finally display the simulation outputs and discuss the implications of the trade-off that are specific to the identification strategy considered. Detailed codes for simulation procedures are available on the project's website.

IV. 1 Controlling for confounders

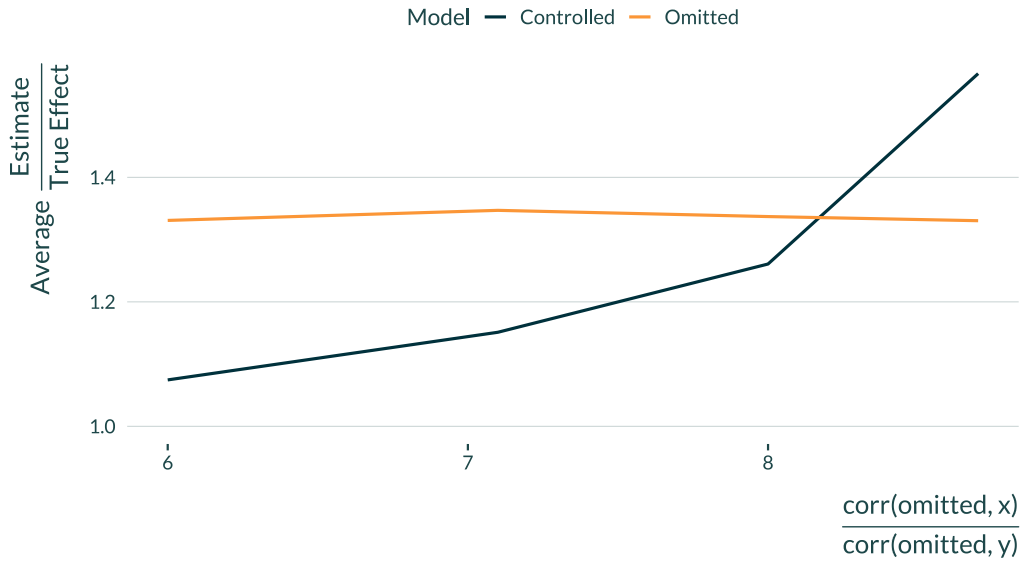
Intuition. To identify a causal effect and avoid the risk of confounders, an "ideal" approach would be to partial them out by directly controlling for them. However, as discussed in the introduction and section III, controlling for an additional variable may increase the variance of the estimator if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

Case-study and simulation procedure. To highlight this trade-off, I consider the extreme case in which we either perfectly control for confounders or do not control for them. I consider a simple setting, with one outcome variable y, one explanatory variable x and an omitted variable w. The data generating process is the same as described in equations 2 and 3. As in section III,

I reason at bias fixed and variances of the defined variables fixed, *i.e.* varying the share of the variance of x and y that is explained by w.

Results. Figure 5 displays the results of these simulations. The more the unobserved variable is linked to the explanatory variable of interest as compared to the outcome variable, *i.e.*, the larger the γ/δ ratio, the larger the exaggeration. When this ratio is large, controlling can cause exaggeration to become larger than the OVB plus exaggeration when the variable is omitted.

Figure 5 – Evolution of the Bias with the Correlation of the omitted variable with x and y, conditional on significance.



Notes: The blue line indicates the average bias for estimates from the control model that are statistically significant at the 5%. The orange line represents the bias of statistically significant estimates from the model with the omitted variable. In this simulation, N = 2,000. For now, the simulations is calibrated with arbitrary numbers but I will modify this in a later version of the project. Details on the simulation and calibration are available at this link.

