Where Are the Factors in Factor Investing?

Marcos López de Prado

Marcos López de Prado is global head of Quantitative Research & Development at Abu Dhabi Investment Authority (ADIA) in Abu Dhabi, United Arab Emirates; professor of practice in the School of Engineering at Cornell University in Ithaca, NY; and professor of practice in the Department of Mathematics at Khalifa University in Abu Dhabi, United Arab Emirates.
marcos.lopezdeprado@adia.ae

KEY FINDINGS

- The factor investing literature is logically incoherent: Factor models have a causal interpretation, and yet factor researchers never disclose the causal structure responsible for the observed associations.
- As a consequence of this logical incoherence, factor models are likely misspecified and the reported anomalies are spurious.
- To address these pitfalls, factor researchers must embrace the modern tools of causal inference.

ABSTRACT

In this article, the author advocates for the use of causal graphs to modernize the field of factor investing and set it on a logically coherent foundation. To do this, first he introduces the concepts of association and causation. Second, he explains the use of causal graphs and the real (causal) meaning of the *ceteris paribus* assumption that is so popular among economists. Third, he explains how causal graphs help us estimate causal effects in observational (nonexperimental) studies. Fourth, he illustrates all of the earlier concepts with Monte Carlo experiments. He concludes that the field of factor investing must embrace causal graphs in order to wake up from its associational slumber.

Ask your favorite factor researcher: What is the meaning of the estimated β in a factor model $Y = \alpha + \beta X + \varepsilon$? You are almost guaranteed to get the same answer enshrined in the most popular econometrics' textbooks: β is the slope of a regression line, and it therefore has a distributional (associational) interpretation, not a causal one (see, for example, Ruud 2000; Kennedy 2008; Woolbridge 2009; Hill, Griffith, and Lims 2011; and Greene 2012).[1]

However, if it is true that β is only an associational attribute of the data, why do researchers choose a particular specification $Y = \alpha + \beta X + \varepsilon$ over $X = \gamma + \delta Y + \zeta$, where Y represents excess returns, and X represents the factors? Associational properties are nondirectional ("X is associated with Y" implies that "Y is associated with X"), thus an associational interpretation of β implies that one model is recoverable from the other: $\hat{\gamma} = -\hat{\alpha}/\hat{\beta}$, $\hat{\delta} = 1/\hat{\beta}$, and $\hat{\zeta} = -\hat{\varepsilon}/\hat{\beta}$. The problem is, this is virtually never the case in the factor investing literature because researchers estimate β with the least-squares method. Under least squares, the estimate of β cannot be recovered from the estimate of δ. By estimating $Y = \alpha + \beta X + \varepsilon$ with least squares, the researcher asserts that the direction of dependence is from X to Y and not the other way around.

[1] For a critical analysis of causality in econometrics textbooks, see Pearl and Chen (2013).

Consequently, the correct interpretation is that β is the linear effect that X has on Y, a causal (nonassociational) concept.

Why do factor researchers use the least-squares method? One reason is, least-squares estimates are unbiased, subject to the model's specification being correct. Unbiasedness is a crucial property for factors researchers: Should the bias be so large that β is estimated with the wrong sign ($\hat{\beta}\beta < 0$), the portfolio would be exposed to a risk with negative expected return. If factor researchers were merely interested in associations, they would minimize the cross-validated errors by exploiting the bias-variance trade-off, like machine learners do. So not only do researchers believe that Y is a function of X (a causal concept), but they are also willing to sacrifice as much predictive power (an associational concept) as necessary to remove all bias from $\hat{\beta}$. The catch is, to monetize the risk premium causally attributed to X, the model must be correctly specified. At the very least, the errors must be exogenous causes of Y, uncorrelated to X (the explicit exogeneity assumption). Assuming exogeneity requires knowledge of the role that each variable plays in the causal graph. Specification setting is the perilous maneuver in which the factor investing ship runs aground, because factor investing researchers justify their specification choice through associational arguments, such as explanatory power, rather than through causal arguments.

There is a logical inconsistency at the heart of the factor investing literature: On one hand, researchers attempt to compute unbiased $\hat{\beta}$ and p-values in a way that is consistent with a causal interpretation of the factors (López de Prado 2023, section 6.1). On the other hand, researchers almost never state a causal graph or falsifiable causal mechanism under which the specification is correct, and the estimates are unbiased (López de Prado 2023, section 6.3). The result of this causal confusion is an academic literature where factors are not really factors (in the causal sense)[2] and where unfalsifiable spurious claims proliferate. Under these circumstances, investors may feel compelled to ask: Where are the factors in factor investing?

In this article, I advocate for the use of causal graphs to modernize the field of factor investing and set it on a logically coherent foundation. In order to do that, first I must introduce the concepts of association and causation. Second, I explain the use of causal graphs and the real (causal) meaning of the *ceteris paribus* assumption that is so popular among economists. Third, I explain how causal graphs help estimate causal effects in observational (nonexperimental) studies. Fourth, I illustrate all of the earlier concepts with Monte Carlo experiments. Fifth, I conclude that the field of factor investing must embrace causal graphs in order to wake up from its associational slumber.

