The Journal of Finance and Data Science 10 (2024) 100127

Contents lists available at ScienceDirect

The Journal of Finance and Data Science

journal homepage: www.keaipublishing.com/en/journals/jfds

Short Communication

Overlooked biases from misidentifications of causal structures

Simone Cenci

Leonardo Centre, Imperial College Business School, Exhibition Rd, London, SW7 2AZ, UK

ARTICLE INFO

Keywords:
Empirical finance
Causal inference
Identification

ABSTRACT

Testing theories and explaining phenomena in empirical finance often requires estimating causal effects from observational data. In this note, we argue that some of the standard practices to address endogeneity concerns in regression-based estimation approaches can, when not correctly implemented and their results not appropriately interpreted, generate additional, often overlooked, problems. We identify three main systemic issues in empirical finance, provide theoretical and numerical examples to illustrate and support our arguments, and propose solutions to overcome these limitations. Overall, we suggest that these issues are caused by a systematic underestimation of the importance of robust ex-ante identification, and interpretation, of causal structures in empirical studies in finance.

Two important objectives of empirical studies in finance are to (1) estimate causal effects and (2) compare the relative importance of different factors in explaining variability in the observed outcome variables. Examples include but are not limited to, studies concerned with the estimation of the effect of corporate policies on firms' performance (Vishwanathan et al., 2020), the impact of climate change on asset prices (Campiglio et al., 2022), the effect of financial and nonfinancial disclosure on stakeholders and markets' participants (Christensen et al., 2021), studies interested in explaining cross-sectional variability of financing choices (Boateng et al., 2022) and equity returns (Giglio et al., 2022). The often implicit objective of the vast majority of these studies is to draw causal conclusions.

Attempts to estimate unbiased causal effects from observational data require careful treatments of endogeneity issues, and a plethora of different methodologies exist to achieve this objective (Heckman, 1979; Rubin, 2006; Pearl, 2009; Cunningham, 2021). In empirical finance, the three most important (albeit not sole) sources of endogeneity are: omitted variables, measurement errors and simultaneity (Roberts and Whited, 2013). Here, we discuss a series of overlooked problems arising from the presence, or weak treatment, of omitted variables, i.e., variables excluded from the control set and that, therefore, induce a correlation between explanatory variables and the error term. Omitted variables are a crucial issue of empirical financial studies for two reasons: (1) it is virtually impossible to observe all sources of endogeneity, and (2) no statistical test exists to fully identify the bias.

In this note, we argue that, even when all sources of bias on the effect of interests are accounted for, standard practices in reporting results can lead to erroneous interpretations and can potentially do a disservice to otherwise well-designed studies. We also show that standard practices to control for omitted variables can introduce an unexpected bias and that greater care should be placed on the interpretation of the comparison of regression coefficients. In particular, we identify three main systemic issues in empirical finance studies that we refer to as (1) over-interpretation of control factors, (2) non-omitted variable bias, and (3) inconsistent comparisons. We discuss the origin of these issues, their potential implications, and suggest solutions to address them. Overall, we argue that the main cause of these systemic issues is an often superficial use of regression to estimate causal effects without proper ex-ante identification and ex-post interpretation of causal structures.

E-mail address: s.cenci@imperial.ac.uk.

Peer review under responsibility of KeAi Communications Co., Ltd.

https://doi.org/10.1016/j.jfds.2024.100127
Received 4 September 2023; Received in revised form 20 February 2024; Accepted 21 February 2024
Available online 29 February 2024

Before proceeding, we would like to make two remarks. First, whilst causal effects from observational data can be estimated through several methodologies, here we focus exclusively on the discussion of issues emerging from regression-based approaches, which play a crucial role in empirical finance (Roberts and Whited, 2013). Second, these issues arise from well-known limitations of causal inference from empirical data (Angrist and Pischke, 2009; Pearl, 2009; Pearl et al., 2016; Cunningham, 2021). These issues are well understood in the natural sciences, see for example (Westreich and Greenland, 2013). They have also been recently discussed at length in several works in the social sciences, most notably in organisation and political studies, see for example (Keele et al., 2020; Hünermund and Louw, 2023). Therefore, none of our conclusions and discussions are new, and we refer to the aforementioned works, and their references, for further insights. Yet, the issues we raise are nevertheless overlooked in empirical finance, so it is important to remark on them, connect them, contextualise them, and propose practical solutions to avoid them. Crucially, these solutions invariably require greater attention to ex-ante identification of causal structures. Hence, we first provide a brief discussion of standard approaches to identification and their limitations.

1. A brief overview of the identification problem

The main limitation of causal inference from observational studies in empirical finance is the presence of omitted variables. Whilst the possibility of omitted confounders cannot be conclusively eliminated from any empirical study, a model's design can be justified if the underlying causal structure is identified and the relevant confounders controlled for in estimation approaches. Identification of causal structures, that is, the identification of how variables in a system relate to one another, is, therefore, a crucial step to any robust estimation of causal effects. Without valid identification of causal structures, we cannot ascertain the validity of virtually any empirical study. Yet, identification is also the most complicated task in the causal inference process.

