Essays on Design of Applied Economics Studies

Vincent Bagilet

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2024

Abstract

Essays on Design of Applied Economics Studies

Vincent Bagilet

Applied economics studies target effects that can be relatively small. This dissertation delves into some statistical obstacles to the accurate estimation of such effects, with a particular focus on the concepts of statistical power and exaggeration—imprecise studies tend to produce inflated estimates of the effect of interest. It explores implications of low power and exaggeration that are specific to applied economics studies and their design. Through the example of studies on the acute health effects of air pollution, the first chapter identifies tangible drivers of exaggeration that extend beyond small effects and a limited sample size. This analysis uncovers an overarching mechanism, studied in Chapter 2, that induces exaggeration when using causal identification strategies. This subsequent chapter emphasizes that causal approaches only focus on a subset of the variation—the exogenous part—reducing the precision of the study and increasing risks of exaggeration. The final chapter further broadens the discussion to analyze design choices in light of the multiple goals of causal inference studies; these studies aim not only to identify an average effect but also differentiated effects across subgroups, as well as producing insights that extend beyond the population considered. Overall, this dissertation underlines the manifold implications of design choices on non-experimental economic studies, with the aim of contributing to more accurate estimations of effects to better inform policymaking.

Table of Contents

List of Figures

List of Tables

Acknowledgements

This section might be the most challenging one to write in my entire dissertation. Although I am not particularly at ease using vehement language, I want to convey how grateful I am to the many people who supported and helped me throughout this PhD. To more accurately capture the intended tone of this section, please read it with a filter enhancing the "thankfulness" of my language (or just ask ChatGPT to rewrite this section; it will certainly be tremendously emphatic).

First and foremost, I would like to thank the entire SusDev community—students, faculty, and staff—for making the whole PhD process enjoyable, challenging, instructive, and engaging.

I extend my heartfelt gratitude to Jeffrey Shrader for his incredible advice, support, and mentorship that enabled me to craft all the projects I undertook during this PhD. Most, if not all, of my projects and their ramifications stemmed from advice you gave me or directions you pointed me towards. Your enthusiasm and energy, as well as your incredible responsiveness and availability were instrumental in keeping me motivated. I am also grateful to Suresh Naidu and Geoffrey Heal for their always judicious comments and advice that significantly improved my work, as well as to Andrew Gelman, John Mutter, and Jeannette Stingone for being part of my committee. I would also like to thank the faculty from the SusDev program for their insightful comments, teaching, and guidance, Tomara Aldrich for her absolutely outstanding assistance with so many aspects of the PhD journey, and Linus Mattauch,

Sylvain Chabé-Ferret, and Emeric Henry for giving me both the opportunity to visit their home institutions and valuable research advice.

I am also grateful to the students in the program for being such a welcoming, supportive, and thought-provoking crowd, with a special note for SLab members. This PhD journey introduced me to extremely bright people who took time and energy to help me in my research and life around the PhD, and who became wonderful friends. As such, I would like to express my gratitude to Anouch, Clara, Claire, Florian, Georgia, Jesse, Marguerite, Nicolas, Pierre, Soraya, Thomas, Tristan, and Zara for the invaluable advice, discussions, support, but also for making New York such a vibrant city to live in, particularly by accompanying me to so many concerts, pizza and ramen places. Together with those from my advisors, your comments amounted to almost the entirety of my dissertation and left me with the only task of arranging them together. I would like to extend a special thanks to Léo. You not only brilliantly helped me build our research projects but also pushed me, kept me motivated, and believed in my work, often more than I did myself. You were and are always available to help, chat, share pastries, great data viz, memes, or nerdy papers. I also extend these sentiments to my other co-authors, most of whom are already mentioned in this section.

Although it may seem like I am thankful to the entire world, I want to use this platform to acknowledge the importance of the tools I use daily, which have helped me conduct better research. I am therefore grateful to anyone who participated in the development of platforms such as R, Posit and associated packages (Tidyverse, distill, fixest, retrodesign, and many others), Github, Zotero, Notion, Feedly, DataWrapper, and Data Is Plural.

Finally, my friends from France were fantastic co-workers during the many peculiar months spent there, away from colleagues. I am immensely grateful to my parents, my sister, and Kenza for supporting me, being there for me, and keeping up with my ever-changing moods.

Thanks to all of you, this PhD has been a fun, challenging, instructive, and overall fantastic experience.

Introduction

My main overall research goal is to study individual and collective environmental attitudes, the factors shaping them and their subsequent influence on consumption behavior, participation in social movements, or demand for policy change. Understanding these determinants and impacts is crucial to address the ecological crises we currently face. The relative priority given to ecological questions in both individual and collective decision making is pivotal to the implementation of the massive societal transformations or stringent policies these crises call for. However, the effects associated with environmental attitudes are often causally diffuse and therefore challenging to capture. As a consequence, in my Ph.D., I undertook the series of projects presented in this dissertation to explore some of the statistical obstacles to the accurate estimation of relatively small effects and to equip myself with tools to rigorously capture effects of environmental attitudes.

This dissertation focuses low statistical power—roughly a low probability of detecting an effect when there is one—and in particular on one of its consequences: exaggeration. Estimates that are statistically significant at the 5% level are located at least 1.96 standard errors away from 0. For relatively imprecise estimators, these estimates are not only far from zero but also from the true effect. They exaggerate the true effect size, even if the estimator is unbiased. This would not be an issue if a vast literature had not shown that significant estimates are favored for publication in economics. This dissertation examines some of the impacts of exaggeration, also called type-M error, a concept introduced by

Gelman and Tuerlinckx (2000) in statistics, on non-experimental economic studies. It focuses on implications for and interactions with settings and methods that are specific to this literature. Overall, this dissertation underlines that their typically large nominal samples do not shield non-experimental economic studies from exaggeration as design choices—i.e., choices regarding the data to consider—and modelling decisions can lead to small *effective* samples.

More specifically, in the first chapter of this dissertation, *Accurate Estimation of Small Effects: Illustration Through Air Pollution and Health*, I highlight tangible drivers of exaggeration that go beyond sample size and effect size. To do so, I use the literature on the acute health effects of air pollution as an illustration. I first document evidence of exaggeration in both the causal inference and the standard epidemiology literature, as well as of its two ingredients—publication bias and low statistical power. Building on actual air pollution and mortality data, I then implement simulations that replicate most prevailing causal and non-causal empirical designs to study the drivers of exaggeration. I show that the *effective* sample size, the number of exogenous shocks, the average count of the outcome or the strength of instruments for air pollution can create exaggeration and thus bias. Despite the data being specific to air pollution and health, these results extend to most other settings where effects are relatively small. Finally, I provide a principled workflow to evaluate the risks of exaggeration when carrying out a non-experimental study.

In *Causal Exaggeration: Unconfounded but Inflated Causal Estimates*, I take a step back and identify an overarching mechanism, core to non-experimental economic studies, that can partly explain the replication failures observed in economics, despite the use of convincing causal identification strategies. I argue that using these strategies can come at a cost: although they allow avoiding bias generated by confounders, they can generate another type of bias, exaggeration. Using a mathematical derivation and realistic Monte Carlo simulations, I show that their use rests on a trade-off between confounding and exaggeration and study its drivers in the case of prevailing strategies: IV, RDD and strategies that rely on fixed

effects or that leverage exogenous shocks. Causal identification strategies inherently increase risks of exaggeration as they only use a subset of the variation–the arguably exogenous part. Limiting the variation used for identification decreases the precision of the resulting estimator. When precision is limited and significant results favored, even if the distribution of the causal estimator is centered around the true effect, the distribution we actually draw estimates from may not be. Consequently, if a causal identification strategy removes too much variation to alleviate the risks of confounding, resulting significant estimates might considerably differ from the true effect.

The third and last chapter, *Design of Empirical Studies Given Their Multiple Goals*, discusses how statistical power and exaggeration interact with the multiple goals of causal inference studies. In these studies, we not only aim to estimate "the" average treatment effect, but also to analyze its variations across individuals and time, assess its impact on multiple outcomes and their conclusions are often generalized to other populations. Anticipation of these goals can change the way we approach study design. Through a series of examples, we discuss both the consequences of not taking them into account and some of the design alterations they imply. These consequences are often related to low statistical power and exaggeration, e.g., with heterogeneity in treatment effects, we need a greater precision to estimate these effects through interactions for instance. We then advocate making substantively-motivated assumptions about both effect sizes and variations in light of the goals of the study, anticipating and allowing for both uncertainty and heterogeneity in effect sizes. Finally, by suggesting a comprehensive workflow and providing accompanying code, we propose practical tools for identifying and addressing design concerns through systematic design calculations and simulations.

To sum up, this dissertation aims to underline some of the implications of design choices in non-experimental economic studies that go beyond concerns ensuring a quasi-random allocation of the treatment. Chapter 1 focuses on concrete parameters affecting design. The second chapter discusses how both modelling and design aspects concerned with guaranteeing

the quasi-random treatment allocation interact with other dimensions of design. The last chapter discusses how the multiple goals of applied studies invite us to modify how we approach design.

Chapter 1: Accurate Estimation of Small Effects: Illustration Through Air Pollution and Health

Abstract

This paper identifies tangible design parameters that might lead to inaccurate estimates of relatively small effects. Low statistical power not only makes relatively small effects difficult to detect but resulting published estimates also exaggerate true effect sizes. Through the case of the literature on the short-term health effects of air pollution, we document the prevalence of this issue and identify its drivers using real data simulations replicating most prevailing identification strategies used in economics. While the analysis builds on a specific literature, it draws out insights that expand beyond this setting. Finally, we discuss approaches to evaluate and avoid exaggeration when conducting a non-experimental study.

1.1 Introduction

Applied economics studies routinely target effects that are relatively small and thus difficult to capture. Studies focusing on such effects are inherently prone to low statistical power; they have a low probability of rejecting the null when it is false. When power is low—or equivalently, precision relatively limited—significant estimates exaggerate true effect sizes (Gelman and Tuerlinckx 2000; Ioannidis 2008; Gelman and Carlin 2014). In imprecise settings, only large effects are significant, *e.g.*, at least 1.96 standard errors away from 0 at the 5% significance level. Such effects are located in the tails of the distribution and do not capture the true effect. When significant results are favored—as it is the case in economics (Andrews and Kasy 2019; Abadie 2020; Brodeur, Cook, and Heyes 2020)—relatively imprecise studies will lead to exaggerated published estimates. This exaggeration is both consequential and widespread in the discipline; nearly 80% of the estimates in a wide array of economics literatures are likely exaggerated by a factor of two (Ioannidis, Stanley, and Doucouliagos 2017). The present paper aims at identifying tangible parameters that drive this exaggeration. It leans on the literature on the short-term health effects of air pollution to explore this matter. While many applied economics literatures target relatively small effects and therefore face a similar challenge, this literature constitutes a particularly well suited setting to study it and to yield insights for other contexts.

The literature on the acute health effects of air pollution is vast, spans multiple disciplines and directly informs policies addressing a crucial public health issue. An evaluation of its reliability and of its characteristics is critical in itself but its analysis also allows producing far-reaching insights applicable to other contexts. Studies in this literature focus on a wide array of settings, involving large variations in sample sizes, underlying effect sizes due to the diversity of populations and types of treatments considered, and in models and identification strategies leveraged. This heterogeneity enables the evaluation, within a single framework, of the impact of a large set of parameters and thus allows consideration of cases regularly

encountered in applied economics. The analysis developed in this paper focuses more on general design characteristics than on aspects specific to studies on the acute health effects of air pollution.

While exaggeration has been discussed and documented, especially in the statistics literature, some of its concrete drivers that extend beyond sample and effect sizes or that are specific to economic approaches remain understudied. We first introduce exaggeration and its link to statistical power and selection on significance. We then show that exaggeration is likely widespread and of non-negligible magnitude by documenting it in the literature of interest. We gather 2692 estimates from a unique corpus of 668 articles based on associations and of 36 articles that rely on causal inference methods. For each of these studies, we run calculations to assess whether the design of the study would allow to accurately capture the true effect if it was smaller than the observed estimate (Gelman and Carlin 2014; Ioannidis, Stanley, and Doucouliagos 2017; Lu, Qiu, and Deng 2019; Timm 2019). However, such calculations do not highlight the causes of exaggeration. Using real data from the US National Morbidity, Mortality, and Air Pollution Study (Samet et al. 2000), we therefore implement simulations replicating most prevailing inference strategies to identify the characteristics of research designs that drive exaggeration. Finally, we discuss a principled workflow to evaluate the risks of exaggeration when carrying out a non-experimental study.

Our literature review results suggest that a substantial share of estimates of the acute health effects of air pollution published in epidemiology and economics could be inflated. To run such calculations, we need to hypothesize true effects sizes. Given that the true effect is never observed, one cannot presume that the obtained estimate differs from it. Reassuringly, for a reasonable share of studies, it does not seem to be the case. However, if the true effect size was equal to half of the obtained estimate, respectively 58% and 89% of the standard epidemiology and causal inference designs would have a power below the conventional 80% target. The median exaggeration factors would be 1.3 and 1.7 respectively. For one quarter of studies, the exaggeration would be larger than 1.9. The assumption regarding the effect size

is based on the fact that Ioannidis, Stanley, and Doucouliagos (2017) and Ferraro and Shukla (2020) find that half of the estimates published in economics and environmental economics are inflated by a factor of at least two. It might thus be reasonable to expect study designs to allow detecting effects of this magnitude. We could not rely on meta-analyses to build better assumptions of true effect sizes because of the wide variety of treatments and outcomes considered in this literature.

The simulation results enable identifying concrete causes of exaggeration. While our simulations are tuned to study the acute health effects of air pollution, their conclusions likely extend to many other literatures. The intuitions behind the impact of each driver can be applied to most settings, even outside health or air pollution studies. In the context studied, we first show that, as expected, exaggeration increases when the sample size decreases. Importantly, we find that for all identification strategies, exaggeration can arise even for large sample sizes. Second, the simulations confirm that the smaller the effect targeted, the larger exaggeration is. They also show that when effect size is small, exaggeration can be substantial. Third, we find using rare exogenous shocks can produce greatly inflated estimates. The number of shocks can represent less than 1% of the observations for some studies leveraging public transportation strikes or thermal inversions, leading to large exaggeration ratios even when sample and true effect sizes are large. Similarly, substantial exaggeration can arise when the instrument only explains a limited portion of the variation in air pollution and that, even when F-statistics are large[1]. Finally, we show that the count of cases of the outcome is a key driver of exaggeration. Estimated effects of air pollution on the elderly or children can be exaggerated due to the small number of daily hospital admissions or deaths for these groups.

This paper makes three main contributions. First, it contributes to a growing literature assessing statistical power and exaggeration issues in various fields (Ioannidis 2008; Gel-

[1]Since this paper focuses on actual implementation of non-experimental studies, it mostly documents tangible drivers. We analyze a theoretical underlying mechanism of exaggeration specific to causal identification strategies in a companion paper presented in Chapter 2 (Bagilet 2023b).

man and Carlin 2014; Ioannidis, Stanley, and Doucouliagos 2017; Ferraro and Shukla 2020; Stommes, Aronow, and Sävje 2023; Arel-Bundock et al. 2022). Existing meta-analyses show that the economics literature is plagued with serious power issues but do not usually discuss the determinants of this lack of power. We overcome this key limitation by coupling our literature review with simulations. Outside of meta-analyses, the drivers of exaggeration in non-experimental studies also remain understudied. To our knowledge, only three papers thoroughly address this critical question (Schell, Griffin, and Morral 2018; Griffin et al. 2021; Black et al. 2022). We complement these studies focusing on difference-in-differences event-study designs by studying the drivers of exaggeration in a wide array of research designs: standard regression, reduced-form, instrumental variable and regression discontinuity design. We identify design parameters that might drive exaggeration in our literature of interest but also in any applied analysis and invite the reader to pay attention to these tangible characteristics in their own work. We expect power, effective sample size, effect size, the number of shocks, the strength of the instrument or the distribution of the outcome to drive the statistical power of most studies, creating important bias in resulting estimates.

Second, this paper contributes to a literature discussing the replicability and credibility of empirical findings in non-experimental studies (Button et al. 2013a; Open Science Collaboration 2015; Camerer et al. 2018; Brodeur, Cook, and Heyes 2020, for instance). We strive to put statistical power at the center of non-experimental analyses; a lack of it can lead to inaccurate published estimates. Well-powered studies on the other hand do not lead to substantial exaggeration, even in the presence of publication bias. We thus provide a reproducible workflow to evaluate and avoid exaggeration issues when running a non-experimental study. It invites to build simulations using fake data or existing datasets before carrying out a study to identify potential exaggeration and its sources. Once the analysis is completed, we advocate running a retrospective power analysis to assess whether the design used would have accurately recovered the true effect if it was in fact smaller than the one estimated, but within a range of reasonable effect sizes. We also propose to report these resulting power

calculations. To ease the adoption of this workflow, we used literate programming to make all replication and supplementary materials accessible, *via* the project's website. We also make the algorithm we developed to automatically review the epidemiology literature readily available to almost instantaneously evaluate publication bias and exaggeration issues in other fields reporting point estimates and confidence intervals in plain text.

Finally, this paper contributes to the literature on acute health effects of air pollution and more generally on the impacts of air pollution by underlining that some studies—but not all—likely suffer from exaggeration issues. We document the presence of a publication bias in this literature and discuss how research design parameters that are specific to this literature can cause exaggeration.

In the following section, we implement a simple simulation exercise to show why statistically significant estimates exaggerate true effect sizes when studies have low statistical power. In section 1.3, we present our retrospective analysis of the literature of interest. In section 1.4, we detail our simulation procedure to replicate empirical strategies. We display the simulation results in section 1.5 and provide specific guidance to avoid exaggeration when running a non-experimental study in section 1.6.

1.2 Background on Statistical Power and Exaggeration

Gelman and Tuerlinckx (2000) and Gelman and Carlin (2014) point out that statistically significant estimates suffer from a winner's curse in under-powered studies. These estimates can largely exaggerate true effect sizes or can even be of the opposite sign. In this section, we illustrate these two seemingly counter-intuitive issues through a simple simulation exercise. For clarity of exposition, this exercise builds on a short-term health effect of air pollution example but can easily translated to more general settings.

1.2.1 Illustrative Example

Let's simulate an "ideal" experiment in which a mad scientist is able to randomly increase the concentration of fine particulate matter ($PM_{2.5}$) to estimate the short-term effects of air pollution on daily non-accidental mortality. The experiment takes place in a major city over the 366 days of a leap year. The scientist increases the concentration of particulate matter by 10 µg/m^3—a large shock equivalent to a one standard deviation increase in the concentration of $PM_{2.5}$. Concretely, the scientist implements a complete experiment where they randomly allocate half of the days to the treatment group and the other half to the control group. They then measure the treatment effect of the intervention by computing the average difference in means between treated and control outcomes. They find a treatment effect of 4 additional deaths that is statistically significant at the 5% level. The statistical significance of the estimate fulfills the scientist expectations.

Contrary to the scientist, we know the true effect of the experiment since we created the data. Table 1.1 displays the pair of potential outcomes of each day, $Y_i(T_i = 0)$ and $Y_i(T_i = 1)$. $Y_i(T_i)$ represents the daily count of non-accidental deaths and T_i the treatment assignment, equal to 1 when unit i is treated and 0 otherwise. We first simulate the daily non-accidental mortality counts in the absence of treatment (i.e., the $Y(0)$ column of Table 1.1), by drawing 366 observations from a negative binomial distribution with a mean of 106 and a variance of 402. We choose these parameters to approximate the distribution of non-accidental mortality counts in a large European city. We then define the counterfactual distribution of mortality by adding the treatment effect, drawn from a Poisson distribution (i.e., the $Y(1)$ column of Table 1.1). We choose its parameter to increase the number of death by 1 on average[2].

Following the fundamental problem of causal inference, the daily count of deaths the scientist observes is given by the equation: $Y_i^{\text{obs}} = T_i \times Y_i(1) + (1 - T_i) \times Y_i(0)$. Considering that

[2]In relative terms, the treatment effect size we set represents a 1% increase in the health outcome. The magnitude of this hypothetical effect is larger than the one found in a recent and large-scale study based on 625 cities. Liu et al. (2019) estimated that a 10 µg/m^3 increase in $PM_{2.5}$ concentration was associated with a 0.68% (95% CI, 0.59 to 0.77) relative increase in daily all-causes mortality.

Table 1.1: Science Table of the Experiment.

Day Index	$Y_i(0)$	$Y_i(1)$	τ_i	T_i	Y_i^{obs}
1	122	124	+2	1	124
2	94	96	+2	1	96
3	96	98	+2	0	96
⋮	⋮	⋮	⋮	⋮	
364	96	97	+1	0	96
365	98	98	+0	0	98
366	143	144	+1	1	144

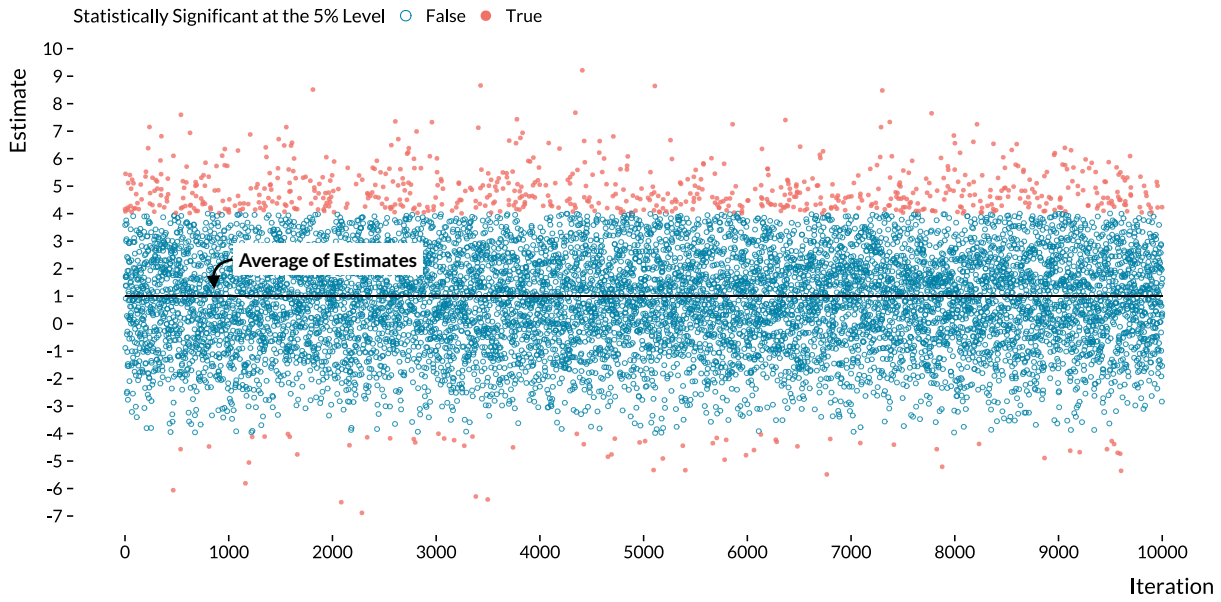
Notes: This table displays the potential outcomes, the unit-level treatment effect, the treatment status and the observed daily number of non-accidental deaths for 6 of the 366 daily units in the scientist's experiment.

the assignment of the treatment was random, how can the statistically significant estimate found by the scientist be 4 times larger than the true treatment effect size? Replicating the experiment a large number of times explains this apparently puzzling result.

1.2.2 Defining Statistical Power, Exaggeration Ratio and Type S Error

Figure 1.1 plots the estimates of 10,000 iterations of the experiment. Even if there is a large variation in the effect size of estimates, their average is reassuringly equal to the true treatment effect of 1 additional death. We can however see that estimates close to the true effect size would not be statistically significant at the 5% level. In a world without publication bias, several replications of this experiment would recover the true treatment effect. Unfortunately, despite recent changes in scientific practices and editorial policies, non-statistically significant estimates and replication exercises are still not valued enough (Brodeur, Cook, and Heyes 2020). In a world with publication bias, statistically significant estimates are more likely to be made public. Out of the 10,000 simulation estimates, about 800 are statistically significant at the 5% level. The *statistical power* of the experiment, which is the probability to reject the null hypothesis when there is actually an effect, is

Figure 1.1: Replicating 10,000 Times the Experiment.



Notes: Each dot represents a point estimates of one of the 10,000 iterations of the randomized experiment ran by the mad scientist. Red dots are statistically significant at the 5% level while blue dots are not. The black solid line represents the average of estimates, equal to the true average effect of 1 additional death.

equal to 8%. The scientist was therefore lucky to get a statistically significant estimate.

With such a low statistical power, statistically significant estimates are however not informative of the treatment of interest. Two metrics, the average type M (magnitude) error and the probability to make a type S (sign) error help assess the negative consequences of a lack of statistical power. The exaggeration ratio, or expected type M error, is defined as the ratio of the absolute values of the statistically significant estimates over the true effect size (Gelman and Carlin 2014). In the present case, with a statistical power of 8%, the scientist could expect their statistically significant estimates to be inflated on average by a factor of 5. We also notice in Figure 1.1 that a non-negligible fraction of statistically significant estimates are of the wrong sign: this proportion is the probability of making a type S error (Gelman and Tuerlinckx 2000). In this experiment, a statistically significant estimate has a 8% probability of being of the wrong sign.

Formally, the statistical power of a test is the probability of rejecting the null hypothesis $H_0 : \beta = 0$, where β is the true effect of the estimand of interest. For $\hat{\beta}$, a normally

distributed unbiased estimate of β with standard error σ, the power of the null hypothesis test at the 5% level is equal to $\Phi\left(-1.96 - \frac{\beta}{\sigma}\right) + 1 - \Phi\left(1.96 - \frac{\beta}{\sigma}\right)$, where Φ is the cumulative distribution function of the standard normal distribution. It increases with β, the true value of the effect and with the precision of the estimate, *i.e.*, when σ decreases. The exaggeration ratio is $\mathbb{E}\left(\frac{|\hat{\beta}|}{|\beta|} \,\middle|\, \beta, \sigma, |\hat{\beta}|/\sigma > 1.96\right)$ and the probability to make a type S error is given by $\Pr\left(\frac{\hat{\beta}}{\beta} < 0 \,\middle|\, \beta, \sigma, |\hat{\beta}|/\sigma > 1.96\right)$. Zwet and Cator (2021) and Lu, Qiu, and Deng (2019) derive closed-form expressions for these quantities. They show that both the exaggeration ratio and the probability of type S error decrease with β and the precision of the estimate and thus with statistical power.

To obtain statistically significant estimates that are informative of the true value of the effect size, the scientist would need to improve the design of their study in order to increase its statistical power.

1.3 Retrospective Analysis of the Literature

From extreme events such as the London Fog of 1952 to the development of sophisticated time-series analyses, a vast epidemiology literature of more than 600 studies has established that air pollution induces adverse health effects on the very short-term. Increases in the concentration of several ambient air pollutants have been found to be associated with increases in daily mortality and emergency admissions for respiratory and cardiovascular causes (Schwartz 1994; Samet et al. 2000; Le Tertre et al. 2002; Bell, Samet, and Dominici 2004; Liu et al. 2019). With this objective in mind, researchers in economics and epidemiology have recently used causal inference methods to improve on the standard epidemiology literature that relied on associations (Dominici and Zigler 2017; Bind 2019). Newly obtained results confirm the short-term health effects of air pollution (Schwartz et al. 2015; Schwartz, Fong, and Zanobetti 2018; Deryugina et al. 2019). Based on these results, environmental protection and public health agencies have designed policies such as air quality alerts to mitigate the burden of air pollution. Accurate estimates of these effects are therefore crucial as they

are directly used to implement and update policies to address this major public issue.

This section reviews this literature, through the prism of statistical power and exaggeration. It describes how we ran retrospective analyses of the causal inference and standard epidemiology literatures as well as steps we undertook to make this analysis easily reproducible in other literatures.

1.3.1 Our Approach

The formulas for power, exaggeration ratio and type S error described in the previous section all depend on the true magnitude of the estimand of interest. The true effect is however never observed in a given study. We can overcome this limitation with a retrospective power analysis. Essentially, it addresses the following question: would the design of our study be reliable enough to retrieve the true effect if it was in fact smaller than the obtained estimate? A retrospective power analysis can be considered as a thought-experiment in which we would exactly replicate the study many times under the assumption that the true effect is different from the observed estimate. The reasoning is analogous to the analysis in the previous section. Concretely, Gelman and Carlin (2014) propose to run simulations in which we draw many estimates from the asymptotic distribution of the estimator, a normal distribution with mean equal to the hypothesized true effect and a standard deviation equal to the standard error we obtained in the study. The statistical power is the proportion of sampled estimates that are statistically significant at the 5% level. The exaggeration ratio is computed as the average ratio of the absolute values of statistically significant estimates over the assumed true effect size. The probability to make a type S error is the proportion of significant estimates that are of the opposite sign of the true value. In this project, we use the **R** package `retrodesign` developed by Timm (2019) that implements the closed-form analogue of these simulations (Lu, Qiu, and Deng 2019).