ASSOCIATION

Every student of statistics, and by extension econometrics, learns that association does not imply causation. This statement, although superficially true, does not explain why association exists and its relation to causation. Two discrete random variables X and Y are statistically independent if and only if $P[X = x, Y = y] = P[X = x]P[Y = y]$, $\forall x,y$, where $P[\cdot]$ is the probability of the event described inside the squared brackets. Conversely, two discrete random variables X and Y are said to be statistically associated (or codependent) when, for some (x,y), they satisfy that $P[X = x, Y = y] \neq P[X = x]P[Y = y]$. The conditional probability expression $P[Y = y \mid X = x] = P[X = x, Y = y]/P[X = x]$ represents the probability that $Y = y$ among the subset of the population

[2] The term *factor investing* is another misnomer. The word *factor* has its origin in the Latin language, with the literal meaning of "doer" or "maker." Semantically, a factor is a cause responsible, in total or in part, for an effect. Ironically, the factor investing literature has not attempted to explain what does or makes the observed cross-section of expected returns.

where $X = x$. When two discrete variables are associated, observing the value of one conveys information about the value of the other: $P[Y = y \mid X = x] \neq P[Y = y]$ or, equivalently, $P[X = x \mid Y = y] \neq P[X = x]$. For example, monthly drownings (Y) and ice cream sales (X) are strongly associated because the probability that y people drown in a month conditional on observing x ice cream sales on that same month does not equal the unconditional probability of y drownings in a month for some (x,y). However, the expression $P[Y = y \mid X = x] \neq P[Y = y]$ does not tell us whether ice cream sales causes drownings. Answering that question requires the introduction of a more nuanced concept than conditional probability: an intervention.

CAUSATION

A data-generating process is a physical process responsible for generating the observed data, where the process is characterized by a system of structural equations. Within that system, a variable X is said to cause a variable Y when Y is a function of X. The structural equation by which X causes Y is called a causal mechanism. Unfortunately, the data-generating process responsible for observations is rarely known. Instead, researchers must rely on probabilities, estimated on a sample of observations, to deduce the causal structure of a system. Probabilistically, a variable X is said to cause a variable Y when setting the value of X to x increases the likelihood that Y will take the value y. Econometrics lacks the language to represent interventions, that is, setting the value of X (Chen and Pearl 2013). To avoid confusion between conditioning by $X = x$ and setting the value of $X = x$, Pearl (1995) introduced the do-operator, $do[X = x]$, which denotes the intervention that sets the value of X to x. With this new notation, causation can be formally defined as follows: $X = x$ causes $Y = y$ if and only if $P[Y = y \mid do[X = x]] > P[Y = y]$.[3] For example, setting ice cream sales to x will not make y drownings more likely than its unconditional probability for any pair (x,y); hence, ice cream sales are not a cause of drownings. In contrast, smoking tobacco is a cause of lung cancer because the probability that y individuals develop lung cancer among a collective where the level of tobacco smoking is set to x is greater than the unconditional probability of y individuals developing lung cancer, for some pair (x,y).[4]

CAUSAL GRAPHS

Variables X and Y may be part of a more complex system, involving additional variables. The causal structure of a system can be represented through a directed acyclic graph, also denoted a causal graph.[5] Although a causal graph does not fully

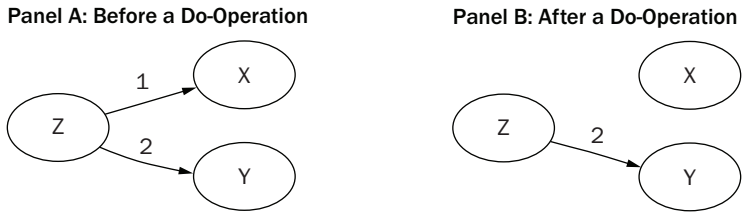
[3] At first, it may seem counterintuitive that causality is defined in terms of a strict inequality (>), in contrast to the difference (≠) used to define association. The reason is there is no need to consider the < case due to complementary probabilities. For example, let $X = 1$ represent receiving a vaccine against COVID-19 and $Y = 1$ represent developing COVID-19. For an effective vaccine, two causal statements are true. First, $P[Y = 1|do[X = 1]] < P[Y = 1]$, which means that receiving the vaccine ($X = 1$) reduces the likelihood of developing the disease ($Y = 1$). Second, $P[Y = 0|do[X = 1]] > P[Y = 0]$, which means that receiving the vaccine ($X = 1$) increases the likelihood of not developing the disease ($Y = 0$). One statement cannot be true without the other, and the redundancy is resolved by picking the latter.

[4] A variable X may be a necessary cause of Y, a sufficient cause of Y, a necessary-and-sufficient cause of Y, or neither a necessary-nor-sufficient cause of Y (also known as a contributory cause). I do not explain the difference in this article because it is not required for the discussion that follows.

[5] The causal structure of a system can also be encoded in a directed cyclical graph; however, cyclicality comes at a heavy cost: The joint probability cannot be factorized as a product of conditional probabilities between ancestors and descendants only. For this reason, directed acyclic graphs are strongly preferred.