IV. 2 Matching

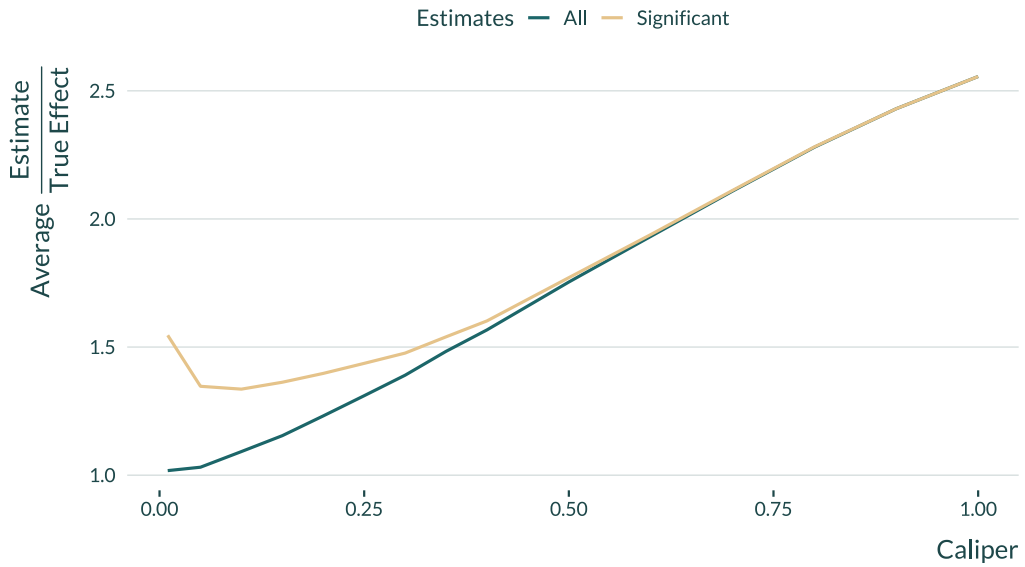
Intuition. Another approach to retrieve a causal effect in a situation of selection on observables is to use matching. This method defines "counterfactuals" for treated units by picking comparable units in the untreated pool. In the case of propensity score matching, treated units are matched to units 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 6 – 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 6 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).

IV. 3 Regression Discontinuity Design

Intuition. To identify a causal effect, a regression discontinuity approach also prunes units that cannot be deemed comparable enough to any units with the opposite treatment status. This method relies on the assumption that for values close to the threshold, treatment assignment is quasi-random. Under this assumption, individuals just below and just above the threshold would be comparable in terms of observed and unobserved characteristics, and only differ in their treatment status. To avoid confounding, the RDD focuses on observations within a certain bandwidth around the threshold and discards observations further away. The effective sample size where the identification of causal effect of the treatment is the most credible is thus smaller than the total sample size, leading to a lower precision. For this method, the confounding-exaggeration trade-off is therefore mediated by the size of the bandwidth.

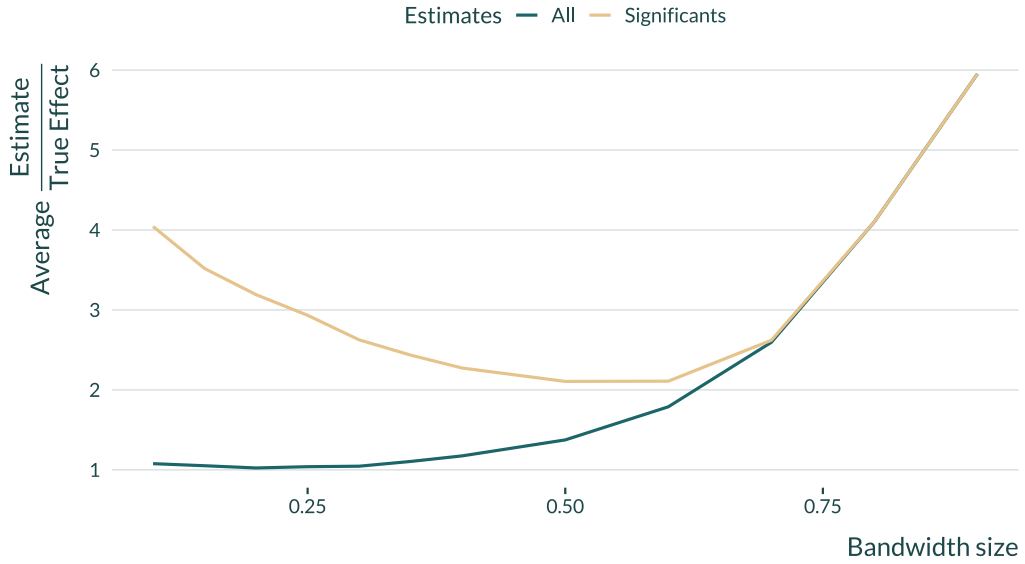
Case-study and simulation procedure. To illustrate this trade-off, I consider a standard application of the sharp RD design in economics of education in which students are offered additional lessons based on the score they obtained on a standardized test. Thistlethwaite and Campbell (1960) introduced 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 just below and just above this threshold.