To get a general overview of power issues in the causal inference and standard epidemiology literatures, we first carry out a simple retrospective analysis for each study. These

computations rely on hypothesized true effect sizes. Yet, since treatments and outcomes vary between studies in this literature, it is not possible to make general aggregated assumptions on true effect sizes. We need to consider specific hypothesized true effect sizes for each study. Since Ioannidis, Stanley, and Doucouliagos (2017) and Ferraro and Shukla (2020) find a typical exaggeration of two in the economics literature, we evaluate the proportion of studies that would have a design reliable enough to retrieve an effect size equal to half of the obtained estimate. On average, by what factor would statistically significant estimates be exaggerated? A well-designed study should be able to detect a range of plausible effect sizes that are smaller than the observed estimate. This method is however by no means ideal but offers some sort of consistency across studies. To overcome this limitation, for a subset of studies, we also use results from a meta-analysis as potential values for the true effect sizes.

1.3.2 Causal Inference Literature

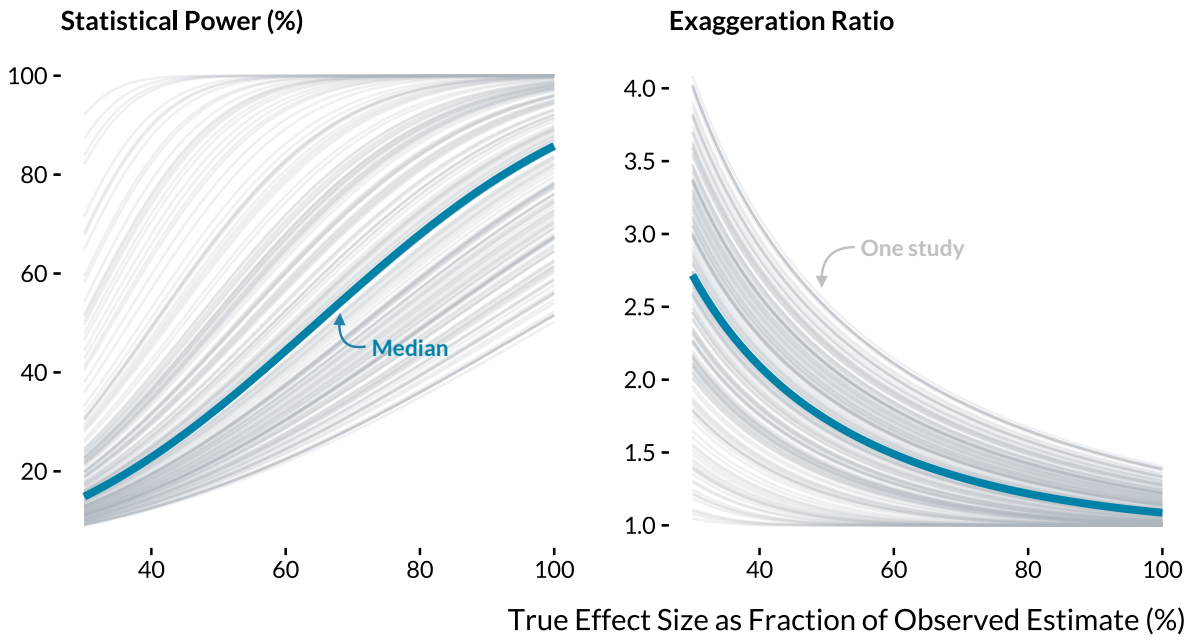
For the causal inference literature, an extensive search strategy on Google Scholar, IDEAS, and PubMed enables retrieving studies that (i) focus on the short-term health effects of air pollution on mortality and morbidity outcomes, and (ii) rely on a causal inference methods[3]. Appendix A.1 displays the list of the 36 articles that match the search criteria. For each study, we manually retrieve the method used by the authors, the health outcome and air pollutant they consider, the point estimate and the standard error of the main specifications.

To evaluate potential statistical power issues in this literature, we follow the same approach as for the analysis of the standard epidemiology literature. Figure 1.2 plots the power and exaggeration curves for 186 specifications which results are statistically significant at the 5% level. 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%. Figure 1.2 also shows that there is

[3]The very recent literature on the effects of air pollution on COVID-19 health outcomes is excluded to gather a relatively homogeneous corpus of studies.

a wide heterogeneity in statistical power issues among studies. Some of them are relatively well powered while others can run into large exaggeration issues. For instance, one quarter of studies would, on average, exaggerate the true effect sizes by a factor greater than 2. This pattern may help explain the very large effect sizes sometimes observed in the causal inference literature.

Figure 1.2: Statistical Power and Exaggeration Curves of Causal Inference Studies.

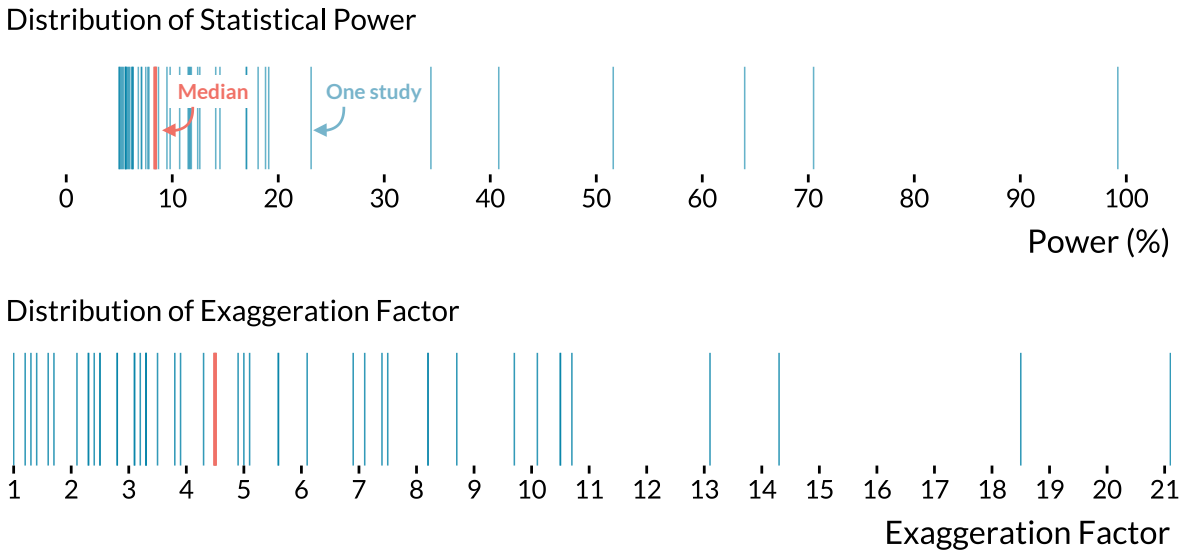


Notes: Each gray line is a power curve or an exaggeration curve of a statistically significant result published in the causal inference literature. The blue lines are the median values. For visual clarity, we drop results for which exaggeration ratios were too large.

Then, for the 49 instrumental variable results that are statistically significant and reporting the corresponding naive regression results, we can evaluate whether the 2SLS specifications would recover a true effect closer to that of the naive estimate. Figure 1.3 displays the distribution of the resulting statistical power and the average exaggeration ratio. The median power is equal to 8.4%. This results in very large exaggeration ratios: half of the studies would exaggerate a true effect of the size of the OLS estimate by a factor of at least 4.5. Such an inflation of statistically significant estimates could explain part of the gap between the standard and causal literature. This discrepancy could also be explained by a

combination of omitted variable bias and attenuation bias caused by classical measurement error in air pollution exposure. It could also come from the fact that the causal estimands targeted by both strategies are different when treatment effects are heterogeneous. Such explanations are not mutually exclusive and the lack of power and inability of the instrumental variables to recover smaller effect sizes remain concerning. In the presence of publication bias, considerable lack of power mechanically causes substantial exaggeration issues.

Figure 1.3: Distribution of Power and Exaggeration Ratio for Instrument Variable Designs, Assuming that the Naive OLS Estimates Are the True Effect Sizes.

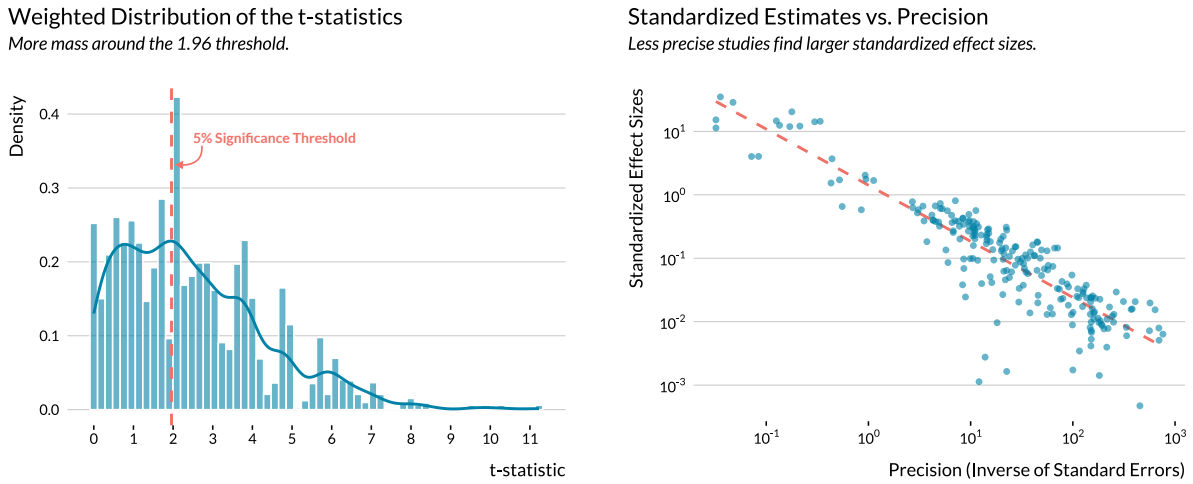


Notes: For 49 statistically significant 2SLS estimates, we define the true values of effect size as the corresponding OLS estimates. Each blue line represents either the statistical power (%) or the exaggeration factor of a study's result. Orange lines are the median of the two metrics. For visual clarity, we do not display three extreme exaggeration ratios.

Exaggeration arises when significant results are favored. The left panel of Figure 1.4 reveals its presence for the causal literature. As in Brodeur et al. (2016) and Brodeur, Cook, and Heyes (2020) for the broader economics literature, there is an excess mass in the t-statistics distribution at the 5% statistical significance threshold. The right panel of Figure 1.4 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 both selection on significance and exaggeration in this literature.

Figure 1.4: 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, Cook, and Heyes (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.

1.3.3 Standard Epidemiology Literature

Hundreds of papers have been published on the short-term health effects of air pollution in epidemiology, medicine and public health journals. A large fraction of articles rely on Poisson generalized additive models, which allow flexibly adjusting for the temporal trend of health outcomes and for non-linear effects of weather parameters. This literature spans over 20 years and has replicated analyses in a large number of settings, providing crucial insights on the acute health effect of air pollution.

To gather a corpus of relevant articles, we use the following search query on PubMed and

Scopus:

```
'TITLE(("air pollution" OR "air quality" OR "particulate matter" OR "ozone"', 'OR
"nitrogen dioxide" OR "sulfur dioxide" OR "PM10" OR "PM2.5" OR', ' "carbon dioxide"
OR "carbon monoxide")', 'AND ("emergency" OR "mortality" OR "stroke" OR "cerebrovascular"
OR', '"cardiovascular" OR "death" OR "hospitalization")', 'AND NOT ("long term"
OR "long-term")) AND "short term"'
```

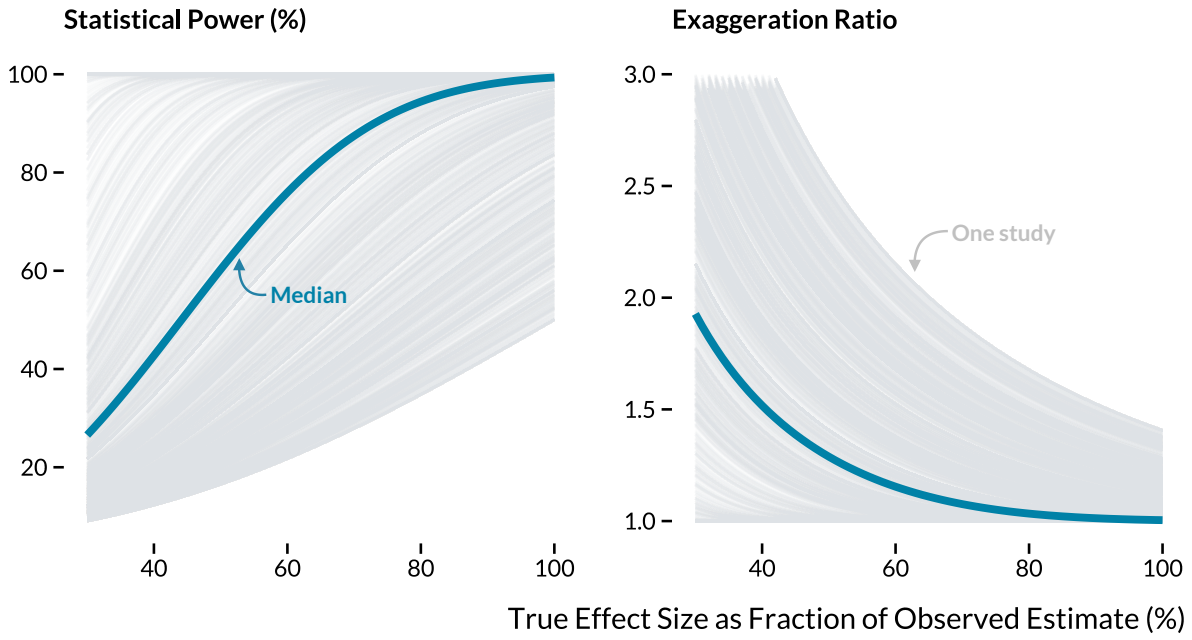
We retrieve the abstracts of 1834 articles. Then, we extract estimates and confidence intervals from these abstracts using regular expressions (regex). Our algorithm available online detects phrases such as "95% confidence interval (CI)" or "95% CI" and looks for numbers directly before this phrase or after and in a confidence interval-like format. We illustrate the outcome of this procedure (in blue) using one sentence of a randomly selected article from this literature review (Vichit-Vadakan, Vajanapoom, and Ostro 2008):

"The excess risk for non-accidental mortality was 1.3% [95% confidence interval (CI), 0.8-1.7] per 10 $\mu g/m^3$ of PM10, with higher excess risks for cardiovascular and above age 65 mortality of 1.9% (95% CI, 0.8-3.0) and 1.5% (95% CI, 0.9-2.1), respectively."

Using this reproducible method, we retrieve 2666 estimates from 784 abstracts. We then read these abstracts and filter out articles whose topic falls outside of the scope of our literature review. The final corpus is thus composed of 668 articles and 2155 estimates. Importantly, the set of articles considered is limited to those displaying confidence intervals and point estimates in their abstracts. We also build regex queries to retrieve other information about the articles such as the air pollutant and health outcome studied, the length of the study and the number of cities considered.

Based on this subset of articles, we first implement a retrospective power analysis to evaluate whether a study could recover an effect size equal to half of the obtained estimate. We carry out this analysis for the 1982 estimates that are statistically significant. Figure 1.5 displays the power and exaggeration curves for each result. They describe how these metrics vary with the hypothetical true effect sizes.

Figure 1.5: Power and Exaggeration Curves for the Epidemiology Literature.



Notes: Each gray line is a power curve or an exaggeration curve of a statistically significant result published in the epidemiology literature. The blue lines are the median values. For visual clarity, we drop results for which exaggeration ratios were too large.

If the true effect size was equal to half of the obtained estimate, 58% of the studies would have a power below the conventional 80% target used in randomized controlled trials. The median exaggeration ratio would be 1.3 and type S error would not be an issue. These figures however hide a lot of heterogeneity across studies. For one quarter of studies, the exaggeration would be higher than 1.9. We therefore try to apprehend the sources of this heterogeneity.

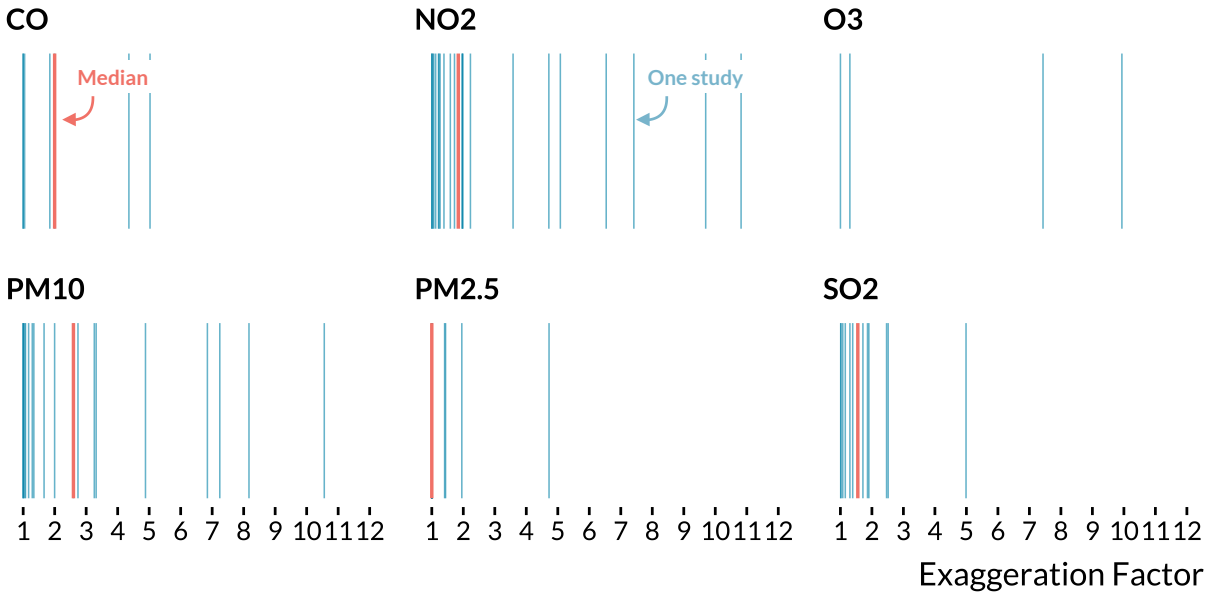
We find that inference issues do not depend on the health outcome and the air pollutant studied. Health science journals appear to be less prone to power issues than other journals. Researchers seem to be aware that they should work with large sample size as they often carry out multi-city studies. They also sometimes explicitly state that they investigate non-accidental mortality causes to increase statistical power since the average daily count is higher than for more specific death causes. Yet, the proportion of low power studies has been stagnating since the 2000s, revealing that practices regarding statistical power have

not evolved.

Studying the ability of each study to detect an effect that would be half of the obtained estimate gives an overview of power issues in this literature. It can however be viewed as arbitrary. Besides, while our approach enables studying the whole literature, it does not allow clearly analyzing the type of pollutants and outcomes considered in each study. As recommended by Gelman and Carlin (2014) and Ioannidis, Stanley, and Doucouliagos (2017), we thus make more informed guesses about potential true effect sizes for a subset of the literature using results from a meta-analysis. Shah et al. (2015) gathered 94 studies on the effects of several air pollutants on mortality and emergency admission for stroke. For each of these studies, we run retrospective power calculations to evaluate their ability to retrieve the meta-analysis estimates. We find that 63% of the studies in Shah et al. (2015) have a statistical power below 80%. The median exaggeration ratio of statistically significant estimate is equal to 1.6. Figure 1.6 plots for each air pollutant, the distribution of the exaggeration ratios (blue lines) and their medians (orange lines). The median exaggeration varies a lot by air pollutant, from 1 for $PM_{2.5}$ up to 13.4 for O_3 (the median is not displayed for visual clarity). More informed guesses about true effect sizes confirm that exaggeration is common in the standard epidemiology literature.

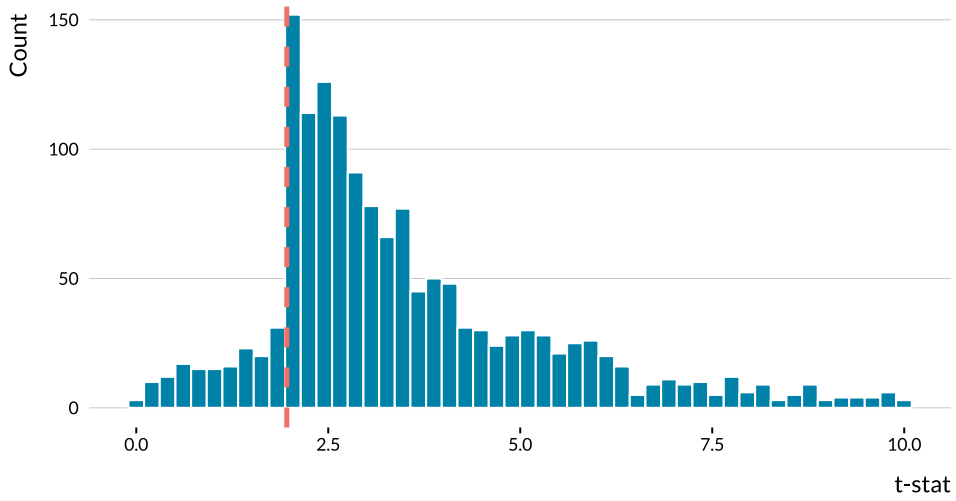
In this corpus and contrarily to the causal inference corpus, we do not observe the entire universe of published estimates but a selected subset: those that are reported in the abstracts. We can however evaluate a particular form of selection on significance, through the selection of headline results. These estimates are the ones that are emphasized and that thus may be used to inform policymaking. As visible on Figure 1.7, there is a striking bunching at the 1.96 threshold. Statistically significant results are favored in this literature, at least in terms of selection of results put forward. Identification approaches being rather uniform in this literature, the results may be particularly central in the assessment of the quality of a study and may play a central role in the decision process for publication.

Figure 1.6: Distribution of Exaggeration Ratios for Studies in Shah et al. (2015)'s Meta-Analysis.



Notes: Each blue line is the exaggeration ratio of a statistically significant estimate retrieved from Shah et al. (2015)'s meta-analysis. We use the meta-analysis estimates as true effect sizes in the retrospective power calculations. Orange lines are the medians. Extreme exaggeration ratios are removed for visual clarity. The median for O_3 is 13.4.

Figure 1.7: Distribution of t-statistics of estimates in abstracts from the standard epidemiology literature



Notes: All estimates come from the abstracts. Estimates with t-stat > 10 are not reported here for readability. The vertical dashed line represents the usual 1.96 threshold.

1.4 Approach for the Prospective Analysis

While a literature review can provide evidence of exaggeration and lack of statistical power, it does not allow us to clearly identify the parameters driving it. We therefore implement a prospective analysis to overcome this limitation (Altoè et al. 2020; Black et al. 2022, for other examples of power simulations). We run Monte-Carlo simulations based on real-data to emulate the main empirical strategies found in the literature. We use real data to avoid the difficult task of modeling the long-term and seasonal variations in health outcomes but also the specific effects of weather variables such as temperature. These simulations are grounded in the literature on the acute health effects of air pollution but share nonetheless many similarities with prevailing applied economics settings; panel data, large data sets, canonical identification strategies.

The present section describes how the simulations are implemented. Before that, we present the causal identification strategies used to measure the acute health effects of air pollution and then briefly describe the data used for the simulations.

1.4.1 Research Designs to Measure the Short-Term Health Effects of Air Pollution

Several empirical strategies have been leveraged to estimate the short-term health effects of air pollution. We simulate the main ones existing in the literature. We consider the usual setting where data on air pollution, weather parameters, and health outcomes are at the daily-city level.

Standard regression approach. The standard strategy consists in directly estimating the dose-response between an air pollutant and an health outcome. In the epidemiology literature, researchers often rely on Poisson generalize additive models where they regress the daily count of an health outcome on an air pollutant concentration, while flexibly adjusting for weather parameters, seasonal and long-term variations. We approximate the workhorse

model used by epidemiologists using linear models estimated via ordinary least squares:

$$\mathrm{Y}_{c,t} = \alpha + \beta \mathrm{P}_{c,t} + \mathbf{W}_{c,t}\lambda + \mathbf{C}_t\gamma + \epsilon_{c,t}$$

where c is the city index and t the daily time index. $\mathrm{Y}_{c,t}$ is the daily count of cases of an health outcome and $\mathrm{P}_{c,t}$ the average daily concentration of an air pollutant and $\epsilon_{c,t}$ an error term. The parameter β captures the short-term effect of an increase in the air pollutant concentration on the health outcome. To address confounding issues, the model adjusts for a set of weather covariates, $\mathbf{W}_{c,t}$, and calendar indicators $\mathbf{C}_t$.

Instrumental variable (IV) approach. The standard strategy could be prone to omitted variable bias and measurement error. A growing number of articles therefore exploits exogenous variations in air pollution. Most causal inference papers rely on IV designs where the concentration of an air pollutant is instrumented by thermal inversions (Arceo, Hanna, and Oliva 2016), wind patterns (Schwartz, Fong, and Zanobetti 2018; Deryugina et al. 2019), extreme natural events such as sandstorms or volcano eruptions (Ebenstein, Frank, and Reingewertz 2015; Halliday, Lynham, and Paula 2019), or variations in transport traffic (Moretti and Neidell 2011; Knittel, Miller, and Sanders 2016; Schlenker and Walker 2016). This approach can be summarized with a two-stage model where the first stage is:

$$\mathrm{P}_{c,t} = \delta + \theta \mathrm{Z}_{c,t} + \mathbf{W}_{c,t}\eta + \mathbf{C}_t\kappa + e_{c,t}$$

where $\mathrm{Z}_{c,t}$ is the instrumental variable. The second stage is then:

$$\mathrm{Y}_{c,t} = \alpha + \beta \widehat{\mathrm{P}}_{c,t} + \mathbf{W}_{c,t}\lambda + \mathbf{C}_t\lambda + \epsilon_{c,t}$$

where $\widehat{\mathrm{P}}_{c,t}$ is the exogenous variation in an air pollutant predicted by the instrument. The causal effect measured by this approach is a weighted average of per-unit causal responses

to an increase in the concentration of an air pollutant (Angrist and Imbens 1995).

Reduced-form approach. A subset of articles directly estimates the relationship between the health outcome and exogenous shocks to air pollution. For instance, articles using this approach exploit public transport strikes or thermal inversion as exogenous shocks (Bauernschuster, Hener, and Rainer 2017; Jans, Johansson, and Nilsson 2018; Godzinski, Castillo, et al. 2019; Giaccherini, Kopinska, and Palma 2021). They estimate a model of the form:

$$\mathrm{Y}_{c,t} = \alpha + \beta \mathrm{D}_{c,t} + \mathbf{W}_{c,t}\lambda + \mathbf{C}_t\gamma + \epsilon_{c,t}$$

where $D_{c,t}$ is a dummy equal to 1 when city c is affected by a shock at time t and 0 otherwise. The parameter β captures an intention-to-treat effect.

Regression-discontinuity design (RDD) approach. The last empirical strategy found in the literature measures the effects of air quality alerts with a regression-discontinuity design (Chen et al. 2018; Anderson, Hyun, and Lee 2022). In this approach, the following model is estimated for observations within an air pollution concentration bandwidth around the alert threshold:

$$\mathrm{Y}_{c,t} = \alpha + \beta \mathbf{1}\{P_{c,t} > P_c^{(a)}\} + \mathbf{W}_{c,t}\lambda + \mathbf{C}_t\gamma + \epsilon_{c,t}$$

where $P_c^{(a)}$ is the air pollution alert threshold for city c. We restrict our simulations to the case of sharp RDD. This model estimates the intention-to-treat effect of air quality alerts. It can both capture the effect of a subsequent decrease in air pollution caused by traffic restriction policies and inhabitants' avoidance behavior.

1.4.2 Data

The simulation exercises rely on a subset of the US National Morbidity, Mortality, and Air Pollution Study (NMMAPS). The dataset is publicly available and has been used in

several major studies in the early 2000s to measure the short-term effects of ambient air pollutants on mortality outcomes (Peng and Dominici 2008). Specifically, we extract data at the city-day level for 68 cities over the 1987-1997 period. It corresponds to 4,018 daily observations per city, for a total sample size of 273,224 observations. We select observations on the average temperature (C°), the standardized concentration of carbon monoxide (CO), and mortality counts for several causes. We focus on CO as it is the air pollutant measured in most cities over the period and its concentration is strongly correlated to that of other pollutants such as particulate matter. Less than 5% of carbon monoxide concentrations and average temperature readings are missing in the initial data set. We impute them using the chained random forest algorithm implemented in the `missRanger` package (Mayer 2019).

1.4.3 Simulations Set-Up

General procedure. Our simulation procedure follows 7 main steps:

1. Randomly draw a study period and a sample of cities.
2. For instrumental variable, reduced-form and regression-discontinuity designs, randomly allocate days to exogenous shocks/air quality alerts.
3. Modify the health outcome, adding a treatment effect that we will try to recover.
4. Estimate the model.
5. Store the point estimate of interest and its standard error.
6. Repeat the procedure 1000 times.
7. Compute the proportion of statistically significant estimates at the 5% level (the power), the average of the absolute value of significant estimates over the true effect size (the exaggeration ratio), and the proportion of significant estimates of the opposite sign of the true effect (the probability to make a type S error).