EXHIBIT 1

Causal Graph of a Confounder (*Z*)



characterize the data-generating process, it conveys topological information essential to estimate causal effects. Causal graphs declare the variables involved in a system, which variables influence each other, and the direction of causality (Pearl 2009, p. 12). Causal graphs help visualize do-operations as the action of removing all arrows pointing toward *X* in the causal graph so that the full effect on *Y* can be attributed to setting $X = x$. This is the meaning of the *ceteris paribus* assumption, which is of critical importance to economists.

The causal graph in Exhibit 1 tells us that *Z* causes *X* and *Z* causes *Y*. In the language of causal inference, *Z* is a confounder because this variable introduces an association between *X* and *Y*, even though there is no arrow between *X* and *Y*. For this reason, this type of association is denoted noncausal. Following with the previous example, weather (*Z*) influences ice cream sales (*X*) and the number of swimmers, hence drownings (*Y*). The intervention that sets ice cream sales removes arrow (1) because it gives full control of *X* to the researcher (*X* is no longer a function of *Z*), while keeping all other things equal (literally, *ceteris paribus*). And because *X* does not cause *Y*, setting $X = x$ (e.g., banning the sale of ice cream, $X = 0$) has no effect on the probability of $Y = y$. As shown later, noncausal association can occur for a variety of additional reasons that do not involve confounders.

Five conclusions can be derived from the exposition. First, causality is an extra-statistical (in the sense of beyond observational) concept, connected to mechanisms and interventions, and distinct from the concept of association. As a consequence, researchers cannot describe causal systems with the associational language of conditional probabilities. Failure to use the do-operator has led to confusion between associational and causal statements in econometrics and elsewhere. Second, association does not imply causation; however, causation does imply association, as evidenced by an intervention (do-operation).[6] Third, unlike association, causality is directional, as represented by the arrows of the causal graph. The statement "*X* causes *Y*" implies that $P[Y = y \mid do\ [X = x]] > P[Y = y]$ but not that $P[X = x \mid do\ [Y = y]] > P[X = x]$. Fourth, unlike association, causality is sequential. "*X* causes *Y*" implies that the value of *X* is set first, and only after that, *Y* adapts. Fifth, the *ceteris paribus* assumption simulates an intervention (do-operation), whose implications can only be understood with knowledge of the causal graph. The causal graph shows what other things are kept equal by the intervention.

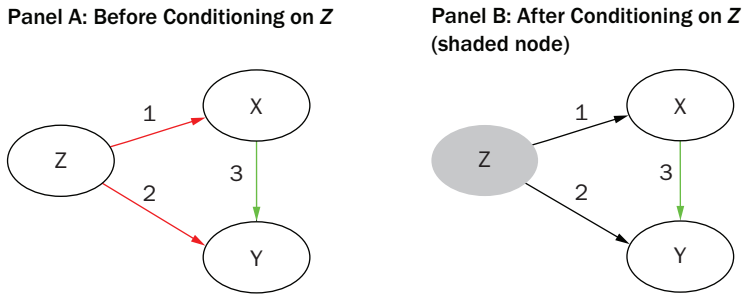
PATHS

In a graph with three variables {*X*, *Y*, *Z*}, a variable *Z* is a confounder with respect to *X* and *Y* when the causal relationships include a structure $X \leftarrow Z \rightarrow Y$. A variable *Z* is a collider with respect to *X* and *Y* when the causal relationships are reversed

[6] Here, I am referring to direct causes (a single link in the causal graph). There are causal structures in which one cause may cancel another, resulting in total causation without association.

EXHIBIT 2

Example of a Causal Graph That Satisfies the Backdoor Criterion



(i.e., $X \rightarrow Z \leftarrow Y$). A variable Z is a mediator with respect to X and Y when the causal relationships include a structure $X \rightarrow Z \rightarrow Y$.

A path is a sequence of arrows and nodes that connect two variables X and Y, regardless of the direction of causation. A directed path is a path where all arrows point in the same direction. In a directed path that starts in X and ends in Z, X is an ancestor of Z, and Z is a descendant of X. A path between X and Y is blocked if either: (1) the path traverses a collider and the researcher has not conditioned on that collider or its descendants, or (2) the researcher conditions on a variable in the path between X and Y, where the conditioned variable is not a collider. Association flows along any paths between X and Y that are not blocked. Causal association flows along an unblocked directed path that starts in treatment X and ends in outcome Y, denoted the causal path. Association implies causation only if all noncausal paths are blocked. This is the deeper explanation of why association does not imply causation and why causal independence does not imply statistical independence.