One of the most well-established approaches for identification of causal structures in the natural sciences, but increasingly so in the social sciences as well, is offered by structural causal modeling (SCM) (Pearl, 2009; Pearl et al., 2016). A full presentation of SCM is beyond the scope of this work and can be found in several causal inference textbooks (Cunningham, 2021; Pearl, 2009; Pearl et al., 2016; Peters et al., 2017). Briefly, a SCM is a system of structural equations of the form $X_i = f_i(\overrightarrow{\mathrm{PA}}_i) + U_i$, where $i = 1, \ldots, d$, d is the number of variables in the system, $\overrightarrow{\mathrm{PA}}_i$ are causes (parents) of X_i, and U_i are exogenous jointly independent noise variables. Each SCM can be represented by a graphical causal model, in particular, a Directed Acyclic Graph (DAG), i.e., a directed graph where each node is a variable in the SCM, and the arrows follow the direction of causation implied by it (Pearl et al., 2016).

For example, the graph $X \rightarrow Y \leftarrow Z$ is the DAG associated with a SCM of the following form: $X = U_X$; $Y = f(X, Z) + U_Y$; $Z = U_Z$, where U_j are independent noise variables and f is an arbitrary function.[1] The DAG in the example is often called a *collider*. Each DAG entails a set of conditional independences among variables that can be directly validated on data. For instance, in the collider structure mentioned above, X and Z are unconditionally independent variables, but they are dependent conditioning on the central node Y, i.e., $X \perp\!\!\!\perp Z$ but $X \not\perp\!\!\!\perp Z|Y$. SCMs offer a series of systematic rules to identify the conditional independences entailed by any arbitrary complex DAG and, crucially, the subsequent variables to select or omit to estimate unbiased causal effects.

The presentation and discussion of those rules are beyond the aim of this work and can be found in (Pearl, 2009; Pearl et al., 2016; Peters et al., 2017). What we want to stress here is that the conditional independences entailed by any arbitrary DAG can, in theory, be tested empirically on observational data, with minimal assumptions on the analytical properties of the SCM associated with the DAG, using nonparametric conditional independence tests, such as the one proposed in (Zhang et al., 2011; Strobl et al., 2019). In the Supplementary Code we provide a simple implementation of these tests. However, it is important to bear in mind that with a finite amount of data, non-Gaussian distributions, and large conditioning sets, there are practical limitations to their validity; see (Peters et al., 2017) for a discussion.

The power of SCM approaches is that hypotheses on causal structures can be directly translated into their corresponding DAG, which, in turn, entails a set of conditional independences that, in principle, can be tested on empirical data ex-ante, that is, before estimating causal effects. These tests can provide empirical evidence to support the omission and inclusion of variables in a given specification and attenuate, albeit not eliminate, the risk of omitted variable bias in the final estimates of the effects. Importantly, expressing hypotheses in terms of DAGs also provides researchers and readers with a clear picture of the structure of the data-generating process, including the role of each variable, not just treatment and outcomes, in the model. Crucially, identification of causal structures from SCM always requires expert knowledge to restrict the space of possible models and choose among equivalent DAGs, i.e. causal graphs with the same implied conditional independences but corresponding to different SCMs (Pearl et al., 2016). Identification of causal structures in empirical finance can rarely be a completely automated process.

Conditional independence tests are the standard approaches for causal discovery in the causal inference literature (Peters et al., 2017; Teymur and Filippi, 2020). However, they are rarely used by empirical researchers in finance. Instead, the vast majority of works in empirical finance rely on the accumulation of expert knowledge to select candidate models within a theory-driven framework without ex-ante validation of the hypotheses. More concretely, if answering a research question requires estimating the effect of a variable X on another variable Y, standard approaches use evidence in the literature, from experiments to other empirical studies, to narrow down the possible confounders to a set of variables whose role is well-understood within the latest theoretical constructs, i.e. variables that are supported by well-established theories. Approaches to causal structure identification based *solely* on extant evidence in the literature, and thus theory-driven, are less powerful than those *also* based on ex-ante validations of the hypotheses concerning the data-generating processes. Indeed, the latter provides a transparent, testable and interpretable picture of the underlying causal structure; the former does not.

[1] Following standard notations we have omitted the noise variables in the DAG.

In the following three sections, we illustrate how underestimating the importance of valid ex-ante identification of causal structures and, importantly, ex-post interpretation of the role of each variable in the structures can potentially lead to serious bias in even well-specified studies.

2. Over-interpretation of control factors

The first issue we discuss in this note concerns a systematic underestimation of the importance of causal structure identification as manifested in the frequent over-interpretation of the causal role of control factors. This issue, which also highlights frequently ignored dangers of omitted variables, has also been discussed at length in (Westreich and Greenland, 2013; Keele et al., 2020) and, more recently, in (Hünermund and Louw, 2023). Here, we stress it further to underscore its importance.