Modeling assumptions. To only capture the specific issues arising due to low statistical power, we build our simulations such that (i) they meet all the required assumptions of empirical strategies and (ii) make it easier—compared to real settings—to recover the treatment

effect. For all research designs, the treatment added to the data is not biased by unmeasured confounders nor measurement errors. For instrumental variable and reduced-form strategies, we only simulate binary and randomly allocated exogenous shocks (*e.g.* the occurrence of a thermal inversion). For the regression discontinuity approach, we only model sharp designs where an air quality alert is always activated above a randomly chosen threshold. The simulations always retrieve on average the true value of the treatment effect.

Two approaches for simulating research designs. For the reduced-form and regression discontinuity designs, we follow the Neyman-Rubin causal framework by simulating all potential outcomes (Rubin 1974). Consider that the health outcome value recorded in the NMMAPS dataset corresponds to the potential outcome $\mathrm{Y}_{c,t}(0)$. To create the counterfactuals $\mathrm{Y}_{c,t}(1)$, we add a treatment effect drawn from a Poisson distribution whose parameter corresponds to the magnitude of the treatment. We then randomly draw the treatment indicators $\mathrm{T}_{t,c}$ for exogenous shocks or air quality alerts. For reduced-form strategies, the treatment status of each day is drawn from a Bernoulli distribution with parameter equal to the proportion of exogenous shocks desired. For air pollution alerts, we randomly draw a threshold from a uniform distribution and select a bandwidth such that it yields the desired proportion of treated observations. We finally express the observed values Y^{obs} of potential outcomes according to the treatment assignment: $\mathrm{Y}^{obs}_{c,t} = (1\text{-}\mathrm{T}_{c,t}) \times \mathrm{Y}_{c,t}(0) + \mathrm{T}_{c,t} \times \mathrm{Y}_{c,t}(1)$.

To simulate standard regression and the instrumental variable strategies, we rely on a model-based approach. For the standard regression strategy, we first estimate the following statistical model on the data:

$$\mathrm{Y}_{c,t} = \alpha + \beta \mathrm{Z}_{c,t} + \mathbf{W}_{c,t}\lambda + \mathbf{C}_t\gamma + \epsilon_{c,t}$$

We then predict new observations of a $\mathrm{Y}_{c,t}$ using the estimated coefficients of the model ($\hat{\beta}$, $\hat{\lambda}$, and $\hat{\gamma}$) and by adding noise drawn from a normal distribution with variance equal to that of the residuals $\widehat{\epsilon_{c,t}}$ (Peng, Dominici, and Louis 2006). We modify the slope of the dose-response

relationship by changing the value of the air pollution coefficient β. For the instrumental variable strategy, we use the same method as for the standard regression approach but first modify observed air pollutant concentrations $\mathrm{P}_{c,t}$ according to the desired effect size θ of the randomly allocated instrument:

$$\widetilde{\mathrm{P}}_{c,t} = \mathrm{P}_{c,t} + \theta \mathrm{Z}_{c,t}$$

We draw the allocation of each day to an exogenous shock from a Bernoulli distribution with parameter equal to the proportion of exogenous shocks. We then estimate a two-stage least squares model (2SLS) and modify the coefficient for the effect of the air pollutant on an health outcome. We finally generate the fake observations of the health outcome by combining the prediction from the modified 2SLS model and noise drawn a normal distribution with variance equal to that of the residuals.

Varying parameters. To understand which parameters affect statistical power issues, we modify one aspect of the research design while keeping other parameters constant. We study the influence of four main parameters. First, we vary the sample size by drawing a different number of cities and changing the length of the study period. Second, we consider different effect sizes of air pollution or of an exogenous shock on the health outcome. Third, we allocate increasing proportions of exogenous shocks/air quality alerts. Fourth, we vary the number of cases in the outcome by considering different health outcomes.

Simulations of Case Studies. The simulations described above help explore the effect of each parameter on statistical power issues. Yet, the resulting set of parameters considered may not be perfectly representative of actual studies. To address this concern, we also calibrate simulation parameters to reproduce three papers published in the literature. We report these analyzes in Appendix A.3.

1.5 Results of the Prospective Analysis

In this section, we describe how statistical power evolves with the treatment effect size, the number of observations, the proportion of exogenous shocks, the average count of the health outcome, and the strength of the instrument. In Appendix A.3, we show that statistical power issues can be substantial for actual parameter values found in the literature on acute health effects of air pollution.

1.5.1 Evolution of Power, Exaggeration Ratio and Type S Error with Study Parameters

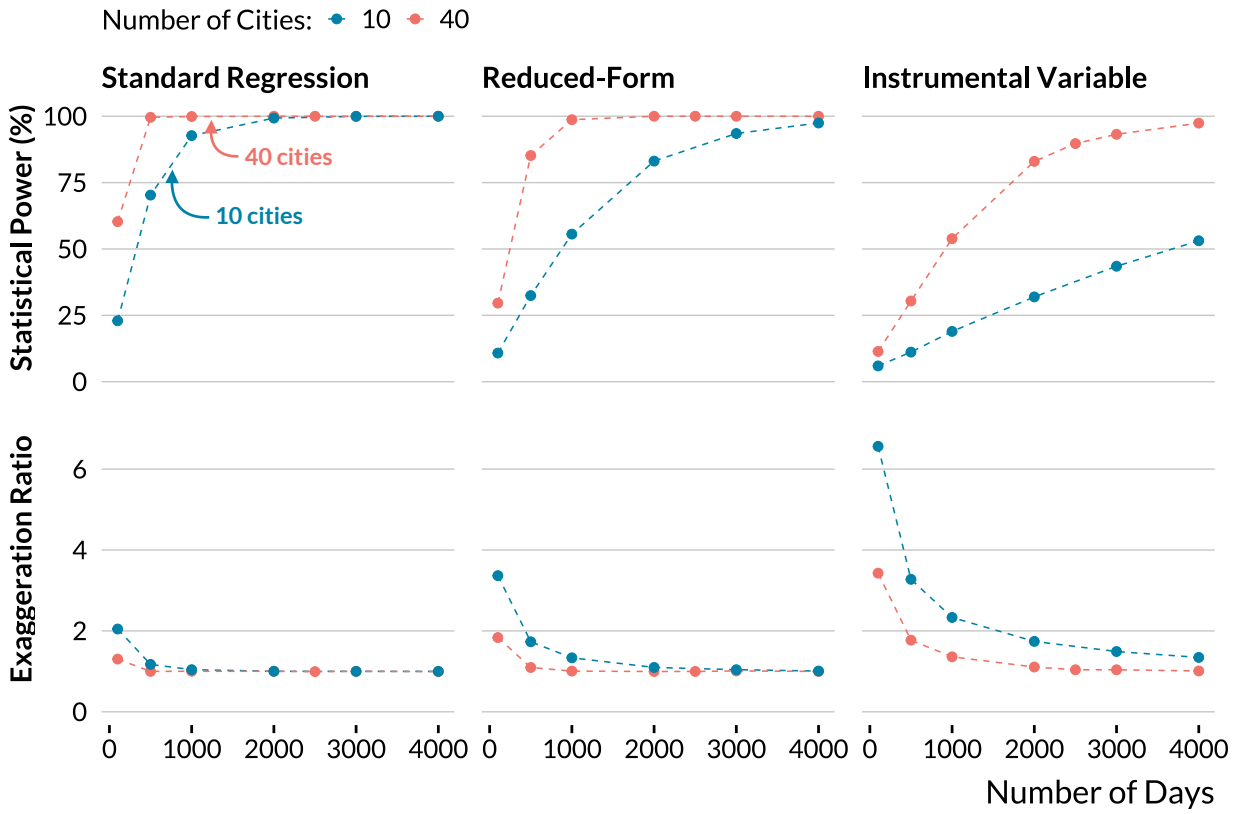
We aim to analyze how statistical power, exaggeration ratio and type S error are affected by the value of different study parameters. To do so, we set baseline values for these parameters and vary the value of each of them one by one. This enables us to get a sense of the impact of each parameter, other things held equal. We consider the following baseline parameters:

- A large sample size of 100,000 observations (2500 days $\times$ 40 cities),
- A 1% effect size, the order of magnitude found in the most precise studies of the literature. A one standard deviation in air pollution or an exogenous shock increases the health outcome by 1%,
- 50% of observations are subject to an exogenous shock. For air pollution alerts analyzed with regression discontinuity designs, we only consider observations close to the threshold, resulting in a smaller proportion of treated units: 10%,
- The health outcome is the total daily number of non-accidental deaths. It is the health outcome with the largest average number of counts (average daily mean of 23 cases).

For all statistical models, we adjust for temperature, temperature squared, city and calendar (weekday, month, year, month$\times$year) fixed effects. We also repeat the simulations for a smaller sample size of 10,000 observations.

Sample Size

Figure 1.8: Evolution of Power and Exaggeration with Sample Size.



Notes: The other parameters are set to their baseline values: a true effect size of 1%, 50% of observations subject to an exogenous shock for instrumental variable and reduced-form designs, and the health outcome is the total number of non-accidental deaths.

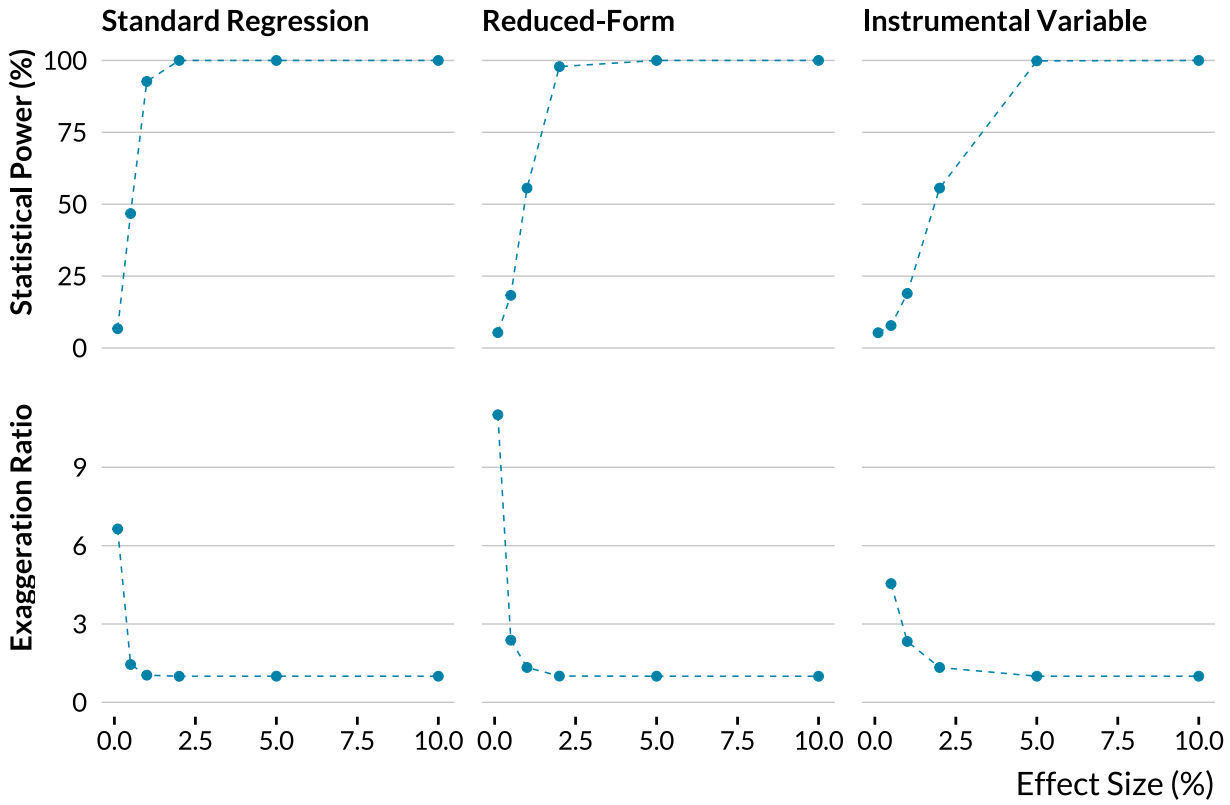
In Figure 1.8, we recover the well-known increasing relationship between the number of observations and statistical power. Conversely, the exaggeration ratio decreases with the number of observations. These results stem from the fact that statistical power decreases and exaggeration increases when the variance of a normally distributed estimator increases (Zwet and Cator 2021; Lu, Qiu, and Deng 2019) and that the variance of common estimators increases as the number of observations decreases.

We also find that statistical power and exaggeration issues can arise even for a large number of observations. For a sample size of 40,000 observations, the instrumental variable strategy only has a statistical power of 54% and exaggerates the true effect by a factor of 1.4.

On the contrary, the standard regression strategy is much less prone to power issues than the instrumental variable strategy. This is explained by the fact that the variance of the two stage least-square estimator is larger than the variance of the ordinary least square estimator. In our simulations, the probability to make a Type S error is null for all identification methods and sample sizes.

Effect Size

Figure 1.9: Evolution of Power and Exaggeration with Effect Size.



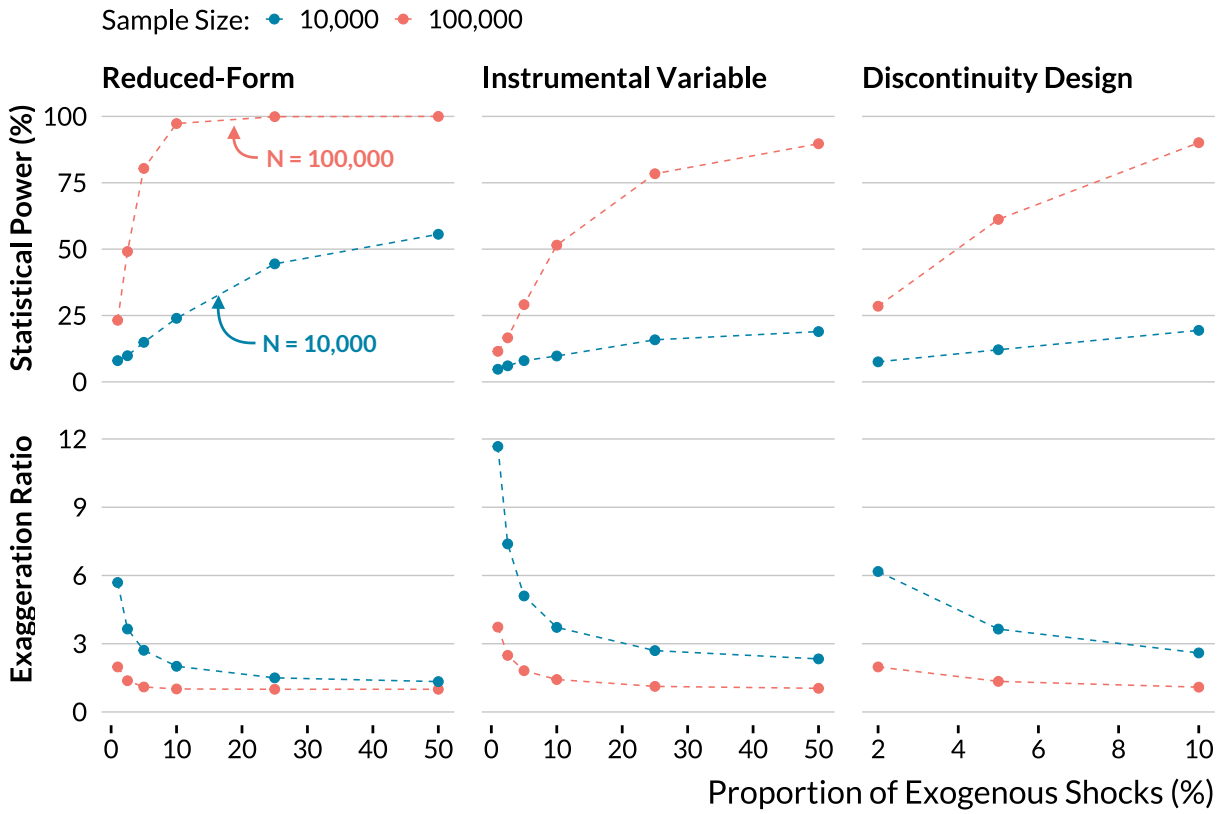
Notes: The sample size is 10,000. The other parameters are set to their baseline values: 50% of observations subject to an exogenous shock for instrumental variable and reduced-form designs, and the health outcome is the total number of non-accidental deaths. For an effect size of 1%, we do not display the exaggeration ratio of the instrumental variable design since it is above 20 and it would distort the graph.

In Figure 1.9, we retrieve another familiar result: the larger the effect size, the larger the power. As expected from Zwet and Cator (2021) and Lu, Qiu, and Deng (2019)'s results, we also find that the exaggeration ratio decreases with the true effect size. Even for our

large baseline sample size, statistical power issues appear for effect sizes routinely found in the epidemiology literature. For instance, for our instrumental variable strategy and an effect size of 0.5%, the average exaggeration ratio is about 1.7. As for results on sample sizes, standard regression and reduced-form strategies are less prone to power issues, even for small effect sizes.

Proportion of Exogenous Shocks

Figure 1.10: Evolution of Power and Exaggeration with the Proportion of Exogenous Shocks.



Notes: The other parameters are set to their baseline values: a true effect size of 1% and the health outcome is the total number of non-accidental deaths. The proportion of exogenous shocks corresponds to the fraction of days in the sample that are allocated to the treatment.

The link between the proportion of exogenous shocks and statistical power might be less widely known. In Figure 1.10, we show that statistical power increases with the proportion of treated units for instrumental variable, regression discontinuity and reduced-form designs.

Conversely, the average exaggeration ratio increases as the proportion of exogenous shocks decreases.

This result can be explained by the fact that exaggeration increases and statistical power decreases with the variance of the estimator (Zwet and Cator 2021; Lu, Qiu, and Deng 2019). Now, as routinely discussed for randomized controlled trials but seldom in the case of non-experimental studies, precision is maximized when half of the observations are exposed to the treatment of interest. The variance of the average treatment effect estimator (ATE) is $\sigma^2/[n \times p(1-p)]$ where σ is the standard deviation of the outcome in the treated and control groups and p the proportion of treated units. This quantity increases when p departs from 0.5. Thus, exaggeration increases when the proportion of exogenous shocks decreases, as long as it was initially smaller than 0.5.

Another way to interpret this result is to consider that a small number of exogenous shocks limits the variation that can be leverage to identify the effect of interest. When the proportion of shocks decreases, the variance of the treatment variable decreases and therefore the variance of the estimator increases. A similar reasoning can be applied to IV strategies.

In practice, air pollution alerts, thermal inversion or transportation strikes are generally rare events. In some studies, they represent less than 5% of the observations. With a dataset of 10,000 observations, our simulations return an average exaggeration ratio of 2.7 for the reduced-form strategy. Despite large sample sizes, air pollution studies exploiting few exogenous shocks might be particularly prone to exaggeration issues.

Average Count of Cases of the Health Outcome

Subgroup analyses are routinely carried out in the literature to evaluate the acute health effects of air pollution on children or the elderly. Yet, the average count of cases can also critically affect statistical power as shown in Table 1.2. For instance, in a setting with only few deaths per day, a 1% increase in the number of deaths will rarely cause additional deaths. The effect will be more difficult to detect. To simulate situations with various number of

Table 1.2: Evolution of Power and Exaggeration with the Average Number of Daily Cases of Health Outcomes.

	Non-Accidental	Respiratory	COPD
Number of Cases	23	2	0.3
Statistical Power (%)	90	16	7.5
Exaggeration Ratio	1	2.4	5.9

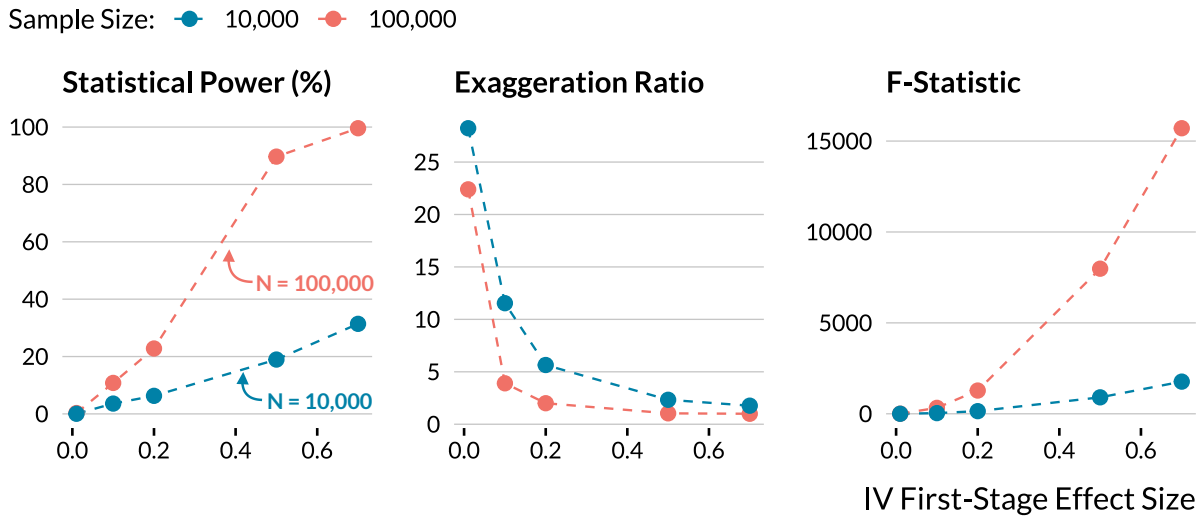
Notes: This table displays the average number of cases, the power and the exaggeration ratio for three health outcomes: non-accidental deaths, respiratory deaths, and chronic pulmonary deaths for individuals aged between 65 and 75. These figures are obtained for the instrumental variable design with a sample size of 100,000 and 50% of observations subject to an exogenous shock. The instrument variable increases the air pollutant concentration by 0.5 standard deviation. A one standard deviation increase in the instrumented air pollutant leads to 1% relative increase in the health outcome considered.

cases, we consider three different outcome variables, with different counts of cases: the total number of non-accidental deaths (daily mean $\simeq 23$), the total number of respiratory deaths (daily mean $\simeq 2$) and the number of chronic obstructive pulmonary disease (COPD) cases for individuals aged between 65 and 75 (daily mean $\simeq 0.3$). Using baseline parameters and in the case of the large dataset, we find that statistical power is close to 100% for a 1% increase in the total number of non-accidental deaths. However, statistical power drops when the average count of cases decreases. For instance, the instrumental variable strategy has only 16% of statistical power to detect a 1% increase in respiratory deaths. The average exaggeration ratio is then equal to 2.4. For chronic obstructive pulmonary deaths—the health outcome with the lowest number of cases—the situation is even worst since the average exaggeration ratio reaches 5.9. When focusing on subgroups such as children or the elderly, one can expect to find larger effect sizes as those populations are more vulnerable to air pollution. While these larger effect sizes attenuate exaggeration concerns, the lower number of cases exacerbates them. It creates a trade-off for power issues.

Issues Specific to the Instrumental Variable Design

In the case of instrumental variable strategies, statistical power is affected by the effect size in the first stage of the IV. In our simulations, we consider a binary instrument (e.g., the occurrence of a thermal inversion or a public transport strike). We define its the effect size of the first stage of the IV as the standardized effect size of the instrument on the air pollutant concentration. A IV first stage effect size of 0.2 means that the instrument increases the concentration by 0.2 standard deviation.

Figure 1.11: Evolution of Power and Exaggeration with the Strength of the Instrumental Variable.



Notes: The true effect size is a 1% relative increase in the health outcome. The health outcome used in the simulations is the total number of non-accidental deaths. Half of the observations are exposed to exogenous shocks. The strength of the instrumental variable is defined as its effect in standard deviation on the air pollutant concentration.

As shown in Figure 1.11, we find that statistical power collapses and exaggeration soars when the "IV first stage effect size decreases. Importantly, this issue even arises for large first-stage F-statistics. In our simulations based the large data set with 100,000 observations, an instrumental variable's strength of 0.2, and an effect size of 1%, we find an average F-statistics of 1278. The statistical power is however only 23% and the average exaggeration ratio 2. A large F-statistic could therefore hide large exaggeration issues.

The relationship between IV strength and exaggeration comes from the fact that the variance of the 2SLS estimator decreases with the correlation between the instrument and the instrumented variable. In the homoskedastic case, the asymptotic variance of the 2SLS estimator is $(\mathbb{E}[XZ']\mathbb{E}[ZZ']^{-1}\mathbb{E}[ZX'])^{-1}\sigma^2$, where σ^2 is variance of the error, X the endogenous variable and Z the instrument. When $\mathbb{E}[XZ']$ and $\mathbb{E}[ZX']$ decrease, the variance of the estimator increases. Again, since Zwet and Cator (2021) and Lu, Qiu, and Deng (2019) show that as the variance of a normally distributed estimator increases, the statistical power decreases and exaggeration increases, we obtain the intuition for the simulation results.

1.6 Discussion

Growing evidence shows the existence of statistical power issues and publication bias towards statistical significance in economics, causing exaggeration (Brodeur et al. 2016; Ioannidis, Stanley, and Doucouliagos 2017; Brodeur, Cook, and Heyes 2020; Ferraro and Shukla 2020). Although this issue appears to be increasingly acknowledged, discussions about the drivers of exaggeration and therefore actionable guidance to tackle it in non-experimental economic research are still lacking (Altoè et al. 2020; Black et al. 2022). In this paper, we highlighted a list of concrete drivers of low power and exaggeration we should pay attention to when carrying a non-experimental study.

While the simulations we ran were specific to studies on the acute health effect of air pollution, we argue that they can provide lessons for other types of non-experimental studies. First, a large literature investigates the short-term impacts of air pollution on different outcomes such as criminality, cognitive skills and productivity (Herrnstadt et al. 2021; Ebenstein, Lavy, and Roth 2016; Adhvaryu, Kala, and Nyshadham 2022, for instance). These studies use data with a very similar structure, only focusing on different outcomes and find effects of comparable magnitude or even smaller than those in our literature of interest. Our results should therefore be directly applicable to these literatures. More broadly, settings with typically low signal-to-noise ratios can by definition be subject to power and exagger-

ation issues. Since as described in Section 1.5, the impact of each driver we identified can be explained theoretically, we expect these drivers to affect power and exaggeration in other settings as well. A limited effect size, effective sample size, number of exogenous shocks, average count of the outcome or strength of the instrument can create exaggeration in many settings. Paying careful attention to these factors should help avoid exaggeration issues when running a non-experimental study.

In addition to these specific factors, when carrying out a study, we propose to systematically run retrospective calculations to gauge the risk of exaggeration. They are easy to implement and force us to discuss credible effect size. They allow evaluating if our research design enables us to confidently estimate a credible range of effect sizes. We implemented and discussed such calculations in our literature review and illustrate this approach in more details in Appendix A.2 by considering the example of Deryugina et al. (2019). As an even simpler first check, we also suggest to consider large confidence intervals verging 0 not only as a sign of uncertainty regarding the exact magnitude of the effect but also of limited power and potential exaggeration of the obtained point estimate.

Then, we advocate conducting prospective simulations before undertaking a non-experimental study. It allows verifying whether our design can detect effects of a credible magnitude in an almost-ideal setting. Unlike a retrospective analysis, enables us to identify factors that drive exaggeration. Fake-data can be simulated from scratch or simulations can build on datasets used in other studies, as we did in the simulation section of this paper. To facilitate the adoption of this practice, we describe the template we use to run our simulations in the replication material. Black et al. (2022) also provide useful recommendations to implement power simulations.

More generally, we advocate paying attention to statistical power in non-experimental studies, even after a statistically significant estimate has been obtained, as insufficient power can lead to exaggeration and inaccurate published estimates. As such, we advocate reporting power calculations to demonstrate the robustness of the design and its ability to accurately

capture smaller effect sizes.

On top of these specific comments, we should not forget that published estimates only suffer from exaggeration in the presence of publication bias. Adopting a different view towards statistically insignificant results would therefore yield important benefits (Ziliak and McCloskey 2008; Wasserstein and Lazar 2016; McShane et al. 2019). It currently dichotomizes evidence according to the 5% significance threshold, disregarding non-significant results (Greenland 2017). Instead, if results were published regardless of their significance, the resulting distribution would be centered around the true effect (Hernán 2022). To replace the null hypothesis testing framework, we advocate focusing on confidence intervals and to interpret the range of effect sizes supported by the data (Amrhein, Trafimow, and Greenland 2019; Romer 2020). Qualifying estimates as "statistically significant" does not acknowledge the actual uncertainty that should be computed and embraced to better help policy-makers. Prospective and retrospective power analyses can help design better studies and improve the interpretation of their results.