BACKDOOR ADJUSTMENT

A backdoor path between X and Y is an unblocked noncausal path that connects those two variables. The term backdoor is inspired by the fact that this kind of paths have an arrow pointing into the treatment (X). For example, Exhibit 2 (Panel A) contains a backdoor path (colored in red, $Y \leftarrow Z \rightarrow X$) and a causal path (colored in green, $X \rightarrow Y$). Backdoor paths can be blocked by conditioning on a set of variables S that satisfies the backdoor criterion. The backdoor criterion is useful when controlling for observable confounders.[7]

A set of variables S satisfies the backdoor criterion with regard to treatment X and outcome Y if the following two conditions are true: (i) conditioning on S blocks all backdoor paths between X and Y, and (ii) S does not contain any descendants of X. Then, S is a sufficient adjustment set, and the causal effect of X on Y can be estimated as

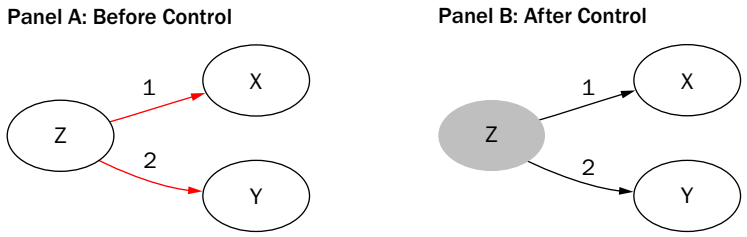
$$P[Y = y|do[X = x]] = \sum_s P[Y = y|X = x, S = s]P[S = s]$$

Intuitively, condition (i) blocks all noncausal paths, whereas condition (ii) keeps open all causal paths. In Exhibit 2, the only sufficient adjustment set S is $\{Z\}$, because conditioning on Z blocks that backdoor path $Y \leftarrow Z \rightarrow X$, and Z is not a descendant of X. The result is that the only remaining association is the one flowing through the

[7] I use here the nomenclature popularized by Pearl (2009); however, this form of adjustment was fully developed by Robins (1986) under the term g-formula.

EXHIBIT 3

Causal Graph with a Confounder *Z*



causal path, thus adjusting the observations in a way that simulates a do-operation on *X*. In general, there can be multiple sufficient adjustment sets that satisfy the backdoor criterion for any given graph.

MONTE CARLO EXPERIMENTS

Factor model specification errors can lead to false positives and false negatives. This section presents three instances of causal structures in which the application of standard econometric procedures leads to mistaking association with causation and ultimately to spurious factor claims. Standard econometric procedures are expected to perform equally poorly on more complex causal structures. The code for these experiments is available at ssrn.com/abstract_id=4205613.

EXPERIMENT 1: FORK

Three variables {*X*, *Y*, *Z*} form a fork when variable *Z* is a direct cause of variable *X* and variable *Y* (see Exhibit 3). Consider a researcher who wishes to model *Y* as a function of *X*. In that case, *Z* is said to be a confounding variable because not controlling for the effect of *Z* on *X* and *Y* will bias the estimation of the effect of *X* on *Y*. Given a probability distribution *P*, the application of Bayesian network factorization on the fork represented by Exhibit 3 yields[8]

$$P[X,Y,Z] = P[Z]P[X|Z]P[Y|Z]$$

which implies a (noncausal) association between *X* and *Y* because

$$P[X,Y] = \sum_Z P[Z]P[X|Z]P[Y|Z] \neq P[X]P[Y]$$

This is an example of noncausal association because *X* and *Y* are associated through the backdoor path $Y \leftarrow Z \rightarrow X$, even though there is no causal path between *X* and *Y*. The effect of conditioning by *Z* is equivalent to simulating a do-operation (an intervention) because it blocks the backdoor path, resulting in the conditional independence of *X* and *Y*,

$$P[X,Y|Z] = \frac{P[X,Y,Z]}{P[Z]} = P[X|Z]P[Y|Z]$$

[8] For an introduction to the calculus of Bayesian network factorization, see Pearl, Glymour, and Jewell (2016, pp. 29–32) and Neal (2020, pp. 20–22).

EXHIBIT 4

False Positive Due to Missing Confounder *Z*

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.254
Model:	OLS	Adj. R-squared:	0.253
Method:	Least Squares	F-statistic:	1697.
Date:	Sat, 21 Jan 2023	Prob(F-statistic):	1.03e-319
Time:	21:36:07	Log-Likelihood:	−8157.8
No. Observations:	5000	AIC:	1.632e+04
Df Residuals:	4998	BIC:	1.633e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0254	0.017	1.450	0.147	−0.009	0.060
X	0.5085	0.012	41.201	0.000	0.484	0.533

Omnibus:	1.899	Durbin-Watson:	2.077
Prob (Omnibus):	0.387	Jarque-Bera (JB):	1.865
Skew:	−0.014	Prob(JB):	0.394
Kurtosis:	2.910	Cond. No.	1.42

NOTES: Results computed using python's statsmodels package. For a description of the terms, consult the package documentation at www.statsmodels.org. The code used in these experiments can be found in López de Prado (2023).

Conditioning by variable *Z* deconfounds $P[X, Y]$ in this causal graph, however, not in other causal graphs. The widespread notion that econometricians should condition (or control) for all variables involved in a phenomenon is misleading. The precise deconfounding variables are determined by do-calculus rules in general and by the backdoor adjustment in this particular example. These conclusions can be verified through the following numerical experiment. First, draw 5,000 observations from the data-generating process characterized by the structural equation model

$$Z_t := \xi_t$$

$$X_t := Z_t + \epsilon_t$$

$$Y_t := Z_t + \zeta_t$$

where $\{\xi_t, \epsilon_t, \zeta_t\}$ are three independent random variables that follow a standard normal distribution. Second, fit on the 5,000 observations the linear equation

$$Y_t = \alpha + \beta X_t + \varepsilon_t$$

Exhibit 4 reports the results of the least-squares estimate. Following the econometric canon, a researcher will conclude that $\hat{\beta}$ is statistically significant. Given the causal content injected by the researcher through the least-squares model specification, a statistically significant $\hat{\beta}$ implies the statement "*X* causes *Y*," not the statement "*X* is associated with *Y*." If the researcher intended to establish association, he should have used an associational model, such as Pearson's correlation coefficient or orthogonal regression. At the same time, Exhibit 3 shows that there is no causal path from *X* to *Y*. The claim of statistical significance is spurious because *Y* is not a function of *X*, as implied by the model's specification. This is the effect of missing a single confounder.