Following standard practices in empirical finance, causal identification is often based on extant evidence in the literature. Outcomes of regressions specifications and robustness tests are then often reported and, most importantly, interpreted based on the sign and significance of the effect relevant to the study, and *also* based on the sign and significance of the other regression parameters without discussions of their role in the underlying causal structures. This practice can be recognised in the tendency to report full regression tables without differentiating the differential causal role of the variables in their interpretation. Yet, what is often not appreciated is that even if, hypothetically, all the relevant confounders are accounted for, and, therefore, the effect of interest has an unbiased causal interpretation, the regression parameters of the confounders can still be severely biased by variables that were correctly omitted in the specification. These coefficients, therefore, should not be assigned an economic interpretation. In other words, without proper consideration of the causal structure of the underlying data-generating process in the ex-ante model identification or ex-post model interpretation, omitted variables can still affect the conclusion of a study even when all the relevant variables were accounted for in the estimation of the effect of interest.

To clarify this point, we run the following numerical example. We generate 5000 realisations of a simple linear SCM associated with the DAG shown in Fig. 1 panel **A**. Specifically:

$$Y = \alpha X + \gamma Z + \psi U + \epsilon_Y; X = \kappa Z + \epsilon_X; Z = \varphi U + \epsilon_Z; U = \epsilon_U \tag{1}$$

where $\alpha = 1, \kappa = 1, \psi = -1, \gamma \sim \mathcal{U}(0,1), \varphi \sim \mathcal{U}(0,.1)$ and $\epsilon_i = \mathcal{N}(0,1)$. We sample the parameters from uniform distributions to retain full control over the boundary of their support and ensure that the parameters are always positive. Our task is to estimate the coefficient

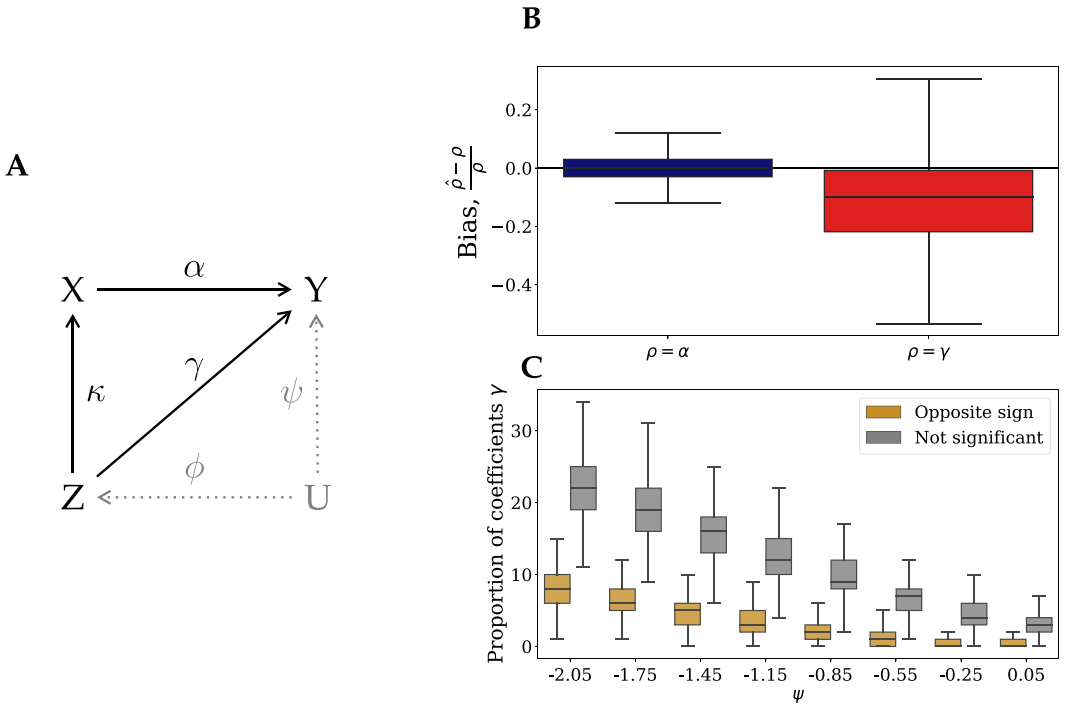

Fig. 1. Over-interpretation of control factors. Panel **A** shows the DAG associated with Eq. (1). Gray nodes denote unobserved variables. Panel **B** shows the bias of the coefficients α (blue) and γ (red) as estimated from Eq. (2). Panel **C** shows the proportion of estimated coefficients $\hat{\gamma}$ that are not statistically significant (grey) and that have the opposite sign to the true γ (yellow), as function of the parameter ψ which regulates the importance of the confounding factor in the model. The proportions are calculated by repeating the simulation of Eq. (1) and the estimation of its parameters with Eq. (2) 100 times. Then, we repeat the process 500 times for each value of ψ. The lines of the box plots are median lines, and the edges of the boxes are the quartile range.

α in the model. To remove sources of bias we control for Z,[2] which is the necessary and sufficient set of variables we need to control for to eliminate the bias in α, namely:

$$y = \alpha x + \gamma z + \epsilon \tag{2}$$

Following standard approaches (Pearl et al., 2016) hereafter we use lower case letters for variables in regressions equations and upper case letters for variables in the structural models.[3] The specification in Eq. (2) does not involve any omitted variable for the estimation of α. Therefore, Fig. 1 panel **B** shows that, as expected, after controlling for z the bias of α is zero. However, the figure also shows that the bias of γ is different from zero. In our example, the bias can be so sever that in approximatively 11% of the realisations the estimated coefficient $\widehat{\gamma}$ was not statistically significant. In 3% of the realisation $\widehat{\gamma}$ and γ even had opposite sign. More generally, Fig. 1 panel **C** shows the proportion of estimated coefficients with opposite sign (dark yellow) or that are not statistically significant (gray) as function of the unobserved effect ψ.