The simulation framework for the RDD 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: $T_i = \mathbb{I}[Qual_i < C]$. Final scores are correlated with qualification score. Further assume that 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 final test score of student i is thus: $Final_i = \beta_0 + \beta_1 T_i + \eta Qual_i + \delta f(w_i) + u_i$, where f a non linear function (here cubic) and $u_i \sim \mathcal{N}(0, \sigma_u^2)$ random noise. β_1 is the causal parameter of interest.

To make the simulations realistic, I derive parameters values from statistics from the Department of Education and treatment effect sizes from a meta-analysis of RCTs in economics of education by Kraft (2020). Given these parameters values, I then generate 1000 datasets with 10,000 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.

Results. Figure 7 displays the results of these simulations. 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 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. As for matching, 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. For other methods such as IV or exogenous shocks, the issue is more dichotomized and the use of the causal identification strategy comes with a drawback.

Figure 7 – Evolution of the Bias with Bandwidth Size in Regression Discontinuity Design, conditional on 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. In this simulation, N = 10,000. The bandwidth size is expressed as the proportion of the total number of observations of the entire sample. Details on the simulation are available at this link.

IV. 4 Instrumental Variables Strategy

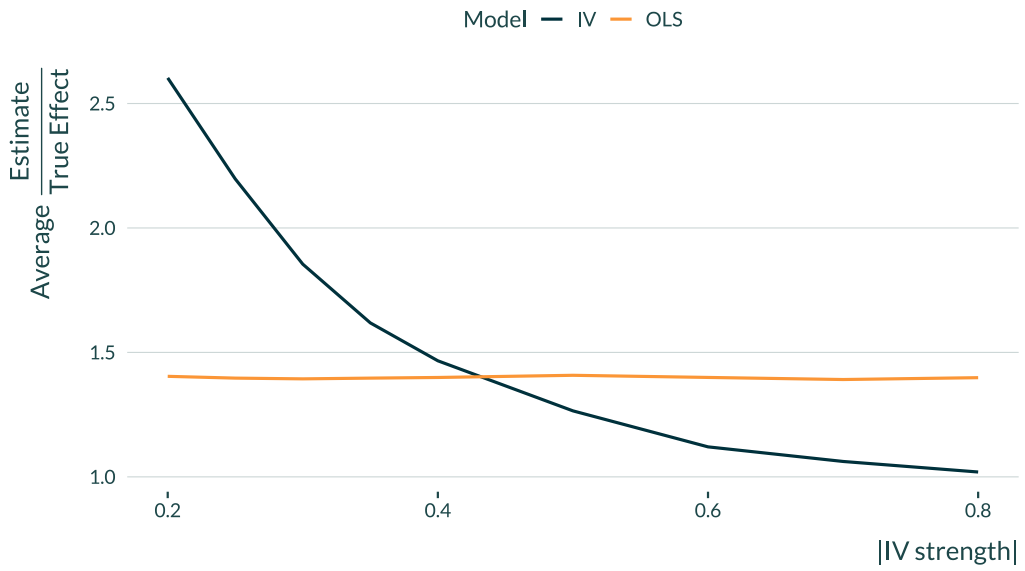
Intuition. Instrumental variables strategies overcome the issue of unobserved confounding by only considering exogenous variation in the treatment, *i.e.* the variation that is explained by the instrument. Even when this exogenous fraction of the variation is limited, the instrument can successfully eliminate confounding on average. However, 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, I focus on the impact of voter turnout on election results. To avoid the threat of confounding in this setting, studies often take advantage of exogenous factors such as rainfall that affect voter turnout. 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 derive parameters values from a set of existing studies using similar variables (Gomez et al. 2007, Fujiwara et al. 2016, Cooperman 2017). For each value of the IV strength considered, I create 1000 datasets. I run both a naive ordinary least squares model and a two-stage least squares model to estimate the impact of voter turnout on the vote share of a party.