EXHIBIT 5

Deconfounding through the Partial Correlations Method

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.509
Model:	OLS	Adj. R-squared:	0.509
Method:	Least Squares	F-statistic:	2593.
Date:	Sat, 21 Jan 2023	Prob(F-statistic):	0.00
Time:	21:37:35	Log-Likelihood:	−7109.0
No. Observations:	5000	AIC:	1.422e+04
Df Residuals:	4997	BIC:	1.424e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0105	0.014	0.740	0.460	−0.017	0.038
X	−0.0100	0.014	−0.699	0.484	−0.038	0.018
Z	1.0291	0.020	51.036	0.000	0.990	1.069

Omnibus:	0.284	Durbin-Watson:	2.064
Prob (Omnibus):	0.868	Jarque-Bera (JB):	0.290
Skew:	−0.018	Prob(JB):	0.865
Kurtosis:	2.993	Cond. No.	2.65

It is possible to remove the confounder-induced bias by adding Z as a regressor (the partial correlations method)

$$Y_t = \alpha + \beta X_t + \gamma Z_t + \varepsilon_t$$

Exhibit 5 reports the result of this adjustment. With the correct model specification, the researcher will conclude that X does not cause Y.

EXPERIMENT 2: IMMORALITY

Three variables $\{X, Y, Z\}$ form an immorality when variable Z is directly caused by variable X and variable Y (see Exhibit 6). Consider a researcher who wishes to model Y as a function of X. In that case, Z is said to be a collider variable.

Colliders should be particularly concerning to econometricians because controlling for the effect of Z on X and Y biases the estimation of the effect of X on Y. Given a probability distribution P, the application of Bayesian network factorization on the immorality represented by Exhibit 6 yields:

$$P[X, Y, Z] = P[X]P[Y]P[Z|X, Y]$$

There is no association between X and Y because

$$P[X, Y] = \sum_Z P[X]P[Y]P[Z|X, Y] = P[X]P[Y]\sum_Z P[Z|X, Y] = P[X]P[Y]$$

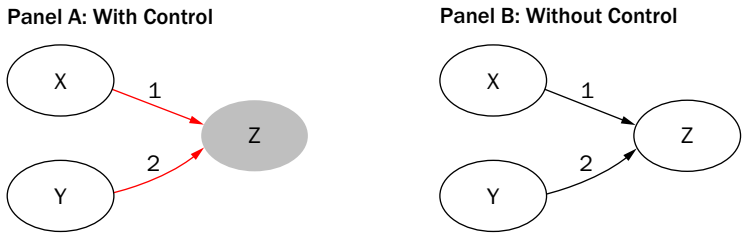
However, conditioning by Z opens the backdoor path between X and Y that Z was blocking ($Y \rightarrow Z \leftarrow X$). The following analytical example illustrates this fact. Consider the data-generating process

$$X_t := \epsilon_t$$

$$Y_t := \zeta_t$$

EXHIBIT 6

Causal Graph with a Collider *Z*



$$Z_t := X_t + Y_t + \xi_t$$

where $\{\xi_t, \epsilon_t, \zeta_t\}$ are three independent random variables that follow a standard normal distribution. Then, the covariance between *X* and *Y* is

$$Cov[X,Y] = E[(X - E[X])(Y - E[Y])] = E[XY] = E[X]E[Y] = 0$$

The problem is, a researcher who (wrongly) conditions on *Z* will find a negative covariance between *X* and *Y*, even though there is no causal path between *X* and *Y* because

$$Cov[X,Y|Z] = -\frac{1}{3}$$

Compare the causal graph in Exhibit 6 with the causal graph in Exhibit 3. Exhibit 3 has a structure $X \leftarrow Z \rightarrow Y$, where not controlling for confounder *Z* results in undercontrolling. The direction of causality is reversed in Exhibit 6, transforming the confounder into a collider. In the structure $X \rightarrow Z \leftarrow Y$, controlling for *Z* results in overcontrolling. This is an instance of Berkson's fallacy, whereby a noncausal association is observed between two independent variables as a result of conditioning on a collider (Pearl 2009, p. 17).