Chapter 2: Causal Exaggeration:
Unconfounded but Inflated Causal Estimates

Abstract

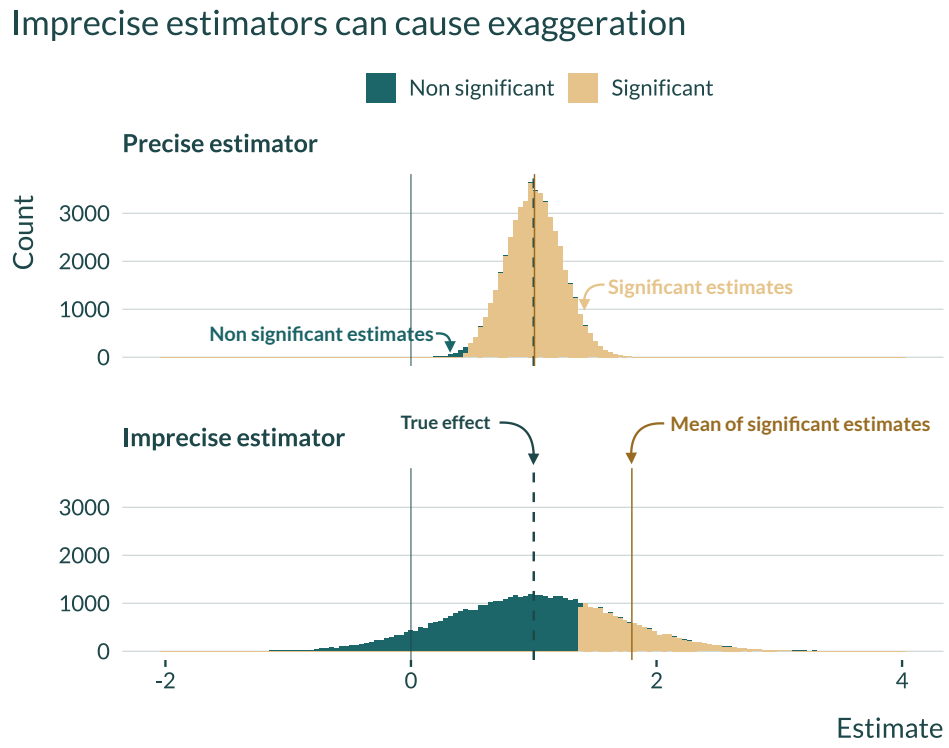
The credibility revolution in economics has made causal inference methods ubiquitous. Simultaneously, an increasing amount of evidence highlights that the literature strongly favors statistically significant results. I show that these two phenomena interact in a way that can substantially worsen the 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 as estimates are routinely turned 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 its prevalence in the literature. I then discuss potential avenues to address this issue.

2.1 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 a portion of the variation in the data. They exploit the exogenous part of the variation in the treatment or decrease the sample size by only considering observations for which the as-if random assignment assumption is credible. 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 (Gelman and Tuerlinckx 2000; Ioannidis 2008; Gelman and Carlin 2014). Only estimates at least 1.96 standard errors away from zero are statistically significant at the 5% level. In under-powered studies, these estimates make up a selected sub-sample of all estimates, located in the tails of the distribution of all possible estimates. The average of these statistically significant estimates differs from the true effect, located at the center of the distribution if the estimator is unbiased. They exaggerate the true effect and the less precise the estimator, the larger exaggeration is. Figure 2.1 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.

Figure 2.1: Significance and distribution of two unbiased estimators with different variances



Notes: 100,000 draws from two normal distributions $\mathcal{N}(1, 0.05)$ and $\mathcal{N}(1, 0.5)$. Significance level: 5%

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 underlines that the economics literature selects results based on statistical significance (Rosenthal 1979; Brodeur et al. 2016; Andrews and Kasy 2019; Abadie 2020; Brodeur, Cook, and Heyes 2020, for instance). Additional studies have highlighted its frequent and substantial lack of statistical power and resulting exaggeration (Ioannidis, Stanley, and Doucouliagos 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, Stanley, and Doucouliagos (2017) finds that the median statistical power in a wide range of areas of economics is no more than 18%. Despite usually large

sample sizes, they show that nearly 80% of estimates are likely exaggerated by a factor of two. As further illustrated in section 2.2, the magnitude of exaggeration can be considerable and in some situations could be on par with that of a bias caused by confounders. Taking exaggeration into account and to understanding its drivers is therefore crucial.

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

In the present paper, I argue that the use of causal identification strategies can increase exaggeration. Reviewing the literature, using a mathematical derivation and Monte Carlo simulations, I show that design choices in quasi-experimental studies can be seen as a trade-off between avoiding confounding and overestimating true effect sizes. To limit the threat of confounding, causal inference methods discard variation and can therefore reduce statistical power. When combined with a statistical significance filter, this results in exaggeration bias. While causal identification strategies are essential to describe causal relationships, the

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

present 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 bias caused by confounding actually introduce another type of bias.

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

Since causal identification strategies can be interpreted as ways of controlling for confounders, this last point actually ties all the strategy-specific arguments together.[2] When these identification strategies absorb more of the variation in the treatment than in the outcome, they increase the variance of the resulting estimator and can cause exaggeration. Considering a simple linear homoskedastic model gives the intuition for this trade-off between exaggeration and omitted variable bias (OVB) for control approaches. Let $y_i = \alpha + \beta x_i + \delta w_i + u_i$, $\forall i \in \{1, .., n\}$, with x the variable of interest, w a potentially unobserved variable correlated with x and u an error term. Under usual assumptions and

[2]Fixed Effects (FEs) based identification strategies such as DiD control for the invariant, unobserved, and arguably endogenous part of the variation in the outcome. An IV approach essentially partials out the variation in x unexplained by the instruments. Fuzzy-RD and propensity score matching can be thought of as control function approaches, of the forcing variable and propensity score respectively. In addition, excluding observations that are outside the bandwidth or unmatched is equivalent to controlling for observation-level fixed effects for these observations.

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

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

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. Since exaggeration increases with the variance of the estimator, controlling for a confounder can increase exaggeration. This has direct implications when using fixed effects: an exaggeration bias can arise from the use of fixed effects if they absorb more of the variation in x than in $y^{\perp x}$. Under some circumstances discussed in sections 2.3 and 3.4.3, controlling can even produce an exaggeration bias larger than the OVB that would result from an absence of controls.

In the remainder of the paper, I first document the magnitude of the trade-off. To do so, I build on existing literature reviews to discuss evidence of exaggeration bias in a large set of causal studies mostly published in top journals (Brodeur, Cook, and Heyes 2020; Young 2022; Bagilet 2023a; Lal et al. 2024). These analyses reveal heterogeneity across analyses: while exaggeration might be limited in some studies, it is likely substantial in many others. In the set of IV papers reviewed in Young (2022), I show that half of the designs would

exaggerate a true effect size of the magnitude of the "naive" OLS estimate by a factor larger than 3.2—and by a factor larger than 2.0 for headline designs. I then directly compare confounding and exaggeration biases for an example study. I show that in this instance the bias of the IV could be on par or even larger than that of the OLS and that exaggeration may explain this difference.

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

Then, I show that this "causal exaggeration" arises for realistic parameter values by further exploring its drivers in realistic settings in which there is no closed form formula for power and exaggeration available. The exaggeration of statistically significant estimates can be defined as the absolute value of the ratio of these significant estimates 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. Monte-Carlo simulations also allow varying the value of the parameter of interest *ceteris paribus*, something that would not be possible otherwise. To make these simulations concrete, I calibrate them to emulate existing studies from environmental, education, labor, health and political economics. I find that causal exaggeration can be substantial in realistic settings where the variation remaining for identification is limited.

Finally, I discuss concrete avenues to address causal exaggeration when carrying out a non-experimental study[3]. A series of tools can be used to evaluate the potential magnitude of confounding and exaggeration issues separately. Sensitivity analyses help with the former while power calculations help with the latter. Considering the attention given to bias avoid-

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

ance in the economics literature, I underline that making power central to non-experimental analyses, even after an effect has been found, would help limiting bias caused by exaggeration. Prospective power simulations help identify the design parameters affecting power and exaggeration by approximating the data generating process (Gelman, Hill, and Vehtari 2020; Black et al. 2022). Retrospective power calculations allow evaluating whether a study would have had enough power to confidently estimate a range of smaller but credible effect sizes (Gelman and Carlin 2014; Stommes, Aronow, and Sävje 2023). Focusing more specifically on the trade-off and its drivers, I consider tools to identify the variation actually used for identification when using causal identification strategies.

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 or focus on specific causal identification methods separately (Ioannidis, Stanley, and Doucouliagos 2017; Schell, Griffin, and Morral 2018; Ferraro and Shukla 2020; Black et al. 2022; Stommes, Aronow, and Sävje 2023; Arel-Bundock et al. 2022). In a companion paper, I highlight tangible design parameters that can cause exaggeration for a wide range of empirical designs (Bagilet 2023a). 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, Stanley, and Doucouliagos 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 the causal economics literature. In section 2.3, 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 3.4.3, 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 2.5 and conclude in section 2.6.

2.2 Causal Exaggeration in the 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 confounding bias these methods allow avoiding.

The literature on the acute health effects of air pollution, that I explore in a companion paper, provides an illustration of potential exaggeration of causal methods (Bagilet 2023a). The historical literature mostly relies on associations (Dominici and Zigler 2017; Bind 2019). A more recent literature based on causal identification strategies confirms the adverse effects of air pollution in the short term (Schwartz et al. 2015; Schwartz, Fong, and Zanobetti 2018; Deryugina et al. 2019). Yet, causal estimates are substantially larger than what would have been predicted by the standard epidemiology literature, estimates being regularly more than 10 times larger. Even within a given setting, my literature review of this literature shows that the median of the ratio of the obtained 2SLS to their corresponding "naive" OLS estimates is 3.8. Similarly, the data in Young (2022) shows that in a set of 30 reproducible instrumental

variables papers published in journals from the *American Economic Association*[4], the median ratio of headline IV estimates over the corresponding OLS is 2.3 and more than 5.4 for 25% of the estimates. A comparable pattern is observed in top political science journals (Lal et al. 2024). What can explain that causal methods yield such large effects sizes, as compared to non-causal methods? They could arguably remove omitted variable bias, reduce attenuation bias caused by classical measurement error or target a different causal estimand. But as I argue in this section, exaggeration and imprecision could also explain part of this difference. In the example case of studies on the short-term health effects of air pollution, causal studies often display a low relative precision, not only because effects are typically small and the data relatively coarse—at the city-day level—but also because they sometimes leverage rare exogenous shocks such as public transportation strikes, take advantage of air pollution alerts to build RDD that limit sample size or use instruments that do not strongly predict air pollution levels. More broadly, published IV papers typically have low power and make most 2SLS estimates statistically indistinguishable from the corresponding OLS (Young 2022).

2.2.1 Quantifying exaggeration

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

To focus more specifically on the causal inference literature, I first leverage data from Brodeur, Cook, and Heyes (2020). This paper reviews of the universe of hypothesis tests reported in papers published in the top-25 economics journals in 2015 and 2018 and using

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

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

The exaggeration is computed by drawing a large number of times from a normal distribution centred on the hypothetical effect size and with a standard deviation equal to the standard error of the estimate found in the study and by then computing the average of the draws that are 1.96 standard errors away from 0. I discuss in details how to compute such power calculations in section 2.5.2. In the data from Brodeur, Cook, and Heyes (2020) the median power and median exaggeration ratio of significant estimates would be 37% and 1.6 respectively. Statistical power would be larger than the usual 80% threshold for only 22% of designs. 28% of the designs would, on average, exaggerate the true effect sizes by more than a factor of 2. While exaggeration is reassuringly limited in some settings, it can be substantial in many others. These results are relatively comparable across methods (IV, DID, RDD and RCT).

The previous hypothesis on the true effect size, despite enabling to get an overview of a wide literature, can seem to some extent arbitrary. To reduce the burden of assumptions, I

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

2.2.2 Illustration of the trade-off

In order to limit hypotheses and to investigate causal exaggeration further, I focus on an example IV study investigating the impact of PM2.5 air pollution on mortality (He, Liu, and Zhou 2020). It allows comparing the bias of the "naive" OLS to that of the IV by computing their distance to an estimate of the "true effect" they target. I define this "true effect" in two ways. First, I use the results of a meta-analysis of epidemiological studies (Shah et al. 2015). By pooling a number of studies carried out in various contexts, this meta-estimate might represent the average effect one may expect from such a study. However, the studies

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

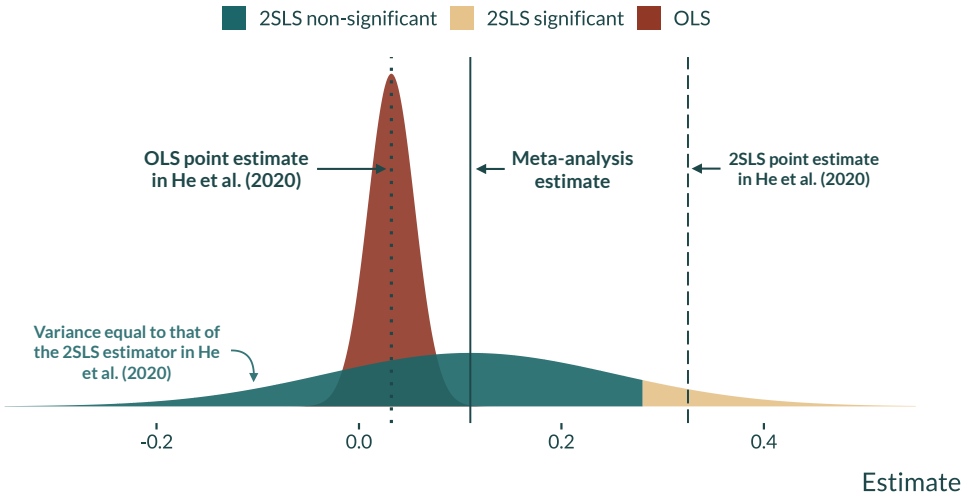
in this meta-analysis do not rely on canonical causal identification strategies used in economics and may be thought of as suffering from counfounding. I thus consider the result of Deryugina et al. (2019)—a precise causal study that may be less exposed to exaggeration—as an alternative estimate of the "true effect". This estimate may be context specific and the "true effect" in He, Liu, and Zhou (2020) may deviate from these. The present discussion is conditional on this true underlying effect being close to these hypothesized true effect sizes.

He, Liu, and Zhou (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, Liu, and Zhou (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, Liu, and Zhou (2020) would still yield exaggerated statistically significant estimates. Figure 2.2 illustrates this point. Even if the 2SLS estimator is unbiased, i.e., centered on the meta-estimate found in Shah et al. (2015), the lack of precision of this design would lead significant estimates to substantially exaggerate the true effect, by a factor 3.2 on average. The 2SLS estimate found in He, Liu, and Zhou (2020) could be one of these estimates.

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

Figure 2.2: Illustration of the Confounding-Exaggeration Trade-off in He, Liu, and Zhou (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, Liu, and Zhou (2020). The distribution for the OLS estimator is centered on the OLS estimate found in He, Liu, and Zhou (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, Liu, and Zhou (2020) respectively. I ignore exaggeration of the OLS for clarity but since the OLS is biased downward, inflating OLS estimates bring them closer to the true effect.

be 3.1 times larger than the OVB (or 1.3 time if the true effect is closer to the one found in Deryugina et al. (2019)).

2.3 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.

2.3.1 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

$$E(\hat{\beta}, \sigma, \beta, z_\alpha) = \frac{\mathbb{E}\left[|\hat{\beta}| \ |\beta, \sigma, |\hat{\beta}| > z_\alpha \sigma\right]}{|\beta|} \tag{2.1}$$

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

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

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

Figure 2.1 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 2.3.1, 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

[6] All the proofs of the lemma and theorems are reported in appendix B.1.

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.

2.3.2 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:

$$y_i = \beta_0 + \beta_1 x_i + \delta w_i + u_i \tag{2.2}$$

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:

$$x_i = \mu_x + \gamma w_i + \epsilon_i \tag{2.3}$$

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* $\mathbb{E}[\mathrm{z}_i u_i] = 0$ and

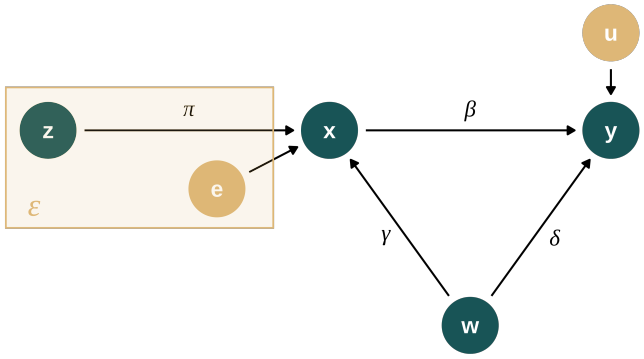
$\mathbb{E}[z_i w_i] = 0$, relevance, *ie* $\text{rank}(\mathbb{E}[\text{z}_i \text{x}_i'])$ $= 2$, and positive-definiteness of $\mathbb{E}[\text{z}_i \text{z}_i']$. The data generating process for x_i becomes:

$$x_i = \pi_0 + \pi_1 z_i + \gamma w_i + e_i \tag{2.4}$$

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 2.3 represents the data generating process.

Figure 2.3: DAG of the data generating process



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

2.3.3 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 $\gamma\delta = \kappa = cst$.

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 2.3.2, for $\hat{\beta}_{\text{OVB}}$ the OLS estimate of β_1 in the regression of y on x, $\hat{\beta}_{\text{OVB}} \xrightarrow{d} \mathcal{N}(\beta_1 + b_{\text{OVB}},\ \sigma^2_{\text{OVB}})$, with*

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

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

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 2.3.1 proves that E_{OVB} does not either.

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*

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

Note that $\sigma^2_x(1-\rho^2_{xw})$ is the part of the variance of x that is not explained by w ($\sigma^2_{x^{\perp w}}$) and σ^2_u 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:

$$\sigma^2_{\text{CTRL}} = \frac{\sigma^2_y - \beta_1^2\sigma^2_x - \frac{\kappa^2}{\gamma^2}\sigma^2_w - 2\beta_1\kappa\sigma^2_w}{n\ (\sigma^2_x - \gamma^2\sigma^2_w)}$$

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 $\lim_{\gamma^2 \to \frac{\sigma^2_x}{\sigma^2_w}} \sigma^2_{\text{CTRL}} = +\infty$.

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*

$$\sigma^2_{\text{IV}} = \frac{\sigma^2_u + \delta^2\sigma^2_w}{n\ \sigma^2_x\rho^2_{xz}}$$

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^2_u, we can rewrite σ^2_{IV} as a function of fixed variances and one varying parameter, π_1:

$$\sigma^2_{\text{IV}} = \frac{\sigma^2_y - \beta^2_1\sigma^2_x - 2\beta_1\kappa\sigma^2_w}{n\ \pi^2_1\sigma^2_z}$$

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

Reduced form

Let us now assume that we want to directly estimate the effect of the instrument on the outcome of interest. Plugging equation 2.4 into equation 2.2 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.

2.3.4 Exaggeration ratios

Combining the results from lemma 2 through 5 regarding the asymptotic distribution of the various estimators with lemma 2.3.1 stating that exaggeration increases with the variance of a normally distributed estimator yields:

Theorem 1 *For the data generating process described in section 2.3.2, 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 2.3.2, 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 2.3.2,* $\forall\ b_{OVB}$*,*

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

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

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

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.

2.4 Simulations

To study the drivers of exaggeration in concrete settings, I build simulations that reproduce real-world examples from economics of education for RDD, political economy for IV, health economics for exogenous shocks and environmental economics for control and fixed effects approaches. I split the simulations by identification strategy. Real-world settings enable clearly grasping the relationships between the different variables and to set realistic parameter values, based on existing studies. I do not reproduce a specific study but instead calibrate my simulations to emulate an archetypical study from each literature to underline that causal exaggeration is not bound to specific studies. It shows that exaggeration can arise for parameter values consistent with existing studies. To check the representativity of my simulations, I compare the Signal-to-Noise Ratios (SNR) from my simulations to the estimate/standard error ratios in studies from the corresponding literature. As underlined by

lemma 1, the SNR is a sufficient statistic for the exaggeration ratio; a SNR consistent with the literature will ensure the representativity of the simulations with respect to the feature of interest, exaggeration. To limit estimation challenges, facilitate the recovery of the effect of interest and focus on power aspects, 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. Extensively documented code for each simulation procedure is available on the project's website.

2.4.1 Regression Discontinuity Design

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

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

The simulation framework is as follows. If a student i has an initial scores $Qual_i$ be-

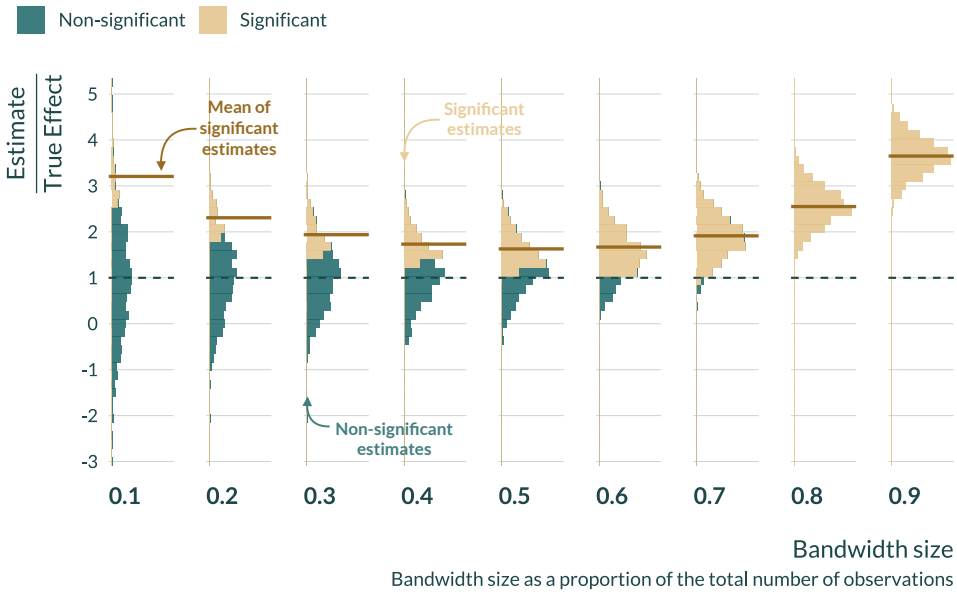
low a cutoff C, they must take additional lessons, making the allocation of the treatment T sharp: $T_i = \mathbb{I}[Qual_i < C]$. Both qualification and final test scores are affected by students' unobserved ability w in a non-linear (cubic) way. A high or low ability has a strong positive impact on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

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

Results. Figure 2.4 displays the results of these simulations. For large bandwidth sizes, the distribution of estimates is far away from the true effect; there is omitted variable bias. As the bandwidth decreases, the identification strategy gets rid of the OVB and the distribution of the estimates becomes centred on the true effet. At the same time, as the bandwidth and the effective sample size decrease, the distribution widens and significant estimates represent a smaller and smaller subset of the distribution, located in the tails of the distribution, creating exaggeration. While the average of all estimates gets close to the true effect as

bandwidth size and thus OVB decrease, in this setting the average of statistically significant estimates never gets close to the true effect. For large bandwidths, the omitted variable biases the effect while for small bandwidths, the small effective sample size creates exaggeration issues. The optimal bandwidth literature describes a similar trade-off but with different consequences (Imbens and Kalyanaraman 2012). They consider a bias-precision trade-off, I consider an omitted variable bias-exaggeration bias trade-off. Here, the parameter mediating the trade-off can directly be adjusted in a continuous way by the researchers and the more we reduce one of these two biases, the more we increase the other.

Figure 2.4: Evolution of the Bias with Bandwidth Size in Regression Discontinuity Design, conditional on significance.



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

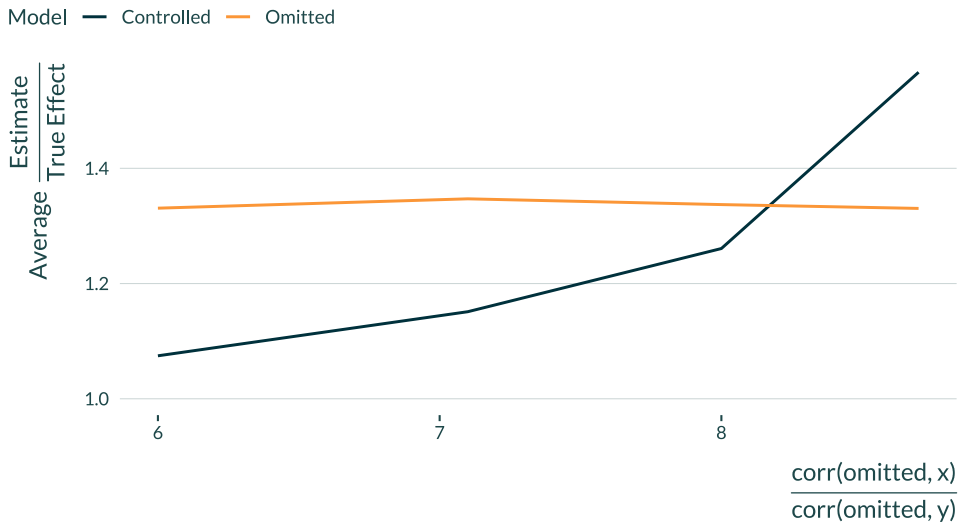
2.4.2 Controlling for confounders

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 2.3, controlling for an additional variable may increase the variance of the estimator if it absorbs more variation in the explanatory variable of interest than in

the outcome variable. The same reasoning applies to a cornerstone of causal identification strategies, 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 and create exaggeration.

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 2.2 and 2.3. As in section 2.3, 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.

Figure 2.5: 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.

Results. Figure 2.5 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.

2.4.3 Instrumental Variables Strategy

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

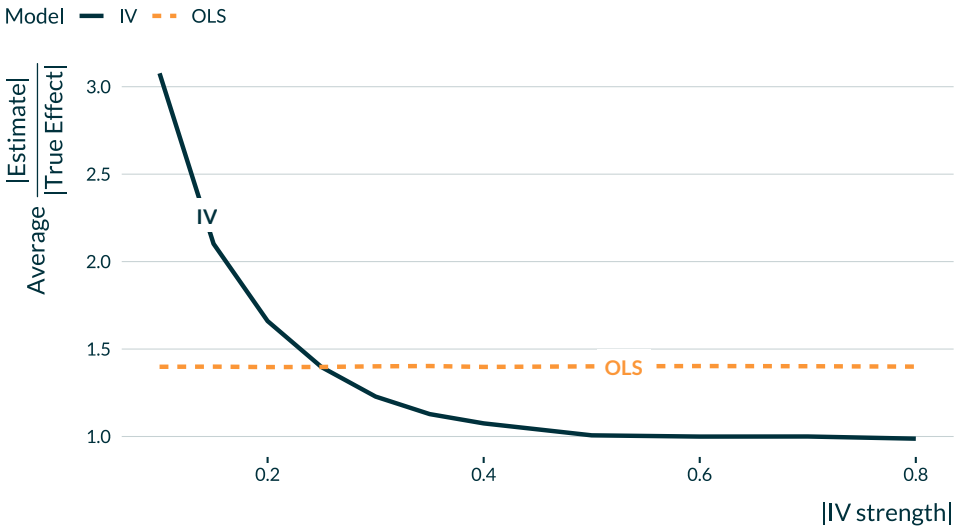
Case-study and simulation procedure. To illustrate this trade-off and its drivers, I consider the example of the impact of voter turnout on election results. To avoid the threat of confounding in this setting, existing studies take advantage of exogenous factors such as rainfall that affect voter turnout[7].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

To make the simulations realistic, I calibrate them on existing studies. I derive sample size, distribution parameters and effect sizes from a set of studies using similar variables (Gomez, Hansford, and Krause 2007; Fujiwara, Meng, and Vogl 2016; Cooperman 2017). Details on the calibration choices are available on the project's website. For each value of IV strength considered, I create 1000 datasets of 30,000 observations. I run both a naive OLS

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

and a 2SLS model to estimate the impact of voter turnout on republican vote share. SNR obtained are aligned with SNR observed in this literature.

Figure 2.6: Evolution of the Bias of Statistically Significant Estimates Against Strength of the Instrument in the IV Case.



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

Results. Figure 2.6 displays, for different IV strengths, the average of statistically significant estimates scaled by the true effect size for both the IV and the naive regression model. When the instrument is strong, the IV will recover the true effect, contrarily to the naive regression model. Yet, when the IV strength decreases, the exaggeration of statistically significant estimates skyrockets. Even if the intensity of the omitted variable bias is large, for limited IV strengths, the exaggeration ratio can become larger than the omitted variable bias. When the only available instrument is weak, using the naive regression model would, on average, produce statistically significant estimates that are closer to the true effect size than the IV. Of interest for applied research, a large F-statistic does not necessarily attenuate this problem. This result complements limitations around the use of first-stage F-statistics with non-iid errors (Young 2022; Lal et al. 2024).