Results. Figure 8 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 the naive regression model. Yet, when the IV strength decreases, the exaggeration of statistical 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.

Figure 8 – 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 this simulation, N = 10,000. Details on the simulation are available at this link.

IV. 5 Exogenous shocks

Intuition. To avoid confounding, strategies such as DiD and event studies take advantage of exogenous variation in the treatment status caused by exogenous shocks or events. 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. Statistical power is maximized when the proportion of treated observations is equal to the proportion of untreated ones, as extensively discussed in the randomized controlled trial literature. In studies using discrete exogenous shocks, a confounding-exaggeration trade-off is thus mediated by the number of observations treated. This issue does not only concern DiD event studies but is particularly salient in this case.

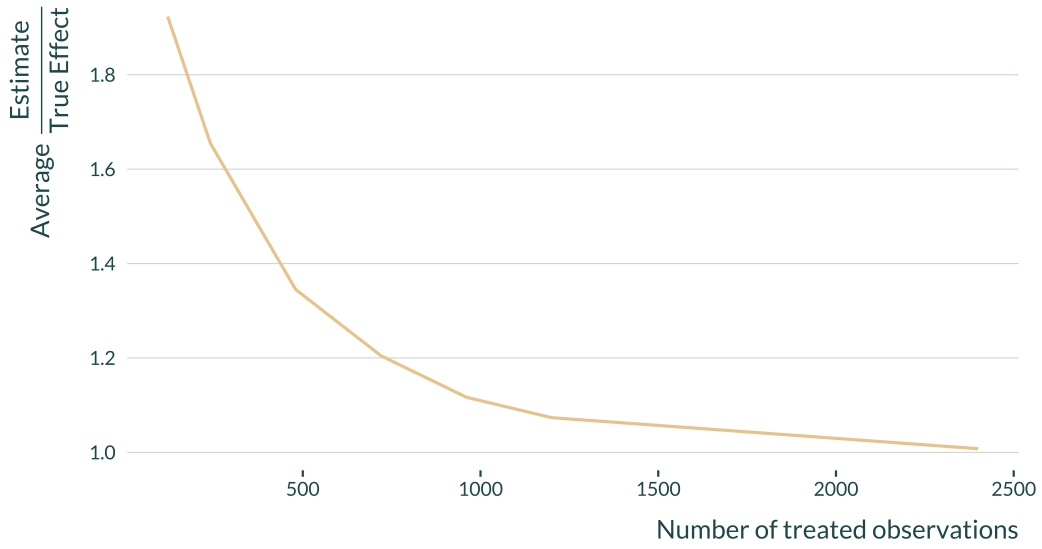
Case-study and simulation procedure. To illustrate this trade-off, I simulate a study of the impact of air pollution reduction on newborn weight of babies. To avoid confounding, one can exploit exogenous shocks to air pollution such as plant closures, creation of a low emission zone or of an urban toll. I simulate the 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 $T_{z,t}$, which is equal to one if a plant is closed in this period and 0 otherwise. The average birth weights bw_{zt} is defined as follows: $bw_{z,t} = \alpha + \beta T_{z,t} + \zeta_z + \tau_t + u_{z,t}$. 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.

The parameters values of the simulations are inspired from Currie et al. (2015) and Lavaine and Neidell (2017). For a fixed sample size of 120,000 observations, I generate 1000 datasets for an increasing number of treated observations and estimate the correct two-way fixed effects model.