The bias shown in the figure is due to the presence of a variable U, whose omission does not induce a bias in the coefficient of interest (α) and that was therefore correctly omitted from the regression in Eq. (2). Put differently, the figure shows that the coefficient of the necessary and sufficient factors to control for to estimate unbiased causal effects (the coefficient of the variable z in this particular example) can be severely biased even when the regression is correctly specified and the effect of interest is unbiased. From a theoretical standpoint, this result is not surprising. What is surprising is that these issues are rarely taken into consideration in empirical studies in finance, and economic interpretations are often assigned to the whole suite of regression coefficients.

The example in Fig. 1 is an over-simplified one. In typical empirical settings there are multiple confounders to control for. Some of these confounders can be confounded by omitted variables, while others might not. These biases would make the interpretation of their relative importance misleading. Indeed, the interpretation of the relative importance different factors within a regression specification is a broader issue which we will discuss later in this note (see section Inconsistent comparisons).

Estimating unbiased causal effects of a treatment or exposure on an outcome requires controlling for confounding factors, but, even when all the relevant factors are accounted for, the associations of these factors with the outcome variable can be biased. This is not an issue *per se*, because the bias does not influence the conclusion drawn from the sign and significance of the treatment effect. Issues can arise when an economic interpretation is assigned to the coefficients on the control set. Therefore, in line with the suggestions of (Keele et al., 2020; Hünermund and Louw, 2023; Westreich and Greenland, 2013), we recommend researchers to refrain from interpreting the full suite of coefficients from regression specifications, unless the full structure of the model is correctly identified ex-ante.

3. Non-omitted variable bias

A well-established methodology to mitigate omitted variable bias is to control for all the possible factors related to the cause and effect variables. However, in empirical finance, and the social sciences in general, controlling for all possible confounders of the effect of interest is virtually impossible since many factors are unobservable. Here, we note that, in the presence of unobservables, adding covariates in regression without accurately identifying their role within the data-generating process could induce a less appreciated, but equally important, *non-omitted* variable bias. This bias is often, albeit not solely, due to the well-known *collider bias* (Pearl et al., 2016; Cunningham, 2021) induced by the inclusion of the central node of a collider structure in the control set (see section A brief overview of the identification problem for a definition of a collider structure). Empirically, collider biases are important as they can lead to substantial divergences in parameter estimates across studies. See (Cenci and Kealhofer, 2022) for a discussion of this form of bias in a series of studies in the capital structure literature.

Fig. 2 provides an example to illustrate how a collider bias can arise in a simple causal structure with two unobservables, and the extent to which the bias can influence the result of a study. The example reproduces the well-known M-bias (Pearl, 2009), which has been a matter of controversy and discussions in the causal inference literature for more than a decade (Ding and Miratrix, 2015). The DAG in Fig. 2 panel **A** is the same as the DAG in Fig. 1 panel **A** with the addition of a variable C which is driven by two unobservables (I and L), one of which (L) in turn drives Y. The path $\mathrm{L} \rightarrow C \leftarrow I$ forms a collider in the DAG where C is a central node. We simulate 5000 realisations of the (linear) SCM associated with the DAG. Specifically:

$$\begin{aligned} Y &= \alpha X + \gamma Z + \psi U + \eta L + \epsilon_Y; X = \kappa Z + \beta I + \epsilon_X; Z = \gamma U + \epsilon_Z; C = \omega I + \delta L + \epsilon_C; \\ U &= \epsilon_U L = \epsilon_L; I = \epsilon_I \end{aligned} \tag{3}$$

where, for simplicity, each coefficient in the structural model is set to one, and $\epsilon_i = \mathcal{N}(0,1)$. Notice that C is not driven by X, in this sense it is a pre-treatment variable. Then, we estimate α without and with realisations of C in the control set. That is we estimate the coefficients of the following regression specifications:

$$y = \alpha x + \gamma z + \epsilon \tag{4}$$

$$y = \alpha x + \gamma z + \nu c + \epsilon \tag{5}$$

[2] For simplicity, we consider Z to be univariate. To extend the example to a multivariate control set, $\vec{Z}$ it is enough to include each variable in $\vec{Z}$ in the control set.

[3] For simplicity we use the same notation for the coefficients.