2.4.4 Exogenous shocks

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

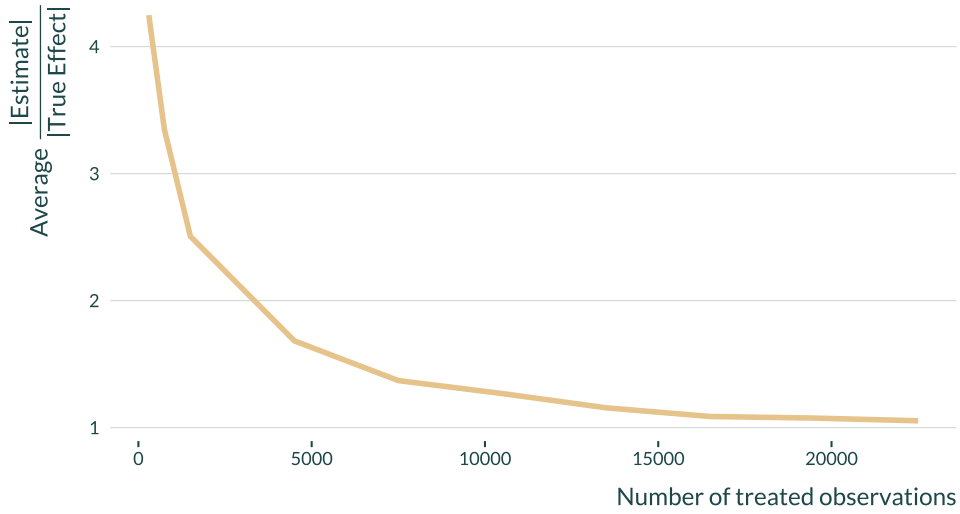
Case-study and simulation procedure. To illustrate this trade-off, I simulate a study of the impact of air pollution reduction on newborn weight of babies. To avoid confounding, one can exploit exogenous shocks to air pollution such as plant closures, creation of a low emission zone or of an urban toll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

I derive parameters values from studies from the literature (Currie et al. 2015; Neidell 2017). The simulations mimic the design of such studies, considering a similar number of observations (150,000), distributions of variables, treatment allocation procedure and treat-

ment effect sizes. I generate datasets with an increasing number of treated observations, varying the proportion of treated units and estimate the correct two-way fixed effects model. The simulation procedure is described in details on the project's website.

Results. Figure 2.7 displays the results of these simulations. As expected, exaggeration is larger for a smaller number of treated observations, keeping every thing else constant. Even though the actual sample size is extremely large in this example, if the number of treated observations is small, as it can be the case in this literature, exaggeration can be substantial. A very large number of observations does not necessarily shield us from exaggeration.

Figure 2.7: Evolution of Bias With the Number of Treated Observations, for Statistically Significant Estimates, in the Exogenous Shocks Case



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

2.5 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.

2.5.1 Gauging omitted variable bias

On one side of the trade-off lies the widely discussed bias caused by confounders. Although it is in essence impossible to measure, tools such as sensitivity analyses are available to gauge its magnitude (Rosenbaum 2002; Middleton et al. 2016; Oster 2019; Cinelli and Hazlett 2020). For instance, the method developed in Cinelli and Hazlett (2020) enables assessing how strong confounders would have to be to change the estimate of the treatment effect beyond a given level we are interested in. It offers bounds for the strength of the association between the treatment and potential omitted variables by weighting it against the measured association between the treatment and observed covariates. A typical conclusion from such an analysis would be: "omitted variables would have to explain as much residual variance of the outcome and the treatment as the observed covariate x (age for instance) to bring down the estimate to a value of β_l". 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

2.5.2 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, Glennerster, and Kremer 2007; McConnell and Vera-Hernández 2015; Athey and Imbens 2016). They are, however, rarely reported in non-experimental studies. Yet, taking publication bias and the threat of exaggeration into account highlights the necessity of running power calculations in non-experimental studies as well. A low power or a relatively large variance not only makes it more difficult to detect an effect or to draw clear conclusions about its magnitude when detected but it can also create a bias. To avoid this bias, I advocate to make power more central to non-experimental analyses. Currently, in causal inference textbooks, very few pages are devoted to statistical power in non-experimental studies (Angrist and Pischke 2009; Angrist and Pischke 2014; Imbens and Rubin 2015; Cunningham 2021). To the best of my knowledge, only two textbooks discuss the matter in depth (Shadish, Cook, and Campbell 2002; Huntington-Klein 2021). Results from power and exaggeration calculations would not only be highly informative but would also be very easy to report in the robustness section of articles.

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, Hill, and Vehtari 2020; Black et al. 2022). Simulating a data generating process from scratch requires thinking about the distribution of the variables, about their relationships and can also help underline the variation used for identification. I implemented such Monte Carlo simulations in Section 3.4.3. The replication material and `R` code I provide can be used as an example to implement simulations for most causal identification strategies, based on data generated from scratch. In situations where the relationships among covariates are too complex to emulate, one can also start from an existing dataset and add a known treatment effect. I implemented real-data simulations in a companion paper and describe their implementation in its replication material (Bagilet 2023a).

Retrospective power calculations

Running post-analysis power calculations can also help getting a sense of the statistical power associated with a research design, as I did in section 2.2. Such *retrospective* calculations allow evaluating whether the design of the study would produce accurate and uninflated statistically significant estimates if the true effect was in fact smaller than the observed estimate (Gelman and Carlin 2014; Ioannidis, Stanley, and Doucouliagos 2017; Stommes, Aronow, and Sävje 2023).

I illustrate how a retrospective analysis works by taking the example of Card (1993) on the relationship between human capital and income. This paper finds that an additional year of education, instrumented by the distance of growing up near a four-year college, causes a 13.2% average increase in wage. 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 detecting 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).[8] Statistically significant estimates (at the 5% level) would on average be roughly equal to 15%, therefore overestimating the true effect 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).[9] When such information is not available, power calculations can be ran for

[8] Timm (2019) and Linden (2019) offer `R` and `Stata` packages that enable easily running these calculations through an extremely short command: `retrodesign(10, 5.5)`.

[9] Note that when such meta-analyses are available, one can use a Bayesian procedure to shrink statistically

a range of smaller but credible effect sizes that can be for instance derived from theoretical findings. It is also possible to evaluate whether the design of our study would be sufficient to detect smaller effects than the point estimate obtained.

Imprecision matters after obtaining a significant estimate

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 no 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 invites to pay attention to the implications of our design and modelling choices on the variance of our estimators, even if a large variance did not prevent us from obtaining a statistically significant estimate.

2.5.3 The variation used for identification

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. The measure I outline here both leverages the interpretation of causal

significant estimates based on the corpus of estimates from prior studies Zwet and Gelman (2021), Zwet and Cator (2021), and Zwet, Schwab, and Senn (2021) .

inference methods as control approaches and builds on a procedure developed in Aronow and Samii (2016a) for a different purpose: evaluating the external validity of standard regressions.

Aronow and Samii (2016a) 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 (2016a) 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.[10] It allows getting 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

[10]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.

and individuals on the y-axis. If the data is geographical, one can directly plot the weights on a map.

2.5.4 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 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, Nelson, and Simonsohn 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 building the actual distribution of the causal estimand of interest. Meta-analyzes would then reduce the uncertainty around the true value of the causal estimand by pooling estimates (Hernán 2022). Finally, the inflation of statistically significant estimates could be limited by

interpreting confidence intervals and not point estimates and thus considering these intervals as compatibility intervals (Shadish, Cook, and Campbell 2002; Amrhein, Trafimow, and Greenland 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.

2.6 Conclusion

The economic literature suffers from an extensive lack of statistical power (Ioannidis, Stanley, and Doucouliagos 2017) and strongly favors statistically significant findings (Rosenthal 1979; Andrews and Kasy 2019; Abadie 2020; Brodeur, Cook, and Heyes 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 $\mathbb{E}[\hat{\beta}] = \beta$ (Gelman and Tuerlinckx 2000; Ioannidis 2008; Gelman and Carlin 2014). It is therefore not surprising that many estimates published in economics have been shown to be considerably exaggerated (Camerer et al. 2016; Ioannidis, Stanley, and Doucouliagos 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 avoiding confounding, they also reduce statistical power and thus increase the risk of exaggeration. The same aspect that makes these methods credible can create another type of bias. A systematic reporting of statistical power calculations and analysis of the variation actually used for identification could help avoiding falling into this exaggeration trap.

Chapter 3: Design of Empirical Studies Given Their Multiple Goals

with Andrew Gelman*, Claire Palandri†, and Yuling Yao‡

Abstract

Causal inference studies generally have multiple goals, aiming not only to estimate "the" average treatment effect but also analyze how it varies across individuals and time, how it impacts multiple outcomes, or how these effects can be extrapolated to other populations. Through a set of examples from multiple literatures in the social sciences, this paper shows how expecting to produce these multiple estimates and the importance of external validity can orient choices at the design stage, to ensure the study—be it a survey, or an experimental or quasi-experimental setting—is set up for success. We advocate making substantively-motivated assumptions about effect sizes and variation in light of the goals of a study, anticipating and allowing for both uncertainty and heterogeneity in effect sizes. We outline a workflow and provide code to implement design calculations and simulations.

*Department of Statistics and Department of Political Science, Columbia University, New York.
†Harris School of Public Policy, University of Chicago.
‡Center for Computational Mathematics, Flatiron Institute, New York.

3.1 Introduction

While econometrics concerned with causal identification typically focuses on estimating "the" average treatment effect, causal inference studies actually produce—or are used as if they produce—multiple estimates of effects. Treatment effects being heterogeneous, we are interested in estimating individual effects or effects on subgroups (we often not only wonder "Does the treatment work?" but also "Where and when does it work, and where and when does it hurt?"); we aim evaluating the impact of the treatment on multiple outcomes; we wish to extend our results to other settings, using them as if having some external validity.[1] In the remainder of the paper, we shall refer to this idea that studies not only aim to estimate "the" average treatment effect, but also to analyze its variations across individuals and time, assess its impact on multiple outcomes and often try to generalize their conclusions to other populations as studies having "multiple goals". This term does not refer to statistical goals—such as the efficiency, robustness, or unbiasedness of an estimator—but to the multiple intended uses of the products of a study.[2]

Being aware of these multiple goals, invites us to be proactive at the design stage to set our empirical studies to be able to meet these various objectives. While the range of possible actions at the design stage differs largely across settings, especially between studies where researchers are directly involved in data collection and measurement (as in experiments or surveys) and studies where they select or leverage existing data (such as in quasi-experiments), in all these different empirical settings, there are steps that can be taken to ensure that the study is set up to measure, model, and estimate what is needed for these multiple goals.

[1] The importance of designs that aim for external validity, notably to inform policy, and the risk of undermining social sciences if they do not, have been highlighted, in the context of field experiments (List 2018), but also of quasi-experimental studies (Bishop et al. 2020). In effect, quasi-experimental papers also often contain claims that would require some external validity.

[2] Products of economics studies are typically numerous. A review, in Brodeur, Cook, and Heyes (2020), of articles published in the top-25 economics journals in 2015 and 2018 and using causal identification strategies reveals that the median number of estimates per articles is 19 and that 19.6% of articles report more than 50 estimates.

Here we use the term "design" to refer to decisions of data collection and measurement. It includes but is not limited to sample size in experimental and survey settings and to choices ensuring the exogeneity of the treatment allocation in quasi-experiments. We use "analysis" to refer to estimation and associated questions of statistical inference such as standard errors, hypothesis tests, and estimator properties. In between is modeling, which includes statistical models such as linear or logistic regression and their generalizations, simultaneous equations, and nonparametric versions of all of these, along with structural models that could involve latent unmeasured variables such as utilities and shadow prices.

In this paper, we show through examples how design choices can be retroactively motivated, given the multiple goals of causal inference studies. We consider the cases of surveys, experimental and quasi-experimental analyses, acknowledging their similarities and differences in terms of design-related concerns. As implications extend to modeling as well (for instance, the incorporation of additional predictors and interactions), we cover those aspects of modeling most directly related to the design issues raised.

We take identification and the existence of asymptotically minimum-variance unbiased estimators as a given, not to diminish the importance of those topics but because they have been comprehensively covered elsewhere. Given that, in settings where we implement data collection and measurement ourselves, design may sound simple: hypothesize an effect size, pick an efficient design to minimize standard error, and then choose a sample size high enough to achieve desired statistical power. But it's not so easy! With multiple goals, no single design will in general maximize efficiency, and it will not be possible to have high power for everything. Similarly, even when leveraging existing data, we need to invest resources to assemble these data and bearing the multiple goals in mind can guide us to best allocate these resources. Directly proceeding with the analysis may still produce results but they would be misleading if the design poor.

Such deceptive results arise because of expectations regarding statistical significance or other forms of effective certainty. Such thresholding can be misleading and increase the

risk of errors in our estimates. The aforementioned variation in effects is also difficult to estimate, and we reasonably have to expect and accept uncertainty around it. Thinking about variation and generalization, and accepting uncertainty, will help avoid the theoretical trap of premature optimization, which arises when selecting on statistically significant effects from noisy experiments and implicitly constraining how a study is reported or followed up.

The remainder of the paper is structured as follows. Section 2 describes multiple channels through which failing to account for multiple goals or selecting on statistical significance generates unreliable inferences, and considers how these problems can be addressed at the design stage. Section 3 illustrates this through concrete examples from the empirical social science literature, covering experiments, surveys, and observational studies. Section 4 provides specific guidance and tools for design calculations, and going further, for performing simulations to anticipate and address these issues at the design stage.

3.2 Statistical implications of multiple goals under uncertainty

3.2.1 The problem with low-power studies and statistical significance

Randomized control trials are often funded under the condition that the design proposed should provide a large enough statistical power—typically 80%—and so require such calculations from applicants; see, e.g., Abdul Latif Jameel Poverty Action Lab (2023). Research that uses existing observational data rarely has this restriction. So why not perform a study with, say, 20% power? Yes, then the study would most likely fail to achieve a conclusive result, but there's still a nontrivial chance of success, and if the study is cheap and the gain is potentially large (as is the case with many social interventions and when leveraging existing data), why not? A simple decision analysis could suggest the trade-off is worth it, if 0.2 times the expected benefit exceeds 0.8 times the cost of the study.

The answer is that, if statistically significant results are favored for publication, a low-power study will most likely have large type S (sign) and type M (magnitude) errors: any statistically significant estimate will be a large overestimate and has a high probability of

being in the wrong direction (Ioannidis 2008; Gelman and Carlin 2014; Lu, Qiu, and Deng 2019; Zwet and Cator 2021). Significant estimates drawn from a wide distribution are not only far from the center of this distribution—arguably the true effect—, when the true effect is relatively small they will also be far from 0, creating these type M and S errors. This systematic overestimation is a form of winner's curse (Button et al. 2013b).

Low power arises when effects are relatively small, which is regularly the case for part of the effects considered in a typical economics study. Given the multiple goals, we often study multiple outcomes or subgroups or break down part of our analyses to study mechanisms, ultimately focusing on some relatively small effects. These concerns invite us to make statistical power a central concern to both experimental and non-experimental studies, even *after* a statistically significant estimate has been obtained.

From an inferential standpoint, it would make sense to regularize the estimate in some way, using a Bayesian analysis or an estimated correction for selection bias (Zwet and Gelman 2022). Stepping back, it is a mistake to select statistically-significant results: it would be better to report everything, without an implicit thresholding that just adds noise and an inappropriate air of certainty to the presentation. Another step into this direction could be to embrace uncertainty and interpret confidence intervals (Amrhein, Trafimow, and Greenland 2019; Romer 2020). The problems with selection on statistical significance, as further illustrated by the disappointing results of replication studies in social sciences—including in economics (Camerer et al. 2016)—motivate increased attention to design, beyond its key role in ensuring exogeneity.

This lack of anticipation at the design stage has adverse consequences on the economics literature. Ioannidis, Stanley, and Doucouliagos (2017) finds that nearly 80% of the estimates in a wide range of areas in economics are likely exaggerated by a factor of 2. This results from a combination of low statistical power and selection on significance. The median statistical power of the designs they consider is 18%. Ferraro and Shukla (2020) finds comparable results in environmental economics. Focusing on the case of nudges, DellaVigna and Linos (2022)

highlights that academic papers find effects that are much larger than those found in large experiments ran by nudges companies and explain this by a combination of low power and selective publication. A broader literature underlines a favouring of statistically significant results in economics (Doucouliagos and Stanley 2013; Brodeur et al. 2016; Andrews and Kasy 2019; Vivalt 2019; Brodeur, Cook, and Heyes 2020; Chopra et al. 2024).

3.2.2 Heterogeneity

Treatment effects in social sciences are seldom homogeneous. The familiar expression "average treatment effect" itself represents an implicit acknowledgment of variation in effect size. Heterogeneity invites to include individual characteristics, contextual characteristics, and geography in analysis, to the extent that theory or data suggest important variation over these factors. The need to measure and adjust for potential confounders is obvious in an observational study where imbalance leads directly to bias, but it is also a concern in clean randomized experiments, given the goal of generalization. Similarly, because of the inevitable need to generalize from past data to future decisions, it is almost always a good idea to model conditional on time. The first step in design, then, is to measure and record these variables so they can indeed be modeled and adjusted for.

The next step in design is to consider how the need for adjustment will affect inferences. A rough guideline is that 16 times the sample size is needed to estimate an interaction relative to that needed to estimate a main effect. This rule of thumb comes from four assumptions (Gelman, Hill, and Vehtari 2020):

1. When estimating a main effect and an interaction from balanced data using simple averages (equivalent to least squares regression), the estimate of an interaction has twice the standard error as the estimate of a main effect.
2. It is reasonable to suppose that an interaction will have half the magnitude of a main effect.

3. From 1 and 2 above, we can suppose that the true effect size divided by the standard error is 4 times higher for the interaction than for the main effect.
4. To achieve any desired level of statistical power for the interaction, one will need $4^2 = 16$ times the sample size that would be needed to attain that level of power for the main effect.

Statements 3 and 4 are unobjectionable, but they somewhat limit the implications of the "rule of 16", which does not in general apply to Bayesian or regularized estimates, and does not consider goals other than statistical power or the equivalent aim of estimating an effect with a desired relative precision. Statements 1 and 2 are a bit more subtle. Statement 1 depends on what is considered a "main effect", and statement 2 represents an assumption regarding the specific context of the problem being studied.

We find the rule of 16 to be a useful guideline in the common scenario in which the primary interest lies in a main effect but it also entails concerns for generalization and thus interaction. We should accept at the design stage that we will not in general have adequate precision to estimate interactions with statistical significance, and thus we should not aim to be able to make causal claims based on interaction, in particular in generalization. Conversely, if the estimation of a particular source of variation is considered crucial to a study, then some aspect of data collection and modeling should be baked into the design. For example, if there is concern about how a historical effect varies over time, one could gather data over a wide span of years so that enough data will be available to accurately fit a theoretically-supported linear or autoregressive model.

Heterogeneity also affects design and sample calculations. Classical calculations are typically framed in terms of estimating an average treatment effect. But real-world effects vary over time and space, which among other things implies that standard errors will not simply scale by $1/\sqrt{N}$. Models of variation are thus essential to design even without considering the goals of generalization. Simulations can also help get a sense of the size of standard errors in complex contexts, as discussed in section 3.4. At a theoretical level, conditioning

on time solves a classical problem of the analysis of sequential designs. As usually stated, the paradox is that an adaptive design rule, in which treatment assignment depends on past data (for example, some sort of play-the-winner rule intended either to increase statistical efficiency or to provide better outcomes for the people enrolled in the experiment), changes the frequency properties of the resulting inference while having no effect on the likelihood function. A satisfying resolution of this paradox is that treatment effects can and do vary over time, and once time is included in the analysis, an adaptive design, with its potential for imbalance, creates the need for conditioning on time and then averaging over that variable in any generalizations of interest. To get a sense of this, consider an experimental study of a policy intervention that is becoming less and less effective over time, conducted using an adaptive design such that the new treatment is more likely to be applied near the beginning of the study and the control is more likely to be applied near the end. A naive likelihood-based or Bayesian estimate of the treatment effect will then be optimistic—that is, positively biased—because of the combination of interaction and expected imbalance in the treatment assignment, but if time is included as a covariate, this imbalance will be adjusted for in the model. A similar principle holds for spatial effects such as spillovers.

3.2.3 Modeling and effective design

We have described how the anticipation of modeling heterogeneity motivates certain choices of data collection. Other anticipations regarding uncertainty at the modeling stage have similar implications; in this section we develop those most directly tied to design.

Efficiency and robustness in design. Two orthogonal goals in design are efficiency and robustness to misspecification of the functional form. Suppose we are running a univariate linear regression $y = \beta_0 + \beta_1 x + \mathcal{N}(0, \sigma^2)$, where we have the freedom to design n sampling points x on an interval $[0, 1]$. Conditional on any x, the standard error of the least squares estimate of β_1 is $\sigma/(\sqrt{n} \cdot \text{sd}(x))$. To minimize the estimation variance, we would like to

maximize the variance of x, which is achieved if and only if x is distributed by two point masses $\Pr(x = 0) = \Pr(x = 1) = 0.5$. If the linear model is correctly specified, then this two-point design is optimal in terms of efficiency, i.e., the standard error of the estimate being small given a fixed sampling budget and given a model. However, if the statistical model is wrong, the prediction obtained using this two-point design will have poor generalization on unseen data.

In contrast, another mode of design is aimed at robustness: the data collection is aimed to mimic a randomized experiment as closely as possible such that complexity and concern of model misspecification are both minimized. For example, with categorical covariates, a factorial block design eliminates the influence of extraneous factors and enables direct sample calculation for estimating treatment effects and performing statistical tests. There is no need to model the interactions among factors, but it is the opposite of sample efficiency. In the context of regression design, Huber (1975) demonstrated that the optimal design that minimizes the integrated error of the response function is a set of Chebyshev points.

As a caveat, the reduced-form model is not automatically immune to model misspecification. A difference-in-differences design—an application of randomized block design—is typically regarded as a quasi-randomized experiment. Seemingly, there is no need for a statistical model, and a direct sample calculation of sample difference-in-differences suffices to estimate the treatment effect. However, it is only valid when the treatment effect is additive. For a more robust analysis, we can consider a regression model where the coefficient of the post-interference outcome on the pre-interference outcome is not necessarily 1; see, e.g., Gelman and Vákár (2021).

In practice, both efficiency and robustness to model misspecification are needed. When the linear model is not correctly specified, the two-point design in the linear regression is dangerous for missing the opportunity to falsify the model, and it makes sense to add samples near the middle point $x = 0.5$ (Gelman 2000) or from a uniform distribution; when the level of factors is big, the factorial block design is nearly impossible for its exponential sample

size requirement.

Interactions between modeling and design. Even after the nominal design has been fixed, i.e., after the data has been collected and the sample defined, modeling choices can still affect the effective design and the products of a study though design related channels. For instance, adjusting for covariates or including fixed effects into the model modifies the variation used for identification by partialling out part of this variation. If for some observations the treatment is well explained by variables adjusted for, these observations will not effectively contribute to the estimation of the effect. When effects are heterogeneous, this can have important implications for external validity, the effective sample being different from the nominal sample (Aronow and Samii 2016b). These model-related alterations of the effective design can create power and type-M error issues by affecting the effective sample size notably in the case of causal identification strategies: they only use the exogenous part of the variation and can reduce the size of the effective sample (Bagilet 2023b).

3.2.4 Interactions between multiple goals

More generally, design implications of the multiple goals interact in non-trivial ways.

Acknowledging heterogeneity in treatment effects is key to address one of the central problems in empirical economics: generalizing from an experiment or observational study to the larger population and to different scenarios. With heterogenous treatments, other populations from the one considered may have different underlying treatment effects, which can limit the external validity of our study and limit its generalizability. Hill (2011) and Wager and Athey (2018) propose nonparametric approaches to generalize estimated treatment effects from sample to population.

Heterogeneity in treatment effects can also affect generalizability indirectly: addressing power related issues often invite to modify the design of the study, by expanding its size, considering additional spatial units or a longer time period for instance. Such expansions

can implicitly change the underlying parameter being studied.

As discussed above, adjusting for covariates in response to heterogeneity or implementing causal identification strategies to ensure exogeneity of the treatment allocation can affect statistical power and our ability to accurately capture the average treatment effect. Similarly, when using other causal identification strategies such as Instrumental Variables, Differences-In-Differences or Regression Discontinuity, aiming for one aspect of design—a quasi-random allocation of the treatment—affects other aspects of design such as the effective sample size and statistical power (Bagilet 2023b).

Altering the design of the study while aiming for one of the multiple goals can have implications for others goals.

3.3 Implications of overlooked multiple goals: case studies

In this section, we illustrate through a set of examples the consequences of failing to account for multiple goals at the design stage.

3.3.1 Experimental studies

It is often possible to anticipate the failure to meet the goals of a study caused by a lack of power coupled with a statistical significance filter. In Gelman (2018) we demonstrate this point in the context of an early-childhood intervention experiment in economics that was powered to overestimate any plausible effect. Under reasonable assumptions, the bias resulting from a lack of power was bound to be several times the size of the underlying effect, a problem that is not resolved in any way by multiple comparisons adjustments. Yet solutions at the design stage exist, and they do not boil down to increasing the sample size, for instance if the limiting factor is the amount of noise in measurement. Noisy measurements have led to misleading literatures, and it is a mistake to think that statistical significance implies retroactively that measurements were sufficiently precise. *We will use this example to cover more in-depth the issues with noise in measurement of the different key variables.*

3.3.2 Surveys

We consider a study of the average effects of social networks on political attitudes to illustrate how anticipating heterogeneous effects, and thinking of the different samples within the population, could have been done at the design stage and reoriented the survey design. Gelman and Margalit (2021) conducted a survey asking approximately 2000 respondents how many people they knew from different groups (e.g., gay people, Muslims, people who were unemployed), along with attitudes on related issues (same-sex marriage, immigration restrictions, unemployment benefits, etc.). The study was done in two waves a year apart, allowing an estimate of the effect of entering the "penumbra" of a social group. For example, we estimated the effect of entering the "gay penumbra" by restricting to respondents who reported knowing no gay people at time 1, and comparing those who knew at least one gay person at time 2 to those who still knew none. When the data came in, we were surprised that the signal-to-noise ratio was so low: estimates for the different issues were on the order of 0.10 with standard error 0.05 for survey responses on a 5-point scale.

Could we have anticipated this before conducting the study? We argue that, yes, we can and should have done so. It is somewhat painful to hypothesize the products of a study before collecting the data, but we think it is possible. In this case, start with the outcome variable. For a contentious political issue, many people will already be at the extreme positions of 1 or 5, without any expectation they will budge. If, for example, one-fifth of the population are *susceptible* to shifting their attitudes on same-sex marriage, one might imagine that knowing a gay person could induce a 1-point shift in 10% of the population, hence an average effect of 0.1. Our point here is not that one could deduce this exact value ahead of time—one could easily imagine an average effect of 0.05 or 0.2—but it does give us an order of magnitude, something to use in designing the study.

Now consider the 2000 respondents. At the time of the survey, it was reasonable to suppose that half of them had a friend or family member whom they knew was gay. One might further suppose that 10% of these 1000 people would enter the penumbra during the

year between the two waves of the survey, hence the standard error of a simple comparison between groups would be $\sqrt{\sigma^2/100 + \sigma^2/900} = 0.10\,\sigma$, where σ is the residual standard deviation of a regression predicting response at time 2 from response at time 1. If $\sigma = 0.5$ (implying that individual responses are mostly stable from one wave to the next), this gives a residual standard deviation of 0.05. Again, this is a rough calculation; our point is that when designing the study we could have anticipated a low signal-to-noise ratio.

Suppose we had conducted this analysis ahead of time; how would that have affected our design? Most simply, we would have recognized that 2000 might not be a sufficient sample size to detect effects of interest. Beyond simply throwing more money at the problem, we could have focused on the challenging aspects of the problem: the individual effect size which we assume would be zero for most people, and the relatively small number of people in the "treatment" or exposed group. One way to get more bang for the buck here would be to study a subset of people who are more susceptible to a change in opinion and to a change in penumbra status, for example focusing on young people. Another direction would be to improve precision in measurement by asking about multiple issue attitudes and creating a combined score.

3.3.3 Quasi-experimental studies

Adverse consequences of the lack of anticipation at the design stage have also been highlighted in non-experimental settings. A large share of the studies in Ioannidis, Stanley, and Doucouliagos (2017), the aforementioned review highlighting a limited statistical power and inflation of results in economics, are non-experimental. (Young 2022) also show that economic studies leveraging Instrumental Variables display a noticeably low power.

Other recent literatures in applied economics have highlighted issues which could be analyzed through the lenses of not adjusting for these multiple goals. We take the example of the recent and growing body of work on "two-way fixed effects" (TWFE) estimators.

Consider the canonical difference-in-differences design. Assuming identification assump-

tions hold, in the 2-groups 2-periods setting of a binary treatment D that only switches on and is assigned at the same time for all units, the difference-in-differences estimator is equal to the TWFE estimator, and both are unbiased for the average treatment effect on the treated (ATET). However, in settings with more variety in exposure to treatment (e.g., with many groups and periods, a staggered treatment, a treatment switching off, or non-binary treatments), naive extensions of these methods estimate weighted averages of effects that can be very far from an informative ATET. Indeed, the TWFE estimator may be biased *if treatment effects are not constant across groups or over time* (e.g., a policy becoming more or less effective). Even with all confounders accounted for, the estimator is not robust to the variation of effects across groups or periods.

Chaisemartin and D'Haultfœuille (2022) summarizes the fast-growing literature on this issue, which has largely addressed it as an analysis problem—and hence proposed alternative estimators. Indeed, the TWFE estimator is a specific weighted sum of the ATEs in each treated cell, whose weights vary when the setting deviates from the textbook 2-group 2-period binary treatment setting. For example, Goodman-Bacon (2021) shows that when D is binary, staggered, in static TWFE regressions, $\hat{\beta}_{\text{FE}}$ is a weighted average of all possible 2-group, 2-period DiD estimators in the data, where each weight is a function of the sample size and the subsample variance of treatment. Some of these 2x2 DiDs misuse an early-treated group as control for a late-treated group, which may induce negative weights if the treatment effect varies over time. Alternatively, this issue could be seen as a modeling problem: not allowing for the heterogeneity, with implications for design. In this vein, Wooldridge (2021) highlights that the problem arises not from the TWFE estimator per se but from its application to a restrictive model: one that does not allow for the heterogeneity in treatment effects. He proposes an "extended TWFE" approach which notably interacts the treatment indicator with time or group-time indicators to allow effects to vary across groups or periods.