This finding is problematic for econometricians because the direction of causality cannot always be solely determined by observational studies (Peters, Janzing, and Scholkopf 2017, pp. 44–45), and solving the confounder-collider conundrum often requires the injection of extra-statistical (beyond observational) information. Causal graphs inject the required extra-statistical information by making explicit assumptions that complement the information contributed by observations.[9] Accordingly, the statistical and econometric mantra "data speaks for itself" is in fact misleading, because two econometricians who rely solely on observational evidence can consistently reach contradicting conclusions from the analysis of the same data. With a careful selection of colliders, a researcher can present evidence in support of any spurious investment factor. The correct causal treatment of a collider is to indicate its presence and explain why researchers should not control for it. A key takeaway is that researchers must declare and justify the hypothesized causal graph that supports their chosen model specification or else submit to the healthy scepticism of their peers.

We can verify the above conclusions with the following numerical experiment. First, draw 5,000 observations from the preceding data-generating process. Second, fit on the 5,000 observations the linear equation

$$Y_t = \alpha + \beta X_t + \gamma Z_t + \varepsilon_t$$

[9] In the absence of an interventional study or a natural experiment, the statement *X* causes *Y* is an assumption, which may be consistent with, however not proved by, observational evidence.

EXHIBIT 7

False Positive Due to Adding Collider *Z*

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.491
Model:	OLS	Adj. R-squared:	0.491
Method:	Least Squares	F-statistic:	2410.
Date:	Sat, 21 Jan 2023	Prob(F-statistic):	0.00
Time:	21:38:49	Log-Likelihood:	−5380.5
No. Observations:	5000	AIC:	1.077e+04
Df Residuals:	4997	BIC:	1.079e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	−0.0077	0.010	−0.768	0.443	−0.027	0.012
X	−0.5017	0.012	−40.421	0.000	−0.526	−0.477
Z	0.4959	0.007	69.420	0.000	0.482	0.510

Omnibus:	1.629	Durbin-Watson:	2.058
Prob (Omnibus):	0.443	Jarque-Bera (JB):	1.614
Skew:	0.010	Prob(JB):	0.446
Kurtosis:	2.914	Cond. No.	2.46

Exhibit 7 reports the results of the least-squares estimate. Following the econometric canon, a researcher will conclude that $\hat{\beta}$ is statistically significant. This claim of statistical significance is spurious because *Y* is not a function of *X*, as implied by the model's specification. This is the effect of controlling for a collider.

We can remove the bias induced by collider *Z* by excluding that variable from the model's specification

$$Y_t = \alpha + \beta X_t + \varepsilon_t$$

Exhibit 8 reports the results of this adjustment. Note that the misspecified model delivered higher explanatory power; hence, specification-searching would have misled the researcher into a false positive. With the correct model specification, the researcher will conclude that *X* does not cause *Y*.

EXPERIMENT 3: CHAIN

Three variables $\{X, Y, Z\}$ form a chain when variable *Z* mediates the causal flow from variable *X* to variable *Y* (see Exhibit 9). Consider a researcher who wishes to model *Y* as a function of *X*. In that case, *Z* is said to be a mediator variable.

Given a probability distribution *P*, the application of Bayesian network factorization on the chain represented by Exhibit 9 yields

$$P[X, Y, Z] = P[X]P[Z|X]P[Y|Z]$$

which implies an association between *X* and *Y* because

$$P[X, Y] = \sum_Z P[X]P[Z|X]P[Y|Z] \neq P[X]P[Y]$$

EXHIBIT 8

Debiasing by Removing Collider *Z*

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.000
Model:	OLS	Adj. R-squared:	-0.000
Method:	Least Squares	F-statistic:	0.2664
Date:	Sat, 21 Jan 2023	Prob(F-statistic):	0.606
Time:	21:38:45	Log-Likelihood:	-7068.5
No. Observations:	5000	AIC:	1.414e+04
Df Residuals:	4998	BIC:	1.415e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0049	0.014	-0.350	0.726	-0.033	0.023
X	0.0072	0.014	0.516	0.606	-0.020	0.035

Omnibus:	4.168	Durbin-Watson:	2.023
Prob (Omnibus):	0.124	Jarque-Bera (JB):	4.157
Skew:	0.056	Prob(JB):	0.125
Kurtosis:	3.086	Cond. No.	1.02

EXHIBIT 9

Causal Graph with Mediator *Z*

There is no backdoor path in Exhibit 9. This is an example of association with causation because *X* and *Y* are associated only through the causal path mediated by *Z*. Like in the case of a fork, the effect of conditioning by *Z* is equivalent to simulating a do-operation (an intervention), resulting in the conditional independence of *X* and *Y*

$$P[X,Y|Z] = \frac{P[X,Y,Z]}{P[Z]} = \frac{P[X]P[Z|X]P[Y|Z]}{P[Z]} = \frac{P[X,Z]}{P[Z]}P[Y|Z] = P[X|Z]P[Y|Z]$$