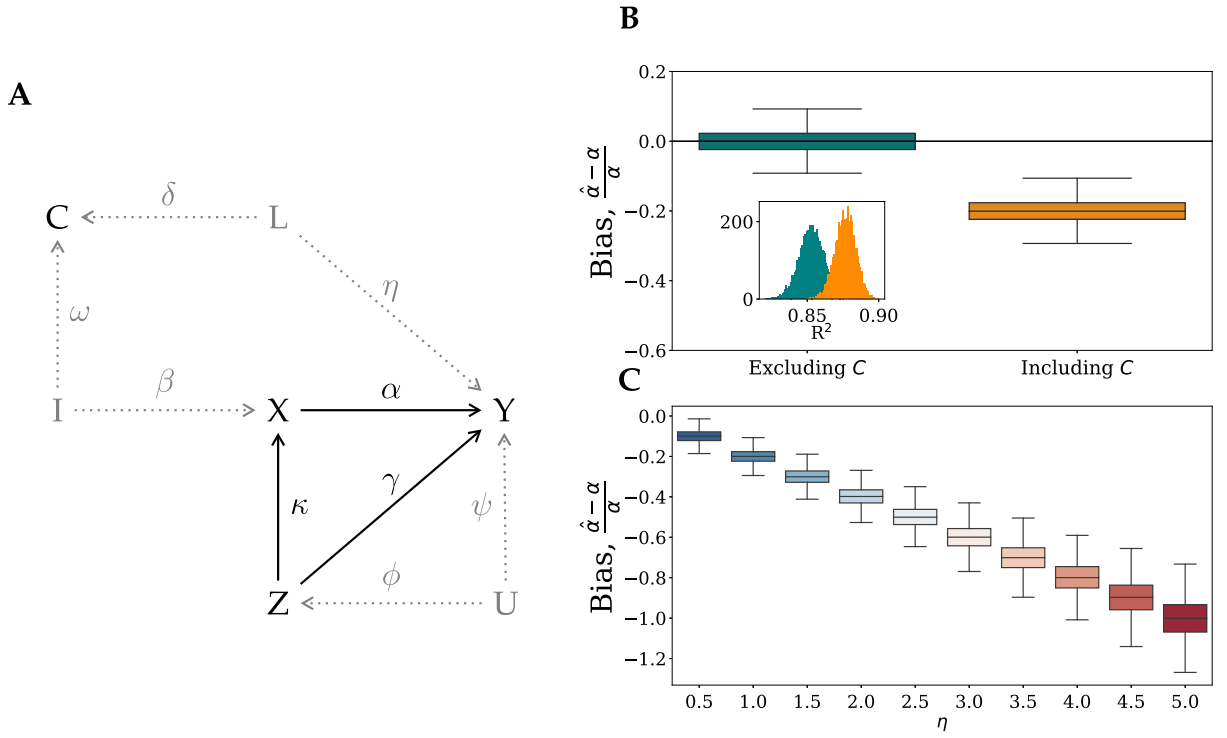

Fig. 2. Non-omitted variable bias in the presence of unobservables. Panel **A** shows an augmented version of the DAG shown in Fig. 1. Gray nodes denote unobserved variables. In this model we include a collider path $L \rightarrow C \leftarrow I$ which includes an unobservable driver of Y, L. The structural equation of the graph is shown in Eq. (3). Panel **B** shows the bias in the coefficient α as estimated from Eq. (4) (teal) and Eq. (5) (orange). The inset shows the R^2 of the two regressions. Panel **C** shows the bias of α as estimated from Eq. (5) as function of the parameter η, which controls the importance of the collider structure in the model. The lines of the box plots are median lines, and the edges of the boxes are the quartile range.

Fig. 2 panel **B**, clearly shows that, as expected, controlling for c induces a bias on α (orange boxplot). The bias would disappear if we could control for l, the realisation of L, since L blocks the path from C to Y in the DAG, but here L is assumed to be unobservable. Clearly the bias could be alleviated by controlling for a proxy for L (which is the standard approach to control for unobservables), but it could be eliminated by not controlling for realisations of C in the first place, i.e., it could be eliminated by *omitting* c from the regressions.

Notice that C is not a "bad control" in the standard econometric interpretation of the term (Angrist and Pischke, 2009), since, as stressed above, it is still a pre-treatment variable fixed at the time at which X was determined (Cinelli et al., 2022). Similarly, C is not a "neutral variable" such as, for example, U after Z is held constant, i.e. including or excluding C does change the value of the parameter α (Cinelli et al., 2022). provides a comprehensive discussion of the identification of "good", "neutral", and "bad" controls using DAGs.

Coming back to the numerical example, it is important to stress that, as shown in the inset of Panel **B** in Fig. 2, the biased model, Eq. (5), has a higher explanatory power (R^2) than the unbiased model, Eq. (4). Therefore, standard approaches for model selection based on quality of fit cannot be used to identify this source of bias and the correct explanatory model for Y. Only a correct ex-ante identification of the causal structure can. Fig. 2 Panel **C** shows the severity of the bias as a function of the dependence of Y on L in the structural model. The figure shows that, as the importance of L increases, so does the bias induced by the wrong control.

In standard econometrics it is often assumed that adding additional *pre-treatment* covariates in a regression helps eliminating the risk of omitted variable bias, and it increases the likelihood that the estimated effect has a causal interpretation. However, in the presence of unobservables, adding covariate without a clear understanding of their role in the model - without a clear ex-ante identification strategy - also increases the probability of inducing additional biases. In line with the recommendation of (Hünermund and Louw, 2023; Keele et al., 2020), a variable in a regression should only be included if there is strong evidence that it would otherwise induce a bias in the effect of interest. If enough evidence cannot be collected, or the structure of the model cannot be identified ex-ante, qualitative reasoning should not be used, or at least should be used with caution, to justify the inclusion of a covariate in the model.