Results. Figure 9 displays the results of these simulations. Even though the actual sample size is extremely large in the example, if the number of treated observations is small, exaggeration can be important. A very large number of observations does not necessarily prevent exaggeration to arise.

Figure 9 – Evolution of Bias With the Number of Treated Observations, for Statistically Significant Estimates, in the Exogenous Shocks Case



Notes: The line indicates the average bias for estimates that are statistically significant at the 5%. In this simulation, N = 120,000. Details on the simulation are available at this link.

V Navigating the trade-off

In the previous sections, I argued that that 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 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 to assess how strong confounders would have to be to change the estimate of the treatment effect beyond a given level we are interested in. 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". In addition, the authors implement graphical tools to facilitate this comparison. I suggest to use such quantitative bias analyses to evaluate the restrictiveness of the causal approach required to limit the threat of unobserved confounding to acceptable levels. In settings where bias caused by confounders is likely to be low, the

V. 2 Evaluating risks of 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-Hernandez 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 could also be reported very concisely 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 first simulate the design of the study (Hill 2011, Gelman 2020, Black et al. 2021). 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. In the replication material, I provide `R` code to be used as examples of how to run such simulations for most causal identification strategies. In situations where the relationships among covariates are too complex to emulate, one can also start from an existing dataset and add a known treatment effect. I implemented 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. Such *retrospective* calculations allow to evaluate 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. 2021).

I illustrate how a retrospective analysis works by taking the example of Card (1993) on the relationship between human capital and income. He 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. The associated standard error is 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. If the existing literature suggests that such effects are likely to be close close to a 10% increase in wage, we may wonder if the design in Card (1993) would allow to detect such an effect. We can thus compute the statistical power of the study under the hypothesis of a true effect size of 10% and for a precision of 5.5% equal to that obtained in Card (1993). [4] Statistically significant estimates (at the 5% level) would on average be roughly equal to 15%, therefore overestimating the true effect

4. Timm et al. (2019) and Linden (2019) offer `R` and `Stata` packages that enable to easily run these calculations through an extremely short command: `retrodesign(10, 5.5)`.

by a factor of 1.5. Statistical power, the proportion of estimates that are significant, would only be 44%. Conditional on a 10% true effect size being a reasonable assumption, this study 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).[5] When such information is not available, power calculations can be ran for a range of smaller but credible effect sizes. Such credible effect sizes can be derived from theoretical findings. It is also possible to evaluate whether the design of our study would be sufficient to detect smaller effects than the point estimate obtained.

V. 3 Driver of the trade-off

V. 3.1 Navigating the bias-variance trade-off

In non-experimental studies, estimator variance is often important to the extent that a large variance may lead to a failure to reject the null of zero effect when it is incorrect. Variance is paramount until a statistically significant estimate is obtained. Yet, exaggeration underlines that variance matters, even once a significant estimate has been obtained.

Obtaining a statistically significant estimate from an imprecise estimator should not necessarily be interpreted as a sign of "success" in getting significance despite a large confidence interval. It could instead be a warning that this estimate may come from the tails of the distribution and would thus inaccurately represent the true effect. Conditional on having obtained a statistically significant estimate, a limited precision can hide a bias: exaggeration. This invites to revisit the well-known bias-variance trade-off: a larger variance can also lead to a larger bias, even in (conditional) expectation. When combined with the existing statistical significance filter, the bias-variance trade-off is in fact a bias-bias trade-off. This paper thus urges us to pay attention to the implications of our design choices on the variance of our estimators, even if a large variance did not prevent us from obtaining a statistically significant estimate.

V. 3.2 The variation used for identifcation

Causal inference strategies only leverage a subset of the variation to avoid confounders. However, when this subset is too small, exaggeration arises. Identifying this variation (and the observations) actually used for estimation can help navigate the confounding-exaggeration trade-off.