3.4 Design calculations and simulations

What can be done at the design stage to best set up our studies to meet these multiple goals? This section discusses concrete steps that can be undertaken across the settings of experiments, surveys, and quasi-experiments.

3.4.1 Design calculations: Hypothesizing an effect size and uncertainty

To evaluate a design, one needs to assume an effect size and a standard error. We recommend the following steps. While they are more easily applicable in simple settings and tuned for surveys and experiments, they also provide intuition, a framework and can be transposed to quasi-experimental settings.

Hypothesizing a standard error. With relatively simple data generating processes, it is often not so hard to come up with a reasonable guess for the standard error: with binary outcomes, one can use the binomial distribution; with continuous outcomes, one typically has some sense of the scale of measurement error and individual variation; and it is often possible using historical data to get a sense of how much precision will be gained from regression adjustments.

1. Start by considering the variance in the outcome, which is a trivial $\sqrt{p(1-p)}$ for binary measurements and can be approximated by more general numerical outcomes by hypothesizing some distribution.
2. Hypothesize the decrease in variance due to the treatment and, especially, pre-treatment predictors.
3. For analyses more complicated than a regression on treatment indicator, it can make sense to perform a simulation using a fixed sample size and assumptions about the distributions of predictors and outcomes and work out the standard error that way. We describe this approach in more details in section 3.4.3.

Hypothesizing an effect size. Guessing an effect size is more challenging. An approach we recommend is to hypothesize an average effect size by embedding it in the larger problem of hypothesizing a distribution of effect sizes. We have recommended the use of graphs to visualize variation in effect sizes Gelman, Hullman, and Kennedy (2023). Average effect sizes are commonly less than one might think, because of an availability bias (Tversky and Kahneman 1982): When thinking about effect sizes, it's natural to visualize the effect in a case where it is large, without thinking about all the people and scenarios for which the treatment will not be activated or will have a near-zero effect.

A reasonable hypothetical distribution of effect sizes can be derived from existing empirical findings, with the aid of theory and awareness of variation. We are recommending a multi-step approach:

1. Consider what proportion of individuals will be affected at all. A promotion might never be seen or noticed, a motivational tool might not motivate, some people will be too unprepared to make use of an intervention or too well prepared to need it, and so on.
2. Among those who are affected, consider a range of effect sizes, including some in the negative direction (counterproductive or substitution effects).

    (a) First, construct hypothetical individual effect sizes based on some model of individual behavior. Economic models are one way to do this. It is rare in economics or policy analysis to not have any theory or expectations about potential behaviors under different inputs. For instance, a reasonable hypothetical minimum wage employment effect can be retrieved through theory, after assuming a labor-labor elasticity of substitution. Marginal willingness to pay (MWTPs) from stated preference studies can bound expectations for reasonable MWTP estimates to obtain from revealed preference analyses using hedonic models.

    (b) The second step is to consider variation in effects across individuals in the sample and in population at large. Variation in behavior is central to economic theory—

differences of valuations, goals, and situations motivate economic activity—and it makes sense to include this in design as well.

3. Multiply the fraction with nonzero effect by the average effect among that group to get an average effect.

In addition to being helpful in checking the design of a study, the above steps can provide insights into how to increase the average effect size (by focusing on scenarios with larger expected effects) and reduce the standard error (by adjusting or controlling for more of the variation in the data).

3.4.2 Choices at the design stage

Evidently, the appropriate choice of next steps will depend on the context of the problem. It is often tempting to start by considering an increase in the sample size, as this requires no conceptual effort. In practice, though, many studies are so noisy and possibly biased that increasing sample size is just throwing away resources. In other settings, as with historical data, it is not possible to increase the sample size without simultaneously changing the scope of the study, for example by going back further in time or by including additional geographic areas, while in yet other settings, sample sizes are limited by cost and practicality.

Yet some choices remain, even in quasi-experimental studies, and anticipating during the design stage allows for identifying limiting factors, and subsequently best allocating resources to increase the prospects of success of the study.

At a general scale, strategies beyond increasing the sample size roughly fall into three categories:

- Increasing the effect size, for example by focusing the study on units where the treatment effect is expected to be higher or by increasing take-up of the treatment;

- Decreasing inferential uncertainty, for example using within-unit comparisons, gathering additional pre-treatment information, and measuring outcomes with more precision;

- Integrating empirical models with substantive theory, measuring intermediate outcomes and possibly adjusting the research question.

If none of these can be done, one should accept the limitations of a study and plan to incorporate inferential uncertainty into reporting and decision making. It is good to be aware of this last option at the design stage, rather than considering it as a final option for an otherwise unsuccessful study.

We will develop examples for each of the three settings (experiments; surveys; quasi-experiments) in the paper.

3.4.3 Going further: Monte Carlo simulations

Approach and usefulness

The design calculations aforementioned enable to assess the impact of design in simple settings, but may be of limited use with a more complex data generating process. In such settings, the outcomes of the study can be jointly affected by several design parameters, making mathematical derivations challenging to implement. Simulations enable us to embrace this multidimensionality and to easily study these joint impacts. While simulations do not provide rigorous closed-form links between design and the conclusions of a study, they do enable us to quickly get a sense of the potential design limitations and of the parameters driving these limitations. They can thus inform us about where to best invest resources while still acknowledging multiple goals, e.g., expanding the sample size, in space or time, or focusing on certain outcomes or sub-groups for which the effect size is larger, or on improving the measurement quality of one of the variables. Simulations, as design calculations, are not meant to rule out the existence of issues but may help reveal them if present. They both invite us to think about realistic values for the underlying parameters, of relations between variables, affected units or more broadly about the impact of design on the products of our study.

Simulations enable to build a distribution of the parameter of interest—such as estimates of treatment effects. Combining knowledge of this distribution and of the assumed value of this treatment effect allows to compute measures such as statistical power, type-M, and type-S errors. Varying design parameters then enables us to assess their impact on these various measures. For clarity, the present discussion focuses on simulations aimed at computing power calculations, even though they can be used for other purposes, not necessarily related to design. For instance, these simulations can be used to evaluate modeling choices and help identify modeling issues. Very simple simulations can for instance highlight the failure of a standard TWFE model to accurately estimate the effect of a staggered heterogeneous treatment in a concrete setting.

There are two main approaches to simulations. First, we can build them on top of an existing data set. We add an effect to the data—according to an allocation mechanism mimicking the one we aim to study—and modify the design according to what needs to be tested. Then, we estimate our model on this modified data set with the objective of recovering the effect we added to the data. We compute the quantities of interest and repeat the process. We then vary the parameters of interest and examine their impact. This approach allows to embrace the complexity of the data structure and can be extremely quick to implement. However, to avoid sampling representativity issues and involuntary data mining, it might be preferable to run these simulations on a different data set or sample than the one of interest. This might however raise limitations in light of multiple goals, in particular regarding external validity.

Alternatively, the data can be generated from scratch, approximating the data generating process of the data set studied. There are pro and cons to this approach. It helps abstract from non-modeled features, allowing to focus on the influence of the parameters of interest. Generating variables also invite to think clearly about their distribution and links between them. These simulations also do not require any material to build on top of. However, the generated data set might be too simple to be representative of the one of interest. These

simulations might also be cumbersome to implement, especially when we want to make them complex enough to attenuate the previous concern. They can therefore be implemented layer by layer, starting from a very simple setting—e.g., one outcome, one explanatory variable of interest, an error term, all normally distributed—and then complexifying the distribution of the variables, their correlation structure, adding heterogeneity in the treatment effect or serial correlation for instance.

Implementation A general workflow to implement design simulation is as follows:

1. Model the relationships between the variables. This modeling can be informed by theory and to some extent by the data set being studied;
2. Set parameter values;
3. Generate the fake-data or modify the existing data set. The shape of the distribution of the variables can be inspired by existing studies describing them or derived from the data set of interest;
4. Run the estimation;
5. Repeat steps 1 through 4 to build the whole distribution for this given set of parameters
6. Compute the measure of interest, e.g., bias, statistical power or exaggeration ratio, for this set of parameters;
7. Vary the parameter values, e.g., the effect size, the correlation between the variables, the sample size,
8. Repeat steps 1 through 7 to understand how the measure of interest evolves with the various parameters.

Example

We will illustrate these points more concretely by developing an example, discussing how to implement simulations to analyze the impact of design choices on the conclusions of an example study. In the Appendix, we detail this workflow and provide accompanying code

to implement such simulations, expanding on the simulation procedures developed in Bagilet (2023b) and presented with the corresponding code here on the project website.

3.5 Conclusion

In econometrics, design is presented in an implicit lexicographic order: identification, then unbiasedness, then minimum variance, with robustness to misspecification somewhere in the mix. But all these conditions depend on how a study will be reported and applied. While causal identification typically addresses internal validity, inferences about new individuals and scenarios require identification of interactions as well as main effects. Such variation in effects is both important and difficult to estimate, which implies that our analyses and summaries should not aim for an unrealistic precision. A well-designed study can yield unbiased estimates of the sample average treatment effect but requires additional assumptions to be unbiased for the population average treatment effect. In real-world studies, the average treatment effect is often defined based on the data collection—for example, a study of a certain set of countries in a certain set of years corresponds to some average over country-years—and this implies, among other things, that changes in the design can often implicitly lead to changes in the underlying parameter being studied, in which case the concept of unbiased inference becomes slippery. Similar issues arise with questions of statistical efficiency: asymptotic theory is not so appropriate in studies of historical data in which an increase in sample size often corresponds to an expansion over space and time, thus violating an implicit stability assumption in the usual asymptotic analysis. In addition to all this, the well-known problems of selection on statistical significance destroy naive notions of unbiasedness and robustness.

Throughout this article we have used the concept of multiple goals, and the acceptance of uncertainty, as principles which encompass the aforementioned concerns.

Given this conceptual understanding, we can do a lot at the design stage, which goes beyond increasing the sample size. In this paper we have recommended an approach to design

that more closely weaves together measurement and statistical modeling with substantive theory. The key implications for design are: (1) hypothesizing an average effect size and its variation in the context of the problem being studied; (2) anticipating uncertainty, especially regarding interactions, and not expecting or demanding effective certainty in inferences; and (3) collecting data on individual and contextual covariates to enable generalization to new people and new scenarios in the future. The resulting workflow has a Bayesian feel in that it combines data with subject-matter knowledge, but does not need to be performed using a formal probabilistic approach. Bayesian reasoning can be helpful for multilevel modeling and for propagating inferential uncertainty into decision analysis, but from the perspective of design our main message is to make substantively-motivated assumptions about effect sizes and variation in the light of the goals of a study, rather than to passively rely on narrow notions of internal validity and statistical efficiency. Opening up design in order to consider the eventual uses of a study, which we have proposed to approach by incorporating these considerations and such practices at the design stage, should help avoid the problems associated with the replication crisis—in part by giving more realistic expectations for future inferences and in part by gathering data that will let us do better—but also foster research efficiency, by focusing the researcher's time on the data collection and cleaning efforts that are revealed as the limiting factors.

References

Abadie, Alberto (2020). "Statistical nonsignificance in empirical economics". In: *American Economic Review: Insights* 2.2, pp. 193–208.

Abdul Latif Jameel Poverty Action Lab (2023). *Proposal Guidelines: RCTs*. https://drive.google.com/file/d

Adhvaryu, Achyuta, Namrata Kala, and Anant Nyshadham (2022). "Management and shocks to worker productivity". In: *Journal of Political Economy* 130.1, pp. 1–47.

Altoè, Gianmarco et al. (2020). "Enhancing Statistical Inference in Psychological Research via Prospective and Retrospective Design Analysis". In: *Frontiers in Psychology* 10, p. 2893.

Amrhein, Valentin, David Trafimow, and Sander Greenland (2019). "Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis If We Don't Expect Replication". In: *The American Statistician* 73.sup1, pp. 262–270.

Anderson, Michael L, Minwoo Hyun, and Jaecheol Lee (2022). *Bounds, Benefits, and Bad Air: Welfare Impacts of Pollution Alerts*. Tech. rep. National Bureau of Economic Research.

Andrews, Isaiah and Maximilian Kasy (2019). "Identification of and correction for publication bias". In: *American Economic Review* 109.8, pp. 2766–2794.

Angrist, Joshua D. and Guido W. Imbens (1995). "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity". In: *Journal of the American Statistical Association* 90.430, pp. 431–442.

Angrist, Joshua D. and Jörn-Steffen Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. 1 edition. Princeton: Princeton University Press.

— (2014). *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press.

Arceo, Eva, Rema Hanna, and Paulina Oliva (2016). "Does the Effect of Pollution on Infant Mortality Differ Between Developing and Developed Countries? Evidence from Mexico City". In: *The Economic Journal* 126.591, pp. 257–280.

Arel-Bundock, Vincent et al. (2022). *Quantitative Political Science Research Is Greatly Underpowered*. Working Paper 6. I4R Discussion Paper Series.

Aronow, Peter M. and Cyrus Samii (2016a). "Does Regression Produce Representative Estimates of Causal Effects?" In: *American Journal of Political Science* 60.1, pp. 250–267.

Aronow, Peter M and Cyrus Samii (2016b). "Does regression produce representative estimates of causal effects?" In: *American Journal of Political Science* 60.1, pp. 250–267.

Athey, Susan and Guido Imbens (2016). "The Econometrics of Randomized Experiments". In: *arXiv:1607.00698 [econ, stat]*.

Baccini, Michela et al. (2017). "Assessing the short term impact of air pollution on mortality: a matching approach". In: *Environmental Health* 16.1, p. 7.

Bagilet, Vincent (2023a). *Accurate Estimation of Small Effects: Illustration Through Air Pollution and Health*. https://vincentbagilet.github.io/inference_pollution/inference_pollution_paper.pdf.

— (2023b). *Causal Exaggeration: Unconfounded but Inflated Causal Estimates*. https://vincentbagilet.github.io/causal_exaggeration/causal_exaggeration_paper.pdf. Accessed: 2024-02-21.

Barwick, Panle Jia et al. (2018). *The Morbidity Cost of Air Pollution: Evidence from Consumer Spending in China*. Tech. rep. w24688. Cambridge, MA: National Bureau of Economic Research, w24688.

Bauernschuster, Stefan, Timo Hener, and Helmut Rainer (2017). "When Labor Disputes Bring Cities to a Standstill: The Impact of Public Transit Strikes on Traffic, Accidents, Air Pollution, and Health". In: *American Economic Journal: Economic Policy* 9.1, pp. 1–37.

Bell, Michelle L, Jonathan M Samet, and Francesca Dominici (2004). "Time-series studies of particulate matter". In: *Annu. Rev. Public Health* 25, pp. 247–280.

Bind, Marie-Abèle (2019). "Causal Modeling in Environmental Health". In: *Annual Review of Public Health* 40.1, pp. 23–43.

Bishop, Kelly C et al. (2020). "Best Practices for Using Hedonic Property Value Models to Measure Willingness to Pay for Environmental Quality". In: *Review of Environmental Economics and Policy* 14.2, pp. 260–281.

Black, Bernard et al. (2022). "Simulated power analyses for observational studies: An application to the Affordable Care Act Medicaid expansion". In: *Journal of Public Economics* 213.C.

Brodeur, Abel, Nikolai Cook, and Anthony Heyes (2020). "Methods matter: P-hacking and publication bias in causal analysis in economics". In: *American Economic Review* 110.11, pp. 3634–3660.

Brodeur, Abel et al. (2016). "Star Wars: The Empirics Strike Back". In: *American Economic Journal: Applied Economics* 8.1, pp. 1–32.

Button, Katherine S. et al. (2013a). "Power failure: why small sample size undermines the reliability of neuroscience". In: *Nature Reviews Neuroscience* 14.5, pp. 365–376.

Button, Katherine S. et al. (2013b). "Power failure: Why small sample size undermines the reliability of neuroscience". In: *Nature Reviews Neuroscience* 14.5, pp. 365–376.

Camerer, Colin F. et al. (2016). "Evaluating replicability of laboratory experiments in economics". In: *Science* 351.6280, pp. 1433–1436.

Camerer, Colin F. et al. (2018). "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015". In: *Nature Human Behaviour* 2.9, pp. 637–644.

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

Chaisemartin, Clément de and Xavier D'Haultfœuille (2022). "Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey". In: *Econometric Journal* 26.3, pp. C1–C30.

Chen, Hong et al. (2018). "Effect of air quality alerts on human health: a regression discontinuity analysis in Toronto, Canada". In: *The Lancet Planetary Health* 2.1, e19–e26.

Chen, Siyu, Chongshan Guo, and Xinfei Huang (2018). "Air Pollution, Student Health, and School Absences: Evidence from China". In: *Journal of Environmental Economics and Management* 92, pp. 465–497.

Cheung, Chun Wai, Guojun He, and Yuhang Pan (2020). "Mitigating the air pollution effect? The remarkable decline in the pollution-mortality relationship in Hong Kong". In: *Journal of Environmental Economics and Management* 101, p. 102316.

Chopra, Felix et al. (2024). "The Null Result Penalty". In: *The Economic Journal* 134.657, pp. 193–219.

Christensen, Garret and Edward Miguel (2018). "Transparency, Reproducibility, and the Credibility of Economics Research". In: *Journal of Economic Literature* 56.3, pp. 920–980.

Cinelli, Carlos and Chad Hazlett (2020). "Making sense of sensitivity: extending omitted variable bias". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.1, pp. 39–67.

Cooperman, Alicia Dailey (2017). "Randomization Inference with Rainfall Data: Using Historical Weather Patterns for Variance Estimation". In: *Political Analysis* 25.3, pp. 277–288.

Cunningham, Scott (2021). *Causal Inference: The Mixtape*. Yale University Press.

Currie, Janet et al. (2015). "Environmental Health Risks and Housing Values: Evidence from 1,600 Toxic Plant Openings and Closings". In: *American Economic Review* 105.2, pp. 678–709.

Deaton, Angus and Nancy Cartwright (2018). "Understanding and Misunderstanding Randomized Controlled Trials". In: *Social Science & Medicine*. Randomized Controlled Trials and Evidence-based Policy: A Multidisciplinary Dialogue 210, pp. 2–21.

Dehejia, Rajeev H. and Sadek Wahba (1999). "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs". In: *Journal of the American Statistical Association* 94.448, pp. 1053–1062.

DellaVigna, Stefano and Elizabeth Linos (2022). "RCTs to Scale: Comprehensive Evidence From Two Nudge Units". In: *Econometrica* 90.1, pp. 81–116.

Deryugina, Tatyana et al. (2019). "The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction". In: *American Economic Review* 109.12, pp. 4178–4219.

Di, Qian et al. (2017). "Association of Short-term Exposure to Air Pollution With Mortality in Older Adults". In: *JAMA* 318.24, p. 2446.

Dominici, Francesca and Corwin Zigler (2017). "Best Practices for Gauging Evidence of Causality in Air Pollution Epidemiology". In: *American Journal of Epidemiology* 186.12, pp. 1303–1309.

Doucouliagos, Chris and T.d. Stanley (2013). "Are All Economic Facts Greatly Exaggerated? Theory Competition and Selectivity". In: *Journal of Economic Surveys* 27.2, pp. 316–339.

Duflo, Esther, Rachel Glennerster, and Michael Kremer (2007). "Using Randomization in Development Economics Research: A Toolkit". In: *Handbook of Development Economics*. Ed. by T. Paul Schultz and John A. Strauss. Vol. 4. Elsevier, pp. 3895–3962.

Ebenstein, Avraham, Eyal Frank, and Yaniv Reingewertz (2015). "Particulate Matter Concentrations, Sandstorms and Respiratory Hospital Admissions in Israel". In: *The Israel Medical Association journal* 17, p. 6.

Ebenstein, Avraham, Victor Lavy, and Sefi Roth (2016). "The Long-Run Economic Consequences of High-Stakes Examinations: Evidence from Transitory Variation in Pollution". In: *American Economic Journal: Applied Economics* 8.4, pp. 36–65.

Fan, Maoyong, Guojun He, and Maigeng Zhou (2020). "The winter choke: Coal-Fired heating, air pollution, and mortality in China". In: *Journal of Health Economics* 71, p. 102316.

Fan, Maoyong and Yi Wang (2020). "The impact of PM2.5 on mortality in older adults: evidence from retirement of coal-fired power plants in the United States". In: *Environmental Health* 19.1, p. 28.

Ferraro, Paul J. and Pallavi Shukla (2020). "Feature—Is a Replicability Crisis on the Horizon for Environmental and Resource Economics?" In: *Review of Environmental Economics and Policy* 14.2, pp. 339–351.

Forastiere, Laura, Michele Carugno, and Michela Baccini (2020). "Assessing short-term impact of PM10 on mortality using a semiparametric generalized propensity score approach". In: *Environmental Health* 19.1, p. 46.

Fujiwara, Thomas, Kyle Meng, and Tom Vogl (2016). "Habit Formation in Voting: Evidence from Rainy Elections". In: *American Economic Journal: Applied Economics* 8.4, pp. 160–188.

Gelman, Andrew (2000). "Should we take measurements at an intermediate design point?" In: *Biostatistics* 1.1, pp. 27–34.

— (2018). "The failure of null hypothesis significance testing when studying incremental changes, and what to do about it". In: *Personality and Social Psychology Bulletin* 44.1, pp. 16–23.

Gelman, Andrew and John Carlin (2014). "Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors". In: *Perspectives on Psychological Science* 9.6, pp. 641–651.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari (2020). *Regression and Other Stories*. Cambridge University Press.

Gelman, Andrew, Jessica Hullman, and Lauren Kennedy (2023). "Causal quartets: Different ways to attain the same average treatment effect". In: *American Statistician*, pp. 1–6.

Gelman, Andrew and Eric Loken (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*. http://stat.columbia.edu/~gelman/research/unpublished/forking.pdf.

Gelman, Andrew and Yotam Margalit (2021). "Social penumbras predict political attitudes". In: *Proceedings of the National Academy of Sciences* 118.6, e2019375118.

Gelman, Andrew and Francis Tuerlinckx (2000). "Type S Error Rates for Classical and Bayesian Single and Multiple Comparison Procedures". In: *Computational Statistics* 15.3, pp. 373–390.

Gelman, Andrew and Matthijs Vákár (2021). "Slamming the sham: A Bayesian model for adaptive adjustment with noisy control data". In: *Statistics in Medicine* 40.15, pp. 3403–3424.

Giaccherini, Matilde, Joanna Kopinska, and Alessandro Palma (2021). "When particulate matter strikes cities: Social disparities and health costs of air pollution". In: *Journal of Health Economics* 78, p. 102478.

Godzinski, Alexandre, M Suarez Castillo, et al. (2019). *Short-term health effects of public transport disruptions: air pollution and viral spread channels*. Tech. rep. Institut National de la Statistique et des Etudes Economiques.

Godzinski, Alexandre and Milena Suarez Castillo (2021). "Disentangling the effects of air pollutants with many instruments". In: *Journal of Environmental Economics and Management* 109, p. 102489.

Gomez, Brad T., Thomas G. Hansford, and George A. Krause (2007). "The Republicans Should Pray for Rain: Weather, Turnout, and Voting in U.S. Presidential Elections". In: *The Journal of Politics* 69.3, pp. 649–663.

Goodman-Bacon, Andrew (2021). "Difference-in-differences with variation in treatment timing". In: *J. Econom.* 225.2, pp. 254–277.

Gray, Wayne B., Ron Shadbegian, and Ann Wolverton (2023). "Environmental Regulation and Labor Demand: What Does the Evidence Tell Us?" In: *Annual Review of Resource Economics* 15.1, pp. 177–197.

Greenland, Sander (2017). "Invited Commentary: The Need for Cognitive Science in Methodology". In: *American Journal of Epidemiology* 186.6, pp. 639–645.

Greenstone, Michael (2002). "The Impacts of Environmental Regulations on Industrial Activity: Evidence from the 1970 and 1977 Clean Air Act Amendments and the Census of Manufactures". In: *Journal of Political Economy* 110.6, pp. 1175–1219.

Griffin, Beth Ann et al. (2021). "Moving beyond the classic difference-in-differences model: a simulation study comparing statistical methods for estimating effectiveness of state-level policies". In: *BMC Medical Research Methodology* 21.1, p. 279.

Guidetti, Bruna, Paula Pereda, and Edson Severnini (2021). ""Placebo Tests" for the Impacts of Air Pollution on Health: The Challenge of Limited Health Care Infrastructure". In: *AEA Papers and Proceedings* 111, pp. 371–375.

Halliday, Timothy J, John Lynham, and Áureo de Paula (2019). "Vog: Using Volcanic Eruptions to Estimate the Health Costs of Particulates". In: *The Economic Journal* 129.620, pp. 1782–1816.

Hanlon, W Walker (2018). "London Fog: A Century of Pollution and Mortality, 1866-1965". In: *The Review of Economics and Statistics*, pp. 1–49.

He, Guojun, Maoyong Fan, and Maigeng Zhou (2016). "The effect of air pollution on mortality in China: Evidence from the 2008 Beijing Olympic Games". In: *Journal of Environmental Economics and Management* 79, pp. 18–39.

He, Guojun, Tong Liu, and Maigeng Zhou (2020). "Straw burning, PM2.5, and death: Evidence from China". In: *Journal of Development Economics* 145, p. 102468.

Hernán, Miguel A and James M Robins (2020). *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Hernán, Miguel A. (2022). "Causal analyses of existing databases: no power calculations required". In: *Journal of Clinical Epidemiology* 144, pp. 203–205.

Herrnstadt, Evan et al. (2021). "Air Pollution and Criminal Activity: Microgeographic Evidence from Chicago". In: *American Economic Journal: Applied Economics* 13.4, pp. 70–100.

Hill, Jennifer L. (2011). "Bayesian nonparametric modeling for causal inference". In: *Journal of Computational and Graphical Statistics* 20.1, pp. 217–240.

Huber, Peter J. (1975). "Robustness and designs". In: *A Survey of Statistical Design and Linear Models*. Ed. by J. N. Srivastava. North Holland, pp. 287–303.

Huntington-Klein, Nick (2021). *The Effect: An Introduction to Research Design and Causality*. 1st ed. Boca Raton: Chapman and Hall/CRC.

Imbens, Guido and Karthik Kalyanaraman (2012). "Optimal Bandwidth Choice for the Regression Discontinuity Estimator". In: *The Review of Economic Studies* 79.3, pp. 933–959.

Imbens, Guido W. and Donald B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge: Cambridge University Press.

Ioannidis, John P. A. (2008). "Why most discovered true associations are inflated". In: *Epidemiology* 19.5, pp. 640–648.

Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos (2017). "The power of bias in economics research". In: *Economic Journal* 127, F236–F265.

Jacob, Brian A. and Lars Lefgren (2004). "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis". In: *The Review of Economics and Statistics* 86.1, pp. 226–244.

Jans, Jenny, Per Johansson, and J. Peter Nilsson (2018). "Economic status, air quality, and child health: Evidence from inversion episodes". In: *Journal of Health Economics* 61, pp. 220–232.

Jia, Ruixue and Hyejin Ku (2019). "Is China's Pollution the Culprit for the Choking of South Korea? Evidence from the Asian Dust". In: *The Economic Journal* 129.624, pp. 3154–3188.

Kasy, Maximilian (2021). "Of Forking Paths and Tied Hands: Selective Publication of Findings, and What Economists Should Do about It". In: *Journal of Economic Perspectives* 35.3, pp. 175–192.

Kim, Moon Joon (2021). "Air Pollution, Health, and Avoidance Behavior: Evidence from South Korea". In: *Environmental and Resource Economics* 79.1, pp. 63–91.

Knittel, Christopher R., Douglas L. Miller, and Nicholas J. Sanders (2016). "Caution, Drivers! Children Present: Traffic, Pollution, and Infant Health". In: *Review of Economics and Statistics* 98.2, pp. 350–366.

Kraft, Matthew A. (2020). "Interpreting Effect Sizes of Education Interventions". In: *Educational Researcher* 49.4, pp. 241–253.

Lal, Apoorva et al. (2024). "How Much Should We Trust Instrumental Variable Estimates in Political Science? Practical Advice Based on 67 Replicated Studies". In: *Political Analysis*, pp. 1–20.

Le Tertre, A et al. (2002). "Short-term effects of particulate air pollution on cardiovascular diseases in eight European cities". In: *Journal of Epidemiology & Community Health* 56.10, pp. 773–779.

Linden, Ariel (2019). *RETRODESIGN: Stata Module to Compute Type-S (Sign) and Type-M (Magnitude) Errors.* Boston College Department of Economics.

Lipsey, Mark W and David B Wilson (2001). *Practical meta-analysis.* SAGE publications, Inc.

List, John A (2018). "In vivo we trust". In: *Science* 361.6400, pp. 339–339.

Liu, Cong et al. (2019). "Ambient Particulate Air Pollution and Daily Mortality in 652 Cities". In: *New England Journal of Medicine* 381.8, pp. 705–715.

Liu, Ya-Ming and Chon-Kit Ao (2021). "Effect of air pollution on health care expenditure: Evidence from respiratory diseases". In: *Health Economics* 30.4, pp. 858–875.

Lu, Jiannan, Yixuan Qiu, and Alex Deng (2019). "A note on Type S/M errors in hypothesis testing". In: *British Journal of Mathematical and Statistical Psychology* 72.1, pp. 1–17.