In this section we focused exclusively to the bias induced by over-controlling for pre-treatment variables in the presence of unobservables and collider bias in the underlying data generating process. In the next section we show that over-controlling can also be problematic in well-designed and correctly identified models if the causal structure of the data-generating process is not correctly identified and interpreted.

4. Inconsistent comparisons

Oftentimes, empirical researchers in finance are interested in comparing the relative effects of different variables or treatments on a specific outcome. As an illustrative example, a large body of empirical corporate finance literature is concerned with identifying the relative importance of different determinants of leverage.[4] These estimations are often performed by regressing leverage values on a series of asset characteristics (e.g., Size) and financial ratios (e.g., Profitability) and comparing the sign and magnitude of these factors. These values are also often compared to those expected by competing theories (e.g., pecking order, trade-off), and the empirical evidence is then used to support the one with greater agreement with the data.

Are these comparisons meaningful? Answering this question through a structural causal modeling lens immediately shows that this is not the case. Comparing the relative importance of different factors within a regression can lead to ambiguous conclusions because the role of the factors within the data-generating process can be significantly different, and the interpretation of these differences is often omitted in the discussion of the comparison.

To illustrate our argument, Fig. 3 panel **A** shows a toy model for the determinant of leverage. We assume that leverage ratios L, depend on three factors: P, S (e.g., Profits and Size), and N, where N is unobserved (but in principle observable). Our objective is to compare the relative importance of S and P in explaining variability in L. Standard approaches suggest including both factors in a regression model. If we regress L on both P and S to compare their relative importance, even if N is unobserved, none of the coefficients is biased by omitted variables. Yet, the coefficient on P measures a direct effect, while the coefficient on S measures a partial effect since part of the effect is blocked by P. The total effect of S on L is biased due to over-controlling for P. If the total effect of S on L is larger than the effect of P on L, but most of the effect of S on L flows through P, then the regression would erroneously lead us to conclude that P is significantly more relevant than S in explaining variability in L. The comparison between the two coefficients is inconsistent, or, better, ill-posed, because the roles of the two variables within the model are fundamentally different.

To illustrate the practical implications of this misleading comparison we run the following analysis. First, we simulate a linear model associated with the DAG in Fig. 3 panel **A**. Specifically,

$$L = \alpha P + \beta N + \epsilon_L; P = \gamma S + \epsilon_P; N = \delta S + \epsilon_N; S = \epsilon_S \tag{6}$$

where $(\alpha,\beta) \sim \mathcal{U}(0,0.5)$, $(\gamma,\delta) \sim \mathcal{U}(0,1.25)$, and $\epsilon_i = \mathcal{N}(0,1)$. Similarly to Eq. (1), we sample the parameter from uniform distributions to ensure that the parameters are always non-negative. Second, we regress L on P and S, and separately, we regress L on S alone, namely,

$$l = \alpha p + \psi s + \epsilon \tag{7}$$

$$l = \sigma s + \epsilon \tag{8}$$

The coefficients from the two regressions are shown in Fig. 3 panel **B**. The panel shows that the effect of P on $L(\alpha)$ is larger than the partial effect of $S(\psi)$ but smaller than the total effect of S[5] (σ). There are no omitted variables in the model and, from a statistical standpoint, there is no bias in the estimation of the coefficients. However, the direct comparison of the coefficients in Eq. (7) is biased if the nature of the coefficients themselves is not accounted for in the interpretation of their differences. In this particular example, the bias only changes the magnitude of the coefficients. However, it is easy to see that if we were also controlling for N in Eq. (7), then the partial effect of S on L would not be statistically significant, making the interpretation of the relative contribution of the two variables (P and S) even more misleading.

The importance of this issue can be understood within the context of interventions. Typically, a researcher is interested in comparing the relative causal effect of a set of variables (P, S) to a given outcome (L) to understand the impact that changes (or interventions) on the former can have on the latter. For this interpretation to be meaningful the comparison between interventions needs to account for the role of the variables in the model, i.e., the nature of the effects. In our example changing (or intervening on) S has a greater impact on L than changing (or intervening on) P. Yet, when partial effects are compared with total effects (i.e., when effects of different nature are compared to one another) these differences are lost.

This issue can be resolved by ensuring that comparisons among variables are interpreted only when the nature of the estimated coefficients is clarified. That is, by running independent regression specifications after careful identification of causal structures, and by only reporting the coefficient of the effect of interest in each of the specification. For example, the problem in the example in Fig. 3 can be resolved by regressing L on P and S (since S is in the backdoor path of P), and L on S independently and only then compare the value of the two coefficients, α and σ.

Generally, running independent regressions provides greater confidence in the treatment or covariate comparison, but there are several circumstances where sets of variables within a regression have similar roles, and variables within the sets can be reasonably compared to one another. For example, if a study requires comparing the effect of different sustainability policies on a given outcome

[4] Here we use the case of Leverage as an illustrative example, but the issue arises in several studies that look for instance, at the determinants of credit spread, stock returns or corporate emissions.