The problem with conditioning on a mediator is that it may disrupt the very causal association that the researcher wants to estimate (an instance of overcontrolling), leading to a false negative. To make matters more complex, conditioning on a mediator can also lead to a false positive. This statement can be verified through the following numerical experiment. First, draw 5,000 observations from the data-generating process characterized by the structural equation model

$$X_t := \epsilon_t$$

$$W_t := \eta_t$$

$$Z_t := X_t + W_t + \xi_t$$

$$Y_t := Z_t + W_t + \zeta_t$$

where $\{\xi_t, \epsilon_t, \zeta_t, \eta_t\}$ are four independent random variables that follow a standard Normal distribution. Exhibit 10 displays the relevant causal graph. Second, fit on the 5,000 observations the linear equation

$$Y_t = \alpha + \beta X_t + \gamma Z_t + \varepsilon_t$$

Exhibit 11 reports the results of the least-squares estimate. Although it is true that *X* causes *Y* (through *Z*), this result is still a false positive because the reported association did not flow through the causal path $X \to Z \to Y$. The reason is *Z* also operates as a collider to *X* and *W*, and controlling for *Z* has opened the backdoor path $X \to Z \leftarrow W \to Y$. This is the reason $\hat{\beta} \ll 0$, despite of all effects being positive. This phenomenon is known as the *mediation fallacy*, which involves conditioning on the mediator when the mediator and the outcome are confounded (Pearl and MacKenzie 2018, p. 315). This experiment also illustrates Simpson's paradox, which occurs when an association is observed in several groups of data, but it disappears or reverses when the groups are combined (Pearl, Glymour, and Jewell 2016, pp. 1–6).

EXHIBIT 10

A Confounded Mediator (*Z*)

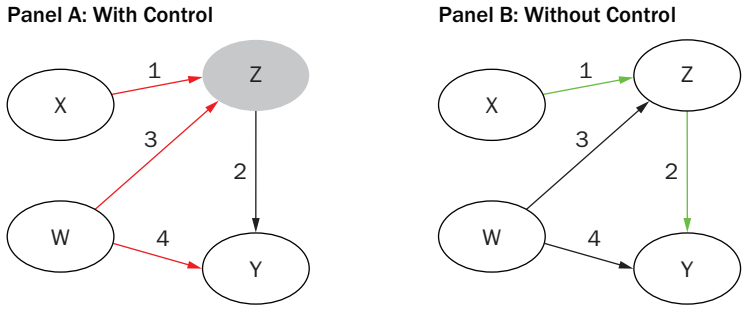


EXHIBIT 11

False Positive Due to Adding a Confounded Mediator *Z*

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.786
Model:	OLS	Adj. R-squared:	0.786
Method:	Least Squares	F-statistic:	9160.
Date:	Sat, 21 Jan 2023	Prob(F-statistic):	0.00
Time:	21:40:59	Log-Likelihood:	−8114.0
No. Observations:	5000	AIC:	1.623e+04
Df Residuals:	4997	BIC:	1.625e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0007	0.017	0.040	0.968	−0.033	0.035
X	−0.5388	0.021	−25.130	0.000	−0.581	−0.497
Z	1.5075	0.012	122.161	0.000	1.483	1.532

Omnibus:	2.003	Durbin-Watson:	2.048
Prob (Omnibus):	0.367	Jarque-Bera (JB):	1.952
Skew:	−0.045	Prob(JB):	0.377
Kurtosis:	3.036	Cond. No.	2.46

EXHIBIT 12

Deconfounding by Removing the Confounded Mediator

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.146
Model:	OLS	Adj. R-squared:	0.145
Method:	Least Squares	F-statistic:	852.2
Date:	Sat, 21 Jan 2023	Prob(F-statistic):	4.03e-173
Time:	21:44:28	Log-Likelihood:	−11571.
No. Observations:	5000	AIC:	2.315e+04
Df Residuals:	4998	BIC:	2.316e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0092	0.035	0.265	0.791	−0.059	0.077
X	1.0083	0.035	29.192	0.000	0.941	1.076

Omnibus:	2.574	Durbin-Watson:	2.024
Prob (Omnibus):	0.276	Jarque-Bera (JB):	2.605
Skew:	−0.026	Prob(JB):	0.272
Kurtosis:	3.098	Cond. No.	1.02

Following the rules of do-calculus, the effect of *X* on *Y* in this causal graph can be estimated without controls. The reason is the noncausal path through *W* is already blocked by *Z*. Controlling for *W* is not strictly necessary to debias $\hat{\beta}$; however, it can help improve the precision of the estimates. The following model specification produces an unbiased estimate of β:

$$Y_t = \alpha + \beta X_t + \varepsilon_t$$

Exhibit 12 reports the results. Note that the correct model specification has much lower explanatory power: The adjusted R-squared drops from 0.786 to 0.146, and the F-statistic drops from 9,160 to 852.2. A specification-searching researcher would have chosen and reported the wrong model because it has higher explanatory power, resulting in a misspecified model that misattributes risk and performance. With the proper model specification, as informed by the declared causal graph, the researcher correctly concludes that *X* causes *Y* and that $\hat{\beta} \gg 0$.

CONCLUSIONS

Scientific theories should matter to investors for at least three reasons: First, theories are a deterrent against selection bias because they force scientists to justify their modeling choices, thus curtailing efforts to explain random variation. A researcher who engages in *p*-hacking or backtest overfitting may build an *ad hoc* theory that explains an observed random variation. However, other researchers will use the theory to design an experiment in which the original random variation is not observed. Second, causality is a necessary condition for investment efficiency. Causal models allow investors to attribute risk and performance to the variables responsible for a phenomenon. With proper attribution, investors can build a portfolio exposed only to rewarded risks and aim for investment efficiency. In contrast, associational models misattribute risks and performance, thus preventing investors from building efficient portfolios. Third, causal models enable counterfactual reasoning, hence the

stress-testing of investment portfolios in a coherent and forward-looking manner (see Rebonato 2010, Rebonato and Denev 2014, Denev 2015, and Rodríguez-Domínguez 2023). In contrast, associational models cannot answer counterfactual questions, such as what would be the effect of *Y* on a not-yet-observed scenario *X*, thus exposing those relying on associations to black-swan events.