Mayer, Michael (2019). *missRanger: Fast Imputation of Missing Values*. Comprehensive R Archive Network (CRAN).

McConnell, Brendon and Marcos Vera-Hernández (2015). *Going beyond Simple Sample Size Calculations: A Practitioner's Guide*. https://www.econstor.eu/bitstream/10419/119557/1/829211055.pdf. Working Paper.

McShane, Blakeley B. et al. (2019). "Abandon Statistical Significance". In: *The American Statistician* 73.sup1, pp. 235–245.

Middleton, Joel A. et al. (2016). "Bias Amplification and Bias Unmasking". In: *Political Analysis* 24.3, pp. 307–323.

Moretti, Enrico and Matthew Neidell (2011). "Pollution, Health, and Avoidance Behavior: Evidence from the Ports of Los Angeles". In: *Journal of Human Resources* 46.1, pp. 154–175.

Mullins, Jamie and Prashant Bharadwaj (2015). "Effects of Short-Term Measures to Curb Air Pollution: Evidence from Santiago, Chile". In: *American Journal of Agricultural Economics* 97.4, pp. 1107–1134.

Neidell, Matthew (2017). "Energy Production and Health Externalities: Evidence from Oil Refinery Strikes in France". In: *Journal of the Association of Environmental and Resource Economists* 4.2, pp. 447–477.

Open Science Collaboration (2015). "Estimating the reproducibility of psychological science". In: *Science* 349.6251, aac4716.

Oster, Emily (2019). "Unobservable Selection and Coefficient Stability: Theory and Evidence". In: *Journal of Business & Economic Statistics* 37.2, pp. 187–204.

Peng, Roger D and Francesca Dominici (2008). "Statistical methods for environmental epidemiology with R". In: *R: a case study in air pollution and health.*

Peng, Roger D, Francesca Dominici, and Thomas A Louis (2006). "Model choice in time series studies of air pollution and mortality". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 169.2, pp. 179–203.

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

Romer, David (2020). "In Praise of Confidence Intervals". In: *AEA Papers and Proceedings* 110, pp. 55–60.

Rosenbaum, Paul R. (2002). *Observational Studies*. Springer Series in Statistics. New York, NY: Springer New York.

Rosenthal, Robert (1979). "The file drawer problem and tolerance for null results". In: *Psychological Bulletin* 86.3. place: US publisher: American Psychological Association, pp. 638–641.

Rubin, Donald B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." In: *Journal of Educational Psychology* 66.5, pp. 688–701.

— (2001). "Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation". In: *Health Services and Outcomes Research Methodology* 2.3, pp. 169–188.

Samet, Jonathan M et al. (2000). "The national morbidity, mortality, and air pollution study". In: *Part II: morbidity and mortality from air pollution in the United States Res Rep Health Eff Inst* 94.pt 2, pp. 5–79.

Schell, Terry L., Beth Ann Griffin, and Andrew R. Morral (2018). *Evaluating Methods to Estimate the Effect of State Laws on Firearm Deaths: A Simulation Study*. Tech. rep. RAND Corporation.

Schlenker, Wolfram and W. Reed Walker (2016). "Airports, Air Pollution, and Contemporaneous Health". In: *The Review of Economic Studies* 83.2, pp. 768–809.

Schwartz, Joel (1994). "What are people dying of on high air pollution days?" In: *Environmental research* 64.1, pp. 26–35.

Schwartz, Joel, Marie-Abele Bind, and Petros Koutrakis (2017). "Estimating Causal Effects of Local Air Pollution on Daily Deaths: Effect of Low Levels". In: *Environmental Health Perspectives* 125.1, pp. 23–29.

Schwartz, Joel, Kelvin Fong, and Antonella Zanobetti (2018). "A National Multicity Analysis of the Causal Effect of Local Pollution, NO2, and PM2.5 on Mortality". In: *Environmental Health Perspectives* 126.8, p. 087004.

Schwartz, Joel et al. (2015). "Estimating Causal Associations of Fine Particles With Daily Deaths in Boston: Table 1." In: *American Journal of Epidemiology* 182.7, pp. 644–650.

Shadish, William R., Thomas D. Cook, and Donald Thomas Campbell (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton Mifflin.

Shah, Anoop S V et al. (2015). "Short term exposure to air pollution and stroke: systematic review and meta-analysis". In: *BMJ*, h1295.

Sheldon, Tamara L. and Chandini Sankaran (2017). "The Impact of Indonesian Forest Fires on Singaporean Pollution and Health". In: *American Economic Review* 107.5, pp. 526–529.

Simeonova, Emilia et al. (2021). "Congestion Pricing, Air Pollution, and Children's Health". In: *Journal of Human Resources* 56.4, pp. 971–996.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn (2011). "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant". In: *Psychological Science* 22.11, pp. 1359–1366.

Stommes, Drew, P. M. Aronow, and Fredrik Sävje (2023). "On the Reliability of Published Findings Using the Regression Discontinuity Design in Political Science". In: *Research & Politics* 10.2.

Thistlethwaite, Donald L. and Donald T. Campbell (1960). "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment". In: *Journal of Educational Psychology* 51.6, pp. 309–317.

Timm, Andrew (2019). *Retrodesign: Tools for Type S (Sign) and Type M (Magnitude) Errors*. Comprehensive R Archive Network (CRAN).

Tversky, Amos and Daniel Kahneman (1982). "The psychology of preferences". In: *Scientific American* 246.1, pp. 160–173.

Vichit-Vadakan, Nuntavarn, Nitaya Vajanapoom, and Bart Ostro (2008). "The Public Health and Air Pollution in Asia (PAPA) Project: Estimating the Mortality Effects of Particulate Matter in Bangkok, Thailand". In: *Environmental Health Perspectives* 116.9, pp. 1179–1182.

Vivalt, Eva (2019). "Specification Searching and Significance Inflation Across Time, Methods and Disciplines". In: *Oxford Bulletin of Economics and Statistics* 81.4, pp. 797–816.

Wager, Stefan and Susan Athey (2018). "Estimation and inference of heterogeneous treatment effects using random forests". In: *Journal of the American Statistical Association* 113, pp. 1228–1242.

Walker, W. Reed (2011). "Environmental Regulation and Labor Reallocation: Evidence from the Clean Air Act". In: *The American Economic Review* 101.3, pp. 442–447.

Wasserstein, Ronald L. and Nicole A. Lazar (2016). "The ASA Statement on p -Values: Context, Process, and Purpose". In: *The American Statistician* 70.2, pp. 129–133.

Wooldridge, Jeffrey M (2021). *Two-way fixed effects, the two-way Mundlak regression, and difference-in-differences estimators*. https://papers.ssrn.com/abstract=3906345.

Xia, Fan et al. (2022). "The short-term impact of air pollution on medical expenditures: Evidence from Beijing". In: *Journal of Environmental Economics and Management* 114, p. 102680.

Young, Alwyn (2022). "Consistency without Inference: Instrumental Variables in Practical Application". In: *European Economic Review* 147, p. 104112.

Zhong, Nan, Jing Cao, and Yuzhu Wang (2017). "Traffic Congestion, Ambient Air Pollution, and Health: Evidence from Driving Restrictions in Beijing". In: *Journal of the Association of Environmental and Resource Economists* 4.3, pp. 821–856.

Ziliak, Stephen Thomas and Deirdre N. McCloskey (2008). *The cult of statistical significance: how the standard error costs us jobs, justice, and lives*. Economics, cognition, and society. Ann Arbor: University of Michigan Press.

Zwet, Erik and Andrew Gelman (2021). "A Proposal for Informative Default Priors Scaled by the Standard Error of Estimates". In: *The American Statistician*, pp. 1–9.

Zwet, Erik, Simon Schwab, and Stephen Senn (2021). "The Statistical Properties of RCTs and a Proposal for Shrinkage". In: *Statistics in Medicine* 40.27, pp. 6107–6117.

Zwet, Erik van and Andrew Gelman (2022). "A Proposal for Informative Default Priors Scaled by the Standard Error of Estimates". In: *The American Statistician* 76.1, pp. 1–9.

Zwet, Erik W. and Eric A. Cator (2021). "The significance filter, the winner's curse and the need to shrink". In: *Statistica Neerlandica* 75.4, pp. 437–452.

Appendix A: Appendices to Chapter 1

A.1 List of Studies Included in the Causal Inference Literature

We display below studies included in the retrospective analysis of the causal inference literature. We group them by research designs:

Instrumental Variable Design: Moretti and Neidell (2011), Ebenstein, Frank, and Reingewertz (2015), Schwartz et al. (2015), Arceo, Hanna, and Oliva (2016), He, Fan, and Zhou (2016), Knittel, Miller, and Sanders (2016), Schlenker and Walker (2016), Sheldon and Sankaran (2017), Schwartz, Bind, and Koutrakis (2017), Zhong, Cao, and Wang (2017), Barwick et al. (2018), Hanlon (2018), Schwartz, Fong, and Zanobetti (2018), Halliday, Lynham, and Paula (2019), Deryugina et al. (2019), Cheung, He, and Pan (2020), Fan and Wang (2020), He, Liu, and Zhou (2020), Giaccherini, Kopinska, and Palma (2021), Godzinski and Suarez Castillo (2021), Guidetti, Pereda, and Severnini (2021), Kim (2021), Liu and Ao (2021), and Xia et al. (2022)

Reduced-Form Design: Bauernschuster, Hener, and Rainer (2017), Jans, Johansson, and Nilsson (2018), Jia and Ku (2019), and Godzinski, Castillo, et al. (2019)

Regression Discontinuity Design: Chen, Guo, and Huang (2018), Fan, He, and Zhou (2020), and Anderson, Hyun, and Lee (2022)

Event-Study Design: Mullins and Bharadwaj (2015) and Simeonova et al. (2021)

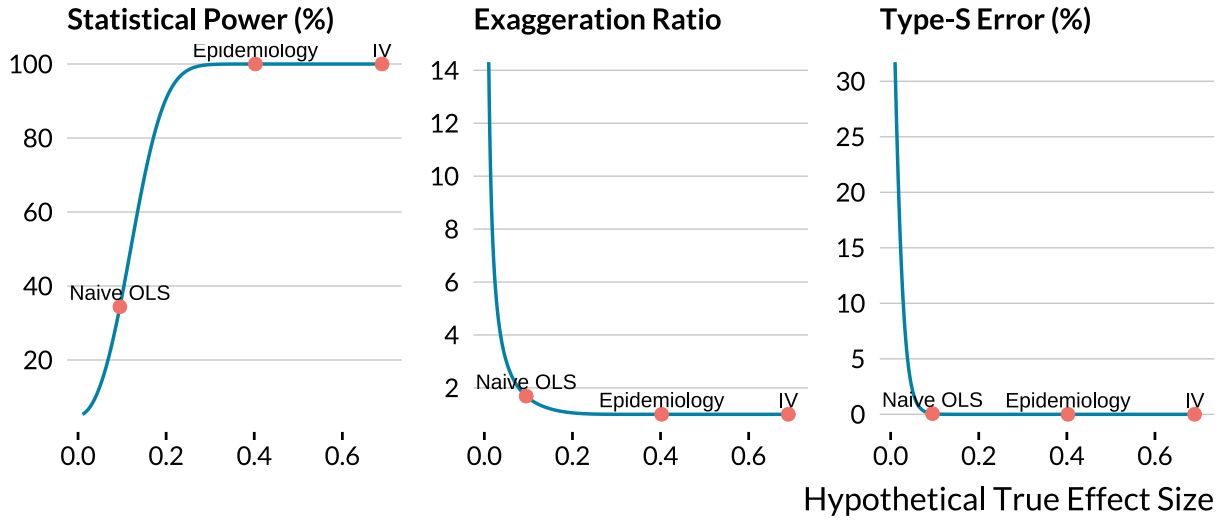
Matching Design: Baccini et al. (2017) and Forastiere, Carugno, and Baccini (2020)

A.2 Implementing a Retrospective Power Analysis

We explain here how we can easily implement a retrospective power analysis once a study is completed. In a flagship publication, Deryugina et al. (2019) instrument $PM_{2.5}$

concentrations with wind directions to estimate its effect on mortality, health care use, and medical costs among the US elderly. They gathered 1,980,549 daily observations at the county-level over the 1999–2013 period; it is one of the biggest sample sizes in the literature. When the authors instrument $PM_{2.5}$ with wind direction, they find that "a 1 $\mu g/m^3$ (about 10 percent of the mean) increase in $PM_{2.5}$ exposure for one day causes 0.69 additional deaths per million elderly individuals over the three-day window that spans the day of the increase and the following two days". The estimate's standard error is equal to 0.061. In Figure A.1, we plot the statistical power, the inflation factor of statistically significant estimates and the probability that they are of the wrong sign as a function of hypothetical true effect sizes.

Figure A.1: Power, Type M and S Errors Curves for Deryugina et al. (2019).



Notes: In each panel, a metric, such as the statistical power, the exaggeration ratio or the probability to make a type S error, is plotted against the range of hypothetical effect sizes. The "IV" label represents the value of the corresponding metric for an effect size equal to Deryugina et al. (2019)'s two-stage least square estimate. The "Epidemiology" label stands for the estimate found in Di et al. (2017), which is the epidemiology article most similar to Deryugina et al. (2019). The " Naive OLS" label corresponds to the estimate found by Deryugina et al. (2019) when the air pollutant is not instrumented.

The estimate found by Deryugina et al. (2019) represents a relative increase of 0.18% in mortality. We labeled it as "IV" in Figure A.1. Is this estimated effect size large compared to those reported in the standard epidemiology literature? We found a similar article to draw a comparison. Using a case-crossover design and conditional logistic regression, Di

et al. (2017) find that a 1 µg/m^3 increase in $PM_{2.5}$ is associated with a 0.105% relative increase in all-cause mortality in the Medicare population from 2000 to 2012. The effect size found by Deryugina et al. (2019) is larger than this estimate labeled as "Epidemiology" in Figure A.1. If the estimate found by Di et al. (2017) was actually the true effect size of $PM_{2.5}$ on elderly mortality, the study of Deryugina et al. (2019) would have enough statistical power to perfectly avoid type M and S errors. Now, suppose that the true effect of the increase in $PM_{2.5}$ was 0.095 additional deaths per million elderly individuals—the estimate the authors found with a "naive" multivariate regression model. The statistical power would be 34%, the probability to make a type S error could be null but the exaggeration factor would be on average equal to 1.7. Even with a sample size of nearly 2 million observations, Deryugina et al. (2019) could make a non-negligible type M error if the true effect size was the naive ordinary least square estimate. Yet, the authors could argue that their instrumental variable strategy leads to a higher effect size as it overcomes unmeasured counfounding bias and measurement error. Besides, for effect sizes down to 0.182 additional deaths per million elderly individuals (a 0.05% relative increase), their study has a very high statistical power and would not run into substantial type M error. A retrospective analysis is thus a very convenient way to think about the statistical power of a study to accurately detect alternative effect sizes.

A.3 Case Studies

The main simulation results help understand how the various parameters influence the statistical power of studies. Yet, these parameters may not perfectly represent actual studies as we made several conservative assumptions: relatively large sample size, proportion of treated units, average outcome counts and instrumental variable strength. For each research design, we therefore consider a realistic set of parameters based on an example from the literature. We then vary the value of key parameters. As we are working with different data, we cannot exactly reproduce the level of precision found in the articles considered. Our goal

is not to claim that the estimates produced by a particular article are inflated, but instead to understand how low power issues could arise for representative parameter values.

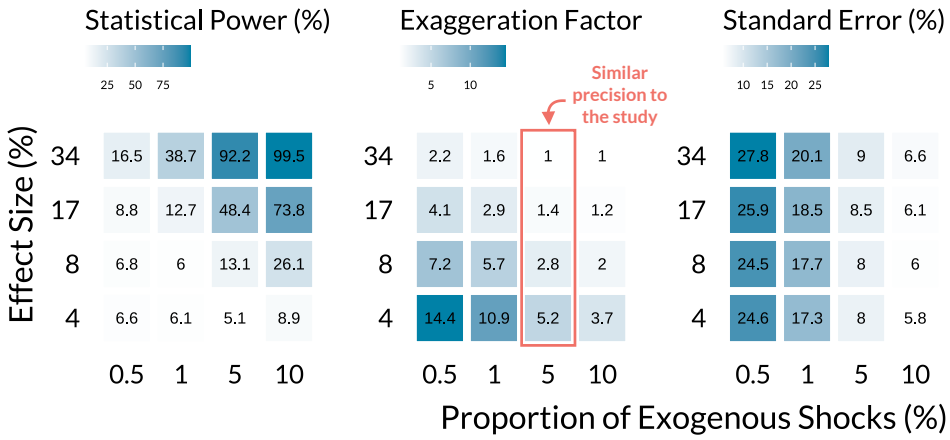
Public Transportation Strikes

Public transportation strikes are unique but rare positive shocks to air pollution as individuals use their cars to reach city centers. Even in a large data set, with several cities and a long study period, the proportion of affected days might be very small. For instance, Bauernschuster, Hener, and Rainer (2017) investigate the effect of public transportation strikes on air pollution and emergency admission in the five biggest German cities over a period of 6 years. Despite a sample size of 11,000, there are only 57 1-day strikes during the study period (0.5% of days are actually treated). The authors find that children hospitalizations for breathing issues increase by 34% (SE=8%) on strike days. On average, 0.22 children per day go to the hospital for breathing issues.

We simulate a similar design with our own data. We first randomly sample 2200 observations for five cities and then vary (i) the proportion of exogenous shocks from 0.5% up to 10%, and (ii) the treatment effect size from a 4% increase up to a 34% increase. We focus on elderly mortality due to chronic obstructive pulmonary disease since it has an average daily count of 0.29 cases.

In Figure A.2, we display our simulation results. The first panel from the left shows that both large effect sizes and a large proportion of exogenous shocks are required to reach adequate power. In the middle panel, we show that a proportion of 0.5% of exogenous shocks is associated with very large exaggeration ratios, from 2.2 for a true effect size of 34% up to 14 for one of 4%. Power issues fade for a combination of a proportion of exogenous shocks above 5% and effect sizes above 17%. In the right panel, we plot the average standard error of the estimates, expressed as a fraction of the average of the health outcome. The standard error of Bauernschuster, Hener, and Rainer (2017)'s is 8%. In our simulations, we recover that specific precision for a proportion of exogenous shocks of 5%. In that case, a true

Figure A.2: Evolution of Power and Exaggeration for Public Transportation Strikes Designs.



Notes: Each panel displays the average value of a metric (power, exaggeration, and standard error) for varying proportions of exogenous shocks and effect sizes. The average standard error of simulations is the raw standard error divided by the mean number of cases of the health outcome. For each combination of parameters, we ran 1000 simulations.

effect size of 34% would not yield inflated estimates. However, if effect sizes are actually smaller and more representative of those found in the literature, the exaggeration would be consequential.

This simulation exercise shows that exaggeration is likely to arise in practice since the proportion of exogenous shocks is low. It occurs even when true effect sizes are relatively large.

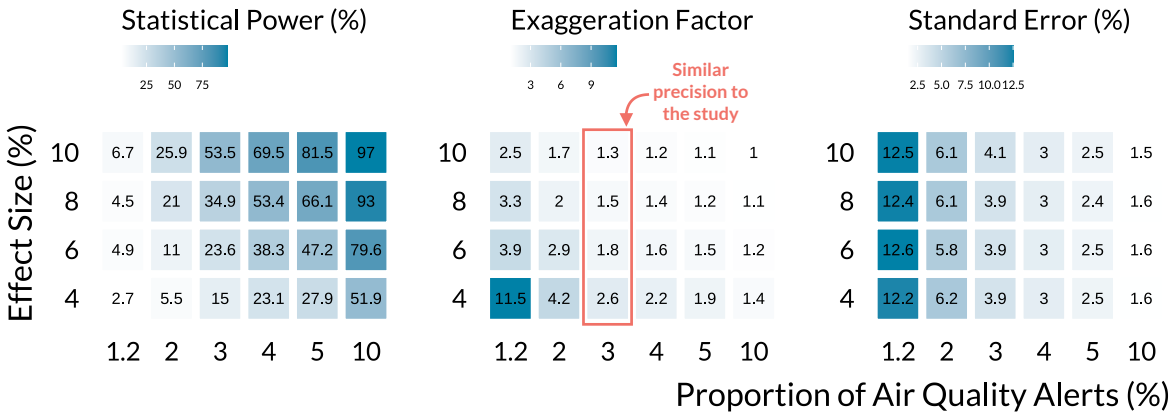
Air Pollution Alerts

Air pollution alerts are also rare events. Their effects are estimated using regression discontinuity designs that restrict the analysis to observations closed to the air quality threshold. As a consequence, the effective sample size may be particularly small. For instance, Chen et al. (2018) investigate the effects of air quality alerts on emergency department visits in Torento, over the 2003-2012 period. While the nominal sample size is 3652, the effective one is only 143 (100 control days and 43 treated days). Only 1.2% of observations are treated. The authors find that eligibility to air quality (the intention-to-treat effect) approximately

reduces emergency visits for asthma by 8% (SE=3.8%). The average daily count of cases of their health outcome is 26.

We approximate the setting of Chen et al. (2018) using our data. We first sample one city for a time period of 3652 days and randomly allocate the treatment. We then repeat the process varying the proportion of alerts and effect sizes. Our outcome variable is the total number of non-accidental deaths since it has a daily mean of 23.

Figure A.3: Evolution of Power and Exaggeration for Air Quality Alerts Designs.



Notes: Each panel display the average value of a metric (power, exaggeration, and standard error) for varying proportions of exogenous shocks and effect sizes. The average standard error of simulations is the raw standard error divided by the mean number of cases of the health outcome. For each combination of parameters, we ran 1000 simulations.

Figure A.3 displays the simulations results. As in Figure A.2, a combination of large effect sizes and many air quality alerts is needed to avoid low power issues. We get a precision similar to Chen et al. (2018) for a proportion of air quality alerts of 3%. For an effect size of 4%, the average exaggeration ratio is equal to 2.6. In that case, the average average of statistically significant estimates is 10%, which is similar to the effect size found by Chen, Guo, and Huang (2018).

Unless true effect sizes are very large, air quality alert designs produce inflated estimates in realistic settings.

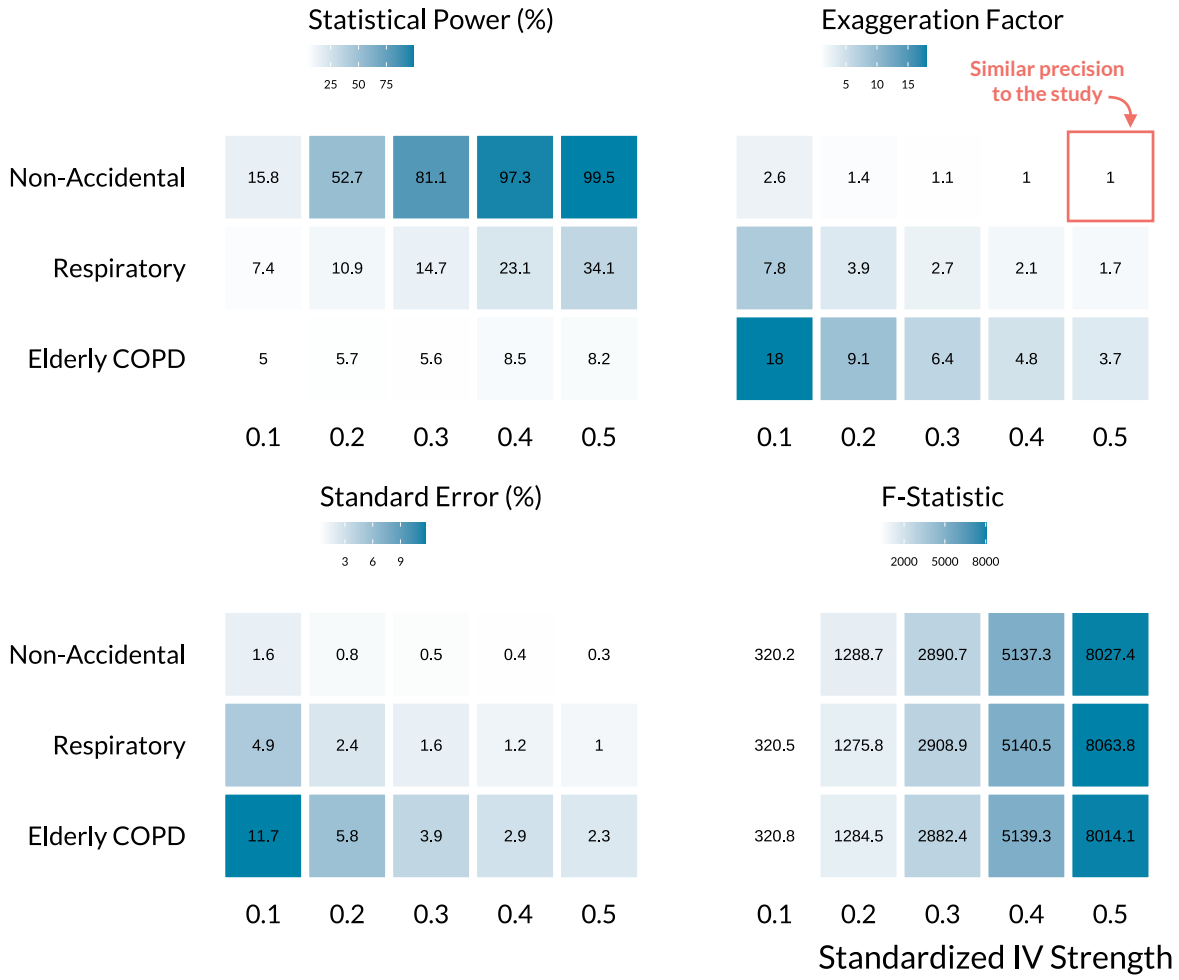
Instrumenting Air Pollution

Finally, we investigate the most commonly used strategy in the causal inference literature, the instrumental variable design. Several studies rely on very large datasets and exploit changes in weather patterns as sources of exogenous variations. For instance, Schwartz, Fong, and Zanobetti (2018) instrument $PM_{2.5}$ concentration with planetary boundary layer, winds speed, and air pressure. Once the effects of seasonal and other weather parameters are accounted for, the combination of their instruments explains 18% of the variation in $PM_{2.5}$ concentration. They find that a 10 $\mu g/m^3$ increase in $PM_{2.5}$ leads to a 1.5% (SE=0.22%) increase in daily non-accidental mortality. There are on average 23 daily deaths in their dataset of 591,570 observations (135 cities with a length of study of approximately 4382 days).

In our simulations, we assess how the strength of the instrumental variable affects power issues for several health outcomes. We consider a binary instrumental variable and vary its effect on air pollution concentration from a 0.1 to a 0.5 standard deviation increase. The 18% correlation in Schwartz, Fong, and Zanobetti (2018) corresponds to a 0.4 standard deviation increase in our case (Lipsey and Wilson 2001). We assume that half of the observations are exposed to exogenous shocks. We set an effect size corresponding to a 1.5% relative increase in three health outcomes with different average number of cases: non-accidental mortality (mean cases of 23), respiratory mortality (mean of 2), and chronic obstructive pulmonary mortality of elderly (mean of 0.3). Our data set being smaller than the one used in Schwartz, Fong, and Zanobetti (2018), we only run simulations for a sample size of 100,000.

In Figure A.4, we see in the top-left panel that power reaches satisfactory level for large instrumental variable strengths but only for non-accidental causes. For respiratory and elderly mortality, exaggeration can be substantial even for large IV strength. While our sample size is large, it is smaller than the one in Schwartz, Fong, and Zanobetti (2018). As a consequence, our simulations only have a precision close to theirs for an instrumental variable strength of 0.5 and non-accidental mortality. Yet, our simulations highlight that

Figure A.4: Evolution of Power and Exaggeration for Instrumental Variable Designs.



Notes: Each panel display the average value of a metric (power, exaggeration, standard error, and first-stage F-statistic.) for varying proportions of exogenous shocks and effect sizes. The average standard error of simulations is the raw standard error divided by the mean number of cases of the health outcome. For each combination of parameters, we ran 1000 simulations.

important exaggeration issues can arise in realistic settings, even for large IV strength. The bottom-right panel of Figure A.4 confirms the result found in the simulations of the previous section: a large first stage F-statistic can be a poor indicator of statistical power issues. For instance, for non-accidental mortality and an IV strength of 0.1, the F-statistic is equal to 320 but the exaggeration factor is 2.6, with an associated power of 16%. Importantly, as the F-statistic does not vary with the number of cases in the outcome it can all the more hide important power issues.