[5] The very same results could have been obtained simply through path analysis without the aid of regression models. Indeed, the partial effect of S on L is simply $\delta\beta$ and the total effect is $(\delta\beta + \gamma\alpha)$. Recalling that $(\alpha,\beta) \sim \mathcal{U}(0,0.5)$, $(\gamma,\delta) \sim \mathcal{U}(0,1.25)$, then $\delta\beta = \gamma\alpha = 0.156$. Therefore the total effect of P on $L(\alpha)$ is 0.25, the partial effect of S on L after controlling for P $(\delta\beta)$ is 0.156 and the total effect $(\delta\beta + \gamma\alpha)$ is 0.312, which are the values shown in Fig. 3.

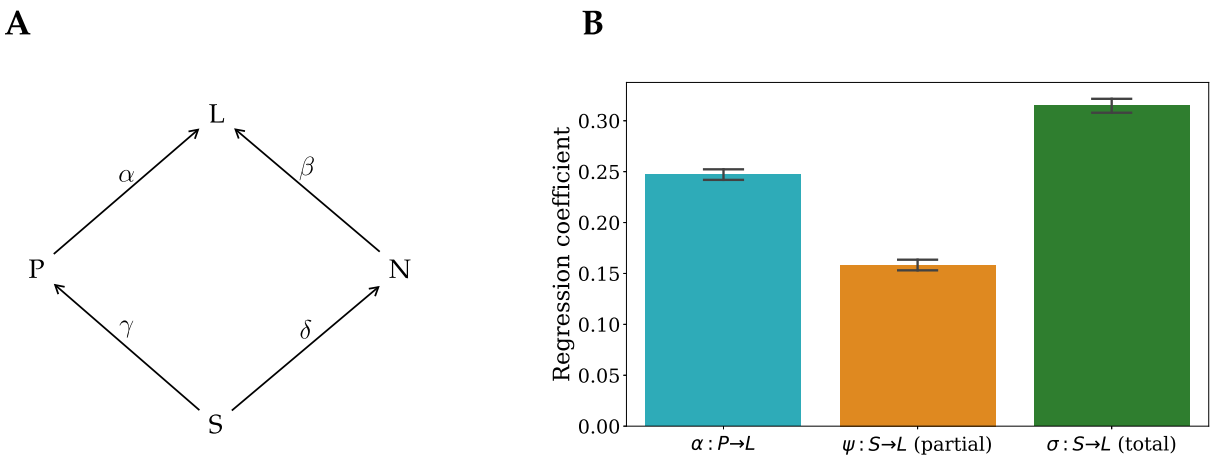

Fig. 3. Inconsistent comparisons. Panel **A** shows the DAG associated with the SCM in Eq. (6), i.e., the toy model for the determinant of leverage ratios (L). Panel **B** shows the regression coefficients estimated from Eq. (7) (blue and orange) and Eq. (8) (green). Error bars are 99% confidence intervals.

(say firms' carbon emissions), conditioned on assets' characteristics (e.g., firms' size, total invested capital), it is reasonable to assume that the paths from the policies to the outcome play approximatively the same role. Similarly, in statistical factor models, orthogonalisation removes the mutual dependency of factors. Therefore, estimating coefficients from independent regression models, or from a single regression model which includes all the orthogonal factors, leads to the same estimates.

Comparing the coefficients of a regression specification to one another can lead to erroneous conclusions due to the different nature of the associations within a given causal structure. Yet, considerations on the role of the coefficients in the underlying data-generating process are often omitted in the interpretations of the regression coefficients. In line with the recommendation of (Keele et al., 2020), to fairly compare different factors, researchers need to run independent regression specifications after correctly identifying the causal structures of the data generating process. Alternatively, researchers can explicitly account for the different roles of the variables in the underlying causal structure, by including these considerations in the interpretations of the coefficients as measuring partial or total effects.

5. Conclusion

In this note, we identified three systemic issues in empirical finance studies that rely on regression approaches to estimate causal effects from observational data. We discussed the practical implications of these issues and suggested strategies to address them. The proposed strategies, that closely align with recommendations in cognate fields (Keele et al., 2020; Hünermund and Louw, 2023), invariably require greater attention to the ex-ante identification of causal structures and the ex-post interpretation of the role of variables in the structure.

Identification is at the core of several approaches to causal inference, most notably structural causal modeling (Pearl, 2009; Pearl et al., 2016). In structural causal modeling, estimation processes of causal effects from observational data start with a mental model, informed by extant research, of how different variables relate to one another. Crucially, these models must then be translated into graphical causal models, or more specifically Directed Acyclic Graphs, which, in turn, entail testable conditional independence conditions (Pearl, 2009; Gow et al., 2016; Pearl et al., 2016; Peters et al., 2017; Deaton, 2020). The conditional independence conditions can be tested empirically and non-parametrically using a variety of different approaches, such as the ones proposed in (Zhang et al., 2011; Strobl et al., 2019). Conditional independence tests in conjunction with expert knowledge can provide the necessary evidence to support the validity of a specific causal structure. Moreover, we believe that the very process of writing down explicit causal graphs for the system under investigation has two additional advantages: (1) it forces researchers to think carefully about how the variables relate to one another and their role within the model, and (2) it provides readers with a clear understanding of the empirical tests used to support the findings.