5. Note that when such meta-analyses are available, one can use a Bayesian procedure to shrink statistically significant estimates based on the corpus of estimates from prior studies Zwet and Gelman (2021), Zwet and Cator (2021), Zwet et al. (2021) .

The measure I outline here both leverages the interpretation of causal inference methods as control approaches and builds on a procedure developed in Aronow and Samii (2016) for a different purpose: evaluating the external validity of standard regressions.

Aronow and Samii (2016) essentially interprets the estimate of the coefficient of the treatment of interest in a simple linear non-causal regression as a weighted average of individual treatment effects. The weight w_i of individual i is simply the squared difference between its treatment status T_i and the value of this treatment status as predicted by the other covariates X: $w_i = (T_i - \mathbb{E}[T_i|X_i])^2$. If treatment effects are heterogenous, the weighting may lead some observations to be disproportionally represented in the average effect of the treatment. In that case, the average of treatment is only representative of a subset of the individual treatments, leading to external validity issues.

The parallel with my setting directly follows from this interpretation regardless of whether the treatment is heterogenous or not: observations whose treatment status is well explained by covariates do not actually contribute to the estimation of the treatment effect. This may lead to a small *effective* sample size and to exaggeration. In the control approach to causal inference strategies, the more variation in the treatment is absorbed when "controlling" for confoundings, the smaller the effective sample, potentially leading to exaggeration. Aronow and Samii (2016) leverages the representativity of the effective sample for fear of external validity issues. I focus on its size for fear of exaggeration.

It might then seem compelling to define this effective sample by proposing a weight value under which the associated observation does not actually contribute to identification. Yet, considering the specificity of each analysis and that exaggeration depends on several factors, including the true effect size, I instead suggest to visualize the individual weights.[6] It allows to get a sense of where the variation comes from and which are the observations that actually contribute to the estimation. Since applied economic analyses often rely on panel data, I propose to use a heatmap as a base for visualization, with time on the x-axis and individuals on the y-axis. If the data is geographical, one can directly plot the weights on a map.

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, even under-powered studies would on average recover the true effect, as long as the estimator is unbiased. The exaggeration issue could

6. In future developments of this project, I will however develop a measure of this effective sample size. I may also need to modify the weights formula in order to account for the variation in y that is absorbed when controlling.

therefore be addressed by tackling publication bias.

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). Forks appear at various stages along the path of research, for instance in data preparation, regarding the inclusion of a given cofntrol variable 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 the end goal of research.

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 to build 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 2021). 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. Confidence intervals will be wide in under-powered studies signaling 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, estimates published from underpowered studies exaggerate true effect sizes, even when the estimators are "unbiased" in the usual sense of $\mathbb{E}[\hat{\beta}] = \beta$ (Ioannidis 2008, Gelman and Carlin 2014, Lu et al. 2019, Zwet and Cator 2021). It is therefore not surprising that many estimates published in economics have been shown to be considerably exaggerated (Camerer et al. 2016, Ioannidis et al. 2017), despite the extensive use of convincing causal inference methods. However, determinants for these exaggeration and power issues have remained understudied. I argue that exaggeration is exacerbated by the foundational component of causal inference: the fact that it only leverages subsets of the variation. Although

causal methods enable to avoid 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. A systematic reporting of statistical power calculations and analysis of the variation actually used for identification could help to avoid falling into this exaggeration trap.

References

Abadie, A. (2020), 'Statistical Nonsignificance in Empirical Economics', American Economic Review: Insights **2**(2), 193–208. 2, 30

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. 30

Andrews, I. and Kasy, M. (2019), 'Identification of and Correction for Publication Bias', American Economic Review **109**(8), 2766–2794. 2, 30

Angrist, J. D. and Pischke, J.-S. (2009), Mostly Harmless Econometrics: An Empiricist's Companion, 1 edition edn, Princeton University Press, Princeton. 27

Angrist, J. D. and Pischke, J.-S. (2014), Mastering 'Metrics: The Path from Cause to Effect, Princeton University Press. 27