Appendix B: Appendices to Chapter 2

B.1 Mathematical proofs

B.1.1 Variation of the exaggeration ratio (lemma 2.3.1)

Proof 1 *Lu, Qiu, and Deng (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.*

B.1.2 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 2 *Since, we do not observe* w, *we consider the projection of* y *on* X *only:*

$$y = X\boldsymbol{\beta}_{OVB} + u_{OVB} \tag{B.1}$$

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

We first compute the bias of the estimator. From equation B.1 we get:

$$\begin{aligned} & X'y = X'X\boldsymbol{\beta}_{OVB} + X'u_{OVB} \\ \Rightarrow \quad & \mathbb{E}[X'y] = \underbrace{\mathbb{E}[X'X]}_{pos.\ def.}\boldsymbol{\beta}_{OVB} + \underbrace{\mathbb{E}[X'u_{OVB}]}_{0} \\ \Leftrightarrow \quad & \boldsymbol{\beta}_{OVB} = \mathbb{E}[X'X]^{-1}\mathbb{E}[X'(X\boldsymbol{\beta} + \delta w + u)] \quad \textit{cf eq. 2.2} \\ \Leftrightarrow \quad & \boldsymbol{\beta}_{OVB} = \boldsymbol{\beta} + \mathbb{E}[X'X]^{-1}\mathbb{E}[X'w]\delta \end{aligned} \tag{B.2}$$

We then compute the asymptotic distribution. We can write:

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

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

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OVB} - \boldsymbol{\beta}_{OVB}) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}[x_i x_i']^{-1}\mathbb{E}[x_i x_i' u_{OVB,i}^2]\mathbb{E}[x_i x_i']^{-1}\right) \tag{B.3}$$

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

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

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

$$\Rightarrow \quad \mathbb{E}[x_i x_i']^{-1}\mathbb{E}[x_i w_i] = \frac{\gamma\sigma_w^2}{\sigma_x^2}\begin{bmatrix} -\mu_x \\ 1 \end{bmatrix} \tag{B.4}$$

Note that $\mathbb{E}[x_i x_i' u_{OVB,i}^2] \overset{LIE}{=} \mathbb{E}[x_i x_i' \mathbb{E}[u_{OVB,i}^2|x_i]]$. *We thus first compute* $\mathbb{E}[u_{OVB,i}^2|x_i]$, *noting that:*

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

And thus,

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

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

$Var(w_i^{\perp})$,

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

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

$$\mathbb{E}[u_{\text{OVB},i}^2|x_i] = \sigma_u^2 + \delta^2\sigma_w^2\left(1 - \frac{\gamma^2\sigma_w^2}{\sigma_x^2}\right) \tag{B.5}$$

Thus, under our set of assumptions, $\mathbb{E}[u_{\text{OVB},i}^2|x_i]$ *does not depend on* x_i *and* $\mathbb{E}[u_{\text{OVB},i}^2|x_i] = \mathbb{E}[u_{\text{OVB},i}^2]$. *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* $\mathbb{E}[x_i x_i' u_{\text{OVB},i}^2] = \mathbb{E}[x_i x_i' \mathbb{E}[u_{\text{OVB},i}^2|x_i]] = \mathbb{E}[x_i x_i' \sigma_{u_{\text{OVB}}}^2] = \sigma_{u_{\text{OVB}}}^2\mathbb{E}[x_i x_i']$. *And thus* $\mathbb{E}[x_i x_i']^{-1}\mathbb{E}[x_i x_i' u_{\text{OVB},i}^2]\mathbb{E}[x_i x_i']^{-1} = \sigma_{u_{\text{OVB}}}^2\mathbb{E}[x_i x_i']$.

Plugin this and equation B.4 into equation B.3, 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 $\rho_{xw} = corr(x, w) = \frac{cov(\mu_x + \gamma w + \epsilon, w)}{\sigma_x \sigma_w} = \frac{\gamma \sigma_w}{\sigma_x}$*, we have:*

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

B.1.3 Asymptotic distribution of $\hat{\beta}_{\text{CTRL}}$ (lemma 3)

Proof 3 *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* $\boldsymbol{\beta}_{CTRL}$*. We know that we have:*

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

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

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

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

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

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

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

$$\hat{\beta}_{CTRL} \xrightarrow{d} \mathcal{N}\left(\beta_1,\ \frac{\sigma_u^2}{n\ (\sigma_x^2 - \gamma^2 \sigma_w^2)}\right) \tag{B.6}$$

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

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

B.1.4 Asymptotic distribution of $\hat{\beta}_{\text{IV}}$ (lemma 4)

Proof 4 *Since* $u_{IV} = u_{OVB} = \delta w + u$*, we have* $\sigma^2_{u_{IV}} = \sigma^2_u + \delta^2 \sigma^2_w$*. Thus, the usual asymptotic distribution of the IV gives:*

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

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

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

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

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

Thus,

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

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

$$\sqrt{n}\left(\hat{\beta}_{IV} - \beta_1\right) \xrightarrow{d} \mathcal{N}\left(0,\ \frac{\sigma_u^2 + \delta^2\sigma_w^2}{n\ \pi_1^2\sigma_z^2}\right) \tag{B.7}$$

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

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

B.1.5 Asymptotic distribution of $\hat{\beta}_{\text{RED}}$ (lemma 5)

Proof 5 *The proof is straightforward: this is the usual univariate, unbiased case, with and error term equal to $(\delta + \beta_1\gamma)w_i + u_i + \beta_1 e_i$. Since w, u and ϵ_{RED} uncorrelated, its variance is $(\delta + \beta_1\gamma)^2\sigma_w^2 + \sigma_u^2 + \beta_1^2\sigma_e^2$.*

B.2 Matching simulations

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

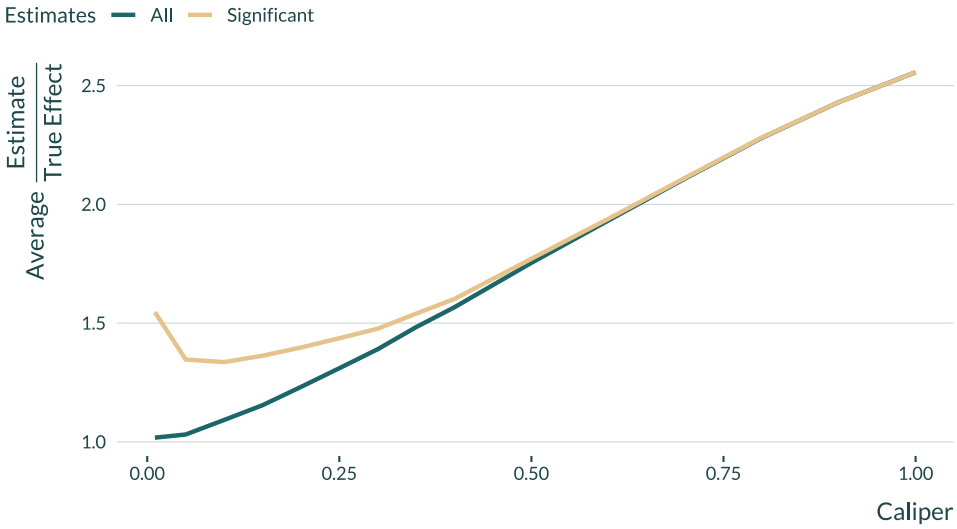
are inflated. In the case of matching, the confounding-exaggeration trade-off is therefore mediated by the value of the caliper.

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

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

Results. Figure B.1 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).

Figure B.1: 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.

B.3 Simulations details for the IV

This section is a reproduction of the analysis for the IV available on the project's website and describing the code in more details.

B.3.1 Intuition

In the case of the IV, the unconfoundedness / exaggeration trade-off is mediated by the 'strength' of the instrument considered. When the instrument only explains a limited portion of the variation in the explanatory variable, the IV can still be successful in avoiding confounders but power can low, potentially leading to exaggeration issues to arise.

B.3.2 Simulation framework

Illustrative example

To illustrate this loss in power, we could consider a large variety of settings, distribution of the parameters or parameter values. I narrow this down to an example setting, considering only one setting and one set of parameter values. I examine an analysis of the impact of voter turnout on election results, instrumenting voter turnout with rainfall on the day of the election. My point should stand in more general settings and the choice of values is mostly for illustration.

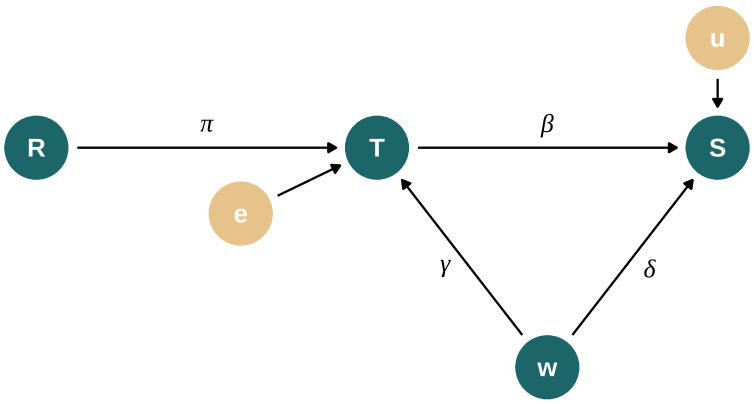
A threat of confounders often arises when analyzing the link between voter turnout and election results. To estimate such an effect causally, one can consider exogenous shocks to voter turnout such as rainfall. Some potential exclusion restriction problems have been highlighted in this instance but I abstract from them and simulate no exclusion restriction violations here.

Modeling choices

For simplicity, I make several assumptions:

- Observations are at the location level,
- Abstract from the panel dimension in this analysis and consider only one time period. This is could be considered as looking at the outcomes of a unique election,
- Only consider the impact of rain on the day of the election,
- Assume no correlation in rainfall between locations. This could be equivalent to considering only a set of remote locations,
- Assume simplify the data generating process and thus do not add any exclusion restriction violations.

The DGP can be represented using the following Directed Acyclic Graph (DAG):



The DGP for the vote share of let's say the republican party in location i, $Share_i$, is defined as follows:

$$Share_i = \beta_0 + \beta_1 Turnout_i + \delta w_i + u_i$$

Where β_0 is a constant, w represents an unobserved variable and $u \sim \mathcal{N}(0, \sigma_u^2)$ noise. β_1 is the parameter of interest. Let's call it 'treatment effect'. Note that parameters names are consistent with the maths section and the other simulation exercises.

The DGP for the turnout data is as follows:

$$Turnout_i = \pi_0 + \pi_1 Rain_i + \gamma w_i + e_i$$

Where π_0 is a constant, $Rain$ is either a continuous variable (amount of rain in location i on the day of the election) or a dummy variable (whether it rained or not) and $e \sim \mathcal{N}(0, \sigma_e^2)$ noise. Let's refer to π_1 as "IV strength".

The impact of voter turnout on election outcome (share of the republican party) is estimated using 2 Stages Least Squares.

More precisely, let's set:

- N the number of observations
- $Rain \sim \text{Gamma}(k, \theta)$, $Rain \sim \mathcal{N}(0, \sigma_R^2)$ or $Rain \sim \text{Bernoulli}(p_R)$ the instrument
- $w \sim \mathcal{N}(0, \sigma_w^2)$ the unobserved variable

- $u \sim \mathcal{N}(0, \sigma_u^2)$
- $e \sim \mathcal{N}(0, \sigma_e^2)$ with σ_e^2 depending on π_1 and defined such that $\sigma_{Turnout}^2$ does not vary when we vary π_1: $\sigma_e^2 = \sigma_{Turnout}^2 - \pi_1^2\sigma_{Rain}^2 - \gamma^2\sigma_w^2$
- For simplicity, I assume that $\delta = -\gamma$. There is no actual basis for that and we may change that in the future. The minus sign is just to get an upward bias, which makes the comparison between OLS and IV easier since the bias and the exaggeration go in the same direction.

If one abstracts from the name of the variable, they can notice that this setting is actually very general.

Data generation

Generating function Let's first write a simple function that generates the data. It takes as input the values of the different parameters and returns a data frame containing all the variables for this analysis.

Note that the parameter `type_rain` describes whether *Rain* is a random sample from a normal or Bernoulli distribution. The distributions of rainfall heights can be approximated with a gamma distribution. The Bernoulli distribution is used if one only consider the impact of rain or no rain on voter turnout. A normal distribution does not represent actual rainfall distributions but is added to run these simulations in other contexts than linking rainfall, voter turnout and election outcomes.

`type_rain` can take the values `gamma`, `bernoulli` or `normal`. `param_rain` represents either σ_R if *Rain* is normal, p_R if it is Bernoulli or a vector of shape and scale parameters for the gamma distribution.

Calibration and baseline parameters' values We can now set baseline values for the parameters to emulate a somehow realistic observational study. I get "inspiration" for the values of parameters from Fujiwara et al. (2016) and Cooperman (2017) who replicates a

work by Gomez et al. (2007).

I consider that:

- **Number of observations**: We consider data at the US county level as in Hansford and Gomez (2010) and Fujiwara et al. (2016). The former use data for presidential elections between 1948 and 2000, restricting their sample of counties to non-Southern ones (2000 per election year). That leads to a sample size of 28000. The latter data for presidential elections between 1952 and 2012 leading to a sample size of about 50000. I thus consider **30000 observations**.
- **Rainfall distribution**: A gamma distribution represents well the distribution of rainfall. Gamma distribution can have two parameters a shape and a scale. The mean is $shape \times scale$ and the variance $shape \times scale^2$. The parameters of the distribution of rainfall are comparable in both Fujiwara et al. and Cooperman (2017) after a conversion into centimeters: mean 2.4 and standard deviation 6.6. I solve the system of mean and variance for shape and scale and set **param_rain to 0.13 and 18**.
- **Standard deviation of the omitted variable** is set to be of the order of magnitude of the error terms. Being conservative, let's set its intensity to be twice as large as the treatment effect.
- **Effect of rainfall on turnout**: Fujiwara et al. find that "The trends specifications suggest that 1 millimeter of rainfall decreases turnout by 0.05–0.07 percentage points" and Gomez et al. (and thus Cooperman) find "a county that receives one inch of rainfall on election day is likely to have approximately 1 percentage point lower voter turnout" which is equivalent to a 1mm increase in rainfall is associated with about a 0.04 percentage points decrease in voter turnout. For simplicity in interpretation, when rainfall is not a dummy, I express in centimeters. So, I consider **pi_1 in the range -0.1 and -1.3**, assuming linearity.
- **Effect of interest** (turnout on vote share): it is subject to intense debate in the literature (cf Shaw and Petrocik (2020) for instance). As underlined by Shaw and

Petrocik and in Fowler (2013), some studies find large effects, others no effects or small effects. Fowler (2013) falls into the large effects category as described by the author himself. The study, for an extremely large shock in voter turnout, compulsory voting, finds "that the policy increased voter turnout by 24 percentage points which in turn increased the vote shares and seat shares of the Labor Party by 7 to 10 percentage points." This correspond to a decrease in Republican vote share of approximately 0.3-0.4 percentage point when turnout increases by 1% (considering that this result is causal and linear). This effect being large, I consider effects that are smaller but of a similar magnitude: I simulate that when turnout increases by 1%, Republican vote share decreases by 0.1 and thus **beta_1 = -0.1**

- **Effect of the omitted variable**. The effect of the omitted variable cannot be observed, I pick it somehow at random such that OVB is substantial without being massive. Since the instrument is valid in these simulations, the IV estimates will not be affected by the intensity of the OVB, only the OLS estimates will (the larger this intensity, the larger the average ratio of the estimate over the true effect).
- **Turnout and vote share** are expressed in percent. I set intercepts and residual standard deviations to produce turnouts and vote shares consistent with Cooperman (2017) and Fujiwara et al. (2016). Voter turnout parameters are roughly similar in both papers (mean 58 sd 14). The mean and standard deviation of Republican vote share are given in Fujiwara et al. (mean 55.3 and sd 14.2). Thus, I set **sigma_share = 14.2 and sigma_turnout = 14**. **pi_0 and beta_0 are manually ajusted** to get the correct mean of turnout and vote share (for pi_1 in the mid-range of its values)

Exploring the distribution of the data I quickly check the standard deviation and means of the variables and at the same time verify that they do not do not change when we vary π_1. They are consistent with what we wanted:

pi_1	share_mean	share_sd	turnout_mean	turnout_sd	rain_mean	rain_sd
0.1	54.07334	14.20147	59.25901	14.06416	2.374203	6.648012
0.2	54.16363	14.35521	59.41936	13.95953	2.323697	6.426635
0.3	54.10888	14.17769	59.74100	14.06694	2.367595	6.484603
0.4	54.04324	14.23257	59.94737	13.97777	2.388056	6.587408
0.5	53.99938	14.13926	60.17718	13.90388	2.236044	6.271868
0.6	53.94911	14.24059	60.41134	14.03049	2.289287	6.400119
0.7	54.05936	14.15616	60.69938	14.09349	2.412667	6.729570
0.8	53.95600	14.15125	60.96088	14.12414	2.386673	6.571114
0.9	53.90390	14.18216	61.16742	14.05391	2.373010	6.649226
1.0	53.86676	14.19336	61.28424	14.01437	2.351829	6.487754

Estimation

After generating the data, we can run an estimation. The goal is to compare the IV and the OLS for different IV strength values. Hence, we need to estimate both an IV and an OLS and return both set of outcomes of interest.

One simulation

We can now run a simulation, combining `generate_data_iv` and `estimate_iv`. To do so I create the function `compute_sim_iv`. This simple function takes as input the various parameters. It returns a table with the estimate of the treatment, its p-value and standard error, the F-statistic for the IV, the true effect, the IV strength and the intensity of the OVB considered (delta). Note that for now, we do not store the values of the other parameters since we consider them fixed over the study.

The output of one simulation, for baseline parameters values is:

estimate	p_value	se	fstat	model	pi_1	delta	param_rain	true_effect
-0.1240442	4e-07	0.0245712	1745.538	IV	-0.5	0.2	0.13, 18.00	-0.1
-0.1465945	0e+00	0.0057604	NA	OLS	-0.5	0.2	0.13, 18.00	-0.1

All simulations

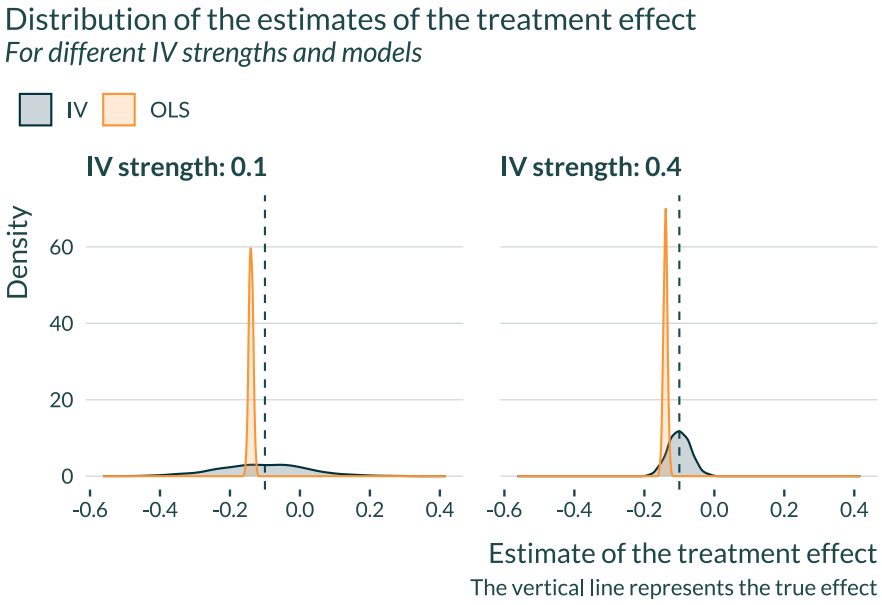
I then run the simulations for different sets of parameters by mapping the `compute_sim_iv` function on each set of parameters. I thus create a table with all the values of the parameters we want to test, `param_iv`. Note that in this table each set of parameters appears `n_iter` times as we want to run the analysis n_{iter} times for each set of parameters.

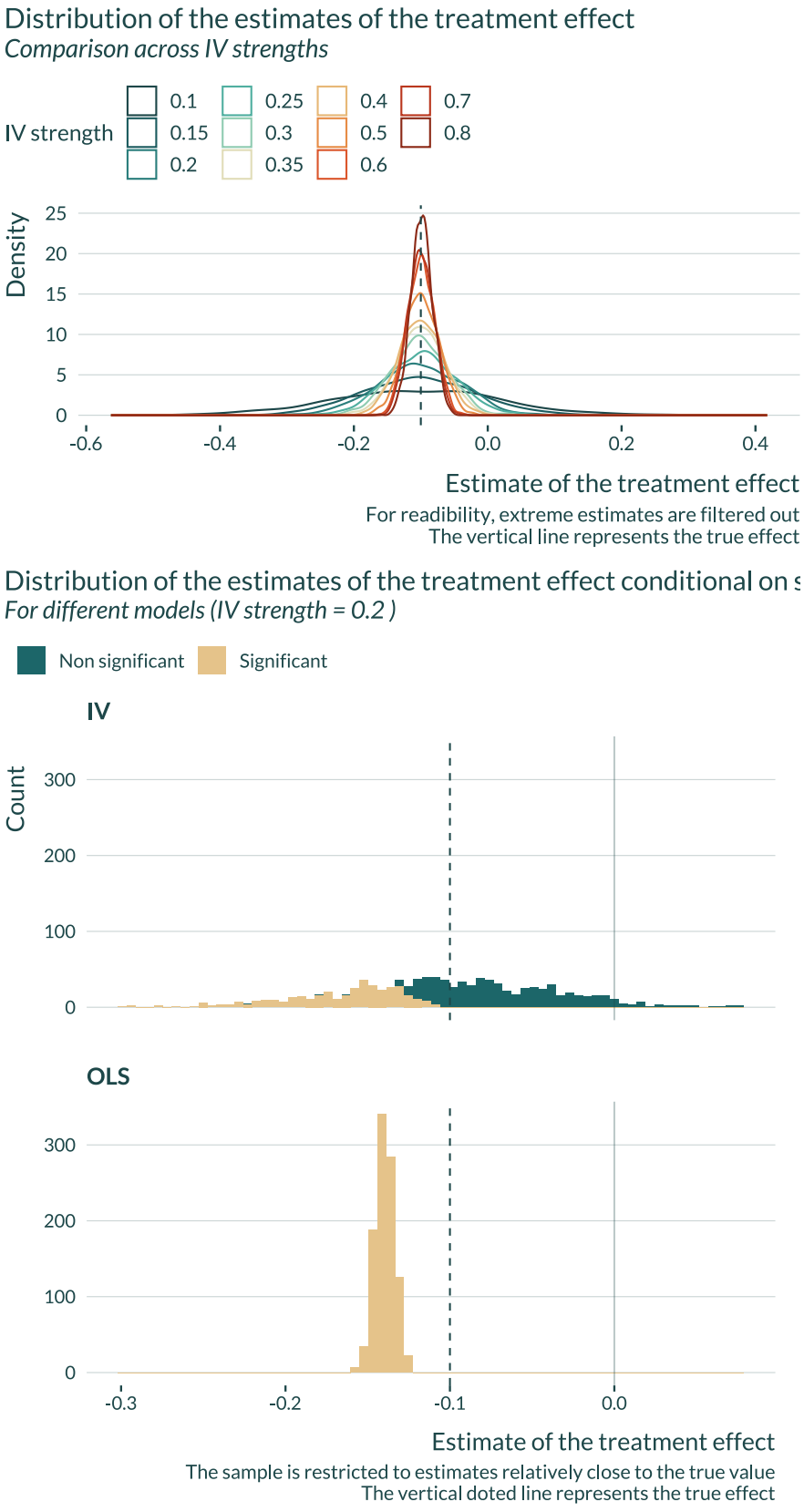
Finally, I run all the simulations by looping the `compute_sim_iv` function on the set of parameters `param_iv`.

B.3.3 Analysis of the results

Quick exploration

First, I quickly explore the results.





We notice that the OLS is always biased and that the IV is never biased. However, for

limited IV strengths, the distribution of the estimates flattens. The smaller the IV strength, the most like it is to get an estimate away from the true value, even though the expected value remains equal to the true effect size.

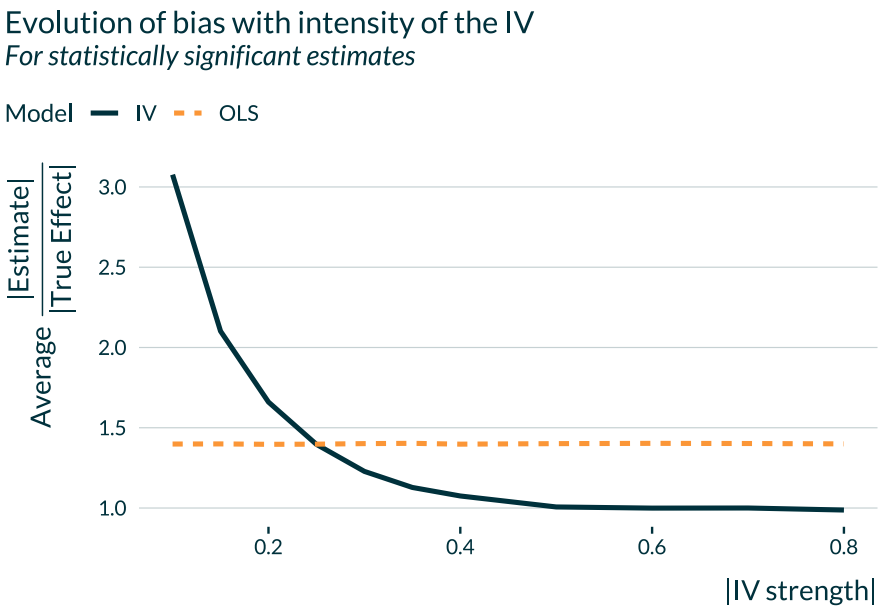
Computing bias and exaggeration ratios

We want to compare $\mathbb{E}\left[\left|\frac{\widehat{\beta_{IV}}}{\beta_1}\right|\right]$ and $\mathbb{E}\left[\left|\frac{\widehat{\beta_{IV}}}{\beta_1}\right| | \text{signif}\right]$. The first term represents the bias and the second term represents the exaggeration ratio.

To do so, I use the function `summmarise_sim` defined in the `functions.R` file.

Graph

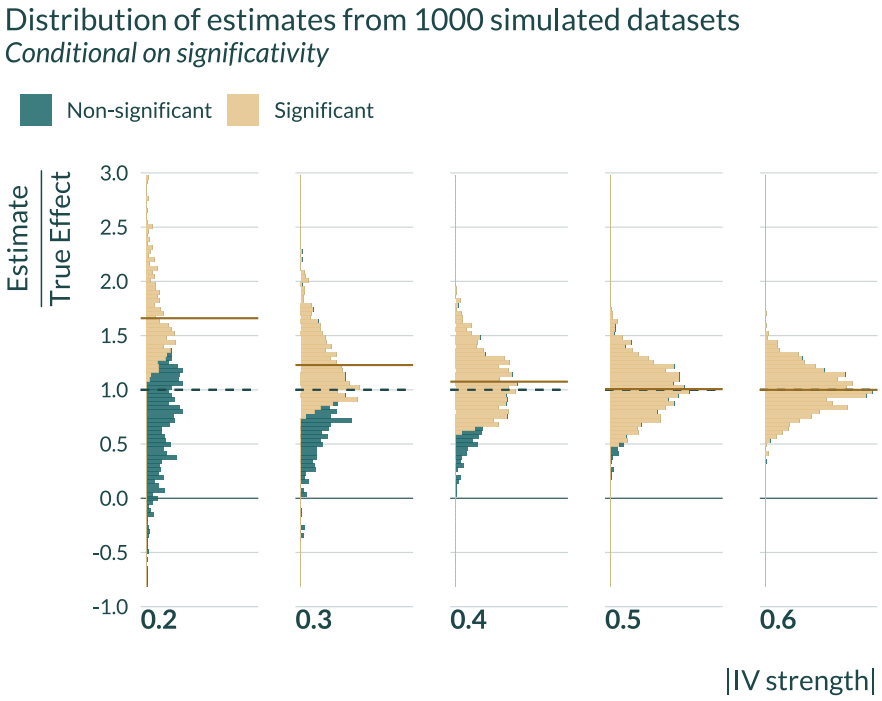
To analyze our results, we build a unique and simple graph:



We notice that, if the IV strength is low, on average, statistically significant estimates overestimate the true effect. If the IV strength is too low, it might even be the case that the benefit of the IV is overturned by the exaggeration issue. The IV yields an unbiased estimate and enables to get rid of the OVB but such statistically significant estimates fall, on average, even further away from the true effect.

Of course, if one considers all estimates, as the IV is unbiased, this issue does not arise.

Distribution of the estimates I then graph the distribution of estimates conditional on significativity. It represents the same phenomenon but with additional information.



B.3.4 Further checks

Representativeness of the estimation

I calibrated the simulations to emulate a typical study from this literature. To further check that the results are realistic, I compare the average Signal-to-Noise Ratio (SNR) of the simulations to the range of SNR of an existing study. The influential study Gomez et al. (2007) have a SNR of about 8. Yet the reanalysis by Cooperman (2017) shows using randomization inference that accounting for spatial correlation in rain patters yield a larger standard error (a point estimate of -1.052 and with a p-value of 0.084).

This would yield a SNR of 1.73. For such an SNR, there is some non-negligible exaggeration in my simulations:

IV strength	Median SNR	Exaggeration Ratio
0.10	0.95	3.08
0.15	1.24	2.10
0.20	1.65	1.66
0.25	1.96	1.40
0.30	2.43	1.23
0.35	2.79	1.13
0.40	3.22	1.08
0.50	3.97	1.01
0.60	4.77	1.00
0.70	5.58	1.00
0.80	6.33	0.99

This result does not mean that Gomez et al. (2007) suffer from exaggeration but rather indicates that my simulations are in line with SNRs that can be observed in actual studies and their calibration is not disconnected from existing studies.