The choice of the appropriate control factors and the subsequent estimation of the magnitude and sign of causal effects should ideally be performed once evidence from conditional independence tests provides some support to the proposed causal structure (Pearl, 2009; Pearl et al., 2016; Gow et al., 2016; Peters et al., 2017; Keele et al., 2020; Cenci and Kealhofer, 2022; Hünermund and Louw, 2023). The interpretations of each estimated effect must then account for the role that the target variables play in the structure. Underestimating the importance of robust ex-ante identification of causal structures can give rise to several overlooked biases in the final estimates, or their interpretation, and potentially undermine the validity of even well-designed empirical studies. We hope that the discussions in this note will encourage the use of identification approaches based on causal graphs and conditional independence tests in empirical finance.

Code availability

The code to generate the figures is available on Harvard Dataverse at https://doi.org/10.7910/DVN/JMQHJN. The code also includes a few examples of conditional independence tests and a function that can be used for identification of structural causal models.

Declaration of competing interest

The author declares no competing interests.

Acknowledgements

The author would like to thank Matteo Burato for feedback on early versions of the manuscript and two anonymous reviewers for their helpful comments. The APC was paid by Imperial College London Open Access Fund.

References

Angrist, J.D., Pischke, J.-S., 2009. Mostly Harmless Econometrics: an Empiricist's Companion. Princeton University Press, Princeton, NJ.
Boateng, P.Y., Ahamed, B.I., Soku, M.G., Addo, S.O., Tetteh, L.A., 2022. Influencing factors that determine capital structure decisions: a review from the past to present. Cog. Busin. Manag. 9 (1), 2152647.
Campiglio, E., Daumas, L., Monnin, P., Von Jagow, A., 2022. Climate-related risks in financial assets. J. Econ. Surv. 37, 950–992.
Cenci, S., Kealhofer, S., 2022. A causal approach to test empirical capital structure regularities. J. Finan. Data Sci. 8, 214–232.
Christensen, H.B., Hail, L., Leuz, C., 2021. Mandatory CSR and sustainability reporting: economic analysis and literature review. Rev. Account. Stud. 26 (3), 1176–1248.
Cinelli, C., Forney, A., Pearl, J., 2022. A Crash Course in Good and Bad Controls. Sociological Methods & Research, 00491241221099552.
Cunningham, S., 2021. Causal Inference: the Mixtape. Yale University Press.
Deaton, A., 2020. Introduction: randomization in the tropics revisited, a theme and eleven variations. In: Randomized Control Trials in the Field of Development: A Critical Perspective. Oxford University Press.
Ding, P., Miratrix, L.W., 2015. To adjust or not to adjust? Sensitivity analysis of M-bias and butterfly-bias. J. Causal Inference 3, 41–57.
Giglio, S., Kelly, B., Xiu, D., 2022. Factor models, machine learning, and asset pricing. Ann. Rev. Finan. Econ. 14 (1), 337–368.
Gow, I.D., Larcker, D.F., Reiss, P.C., 2016. Causal inference in accounting research. J. Account. Res. 54 (2), 477–523.
Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica 47 (1), 153–161.
Hünermund, P., Louw, B., 2023. On the nuisance of control variables in causal regression analysis. Organ. Res. Methods, 10944281231219274.
Keele, L., Stevenson, R.T., Elwert, F., 2020. The causal interpretation of estimated associations in regression models. Polit. Sci. Res. Methods 8 (1), 1–13.
Pearl, J., 2009. Causality: Models, Reasoning and Inference, second ed. Cambridge University Press, Cambridge.
Pearl, J., Glymour, M., Jewell, N., 2016. Causal Inference in Statistics: A Primer. Wiley.
Peters, J., Janzing, D., Schlkopf, B., 2017. Elements of Causal Inference: Foundations and Learning Algorithms. The MIT Press.
Roberts, M.R., Whited, T.M., 2013. Chapter 7 - endogeneity in empirical corporate finance. In: Constantinides, G.M., Harris, M., Stulz, R.M. (Eds.), Handbook of the Economics of Finance, vol. 2. Elsevier, pp. 493–572.
Rubin, D.B., 2006. Matched Sampling for Causal Effects. Cambridge University Press, Cambridge.
Strobl, E.V., Zhang, K., Visweswaran, S., 2019. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. J. Causal Inference 7 (1), 20180017.
Teymur, O., Filippi, S., 2020. A Bayesian nonparametric test for conditional independence. Found. Data Science 2 (2), 155–172 (EP –).
Vishwanathan, P., van Oosterhout, H., Heugens, P.P.M.A.R., Duran, P., van Essen, M., 2020. Strategic CSR: a concept building meta-analysis. J. Manag. Stud. 57 (2), 314–350.
Westreich, D., Greenland, S., 2013. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. Am. J. Epidemiol. 177 (4), 292–298.
Zhang, K., Peters, J., Janzing, D., Schölkopf, B., 2011. Kernel-based conditional independence test and application in causal discovery. In: Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, UAI'11. AUAI Press, Arlington, Virginia, USA, pp. 804–813.