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

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

Bagilet, V. (2023), 'Accurately Estimating Relatively Small Effects: Air Pollution and Health'. 3, 5, 7, 27

Bind, M.-A. (2019), 'Causal Modeling in Environmental Health', Annual Review of Public Health **40**(1), 23–43. 8

Black, B. S., Hollingsworth, A., Nunes, L. and Simon, K. I. (2021), Simulated Power Analyses for Observational Studies: An Application to the Affordable Care Act Medicaid Expansion, SSRN Scholarly Paper ID 3368187, Social Science Research Network, Rochester, NY. 6, 7, 27

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, 8, 9, 30

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. 8

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

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, 30

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

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

Christensen, G. and Miguel, E. (2018), 'Transparency, Reproducibility, and the Credibility of Economics Research', Journal of Economic Literature **56**(3), 920–980. 7, 30

Cinelli, C. and Hazlett, C. (2020), 'Making sense of sensitivity: Extending omitted variable bias', Journal of the Royal Statistical Society: Series B (Statistical Methodology) **82**(1), 39–67. 6, 26

Cooperman, A. D. (2017), 'Randomization Inference with Rainfall Data: Using Historical Weather Patterns for Variance Estimation', Political Analysis **25**(3), 277–288. 23

Cunningham, S. (2021), Causal Inference: The Mixtape, Yale University Press. 27

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. 25

Deaton, A. and Cartwright, N. (2018), 'Understanding and misunderstanding randomized controlled trials', Social Science & Medicine **210**, 2–21. 7

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. 20

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. 8, 10, 11, 12

Dominici, F. and Zigler, C. (2017), 'Best Practices for Gauging Evidence of Causality in Air Pollution Epidemiology', American Journal of Epidemiology **186**(12), 1303–1309. 8

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, 3, 7, 9

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. 23

Gelman, A. (2020), Regression and Other Stories, Cambridge University Press, Cambridge New York, NY Port Melbourne, VIC New Delhi Singapore. 6, 27

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, 30

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'. 30

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. 23

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, 4

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. 4

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, 12

Hernán, M. A. (2021), 'Causal analyses of existing databases: No power calculations required', Journal of Clinical Epidemiology p. S0895435621002730. 30

Hernán, M. A. and Robins, J. M. (2020), Causal Inference: What If, boca raton: chapman & hall/crc edn. 7

Hill, J. L. (2011), 'Bayesian Nonparametric Modeling for Causal Inference', Journal of Computational and Graphical Statistics **20**(1), 217–240. 27

Huntington-Klein, N. (2021), The Effect: An Introduction to Research Design and Causality, 1 edn, Chapman and Hall/CRC, Boca Raton. 27

Imbens, G. and Kalyanaraman, K. (2012), 'Optimal Bandwidth Choice for the Regression Discontinuity Estimator', The Review of Economic Studies **79**(3), 933–959. 7, 22

Imbens, G. W. and Rubin, D. B. (2015), Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction, Cambridge University Press, Cambridge. 27

Ioannidis, J. P. A. (2008), 'Why Most Discovered True Associations Are Inflated', Epidemiology **19**(5), 640–648. 2, 30

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, 7, 9, 27, 30

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, 30

Kraft, M. A. (2020), 'Interpreting Effect Sizes of Education Interventions', Educational Researcher **49**(4), 241–253. 22

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. 25

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

Lu, J., Qiu, Y. and Deng, A. (2019), 'A note on Type S/M errors in hypothesis testing', British Journal of Mathematical and Statistical Psychology **72**(1), 1–17. 2, 12, 30, 38

McConnell, B. and Vera-Hernandez, M. (2015), Going beyond simple sample size calculations: A practitioner's guide, Technical report, Institute for Fiscal Studies. 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. 30

Middleton, J. A., Scott, M. A., Diakow, R. and Hill, J. L. (2016), 'Bias Amplification and Bias Unmasking', Political Analysis **24**(3), 307–323. 26