Financial economists' adoption of causal inference methods has the potential to transform investing into a truly scientific discipline. Economists are best positioned to inject, make explicit, and argue the extra-statistical information that complements and enriches the work of statisticians. Machine learning methods can assist economists in identifying the set of variables involved in the data-generating process, from which a causal graph can be postulated, which in turn will support a chosen specification (López de Prado 2022).

The new discipline of causal factor investing will be characterized by the adaptation and adoption of tools from causal discovery and do-calculus to the study of the risk characteristics that are responsible for differences in asset returns. Every year, new alternative datasets become available at an increasing rate, allowing researchers to conduct natural experiments and other types of causal inference that were not possible in the 20th century. Causal factor investing will serve a social purpose beyond the reach of (associational) factor investing, helping asset managers fulfill their fiduciary duties with the transparency and confidence that only the scientific method can deliver. To achieve this noble goal, the dawn of scientific investing, the factor investing community must first wake up from its associational slumber.

ACKNOWLEDGMENTS

I thank Alexander Lipton, Jean-Paul Villain, and Vincent Zoonekynd for numerous comments and contributions. I also benefited from conversations with more ADIA colleagues than I can cite here, as well as David H. Bailey (University of California, San Diego), David Easley (Cornell University), Campbell Harvey (Duke University), Miguel Hernán (Harvard University), John Hull (University of Toronto), Maureen O'Hara (Cornell University), Riccardo Rebonato (EDHEC), Alessio Sancetta (Royal Holloway, University of London), Horst Simon (Berkeley Lab), Sasha Stoikov (Cornell University), and Josef Teichmann (ETH Zurich). This article is an abridged version of the monograph "Causal Factor Investing," available at ssrn.com/abstract_id=4205613.

REFERENCES

Chen, B., and J. Pearl. 2013. "Regression and Causation: A Critical Examination of Six Econometrics Textbooks." *Real-World Economics Review* 65: 2–20.

Denev, A. 2015. *Probabilistic Graphical Models: A New Way of Thinking in Financial Modelling*. London, UK: Risk Books.

Greene, W. *Econometric Analysis*, 7th ed. Hoboken, NJ: Pearson Education, 2012.

Hill, R., W. Griffiths, and G. Lim. *Principles of Econometrics*, 4th ed. New York, NY: John Wiley & Sons, 2011.

Kennedy, P. *A Guide to Econometrics*, 6th ed. Cambridge, MA: MIT Press, 2008.

López de Prado, M. 2022. "Machine Learning for Econometricians: The Readme Manual." *The Journal of Financial Data Science* 4 (3): 10–30.

——. *Causal Factor Investing. Elements in Quantitative Finance*. Cambridge University Press, 2023 (forthcoming). Pre-print available in https://ssrn.com/abstract_id=4205613.

Neal, B. "Introduction to Causal Inference: From a Machine Learning Perspective." Course lecture notes, December 17, 2020, https://www.bradyneal.com/causal-inference-course.

Pearl, J. 1995. "Causal Diagrams for Empirical Research." *Biometrika* 82: 669–710.

——. *Causality: Models, Reasoning and Inference,* 2nd ed. New York, NY: Cambridge University Press, 2009.

Pearl, J., M. Glymour, and N. Jewell. *Causal Inference in Statistics: A Primer*. Hoboken, NJ: John Wiley & Sons, 2016.

Pearl, J., and D. MacKenzie. *The Book of Why*. New York, NY: Basic Books, 2018.

Peters, L., D. Janzing, and B. Scholkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA: MIT Press, 2017.

Rebonato, R. *Coherent Stress Testing*. New York, NY: John Wiley & Sons, 2010.

Rebonato, R., and A. Denev. *Portfolio Management under Stress: A Bayesian-Net Approach to Coherent Asset Allocation*. Cambridge, UK: Cambridge University Press, 2014.

Robins, J. 1986. "A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period: Application to Control of a Healthy Worker Survivor Effect." *Mathematical Modelling* 7 (9–12): 1393–1512.

Rodríguez-Domínguez, A. 2023. "Portfolio Optimization Based On Neural Networks Sensitivities From Asset Dynamics Respect Common Drivers." *Machine Learning with Applications* 11: 100447.

Ruud, P. *An Introduction to Classical Econometric Theory*. Oxford, UK: Oxford University Press, 2000.

Wooldridge, J. "Should Instrumental Variables Be Used as Matching Variables?" Technical report, Michigan State University, 2009, https://www.msu.edu/~ec/faculty/wooldridge/current%20research/treat1r6.pdf.

Disclaimer

The views expressed in this article are the author's and do not necessarily represent the opinions of the organizations with which he is affiliated.