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

Oster, E. (2019), 'Unobservable Selection and Coefficient Stability: Theory and Evidence', Journal of Business & Economic Statistics **37**(2), 187–204. 26

Peng, R. D. and Dominici, F. (2008), Statistical Methods for Environmental Epidemiology with R: A Case Study in Air Pollution and Health, 2008th edition edn, Springer, New York ; London. 8

Peng, R. D., Dominici, F. and Louis, T. A. (2006), 'Model Choice in Time Series Studies of Air Pollution and Mortality', Journal of the Royal Statistical Society Series A: Statistics in Society **169**(2), 179–203. 8

Ravallion, M. (2020), Should the Randomistas (Continue to) Rule?, Working Paper 27554, National Bureau of Economic Research. 7

Romer, D. (2020), 'In Praise of Confidence Intervals', AEA Papers and Proceedings **110**, 55–60. 30

Rosenbaum, P. R. (2002), Observational Studies, Springer Series in Statistics, Springer New York, New York, NY. 26

Rosenthal, R. (1979), 'The file drawer problem and tolerance for null results', Psychological Bulletin **86**(3), 638–641. 2, 30

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. 21

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

Schwartz, J., Austin, E., Bind, M.-A., Zanobetti, A. and Koutrakis, P. (2015), 'Estimating Causal Associations of Fine Particles With Daily Deaths in Boston', American Journal of Epidemiology **182**(7), 644–650. 8

Schwartz, J., Fong, K. and Zanobetti, A. (2018), 'A National Multicity Analysis of the Causal Effect of Local Pollution, NO2, and PM2.5 on Mortality', Environmental Health Perspectives **126**(8), 087004. 8

Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002), Experimental and Quasi-experimental Designs for Generalized Causal Inference, Houghton Mifflin. 27, 30

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. 30

Stommes, D., Aronow, P. M. and Sävje, F. (2021), 'On the reliability of published findings using the regression discontinuity design in political science', arXiv:2109.14526 [stat] . 6, 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. 21

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

Weidmann, B. and Miratrix, L. (2021), 'Lurking Inferential Monsters? Quantifying Selection Bias in Evaluations of School Programs', Journal of Policy Analysis and Management **40**(3), 964–986.

Young, A. (2021), 'Leverage, Heteroskedasticity and Instrumental Variables in Practical Application', p. 43. 7

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. 28

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. 28

Zwet, E. W. and Cator, E. A. (2021), 'The significance filter, the winner's curse and the need to shrink', Statistica Neerlandica **75**(4), 437–452. 2, 12, 28, 30, 38

A Mathematical proofs

I. 1 Variation of the exaggeration ratio (lemma III. 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:

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

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 $\mathbb{E}[\mathrm{x}_i\mathrm{x}_i' u_{\text{OVB},i}^2]$.

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

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

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

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

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

Notice that, by the law of total variance, $\mathbb{E}[(w_i^\perp)^2|x_i] = \text{Var}(w_i^\perp|x_i) + \mathbb{E}[w_i^\perp|x_i]^2$. 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 $\text{Var}(w_i^\perp|x_i) = \text{Var}(w_i^\perp)$,

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

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

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

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

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

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:

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

□

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:

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

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

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

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

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

and the "central" component of C, σ_w^2. Thus the central component of interest of $\mathbb{E}[\text{x}_{w,i}\text{x}'_{w,i}]^{-1}$ 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 $\rho_{xw} = \text{corr}(x, w) = \frac{\text{cov}(\mu_x + \gamma w + \epsilon, w)}{\sigma_x \sigma_w} = \frac{\gamma \sigma_w}{\sigma_x}$, we have:

$$\sigma^2_{\text{CTRL}} = \frac{\sigma^2_u}{n\ \sigma^2_x(1 - \rho^2_{xw})}$$

□

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

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

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

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.

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

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

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

Thus,

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

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

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

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 $(\delta + \beta_1\gamma)^2\sigma_w^2 + \sigma_u^2 + \beta_1^2\sigma_e^2$. □